Contents

iGenome 2017: An updated interactive view on ’The Regulatory Genome’

Ralf Zimmer (editor)

Department of Informatics, Ludwig-Maximilians-Universität München,
Amalienstr. 17, 80333 München, Germany

1 Introduction

The current 2017 update on ’The regulatory genome’ provides an in-depth update on various aspects of the regulation and control of the genome with particular emphasis on state-of-the-art high-throughput methods.

The book contains many figures which are accompanied with scripts and data to understand and reproduce the figures in any detail. Where possible the figures are also interactive allowing for different parametrisation and the access of the raw data and evidence underlying the figures.

The first three chapters focus on next generation sequencing methods to measure and predict gene expression, to derive transcription factor binding site, to determine open chromatin regions and, thus, context specific regulation.

Chapter 5 reviews the impact of alternative splicing on the regulation by the genome.

Chapter 12 investigates the role of regulatory networks.

Chapter 13 and 14 describe regulatory functions of long and short non-coding RNA.

The new insights on the impact of genomic 3D interactions on regulation are covered in chapter 15.

Finally, chapter 16 introduces protein translation and protein expression to cover the impact of the regulatory genome on the protein level.

And in chapter 17 an application of gene expression profiling on breast cancer subtype-classification is discussed.

Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2 Gene Expression Prediction

by Nicola Palandt, Katharina Schmid and Markus Gruber

2.1 Project Idea

The effect of histone modifications on gene regulation is shown in many studies, yet detailed information about the role of individual histone modifications, their combinatorial influence and the underlying biological mechanisms, such as how the histones are modified and how they affect the gene expression, are not known in detail in many cases until now (see Dong et al. [1]). There are several studies to analyze the role of histone modifications for transcription, such as Koch et al. [2] and Barski et al. [3].

A bioinformatical approach to study the relationship between both is the attempt to predict gene expression using histone modifications. There are two different ways, either predicting the gene expression level directly using regression methods or predicting classes of highly and lowly expressed genes using classification. From both similar insights can be drawn.

The predictability of gene expression with histone modifications shows how tight the relationship between both is. A good performance would show a tight connection between histone modifications and gene expression, indicating that the histone modifications influence the gene expression level. So biological knowledge can be gained by the machine learning strategy.

Another interesting question which can be addressed by those methods is the relative importance of individual modifications. Are all histone modifications required to predict the gene expression accurately or do they contain partly redundant information and can the right subset of modifications predict the expression as good as all modifications?

At last, differences between cell types, species and RNA types, such as mRNA, microRNA and long non-coding RNA, can be evaluated using the prediction method. It can be investigated how well a model performs on a specific data set, especially, whether the prediction accuracy drops drastically, when training on one data set and predicting another data set, e.g. from a different cell type. If the performance remains good for different data sets, that would indicate general mechanisms of expression regulation by histone modifications, otherwise cell type, species and RNA type specific differences could be observed.

Several methods investigating these aspects have already been published, but they all have a huge drawback. For most of the methods, only one machine learning strategy was implemented and different publications use different cell types and features sets. For example, some authors use only histone modifications, others include features such as DNase hypersensitivity sites or transcription factor binding sites. This makes it very difficult to compare the results. Often, very different results are shown by the papers, with different performance values and different histone modifications which are important for the prediction. But it is hard to say if the differences are due to the different methods or if they dependend on other factors such as the data set or the factor set. This makes it hard to interpret the results correctly in context with the results of the other papers.

The goal of our project is the implementation of different classification and regression methods which predict gene expression using histone modifications as factors. We will evaluate the performance differences of the individual methods and analyze for each method the relative importance of individual histone modifications and the performance on different data sets to detect cell-type specific differences. Testing different data sets and methods will show us, what is responsible for the big published differences: the methods or the different data sets and data processing.

Before implementing the methods, we first investigated the biological background, the published methods, which predict gene expression, and available data sets. This is summarized in the first part of our chapter, afterwards the results of our project are described.

2.2 Biological Background

Histones are proteins in eukariotic cells, which are like spools for the DNA. DNA is wound around histones to save a lot of space. There are five different families of histones: H1, H2A, H2B, H3 and H4 and all have several subfamilies. H2A, H2B, H3 and H4 are the four core canonical histones, that build the histoneoctamer in the nucleosome, the basic unit of DNA packaging. This octamer usually consists of two H2A-H2B dimers and two H3-H4 dimers, see figure 1.

In some cases there are special variants of the canonical histones, which substitute them in the octamer. These variants can play an important role for gene regulation. One example is the H2A.Z variant which is found in flowers. Nucleosomes that contain the H2A.Z variant wrap the DNA more tightly and a transcription at this part of the DNA is not possible. H2A.Z is thermosensitive. With rising temperature, the nucleosome with H2A.Z gets evicted and the gene is free for transcription. Flowers use this mechanism to produce proteins, which are necessary for flowering [4].

However, this is not the only way how histones are able to play a role in the regulation of gene expression. Jenuwein et al [5]. propose the "histone code" hypothesis which states that different histone amino-terminal modifications and their combinations are able to up- or downregulate gene expression [5].

Each histone can be modified by one or more of the following four different types of histone modifications:

- Methylation
- Acetylation
- Phosphorylation
- Ubiquitination

For some of the histone modifications the effect on gene expression is already known. Acetylations such as H3K9ac and H3K14ac have an activating influence [2], because acetylation tends to define how tight the DNA is wound. Usually, DNA with negatively charged phosphor groups and positively charged histones are bound tight, but due to acetylation, the positive charge of histones is removed and DNA is wound less tight. Therefore polymerase is able to transcribe the genes.

It is difficult to be sure about the influence of histone modifications, because it is also possible that the modification is an effect of the gene expression or the effect of an other modification. If so, it appears as directly regulating the gene expression, but in reality it is not causal for the expression. This problem makes it very difficult to decode the histone code.

2.3 Different Methods

Various methods have been published in the last years which predict gene expression using histone modifications and other chromatin factors. In the following, we explain a few of them in more detail to get a better understanding of the general approach and critical aspects of our project.

To better distinguish the papers, we choose to categorize them according to their mainly used machine learning technique, but there are of course other important differences between them such as the data sets and features they used. As mentioned before, the main goal

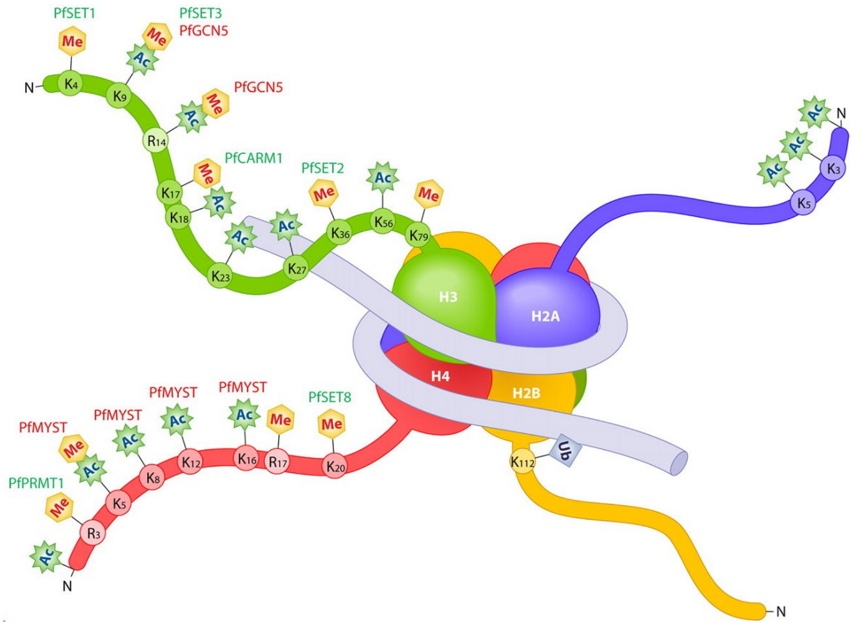


Figure 1 Nucleosom with the 4 histone families H2A, H2B, H3 and H4 and the different histone modifications [6].

of implementing such methods is not to simply predict gene expression, but to gain more insights about the biological background. Therefore it is also interesting what kind of results the methods show, e.g. the prediction on different feature subsets or the prediction using data sets from different cell types.

Karlic et al. were one of the first to publish a machine learning approach for the prediction of gene expression values using simple linear regression models [7]. The features for the regression were the average values of histone modification coverage per modification and gene. Over time, also more sophisticated methods were used, such as Support Vector Machine Regression [8] and Multivariate Adaptive Regression Splines [9]. Nevertheless, the simple approach of linear regression was adapted by different others, such as [9], [10], [8] and [11]. Dong et al. even wrote that the linear regression showed similar performance as Random Forrest Regression and Multivariate Adaptive Regression Splines [11].

We want to concentrate on classification methods in the first part of our project, as the newest published methods to predict gene expression are classification methods. The major drawback when using classification is that a cut-off value needs to be chosen to define highly and lowly expressed genes or even more classes. Natarjan et al. e.g. separated the genes in five classes [12], a more detailed description of his approach can be found in the following section. The specification of such a threshold to classify the genes is to some extent a subjective decision, which may be hard to justify sensible.

Comparing the different classification strategies, the Support Vector Machines of Cheng et al. are interesting, as they used a new strategy for feature selection by binning the regions around the transcription start site and transcription termination site [8]. Dong et al. further developed this binning approach and combined a Random Forrest classifier with different regression methods to exceed the performance of gene expression [11]. In contrast, Natarajan et al. used a completely different set of features, DNase hypersensitivity sites, and Sparse Logistic Regression Classifier [12]. The most recent classification method to predict gene expression with histone modifications and machine learning method in general, is from 2016 and uses Deep Learning [13].

In the following we will present five publications in more detail, Karlic et al. with a simple linear regression model [7], and four classification strategies, [12], [8], [11] and [13].

2.3.1 Karlic et al, 2010 (Linear Regression)

As mentioned before, Karlic et al. used a simple linear regression model to predict gene expression [7]. Their data set contained 38 histone modifications and one histone variant in human CD4+ T-cells, measured by ChIP-seq, as well as gene expression values, measured by microarrys. For feature generation, they counted the number of tags for each histone modification in the promoter of each gene, defining a window of 4,0001 bp around the TSS as the promoter region.

The main focus of the publication was to investigate how redundant the information content of histone modifications is. So beneath the full model, trained as a linear regression of all features, also models with only a subset of features were trained. They trained One-modification models, where a linear regression with one factor was built. Doing this with each of the factors, they got 41 One-modification models in total. Equivalently, they built 820 Two-modification models with every possible combination of two histone modifications and 10,660 Three-modifications models with every possible combination of three modifications. Models with combinations of more than three modifications were not doable, because the number of models becomes too large if all combinatorial options would be considered.

All models were evaluated calculating the correlation between predicted expression values and measured expression values. The full-model performed best with a correlation of 0.77, but the best One-modification model got a score of 0.72, the best Two-modification model of 0.74 and the best Three-modification model even 0.75. So, when taking the right combination of three histone modifications nearly the same predictive power can be established as taking all histone modifications when using this approach of linear regression.

The occurrence of all histone modifications in Three-modification models which got a correlation of at least 95% the correlation of the full model were counted to gain the relative importance of individual modifications. 4 modifications were very significantly overrepresented in these models, H4K20me1 in 58% of all models, H3K27ac in 54%, H3K79me1 in 43% and H2BK5ac in 41%. So according to Karlic et al., only few histone modifications are necessary for the prediction of gene expression, most of the others contain only redundant information.

2.3.2 Natarajan et al, 2012 (Sparse Logistic Regression Classifier)

Natarajan et al. used Sparse Logistic Regression Classifier to classify genes according to their gene expression [12]. The feature selection of the method is interesting, as it shows that not only histone modifications, but a variety of other features can be used for the machine learning methods. Natarajan et al. chose a dataset containing DNase hypersensitivity sites (DHS) in 19 human cell lines with the corresponding gene expression levels measured by microarries.

Since they used a classification method, the genes needed to be assigned to different classes before training the model. In contrast to the other groups, who separated the genes in only two classes of highly and lowly expressed genes, choosing the mean as a cut-off value, Natarajan et al. created a more complex definition. They separated the genes in five classes, after normalization of the values by Z scores:

- **UR:** top 200 genes (after sorting according to the Z score) that have a high expression level (large Z score) in one cell lines and a expression close to the mean in all other cell lines
- **DR:** last 200 genes (after sorting according to the Z score) that have a low expression level (small Z score) in one cell lines and a expression close to the mean in all other cell lines
- **constitutive regulated:** all genes that are not in the UR and DR genes and have a absolute Z score smaller than 1.7 in all cell lines.
- **UR-Other:** UR gene expression and DHS in all other cell lines where they are not specially up-regulated
- **DR-Other:** DR gene expression and DHS in all other cell lines where they are not specially down-regulated

To get a feature matrix out of the DHS, they were classified in intergenic DHS, TSS DHS and gene body DHS based on their genomic position. 789 Position Weighted Matrices (PWMs) of transcription factor binding motives were downloaded from different sources and applied to each DHS in a sliding window approach to get transcription factor binding scores (TFBS) for the specific TF for each position in the DHS. Then in another sliding window approach the maximal sliding window score for each TFBS was calculated to associate each DHS with one value per TF. Each DHS was associated with the nearest gene, because of lack of more information about the long-range interactions. The DHS were further categorized into distal DHS sites, containing Intergenic and Gene Body DHS, and TSS DHS.

Then three different features sets were created to evaluate the influence of promoter and distal regulatory elements on the gene expression. The first set, called promoter, contained only the score of the TSS DHS for each gene. The second set, called Closest Gene DHS, consisted of one maximal score of all DHS, which were associated with the gene, distal and proximal DHS together. The third set, called Split DHS, assigned two values per gene and TF, the score of the TSS DHS and the maximal score of a distal element.

The model in the end was trained to distinguish two classes of genes, e.g. UR vs. DR, UR vs. UR-Other or UR vs. constitutive. Comparing the different features sets with AUROC scores, the promoter feature set had the worst performance with median values for each classification task between 0.5 and 0.6, close to a random predictor. The performance increased considerably when using also information from the distal DHS with the Closest Gene DHS and Split DHS. The Split DHS set showed the best performance of all methods with a median AUROC of 0.73. Adding the CG content of each DHS as an additional feature improved the prediction in some cases slightly, especially when using only Promoter DHS.

The predicted significant TFBS were then validated using TF binding sites, measured with ChIP-seq, and DHS footprints. It could be shown for several transcription factors that if a significant TFBS score was predicted at a DHS for one cell type they really bind at this DHS in this cell type and do not bind at this position in other cell types.

In general, Natarajan et al. showed that a variety of features can be used to predict gene expression beneath histone modifications, such as DHS and CG content, and that the inclusion of distal regulatory elements improves the prediction significantly.

2.3.3 Cheng et al, 2011 (Support Vector Machine)

To get position specific information as well and not only one score per gene, Cheng et al. introduced a binning approach to calculate average histone modification signals [8]. They used data from the modENCODE project of the model organism C.elegans and a variety of different features. Beneath 10 histone modifications and the occupation pattern of histone H3, they also chose information about Polymerase II binding and binding sites of 8 X-inactivation factors and 5 transcription factors. For expression values, they compared protein coding transcript expression with microRNA expression, both measured by RNA-seq.

To bin the regions around the TSS and the TTS, they created 80 bins of length 100 bp around each, 40 before the TSS and 40 behind, for the TTS equivalently. In total, they got 160 bins and calculated the average signal of each feature in each bin. They were most interested in how good the information of an individual bin was able to predict the gene expression. They trained their machine learning strategies, which were Support Vector Machine classification and regression as well as linear regression, on each bin individual and compared the performance of the bins.

The features in the bin around the TSS could predict the gene expression best. The farther away bins were located before the TSS in the intergenic region, the worse the AUROC scores got. Around the TTS a similar pattern with a peak at the TTS could be observed.

Next, they evaluated the prediction using different subsets of their features. After separating the total feature subsets in groups of using only the H3 occupation, all histone modifications, all X- inactivation factor binding sites and the polymerase II binding profile, using only histone modifications had the best performance with an AUROC nearly as good as when using all features, both with AUROC values around 0.9. When predicting with just Polymerase II, which is only one feature, still an AUROC value up to 0.8 could be achieved, while the H3 features performed worst with values up to 0.6, close to a random prediction.

Training the SVM on EEMB cells, embryonic stem cells of C.elegans, and predicting gene expression of cells from other development stages from L1 up to adult cells, all predictions obtained good validation scores. The prediction on EEMB cells works of course best with an AUROC of 0.905 for bin 43, but for all other cells the values were between 0.82 to 0.80. So the machine learning model is robust against predicting cell types from different developmental stages.

2.3.4 Dong et al, 2012 (Random Forest)

Dong et al. refined the algorithm of Cheng et al., using a similar binning approach [11]. Instead of the C. elegans data they used ENCODE data from 7 human cell lines. They also restricted their feature set to 11 different histone modifications and one histone variant, together with DNase I hypersensitivity sites. For gene expression data, they were the first to compare different measurement strategies, i.e. CAGE, RNA-PET and RNA-SEQ, to see if the measurement strategy would influence the result. The binning strategy is similar to Cheng et al., they used also 80 bins around the TSS and 80 bins around the TTS, each of length 50bp, instead of 100bp, and with an additional bin in the gene body, the part, which was not covered by the other bins. Because of that only transcripts longer than 4,001 bp were analyzed, so that the bin in the gene body could be calculated. Instead of training on each bin and comparing the performances, like Cheng et al. did, Dong et al. used a best binning strategy to preselect the best bin. Therefore they calculated the correlation between the bins and the expression value and chose the bin with the highest absolute correlation for the classification. Because the chromatin features were log2-transformed, an optimized pseudo count was added to avoid problems with the logarithm of 0. The pseudo count was optimized, so that it maximizes the correlation between the logarithmized features signal plus the pseudo count and the logarithm of the measured expression values. This preselection was done on one third of the genes and this part of the data was then not further used for creating the prediction model.

The idea of Dong et al. was to combine the classification and the regression approach to improve the performance of the regression. The problem when predicting gene expression levels is that many gene are not expressed at all, i.e. many genes in the training set have a value of 0. This can complicate a good fit of the regression curves. So Dong et al. first tried to identify genes without expression, by classification of the genes in two classes, expression and no expression. Afterwards, only for the genes with a predicted expression value the regression is performed to estimate the exact level of expression.

Interestingly, when evaluating which histone modification is most important for the prediction, the regression methods have different most important histones compared to the classification methods. This indicates that different histones are responsible to activate and deactivate the expression, which is analyzed by the classification, and to determine the actual level of expression, which is analyzed by the regression.

Dong et al showed that results differ greatly between data sets. The method worked better when using CAGE to estimate the expression values compared to RNA-PET and RNA-seq. Because CAGE measures only the 5'end of the transcript, with this method only the transcription initiation can be measured, not the elongation or the termination. So, it seems like the initiation is correlated best with the histone modifications.

Expression values of different RNA types and in different locations show also different performance values for prediction. For example, PolyA+ RNA is better predictable than PolyA- RNA and the cytosolic PolyA+ RNA better than the nucleus PolyA+ RNA.

2.3.5 Singh et al, 2016 (Deep Learning)

One of the newest methods is DeepChrome, a deep learning algorithm by Singh et al., that uses convolutional neural networks (CNNs) to automatically learn combinatoric interactions of histone modification marks to predict the up- or downregulation of gene expression [13]. CNNs were created similar to the biological process of neurons in the visual cortex. First, CNNs were used to classify pictures, Singh et al. used the signal of histone modifications instead of pictures and tried to find hidden patterns that normal machine learning approaches were not able to find.

Equal to the prediction methods of Cheng et al. [8] and Dong et al. [11], Singh et al. divided the base pairs into bins of 100 bp. However, DeepChrome took only the 10,000 base pairs around the transcription start site (TSS) into account. The reason is that Cheng et al. showed that the signals around the TSS are the most informative ones. One big difference to Cheng et al. and Dong et al. is that the deep learning algorithm is able to handle all bins together, which has the advantage to capture neighboring range and long range interactions. Besides DeepChrome is able to extract automatically important features. Singh et al. used data of five core histone modifications for 56 different cell types from the Epigenetic Roadmap database [14]. After the binning as explained above, they got a 5x100 matrix where the columns are the bins and the rows the different histone modifications. The DeepChrome method has five different steps:

1. Convolution:

We have N_{out} different filters of length k. This means we always look at k different bins at the same time. This builds a sliding window operation across all bin positions. With the following formula we calculate the feature map Z in the convolution step:

$$Z_{p,i} = B_i + \sum_{j=1}^{N_f} \sum_{r=1}^{k} W_{i,j,r} X_{p+r-1,j}$$

N_f is the number of features, which is the number of histone modifications for the first layer, X is the matrix of bins and W is the filter matrix.

2. Rectification

This is the activation function where we set all negative values of the feature map Z to 0. The advantage of rectified linear units (ReLU) is that it induces sparsity in the hidden units and deep networks can be trained efficiently with ReLU without pre-training [15].

3. Pooling

In this step the algorithm learns translational invariant features. Max pooling selects for each row the maximum of m values. This step reduces the complexity of the matrix and the number of features and prevents overfitting.

4. Dropout

In this layer we randomly set a value with a 50% chance to 0. It is a form of regularization and can be used instead of common alternatives such as bagging or averaging. The regularization is important to prevent overfitting. It is always the risk during training that the network learns the input data too well, so it can predict the input with a very high performance but looses its generality to predict other data as well. With setting randomly some part of the input data to 0, or rather some of the nodes in the network, which represent the input data, the input data is changed each training step and not the totally same data is observed each time. So it should pervent to learn the input data by heart and instead keep the network more general.

5. Classical feed-forward neural network layers

In this step the output of all earlier steps is used as input for a multilayer perceptron classifier to learn a classification function, which maps on the gene expression levels. This fully connected multilayer perceptron network has multiple alternating linear and non-linear layers. The input of each layer is learned to map to a hidden feature space. The output layer learns the mapping from this hidden feature space to the classes +1 and -1. -1 means the features set of histone modifications depresses gen expression and +1 means it activates gen expression.

Results

Singh et al. evaluated their new classification method against the existing methods of Cheng et al. and Dong et al. They compared the performance of Deep Chrome with the performance of the Support Vector Machine, once with average binning and once with best binning, and with the performance of the Random Forest method. The results show that in their approach the deep learning method is the best with an maximal AUC score of 0.92, while the SVM has an maximal AUC of 0.87 with the best bin method and 0.79 with the average bin method and the Random Forest algorithm has only a maximal performance of 0.71.

We have to be careful with these values because of various factors. It is possible that the performance is worse than in the original papers because Singh et al. used less histone modifications and only the transcription start site. These are all influences that show that it is extremely difficult to have a good comparison of the methods. We implemented all three classification methods and tried to get a better comparison.

2.4 Data sets

There are several different databases with many possible datasets that we could use for our project. Out of the big variety of different data we have to choose which cell lines and species we want to use. The data is stored in different formats from raw data files to files where a part of the preprocessing was already done. We also have to select which kind of features we are using. In the literature histone modifications, histone variants and DHS are used as well as transcription factor binding sites, DNA methylation and several other possible features. We decided for our project to concentrate only on histone modifications as features and analyze those in more detail. We still expect good performance results, when using only histone modifications. Different publications claimed that histone modifications alone are enough to predict gene expression well and additional features, such as transcription factor binding, bring mostly only redundant information without further improving the results [10], [8].

We used the already processed tsv files with the gene expression per gene in FPKM and bigWig files, which contain the signals at every base for a histone modification. For the cell lines we decided to choose two immortalized cell lines (K562, SK-N-SH), two tissues (Gastrocnemius medialis (in the following also referred to as Gastro), Thyroid gland) and two primary cell lines (Endothelial cell of umbilical vein (in the following also referred to as Endo), Keratinocyte). K562 is an immortalized leukemia cell line and SK-N-SH an immortalized neuroblastoma cell line. Of the two primary cell lines, the endothelial cells of the umbilical vein are cells from a vein present during fetal development, connecting placenta and fetus, and the keratinocytes are a cell type in the epidermis. The tissue gastrocnemius medialis is a muscle of the lower leg. These datasets are all from the ENCODE project [16]. Some other databases have also datasets with several histone modifications. The Epigenetic Roadmap [14] is the second biggest database for histone modifications. DeepBlue [17] is a meta database that has data from many databases and projects. Currently, different projects of the International Human Epigenome Consortium (IHEC) are running, such as Blueprint of the European Union [18], CEEHRC in Canada [19] and CREST in Japan [20], and the total amount of epigenetic data should increase significantly in the next years.

However at the moment, it is very difficult to find many data sets where the same set of histone modifications is available. So we used for the first steps of the analysis only three of the six data sets, K562, the Endothelial cells and the kerationcytes with six histone modifications (H3K27ac, H3K27me3, H3K4me3, H3K79me2, H3K9me3 and H3K36me3) and the histone variant H2AFZ. In the following chapter, we will reference all 7 features as histone modifications, including the histone variant, due to reasons of readability. Only for the last step when comparing the performance on different data sets, all data sets are used with a reduced set of five histone modifications (H3K27ac,H3K36me3, H3K9me3, H3K4me3, H3K27me3). Additionally, we created a big dataset were we merged the data of K562, the Endothelial cell and Gastrocnemius medialis to see if more general features between cell types could be learned.

To analyze the influence on the parameter fitting and feature selection when using only 5 instead of all 7 histone modifications, we did the analysis on the 3 bigger data sets with 7 histone modifications and additionally for K562 also on the shorter variant, containing only the 5 histone modifications. In the plots they are labeled as K562 (big dataset with 7 histone modifications) and K562_short (small dataset with 5 histone modifications).
We restricted our project to protein coding genes because of various reasons. The most important one is that all the publications used only protein-coding genes as their main data set, even if some tested also other types, such as micro RNA genes (see [8]). Using a data set as similar as possible makes our results better comparable to the papers. It is also beneficial for the runtime to use less genes, especially for methods with non-linear time complexity such as the Support Vector Machines. Doing a short analysis on the total RNA data set, we also observed that the performance of the protein-coding genes alone is much better. That could have various reasons. There could be different regulatory histone patterns for the different gene types, but also that our method of binning works not so good, if the genes get too short, as could be the case for the microRNA genes, because then bins behind the TSS and before the TTS overlap.

2.5 Project Implementation

We reimplemented some of the ideas for predicting gene classes using classification and predicting gene expression using regression, concentrating mainly on the publications of [8], [11] and [13]. Applying all methods to the same data sets and features, we tried to make the methods comparable and see whether differences in the performance of the methods, the relative importance of individual histone modifications and other interesting aspects could be shown. Another important result is to see if we can reproduce some of the published plots, using our methods and data. Only if the published results can be reproduced, in particular with other data sets and methods, the results can be verified.

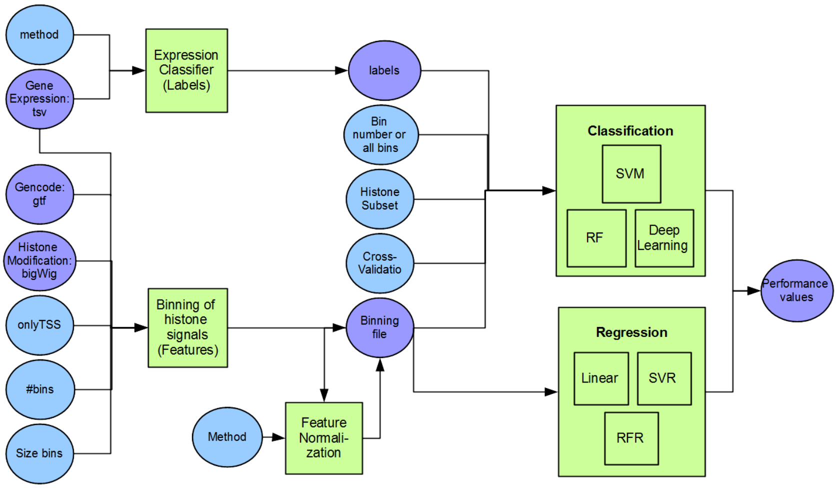


Figure 2 Core elements of our project pipeline, visualized as a Petri net.

The core of our project pipeline contains five modules, as shown in figure 2. First, the different input files need to be processed. The bigwig files containing the histone modification signal along the genome are converted into features according to the method of Cheng et al. [8] in the module **Binning of histone signals (Features)**. Therefore, the expression file is required to get the measured gene IDs and the GENCODE file to get the position of these genes on the genome. For the classification task, labels need to be calculated using the gene expression values, according to different methods, using the module **Expression Classifier (Labels)**. Before running the actual prediction, it is beneficial to normalize the features with the module **Feature Normalization**, again different methods are possible.

The **Classification** and **Regression** module get both the feature file, the **Classification** also the labels. We implemented three classification methods - Random Forest, Support Vector Machine and Deep Learning - and three regression methods - Linear Regression, SVM Regression and Random Forest Regression.

One of the most interesting results beneath changing the input data set is changing some of the input features to see how the results change. The hope is that biological insights can be drawn when investigating the important factors for a good performance in more detail. We implemented the possibility to change the input bin (all bins or a specific bin), the subset of used histone modification and to do either cross-validation or add two different data sets, one training and one test set. All tests and the modules in more detail are described in the sections below.

2.5.1 Preprocessing

For feature selection, we oriented ourself on the binning approach of Cheng et al. [8]. So the histone modification signal is binned around each transcription start site and transcription termination site of a gene in bins of a specific size and number. Both parameters can be adjusted. The default values we used most of the time are 40 bins before and 40 bins behind the TSS and the same number around the TTS, so 160 bins in total, each of size 100 bp.

In total, three different kinds of data are necessary to run the preprocessing part of our pipeline: First, the bigwig files need to be provided, one for each histone modification. To get the position of the TSS and the TTS of the genes in the binning step a GENCODE annotation file is required. Furthermore, the measured gene expression values are required. We chose to use the FPKM values from the tsv format which is already preprocessed by ENCODE. For the regression methods the expression values can be used directly, for the classification methods another preprocessing step is necessary.

To classify the genes, first a definition of the class labels is needed. Even if there are many possibilities to assign the genes into classes, we restricted our analysis on three common ones. Each time we split the genes into two classes. Genes were assigned to class 1 if their gene expression was above a specific threshold else to class 2. For the threshold we tried the mean of the expression values, the median of the expression values (according to Cheng et al. [8]) and a dataset-independent cut-off of zero (according to Dong et al. [11]). During the analysis, it was also assessed how the different labeling methods influenced the performance.

At last, also a module was added to normalize the binning features. Some machine learning methods such as Support Vector Machines are known to perform significantly better, if the input data is normalized. We thought about many different normalization methods, but restricted our detailed analysis on two variants where the performance increased significantly for some of the methods.

Both methods normalize the data over all genes for each bin and histone modification. The first approach, which we call scale in the following, uses Z scores for normalization. The mean of the column is subtracted from each data points and the result divided by the standard deviation of the column. In the second approach, which is referred as normalize in the following, the data is scaled to have unit norm using the L1 norm.

2.5.2 Unsupervised machine learning

Before we started with the classification and regression methods, we applied some unsupervised machine learning methods on the data. This is a typical step before implementing the more complex tasks to get a better view of the data in general. As a first step, we tried to find signal pattern and correlation pattern in the binning features, reimplementing some strategies of Cheng et al. [8].

The signal pattern shows the average distribution of each histone modification over the bins. Therefore, simply the average feature value for each bin and histone modification over all genes is calculated. To be able to compare the different histones, which have partly signal values in different amplitudes, the result is scaled over each histone row, using Z scores. Additionally, the correlation between the binning values and the expression values is calculated. For each bin and histone modification the Spearman correlation of the feature vector to the expression vector is calculated. Creating the same plots as Cheng et al. was also our first attempt to reproduce some of the published plots.

2.5.3 Classification

First, we reimplemented the two simpler classification methods Support Vector Machines and Random Forests according to Cheng et al. [8] and Dong et al. [11]. In the papers, Dong et al. used our so called zero approach to split the genes into two sets, while Cheng et al. used the median approach. To make our results better comparable, we tested each classification method with each of our labeling method and used the same feature set of binned histone signals, also described above.

The classification was first evaluated with a 10-fold cross-validation using only one data set each time and comparing afterwards the results for the different data set. Then the transferability of the learned models between the data sets was tested, by training on one dataset and predicting another. Each time, the AUC score was chosen as the evaluation measure. The AUC score is the area under the Receiver Operating Characteristic Curve, in which the sensitivity is plotted against the false positive rate. The values range from 0 to 1. 1 is the best score, 0.5 the score of a random predictor and values smaller than 0.5 indicate some specification problem such as inverted labels.

Classification and the following regression were both implemented using the Python package scikit-learn [21]. The package can be used to do 10-fold cross-validation and to calculate different evaluation scores, such as the AUC score.

2.5.4 Regression

Additionally to the different classification methods, we implemented the Random Forest Regression, the Support Vector Machine Regression and the Linear Regression, as used by Karlic et al. [7], Cheng et al. [8] and Dong et al. [11]. Dong et al. described differences when predicting the expression with and without zero expression values, because for many genes no expression was measured at all. If many of these zero values can be predicted correctly, the overall performance would already look good independent of the regression results for non-zero values. So we tested both, regression only on the non-zero values and on all values.

Analogously to the evaluation of the classification methods, the regression was performed with a 10-fold cross-validation of one data set and with training on one dataset and predicting another. The R2 score was used to evaluate the results. It is a typical score for the evaluation of regression methods and describes which proportion of variance in the outcome variable can be explained by the feature variables with the model. It is the square of the correlation coefficient r. The values range between 0 and 1, with 1 as the best value.

2.5.5 Parameter fitting

Many factors influence the performance of the classification and the regression models. To establish how good the prediction of gene expression values using histone modifications can get, we tried to fit various parameters of the models themselves and the preprocessing.

To optimize the performance of the methods on our datasets we tested different number of trees for Random Forest and we tried different kernel types for the Support Vector Machine. As described above, for some methods the normalization of the data can improve the performance, so we compared also our different normalization strategies. The classification methods are furthermore influenced by the chosen labeling methods, while the regression is influenced by using all expression values or only non-zero values.

2.5.6 Feature evaluation

We tried to assess how big the performance on each bin varies, so whether the spatial information of the histone modifications is important. It is known that some histone modifications show a stronger signal at the TSS, others in the gene body or at the TTS. Therefore we trained the methods using only one bin for each of the 160 bins and compared the performances.

Another interesting question beneath, which bin is more important, is, which histone modification is more important. Different papers, such as Karlic et al. [7] and Dong et al. [11], state that some histone modifications are significantly more important for the gene expression prediction, probably because they have a biological function in the transcription process. To test their hypothesis we run the models using all bins, but always only a subset of histone modifications. We tried first a leave-one-out approach, running the model on all but one histone modification. In this approach the performance influence of the histone modification which was left out can be estimated, comparing the performance of the whole model with the performance of the model without the histone modification. Because the differences between the performance measurements were not very clear, we tried afterwards to run the model using always only one histone modification and each pair of modifications.

2.5.7 Comparison of data sets

An important criteria which made the comparison of the published methods difficult was that each method was run on a different data set. We tested the models in the steps before always on the same dataset, using the same preprocessing, but still it can be possible that one method performs significantly better than the other methods on a special dataset and for another dataset it is the other way round. So different datasets must be tested to see if one method performs better independent of the data set and if there are big performance differences between the datasets, as shown by Singh et al. [13]. The paper suggested that the prediction works on some datasets much better compared to others.

With the goal to compare as much data sets as possible, we now use our bigger data set with 6 cell types and 5 histone modifications, while we use in all other sections the smaller data set with 3 cell types and 7 histone modifications. Furthermore, to see if more general features can be extracted when training on more than one cell type, we created a merged big data set, containing the cell lines K562, keratinocytes and gastrocnemius medialis. Using

this dataset and the 6 single cell lines, we did again a 10-fold cross-validation on each of them and compared the performance differences for different methods.
We also wanted to see how good the learned models are transferable to other cell types. Therefore, we created models, training a prediction method on each data set, and then tried to predict the other data sets with these models. The results are shown in a matrix for each prediction method, where the training sets are one axis and the test sets on the other.

2.5.8 Deep learning

To see if the more sophisticated approach of deep learning performs even better, when predicting gene expression using histone modifications, we reimplemented the convolutional neuronal network of Singh et al. [13]. Based on their description, our network contains first multiple layers of alternating convolution and max-pooling layers, followed by a drop-out layer, multiple fully connected layers and a soft-max layer. A schematic drawing of our network is shown in figure 3.

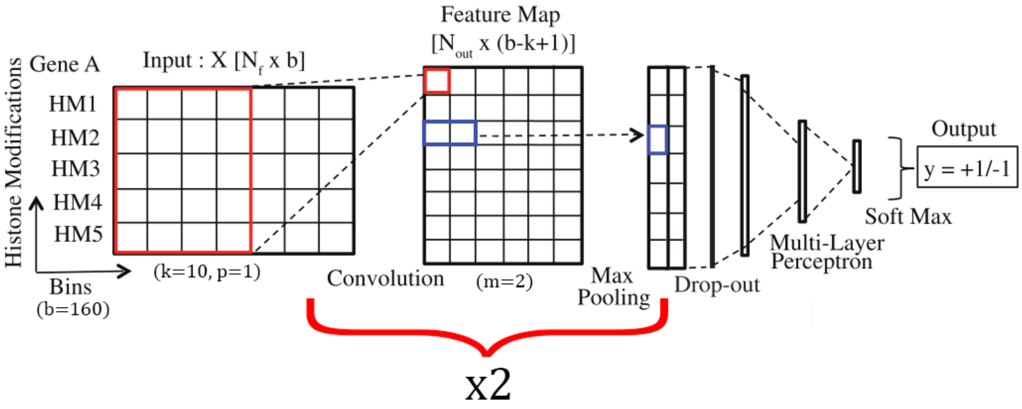


Figure 3 This shows the convolutional neural network as we implemented it. It contains two times a convolution and a max pooling layer, a dropout layer afterwards, a fully connected layer and the soft max function in the end, but we tested also other graph layouts.

The optimization of the network was done using the Adam Optimizer and the loss calculated using the cross-entropy. Before starting the training, the complete dataset was split into three parts: 81% were used for training, 9% for validating the training curve on an independent test set, called validation set, and the remaining 10% to assess the performance of the model after the completed training, called test set. To evaluate the results on the different sets, the accuracy and the AUC scores were calculated.
Among the many parameters which can be adapted in the network, the influence of the number of convolution layers, the number of hidden layers and different learning rates were tested to see how the performance changes. Furthermore, the drop-out layer was replaced by batch-normalization to see if an effect can be observed. Examples for our tested variable combination can be observed in the table 1. We tested three different learning rates, 0.05, 0.005 and 0.0005, to find the rate which produces the best learning curve. The curve should rise clearly with increasing number of steps, but do not oscilate too much.
Based on the description of Singh et al., we created first a deep neuronal network with a convolution layer with kernel size 10 and 50 output channels, followed by a maxpooling layer with a pool size of 2. Furthermore, we increased the number of convolution layers to 2 and

5, each layer with a kernel size of 10 and alterating with maxpool layers of pool size 2. When implementing 2 convolution layers, the output channels were 20 and 50, when implementing 5 convolution layers, the output channels were 20,30,40,50,60. Each model was of course tested on different data sets.
We used the framework TensorFlow [22] to implement the convolutional neural network.

2.6 Results

2.6.1 Unsupervised machine learning

First of all we looked at the signal and correlation pattern of all bins to see where the histone modifications mainly occur and how they correlate to the different expression values. Figure 4 shows the results exemplary for the data set of K562, but the plots show very similar signal and correlation pattern of the histone modifications in different data sets.

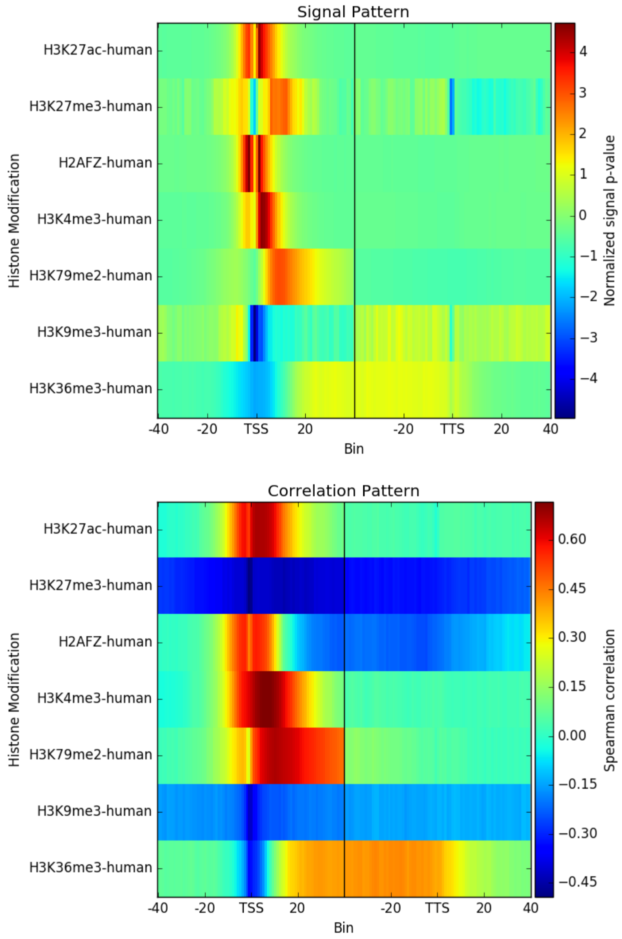


Figure 4 Signal and correlation pattern of the data set of K562. Top: Distribution of the different histone modification signals on the bins. Bottom: Correlation between the different bins and gene expression values per histone modification.

Looking at the signal values, it can be seen that the histone signal is not equally distributed in the region around the TSS and in the gene body, but that there are clear differences between histone modifications. H3K4me3 and H3K4me1 show a very strong signal in the region of the TSS, while H3K36me3 and H3K9me3 seem to be more present in the gene body and especially towards the TTS. H3K27me3 occurs mostly in the gene body behind the TSS.

In the correlation pattern we recognize a strong positive correlation of H3K4me3 around the transcription start site to the expression values, while H3K9me3 and H3K27me3 show both a very negative correlation to the expression values throughout the whole gene body. H3K4me1 and H3K36me3 are both also positive correlated with the gene expression, but in contrast to H3K4me3 the correlation gets stronger for bins in the gene body, for H3K4me1 at most in the gene body behind the TSS, for H3K36me3 in the gene body before the TTS.

2.6.2 Parameter fitting

The very first results after implementing the machine learning approaches was an AUC score of 0.81 for the Random Forest and 0.8 for the Support Vector Machine on bin 40 (the transcription start site). To improve these results we tried to optimize the different parameters starting with the parameters of Random Forest and the Support Vector Machine. In the Random Forest algorithm you can change the number of trees, so we ran the algorithm with different number of trees, using only bin 40. In figure 5 you can see that more trees are rising the AUC score, but starting around a number of 12 trees the difference is not really significant anymore. For this reason we chose 12 as the parameter for later runs. For the Support Vector Machine we tried different kernels. As you can see in figure 5, the linear kernel is working as good as the radial basis function kernel (rbf) and the polynomial kernel is only very slightly worse than these two. The sigmoid did not work at all. We decided to use the rbf kernel for our project, because it runs the fastest of the three best performing kernels.

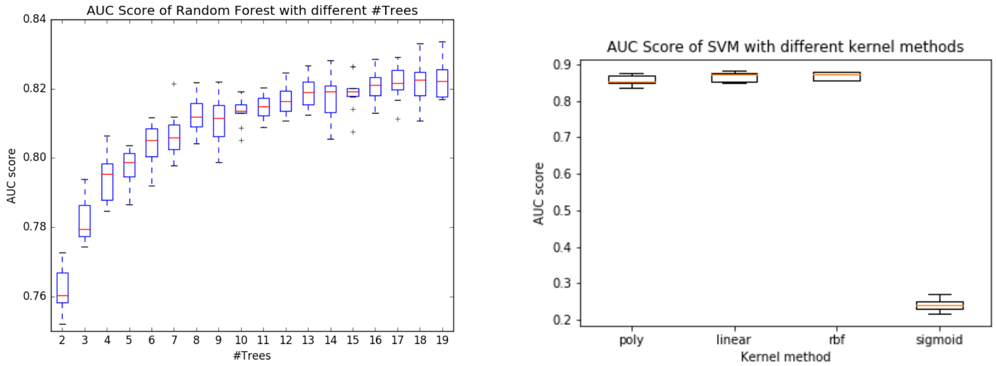


Figure 5 Parameter fitting of the classification algorithms, using the small version of the dataset K562 and bin 40. Left: Performance of Random Forest with different number of trees. Right: Performance of Support Vector Machines with different kernels.

To further improve this we tried different normalizations of the features, see figure 6. The different methods react differently on the normalization. For Random Forest classification and regression, no differences can be observed after normalization. However, the Support Vector machine classification performs significantly better after normalization. Both the unit norm normalization, but especially the scaling made a big difference in the performance. In contrast, the linear regression works better with normalized data than with scaled data.

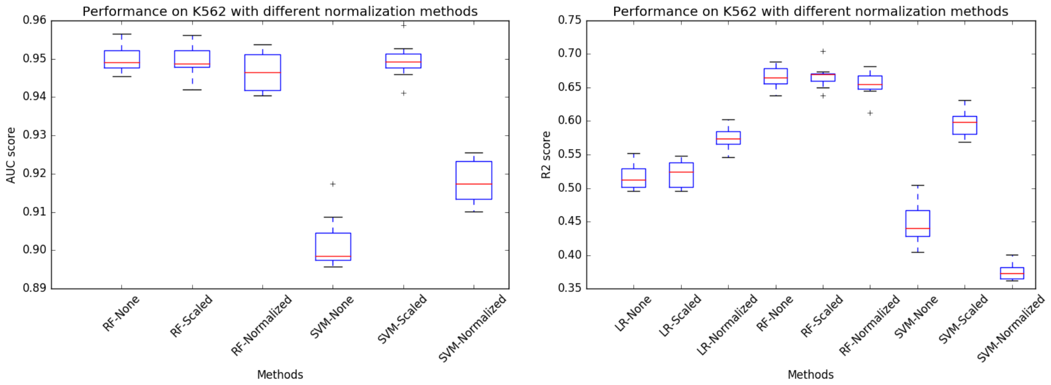


Figure 6 Performance on the small data set of K562 and on bin 40 with different normalizations of the histone modifications. Left with the different classification methods, right with the different regression methods.

We also compared different bin sizes and different number of bins, but the results showed that the performance did not change significantly for different bin sizes or a different total number of bins.

For the classification tasks, we split the genes in the two classes of low and high expression. For our first results, shown above, we always used the split at the expression value 0. However, our analysis of the different class labeling methods shows that the best performance is always with a median split where the two classes are equally divided, as shown in figure 7.

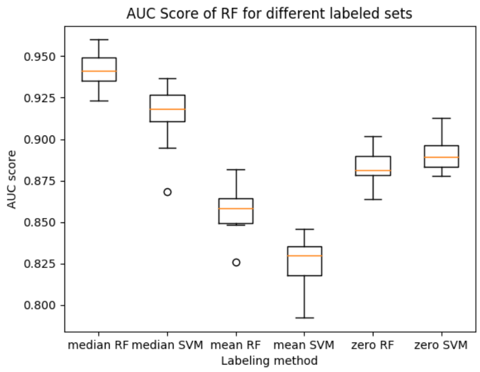


Figure 7 Performance of Random Forest and Support Vector Machine classification, comparing different labeling methods. The dataset K562_short was used and normalized before the analysis.

As mentioned above, Dong et al. describes large performance differences in the regression, when using all expression values or only expression values above 0 [11]. When comparing the two approaches, we observe small improvements (2%) in figure 8 because of the zero values, but not as huge as Dong et al.

2.6.3 Feature evaluation

The two dimensions of the input feature matrix, which are the different bins and the different histone modifications, were analyzed in detail to see how important individual bins

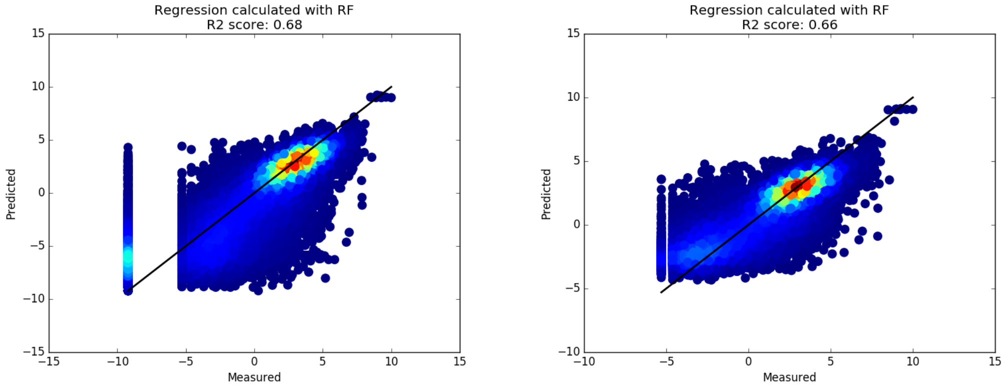


Figure 8 Performance of Random Forest Regression on the data set K562. Left: with zero expression values. Right: Without the zero values.

and individual histone modifications are. Both results can help to interpret the biological connection of histone modifications with the gene expression.

First, we trained the model on each bin individually. When using the data set K562, we can clearly observe a peak with the absolute best performance a few bins after the TSS. This pattern is visible in all classification and regression methods (see figure 9). Comparing the results with the smaller variant of K562 with 5 instead of 7 histone modifications, the peak is better visible in the smaller dataset than in the bigger data set before, reaching nearly the same performance. However, the performance drops in the gene body for the small data set. The third data set of the Endothelial cells shows in general a worse performance, but the curve is more similar to the curve of the big K562 data set.

For all data sets and methods the performance when all bins are used as features is the best, though only 2% better than the best performing single bin.

The role of individual histone modifications for the gene expression seems to be even more interesting, as histone modifications which are important for gene expression could possibly be identified. The results in the publications differ quite a lot in this topic. Looking at the performance of single and pairs of histone modifications compared to all histone modifications, only small differences can be observed in figure 10.

The full model with all histone modifications reaches a performance of 0.949, when using Random Forest, while the best two modification model 'H3K79me3-H3K36me3' reaches a nearly equal performance of 0.948. The differences between the pairwise histone modifications is also very small, all models are at most 10% worse than the full model and 9 of 21 pairs have a performance better than 0.93. In general, using only one modification performs worse than using two, but still there are single histone modifications which show a better AUC score than some pairs. Especially, H3K79me3 excels with an AUC score of 0.9 in the figure.

To compare the histone importance of different datasets in a convenient way, we included a table in our website that calculates the occurrence of each histone modification in the top performing histone pairs and single histones, given a specific threshold. In the figure 11 shown below, for the two data sets K562 and Endo and for all tested classification and regression methods, the occurrence of each histone modification in the top 33% performing pairs and single modifications is shown. The first number in the table shows always the total number of occurrences, the number in the brackets indicates how often the modification was counted as a top performing single histone modification, not in a pair of two modifications

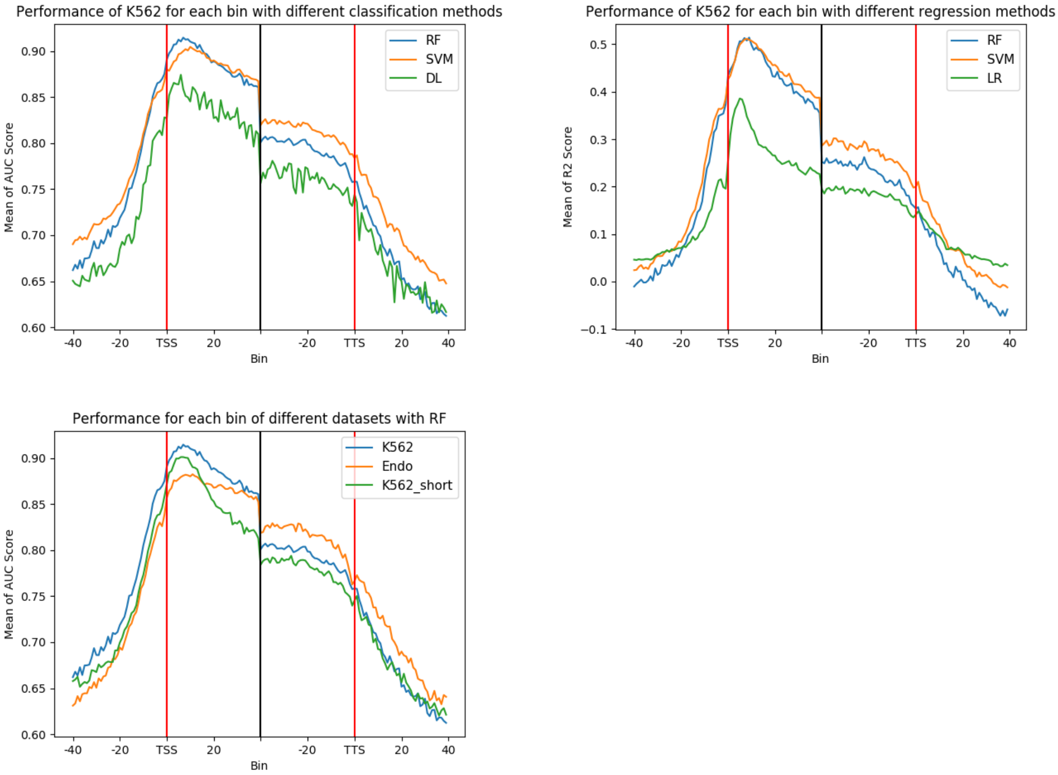


Figure 9 Comparing the performance of each individual bin. Top left: Performance for the dataset K562 and the three classification methods (RF: Random Forest, SVM: Support Vector Machine, DL: Deep Learning). Top right: Performance for the dataset K562 and the three regression methods (RF: Random Forest, SVM: Support Vector Machine, LR: Linear Regression). Bottom: Performance for three different data sets and Random Forest.

(so possible values are 0 and 1).

The table shows clearly that the occurrences of the histone modifications in the top performing subset varies more between the methods than between the data sets. The modification with the most occurrences for each data set and method is marked with a red square. Except for deep learning, the most occurring modification is always the same for each dataset tested with the same method. The table shows only two data sets, but it can also be observed for the third tested data set, the keratinocytes. No general differences between classification and regression can be observed: interestingly, the most often occurring histone modification is the same for Random Forrest classification and regression and the same for Support Vector Machine classification and regression.

2.6.4 Comparison of data sets

After fitting all the parameters we compared different data sets. As described above, we took the six smaller data sets and the merged data set of the cell lines K562, keratinocytes and gastrocnemius medialis.

Looking at the results of the 10-fold cross-validation individually for each data set (see figure 12), some differences between the data sets can be observed. K562 performs best with an

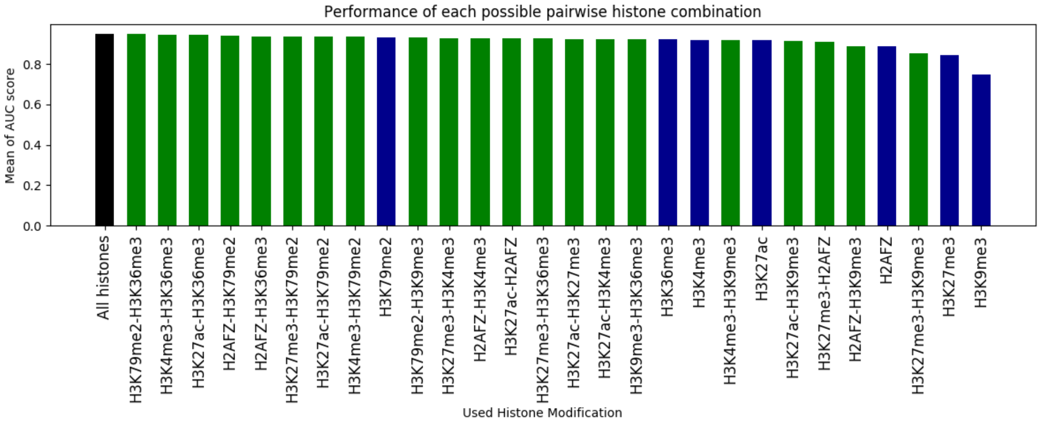
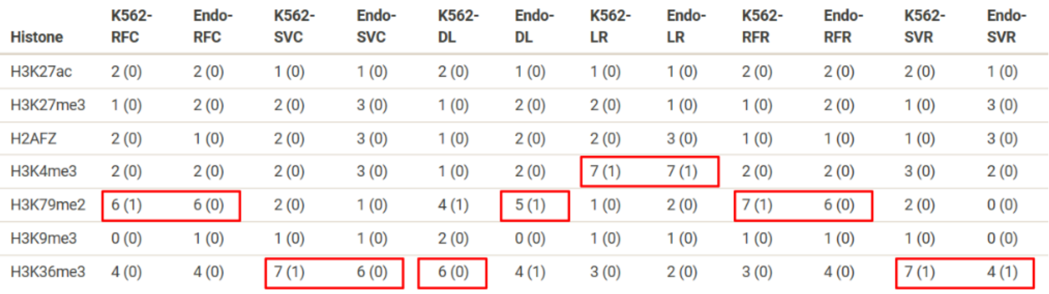


Figure 10 Comparison of performance of the data set K562 with all histone modifications to the the performances when using each histone modification alone or each pair of histone modifications. Calculated for Random Forest and a 10-fold cross-validation.



Histone	K562-RFC	Endo-RFC	K562-SVC	Endo-SVC	K562-DL	Endo-DL	K562-LR	Endo-LR	K562-RFR	Endo-RFR	K562-SVR	Endo-SVR
H3K27ac	2 (0)	2 (0)	1 (0)	1 (0)	2 (0)	1 (0)	1 (0)	1 (0)	2 (0)	2 (0)	2 (0)	1 (0)
H3K27me3	1 (0)	2 (0)	2 (0)	3 (0)	1 (0)	2 (0)	2 (0)	1 (0)	1 (0)	2 (0)	1 (0)	3 (0)
H2AFZ	2 (0)	1 (0)	2 (0)	3 (0)	1 (0)	2 (0)	2 (0)	3 (0)	1 (0)	1 (0)	1 (0)	3 (0)
H3K4me3	2 (0)	2 (0)	2 (0)	3 (0)	1 (0)	2 (0)	7 (1)	7 (1)	2 (0)	2 (0)	3 (0)	2 (0)
H3K79me2	6 (1)	6 (0)	2 (0)	1 (0)	4 (1)	5 (1)	1 (0)	2 (0)	7 (1)	6 (0)	2 (0)	0 (0)
H3K9me3	0 (0)	1 (0)	1 (0)	1 (0)	2 (0)	0 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	0 (0)
H3K36me3	4 (0)	4 (0)	7 (1)	6 (0)	6 (0)	4 (1)	3 (0)	2 (0)	3 (0)	4 (0)	7 (1)	4 (1)

Figure 11 Occurrence of each histone modification in the top 33% best performing histone pairs and single histones, for the data sets K562 and Endo and all implemented classification and regression methods. In each entry the first number shows the total number of occurrences of the histone modification, the second number in brackets the number of occurrences as a top performing single histone modification. The modification with the most occurrences in each column is marked with a red square.

AUC score median of 0.95, the worst small data set is the thyroid gland with a median score of 0.89. The performance of the merged big data set is always between the performance of the datasets that were merged in the big one (always between 0.92 and 0.95).

In contrast, the regression methods in figure 12 show an other tendency. For linear regression, the data set of the keratinocytes performs best with a median R2 score of 0.51. The performance of the big dataset is extremely variable between 0.32 and 0.52 for the linear regression, the other regression methods show a better performance of the merged data set but all have a big variance in the performance.

We trained a model with each data set and tested it with all the others. In all matrices, where the performance of different datasets on models with other trainings sets is shown, you can see that the performance on the diagonal, when predicting the training set, is clearly the best. In figure 13 on the right side you can see that the Support Vector Machines diagonal values are between 0.96 and 0.97. SVM trained and evaluated on two different data sets still perform quite well. All AUC score values are above 0.85.

Figure 13 on the left side shows the comparison of the different datasets using the Random

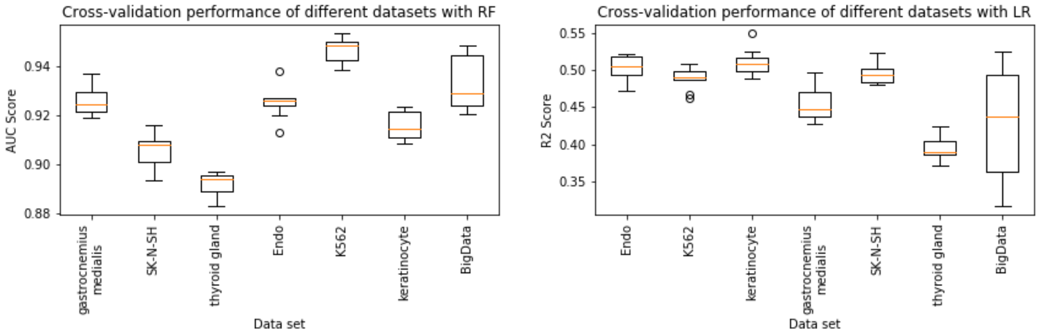


Figure 12 Performance of the 6 small data sets with 5 histone modifications and the big merged data set each for a 10-fold cross-validation. Left: Cross-validation with Random Forest Classification. Right: Cross-validation with Linear Regression.

Forest Classification. Here, the AUC score is even higher when predicting on the training set with always a value above 0.999, but not that high when you test the model on another dataset than you trained. The best value is at 0.93, but a few are only as good as a random classifier with values around 0.5.

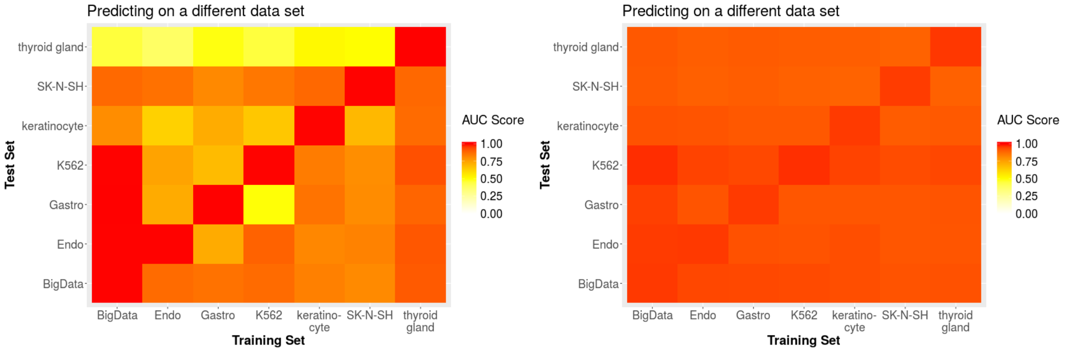


Figure 13 Performance of the classification methods when training on one data set and prediction the other datasets. Left: for Random Forest. Right: for Support Vector Machine

For the regression prediction the results were similar, as shown in figure 14. For the Random Forest Regression we see that the prediction on another dataset is around 0.4, the prediction on the same dataset is around 0.92. The Linear Regression has a very bad performance for most of the datasets and the Support Vector Machines has better values for the performance on other datasets, but the performance on the same dataset is around 0.66.
We used the merged data sets as well. The results for all methods show that the performance of the three datasets that are in the big one is nearly as good as when you test it on itself. We observed for all methods that the performance on other datasets is not better and the prediction of the dataset is not easier.

2.6.5 Deep learning

In our deep convolution network, many parameters can be adapted, which possibly have an influence on the learning curves and performance results. We tested different number of convolution layers, different learning rates and batch-normalization instead of a dropout-

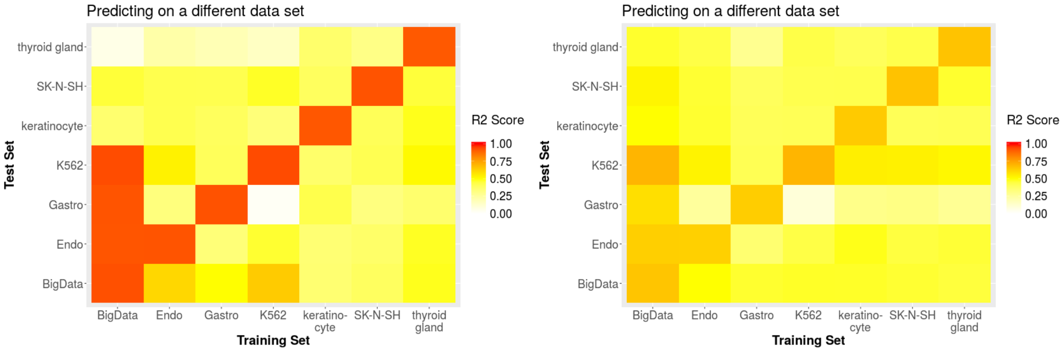


Figure 14 Performance of the regression methods when training on one data set and prediction the other datasets. Left: for Random Forest. Right: for Support Vector Machine

layer.
Independent of our tested parameter, the network shows a quick increase of the AUC scores both for the training and the validation set during training, but also a quick saturation of the accuracy and AUC curves of the validation set, see figure 15. Already after 500 iterations the validation set performance has its maximum and does not rise anymore. This development was always visible, no matter which parameter we used.

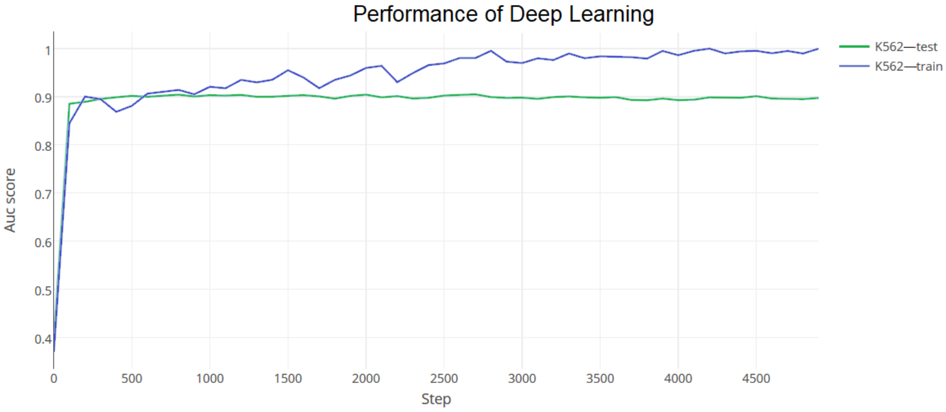


Figure 15 Learning curves of the training and the validation set of K562 with two convolution layers and a learning rate of 0.0005. In total 5,000 iterations were made (each with a batchsize of 200 genes) and after each 100 iterations the AUC scores of the training and the validation set were calculated.

Also the final performance values on the test set showed no big differences, see table 1. We tested three different learning rates: a rate of 0.05 is too big training the model, the curve oscillates too much. So we discarded that learning rate immediately. The learning rates of 0.005 and 0.0005 differ not significant in the AUC scores on the test set, both can train a model with a similar performance. So we chose the threshold of 0.005, as here the learning is steeper in the first steps, because it learns faster.
We tested also different network layouts, such as different numbers of convolution layer. The AUC scores on the test set differ also here not significantly, independent if we use 1, 2 or 5 convolution layers, again shown in table 1. It can only be observed that training 5 convolution layers takes more time. So we chose 2 layers for the following analysis. Also

taking batch normalization instead of a dropout-layer with 50% drop-out rate could not change the performance.

Learning rate	Number convolution layers	Convolution kernel size	Maxpooling pool size	Output channels after each convolution	Dropout/ Batch-Normalization	Performance on test set (AUC)
0.005	2	10 (2x)	2 (2x)	20,50	Drop-out	0.897
0.0005	2	10 (2x)	2 (2x)	20,50	Drop-out	0.901
0.005	1	10	2	50	Drop-out	0.902
0.005	5	10 (5x)	2 (5x)	20,30,40,50,60	Drop-out	0.902
0.005	2	10 (2x)	2 (2x)	20,50	Batch-Norm	0.901

Table 1 Performance of the test set of data set K562 after 1,000 learning iterations, after changing different parameters of the convolutional network.

2.6.6 Comparison of the methods

Finally, we compared the different classification and regression methods as Singh et al. did [13]. For each method and the three bigger data sets, we analyzed the result values of the 10-fold cross-validation with optimized parameters. In figure 16 we see on the left site the results of the comparison of the different classification methods. Deep Learning performs worst for all cell lines, the highest AUC score is 0.89 for K562 and the lowest 0.84 for the keratinocyte. Random Forest and the Support Vector Machines have very similar results and the performance is between 0.95 and 0.92.

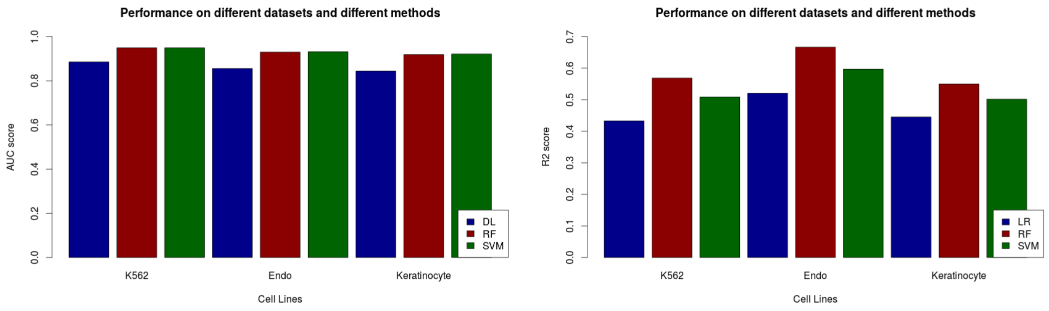


Figure 16 Comparison of the different methods we used on the different datasets we had, using the bigger data sets with 7 histone modifications and a 10-fold cross-validation for each method. Left: the different classification methods (DL - Deep Learning, RF - Random Forest, SVM - Support Vector Machine). Right: the different regression methods (LR - Linear Regression, RF - Random Forest, SVM - Support Vector Machine)

We also compared the different regression methods, see figure 16 on the right site. The Random Forest Regression is the best predictor with a maximal score of 0.67 and a minimum score of 0.55. The Linear Regression is the worst regression method with scores between 0.43 and 0.52.

2.7 Discussion

Our signal and correlation patterns in figure 4 show similar results as in the paper of Cheng et al. [8]. They analyzed even more histone modifications, which were not measured for our data sets, but they identified very similar signal and correlation pattern for the four histone modifications which were in both our and their data set. In general, the positive correlation of H3K4me3 around the TSS, the positive correlation of H3K36me3 in the gene body and negative correlation of H3K27me3 and H3K9me3 throughout the whole region match the current biological knowledge about these histone modifications. H3K4me3 and H3K36me3 are two known gene expression activators, while H3K9me3 and H3K27me3 are expression suppressors (see [2] and [3]).

So the basic biological connections between single histone modifications and gene expression can be identified easily, when using our data and the binning approach of Cheng et al. [8]. The correlation signal differs in the gene region and different spatial patterns can be observed for different histone modifications. This indicates that this binning approach captures more information than calculating only one threshold per gene (as for example Karlic et al. did [7]) and might also improve the predictions.

When assessing the performance of each classification and regression method, it is important to know that many factors have an impact on the exact performance quality of each method. We showed that the labeling method has a very big influence on the performance and it is the best to make a median split, see figure 7. The reason is that the classification is working better when you have two classes of an equal size.

The normalization method has a big influence on the Support Vector Machine classification and the SVM regression. While the performance of the Random Forest classification and Random Forest regression does not change very much between different normalization methods, we see clearly that scaling the signals of the histone modifications makes the performance of the SVMs much better. The reason can be explained by the algorithms behind both methods. The decision trees of the Random Forest methods split the data into different branches, choosing a distinct threshold. When scaling the data, simply the same now scaled threshold is chosen, but this does not improve the performance. During the Support Vector algorithms, the optimal support vectors need to be found in the feature space. It is highly recommended to scale the data before, because that improves the speed and the quality to find those support vectors.

This shows one problem when investigating which method performs better. Using unnormalized data, the Random Forrest algorithms work better, but it would be wrong to say that they are better in general, because when using normalized data the Support Vectors can be even better for some data sets.

Other parameters, which can be adapted in our process to generate features are the number and size of bins. When we analyzed those two parameters, the different bin sizes and different total number of bins did not have an influence on the performance. Our hypothesis is, that the clearest signal is around the transcription start site and in the gene region, as the analysis of the individual bins showed in figure 9. Increasing the number of bins adds a few additional information at the beginning, but we could see no more performance difference, when using 120, 160 or 200 bins. To make our results as good as possible comparable with the publications we decided to use always the same bin size (100 bp) and number of bins (160 bins) as in literature [8].

Many of our results are very similar to literature. We were able to show that the bin that has the best performance is a few bins after the TSS, see figure 9. This was also shown by Cheng et al. [8] and Sing et al. [13]. It is in accordance with the distribution of the histone modifications as shown in figure 4. It is very interesting that the clear peak disappears when we used a bigger set of histone modifications. The reason is in the choice of the used histone modifications. When we look at the signal distribution of the bigger set of histone modifications we see that especially H3K79me2 appears very often in the gene body and has a high correlation with the expression in this region. This modification is only part of the bigger data set of K562, but not of the small one. It is the same histone modification that is the most important modification for the data set K562, see figure 10. So H3K79me2 seems to be a well suited histone modification for the gene expression prediction and it occurs mainly in the gene body. So the prediction of bins in the gene body improves when data from this histone modification is used. This can be an explanation why the performance the big data set including H3K79me2 is significantly better in the gene body than the performance of the small data set in this region, while it is very similar around the TSS.

Even if the plot over the individual bins of Cheng et al. [8] is very similar, it showed better performance values in the region around the TTS. Cheng et al. did not only use histone modifications, but also Polymerase II binding sites, which showed a strong signal in the TTS region. So it is probably the influence of the polymerase II which improves the prediction in this region in the plots of Cheng et al.

In general, the best performing region for the prediction is close to the TSS in the gene body, where most of the histone modifications are found, but depending on the set of used histone modifications (or other features) also bins in other regions, such as the gene body and the TTS can be used for predicting with high AUC scores. To better identify the most informative bin for prediction, a as complete as possible set of measured histone modifications is required.

For identification of the most important histone modifications, we have a very similar problem that only a small part of all possible histone modifications are currently measured. Our results in figures 10 and 11 showed that the performance between the data sets seems to differ less than between the methods. Interestingly, the most important histone modification is the same for Random Forrest Classification and Regression and the same for Support Vector Machine Classification and Regression. That contradicts the result of Dong et al. [11] who showed that different modifications are most important for classification and regression. However, Dong et al. compared the most important histone modifications of the Random Forest classification with the most important histone modifications of the Linear regression, so two different methods. These show different results in our analysis, too, but not Random Forest classification and regression, so it would be wrong to say that there are general differences between classification and regression with our results. The problem is that Dong et al. tested only one regression and one classification method.

Furthermore, setting a cut-off for the most important histone modification is quite difficult, as many pairs of histone modifications show a performance very similar to the total performance. Still, our most important histone modifications are similar to the ones of Dong et al. [11], which were H3K4me3, H3K79me2 and H3K36me3, even though not the same differentiation between classification and regression methods could be made. This might suggest that even if it is not possible to find the one most important histone modification, still a subset can be found which shows a better performance. Or thinking the other way round, some of the histone modifications, in our case H3K9me3, perform significantly worse than others, independent of the method and data set. This histone modification seems to be at least not alone sufficient for predicting gene expression, but could of course influence the expression in connection with other histones.

Karlic et al. [7] found a completely different subset of histone modifications significant, which were except for H3K27ac not measured in our data set. As he used a much bigger data set of 41 histone modifications, of which most did not exist in our data set, the results are difficult to compare. That shows again the data problematic.

All this results lead us to the hypothesis that you can not clearly identify the most important histone modification, but that there is a set of many histone modifications which can be used for a good prediction of the gene expression. There is a high redundancy in the data, so that different subsets of these modifications can accomplish nearly the same performance as the case model. That different modification are found to be most important could possibly be explained by differences in the algorithms. At least between the different cell types the important modification do not change, suggesting the same or a very similar biological role of the modifications in the different cell types for the expression.

The cross-validation shows that there are only small differences in the performance of the seven different data sets, independent of the method used. This indicates that in general the method seem to work for each data set. That the performance is worse on some data sets can have multiple reasons: either the relationship between the expression values and the histone modifications is harder to predict for them, as the pattern is more complex, or simply the measurement results for those data sets were not as accurate as for the other data sets. Because the differences are only so small, they should not be over-interpreted.

We compared different datasets with each other by training the model on one dataset and testing it on another. The results show very clearly that the prediction is always very good after training the dataset with a Support Vector Machine. It shows that a Support Vector Model is not that specific on a dataset as the Random Forest. The prediction with a Random Forest Model gives always a score of nearly 1 on the training dataset, but on other datasets the AUC score is sometimes as bad as a random prediction with values around 0.5. The Support Vector Machine has a worse performance on the trainings set, but is much better in the prediction for the other datasets. This shows that the SVM learns more general features than Random Forest. When using the right method, such as the Support Vector Machine, it is possible to predict also genes from another dataset than the training set with a high accuracy.

Very interesting is that figure 13 shows that the performance of thyroid gland is always very bad when you trained the model with another data set, but the performance on other datasets after training it with thyroid gland is always very good. We also see that it SK-N-SH is always easy to predict.

We did the same comparison by using the regression methods. Here the prediction of other datasets did not work very well, especially the linear regression gave only very bad results. It means that the linear regression models are very specific to the dataset and the regression is in general more complicated then the classification.

The big merged data set did not perform better than the small ones, neither for classification nor for regression, but the performance values were always in the same range. Our hope that more general features could be captured with more data, containing more than one cell type, was not fulfilled. The same generality of features could already be captured by the smaller data sets. Maybe another method or another configuration of the merged data set, e.g. taking other data sets or simply more data sets, can increase the performance on other data sets with a merged data set. Unfortunately, we were not able to investigate this further.

The Deep Learning results show that we get a good performance very easy on the test data but already after 1000 iterations the AUC score does not rise anymore. Different learning rates and different numbers of convolution layers could not change that result. So a relatively good predicting model can be learned very fast, but then the performance does not increase anymore with continued training.

Maybe we could improve the results of Deep Learning by trying to improve the algorithm but it is possible that you would simply need much more data than we had. Unfortunately, there was not enough time to investigate that further.

Finally, we see in the comparison of the classification methods that the Deep Learning as we implemented it is always worse than the much simpler methods Random Forest and Support Vector Machines. Our results are contradicting the method comparison of Sing et al. [13]. In his paper he had values between 0.92 and 0.69 for deep learning. With AUC scores between 0.85 and 0.90 we are similar good with our convolutional network, but for the Random Forest he got an average AUC of 0.72 and for the Support Vector Machine an average AUC of 0.87. This is much worse than the results we got without fitting the parameters. Our AUC performance values of the fitted models are between 0.92 and 0.96.

The deep neural network had the worst performance values of our methods, even if they were still good, Support Vector Machine and Random Forest worked better. For Singh et al. it is the other way round, the deep learning network showed the best performance. Moreover the performance of Support Vector Machine and Random Forest was for us very similar, Singh at al. showed that the Support Vector Machine worked better. A possible reason, why SVMs and Random Forest performed better in our tests, could be that Singh et al. did not take much time in implementing the other methods as good as possible, as they wanted to promote their method as the best new method. Additionally, our results are better comparable than the results of Singh et al. because he used for the Support Vector Machines and the Random Forest only one bin and for the Deep Learning all bins. We trained each method on the same features set, using always all bins.

Our performance results fit also much better to the performance values in the original publications of Cheng et al. [8] and Dong et al. [11]. Cheng et al. proposed that their SVM got AUC values in the cross-validation of 0.90, Dong et al. for their Random Forest even of 0.95. This indicates also that Singh et al. did not implement the best performing variants of the two classification methods.

Furthermore, we also never saw such a big variance between the different datasets as Singh et al. Their results deviate from values of 0.92 to 0.68 even for the deep learning implementation. At least for our six tested data sets we saw only very little variance. This could be by chance, as Singh et al. tested 56 cell types. But it could also be possible that our methods perform more robust. Further data sets needed to be tested to determine the reason for the different results.

The difference between the regression methods is greater, here we see that Random Forest gives the best results in all datasets, and also the differences between the data sets is a bit more significant compared to the regression methods. Our linear regression performs a bit worse with R2 scores between 0.43 and 0.52 compared to the implementation of Karlic et al. [7] and Dong et al. [11] with both a score of 0.59. Our Support Vector Regression has R2 values between 0.55 and 0.63, while Cheng et al. [8] published a R2 score of 0.56. For the Random Forest regression we have unfortunately no published scores for comparison. Even if the performance results are not identical, they differ not very much and could probably be explained by the use of different data sets or slightly different preprocessing.

2.8 Conclusion

All results show that there is a close connection between gene expression and histone modifications. The performance values for the prediction are in general very good, independent of the used data set and method. This was already shown in many papers and fits to the biological background.

Comparing different methods, such as Singh et al. [13], seems to be a complex task, because the exact performance values depended on many factors such as the normalization of the features. Each method shows good performance, when preprocessing the data correctly, so the connection between gene expression and histone modifications can be robustly detected by all methods. The classification methods work better than the regression methods, which is not surprising, as predicting the exact value is more error-prone than predicting only a class.

The Support Vector machines show a better performance, when predicting another data set, and a worse, when predicting the training dataset, so they may extract more general features than Random Forest. This is true for both classification and regression. The linear regression, which is the least sophisticated method, shows clearly worse results than the other methods.

For Deep Learning, there was unfortunately not enough time, to calculate all the other results after pruning the parameters. Still, the first results suggested, that the performance is worse than with the other classification methods, contradicting to Singh et al. [13]. Possible reasons could be that we had too less data or that another graph layout would perform better as Singh et al. did not describe his network in detail.

Looking at, where the histone signal is the most informative for prediction gene expression, the region shortly after the TSS shows the best performance values, as shown by Cheng et al. [8]. Still, we could show that results of this analysis depend on the used histone modification, as for some modifications also a predictive signal in the gene body was found. For a more universal statement, more histone modifications need to be measured and then the analysis repeated.
The histone modifications seem to be quite redundant, as only a small subset of them is enough for a good performance. Also many histone modifications can be used in the subset, so it seems that they are strongly correlated and it is not possible to find the one most important histone modification for classification. This is also supported by our results, which show that the different methods differ more in the most important histone modification than between different data sets. This may also be an explanation why different papers identify different histone modifications, when using only one data set and one method.

The comparison of different datasets showed that some are easier to predict and some are more difficult but in any case the prediction gets never as bad as in the comparison of the 56 cell lines of Singh et al. [13]. Using SVM, even data sets from different cell types as the training set cell type can be predicted with very high AUC scores. This indicates that there are general mechanisms of the interaction between histone modifications and gene expression across different cell types in human.

There are many more different aspects that can be analyzed in context with gene expression prediction using histone modifications. As we restricted our analysis only to protein-coding genes, it would be interesting to test also other kind of RNAs. Furthermore, we concentrated our data set comparison to cell types of human, but also cell types of different species could be compared. So further analysis is possible and may reveal more interesting results in context of the biological connection between gene expression and histone modifications. But for this, first more data sets need to be created, especially measuring much more different histone modifications for one data set. This would improve the explanatory power of all analyses significantly.

2.9 Our website

The plots shown in our chapter visualize only a part of the data we ran during our project. A more comprehensive summary of all our plots can be found at our website. Beneath many interactive plots visualizing the results our data analysis, there is also the opportunity to upload your own data set and try to predict it with one of our tested methods.
To run the interactive website in R, the following packages need to be installed:

- shiny
- shinythemes
- plotly
- reshape2
- ggplot2

Additionally, all our scripts are available in our github project *https://github.com/seidenfeder/neap.* It is possible to run all scripts and reproduce all shown plots with new data or other parameters. The following Python packages need to be installed therefore:

- numpy
- sklearn
- optparse
- math
- matplotlib
- mpl_toolkits
- scipy
- __future__
- os
- argparse
- tensorflow
- time
- pyBigWig

References

[1] Xianjun Dong and Zhiping Weng. "The correlation between histone modifications and gene expression". In: *Epigenomics* 5.2 (2013), pp. 113–116. DOI: 10.2217/epi.13.13. arXiv: NIHMS150003. URL: http://www.futuremedicine.com/doi/10.2217/epi.13.13.

[2] Christoph M. Koch et al. "The landscape of histone modifications across 1% of the human genome in five human cell lines". In: *Genome Research* 17.6 (2007), pp. 691–707. DOI: 10.1101/gr.5704207.

[3] Artem Barski et al. "High-Resolution Profiling of Histone Methylations in the Human Genome". In: *Cell* 129.4 (2007), pp. 823–837. DOI: 10.1016/j.cell.2007.05.009. arXiv: NIHMS150003.

[4] S. Vinod Kumar and Philip A. Wigge. "H2A.Z-Containing Nucleosomes Mediate the Thermosensory Response in Arabidopsis". In: *Cell* 140.1 (2010), pp. 136–147. ISSN: 00928674. DOI: 10.1016/j.cell.2009.11.006.

[5] T. Jenuwein. "Translating the Histone Code". In: *Science* 293.5532 (2001), pp. 1074–1080. DOI: 10.1126/science.1063127. arXiv: arXiv:1011.1669v3. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1063127.

[6] http://www.amsbio.com/images/featureareas/nucleosomes-and-histone-proteins/nucleosomes.jpg.

[7] "Histone modification levels are predictive for gene expression". In: *Proceedings of the National Academy of Sciences* 107.7 (2010), pp. 2926–2931. DOI: 10.1073/pnas.0909344107. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.0909344107.

[8] Chao Cheng et al. "A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets". In: *Genome Biology* 12.2 (2011), R15. URL: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-2-r15.

[9] Xiaojiang Xu et al. "Application of machine learning methods to histone methylation ChIP-Seq data reveals H4R3me2 globally represses gene expression." In: *BMC bioinformatics* 11 (2010), p. 396.

[10] Ivan G Costa et al. "Predicting gene expression in T cell differentiation from histone modifications and transcription factor binding affinities by linear mixture models". In: *BMC Bioinformatics* 12.Suppl 1 (2011), S29. URL: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-S1-S29.

[11] Xianjun Dong et al. "Modeling gene expression using chromatin features in various cellular contexts". In: *Genome Biology* 13.9 (2012), R53. URL: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-9-r53.

[12] Anirudh Natarajan et al. "Predicting cell-type - specific gene expression from regions of open chromatin the genome". In: *Genome Research* (2012), pp. 1711–1722. DOI: 10.1101/gr.135129.111.

[13] Ritambhara Singh et al. "DeepChrome: Deep-learning for predicting gene expression from histone modifications". In: *Bioinformatics* 32.17 (2016), pp. i639–i648. ISSN: 14602059. DOI: 10.1093/bioinformatics/btw427.

[14] Wouter Kundaje, Anshul Meuleman et al. "Integrative analysis of 111 reference human epigenomes". In: *Nature* 518.7539 (2015), pp. 317–330. ISSN: 0028-0836. DOI: 10.1038/nature14248.

[15] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. "Rectifier Nonlinearities Improve Neural Network Acoustic Models". In: *Proceedings of the 30 th International Conference on Machine Learning* 28 (2013), p. 6. URL: https://web.stanford.edu/{~}awni/papers/relu{_}hybrid{_}icml2013{_}final.pdf.

[16] Natalie de Souza. "The ENCODE project". In: *Nature Methods* 9.11 (2012), pp. 1046–1046. ISSN: 1548-7091. DOI: 10.1038/nmeth.2238. URL: http://www.nature.com.ezp.lib.unimelb.edu.au/nmeth/journal/v9/n11/full/nmeth.2238.html{\%}5Cnhttp://www.nature.com.ezp.lib.unimelb.edu.au/nmeth/journal/v9/n11/pdf/nmeth.2238.pdf.

[17] Felipe Albrecht et al. “DeepBlue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets.” In: *Nucleic acids research* 44.i (2016), gkw211–. ISSN: 1362-4962. DOI: 10.1093/nar/gkw211. URL: http://nar.oxfordjournals.org/content/early/2016/04/15/nar.gkw211.full.

[18] *BLUEPRINT - A BLUEPRINT of Haematopoietic Epigenomes.* http://www.blueprint-epigenome.eu.

[19] *Canadian Epigenetics, Environment and Health Research Consortium (CEEHRC) initiative.* http://www.epigenomes.ca.

[20] *CREST program Development of Fundamental Technologies for Diagnosis and Therapy Based upon Epigenome Analysis (Disease Epigenome).* http://crest-ihec.jp/english/index.html.

[21] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[22] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.* Software available from tensorflow.org. 2015. URL: http://tensorflow.org/.

3 Transcription Factor Binding

by Pia Kirchmeier and Gergely Csaba

3.1 Transcriptional Regulation

Transcription factors (TFs) are proteins which are assumed to bind DNA in a sequence-specific manner, thereby controlling the rate of transcription alone or in cooperativety with other TFs by promoting or blocking the recruitment of RNA polymerase. The human genome is estimated to contain about 2000 to 3000 TFs, but most remain unannotated and only few have been experimentally verified to date [23]. For the up- or downregulation of genes adjacent to the targeted enhancer or promoter regions, a variety of mechanisms have evolved, including the stabilizing or blocking of RNA polymerase as well as recruiting coactivator or corepressor proteins to the transcription factor DNA complex. Alternatively, the acetylation or deacetylation of histone proteins may be catalyzed to induce the weakening of the association of DNA with histones to make the DNA more accessible to transcription and vice versa [24, 25]. By ensuring the correct expression of specific genes, a wide range of biological processes is controlled. Many TFs are involved in cell fate determination and cellular differentiation as well as the maintenance of intracellular metabolic and physiological balance and participate in intercellular signal cascades and response to environmental stimuli [26, 27]. Regarding the evolutionary history of human TFs, the proliferation of new regulatory genes seems to coincide with the emergence of increasing organismal complexity and TF expansions are suggested to have continued until recently in human evolution. It has been shown that TF-coding genes tend to underlie great positive evolutionary selection and minor differences in the underlying nucleotide sequence may induce profound effects on regulatory function, especially if the mutations concern DNA binding sites [23].

3.2 ChIP-sequencing

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) has become the most commonly used technique for mapping TF binding sites (TFBS) in living cells on a genome-wide scale. Although TFs recognize short sequence motifs with a length of 8-21 base pairs (bp) in the genome, they only associate with a small subset of the sites matching its binding motif [28, 29]. ChIP experiments aim to map those binding sites with maximal signal-to-noise ratio and completeness across the genome.As the first step, the proteins are covalently cross-linked to DNA by treating the cells with a chemical agent. The chromatin is then fragmented by cell disruption and sonication or enzymatic digestion, and the protein of interest (e.g. one specific TF) which is bound to a DNA fragment is enriched relative to the starting chromatin. This is done by using an antibody or an epitope tag specific for the TF. Subsequently, the cross-links are reversed and the enriched DNA is purified. Finally, the sequence is identified by DNA hybridization to a microarray (ChIP-chip), or more recently by DNA-sequencing (CHIP-seq) [30, 31].

The quality of different ChIP experiments may vary significantly and depends heavily on the specificity of the antibody and the enrichment degree achieved in the affinity precipitation step. Generally, poor reactivity against the target or cross reactivity with other DNA-associated elements constitute the main deficiencies of antibodies. Thus, one cannot assume a strong relationship of ChIP signal strength to biological regulatory activity. As even some highly transcriptional enhancers reproducibly show modest ChIP signals, it is hard to a priori set a specific target threshold for ChIP signal strength assuring inclusion of all functional sites

[32]. In order to balance data quality with practical attainability, the ENCODE Consortium has established a set of quality thresholds. To assess replicate agreement, the irreproducible discovery rate (IDR) analysis methodology is used which quantifies the peaks present in both replicates with a comparable binding signal [33].
After the sequenced reads have been mapped to the genome, peak calling software is used to obtain regions of ChIP enrichment. For this task, several software packages are available, with the most widely used being SPP, PeakSeq and MACS [34, 35, 36]. The output of these methods generally ranks called peaks by absolute signal (i.e. read number) or by significance of enrichment (e.g. p-values). However, significance values from different software packages are not directly comparable as they rely on different statistical models and the composition of the final peak list depends significantly on the specific parameter setting [30].

3.3 How complex is gene regulation?

The main question we want to address is the complexity of gene regulation. In this chapter, we are specifically interested in the impact of TF binding and want to identify the level of dependency of gene regulation and specific TF binding patterns. This issue is a major topic in current research, and a great number of different approaches trying to predict gene expression from TF regulation are available. In the following section, we describe some of the available methods predicting gene regulatory networks (GRNs) and TF-target gene interactions. Subsequently, we present an extensive analysis of TF binding dynamics in human embryonic stem cell (HESC) differentiation. Based on the data of this approach, we designed our project which is presented in the last section of this chapter. Our aim was to find out if it is possible to identify clear dependencies of TF binding and gene regulation based on differential binding data of a small number of TFs and genome-wide differential gene expression data. The concept of differential expression helps to identify dependencies when the binding of one or multiple TFs changes between two states and the expression of the gene regulated by those TFs also changes (see figure 17).

3.4 Existing methods for predicting gene regulation

In contrast to RNA-seq, ChIP-seq needs to be performed individually for each TF, which makes it both time-consuming and costly. Often, there is no ChIP-seq data available for the organism, tissue or cell line of interest. The reconstruction of GRNs based on gene

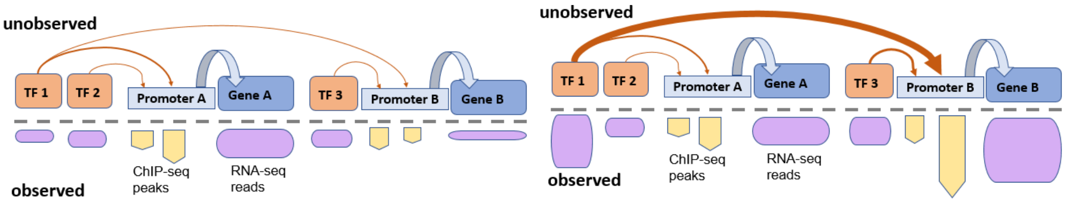


Figure 17 Exemplary representation of the regulation of two genes A and B by three TFs in two different states. The actual dependencies (depicted by the arrows) are unknown, we can only observe the expression of genes and the binding of TFs in the promoter region of genes (visualized at the bottom in purple and yellow, respectively). We can see an increase in expression of TF1 in state 2, however only the expression of gene B is affected although both genes are regulated by TF1. When considering the ChIP-seq data, we observe differential binding of TF1 in the promoter of gene B, but not in gene A, which is consistent with gene expression changes.

expression data only is thus of great interest and has been one of the most widely studied problems in the last decade [37].

3.4.1 Estimation of transcriptional regulation using expression data

Algorithms exploiting coexpression of genes are popular for mapping TF networks from gene expression data. A simple approach based on mutual information between the expression profiles of TFs and potential target genes was introduced by Faith et al. in 2007 [37]. The group used 445 E.coli Affymetrix microarray expression profiles collected under various conditions to identify transcriptional regulatory interactions which were then validated with the RegulonDB database and ChIP experiments. The predictions were performed using the context likelihood of relatedness (CLR) algorithm, which is an unsupervised network inference method and uses mutual information to score the similarity between the expression levels of two genes in a set of microarrays. When this score is above some set threshold, a gene and a transcription factor are predicted to interact. Mutual information is a metric that detects statistical dependence between two variables and is thus similar to correlation. However, it does not assume linearity and is therefore able to detect regulatory interactions that might be missed by linear measures such as the correlation coefficient. An important step of the CLR algorithm is to apply a background correction step in order to eliminate false correlations and indirect influences. To achieve this, the statistical likelihood for each mutual information between a TF and a potential target gene is calculated within the network context by comparing the mutual information value to the background distribution of mutual information scores for all possible TF-gene pairs including either the TF or its target gene. The most probable interactions are thus the ones whose mutual information scores are significantly higher than the background distribution of mutual information scores. Faith et al. applied the CLR algorithm to the 4,345 E.coli genes on the 445 microarrays, which included 328 known or predicted TFs. They predicted 1,079 regulatory interactions with 60% precision; 426 of those were also predicted at 80% precision. Of the 1,079 predicted interactions, 338 were contained in RegulonDB. To validate some of the 741 de novo predictions, the authors performed ChIP-seq on three TFs (Lrp, PdhR, FecI) with substantial connectivety in the network mapped by CLR. In total, 244 genes were tested for interactions with at least one of the TFs, 21 of these could be validated as regulatory interactions. However, there is still a large fraction of predicted interactions that could not be validated. Furthermore, only a small part of the interactions present in RegulonDB were detected, which can be explained by the fact that high recall can only be obtained if both the transcription factor and its target gene are adequately perturbed in the dataset. The number and diversity of available expression profiles has thus a great influence of the performance of the algorithm [38] to predict the network. Moreover, the approach is very context-dependent. As we will see later, even different human cell lines show big differences in the TF binding dynamics, thus it is difficult to predict one network for an organisms that contains all valid regulations.

3.4.2 Exploiting DE for investigating transcriptional regulation

Another class of algorithms for mapping TF networks from gene expression data are those exploiting differential expression (DE). These algorithms typically construct networks with regulatory edges from each TF to all genes being DE when the TF is knocked out or otherwise perturbed. Compared to algorithms based on coexpression, this approach has the advantage of less relying on chance when predicting interactions as DE creates variation in the TF expression by direct experimental intervention. Thus, one would be more confident that observed dependencies are actually causal and not just a correlation of expression levels [37]. Pinna et al. published a method based on DE in 2010 and presented their results using data of the DREAM4 *In Silico* Network challenge. The Dialogue for Reverse Engineering Assessments and Methods (DREAM) is an international initiative aiming to evaluate methods for biomolecular network inference in an unbiased way and organises yearly competitions for teams all over the world [39]. Pinna et al. participated in the second challenge of 2009, which asked to infer directed GRNs from simulated data. The data comprised a gene subnetwork including 100 genes from E.coli and S.cerevisiae and includes both a *wild-type* file and a set of *knockout* files with simulated biological and experimental noise. While the *wild-type* data contained the expression levels of all genes in an unperturbed network, each *knockout* set contains expression values for the same set of genes where exactly one gene in the network is deleted. The aim of the challenge was then to predict the directed network structure. The authors describe that was very easy to infer a causal influence network from this kind of data, as genes whose expression changes as a result of single-gene knockout are likely to be downstream of the perturbed gene. However, some of the identified edges of the network may be indirect, which happens when one TF (tf1) regulates another TF (tf2) that is regulating some target gene (tg). An edge directly from tf1 to tg might be predicted, although it does not exist in reality. To distinguish direct from indirect relationships, the group developed an algorithm which estimates the confidence of each possible edge directly from the knockout data (i.e. quantifies how likely it is for the gene to be downstream of the perturbed gene) and then refines the prediction by down-ranking the feed-forward edges. The authors tested four different confidence matrices W in which elements (i,j) reflect the confidence in the existence of the edge (i.e. regulation) from gene i to gene j. The deviation matrix W^D contains the confidences estimated by the absolute value of the deviation from wild-type expression of gene j after the knockout of gene i, where higher deviations induce higher confidence of j being downstream from i. The normalised deviation matrix W^{ND} contains relative deviations, which are derived by normalising each column of W^D by the corresponding wild-type. Matrix W^{ZD} contains the Z-scores of the deviation matrix. Finally, W^{ZR} contains the Z-scores of the raw steady-state values after each knockout instead of the deviation from wild-type. After obtaining the initial network described by the deviation matrices, edges are removed if there is at least one additional path between the genes. This is achieved by reducing the confidence in the existence of a direct edge if the causal effect can also be explained by additional paths (see figure 18). Table 2 shows the results of the network prediction on the DREAM4 data set as average AUROC values. W^{ZR} could best predict the network structure and has an AUROC of 0.83. The algorithm of Pinna et al. was awarded the best overall performance at the 100-gene network subchallenge, thereby showing that the method is effective in inferring medium-size GRNs [40]. However, it is not clear how such an algorithm would perform on a genome-wide level, especially for organisms like human, which has many more genes than E.coli and S.cerevisiae, making it very hard to obtain knockout samples for each gene. Moreover, the algorithm was used with an *In Silico* similuted data set, i.e. the simulated regulation rules had to be fairly simple. We can not expect that these rules also hold for actual gene regulation: this is the issue we wanted to investigate in our project which is described in detail in the following.

3.4.3 Combining RNA-seq and ChIP-seq to estimate gene expression

The inference of GRNs on the basis of gene expression data is a widely studied problem, with several approaches available. However, the integration of ChIP-seq data for mapping TFBS and mRNA expression data as well as data from other transcriptional and post-transcriptional regulators can improve methods for inferring GRNs. ChIP-seq offers a great advantage for obtaining information about TFs compared to RNA-seq: With the latter, we only know the expression value of a TF, but we do not know which gene promoters are bound by this TF. Instead, ChIP-seq provides genomic reads instead of transcriptomic reads; these can be mapped to the binding sites of the respective TF and we therefore obtain information about which genes are regulated by the TF, together with a signal of the binding strength. In addition, the binding motifs of TFs can be identified, which provides further information about the TF-gene interaction. The task of combining RNA-seq and ChIP-seq in order to induce GRNs is especially challenging for complex organisms and to date, we are far from inferring realistic quantitative models of genome-wide regulatory networks. However, one can reveal the main interactions and some of the most relevant players in order to refine sub-networks for specific functions. In 2014, Dunn et al. generated all possible networks that could explain stem cell self-renewal and used formal verification procedures and Boolean network formalisms to select a core network consisting of only 12 TFs and 16 interactions [41]. This suggests that stem cell self-renewal relies on a relatively low number of TFs and regulatory interactions. Still, it remains a challenge to explain and predict gene expression on the basis of the coordinated binding of TFs [42]. In the following paragraphs, two different algorithms for the prediction of gene expression on the basis of ChIP-seq data of TFs are described. Additionally, present limitations and future perspectives of the general approach are discussed.

Prediction of absolute and differential gene expression using log-linear regression One of the first attempts to address the prediction of gene expression on the basis of ChIP-seq and RNA-seq is represented by the work of Ouyang et al in 2009. The group uses log-linear regression where gene expression is regarded as the response variable and TF-related features serve as predictors [42]. They used mouse ESCs with genome-wide RNA-seq data as well as ChIP-seq data for 12 TFs. First, Ouyang et al. constructed the TF association strength (TFAS) for each gene, which is the weighted sum of the corresponding ChIP-seq signal. The closer the peak is to the TSS of the gene, the higher the associated weight. The TFAS are used as predictor variables, however this approach involves a problem: If multiple TFs cooperate to regulate many genes, one can expect that the respective TFAS vary in a coordinated manner across different genes, meaning they represent correlated predictor variables which could lead to highly unstable coefficients in the fitted model. The authors address this issue by presuming several types of coordinated interactions among TFs that are relevant for the regulation of many genes. This means that the TFAS for the target genes

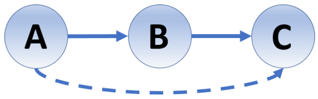


Figure 18 Feed-forward loop consisting of three genes. The edge between A and C might be predicted but should be removed, as the causal effect of gene A on gene C could be predicted by the indirect path through gene B.

	W^D	W^{ND}	W^{ZD}	W^{ZR}
AUROC	0.7844	0.7927	0.8275	0.8297

Table 2 Average AUROC for the prediction of the different 100-gene networks from the DREAM4 *In Silico* benchmarks, calculated with four different confidence matrices by Pinna et al.

under such a type of coordinated regulation should show a characteristic pattern that can be extracted through unsupervised learning from the set of TFAS vectors. In this approach, each gene has a TFAS vector whose ith component is the TFAS between this gene and the ith TF. The group then used principal component analysis (PCA) to extract uncorrelated patterns of the TFAS vectors, called TF principal components (TFPCs). This resulted in 12 TFPCs, which were subsequently used as covariates for log-linear regression on gene expression. The following model was used to predict the absolute expression Y_i of gene i:

$$logY_i = \mu + \sum_{j=1}^{M} \beta_j Xij + \epsilon_i \tag{1}$$

Here, μ is the basal expression, Xij is the score of the jth TFPC on gene i, β_j is the regression coefficient of the jth TFPC and ϵ_i is a gene-specific error term. An important advantage of this approach is that the same TF can have a positive coefficient in one selected TFPC and a negative coefficient in another. This makes it possible to model different regulatory effects on different genes. After predicting gene expression, the predicted values were compared to the expression values obtained by RNA-seq (see figure 19A). This gives a Pearson correlation coefficient (PCC) of r=0.806. The square of PCC (R^2, coefficient of determination) describes the proportion of gene expression variation that can be explained by the model; here, R^2 is 0.65 thus it is possible to explain 65% of the variation in gene expression by just 12 TFs.
The group also studied how DE genes are regulated by TFs by comparing the ESCs to embroid bodies (EBs). The experimental measurements detected 668 genes highly expressed in both cell types (Uniform High), 838 genes lowly expressed in both (Uniform Low), 782 gene upregulated in ESCs (ES Up), and 831 genes downregulated in ESCs (ES Down). The aim was to identify quantitative rules of TF binding governing the regulation of differential gene expression. To achieve this, they first plotted the genes in the TFPC1-TFPC2 plane as shown in figure 19B. Although there is a great extent of overlap for the different gene categories, Ouyang et al. stated that they see clear clusters, suggesting that the 4 gene sets are regulated by different TF combinations. To identify rules for these TF combinations, the authors first selected the 3 TFPCs most explaining gene expression variation (which are the TFPCs 1, 2, and 11) and then applied the Classification and Regression Tree (CART) algorithm based on these 3 TFPCs to discriminate the 4 gene sets. This resulted in a binary classification tree which implements the regulatory rules as combinations of TFPCs (see figure 19C). The misclassification error rate of random guessing the class of a gene is 75%. With the learned tree, this rate could be reduced to 37.1%. Although the tree manages to shrink the misclassifiaction error to about a half compared to random guessing, more than a third of the genes are still not correctly classified. Table 3 shows the predicted versus actual class labels for the genes. While uniform high and uniform low genes could be separated almost perfectly, the algorithm had big difficulties for discriminating up- and downregulated genes [43].

Prediction of gene expression using a nonlinear model Another approach for the prediction of gene expression on the basis of TF binding was presented by Cheng and Gerstein in 2012. In addition, they also tested their model for histone modifications (HMs) as predictors and concluded that both approaches are highly predictive of gene expression. In their paper, they also referred to the approach of Ouyang et al. but criticised the application of a linear model. They also worked with mouse ESC cells and used the same 12 TFs as Ouyang et al., making it easier to compare their results. As Ouyoung, Cheng and Gerstein also took the spatial effect of TF binding into account, however they also incorporated the transcription terminal site (TTS) into their analysis. First, the DNA regions 4kb upstream and downstream from the TSS and TTS were separated into bins of 100 bp size, resulting in a total of 160 bins per gene. Then, the coverage of the ChIP-seq signal was calculated for each nucleotide and then averaged over the 100bp for each bin. This was done for each TF, resulting in a TF-binding matrix for each bin, containing the TF-binding signals for all genes in the corresponding bin. For each TF, the signal was averaged over all genes separately

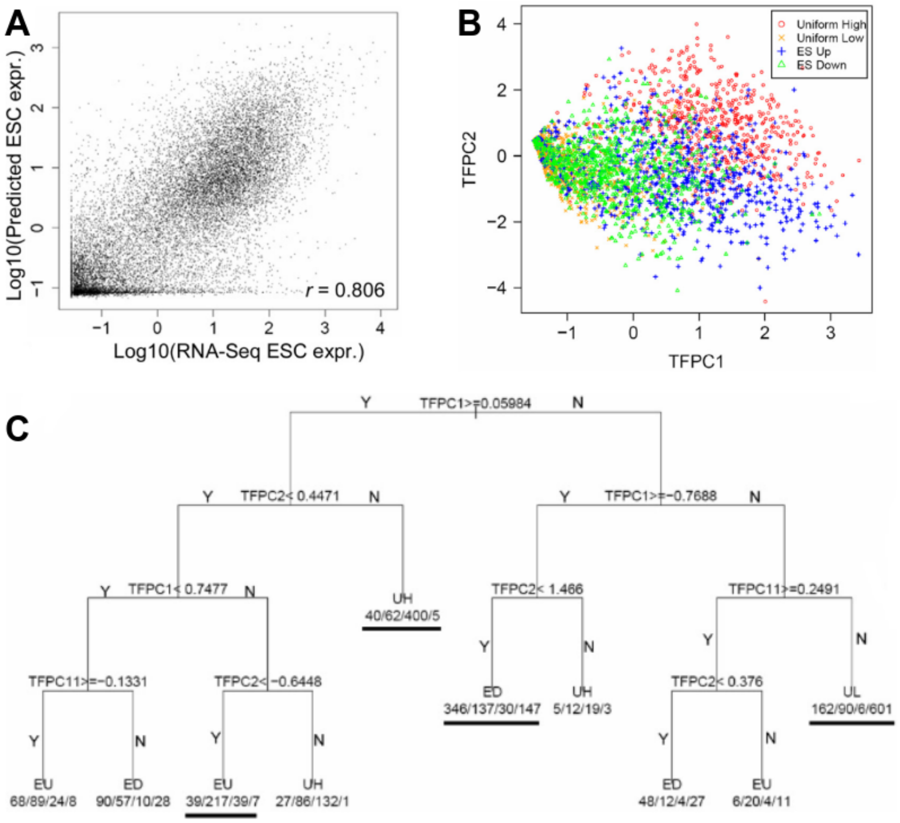


Figure 19 (A) Experimentally measured mouse ESC gene expression versus expression predicted by Ouyang et al. (B) 4 defined gene sets (uniform high expressed, uniform low expressed, upregulated in ES, downregulated in ES) are plotted in the first two TFPCs previously identified. (C) Binary classification tree for predicting differential gene expression based on the combination of 3 TFPCs. At each split, the classification criterion is given, where Y indicates Yes and N indicates No. At each terminal node, the predicted category and the actual numbers of genes in the four categories are indicated according to the order of ES Down (ED), ES Up (EU), Uniform High (UH), and Uniform Low (UL).

	ESC Down	ESC Up	Uniform High	Uniform Low
ESC Down	*484*	113	72	162
ESC Up	206	*326*	160	90
Unifom High	44	67	*551*	6
Uniform Low	202	26	9	*601*

Table 3 Predicted by Ouyang et al. (on the upper side) versus actual (on the left side) gene expression sets. The misclassification error rate is 37.1%.

in each bin in order to obtain the signal profile of the TF in the 160 bins. To predict gene expression, the authors applied support vector regression (SVR, a supervised machine learning method allowing also nonlinear models) to each of the 160 bins. In each bin the signals (i.e. the mean coverage of the ChIP-seq signal of the 100 bp in the bin) for the 12 TFs were taken as predictors for the gene expression levels measured by RNA-seq. They calculated the PCC between the predicted and experimental measured expression values with 10-fold cross validation. Figure 20A shows that the highest predictive power was achieved at the TSS, which individually accounts for about 50% of the variation of gene expression. With growing distance from the TSS, predictive power decays quickly. The authors described that the performance of their model was higher for the integration of binding signals of the TFs at different locations and not separately for each bin. This achieved a PCC of 0.77 (R^2=0.59) between predicted and measured data (see figure 20B). When using microarray data, the prediction accuracy dropped to a PCC of 0.69. Although Cheng and Gerstein claimed that their model predicts gene expression better than the approach of Ouyang et al., the PCC of their prediction (0.77) is lower than the one of Ouyang et al. (0.81). However, when comparing the scatter plots of both models (figure 19A and figure 20B), the one for Cheng and Gerstein looks much more reasonable. Ouyang's plot seems to contain much more variation, a possible explanation of the high PCC is that their plot contains a big amount of points in the lower left corner, which means that the genes are not expressed. Not eliminating these points might bias the result and significantly improve the correlation.

The authors also used a set of 7 HMs to predict gene expression using the same SVR model. Both approaches (based on TFs and HMs) seem to be predictive for gene expression. However, Cheng and Gerstein stated that the combination of both signals could not lead to a big improvement of the predictions, suggesting a redundancy of the models to predict gene expression. Figure 20C shows the correlation between TF-binding signals and HM signals (PCC=0.85). Interestingly, the models show distinct differences in the spatial patterning of their predictive strength: while TFs best predict gene expression in a small region around the TSS, HMs have high predictive powers across a wide region around genes.

The authors also investigated the possibility of predicting differential gene expression between cell lines. As they did not have enough data available to investigate this for TF binding, they only conducted this analysis based on the HM data. They first calculated the difference of the HMs between the ESCs and neural progenitor cells (NPCs) and used them to predict the DE as a log2 fc. They also used the previously described SVR approach and achieved a correlation of 0.58 between the predicted and experimental determined log ratios (see figure 20D. Thus, DE is more challenging to predict compared to the expression level: While HM-based models could explain 68% of the variation in expression levels in ESC, only 34% of variation could be explained for DE between ESC and NPC [8].

Possible limitations and future perspectives Genome-wide omic data have helped researchers to acquire a deeper understanding of many biological aspects. However, to date there is still limited use of the available data, the association between different epigenetic features and genes is still mainly done based on their proximity with respect to the TSS and existing methods only account for local interactions. Although methods for generating genome-wide maps of long-range chromatin interactions do exist (e.g. Hi-C), they have not been integrated in the previously described inferential models. Angelini et al. also suggest integrating chromatin accessibility data such as DNase-seq, DNA regions associated with regulatory activity (FAIRE-seq) and DNA methylation data in order to improve the predictions and reduce the number of false positive relations.

Interestingly, Ouyang et al. and Cheng and Gerstein demonstrated that relatively few TFs are sufficient to explain gene expression quite accurately. This apparent redundancy has only been described with regard to gene expression levels, without considering alternative splicing and differential isoform abundance. Both approaches described previously only use one TSS per gene, although there often exist multiple TSS if transcripts have alternative start sites. Thus, the observed redundancy could partially account for a different layer of complexity that has only been poorly explored until now.

The non-differential approach also has the following disadvantage: Some TFBS might always be occupied and have thus no influence on the change of gene expression between two states. This issue does not depict a problem in the differential approach, as these sites will not be included in the model.

Finally, it is important to note that despite the possibility to predict gene expression using few epigenetic features, we cannot infer causal relationships directly from such methods. To determine whether causal relationships exist or TFs only represent a code requires developing causal inference that has only received limited attention until now [42].

3.5 Transcription factor binding dynamics during human ES cell differentiation

Tsankov et al. (2015) presented a report describing their integrative analysis of genome-wide binding data of 38 TFs and extensive epigenetic and transcriptional data. The project aimed to investigate regulatory interactions and dynamics across the differentiation of human embryonic stem cells (HESCs) to the three germ layers endoderm, mesoderm and ectoderm

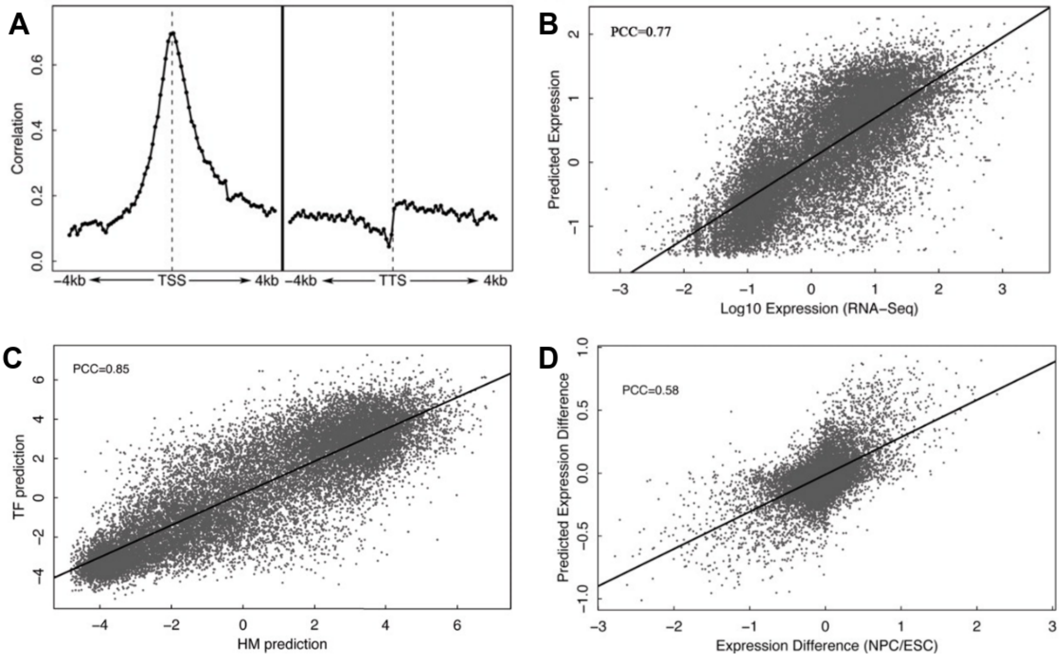


Figure 20 (A) Prediction accuracy of each of the 160 bins around TSS or TTS based on SVR on the binding signal of 12 TFs. (B) Experimentally measured mouse ESC gene expression versus expression predicted by Cheng and Gersetin. (C) Comparison of gene expression values predicted on the basis of TFs and HMs. (D) Actual expression difference between ESC and NPC versus fold changes predicted by Cheng and Gerstein.

(see figure 21A). In addition to the description of core regulatory dynamics, lineage-specific behaviour of individual TFs was investigated. Tsankov et al. also included ChIP-seq data for four core HMs (H3K4me1, H3K4me3, H3K27Ac, H3K27me) and RNA-seq of polyadenlated transcripts. In the following, we summarize their key findings showing the context-dependent rewiring of TF binding and the epigenome during HESC differentiation [44].

Classes of TF dynamics Tsankov et al. analysed the TF binding dynamics by grouping them into four classes, where static TFs do not change their binding patterns in two states, dynamic TFs show a similar amount of binding sites but at different locations, enhanced binding TFs bind more sites in the first state than in the second, and suppressed TFs show a contrary pattern. Although the germ layers exhibit unique expression signatures, they show overall only limited transcriptional dynamics. However, a small number of TFs shows dynamic binding between two (e.g. GATA4) or more (e.g. SMAD4) states. The classes of TF binding dynamics were further subdivided as temporal (between successive timepoints) or cross-lineage (between germ layers), showing that many TFs exhibit temporal and cross-lineage dynamics [44].

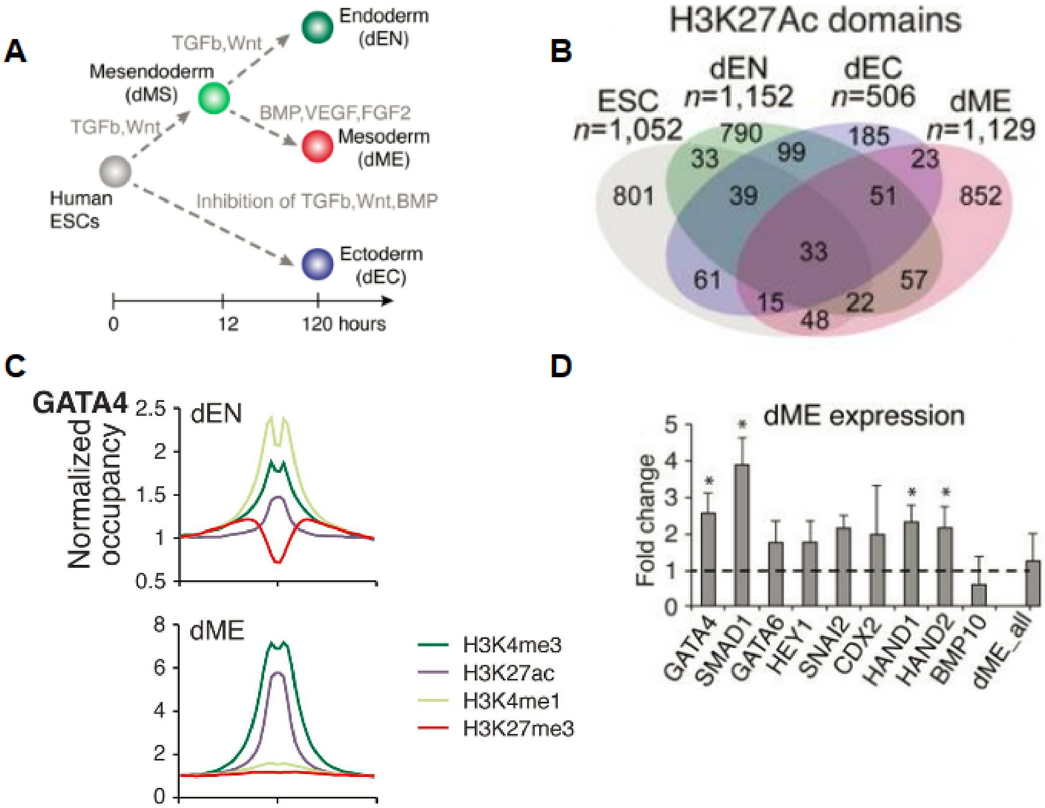


Figure 21 (A) Human ESC differentiation into the three germ layers endoderm, mesoderm and ectoderm. (B) H3K27Ac domains are predominantly unique to each cell type and can possibly be used to identify super-enhancers enriched for the binding of master regulators. (C) GATA4 is associated with dynamics of H3K4me1 in endoderm and H3K27Ac in mesoderm. (D) As a result GATA4 knock down, seven mesodermal key factors were downregulated.

Super-enhancers Super-enhancers are defined as extended H3K27Ac domains and have been used to describe regulatory regions enriched for binding sites of master TFs in specific cell lines [45]. Tsankov et al. found that GATA4 binding in mesoderm indeed coincides with long regions of H3K27Ac close to several mesodermal genes. They ranked identified H3K27Ac regions in order to identify such super-enhancers, which turned out to be predominantly unique to each cell type (see figure 21B). Furthermore, core regulators of HESC such as OCT4, SOX2, OTX2 and NANOG binding is highly enriched at super enhancers. Thus, the group used enrichment of binding at super-enhancers for identifying possible master regulators in the germ layers, which resulted in the identification of several core regulators of ESC as well as some additional TFs (e.g. EOMES, T and FOXA1) [44].

Poised enhancers Tsankov et al. could hardly identify any known endoderm TFs in H3K27Ac domains, thus they analysed if such regulators are instead present at regions being enriched for another HM. H3K4me1, which is also known to form extended enhancer domains not overlapping with H3K27Ac, was used as in the previously described approach to identify TF binding in endoderm cells. Extensive H3K4me1, which are called poised enhancers, were enriched for a different set of TFs as in ESC, including GATA4, GATA6 and SOX17, which are known early endodermal regulators [44].

Influence of GATA4 on histone modifications GATA4 is associated with dynamics of H3K27Ac in mesoderm and H3K4me1 in endoderm (see figure 21C). A reasonable explanation for this finding is that GATA4 motif instances are higher enriched at SMAD1 binding sites in mesoderm than in endoderm, as SMAD1 is known to interact with histone acetyltransferases. For further analysis, shRNA knockdown (KD) was performed for GATA4 prior to gene expression measuring in mesoderm and endoderm. Although the KD in endoderm did not greatly affect any measured endodermal TFs, it leads to a 1.7-4-fold reduction in the expression of 7 key factors in mesoderm (see figure 21D). Interestingly, H3K27Ac super-enhancers in mesoderm were largely unaffected by the KD [44].

Dependency of TFs and DNA methylation Some TFs can modulate DNA methylation levels, which leads to silencing of genomic regions. However, it is not generally known which TFs can alter methylation and which ones are sensitive to its presence [46]. Tsankov et al. also observed the association of DNA methylation loss with lineage-specific binding of several TFs and subsequently performed global enrichment analysis for TF binding at regions with a change of DNA methylation. This revealed that many target sites of NANOG, SMAD1 and TCF4 show a gain of methylation in all germ layers consistent with silencing of their pluripotency-related target genes. Moreover, the found a frequent reciprocal gain in DNA methylation in the different germ layers of key endoderm and ectoderm TFs [44].

3.6 Project

Tsankov et al. described the dynamics of TF binding and the interplay of TFs with epigenetic modifications [44]. However, they did not investigate the dependency of gene expression and TF binding. Using their data, we want to extend their analysis by incorporating RNA-seq for the HESC and germ layers and combining it with the information obtained from ChIP-seq. To do this, we first try to replicate part of their preprocessing steps for the ChIP-seq data of the 38 TFs in order to achieve consistent results. As we are interested in the regulatory dependencies between genes and TF, we use a differential approach here: we first determine differentially binding TF peaks for all combinations of 2 states and identify DE genes for the same state pairs. Then, we combine the data to analyse the relation between DE genes and differentially binding TFs regulating these genes. Pluripotent stem cells constitute a powerful system to learn to understand the molecular mechanisms underlying cell differentiation. It is known that TFs orchestrate the overall remodelling of the epigenome, including loci that will change expression only at later states. In this manner, lineage-specific TFs play an important role in exiting pluripotency and activating cellular specification [47]. On the basis of this data set, our aim is to find out how complex gene regulation is. To explore this issue, we investigate how big the dependencies between differential TF binding and differential gene expression is, where small or absent dependency means that gene expression cannot be well explained on the basis of TF binding alone. In contrast to the different approaches described in the beginning, we do not try to predict gene expression based on TF binding. Instead, we develop an interactive visualisation of the detected dependencies and inconsistencies in a detailed and reproducible manner. Such a tool could be used for validation of prediction methods by providing an interactive and user-friendly representation of the data which not only includes the final results, but also the raw data underlying each finding. In the following, the different parts and results of the project are described in detail. The used data and respective results of each step, including much more detailed information, can also be viewed in the interactive view (IV) of this chapter (see the overview page IV0 and figure 22).

3.6.1 Available data

ChIP-seq data was obtained from Tsankov et al. over the Sequence Read Archive (SRA, id: SRP047193) [44]. The data set comprised 200 ChIP-seq experiments profiling 38 TFs and 6 chromatin marks in 5 human cell types (male human embryonic stem cell line HUES64, directed differentiation of HUES64 towards mesoendoderm (dMS), endoderm (dEN), mesoderm (dME), and ectoderm (dEC) and contains both single-end and paired-end samples. In addition, some cell lines were derived with shRNA mediated knockdown of GATA4. As we did not include the samples for the chromatin marks and GATA4-knockdown, 152 ex-

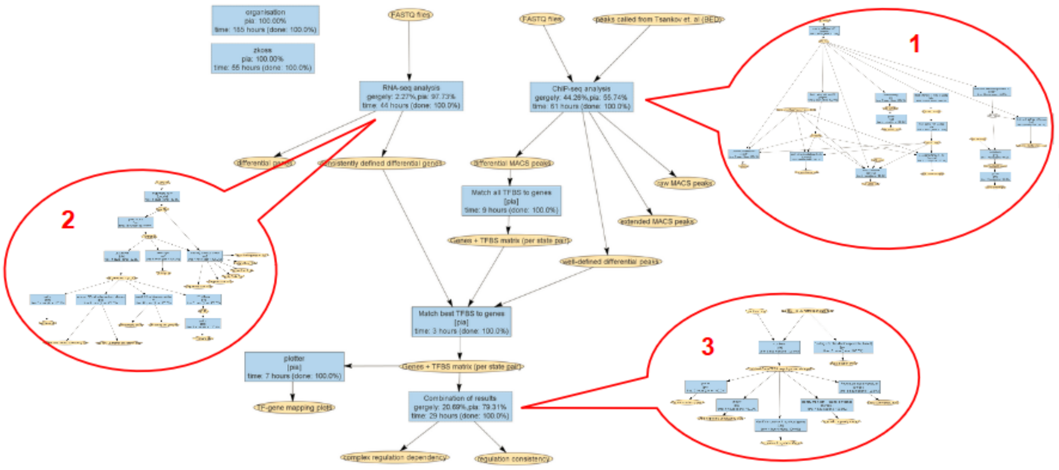

organisation
pia: 100.00%
time: 185 hours (done: 100.0%)
zkoes
pia: 100.00%
time: 55 hours (done: 100.0%)
FASTQ files
FASTQ files
peaks called from Tsankov et. al (BED)
RNA-seq analysis
gergely: 2.27%,pia: 97.73%
time: 44 hours (done: 100.0%)
ChIP-seq analysis
gergely: 44.26%,pia: 55.74%
time: 61 hours (done: 100.0%)
differential genes
consistently defined differential genes
differential MACS peaks
raw MACS peaks
Match all TFBS to genes
[pia]
time: 9 hours (done: 100.0%)
extended MACS peaks
Genes + TFBS matrix (per state pair)
well-defined differential peaks
Match best TFBS to genes
[pia]
time: 3 hours (done: 100.0%)
plotter
[pia]
time: 7 hours (done: 100.0%)
Genes + TFBS matrix (per state pair)
TF-gene mapping plots
Combination of results
gergely: 20.69%,pia: 79.31%
time: 29 hours (done: 100.0%)
complex regulation dependency
regulation consistency
1
2
3

Figure 22 Snapshot of the interactive workflow of our project (see IV0) containing three sub networks: (1) ChIP-seq preprocessing and analysis for TFs (2) RNA-seq preprocessing and analysis for genes (3) Combination of the two previous parts and analysis of gene regulation complexity. All orange states are linked to html sites visualizing the analysis results.

periments were used for our analysis. In addition to the raw data in fastq format, we also downloaded the peaks called by Tsankov et al. from Gene Expression Omnibus (GEO, id: GSE61475), which were called by the authors after preprocessing of the raw ChIP-seq data. As there was no expression data available from Tsankov et al., RNA-seq data was obtained from the Roadmap Epigenomics Project over SRA (id: SRA009256). We selected samples for ESC and the three differentiated germ layers (endoderm, mesoderm, and ectoderm) to get expression data for the same states as for ChIP-seq.

3.6.2 Preprocessing of ChIP-seq data

A major issue concerning the available ChIP-seq data is that for most TFs, there were not sufficient replicates available: For 82% of the TF-state combinations, there was only 1 replicate. For 14%, 2 replicates were available and only 5 combinations had 3 or 4 replicates. Furthermore, there were 62 combinations of a TF and state with no available experiment; four TFs (CDX2, HAND2, HEY1, PAX6) were only analyzed for one of the 5 states (see IV1). Obviously, Tsankov et al. also had to deal with this problem, however they did not mention the issue concerning the low number of available replicates and the resulting uncertainty of the results at all.

3.6.2.1 Read mapping

The ChIP-seq reads from the available fastq files were mapped to hg19 with Bowtie2. The number of reads differs significantly between the different samples and ranges from 2.8 million to 61.5 million. The same holds for the percentage of mapped reads, which ranges from 31.1% to 98.9%, however the vast majority of samples shows a percentage of at least 75% for mapped reads. To analyze the position of mapped reads in a genomic context, the reads were annotated to promoter, exonic, intronic and intergenic regions based on their positions. Although we would expect that most of the TF bindings take place in the promoter region, and thus most of the reads measuring these binding sites would map to promoters, less than 10% of the mapped reads were annotated to this region, which is still slightly more than the portion of reads mapped to exonic regions, but significantly less than the percentage of reads being mapped to intronic and intergenic regions (see figure 23A). However, one has to consider that promoters only constitute a very small fraction of the whole genome. When normalizing for this trait and computing the log fc to the background distribution, reads in promoter regions are enriched, while exonic reads are slightly underrepresented and reads in intronic and intergenic regions are significantly underrepresented (see figure 23B). When investigating intergenic reads, it becomes apparent that the majority of those reads are not in the proximity of a gene. Less than 15% of intergenic reads have a distance of maximal 10kb, while reads with a distance of at most 50kp make up 30%. Almost 60% of the intergenic reads have distance greater than 200kb to the closest gene, making it very difficult to annotate peaks called from those reads to specific genes (see figure 23C). Note however, that not all of the reads will later be contained in peaks, as there is high dispersion of noise throughout the whole genome, making peak calling necessary. For detailed information about the samples, see IV1.

3.6.2.2 Read extension

While reads from ChIP-seq experiments usually have a length of 200bp, the reads in the data set from Tsankov et al. only show a length of 35bp. This is because the group performed MNChIP-seq, which is a sightly modified version of conventional ChIP-seq using a micrococcal nuclease (MNase) instead of sonification to cut the fragments, resulting in shorter reads [44]. To compensate for the read length, the authors extended the reads to 200 bp, however they did not describe the exact procedure and how they coped with single- and paired-end experiments. As we wanted to achieve comparable results, we also performed a read extension step after mapping them to the genome. Since it is not clear how to extend reads containing mismatches, we first analyzed the extend to reads which could not be perfectly aligned. See IV2 for images showing the number and percentage of reads having 0 to 5 mismatches (with 5 mismatches being the maximum number of mismatches in the whole data set) for each experiment. As only 0.5% or less of the reads per experiment were not perfectly matched, we excluded those reads from downstream analysis. For the extension of reads, we had to decide for a consistent and straightforward procedure, especially when dealing with experiments that are either single-end or paired-end. While the reads resulting from single-end sequencing could simply be extended to a total length of 200bp, we decided to transform the paired-end data to single-end data by just extending the first read of each pair.

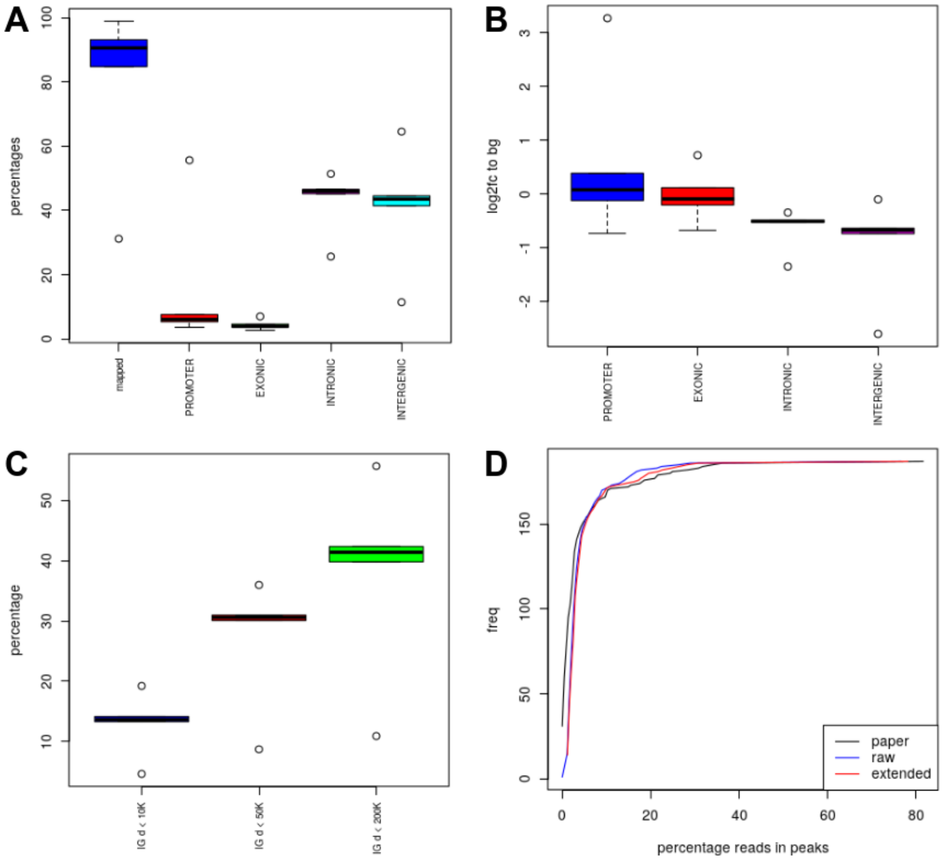


Figure 23 (A) Percentage of ChIP-seq reads mapped to different regions within and around genes. (B) Occupancy of ChIP-seq reads at different regions within and around the gene, normalized with the background distribution of mapped reads. (C) Distance of intergenic ChIP-seq reads to the closest gene. (D) Percentage of mapped reads in peaks for raw reads, extended reads and peaks called by Tsankov et al.

3.6.2.3 Peak calling

The next crucial step in ChIP-seq data analysis was to obtain the TF binding sites (TFBS) in the genome by calling the peaks from the mapped reads. Like Tsankov et al., we used MACS2 for this task with a FDR cutoff of 0.05 to call significant regions [44]. For subsequent consistency analysis, we called peaks from the raw bam files as well as from the bam files after read extension.

3.6.2.4 Consistency analysis

In addition to the peaks called from raw bam files and after read extension, the peaks called by Tsankov et al. were also available. To control for consistency among the three different peak sets, we analyzed the data by different criteria.

Percentage of reads in peaks. Figure 23D shows the percentage of reads in peaks, which is generally very low. For more than 80% of the samples, maximal 5% of the mapped reads were contained in peaks. The distribution is very similar for the raw and extended type, while the percentage of reads in peaks is even smaller for the peaks mapped by Tsankov et al. (see IV3 for detailed information of the percentages for individual samples).

Peak consistency for different preprocessing types. As we have both single-end and paired-end data in our set, we also wanted to analyze if this has an effect on MACS2 when performing peak calling. To test this, we also performed peak calling with MACS2 using only the forward reads of the raw data (for paired-end samples) and then repeated the procedure using only the reverse reads. When comparing the peaks called with forward reads and raw read data, more than 90% of the peaks are overlapping. However, when comparing peaks resulting from forward reads with those coming from reverse reads, the percentage of overlapping reads drops to 60-75%. For the pairs reverse vs. raw, forward vs. extended, reverse vs. extended and raw vs. extended, the percentage of overlapping peaks is similar, with a mean of 65-70% (see figure 24A). The results suggest that the algorithm of MACS2 is very susceptible to the type of sequencing, which results in highly inconsistent results. Moreover, we would expect a high percentage of overlapping peaks between peaks from forward and extended reads, as only forward reads were used for the extension step. For the intersection of peaks called in two different modes, we are also interested if the peaks were identified with a similar significance. Figure 24B shows the rank-correlation of mapped peaks between two types for overlapping peaks. While the correlation between forward and raw peaks is nearly 100%, the correlation of reverse read peaks to either forward or raw read peaks as well as the correlation between extended and either of the 3 other types is again significantly lower and has a mean of around 0.75.

We also analyzed the percentage of overlapping peaks for the peaks called by Tsankov et al. compared to peaks called on raw, forward, and reverse reads (see figure 24A). For all 4 combinations, the overlap is very small, with a mean of just 20%. Interestingly, the overlap with peaks from Tsankov et al. is not higher for peaks based on extended reads, although we tried to reproduce their data by implementing the same preprocessing procedures. When computing the rank-correlation of mapped reads from Tsankov et al. compared to the 3 types of peak data sets we generated, the mean is about 0.4-0.5 for peaks from raw, forward and reverse reads and rises to about 0.6 for extended reads (see figure 24B). This suggests that read extension might actually have an influence on the performance of MACS2. In general, both the percentage of overlapping peaks and the rank-correlation of mapped peaks is relatively small for the majority of the data. Figure 24C shows the percentage of overlapping peaks for all pairwise combinations of the available peak data sets. As the same ChIP-seq data was used for all peak sets, we would expect a high similarity of all peak sets, however 50% of the samples show an overlap of peaks of at most 60%. The overall rank-correlation of mapped peaks is visualized in figure 24D, for detailed information and plots showing the correlation for each TF, see IV4. A reason for the low percentage of overlapping reads between the peaks called on the extended reads compared to the peak set from Tsankov et al. could be that it was hard to reproduce the preprocessing steps performed by Tsankov et al. in a consistent manner, as the group did not exactly specify how the reads were extended. In addition, the authors discarded peaks if they overlapped with regions that MACS2 detected as peaks in 4 different whole cell extract (WCE) samples, as those regions may cause false-positive peaks due to unannotated high copy number regions [48, 44]. As those WCE samples were not publicly available, we were not able to perform this preprocessing step. From our analysis, we assume that all steps in preprocessing as well as the peak calling settings may cause significant changes in the called peaks, which makes ChIP-seq results irreproducible to a certain extend.

Replicate consistency. After analyzing the consistency for different types of read preprocessing, we were also interested in the consistency of peaks among replicates. As mentioned previously, for most TFs and states there was just one replicate available. Here, we analyze

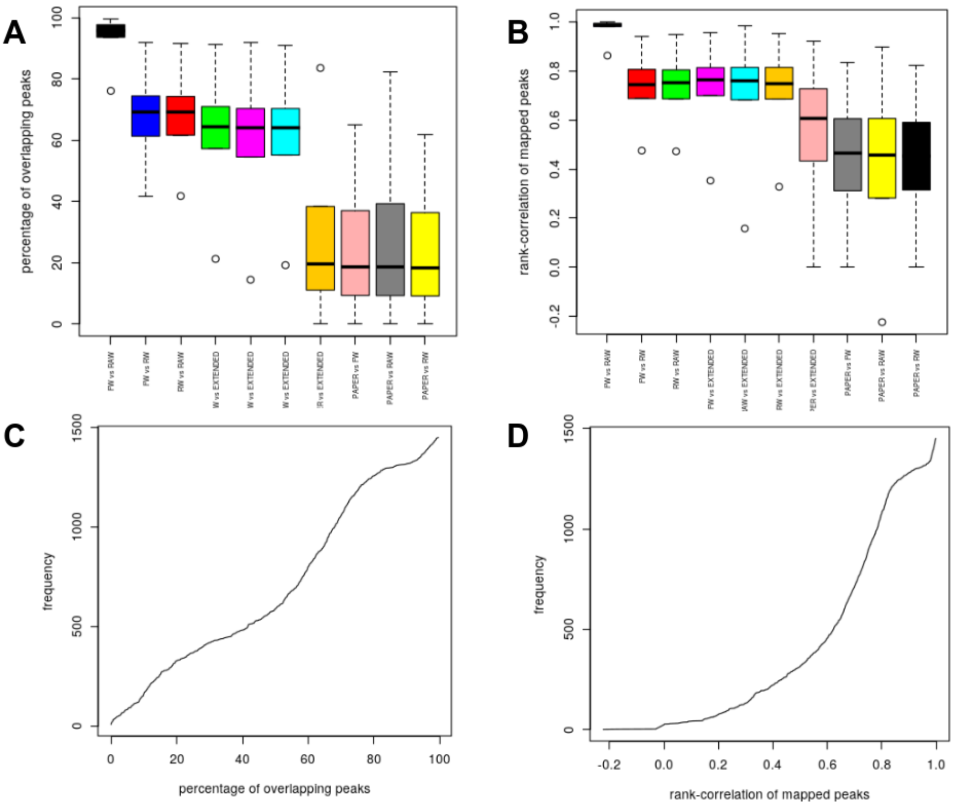


Figure 24 (A, C) Percentage of overlapping ChIP-seq peaks for all combinations of peaks called from raw reads, forward reads, reverse reads, extended peaks and peaks called by Tsankov et al. (B, D) Rank-correlation of overlapping peaks for all combinations of read types.

the TF-state combinations with at least 2 replicates, again including peaks called from raw reads and extended reads as well as the peaks called by Tsankov et al. Figure 25A shows the percentage of overlapping peaks between replicates for the three peak sets. There is even less consistency than between different preprocessing types: For more than 80% of the pairs, the percentage of overlapping peaks is below 40%, for peaks from raw and extended reads it is even lower than for peaks from Tsankov et al. Moreover, the rank-correlation of mapped peaks is at most 0.5 for 80% of the pairs (see 25B). Check IV5 for detailed information and plots for each TF and state. The inconsistency of replicates represents an essential problem for ChIP-seq analysis, as it is not trivial how to define peaks. Even for peaks with a high signal in one replicate, there is often no signal at the same location for the other replicate. To cope with the high noise in ChIP-seq data, it is recommended to use a control data file containing background signal for peak calling with MACS2. Unfortunately, we did not have an available control file. Tsankov et al. used the WCE samples for this purpose and achieved a slightly higher percentage of overlapping peaks, however the rank-correlation of those peaks was lower compared to our peaks based on extended reads.

3.6.2.5 Differential Peak calling

After mapping the peaks individually for each state, we want to identify differential binding for each TF between two states. A popular tool for this task is DiffBind, which is an R package for computing differential bound sites from multiple ChIP-seq experiments [49]. However, we cannot use this tool as we do not have enough replicates for most TFs. To overcome this problem, we decided to use MACS2 to identify differential binding. As described earlier, the method has an option to include a control bam file in the analysis of mapped peaks, where regions with reads in the control bam file are subtracted from the ChIP-seq data to be mapped. We can exploit this option by giving the bam file of one state as input and the bam file of the other state as control. Thus, the peaks in the second file will be subtracted from the peaks of the first file, leaving us with only those peaks that are upregulated in the first state. Of course, we need to repeat the procedure with the bam file of the second state as input and the bam file of the first state as control, to get the downregulated peaks as well. For the TFs with multiple replicates per state available, we

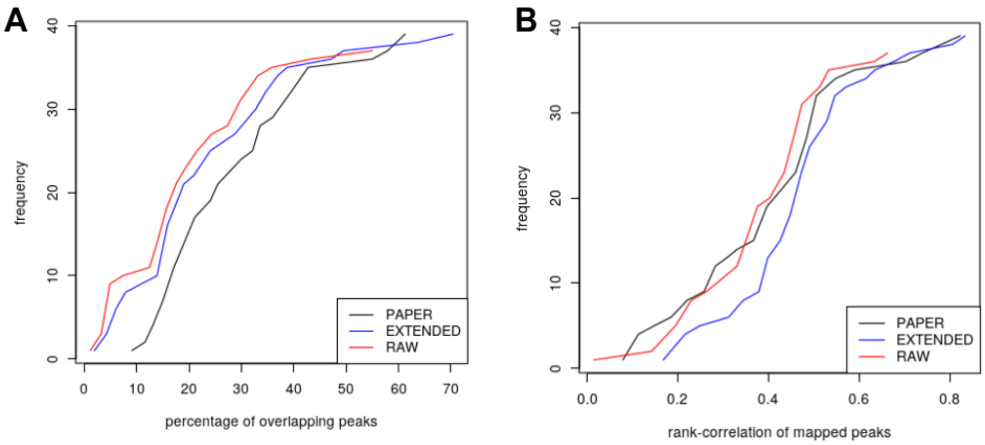


Figure 25 (A) Percentage of overlapping peaks among replicates for the same TF and state, for peaks mapped from raw end extended reads as well as peaks mapped by Tsankov et al. (B) Rank-correlation of overlapping peaks among replicates.

computed differential binding for all combinations of a replicate from the first state with a replicate of the second state. As read extension did not seem to give a reasonable advantage and it is unclear why Tsankov et al. chose to perform this step, we used the raw data for differential binding identification. As we are interested in the difference between the binding in two states, it is likely that the peaks in the raw data are more precisely defined and we avoid possible bias incorporated by read extension.

3.6.2.6 Annotation of differential TF peaks to gene promoters

In order to be able to combine the results from ChIP-seq data and RNA-seq data in downstream analysis, it was necessary to map the obtained differential TFBS (DTFBS) to a promoter region of a specific gene. However, there is no trivial definition of the exact location of the promoter that holds for every gene. In general, one supposes that the promoter region spans the transcription start site (TSS) as well as a small region upstream and downstream from the TSS, e.g. -1000bp to +200bp relative to the TSS. As there is not a predefined length for this region, we conducted our subsequent analysis for four different promoter location definitions: (1) -700bp to +150bp (2) -1000bp to +200bp (3) -2000bp to +200bp (4) -4000bp to +400bp. A further challenge is represented by the principle of alternative splicing, as there are multiple transcripts for many genes, which sometimes have different TSS. This leads to the question, if one promoter is sufficient for the regulation of all isoforms. This is unlikely if the respective TSS are not close to each other, but makes it difficult to map the DTFBS to the promoter region of a gene, as there is no single clearly defined promoter region. Another question related to this issue is, if it is possible to infer the correct promoter for a transcript based on only the expression data. This might be hard, as transcripts often overlap consistently, thus many reads are not unique for a specific isoform. We defined the promoter of a gene as the merged promoter regions for each transcript for the 4 different promoter definitions separately. Figure 26A shows the effect of the different promoter definitions, i.e. the size of the promoter on the number of DTFBS mapped to these regions. While the percentage of DTFBS that could be mapped to a unique promoter of a gene ranges from 14% to 24% depending on the promoter size, the portion of DTFBS that maps to a region which could be the promoter for more than one gene is 4-8%. This percentage varies for different TFs: For SMAD23, less than 5% of DTFBS are mapped to a promoter region, while this portion is up to 50% for SRF and THAP11 in some state pairs (see IV6 for detailed visualization of the DTFBS in promoter regions for all TFs and states). Subsequently, we analysed the effect of multiple TSS on peak mapping. For genes with at least 2 transcripts, figure 26B shows the number of transcripts per gene as well as the number of different TSS and the number of clustered TSS (i.e. proximal TSS with overlapping promoter regions). Half of those genes (about 12,000) have at least 10 transcripts, with mostly a unique TSS. Figure 26C shows why clustering of TSS often is not possible: For more than 50% of genes with multiple transcripts, the maximum number of bases between TSS is at least 10,000, with distances of up to one million bp. When investigating how many of the different TSS have associated ChIP-seq signals, it becomes apparent that half of the genes with multiple transcripts have peaks mapped to at least 5 different TSS, with up to 18 active TSS (see IV7). The ChIP signal of the second TSS is lower compared to the signal of the first TSS of a gene, however the difference is not very big (see figure 26D). This strongly suggests that one should not simplify the analysis and just consider the first TSS of a gene, as a significant amount of information were lost. IV7) contains detailed information about multiple-TSS analysis for all genes with at least two replicates. For each gene, we also provide a visualization of the transcript structure and the TF peaks annotated to the respective genomic location. This includes ChIP-seq data from raw reads, extended reads and peaks called by Tsankov et al. Figure 27A shows an exemplary visualization for a gene with multiple transcripts having different TSS. The picture also shows a general observation: While the peaks mapped from raw and extended peaks are quite similar, Tsankov et al. often identify peaks for more TFs compared to us. In addition, even peaks for the same TFs vary significantly in length and location. They generally map more peaks compared to us and identify a stronger activity for alternative TSS.

3.6.2.7 Selection of appropriate ChIP-seq sample pairs

As described previously, all DTFBS were mapped to unique gene promoter regions. IV8 contains the mapping files for each state pair and each promoter size definition. For each DTFBS, the summit and signal of the peak is annotated. As signals are not comparable for different TFs and experiments, we also computed two normalized scores for each peak (signal/max(signal) and zscore of the signal). For the TFs with multiple replicates available, we used all combinations between replicates in one state and replicates in the other state, thus generating multiple sets of DTFBS for a TF. Unfortunately, these DTFBS are often very inconsistent and it is unclear how to best merge this information. A reasonable explanation for this is the low degree of consistency between peaks among replicates, as described previously. To cope with this problem, we decided to use only the DTFBS obtained by combining

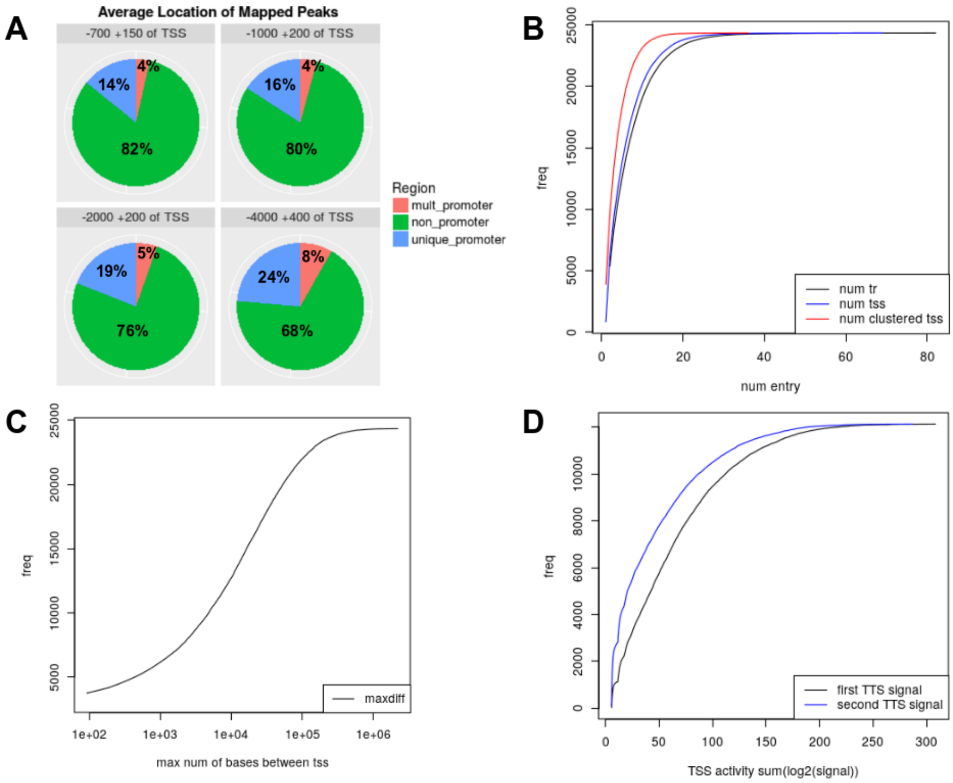


Figure 26 (A) percentage of differential peaks mapped to promoter regions unique for a gene, to promoter regions shared by multiple genes, and to non-promoter regions for 4 different promoter size definitions. (B) Number of transcripts, TSS and clustered TSS per gene for genes with at least 2 transcripts. (C) Number of genes with a certain maximum distance between different TSS. (D) ChIP signal for peaks mapped to the first and second TSS of genes with at least 2 transcripts.

the “best” replicates from each state. The appropriate pair of replicates is defined as follows: If possible, use two samples from paired-end sequencing. If this is not possible, prefer two samples from single-end sequencing over a combination of single- and paired-end sequencing. If there are still multiple possible combinations, take the sample with the highest read count from each state. This way, we obtain a unique mapping of DTFBS to the gene promoters.

3.6.2.8 Motif identification

TFs recognize short degenerate sequence motifs, which enables a context-specific binding to the regulatory region of genes. In general, the different binding sites of a TF show slight variations of the motif instead of perfect matches of a specific sequence. We would like to find out if a difference in the binding site could imply a different regulation of the respective gene. To analyze the impact of different TF motifs, we first need to obtain the motif sequences for the TFs we are investigating. JASPAR is a popular open-access database

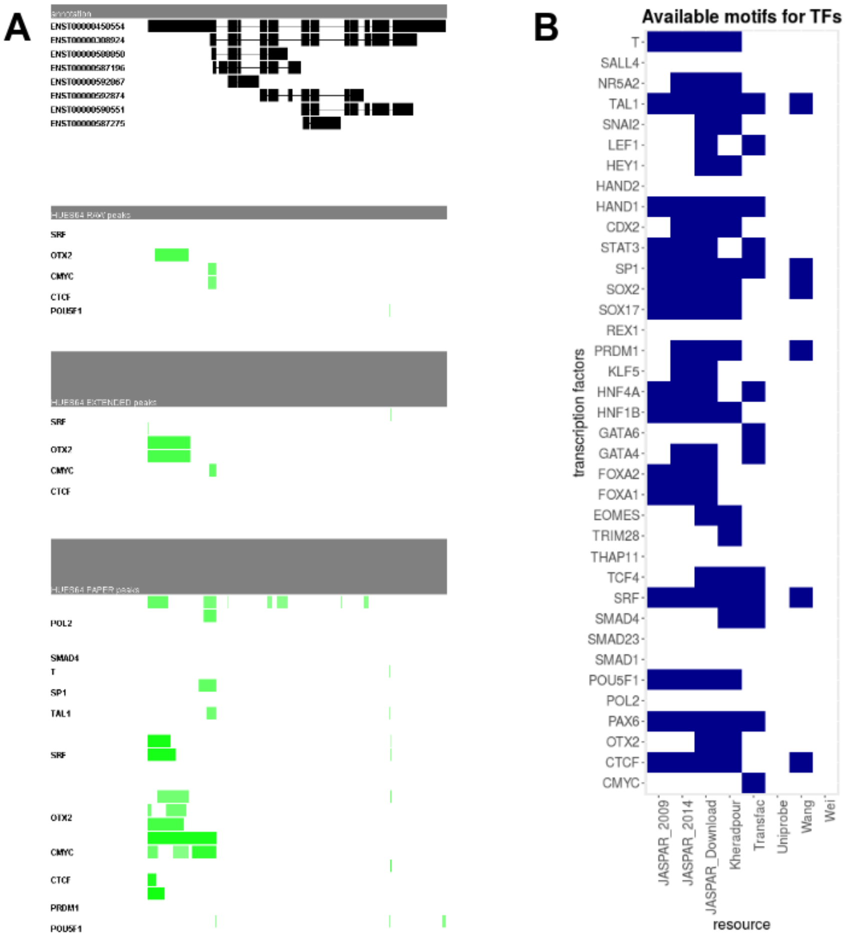


Figure 27 (A) Transcript structure and mapped peaks for the gene U2AF65. At the top, the transcripts are visualized with exons (black blocks) and introns (thin lines connecting the blocks). Beneath, the mapped peaks (green blocks) are annotated relative to the location within the gene for all TFs with at least one mapped peak. Lighter green denotes a weaker signal. Top: Peaks called from raw reads, middle: peaks called from extended reads, bottom: peaks called by Tsankov et al. (B) Available motif matrices for a specific TF in a source are marked in blue.

containing curated, non-redundant matrix models describing DNA-binding preferences for TFs [50]. The data are derived from published collections of experimentally defined TFBS for eukaryotes and are subjected to high quality standards [28]. Other databases containing TF motifs are TRANSFAC and UniPROBE. In addition to experimental validated TFBS, binding sites can be predicted with special algorithms like MEME (Multiple EM for Motif Elicitation), which is one of the most widely used tools for identifying novel signals in sets of biological sequences. Based on DNA or protein sequences provided by the user, MEME searches for repeated, ungapped sequence patterns and discovers TFBS motifs. However, this is a difficult task as binding sites are generally short and degenerate. Moreover, promoter regions are often difficult to identify precisely, especially in eukaryotes because TFBS tend to be even shorter and more variable. [51].

We consulted different resources in order to obtain sequence motifs for our TFs of interest. In addition to the current version of JASPAR, we used the versions of 2009 and 2014, respectively, as well as the data from the TRANSFAC database and UniPROBE [52, 53]. Additionally, we used the data sets of Kheradpour, Wang and Wei, all of which used MEME for the identification of TFBS [54, 29, 55]. Figure 27B shows a matrix denoting the availability of TFBS motifs in the described resources. For some TFs, there are motifs available in multiple resources (e.g. HAND1, TAL1, SRF, PAX6, CTCF). Although the current JASPAR set contains the majority of sequence motifs, certain motifs are only available in one of the other data sets (e.g. CMYC in TRANSFAC, TRIM28 in Kheradpour). Unfortunately, for 7 of the 38 TFs there is no motif available in the consulted resources (e.g. HAND2, SMAD1, POL2).

The analysis of the effect of different motifs on gene regulation was not carried out in this project and represents an interesting topic for future investigations.

3.6.3 Preprocessing of gene expression data

In addition to the preprocessing of ChIP-seq data to obtain differential binding sites of TFs, preprocessing of gene expression data constituted a major part of the project. In contrast to the ChIP-seq data, sufficient RNA-seq replicates were available to achieve reliable estimates of differential expression of genes (see table 4) . As mappers for transcriptomic data like RNA-seq generally apply more complex methods compared to genomic mappers like Bowtie2, we used 4 different mappers (ContextMap, HISAT, STAR, and TopHat2) to map the reads to hg19 in order to be less dependent on specific implementations [56, 57, 58, 59].

3.6.3.1 Mapping statistics

In order to compare the available data for different replicates, states and mappers, feature statistics were computed for each sample, including general information about mapping, data quality and prevalence of biotypes for mapped reads. Detailed visualization of diverse criteria can be found at IV9, where all samples mapped by the different mappers are compared to

State	ContextMap	HISAT	STAR	TopHat2
HESC	6	6	6	6
Endoderm	4	4	4	4
Mesoderm	6	6	6	6
Ectoderm	10	10	12	7

Table 4 Number of RNA-seq replicates per state for each mapper.

each other by the selected criterion and are sorted by either the total count of reads for this criterion or the percentage of reads mapped for this sample.
The total read count for the replicates is between 17 and 42 million, however the number of mapped reads seems to be dependent on the mapper. While STAR and ContextMap could map 59-81% of the reads, HISAT and TopHat2 only mapped 28-65%. Similar observations can be made when comparing the percentage of transcriptomic reads: ContextMap and STAR consistently mapped a higher percentage of reads to transcriptomic regions (up to 27%), while TopHat2 only mapped 10-22% transcriptomic reads. The choice of mappers also appears to have an effect on identified biotypes for the mapped reads. While 76-82% of the reads mapped by TopHat2 are for protein-coding genes, the majority of samples mapped with STAR have only about 45-60% reads for protein-coding genes, but show a large fraction of reads mapped to rRNA genes compared to other mappers. A variety of additional criteria for comparing the samples are implemented in the interactive version of this book chapter. To view the combination of different criteria for a single sample, check IV10, IV11, IV12, IV13, and IV14.

3.6.3.2 Read counting

After mapping the reads to the genome, read counts per gene were obtained in order to compute the fragments per kilobase of exon per million reads mapped (FPKM). For each mapper and state, we analysed the correlation of FPKM values for all replicates. This was done by computing the Pearson correlation coefficient (PCC) for each pair of replicates within a state for a specific mapper. Independently from the used mapper, the PCC is predominantly consistent for all pairs of replicates and ranges from 47-50% (mesoderm) to 46-59% (ectoderm). However, the PCC becomes much higher (> 90%) if genes are excluded which have a FPKM value >0 in only one of the two replicates. For a visualization of the PCC per state and mapper as well as detailed plots with all FPKM values for each pair of replicates, check IV15. The site also shows an interactive plot comparing the cumulative FPKM values for groups of 1 to n replicates (with n being the total number of replicates per state) to visualize the consistency of all available replicates. See figure 28A for the replicates of endoderm mapped with ContextMap. For small FPKM values, there is a noticeable difference between the 4 replicates; when only taking into account genes having a non-zero FPKM for all replicates the number of genes shrinks from 27.800 to about 19.600. However, for FPKM values > 1 the replicates show predominately high consistency.

3.6.3.3 DE Analysis

The next important step in the analysis of gene expression data is the identification of DE genes. Again, we used different analysis methods (DESeq, edgeR, and limma) for this task to avoid bias by a specific implementation [60, 61, 62]. DE genes were identified for each combination of two states for each mapper separately. The number of DE genes with a log2 fc ≥ 1 on a significance level of 0.05 is predominantly consistent for different mappers and DE analysis methods but differs between state pairs: about 1500 genes are DE between endoderm and ectoderm, while there are more than 4000 DE genes from HESC to mesoderm. However, in samples mapped with TopHat2, there are generally more DE identified (about 300 genes more compared to other mappers). Check IV16 for the volcano plots for each mapper and state pair. To compare the consistency of the different DE analysis methods with respect to the identified genes, we made Venn diagrams comparing the portions of genes identified by one or more methods based on the same mapper (see IV17). All three tested methods show a high level of consistency for all state pairs and mappers, especially DESeq and edgeR. Only limma tends to predict slightly more genes as DE in state pairs including ectoderm (about 140-170 additional genes, see figure 28B). Similar to DE analysis methods, we wanted to analyze the consistency of mappers. IV17 also shows Venn diagrams comparing the four mappers ContextMap, HISAT, STAR and TopHat2 for each DE analysis method separately. Compared to the DE analysis methods, the mappers show a higher level of inconsistency. In almost each state pair, the usage of a specific mapper leads to at least 100 genes that are not detected with any other mapper. TopHat2 seems to be by far the most inconsistent with respect to the other mappers, leading to a prediction of up to 665 additional DE genes (see figure 28C).

3.6.3.4 Consistency of gene regulation

After analyzing the consistency among mappers and DE analysis methods separately, the next step was to identify genes that are predicted as DE for preferably all mappers and DE methods. In addition, we are also interested in genes that are not DE, but show a constant expression rate in two states on a high significance level (thus being defined by a small fc and small p-value). In this context, we define a gene to be ’save regulated’ if it is either DE or constantly expressed with a fc $\leq$ 0.5 for all mappers and DE analysis methods. To take into account the effect of a rather arbitrary choice of a specific p-value and fc to decide if a gene is DE, we conducted our analysis with different parameters: (1) p-value $\leq$0.01, fc $\geq$1.0, (2) p-value $\leq$0.05, fc $\geq$1.0, (3) p-value $\leq$0.05, fc$\geq$ 1.5. For constantly expressed genes, the fc threshold always stays at 0.5 while the p-value threshold is defined

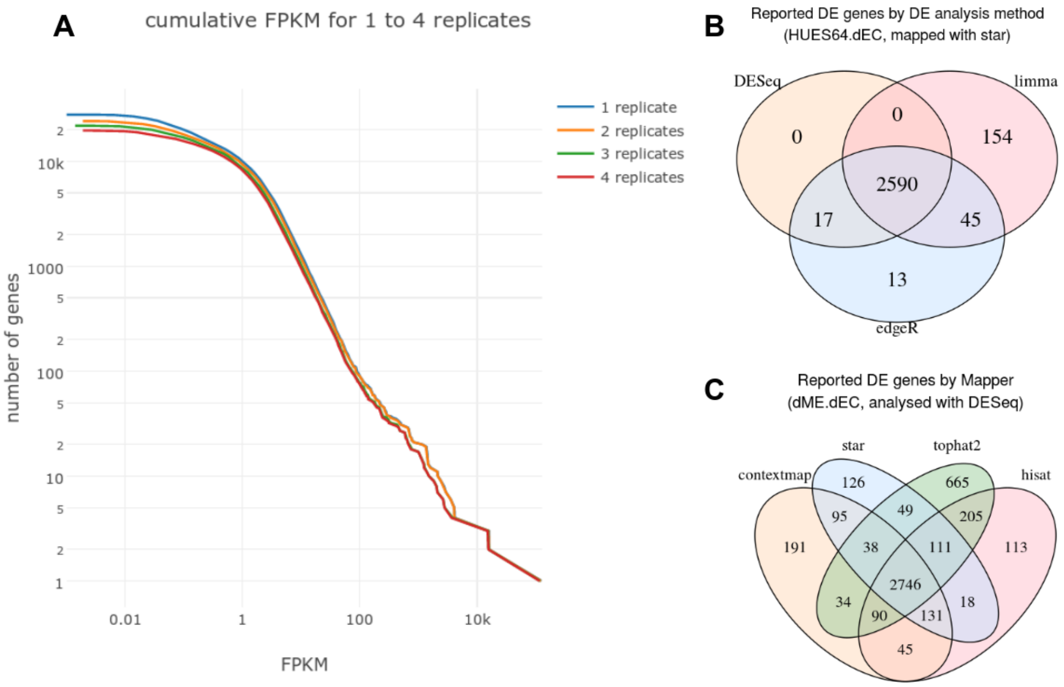


Figure 28 (A) Cumulative plot of the maximal FPKM in 1 to 4 replicates for the endoderm cell line (mapped with ContextMap). (B) Number of DE genes as predicted by the DE analysis methods DESeq, edgeR and limma (mapped with STAR). (C) Number of DE genes as predicted by DESeq, comparison of the mappers ContextMap, HISAT, STAR and TopHat2.

as for DE genes. We are interested in how consistent the predictions for gene regulations are for all 12 combinations of our four mappers and three DE analysis methods. Figure 29A shows the number of constant and DE genes for parameter set 2 for all state pairs. Genes, which are not save regulated in any state pair were excluded from the plot. For all state pairs, the number of genes identified by 1-11 mapper-DE method combinations is comparably low (mostly < 500), while there are significantly more genes identified by either 0 or 12 combinations (about 2000-4000). This suggests two properties: first, genes that are save regulated for some state pairs (i.e. constant or DE for all 12 combinations) are neither constant nor DE for any mapper or DE analysis method in other state pairs and second, if a gene is predicted constant or DE for at least one combination, it is very probable that it is save regulated. Another observation is that for all state pairs, there are slightly more genes identified for 3, 6 and 9 combinations. As these numbers are all multiples of 3, it seems that this increase of genes is caused by one of the four mappers. As Tophat2 showed to be the most inconsistent compared to other mappers in the previous paragraph, this is a reasonable explanation. The figures also show a significant difference in the number of save regulated genes per state pair: the highest number of DE genes is in the differentiation from HESC to mesoderm. Figure 29B compares the number of save genes per state pair to the total number of genes predicted by 1-11 combinations, i.e. all genes that were predicted for this state pair for at least one mapper or DE analysis method but are not save regulated. For all state pairs except for endoderm-ectoderm, the number of save genes was higher than the number of 'non-save' genes. The proportion between these two classes of genes were highly consistent for the three different parameter sets (see IV18), so we chose the most conservative approach and only used save regulated genes for downstream analysis. For genes identified by at least 10 combinations, figure 29C visualizes the number of constant expressed, upregulated and downregulated genes for parameter set 2. The proportions differ for each state pair and parameter set, with the latter being expactable as the parameters define whether a gene is assigned DE, constant, or non of both.

3.6.3.5 Regulation of TFs

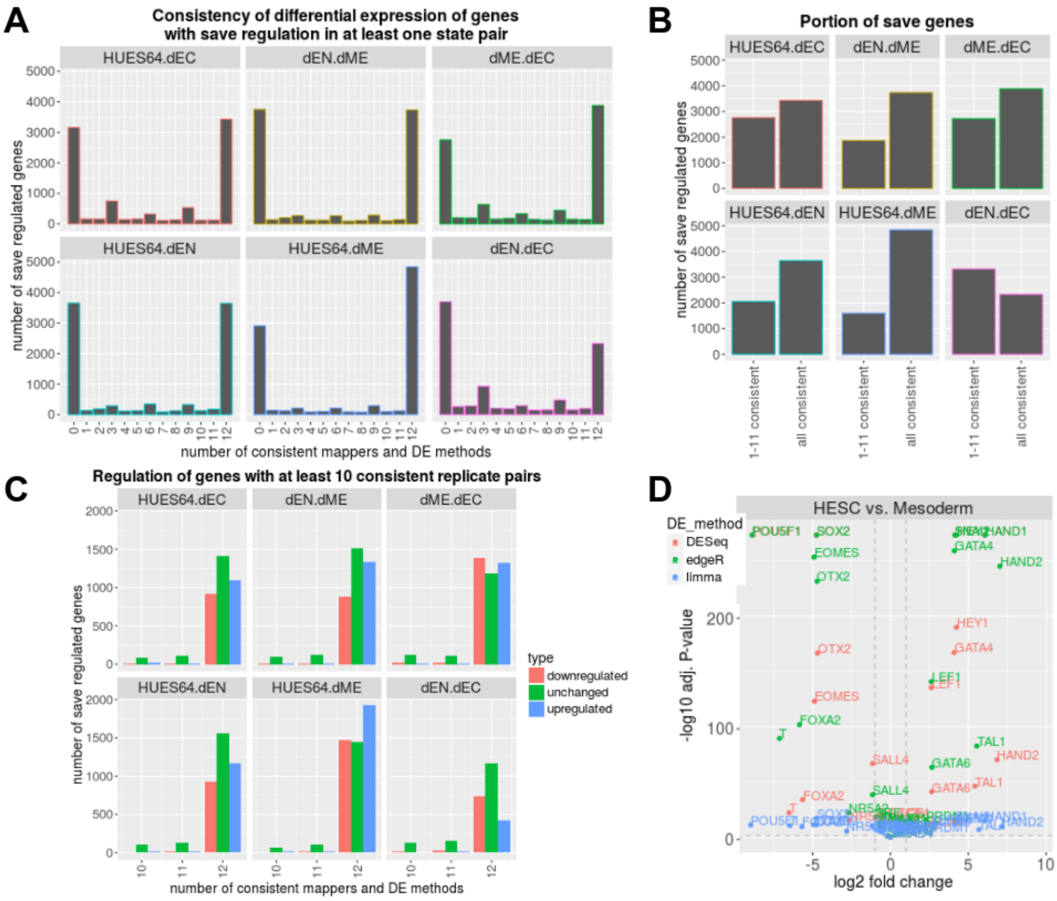
Figure 29D show a volcano plot including only our TFs of interest (i.e. those with available ChIP-seq data) for ContextMap, comparing the states HESC and mesoderm (see IV19 for all mappers and state pairs). DE regulated TFs are highly consistent for the different mappers and DE analysis methods. Although limma has significantly higher p-values, the genes are still significantly DE; DESeq and edgeR just show extreme p-values of up to 300 on the -log10 scale. Again, the the highest fold changes can be found for the differentiation from HESC to mesoderm. The TF POU5F1 is highly DE in the differentiation into all 3 germ layers, and is known to play a key role in embryonic development and stem cell pluripotency [63]. In general, we can see that some of the TFs that are described to be very important for lineage differentiation by Tsankov et al. are also DE between the states (e.g. OTX2, PAX6 in dEC, HAND1 in dME). While there are some TFs that show DE in the differentiation to more than one lineage (e.g. POU5F1, HEY1, OTX2), the combination of DE TFs seems to be lineage specific.

3.6.4 Combination of RNA-seq data and ChIP-seq data

After preprocessing and analysing the data from ChIP-seq and RNA-seq individually from each other, we need to combine the information we obtained from both parts to be able to identify dependencies between DTFBS and corresponding DE genes.

3.6.4.1 Annotation of differential TF peaks to DE gene promoters

This step is an extension of what we did previously in the analysis of ChIP-seq data. In section 3.6.2.6 we mapped the identified DTFBS to unique promoter regions of genes, with 4 different definitions of the promoter region size. Now that we have identified save regulated genes for all state pairs, we can extend the mapping files generated earlier by this new information. For each gene that is save regulated for at least one state pair, we annotated the type of regulation to the mapping files for each state pair (see IV20 for all extended mapping files). The column *regulation* contains 4 different symbols: '*+*' denotes a gene safely upregulated from state 1 to state 2, '-' denotes save downregulation, '*0*' means a gene is constantly expressed in both states and '.' shows that the gene is not save regulated. However, this gene is save regulated in at least one other state pair. DTFBS mappings to promoters whose corresponding gene is not save regulated in any state pair were removed, as those mappings do not provide any reliable information about possible dependencies. Moreover, we only used the optimal replicate pairs for TFs to obtain DTFBS, which were defined in section 3.6.2.7. The mapping file also contains diverse information about the gene,



■ **Figure 29** (A) Number of constant and DE genes for all combinations of a mapper and a DE analysis method. (B) Number of constant and DE genes in 1-11 combinations compared to the number of genes in all 12 combinations. (C) Number of constant expressed, upregulated and downregulated genes, excluding genes identified by fewer than 10 combinations. (D) Volcano plot of 38 TFs for the states HESC and mesoderm for all DE analysis methods (mapped with ContextMap).

the peaks and the mapping, including the gene biotype, genomic location of the merged promoter region, maximum p-value and minimum log2 fc of different mapper-DE analysis method combinations and normalized signal values for each peak. We processed the mapping files for each promoter size definition and for each of the 3 DE parameter sets (i.e. different p-value and fc thresholds), which resulted in a total of 12 different modes.

Figure 30 shows the number of different DTFBs in the promoter region of a save regulated gene for all state pairs and promoter definitions for the DE parameters p-value ≤ 0.05 and fc ≥ 1 (see IV21 for images of the distribution for individual state pairs). In general, we observe that the majority of save genes is regulated by more than one DTFBS. While the promoter region size does not seem to have a big influence on the distribution of DTFBS per gene, we observe a difference between state pairs. In general, genes seem to be regulated by more DTFBS during the differentiation from ESC to the germ layers (especially to endoderm and mesoderm) in contrast to the comparison of two germ layers. For the differentiation into endoderm and mesoderm, up to 15 DTFBS contribute in the regulation of a gene, while the maximum is 8 for all other state pairs. However, there is a serious issue, which was already mentioned in section 3.6.2: For several TFs, we do not have any data for some of the states. Thus, the number of DTFBS shown in the plot has an underlaying bias, as there were generally more data available for ESC than for the germ layers. For ESC, there is data for 20 TFs, while we only have data for 11 TFs in mesoderm. However, if we would filter our mapping and just use the TFs with available data for each state, a huge amount of information would be lost.

We are also interested in the number of genes that are regulated by a certain TF. Figure 31A shows that there are big differences both among TFs and state pairs. Some TFs only regulate a small set of genes in all state pairs (e.g. STAT3, GATA4, THAP11). In contrast, SRF regulates less than 1000 genes between mesoderm and ectoderm, but nearly 8000 genes between endoderm and mesoderm.

3.6.4.2 Aggregation of peak annotation over state pairs

This step constitutes one of the most important parts in our project: we want to find out whether it is possible to identify clear dependencies between differential TF binding and DE

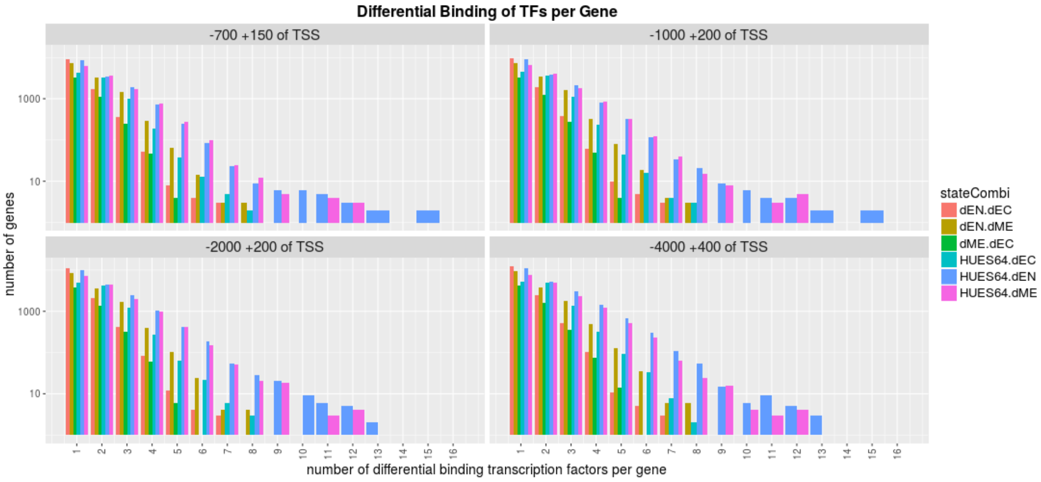


Figure 30 Number of different DTFBs in the promoter region of a save regulated gene for all state pairs and promoter definitions for the DE parameters p-value ≤ 0.05 and fc ≥ 1.

genes. We gain confidence in the dependency of a DE gene on a specific TF or a combination of multiple TFs, when we observe the same type of gene regulation together with the same TF signature in multiple state pairs. Contrarily, we lose confidence if we observe a gene with the same type of regulation in several state pairs that shows a different TF signature. To obtain this type of information, we needed to aggregate the mapping of save regulated genes and DTFBS over all state pairs for each gene. This was again done separately for each promoter size and DE parameter set (see IV22 for the resulting files). Figure 31C shows information about the distribution of different TF signatures for all 6000 save regulated genes (promoter region -1000 and +200 of TSS, pval≤0.05, fc≥1.0). Note that these genes are not save regulated in all 6 state pairs: about 2,500 genes are save regulated in only 1 state pair, another 2000 genes are save regulated in 2 state pairs. The number of different TF signatures (i.e. a specific combination of TFs at a specific genomic location) per save regulated gene is almost the same as save regulated genes for the number of state pairs,

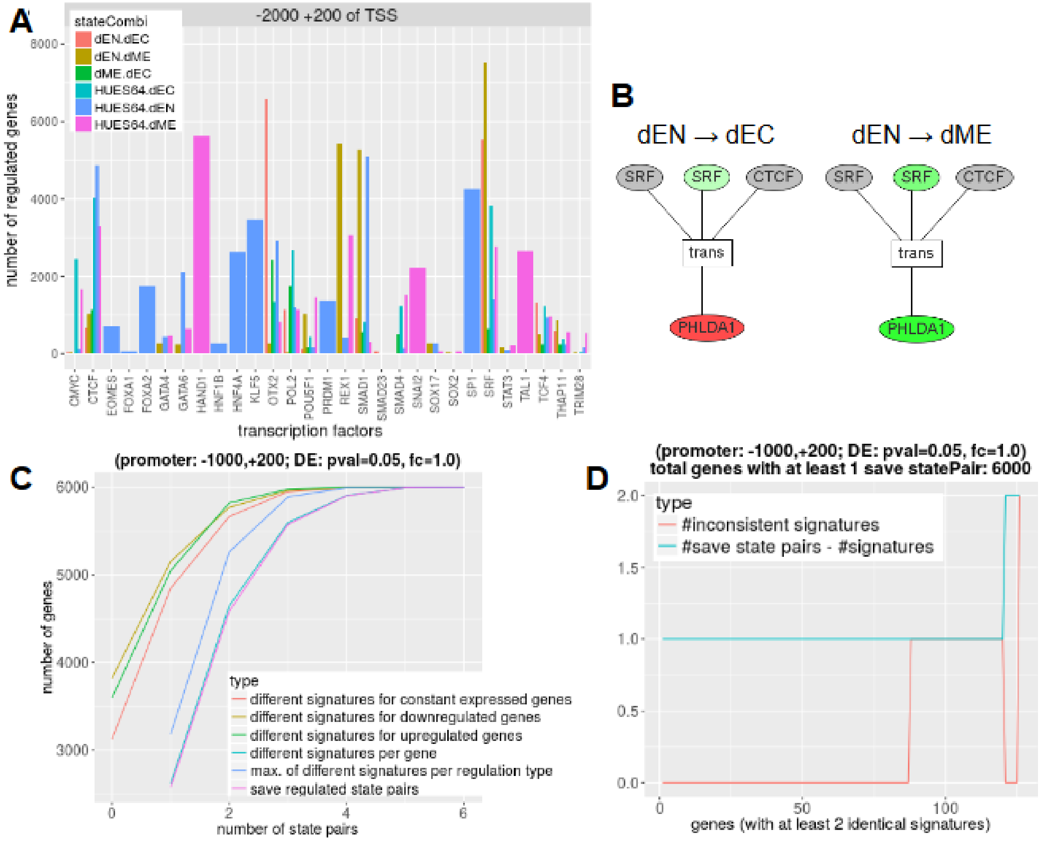


Figure 31 (A) Number of save regulated DE genes per TF (promoter region -2000 and +200 of TSS, pval≤0.05, fc≥1.0). (B) Example for a gene with an inconsistent TF signature. The gene PHLDA1 (at the bottom) is regulated by 2 TFs with a total of 3 DTFBS (at the top). While the second binding site of SRF is upregulated in both states (shown in green), the gene is downregulated in the first state pair (shown in red) and upregulated in the second. The intensity of the colours represents the measures binding signal for TFs and the log2 fc for genes. Inactive TFBS are shown in grey. (C) Distribution of different TF signatures for save regulated genes over the number of state pairs. (D) Number of TF signatures occurring several times for a save gene (blue) and portion of inconsistent signatures (red). Genes were first sorted by the number of multiple TF signatures and then by the number of inconsistent signatures.

which means that even if a gene is save in multiple state pairs, it mostly has a different TF signature for each of those state pairs. This is reasonable if the gene is upregulated in one state pair, and downregulated or constantly expressed in another pair. To analyse this behaviour for save genes with the same regulation type in multiple state pairs, the figure also shows the number of different signatures for constant expression, upregulation and downregulation separately. Although there are significantly more cases where the gene has only one TF signature for a certain regulation type, there are still 100-400 cases were a gene has at least two different TF signatures for the same type of gene regulation. However, this number is computed separately for each regulation type. Most genes do not show all of the 3 different states in some state pair. Figure 31C also shows the maximum of different signatures per regulation type: here we see that about 750 genes have at least 2 different TF signatures for the same regulation type. See IV23 for plots for all promoter and DE parameter sets.

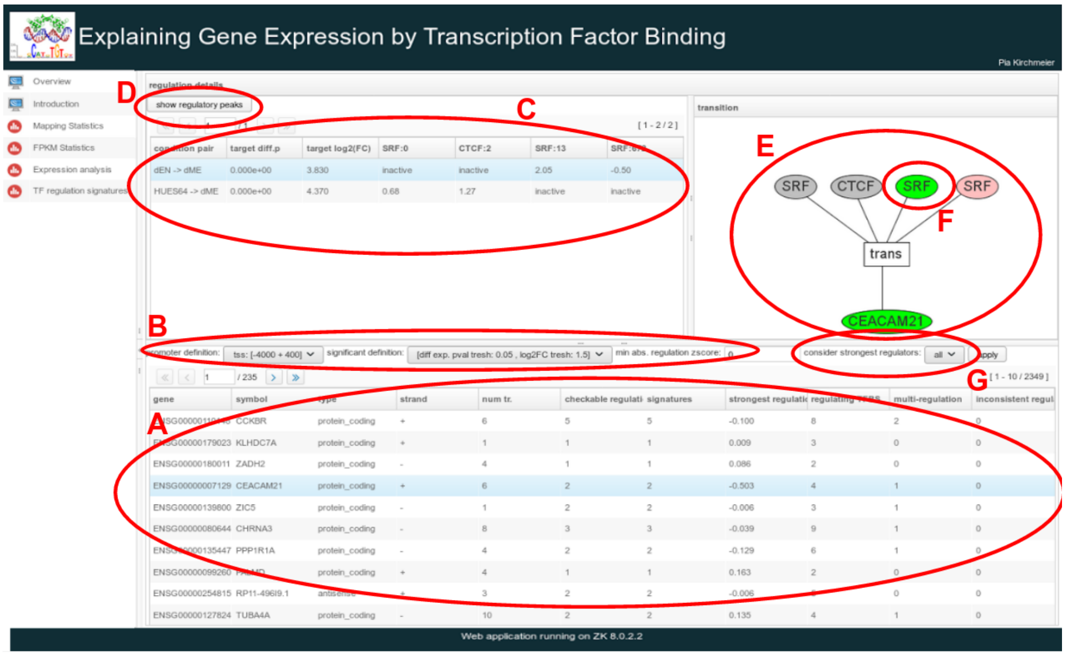
3.6.4.3 Inconsistent TF signatures

We are especially interested in TF signatures that appear multiple times for a save regulated gene. If the gene also has the same regulation type for these state pairs, the regulation of this TF signature is confirmed as we observe a consistent regulation. We define a TF signature as inconsistent, if the save regulated gene has different regulation types in the respective state pairs. For example, the gene PHLDA1 is regulated by 2 different TFs in our setup; the promoter contains 2 binding sites for SRF and one binding site for CTCF (see figure 31B). Although the second binding site of SRF is upregulated in both dEN to dEC and dEN to dME (with the two other binding sites being inactive in both state pairs), PHLDA1 is significantly downregulated (log fc = -2.12) for the first pair, but upregulated (log fc = 2.4) for the second pair. This indicates that the signature alone is not enough to explain the gene regulation for this gene. Figure 31D shows the number of TF signatures occurring several times for a save gene and the portion of inconsistent signatures (promoter region -1000 and +200 of TSS, pval$\leq$0.05, fc$\geq$1.0). About 140 signatures are observed twice for a save regulated genes, 10 signatures are even observed three times. Of these, 35 signatures are inconsistent, i.e. the gene is differently regulated in the respective state pairs. Generally, about 25% of TF signatures that are observed more than once for a save regulated gene are inconsistent for different promoter and DE parameter sets, however for a fc threshold of 1.5 this portion gets smaller (see IV24). Files containing detailed information about inconsistent genes can be obtained from IV25.

3.6.4.4 Interactive visualisation of results

The main page to visualize our results constitutes IV26. The site provides an interactive and browsable view to navigate through all genes and state pairs. For each save regulated gene, extensive information is available in a table at the bottom of the page, including the Ensembl gene ID, gene symbol, biotype, strand, number of transcripts, number of checkable regulations (i.e. the number of states where the gene is save regulated), number of different TF signatures, the strongest regulation (z-score) for a TF in any signature associated to the gene, the number of DTFBS involved in the regulation, and finally the number of same TF signatures for a gene and the number of inconsistent TF signatures (see figure 32A). Each column is sortable to facilitate a user-friendly layout which helps to quickly obtain information for a certain category of genes (e.g. genes with a lot of regulating TFBS, genes with multiple inconsistent regulations etc.). In addition, the user can select one of the 4 predefined promoter regions, the DE parameters (p-value and log2 fc) and a threshold for the binding strength of DTFBS (see figure 32B). When a column for a certain gene is clicked by the user, the regulation information for each state pair with a save regulated gene is displayed (see figure 32C). For each state, the p-value and log fc for the gene is displayed as well as the binding signal and distance from the first TSS for all DTFBS. To visualize the location of all differential binding peaks for this gene, the user can click a button on the top left (see figure 32D) which generates an image visualizing the transcript structures of the gene with annotated TF peaks similar to figure 27A. The difference to the visualization described earlier is that now differential binding peaks are shown instead of all mapped peaks and the colour of the peaks represents the binding strength (stronger signals are less transparent) and type of regulation (red for downregulation, green for upregulation). The user can also click a column for a certain state pair for the gene in the table, which generates a visualization of the regulation on the top right of the website (see figure 32E). For the selected gene and state pair, all DTFBS are shown. Again, the binding signal of the DTFBS and the log2 fc of the regulated gene are represented by colour. When one of the regulating DTFBS is clicked in the image (see figure 32F), the underlying ChIP-seq data are displayed for the respective location of the binding site. This is helpful to see how a peak is defined on the basis of the mapped reads.

The gene EBF3 is an interesting example to show some of the properties of the annotated



[1 - 2 / 2]

condition pair	target diff.p	target log2(FC)	SRF:0	CTCF:2	SRF:13	SRF:[illegible]
dEN -> dME	0.000e+00	3.830	inactive	inactive	2.05	-0.50
HUES64 -> dME	0.000e+00	4.370	0.68	1.27	inactive	inactive

promoter definition: tss: [-4000 + 400] significant definition: [diff exp. pval tresh: 0.05 , log2FC tresh: 1.5] min abs. regulation zscore: 0 consider strongest regulators: all apply

1 / 235 [1 - 10 / 2349]

gene	symbol	type	strand	num tr.	checkable regulati	signatures	strongest regulatic	regulating TFBS	multi-regulation	inconsistent regul
ENSG00000110148	CCKBR	protein_coding	+	6	5	5	-0.100	8	2	0
ENSG00000179023	KLHDC7A	protein_coding	+	1	1	1	0.009	3	0	0
ENSG00000180011	ZADH2	protein_coding	-	4	1	1	0.086	2	0	0
ENSG00000007129	CEACAM21	protein_coding	+	6	2	2	-0.503	4	1	0
ENSG00000139800	ZIC5	protein_coding	-	1	2	2	-0.006	3	1	0
ENSG00000080644	CHRNA3	protein_coding	-	8	3	3	-0.039	9	1	0
ENSG00000135447	PPP1R1A	protein_coding	-	4	2	2	-0.129	6	1	0
ENSG00000099260	PALMD	protein_coding	+	4	1	1	0.163	2	0	0
ENSG00000254815	RP11-496I9.1	antisense	+	3	2	2	-0.006	[illegible]	0	0
ENSG00000127824	TUBA4A	protein_coding	-	10	2	2	0.135	4	1	0

Figure 32 Snapshot of the main result page of the interactive web application. (A) Table containing detailed regulatory information about all genes and the respective TF signatures. (B) Selection of parameters for DE of genes and differential peak annotation. (C) Table containing information about the state pairs where a gene is save regulated. (D) Button for displaying the transcript structures and location of annotated DTFBS for the selected gene. (E) Regulation model of a gene for a specific state pair. See figure 31B for a detailed description of the model. (F) A click on a certain DTFBS generates a figure showing the underlying read data for this location. (G) Selection of the maximum number of DTFBS to be included in the regulation model. The peaks with the strongest signals are selected.

DTFBS we observed for different genes. Figure 33A-C shows the regulation model for the gene in three state pairs. Interestingly, the TF SRF has 5 DTFBS in the promoter region of the gene, which change their differential binding independently from each other: in the state pairs shown in figure 33A and C, 2 TFBS are upregulated for both pairs while another TFBS is just upregulated in the first state pair. In the second state pair, again another TFBS is downregulated, which is inactive in the first state pair. Both of the latter TFBS are inactive in the state pair depicted in figure 33C. To investigate the underlying cause of this principle of TF binding, it would be interesting to know the binding motifs for the TFBS. Different motifs for the binding sites could help to explain why some DTFBS show the same behaviour in different state pairs while others do not. We also investigated some of the underlying read data for the peaks. Figure 33E shows the mapped reads for one binding site of REX1 in the pair endoderm-mesoderm, where it is downregulated. The read data shows that REX1 has a rather weak signal for the binding in endoderm, but was not detected at all in mesoderm. In contrast, the read data for a binding site of SRF in the state pair ESC-endoderm shows a quantitative difference: SRF binds in both states, but the signal is much higher for the binding in endoderm.

3.6.4.5 Identification of a main TF per gene

After identifying a complex regulation model for each save regulated gene including several DTFBs, we wanted to find out how much dependency between differential gene expression and differential binding TFs we can see if we only consider a subset of the annotated DTFBS. Of course, our complex regulation model is already a reduced model of the actual regulatory network, as we only included 38 TFs in our study, which is a small subset of all known TFs. We are hence interested how the observed dependency changes with a step by step reduction of the DTFBS included in the regulatory model. We first need to define a rule how to select the main TFBS from the whole set of annotated peaks. We chose to use the binding sites with the strongest signals, however there are also alternative ways to select the main binding sites, e.g. by choosing the peaks which are the closest to the TSS. The problem with the latter approach is that for many genes, there is more than just one TSS and it is not trivial which TSS to use or if all TSS should be used. Figure 34A shows the number of checkable and inconsistent genes for a reduction to 1, 2, 3 and 4 main DTFBS compared to the complex model including all DTFBS for the different promoter sizes for p-value$\leq$0.05 and log fc$\geq$1.0 (see IV27 for images for different DE parameters). Here, 'checkable' genes mean that we have the same TF signature for at multiple state pairs. If the gene regulation in these state pairs differs, the regulation of this gene is inconsistent. The number of checkable genes does not change significantly for different promoter sizes and is about, but is greatly affected by the number of main DTFBS in the model. While there is only a very small difference between the number of checkable genes for 4 DTFBS compared to all DTFBS, the number of checkable genes rises from about 120 to 350 between 2 and 1 DTFBS per gene. We are also interested in the percentage of inconsistent genes. Here, we also observe a significant influence of the number of DTFBS included in the model: When using 4 or more DTFBS, 50-75% of checkable genes are inconsistent. With 3 DTFBS, we still observe 50-65% ionconsistent genes. This percentage drops to about 35% when just considering the DTFBS with the strongest signal (see figure 34B). Although we would assume that we could better understand gene regulation using more DTFBS to get closer to the complete regulatory model, it seems that we can better explain gene expression based on only very few DTFBS. A reason for this might be that the DTFBS having the highest signal have a lower chance of being the result of experimental errors. However, we need to keep in mind that we observed about 5000-6000 save regulated genes and only identify about 50 checkable genes for using all DTFBS, thus we can only check consistency for about 10% of save regulated genes, which makes it hard to evaluate the real extent of inconsistency.

3.7 Discussion

In this chapter, we presented different approaches to predict gene expression based on TF binding both in a single-state and differential setup. While simple methods just use expression data obtained with RNA-seq or microarrays, others use ChIP-seq data of TFs to build GRNs or predict gene expression of target genes. In addition, we described an integrative analysis of the regulatory interactions and TF dynamics across the differentiation of HESCs to the three germ layers by Tsankov et al., which revealed complex interactions of TFs with HMs and DNA methylation in the context of differentiation. We extended this analysis by using their data to investigate the dependency of differential binding TFs and DE genes, which turned out to be a very complex task depending on a variety of parameters and non-

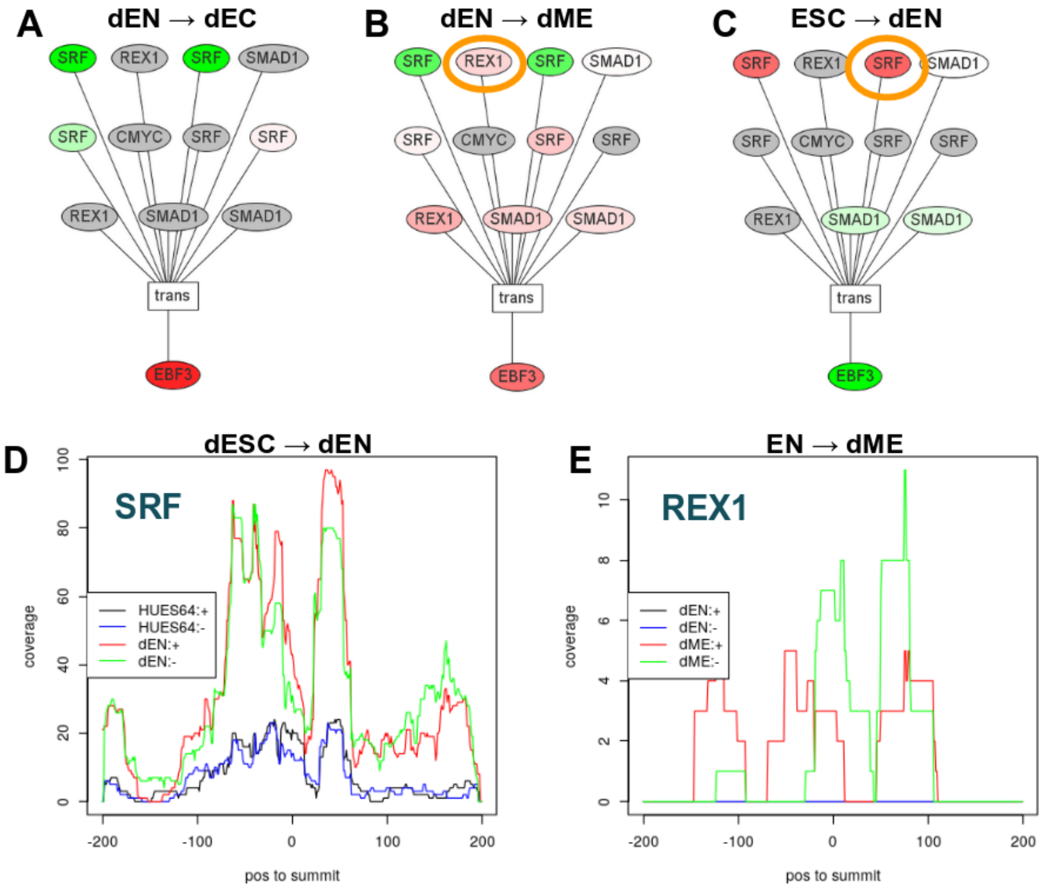


Figure 33 (A-C) Regulation models for different state pairs for the gene EBF3. Some TFs have multiple DTFBS in the promoter region, these often show different binding patterns. (D) Mapped reads for one binding site of SRF in the state pair ESC-endoderm (labeled in C). The TF binds in both state pairs, but the signal is much higher in endoderm, thus we observe a quantitative binding difference. (E) Mapped reads for one binding site of REX1 in the state pair endoderm-mesoderm (labeled in B). This TF only binds in endoderm, but not in mesoderm.

trivial definitions of regulatory principles. One of this difficulties represents the definition of the TSS of a gene, which determines the assumed promoter region. Ouyang et al. and Cheng and Gerstein just used one TSS per gene. However, genes often have transcripts with different TSS, often laying far apart from each other (see section 3.6.2.6). Considering only one TSS may oversimplify the model as we found out that regions around alternative TSS often show a significant amount of TF binding. Moreover, it is not trivial how to choose an appropriate TSS for a gene: possible selection criteria could be the number of TFs binding to that region, the location of the TSS in the gene (e.g. one could just take the first TSS) or the TSS which corresponds to the most abundant isoform of the respective gene. We also observed that some TFs binding at multiple promoter regions within the gene show a different regulation pattern for these binding sites, suggesting that just considering one TSS might bias the result (see section 3.6.4.4). We suppose that each isoform is regulated by the combination of TFBS that are close to its TSS. Thus, it would be reasonable to conduct the analysis on a transcript level instead of considering only the genes, which also requires obtaining read counts from RNA-seq on the transcript level. This is a difficult task, as reads are often hard to assign to a unique transcript, especially if the coverage is rather low.
Another issue affecting our project was the fact that it is not trivial how to obtain differential peaks, especially when only one replicate is available for each of the two states to be compared. As tools for the identification of DTFBS like DiffBind require at least two replicates per state, we could not use them and instead used MACS2 for identifying peaks in one cell line that were not present in the other cell line or had significantly fewer mapped reads. In some cases, these identified peaks did not look very reasonable when looking at the underlying read data, also it was not possible to obtain fold changes or p-values with this method. An option to overcome this difficulty would be to implement a method to extract peaks from MACS2 output that meet our assumptions and use LFC for obtaining differential binding peaks in order to get fold changes and p-values [64].
When comparing TF signatures for different state pairs for the same gene, we only define them to be equal if they (1) show the same type of regulation for all binding sites and (2)

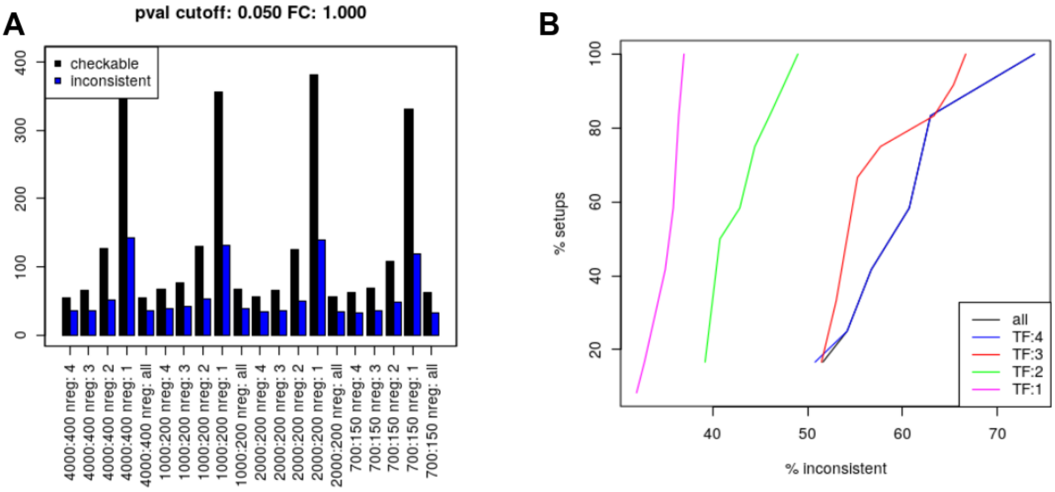


Figure 34 (A) Number of checkable and inconsistent genes when considering only the 1, 2, 3, or 4 DTFBS per gene with the strongest signal compared to our complex regulation model with all mapped DTFBS. (B) Percentage of inconsistent genes of the checkable genes for different numbers of main DTFBS. With fewer DTFBS in the model, the inconsistency decreases.

have the same binding sites for all TFs at the exact same position. The latter criterion is justified as TFs bind a specific DNA pattern, so a shift of the binding site by a few bases for the same TF is not reasonable. However, we often observe small distances of only a few bp between summits of the same TF in different states. This might also be caused by experimental inaccuracy, suggesting that overlapping binding sites with almost identical summit locations could also be merged. Still, the choice of an appropriate threshold for the maximal distance between summits needs to be made.

Both Ouyang et al. and Cheng and Gerstein considered the spatial effect of TFBS to the promoter in their analysis. Cheng and Gerstein also showed that binding sites offer more predictive power the closer they are to the TSS. Thus, it is possible that we could also benefit from this investigation and use it as a criterion to select the main TFs for a gene (see section 3.6.4.5). However, this again rises the question which TSS to use if there are multiple TSS for a gene.

As discussed previously, the association between TFs and target genes is still mainly done based on proximity (see section 3.4.3). However, more than half of the identified TF peaks are located in intergenic regions, most of them being far away from a genic region. The integration of methods generating genome-wide maps of chromatin interactions (e.g. Hi-C) could help to identify TFBS regulating a certain gene that would be missed with the proximity-based approach.

Still, one of the most profound problems we are dealing with is the low consistency of mapped peaks for replicates of the same cell line as discussed in section 3.6.2.4, suggesting that ChIP-seq does not lead to highly reliable results. We do not know to how much extent our findings are biased by experimental inaccuracy. However, this does not only concern our data: out of 3,000 experiments with replicates available on ENCODE, about 2000 are called with ≤10% IDR (i.e. at most 10% of the peaks in both replicates are consistent), 500 of those even with ≤1% IDR [65]. This shows, that it is very important to have multiple replicates for ChIP-seq analysis.

Finally, note that we only had available ChIP-seq data for 38 TFs, which is a small subset of known human TFs. Although the inconsistency decreased when we used less DTFBS for each gene (see section 3.6.4.5), the dependency of differential gene expression and TF binding might be higher if we could use data for all existing TFs for our model to represent the actual complexity of gene regulation.

References

[8] Chao Cheng et al. “A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets”. In: *Genome Biology* 12.2 (2011), R15. URL: `http://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-2-r15`.

[23] J. M. Vaquerizas et al. “A census of human transcription factors: function, expression and evolution”. In: *Nat. Rev. Genet.* 10.4 (Apr. 2009), pp. 252–263.

[24] G. J. Narlikar, H. Y. Fan, and R. E. Kingston. “Cooperation between complexes that regulate chromatin structure and transcription”. In: *Cell* 108.4 (2002), pp. 475–487.

[25] L. Xu, C. K. Glass, and M. G. Rosenfeld. “Coactivator and corepressor complexes in nuclear receptor function”. In: *Curr. Opin. Genet. Dev.* 9.2 (1999), pp. 140–147.

[26] C. K. Osborne et al. “Estrogen receptor: current understanding of its activation and modulation”. In: *Clin. Cancer Res.* 7.12 Suppl (2001), 4338s–4342s; discussion 4411s–4412s.

[27] T. Pawson. "Signal transduction–a conserved pathway from the membrane to the nucleus". In: *Dev. Genet.* 14.5 (1993), pp. 333–338.

[28] J. C. Bryne et al. "JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update". In: *Nucleic Acids Research* 36.Database (2007), pp. D102–D106. DOI: 10.1093/nar/gkm955.

[29] J. Wang et al. "Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors". In: *Genome Research* 22.9 (2012), pp. 1798–1812. DOI: 10.1101/gr.139105.112.

[30] S. G. Landt et al. "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia". In: *Genome Research* 22.9 (2012), pp. 1813–1831. DOI: 10.1101/gr.136184.111.

[31] B. Ren. "Genome-Wide Location and Function of DNA Binding Proteins". In: *Science* 290.5500 (2000), pp. 2306–2309. DOI: 10.1126/science.290.5500.2306.

[32] A. Ozdemir et al. "High resolution mapping of Twist to DNA in Drosophila embryos: Efficient functional analysis and evolutionary conservation". In: *Genome Research* 21.4 (2011), pp. 566–577. DOI: 10.1101/gr.104018.109.

[33] Qunhua Li et al. "Measuring reproducibility of high-throughput experiments". In: *The Annals of Applied Statistics* 5.3 (2011), pp. 1752–1779. DOI: 10.1214/11-aoas466.

[34] Peter V Kharchenko, Michael Y Tolstorukov, and Peter J Park. "Design and analysis of ChIP-seq experiments for DNA-binding proteins". In: *Nature Biotechnology* 26.12 (2008), pp. 1351–1359. DOI: 10.1038/nbt.1508.

[35] Joel Rozowsky et al. "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls". In: *Nature Biotechnology* 27.1 (2009), pp. 66–75. DOI: 10.1038/nbt.1518.

[36] Yong Zhang et al. "Model-based Analysis of ChIP-Seq (MACS)". In: *Genome Biology* 9.9 (2008), R137. DOI: 10.1186/gb-2008-9-9-r137.

[37] B. C. Haynes et al. "Mapping functional transcription factor networks from gene expression data". In: *Genome Res.* 23.8 (2013), pp. 1319–1328.

[38] J. J. Faith et al. "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles". In: *PLoS Biol.* 5.1 (2007), e8.

[39] G. STOLOVITZKY, D. MONROE, and A. CALIFANO. "Dialogue on Reverse-Engineering Assessment and Methods: The DREAM of High-Throughput Pathway Inference". In: *Annals of the New York Academy of Sciences* 1115.1 (2007), pp. 1–22. DOI: 10.1196/annals.1407.021.

[40] Andrea Pinna, Nicola Soranzo, and Alberto de la Fuente. "From Knockouts to Networks: Establishing Direct Cause-Effect Relationships through Graph Analysis". In: *PLoS ONE* 5.10 (2010). Ed. by Mark Isalan, e12912. DOI: 10.1371/journal.pone.0012912.

[41] S.-J. Dunn et al. "Defining an essential transcription factor program for naive pluripotency". In: *Science* 344.6188 (2014), pp. 1156–1160. DOI: 10.1126/science.1248882.

[42] C. Angelini and V. Costa. "Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems". In: *Front Cell Dev Biol* 2 (2014), p. 51.

[43] Z. Ouyang, Q. Zhou, and W. H. Wong. "ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells". In: *Proceedings of the National Academy of Sciences* 106.51 (2009), pp. 21521–21526. DOI: 10.1073/pnas.0904863106.

[44] A. M. Tsankov et al. "Transcription factor binding dynamics during human ES cell differentiation". In: *Nature* 518.7539 (2015), pp. 344–349.

[45] Denes Hnisz et al. "Super-Enhancers in the Control of Cell Identity and Disease". In: *Cell* 155.4 (2013), pp. 934–947. DOI: 10.1016/j.cell.2013.09.053.

[46] Michael B. Stadler et al. "DNA-binding factors shape the mouse methylome at distal regulatory regions". In: *Nature* (2011). DOI: 10.1038/nature10716.

[47] Matt Thomson et al. "Pluripotency Factors in Embryonic Stem Cells Regulate Differentiation into Germ Layers". In: *Cell* 145.6 (2011), pp. 875–889. DOI: 10.1016/j.cell.2011.05.017.

[48] Joseph K. Pickrell et al. "False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions". In: *Bioinformatics* 27.15 (2011), pp. 2144–2146. DOI: 10.1093/bioinformatics/btr354.

[49] R. Stark and G. D. Brown. "DiffBind: Differential Binding Analysis of ChIP-Seq Peak Data." In: *Bioconductor* (2011). URL: http://bioconductor.org/packages/release/bioc/html/DiffBind.html.

[50] Anthony Mathelier et al. "JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles". In: *Nucleic Acids Research* 42.D1 (2013), pp. D142–D147. DOI: 10.1093/nar/gkt997.

[51] T. L. Bailey et al. "MEME: discovering and analyzing DNA and protein sequence motifs". In: *Nucleic Acids Research* 34.Web Server (2006), W369–W373. DOI: 10.1093/nar/gkl198.

[52] V. Matys. "TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes". In: *Nucleic Acids Research* 34.90001 (2006), pp. D108–D110. DOI: 10.1093/nar/gkj143.

[53] K. Robasky and M. L. Bulyk. "UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions". In: *Nucleic Acids Research* 39.Database (2010), pp. D124–D128. DOI: 10.1093/nar/gkq992.

[54] P. Kheradpour et al. "Reliable prediction of regulator targets using 12 Drosophila genomes". In: *Genome Research* 17.12 (2007), pp. 1919–1931. DOI: 10.1101/gr.7090407.

[55] Gong-Hong Wei et al. "Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo". In: *The EMBO Journal* 29.13 (2010), pp. 2147–2160. DOI: 10.1038/emboj.2010.106.

[56] Thomas Bonfert et al. "ContextMap 2: fast and accurate context-based RNA-seq mapping". In: *BMC Bioinformatics* 16.1 (2015). DOI: 10.1186/s12859-015-0557-5.

[57] Daehwan Kim, Ben Langmead, and Steven L Salzberg. "HISAT: a fast spliced aligner with low memory requirements". In: *Nature Methods* 12.4 (2015), pp. 357–360. DOI: 10.1038/nmeth.3317.

[58] Alexander Dobin et al. "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1 (2012), pp. 15–21. DOI: 10.1093/bioinformatics/bts635.

[59] Daehwan Kim et al. "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions". In: *Genome Biology* 14.4 (2013), R36. DOI: 10.1186/gb-2013-14-4-r36. URL: https://doi.org/10.1186/gb-2013-14-4-r36.

[60] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12 (2014). DOI: 10.1186/s13059-014-0550-8.

[61] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics* 26.1 (2009), pp. 139–140. DOI: 10.1093/bioinformatics/btp616.

[62] M. E. Ritchie et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic Acids Research* 43.7 (2015), e47–e47. DOI: 10.1093/nar/gkv007.

[63] M. D. Bethesda. “National Library of Medicine (US), National Center for Biotechnology Information. Gene ID: 5460, Homo sapiens POU class 5 homeobox 1 (POU5F1)”. In: *Internet* (1988). [Online; accessed 19-July-2017].

[64] Florian Erhard and Ralf Zimmer. “Count ratio model reveals bias affecting NGS fold changes”. In: *Nucleic Acids Research* (2015), gkv696. DOI: 10.1093/nar/gkv696.

[65] Stanford University. *ENCODE: Encyclopedia of DNA Elements*. https://www.encodeproject.org/. Accessed: 2017-07-25.

4 Open Chromatin

by Tatjana Ammer, Nathalie Gerstner and Gergely Csaba

4.1 Motivation

On our way to a better understanding of gene regulation in yeast, the knowledge of sites bounded by transcription factors is of major importance. A large component of the regulation of gene expression is the direct binding of transcription factors (TFs) to distal and proximal regulatory regions of the genes. The understanding of those mechanisms is complicated because different TFs bind in various combinations and generate complex patterns of gene expression. [66]
One possibility to understand these pattern is to identify single transcription factors of interest and their associated binding sites in the genome through ChIP-Seq [67] experiments, as described in the previous chapter. The problem of ChIP-Seq experiments is their limitation to a single TF per experiment. In order to derive all binding sites for various factors, a vast number of experiments would be needed.
Since the activation of transcription is linked to chromatin remodeling, an alternative approach to obtain information about regulatory elements like transcription factor binding sites would be to use open chromatin data. [66],[68] Such datasets allow to search genome-wide for transcription factor binding sites (TFBSs) because TFs are only able to bind DNA in accessible and therefore nucleosome-free regions. Accordingly, better predictions of TFBSs can be reached by searching for motif matches in open chromatin regions, than by searching for them only in gene proximal promoter regions.

In this book chapter, the focus is on the open chromatin detection method *ATAC-Seq*. [69] ATAC-Seq is a relatively new method that promises to enable a highly accurate prediction of nucleosome occupancy and transcription factor binding sites. Whether ATAC-Seq holds this promise is investigated here using open chromatin and gene expression data from timecourse experiments during osmotic stress in *Saccharomyces cerevisiae*. The goal is to identify differentially expressed genes at each of the measured timepoints and the associated transcription factors that are responsible for the differences.

4.2 Methods for open chromatin detection

There are several methods capable of finding open chromatin regions. These differ in various aspects, like the fragmentation of the DNA, the amplification approach used and the subsequent analysis of the output. Further, these sequencing routines have different limitations. ChIP-Seq can be used a reference method to detect open chromatin regions. Here, a protein of interest (e.g. a TF) is enriched by immunoprecipitation from cross-linked cells along with bounded DNA fragments and afterwards the DNA is sequenced. This appraoch identifies precise TFBSs across the whole genome but it needs an antibody and for each TF an individual experiment is necessary. [67] Therefore, to identify pattern of TF binding, many individual experiments are needed. This disadvantage induced the development of other methods, that measure the whole open chromatin region. The different outputs of these approaches are shown in figure 35 and described in the following.

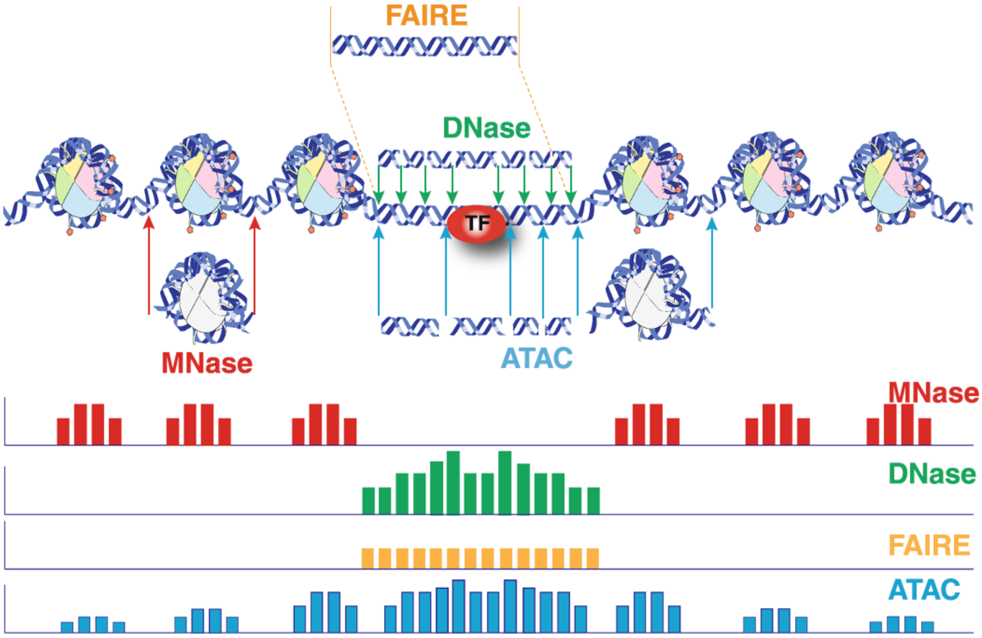


Figure 35 Overview of the different methods and there produced signal to detect open chromatin regions. (Source: [70])

4.2.1 FAIRE-Seq

FAIRE-Seq, which stands for **F**ormaldehyde-**A**ssisted **I**dentification of **R**egulatory **E**lements, is one of those methods to find regions of open chromatin. In this approach, histones are crosslinked with formaldehyde and afterwards the protein-bound DNA and the free DNA are separated with phenol-chloroform. Here, the DNA of accessible regions is in the aqueous phase whereas the DNA which is bound by proteins is in the interphase. The nucleosome-free DNA out of the aqueos phase gets sequenced. The result is a signal in the region of open chromatin, shown in 35 in orange. One advantage of this approach is, that as there is no enzymatic cleavage necessary, there is also no sequence-specific separation bias. But the approach has a lower signal-to-noise ratio than other assays and thus, a high background signal which makes the interpretation difficult. This can also be seen in figure 35, as the signal is equal across the whole open chromatin region. [71], [70]

4.2.2 MNase-Seq

MNase-Seq uses the **mi**crococcal **n**ucle**ase** (MNase), which is a single-strand specific endo-exonuclease, that digests unprotected DNA until it hits an obstacle, e.g. a nucleosome. Thus, it makes use of the MNase digestion resistance of nucleosomes. As the open chromatin regions are digested by MNase, the DNA, that is occupied by nucleosomes can be sequenced. Therefore, MNase-Seq is considered as an indirect measurement for open chromatin regions as the signal is missing in those regions and present in the occupied ones. One problem about MNase-Seq is, that it needs many cells. Another disadvantage is, that due to the fact that the method uses enzymatic digestion, there is a sequence bias. [70]

4.2.3 DNase-Seq

As the open chromatin regions are hypersensitive to DNase I, this method uses this characteristic of those areas. DNase I, a double-strand endonuclease, cuts preferentially the chromatin at nucleosome-free regions, which are, as already mentioned, considered to be DNase I hypersensitive sites (DHS). Thus, the emerging DNase-Seq signal is enriched at those DHS, so it reveals open chromatin regions. In addition, digital genomic footprinting (DGF) with DNase I enables the location of TFBS in these accessible regions. But on the contrary it has been discovered, that DNase I has a sequence preference, that's why this method also has a cleavage bias. Further, many steps and many cells are required to follow the protocol. [72], [70]

4.2.4 ATAC-Seq

One of the newest methods, that currently makes the best promises is ATAC-Seq, which stands for **A**ssay for **T**ransposase **A**ccessible **C**hromatin using sequencing. This method uses a hyperactive Tn5 transposase. The transposase is loaded with adapters, fragments the accessible DNA and integrate its adapter load into those regions. The regions that are not or less accessible incorporate a steric hindrance (e.g. a nucleosome) and transposition becomes less probable there. As the transposase can only insert in the regions of open chromatin, the fragment length of the ATAC-Seq reads differ essentially, in so far as the fragments that are within nucleosomes are about 140bp long whereas the reads in the open chromatin regions are shorter. Further, ATAC-Seq not only reveals the information where open chromatin is, but also points out the varying degrees of DNA accessibility within those regions. ATAC-Seq resolves even from few cells the chromatin structure with high resolution and sensitivity and thus enables analyses of different aspects of it. Besides deriving nucleosome positioning, one can use the occupancy information to map TFBS. [69],[70]

4.3 Peak Calling with MACS2

To obtain the regions where there are lots of ATAC-Seq reads, the first step to analyze the open chromatin regions, is to call the peaks. For this purpose, MACS [73] is a well-established tool. The abbreviation stands for **M**odel-based **A**nalysis of **C**hIP-**S**eq. So as the name reveals, it was initially designed for ChIP-Seq data but can also be used with other methods, like ATAC-Seq. In principle MACS identifies the genomic regions, where there are statistically significant more reads than in other parts of the genome, therefore, these are considered to be enriched regions. Considering the case, where MACS is called with ATAC-Seq data, these are the regions, where there are more transposition events, so those are the accessible chromatin regions. Further, the peaks are needed as an input for NucleoATAC.

4.4 Nucleosome positioning and occupancy - Nucleo-ATAC

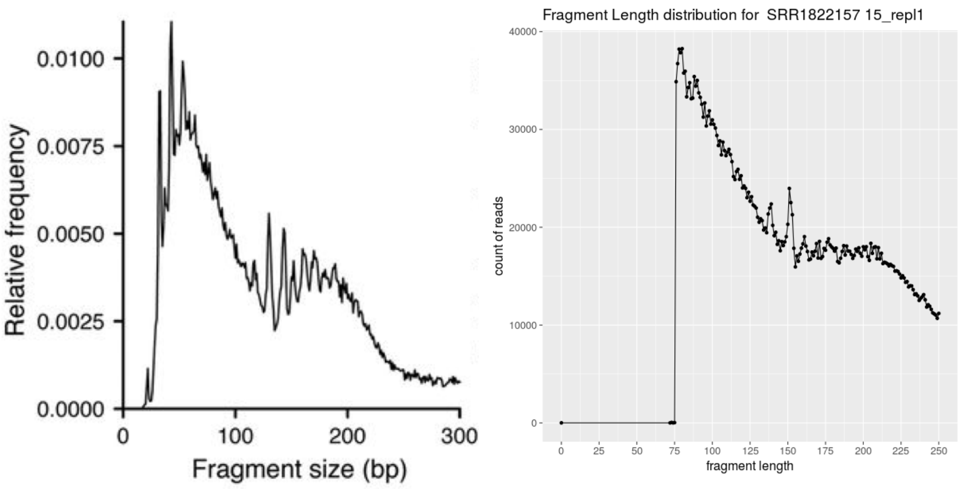
Nucleosome positions and thus the accessibility of the chromatin is of great importance to the regulation of gene expression. The reason for this is, that transcription factors compete with nucleosomes in so far as they both want to bind to the chromatin. Therefore, we want to know, how the changes in accessibility of the chromatin influence the transcriptional regulation of the genes. In detail, it is interesting to understand, how changes in nucleosome positioning and occupancy between different conditions are linked with differences in gene expression between these states. More concrete, we investigated this issue in our project

with data from a time course experiment with yeast under osmotic stress.
Different methods are available to examine the positions of nucleosomes but there are several limitations. For example one method based on MNase-Seq data, tries to deduce the nucleosome positions from the MNase-Seq reads. But, as the signal-to-noise of those reads is already high, the resolution of the positions is also poor. Another problem with this method is, that it cannot determine quantitative measurements of the nucleosome occupancy. [70]
A distinct approach is to determine the nucleosome positions with chemical mapping. Although the resolution of this method is very high and the positions are also very accurate, this method is very costly and therefore not practicable for a large-scale experiment. [74]
Hence, another method, NucleoATAC [75], promises to identify the positions with up to base-pair resolution and also to provide quantitative measurements of the occupancy. This method enables to detect changes in the nucleosome positions and occupancy during a dynamic cellular response, e.g. yeast under osmotic stress, as it uses ATAC-Seq reads.
In the following, the different steps of the NucleoATAC workflow are described in detail. [75]

4.4.1 NucleoATAC - workflow

NucleoATAC uses as basis for it's algorithm the specific fragment length distribution of ATAC-Seq data. This fragment length distribution is shown in figure36. The plot shows the fragment size on the x-axis and the relative frequency of the corresponding fragments on the y-axis. Therefore, it can be seen, that there are lot of short fragments (around 75bp long) which correspond to the nucleosome free regions (NFR). There is another peak between 140 and 200bp, that represents the fragments of nucleosome-associated regions. This specific length distribution origins from the fact, that the transposase Tn5 inserts itself at those sites where the chromatin is open and can not insert itself in those DNA regions where there is a nucleosome, because the nucleosome protects the DNA. In the NFR there are lots of Tn5 insertions, that's why there is a peak at about 75 bp accounting for those shorter fragments. This fragment length distribution was also examined in the ATAC-Seq data which was used in the project. Figure 36b also shows that there are lots of short fragments and that there is another small peak at around 150bp accounting for the nucleosome-protected regions.
Furthermore, the distribution was investigated in combination with the position of the fragment relative to a nucleosome. The nucleosome location was inferred from the dyad position determined by chemical mapping. The dyad of a nucleosome is a cystein-modification of a histone at the center of the nucleosome. Therefore, it represents the nucleosome midpoint. The resulting plot is shown in figure37a. To represent this relationship between the fragment size and position, a V-Plot is used, which maps the density of the sizes against the midpoint location relative to a genomic feature, in this case, the dyad position. Thus, it can be used to deduce the characteristics of the fragment length distribution of the ATAC-Seq data relative to a nucleosome. The V-plot shows that the fragments positioned at a nucleosome, about 60bp upstream and 60bp downstream, are longer than those that are in nucleosome-free regions, which are located at about 100 to 150bp upstream and downstream. Further, it can be seen that there are almost no short fragments in the region that is protected by a nucleosome. This highly structured pattern of the fragment size distribution around nucleosome positions (i.e. between 60bp up- and downstream of the dyad position for sizes between 105 and 250bp) is then used to determine the nucleosome positions and their occupancy by NucleoATAC. The V-plot is also displayed for the ATAC-Seq reads used in the project (37b).

The corresponding workflow of NucleoATAC is shown in figure 38 exemplary for one genomic



(a) Fragment length distribution of ATAC-Seq reads for yeast.

(b) Fragment length distribution for ATAC-Seq data used in the project for 15 minutes, replicate 2.

Figure 36 Fragment length distribution.

region of chr I (i.e. 76000bp to 76800bp with two genes).
First of all, for each genomic position x a matrix F is computed, for the fragments with their centers between x−60 and x+60bp and sizes between 105 and 250bp, showing the length distribution for each genomic region according to the ATAC-Seq data. This specific matrix is then cross-correlated against the average V-plot matrix in step one. Thereafter, the result of the correlation is normalized by a background model, the nucleosome occupancy is determined and the positions are called. The individual steps, which can be seen in figure 38, are described in detail hereafter. [75]

4.4.1.1 V-Plot cross-correlation

The first step of the NucleoATAC algorithm is the cross-correlation of the V-plot matrix against the matrix F, which contains the fragment center and sizes for a specific genomic region, as described above. The cross-correlated signal is calculated as follows:

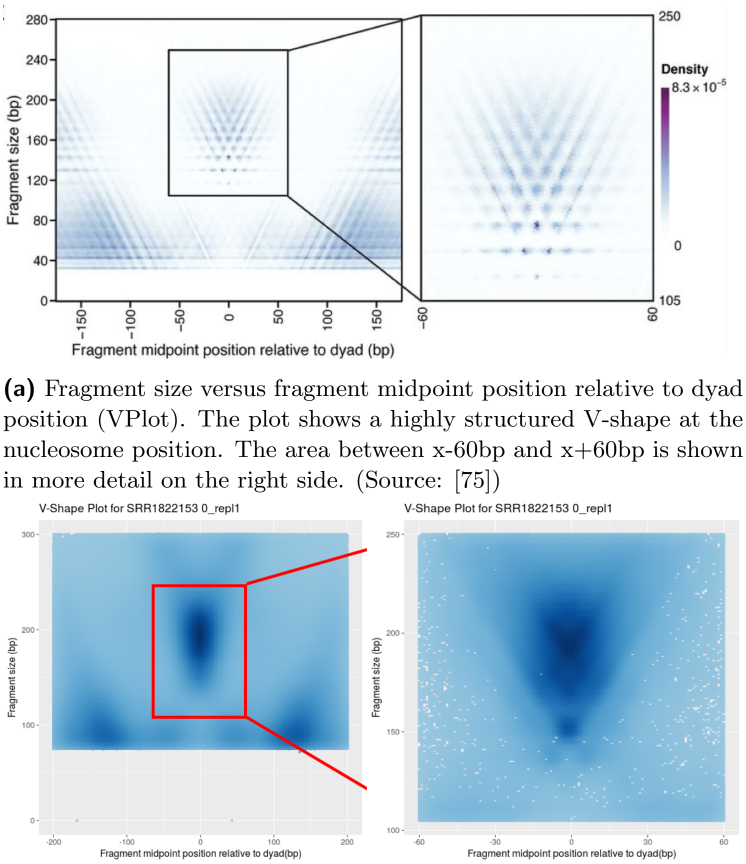
$$Signal(x) = F \cdot V \tag{2}$$

This dot - product shows how good the ATAC-Seq data match with the expected pattern around a nucleosome dyad position represented by the V-plot. Therefore, the signal is high, if the corresponding values in both matrices are similar to each other and are both high, and the signal is low when there is a huge difference between those values or when both values are low.

4.4.1.2 Normalization

The cross-correlated signal is then normalized using a background model. The normalization is needed as the Tn5 has an insertion sequence bias and the signal can vary due to differences in the openness of the chromatin. The background signal is calculated as shown in 3.

$$Background(x) = B \cdot V * \sum F \tag{3}$$



(a) Fragment size versus fragment midpoint position relative to dyad position (VPlot). The plot shows a highly structured V-shape at the nucleosome position. The area between x-60bp and x+60bp is shown in more detail on the right side. (Source: [75])

(b) Vplot for ATAC-Seq data at 0 minutes for replicate 0.

Figure 37 V-Plot for ATAC-Seq data.

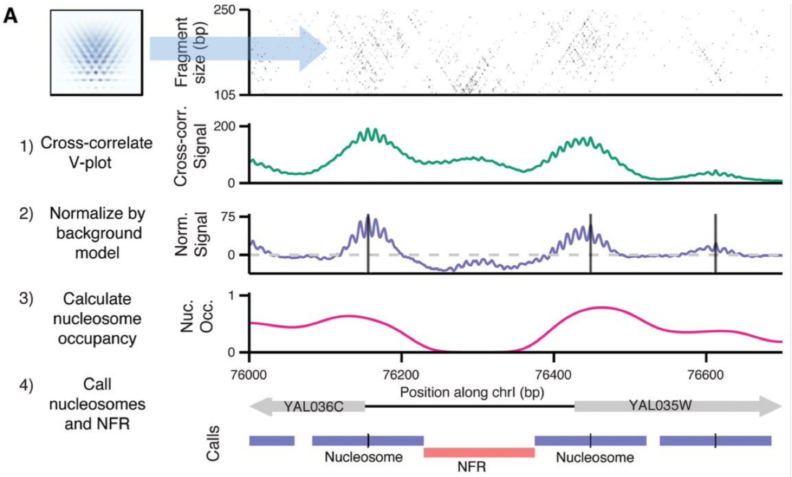


Figure 38 Schematic workflow of NucleoATAC. The x-axis shows an exemplary genomic region (chr I, 76000bp to 76800bp) with two genes. (Source: [75])

It uses the matrix B, the V-plot matrix and the sum over matrix F, which acts as a scaling factor to ensure that the background model represents the anticipated signal for the observed number of fragments in this area. Matrix B on the contrary takes into account the sequence bias of Tn5 and the fragment size distribution. Therefore, it contains the relative probabilites of observing fragments of different sizes and midpoint positions. In order to account for the sequence bias of Tn5, a one-dimensional sequence preference is calculated. This is done, as the sequence preference of Tn5 is about 21bp long, so a Positional Weight Matrix (PWM) is calculated for ± 10bp to the Tn5 insertion site. Then this PWM is used to calculate relative probabilites for each genomic region which results in the one-dimensional sequence preference. This 1D-preference is then needed to calculate the probability for the necessary Tn5 insertions for a specific fragment, as one needs one Tn5 insertion at the beginning and one at the end to define a fragment. Moreover, the probability to observe a fragment of this particular size is calculated using the fragment length distribution. After that these three probabilites are multiplicated, resulting in one entry for matrix B. After B has been calculated, the background signal 3 can be determined.
The normalized signal is then computed by subtracting the background signal from the cross-correlated signal, in order to get rid of the background noise.

$$NormalizedSignal(x) = Signal(x) - Background(x) = F \cdot V - B \cdot V * \sum F \tag{4}$$

4.4.1.3 Calculation of the nucleosome occupancy

In order to calculate the nucleosome occupancy, NucleoATAC uses the fragment-size distribution. This distribution is modeled as a mixture of the nucleosome-free and the nucleosome-associated fragment size distribution. The different distributions are shown in figure 39.
Fragments smaller than 115bp are more likely to origin from the nucleosome-free distribution, that's why the nucleosome-free distribution below 115bp is parameterized as an exponential distribution, in detail, with a two-parameter distribution, the gamma distribution. To imbed the fragments bigger than 115bp into the nucleosome-free distribution, the gamma distribution is extrapolated. This is shown as the NFR fit model by the green (until 115bp) and the turquoise (from 115bp) line (39). The model for the nucleosome-free fragment size distribution is then presented by a combination of the observed lengths until the cut-off and

the extrapolated model from 115bp on (turquoise line). The nucleosome-associated model is then computed as the difference between the observed sizes and the NFR-model. Because these two values are the same until the cut-off, the nucleosome-associated model is zero until there. After this point the nucleosome-model is shown with the red line.
The concluding fragment-size distribution can then be modeled as displayed in 39.

$$P(i) = \alpha * P_{nucleosomal}(i) + (1 - \alpha) * P_{nucleosome-free}(i) \tag{5}$$

To get the α, which is between 0 and 1, all fragments across the genome are traversed with a 121bp window with a 5bp interval. For those fragments centered in such a window the maximum-likelihood (ML) estimate is calculated, that the fragment arises from the nucleosomal distribution. Therefore α is the fraction of fragments originating from the nucleosome-associated distribution and thus $(1-\alpha)$ the fraction of fragments arising from the nucleosome-free distribution. After that nucleosome occupancy tracks can be determined. Moreover, these tracks are then smoothed with a 121bp Gaussian-window and confidence intervals are computed.

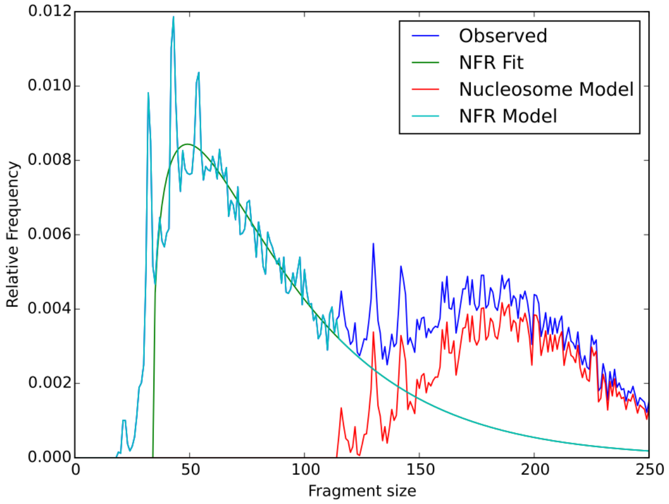


Figure 39 Fragment length distribution. Observed and modeled distributions displayed in different colours.

4.4.1.4 Nucleosome and NFR call

The last step of the NucleoATAC workflow is the determination of nucleosome positions, more precisely the dyad positions, and the nucleosome free regions. For this step, NucleoATAC makes use of the normalized signal. As this signal shows lots of local maximas with high periodicity, it is smoothed with a 25bp Gaussian window. In order to find maxima, that represent nucleosome positions, the sum of the smoothed normalized signal and the normalized signal is built. Thereafter local maxima are identified in this overall signal and these are considered to be potential nucleosome positions.
Another goal of NucleoATAC, besides determining the nucleosome occupancy, is to generate a non-redundant map of nucleosome positions. In order to fulfill this intention, the

candidate nucleosome positions are registered by a greedy algorithm. This algorithm first includes the potential nucleosome position with the highest signal. After that the position with the next highest signal and a distance to all other positions in the map of at least 120bp is imbedded in the map. This procedure continues until there is no position left with a distance over 120bp to any position in the map. The resulting map is considered as the non-redundant map of nucleosome positions. For each such position in the map a Z-score and a log-likelihood ratio is computed. The positions of nucleosomes can also be observed in the last line of figure 38, indicated by the line in the middle of each nucleosome.
Further, figure 38 recaps the whole consistent workflow in so far, as the nucleosome positions are indicated where the maximas of the normalized signal and the cross-correlated signal lie and also where there is a light V-shape in the first plot. Moreover, the nucleosome occupancy is high in those regions where there is a nucleosome and low (0) in the NFR.

4.4.2 Validation

To validate NucleoATAC and to show its ability to deduce the competition between TFs and nucleosomes, the relation between the nucleosome occupancy and the TF binding can be examined. This is also shown in figure 40. There the NucleoATAC signal (first column) as well as the insertion profile (middle column) and the ChIP-Seq signal (last column) for five NFKB subunits is investigated. The figure demonstrates that sites with a very low nucleosome occupancy score (first row) show a clear deficiency of the NucleoATAC signal and a TF-footprint, indicated in the transposase insertion profile. Further, there is a high ChIP-Seq signal for all NFKB subunits, suggesting that these TF are bound and the chromatin is not accessible in this region. On the other hand, sites with a high nucleosome occupancy score (last row) do have a peak in the NucleoATAC signal, indicating that these sites are nucleosome-associated regions. Moreover, there is no such clear footprint in the insertion profile, denoting that the region is protected with nucleosomes and not bound by a TF. Besides, the ChIP-Seq signal is lower for these regions compared to those areas with a lower nucleosome occupancy score. This signalises, that these sites are less occupied by a TF than the other regions. All in all, this shows that NucleoATAC is able to distinguish between TF binding and nucleosome occupancy. [75]

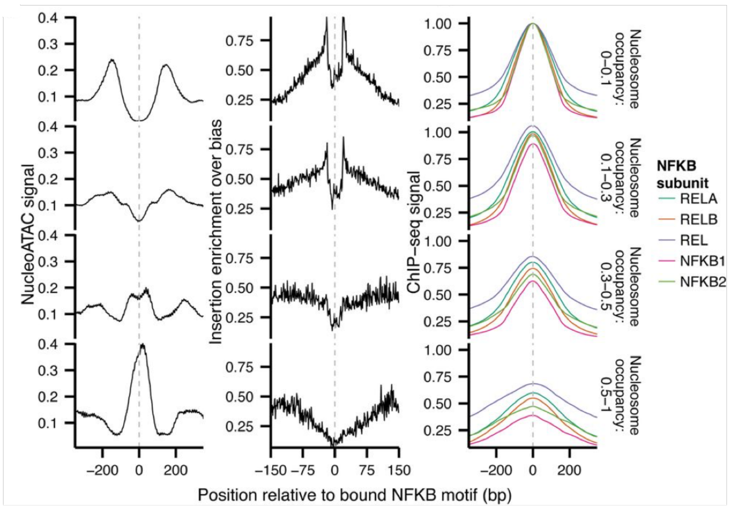


Figure 40 Comparison of NucleoATAC signal, ATAC-Seq insertion profile and ChIP-Seq signal. (Source: [75])

4.4.3 Summary

It can be summarized that the use of the specific fragment length distribution of the ATAC-Seq data leads the prediction of nucleosome positions with very high resolution. This is the case, as NFR are defined at those regions where short fragments are enriched, whereas in other methods, NFR are identified just by a lack of signal. Furthermore, by combining data of chromatin accessibility, nucleosome positioning and nucleosome occupancy, the chromatin architecture and especially changes in this architecture can be analyzed in detail. This is a great advantage, because with this approach, changes during a dynamic cellular response (e.g. osmotic stress in yeast) can be investigated. Due to the fact that also the competition between TF and nucleosome binding can be detected, NucleoATAC can be utilized to understand changes in gene expression data (e.g. obtained through RNA-Seq) by considering the changes on the chromatin level.

4.5 Transcription factor binding prediction — CENTIPEDE

Besides the calculation of nucleosome occupancy scores, it is also of great interest to predict transcription factor binding sites (TFBSs) across the whole genome precisely. The investigation of differences in the binding properties of transcription factors between various experimental conditions is a major step in understanding gene regulation.
Typically, TFBS prediction methods make use of known sequence motifs from databases (e.g. JASPAR) that are preferentially bound by the specific transcription factor of interest. [76], [77] For every site across such a motif, each of the four bases receives a score, according to the fact how likely this base would be at this position of a TFBS. These scores are collected in a matrix, called the *Position-Weight-Matrix* (PWM). Each site across the whole genome is scored according to the known PWM and positions with a score above a certain threshold are predicted as bound. An obvious drawback of these methods is that they incorporate only the DNA sequence, but neglect its accessibility. Since transcription factors compete with nucleosomes, the DNA mustn't be wrapped around nucleosomes to enable TF binding.
One method that takes also information about DNA accessibility into account is CENTIPEDE. [78] In addition to genomic information which are independent of cell- or tissue-type, CENTIPEDE considers also cell-specific experimental data (like open chromatin data or histone marks). CENTIPEDE requires the knowledge of a PWM to scan the entire genome for possible binding sites. All with a PWM score above a certain threshold are investigated. The method uses a hierarchical bayesian mixture model to calculate a posterior probability for each motif match in the input, that expresses the likelihood of this motif match site of being bound by the factor of interest. Parameter estimation of the model is done by an expectation maximization (EM) algorithm.

4.5.1 CENTIPEDE — the model

The hierarchical bayesian mixture model of CENTIPEDE calculates a posterior probability of being bound for each site, incorporating the likelihood of the experimentally data D and the prior probability of the site of being bound which is based on the genomic information of the site G. [78]

Figure 41 shows exemplary the different data types used by the CENTIPEDE model. The prior information for each motif match is consisting of a PWM score, the distance of the next transcription start site and a score which resembles the evolutionary conservation of

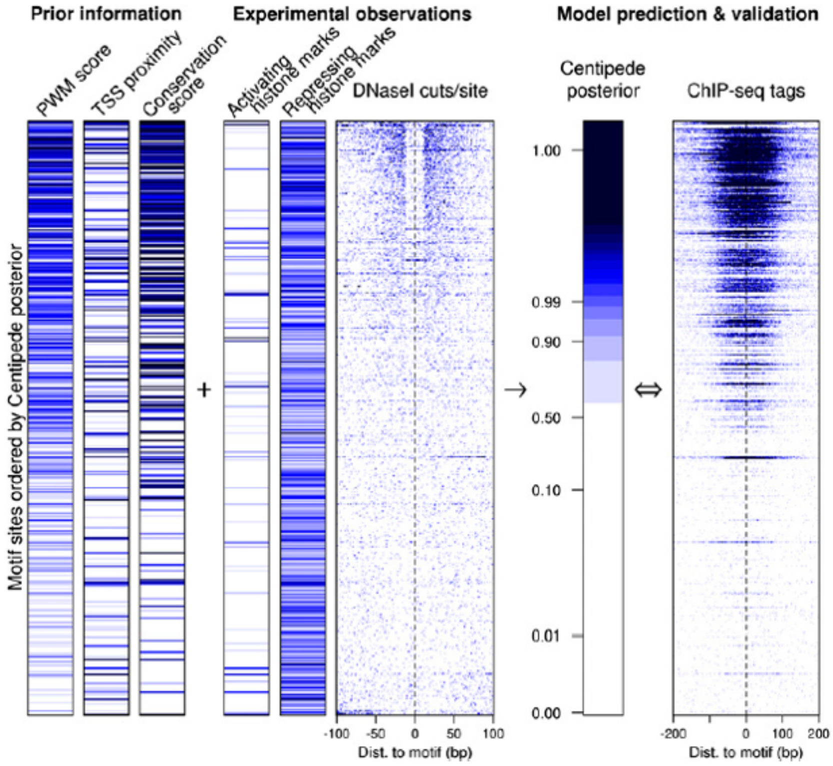


Figure 41 Schematic illustration of the CENTIPEDE model for the TF *REST* in human. (Source: [78])

the site. In addition to the genomic information, CENTIPEDE incorporates different cell-specific experimental datasets, such as activating histone marks, repressing histone marks and DNase-Seq data in this example. To the right of the various input data sets, the resulting posterior probabilities are shown and compared to ChIP-Seq data for the TF *REST* in human. Generally, the ChIP-Seq data correlates very well with the posterior probabilities calculated by CENTIPEDE.

$$P(Bound|D) = \frac{P(D|Bound) * P(Bound|G)}{P(D|G)} = \frac{P(D|Bound) * P(Bound|G)}{P(D|Bound) * P(Bound|G) + P(D|Unbound) * P(Unbound|G)} \tag{6}$$

As formula 6 shows, the CENTIPEDE model uses the formula of the Bayes theorem, whereby the denominator of the formula is a mixture of the probability for data D in the bound condition and the unbound condition, since it is assumed that data D is generated from one of the two underlying distributions (bound or unbound).

The prior probability π_l which corresponds to $P(Bound|G_l)$ for a motif match l, is modeled using a logistic function.

$$log(\frac{\pi_l}{1-\pi_l}) = \beta_0 + \beta_1 * PWMScore + \beta_2 * Cons.Score + \beta_3 * TSSproximity \tag{7}$$

This logistic function (formula 7) weighs each of the genomic input values with a factor β corresponding to the predictive power of the specific data type. In case that no genomic

information is entered into the model, the prior probability would be equal across all motif matches.
The likelihood of the experimentally data for a single motifhit in the bound or unbound condition is estimated, using a 200 bp region around the site for e.g. DNase-Seq or ATAC-Seq data. Using the fragment start sites in the surrounding region enables to model the total number of reads in the motif match area and also the spatial distribution of reads around the possible binding site, each for the bound and the unbound condition. Assuming that the fragment start counts are given in the matrix X with L rows for the L considered motif matches and S columns for the S different positions around the TFBSs, the total number of reads for each of those sites can be calculated by summing over the S columns in the specific row (see formula 8).

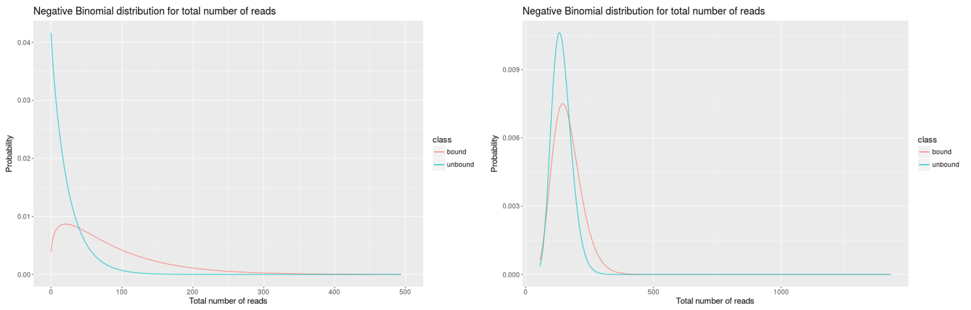
$$R_l = \sum_{s=1}^{S} X_{l,\cdot} \tag{8}$$

In order to model the distribution of the total number of reads in the motif match regions, the real distribution is used to fit a negative binomial distribution.

$$P(R_l|Bound) = NegativeBinomial(R_l|\alpha_1, \tau_1) = \frac{\Gamma(\alpha_1 + R_l)}{R_l!\Gamma(\alpha_1)}\tau_1^{\alpha_1}(1-\tau_1)^{R_l} \tag{9}$$

$$P(R_l|Unbound) = NegativeBinomial(R_l|\alpha_0, \tau_0) = \frac{\Gamma(\alpha_0 + R_l)}{R_l!\Gamma(\alpha_0)}\tau_0^{\alpha_0}(1-\tau_0)^{R_l} \tag{10}$$

Formula 9 shows the negative binomial distribution for the bound condition and formula 10 the same for the unbound condition. The parameters α and τ are different between the bound and the unbound state. For DNase-Seq data, the bound distribution has its peak at a higher total number of reads in the region than the unbound has, as figure 42(a) shows. Unfortunately, the difference between bound and unbound distribution is not that clear in the ATAC-Seq example which can be seen in figure 42(b).



(a) DNase-Seq data for *REST* in human **(b)** ATAC-Seq data for *FHL1* in yeast

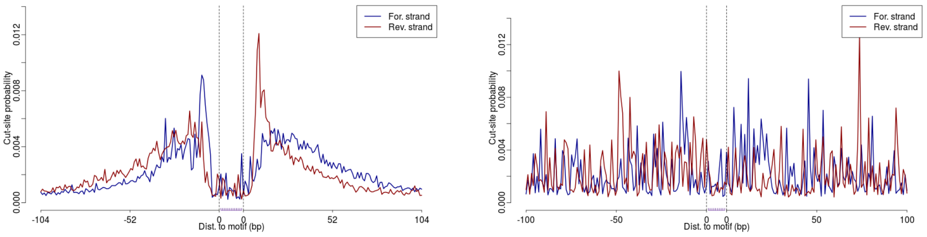
Figure 42 Negative Binomial distribution for total number of reads in the bound and the unbound class.

The two distributions allow to distinguish between open and closed chromatin according to open chromatin measurements. If the positional distribution of the reads inside the motif match region has no predictive power (e.g. for histone marks), it can stay unspecified. For DNase-Seq data it has been shown that the spatial distribution is very informative, since a specific footprint reflects the DNase cleavage pattern around the TFBS. This spatial distribution is modeled using a multinomial distribution which depends on as many parameters as there are positions taken into consideration around the motif match.

$$P(X_{l,\cdot}|Bound, R_l) = Multinomial(X_{l,\cdot}|R_l, \{\lambda_1, ..., \lambda_S\}) = R_l! \prod_{s=1}^{S} (\frac{\lambda_s^{X_{l,s}}}{X_{l,s}!}) \tag{11}$$

$$P(X_{l,\cdot}|Unbound, R_l) = Multinomial(X_{l,\cdot}|R_l, \{1/S, ..., 1/S\}) = R_l! \prod_{s=1}^{S} (\frac{S^{-X_{l,s}}}{X_{l,s}!}) \tag{12}$$

Formula 11 shows the multinomial distribution for the bound state which depends on the parameters λ_1 to λ_S, where each λ_i describes the probability of obtaining a read from the position with index i. Accordingly, $\lambda_i R_l$ gives the expected number of reads at position i in region l. In case of the unbound state, there is no specific footprint around the motif match expected. Therefore, the parameters λ_1 to λ_S are all set to $1/S$, describing a uniform cut-site distribution next to the motif match (see formula 12).



(a) DNase-Seq data for *REST* in human **(b)** ATAC-Seq data for *FHL1* in yeast

Figure 43 Cut-site probabilities for the positions surrounding a TFBS on forward and backward strand.

Figure 43(a) shows how likely a DNase cut is at the positions surrounding a specific TFBS, in accordance with the underlying data. Since the TFBS is bounded and therefore not accessible, no cuts are expected at those positions, while directly next to the TFBS the probability for the positions to be cut is extremely high. Not equally clear, but still visible, this signal is also detectable in the ATAC-Seq example in figure 43(b).

$$\begin{aligned} P(Bound|D) = & \frac{P(D|Bound) * P(Bound|G)}{P(D|Bound) * P(Bound|G) + P(D|Unbound) * P(Unbound|G)} = \\ & \frac{1}{1 + \frac{P(D|Unbound)*P(Unbound|G)}{P(D|Bound)*P(Bound|G)}} = \\ & \frac{1}{1 + (\frac{NegativeBinomial(R_l|\alpha_0,\tau_0)*Multinomial(X_{l,\cdot}|R_l,\{1/S,...,1/S\})}{NegativeBinomial(R_l|\alpha_1,\tau_1)*Multinomial(X_{l,\cdot}|R_l,\{\lambda_1,...,\lambda_S\})}) * (\frac{1-\pi_l}{\pi_l})} \end{aligned} \tag{13}$$

The complete model is shown in formula 13. This equation shows how the different distributions are merged together to build the full model (e.g. if only DNase-Seq or ATAC-Seq data is used as experimental data).

4.5.2 EM algorithm

Since the described model depends on many different parameters, a method to fit all of them is needed. This job is done by an expectation maximization (EM) algorithm which maximizes the likelihood function. The EM algorithm consists of 4 major steps: the initialization starts with the fitting of the βs in the logistic model by replacing the posterior probability by a binary variable that depends on the 90th%-tile of the DNase-Seq distribution. Thereupon, it follows the second step which comprises the first calculation of posterior probabilities by using only the prior model with the fitted βs. According to those initial posterior probabilities, all parameters are updated and the initialization steps are followed by a regular EM algorithm, until the change of the parameters in one iteration is below a certain threshold.

4.5.3 Validation

Pique-Regi et al. [78] also validate their method CENTIPEDE with various approaches. Figure 44 shows the comparison between CENTIPEDE posterior probabilities and ChIP-Seq positives and negatives. ChIP-Seq positives are defined as those sites where the motif match falls within a ChIP-Seq peak, while ChIP-Seq negatives are defined as those sites for which the fraction of reads from a control experiment is lower or equal compared to the number of reads in the ChIP-Seq experiment.

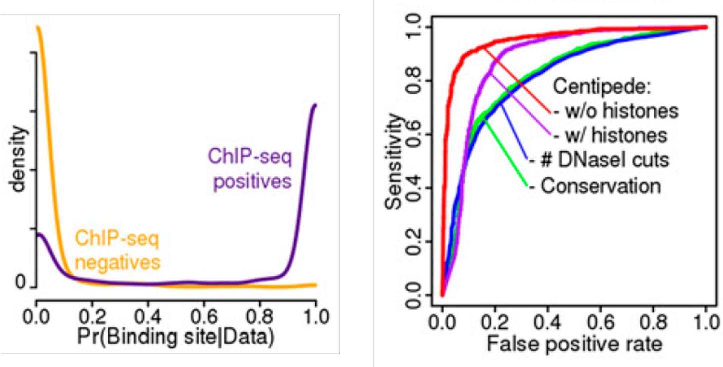


Figure 44 Validation of the CENTIPEDE model for the TF *REST* using ChIP-Seq data. (Source: [78])

The left plot in figure 44 shows the distribution of posterior probabilities for ChIP-Seq positives in yellow and that for ChIP-Seq negatives in purple. For ChIP-Seq negatives the posterior probabilities are generally small and for most of the ChIP-Seq positives they are close to 1, validating the predictive power of the CENTIPEDE model. On the right side of figure 44 ROC curves for different CENTIPEDE models for the TF *REST* are shown. The AUROC reaches its maximum for the CENTIPEDE model with DNase-Seq information, followed by that with DNase-Seq and histone mark data. The model that incorporates only the total number of DNase cuts —but no footprint information— is the third best and the model that uses only the conservation score is the worst of the four options in this example.

4.5.4 Summary

CENTIPEDE enables to predict TFBS across the whole genome for various transcription factors in a broad range of tissues and experimental conditions without the need of expensive ChIP-Seq experiments. However, the authors themselves [78] state that the method should not be used instead of ChIP-Seq experiments, but as a complement. Transcription factor binding maps are a major step in understanding gene regulation, also between various conditions. Incorporating open chromatin data into the prediction of these sites saves a lot of experimental effort.

4.6 Project

The objective of our project was the holistic explanation of differential gene expression by differential regulation. In particular, differential TFBS in the promoter regions of differential expressed genes should be identified by the analysis of ATAC-Seq data. The biological setup of the deployed gene expression datasets as well as the ATAC-Seq dataset was a timecourse experiment in yeast during osmotic stress. To what extent ATAC-Seq meets the expectations and the promoter regions of differentially expressed genes include differential binding sites is discussed in the following.
A detailed workflow of this project with linked inputs, outputs and code is provided on our project website which includes also interactive figures and results. [see Workflow],[see Website]

4.6.1 Data Retrieval

In order to achieve the goal of getting a better understanding of the impact of open chromatin on gene regulation we analyzed one of the best-studied organisms, Saccharomyces cerevisiae. To obtain differentially expressed genes at each timepoint of the experiment we used RNA-Seq and microarray data, whereas ATAC-Seq data was devoted to infer changes in nucleosome positioning and thus in the open chromatin regions. This is necessary to detect the associated transcription factors that are responsible for the changes in the gene expression. To get those TFs, we made use of motifs of four different databases and their related motifhits, which are analyzed further in the section 4.6.4. The four devoted resources were: JASPAR [76], Macisaac [77], matrix yeast and scpd [SCPD].

The RNA-Seq timecourse data was obtained from Babazadeh et al. [79]. During the experiment NaCl was added to the medium until a concentration of 0.4M was met and samples of the medium were gathered at 0min (just before the NaCl-induction) and at 15, 30, 60 and 90 min after the stress exposure. Further there are three replicates available at each time point. The data was then mapped with Contextmap [80], Tophat2 [81], Star [82] and Hisat [83] against the Saccharomyces cerevisiae R64-1-1.75 reference genome.

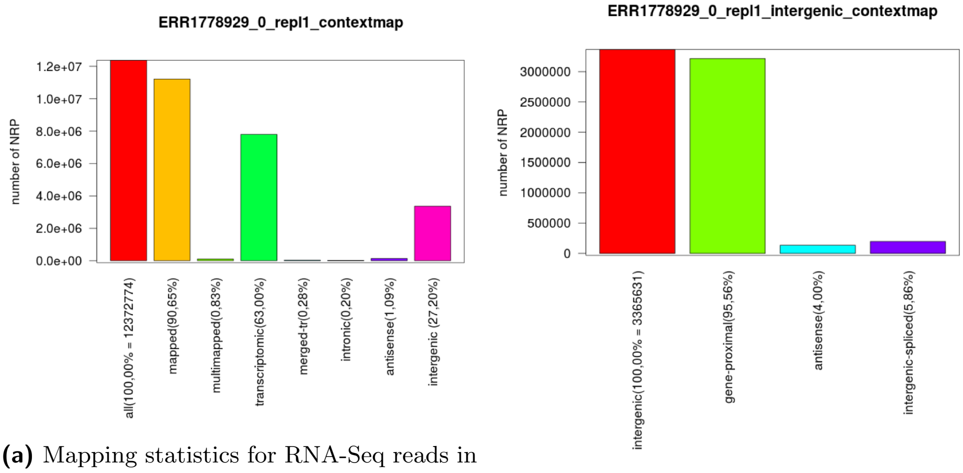
The microarray data was obtained from Ni et al. [84]. Its samples were taken at different timepoints, in particular, 0, 7.5, 15, 22.5, 30, 45 and 60 min after treatment with 0.6M NaCl and a two-channel hybridization was performed. For further analysis only the timepoints corresponding with the RNA-Seq times were considered.

The ATAC-Seq data was received from Schep et al. [75] and then mapped with bowtie2 [85] against the same reference genome as the RNA-Seq data. Here the time points of the chosen samples were again 0min, so just before the stress application and 15, 30, 45 an 60 min after NaCl was added so that a NaCl concentration of 0.6M was met. There are four replicates for the first time point and two for each of the following ones.

All three data sets were then first fundamentally analyzed and after that tried to infer the relationship between the different measured features.

4.6.2 Differential Expression and fundamental data analysis

First of all the RNA-Seq reads were analyzed in so far as it was calculated in which genomic feature each read resides. The resulting distribution of reads is shown exemplary in figure 45(a). Here we can see that the majority of reads lies in the genomic, specific in the transcriptomic area. But nevertheless almost one third of the reads are situated in the intergenic region, which first seems kind of odd. In further analysis, it turned out that those intergenic reads are almost all in the gene proximal region (Fig. 45(b)). This indicates that the gene start does not equal the transcription start site which is not annotated in Saccharomyces cerevisiae. Since the TSS is needed to define the promoter region accurately, the determination of it is another step in the project, described in 4.6.3.



(a) Mapping statistics for RNA-Seq reads in different features.

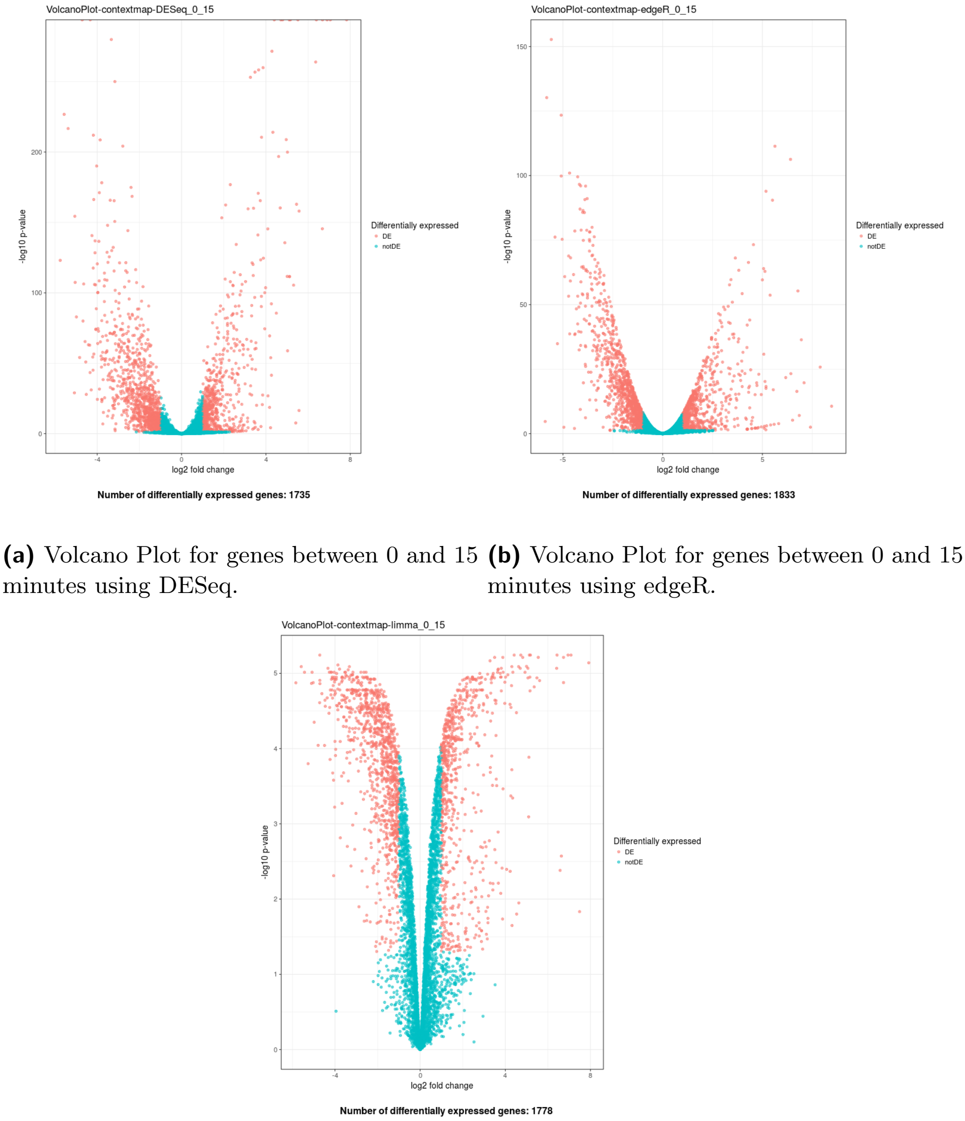
(b) Mapping statistics for intergenic reads.

Figure 45 Mapping statistics.

Another thing learned from this analysis is, that the different mapper seem to provide similar results.

Further there was a differential expression analysis of the RNA-Seq data and the microarray data conducted. The investigation of the RNA-Seq data was done with three different tools: edgeR [86], DESeq [87] and limma [88]. As shown in figure 46 those methods calculate very different p-values and fold changes but if only the differentially expressed genes are considered without regard to those values, the majority of genes is consistent over the methods. In the shown case between the condition 0 and 15min 1707 of 1735(DESeq), 1833 (edgeR) and 1778 (limma) are consistent. Between the conditions 0 and 15 minutes far more genes are differentially expressed than between the conditions 0 and 90 min, more accurate 1707 (0-15) compared to 585 (0-90). This coincides with the fact that yeast reacts strongly to stress conditions in the first interval, reaching a peak at 15min and after that it slowly eases down towards the starting state. [79]

The analysis of the microarray dataset was carried out using limma, as this seemed the most convenient tool for working with microarray data. As already mentioned, here only the timepoints corresponding to the RNA-Seq points are considered, so only the data for 0, 15, 30 and 60 minutes was used. This data likewise shows that between 0 and 15 minutes 1329 genes are significantly up- or downregulated whereas between 0 and 60 minutes only 286 genes are differentially expressed aiming at the same characteristic of the yeast's stress response.



(a) Volcano Plot for genes between 0 and 15 minutes using DESeq.

(b) Volcano Plot for genes between 0 and 15 minutes using edgeR.

(c) Volcano Plot for genes between 0 and 15 minutes using limma.

Figure 46 Differential expression analysis. [see interactive figures]

As we want to further work with the RNA-Seq data but the microarray data was used in the reference paper ([75]), the consistency between differentially expressed genes in those two datasets was determined. Figure 47 shows the result. As can be seen, there is a high correlation between the differentially expressed genes (p-value cut off: 0.05) between both methods (red points). Nevertheless, in RNA-Seq data more genes are differentially expressed, which is also shown by the different number of DE genes between microarray and RNA-Seq (1329 vs. 1707) but also some of those indicated by the microarray dataset are not classified as significantly up- or downregulated in the RNA-Seq data as pictured by the green points. In the downstream analysis the differentially expressed genes governed by limma were used as this is also the method used for microarray data analysis.

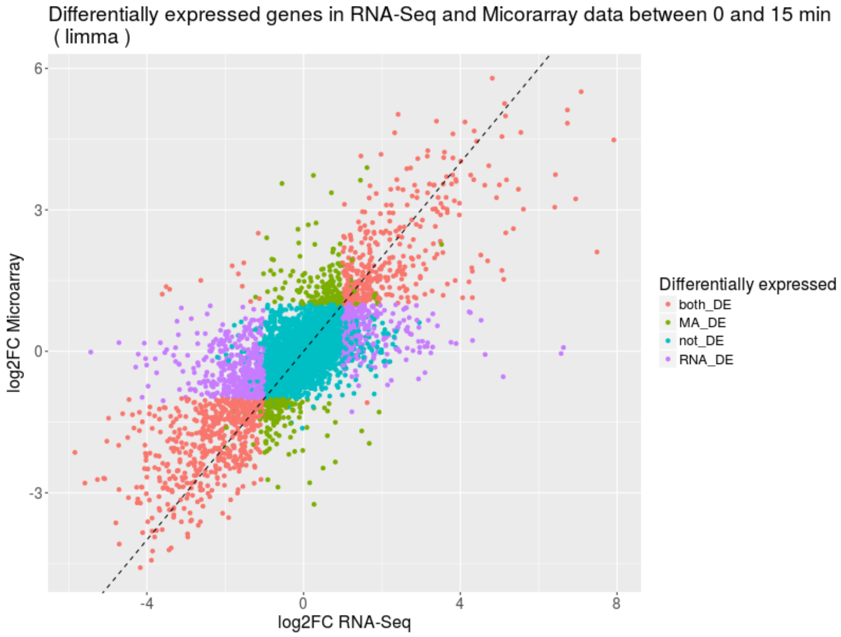


Figure 47 Consistency between microarray and RNA-Seq data between 0 and 15 minutes using limma. [see interactive figures]

4.6.3 Definition of the promoter region in yeast

Due to the fact that transcription factors usually bind in the promoter region of a gene, the definition of this feature was needed in order to analyze the open chromatin state of this property of the respective genes. Therefore we first had to determine the transcription start sites (TSS) as these are not annotated in Saccharomyces cerevisiae.
One fact that could be seen in the fundamental data analysis was that lots of the reads were intergenic and the majority of those reads was situated in the gene proximal area. Therefore those proximal intergenic reads were investigated using a two-window sliding technique. For each window the number of starting fragments was counted to obtain the coverage of each window and after that the ratio between these coverages was calculated. The TSS was then defined as the starting point of the second window, simultaneously the end of the first window, of those windows that had the maximal ratio. The technique is schematically shown in figure 48.

$$TSS = window_{2,start}(max(\frac{cov(window_2)}{cov(window_1)})$$

First the TSS was defined for each replicate in each condition and the consistency of the replicates was analyzed. This can also be seen in figure 49. The plot shows that the majority of TSS are consistent between the replicates, differing smaller than 10bp. But there are also some outliers with a difference greater than 100bp. That's why for constructing the overall consensus TSS the sites had to be merged. The merging process was divided into two stages. First the replicates were analyzed due to their consistency and the most consistent condition was then selected to define the TSS in the second step. The most consistent condition was defined as the one with the least difference of the TSS between the replicates for this condition. Further the final definition was carried out in so far as the mean of the sites between the replicates was calculated, except if one replicate defined a start site far

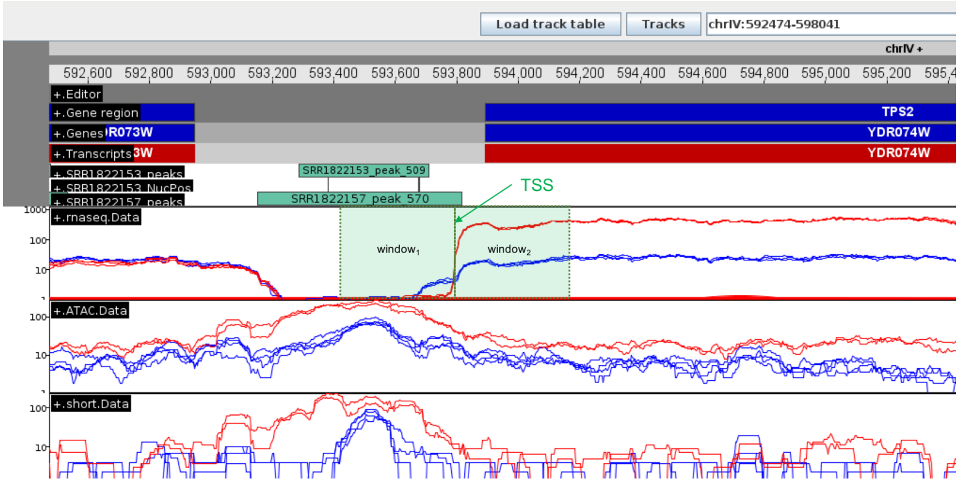


Figure 48 Schematic TSS definition with two sliding windows.

away from the site of the other two replicates then only the mean of the two more concurrent replicates was taken as the final TSS.

The promoter region of yeast, which is later used to analyze the changes in nucleosome positioning for the transcription factor binding sites (TFBS), was then defined as -600 and +100bp relative to the TSS.

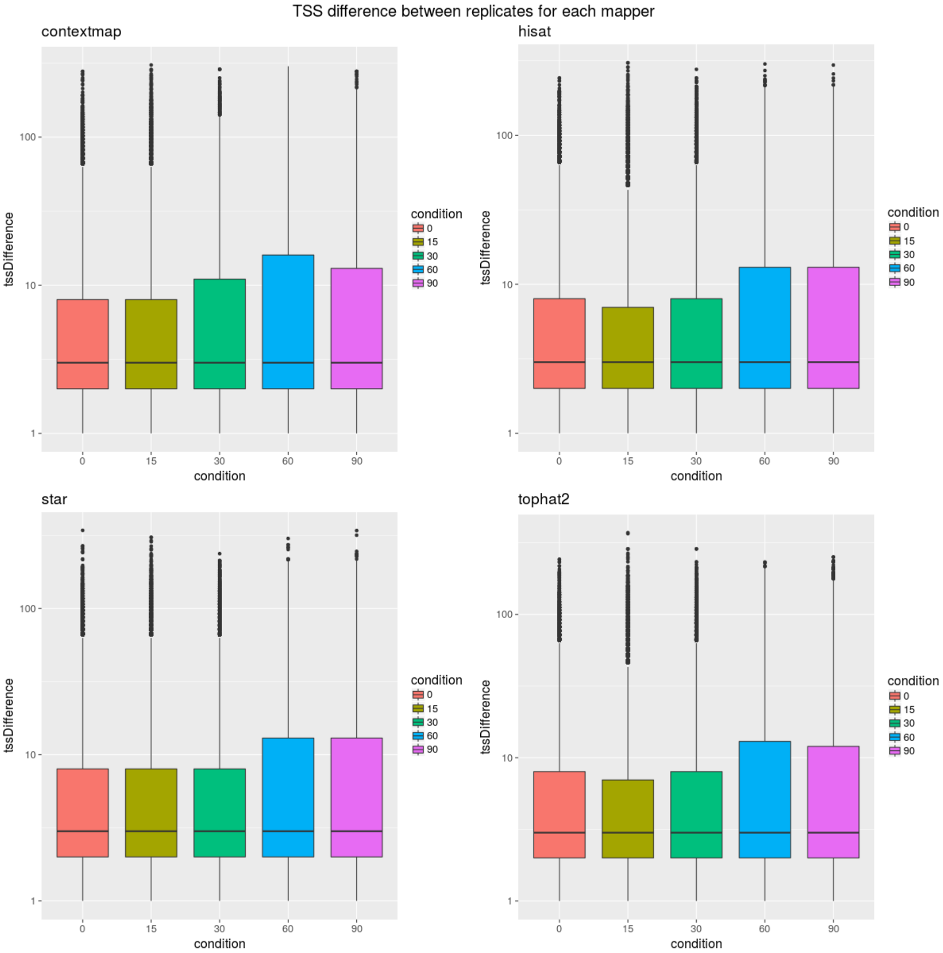
4.6.4 Motifs and motifhits

For the analysis of various transcription factor binding motifs, four different databases were exploited, namely $JASPAR\_CORE\_2009\_fungi$ [76], $macisaac\_yeast.v1$ [77], $matrix\_yeast$ and $scpd\_matrix$ [SCPD], respectively containing 177, 124, 38 and 24 motifs. Motifhits in the yeast genome were searched with FIMO (Find Individual Motif Occurences) [89]. The first step of this analysis was to compare the motifnames across the databases to see what amount of similarity between the motifs could be expected.

As figure 50 shows, just a fraction of the motifs is located in intersectional areas of the venn diagram. The reason for this may be that the experiments which led to the respective databases, revealed mainly distinct sets of TF binding motifs or that they were named only differently.

In order to select a unique set of TF motifs, the similarity between the respective Position-Weight-Matrices (PWMs) as well as the overlap between the motifhits of the motifs across the whole genome were analyzed. Concerning the motif-motif similarity, the pearson correlation between each pair of motifs — within a database and between the databases — was calculated and the motif comparison tool TomTom [90] was used to quantify the similarity measure between motifs. Generally, the pearson correlation revealed high values for the motif pairs with the same name, but there are also several cases where completely different named motifs have a correlation greater than 0.9. TomTom leads to a similar result: the motif pairs with a resulting p-value below a certain threshold are often those with the same motif name, but also pairs from whom one wouldn't expect such a similarity.

A meaningful example is the comparison of CAD1 from $macisaac\_yeast.v1$ and YAP7 from $JASPAR\_CORE\_2009\_fungi$ which is shown in figure 51(a). Although the motifs are rather long, they are significantly correlated and their similarity is clearly visible. By contrast, there are also motifs with the same name in different databases but completely different



■ **Figure 49** TSS difference between the different replicates for each condition and for each mapper. [see interactive TSS statistics]

motif logos for which an example can be seen in figure 51(b).
Even though there are many motif pairs with a high correlation which could be summarized according to their motif similarity measure, the overlap between their hits is not always as high as expected. Consequently, the aspired distinct set of motifs across the four databases contains hardly less motifs than the union of them would.
As figure 52 demonstrates, the usage of very strict cutoffs (e.g. Pearson correlation = 1, Motifhits overlap = 1 and TomTom=0.00001) would lead to the exclusion of only 2 motifs from the whole set of 363. Less stringent cutoffs, especially for the motifhit overlap, lead to the exclusion of 24 motifs which is significantly more but still only a fraction of the motif pairs whose names are the same. Consequently, the less stringent cutoffs were used for the final unique motif set and additionally, 42 motifs for which no motifhits were outputted by FIMO were excluded for further processing.

4.6.5 Differential motif binding

The application of CENTIPEDE to every motifhit in all experimental conditions enabled to detect differential motif binding across the measured timepoints during osmotic stress in

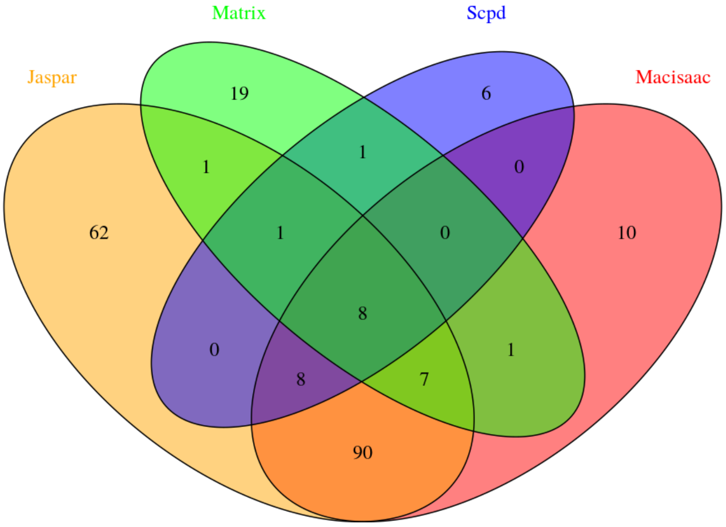
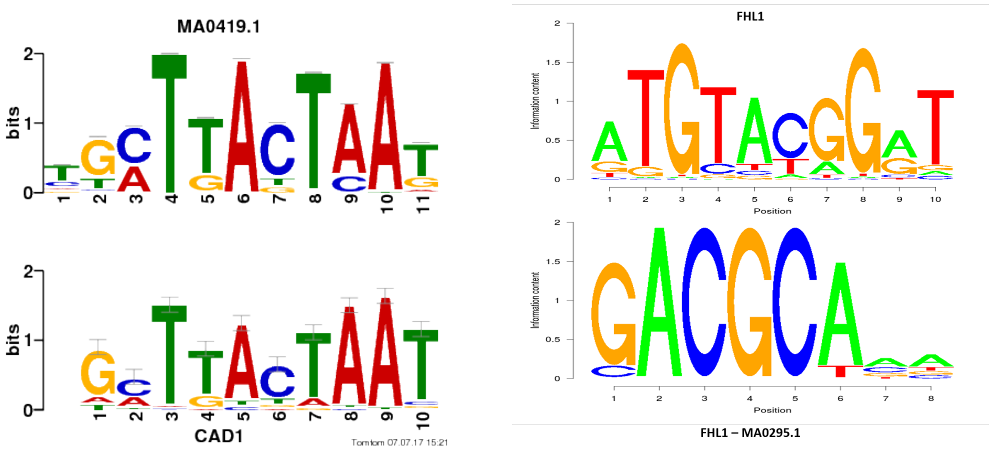


Figure 50 Overlap of motifnames between the 4 databases.



(a) Alignments of motifs CAD1 from *macisaac_yeast.v1* and YAP7 from *JASPAR_CORE_2009_fungi*

(b) Motifs of FHD1 in *JASPAR_CORE_2009_fungi* and *macisaac_yeast.v1*

Figure 51 Motif comparison between databases. [browse motif comparison *Motif analysis* → *Motif Comparison*]

yeast. [78] Additionally, we summed up the occupancy scores calculated by NucleoATAC [75] for each motifhit and analyzed them differentially. The problem with this approach is that occupancy scores are only calculated for positions within the peaks, that were called with MACS [73] previously. Positions outside of those peaks are expected to lie within rather closed chromatin, so that we used a default occupancy score of 0.8 for them.
In figure 53, volcano plots of the motifhits according to the CENTIPEDE method as well as the occupancy score method can be seen for the contrast 0 against 15 minutes. The plots

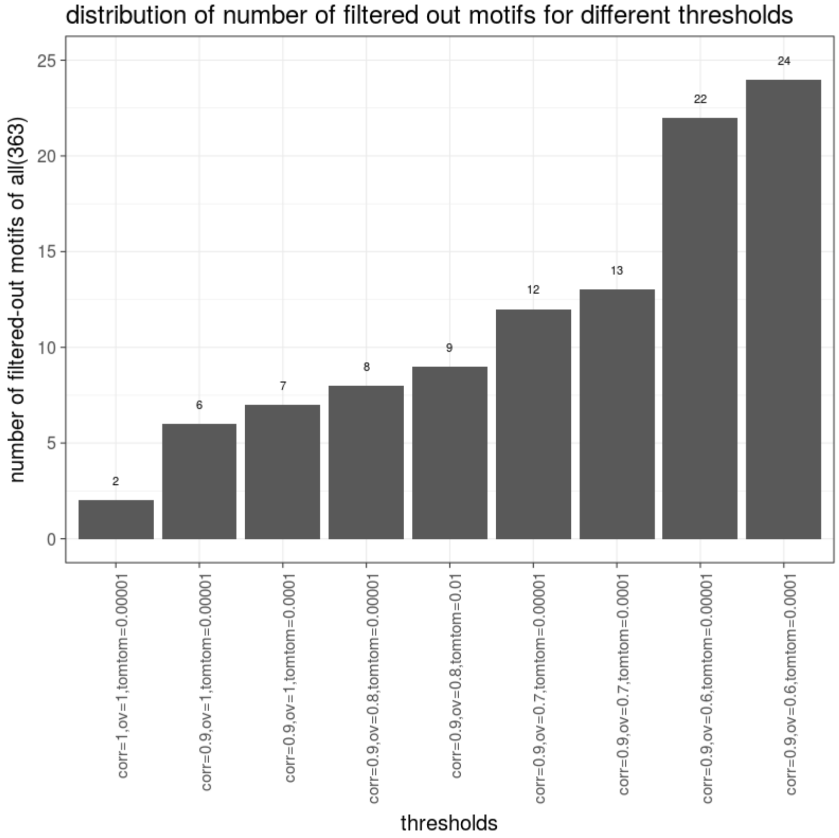
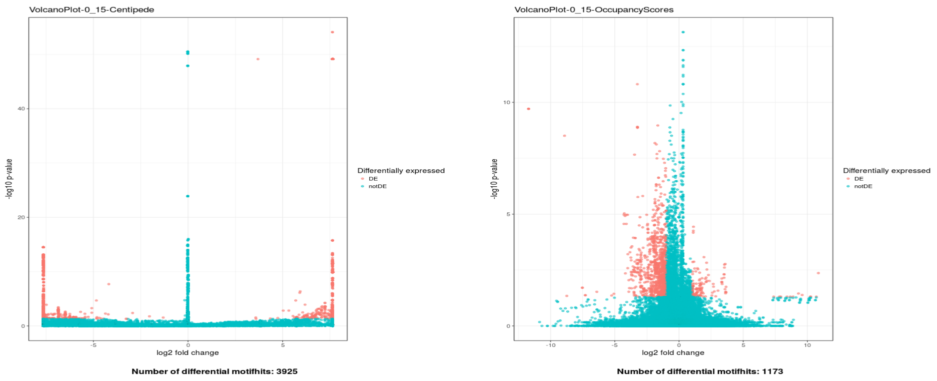


Figure 52 Number of excluded motifs corresponding to the specified cutoffs.



(a) Centipede

(b) Occupancy Score

Figure 53 Volcano plots of motifhits after 15 minutes according to Centipede and Occupancy scores. [see interactive figures]

show the respective differential motifhits across the whole genome for all motifs contained in the unique set with their fold change and p-value. Only those motifhits with a log fold change greater than 1 and an adjusted P-value smaller than 0.05 were declared as differential, incorporating all available replicates for each timepoint respectively. After 15 minutes of osmotic stress a maximal amount of differential motifhits of 3925 with CENTIPEDE and 1173 with the occupancy scores is reached. The maximal amount of differentially expressed genes was about 1800 after 15 min as well, implying that according to CENTIPEDE each of those genes could be controlled by several differential motifhits in its promoter region. How the distribution of differential motifhits on the genes really looks like, was investigated in the downstream analysis.

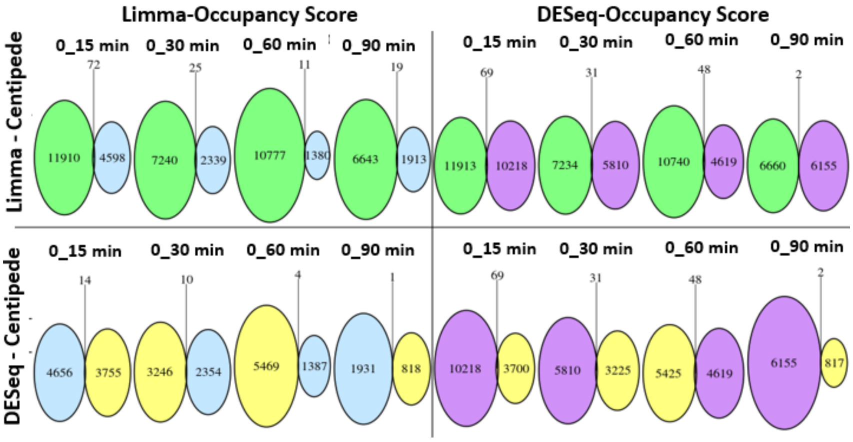


Figure 54 Overlap between the differential motifhits found with CENTIPEDE and with Occupancy scores.

Figure 54 illustrates the overlap between the differential motifhits identified by CENTIPEDE and the occupancy scores calculated by NucleoATAC. Apparently, the results of both tools differ significantly, since the overlap at all timepoints with all combinations of differential analysis methods is close to zero.

4.6.6 ATAC-Seq signal and occupancy scores surrounding motifhits

The assumption was that ATAC-Seq signals and accordingly the calculated occupancy scores from NucleoATAC indicate precisely whether a motifhit on the DNA is accessible for TF binding. In order to investigate this hypothesis, the distribution of ATAC-Seq fragment start positions and the occupancy scores surrounding motifhits were pictured.

The transcription factor FHL1 was identified to show significant enrichment among the differentially expressed genes after 15 and 30 minutes of osmotic stress by Babazadeh et al. [79]. Figure 55 shows the distribution of ATAC-Seq start positions and occupancy scores for the motifhits of FHL1 after 30 minutes of osmotic stress in yeast. The left column of the figure shows the mean count of starting ATAC-Seq fragments in windows of 100 bp upstream and downstream surrounding the motifhits, diagrammed as the percentage of all fragment start positions in those regions. This distribution is grouped according to the occupancy score at the start position of the motifhit. Occupancy scores in +/- 50 bp windows surrounding the motifhits are shown as boxplots in the adjacent column, also grouped by the occupancy score of the motifhit start itself. The total ATAC-Seq fragment start counts in the regions surrounding motifhits can be seen in the red heatmaps in the second column from the right, whereby the x axis indicates again the distance to the motifhit and the different motifhits are layed on the y axis. In the upper heatmap those hits are sorted according to the occupancy score at their start position and in the lower heatmap the order

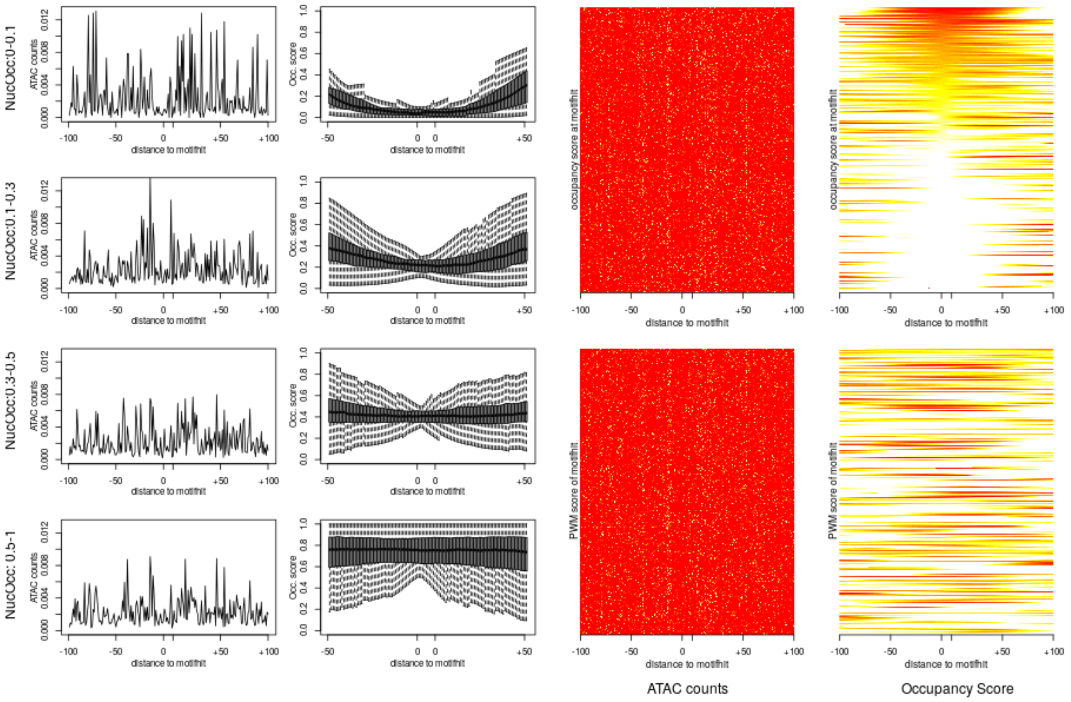


Figure 55 ATAC-Seq signal and occupancy scores for motifhits of FHL1 from *macisaac_yeast.v*1. [see this figure for all motifhits and replicates *Motif analysis* → *Motifhit analysis*]

of motifhits follows their respective PWM score. The same two heatmaps for occupancy scores are displayed in the right column. Obviously, there is a continuous color gradient visible in the plot where the motifhits are sorted by their occupancy score.

Although the trends of the ATAC-Seq data are heading in the right direction, we were actually expecting to see more precise signals in the raw data. The promised base-pair accurate resolution of TF binding sites don't seem to be achievable by this data.

4.7 Conclusion and outlook

The current results of the open chromatin project meets our previous expectations of an explanation of differential gene expression by differential motifhits in their promoter regions partially. Overall, there are only few differential motifhits identified which are located in the promoter region of differentially expressed genes.

Figure 56 shows for each gene that is differentially expressed after 15 min of osmotic stress, the log fold change of the respective most differential motifhit in the promoter region relative to the log fold change of the gene. Apparently, in most of the cases the logFC of the respective motifhit is close to 0, meaning that no differential motifhit induces the strong fold change in the expression of the gene. In the other cases, where a differential motifhit was idenified there is no great variance in the logFC which goes hand in hand with the fact that Centipede outputs mainly probabilities to be bound of 0 and 1.

The percentage of DE genes that can be explained with a regulation by x distinct TFs is shown in figure 57. On the left, the percentage of genes is plotted for differential motifhits identified with Centipede, while it is drawn for differential hits according to the summed up occupancy scores of the motif positions on the right. Not even 50% of all differentially

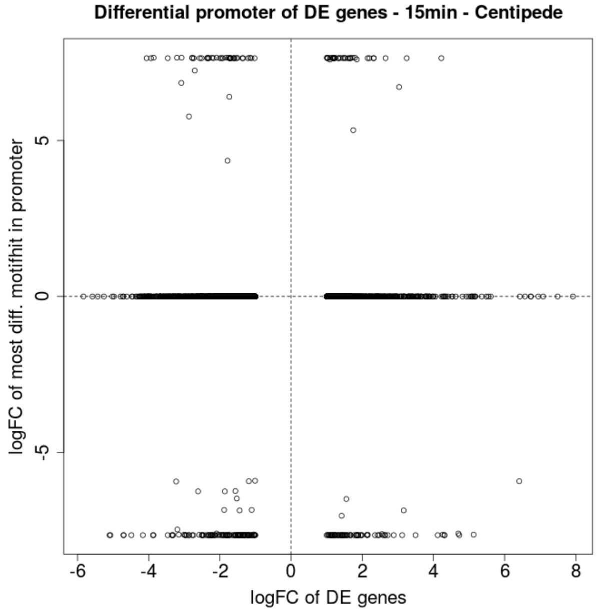
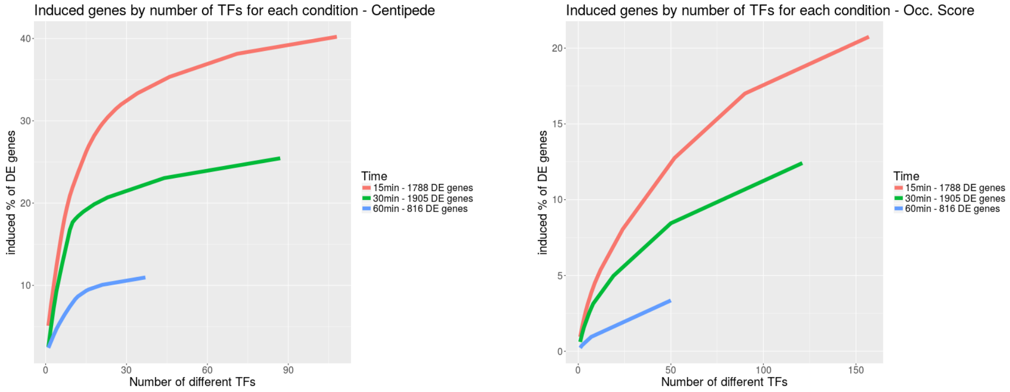


Figure 56 LogFC of most differential motifhits in promoter regions of DE genes relative to logFC of genes. [see interactive figures]



(a) Centipede

(b) Occupancy Score

Figure 57 Necessary number of TFs to explain differential expression in the different conditions. [see interactive figures]

expressed genes can be explained by differential motifhits in any case of condition or exploited motifhit binding prediction method. The maximal percentage of induced genes is reached with Centipede after 15 minutes of osmotic stress, where about 100 TFs induce 40% of all DE genes.

For each TF that has a differential motifhit in the promoter region of a gene in any conditions, the number of induced up- and downregulated genes is illustrated for all timepoints (15 min, 30 min, 60 min) in figure 58. Light colours indicate many induced genes in a condition, while red indicates only few genes. It can be seen that few TFs regulate many genes and many TFs regulate only very few genes. It is conspicious that the TFs which regulate many genes seem to be time specific, meaning that the motifs with many induced genes at one

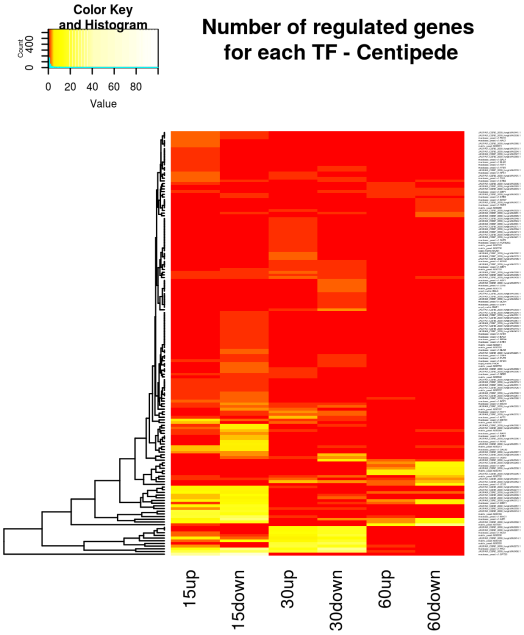


Figure 58 Number of regulated genes for different motifs. [see interactive figures]

timepoint often differ from those motifs with many induced genes at another timepoint, but that the same TFs regulate up- and downregulated genes simulataneously.
As already mentioned, a general problem of our results is that only a small number of differential motifhits are found in promoter regions of differentially expressed genes. A possible reason might be replicate inconsistency that could lead to lower fold changes and higher p-values for actual differential motifhits. Furthermore, a comparatively small portion of all identified differential motifhits are located within the promoter regions of DE genes: from the 3925 differential Centipede motifhits after 15 min only 719 lie within such promoter regions.

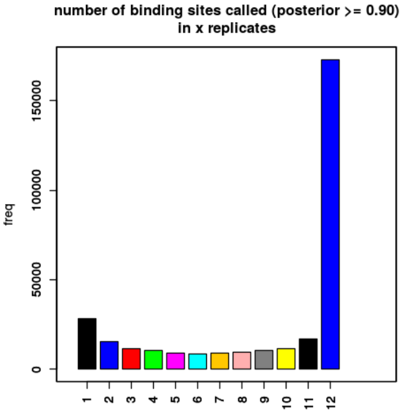


Figure 59 Number of motifhits predicted to be bound in x replicates

Beyond, a reason might be that there are many motifhits in the promoter regions, which are, however, not differential. As figure 59 shows, most of the motifhits in the genome which are predicted to be bound in any replicate are predicted to be bound in all samples, i.e. regardlass of the stress condition, many TFs are predicted to be bound.

In our project setup, the motifhits were searched in the whole genome with FIMO. Consequently, interesting binding sites within promoters were maybe not detected due to very low p-values. An opportunity for further investigations would be to search only in differential regions of the genome for motifhits to reach a higher sensitivity for the sites of interest. Moreover, improvements may be possible regarding the NucleoATAC predictions. Until now, the occupancy scores were calculated only within the peaks called by MACS. To gain precise occupancy information for all motifhits, NucleoATAC could be called for the whole genome without any restrictions.

With a few improvements in the sensitivity of the motifhit detection and their differential analysis, it is possible that ATAC-Seq enables a precise explanation of differential expression by differential transcription factor binding.

References

[66] Robert E. Thurman et al. “The accessible chromatin landscape of the human genome”. In: *Nature* 489.7414 (2012), pp. 75–82. ISSN: 0028-0836. DOI: 10.1038/nature11232.

[67] Peter J. Park. “ChIP-seq: advantages and challenges of a maturing technology”. In: *Nature reviews. Genetics* 10.10 (2009), pp. 669–680. ISSN: 1471-0064. DOI: 10.1038/nrg2641.

[68] Anirudh Natarajan et al. “Predicting cell-type-specific gene expression from regions of open chromatin”. In: *Genome research* 22.9 (2012), pp. 1711–1722. ISSN: 1088-9051. DOI: 10.1101/gr.135129.111.

[69] Jason D. Buenrostro et al. “Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position”. In: 10.12 (2013), pp. 1213–1218. ISSN: 1548-7091. DOI: 10.1038/nmeth.2688.

[70] Maria Tsompana and Michael J. Buck. “Chromatin accessibility: a window into the genome”. In: *Epigenetics & chromatin* 7.1 (2014), p. 33. DOI: 10.1186/1756-8935-7-33.

[71] P. G. Giresi et al. “FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin”. In: *Genome research* 17.6 (2007), pp. 877–885. ISSN: 1088-9051. DOI: 10.1101/gr.5533506.

[72] Alan P. Boyle et al. “High-Resolution Mapping and Characterization of Open Chromatin across the Genome”. In: *Cell* 132.2 (2008), pp. 311–322. ISSN: 00928674. DOI: 10.1016/j.cell.2007.12.014.

[73] Yong Zhang et al. “Model-based analysis of ChIP-Seq (MACS)”. In: *Genome biology* 9.9 (2008), R137. DOI: 10.1186/gb-2008-9-9-r137.

[74] Kristin Brogaard et al. “A map of nucleosome positions in yeast at base-pair resolution”. In: *Nature* 486.7404 (2012), pp. 496–501. ISSN: 0028-0836. DOI: 10.1038/nature11142.

[75] Alicia N. Schep et al. “Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions”. In: *Genome research* 25.11 (2015), pp. 1757–1770. ISSN: 1088-9051. DOI: 10.1101/gr.192294.115.

[76] Albin Sandelin et al. “JASPAR: an open-access database for eukaryotic transcription factor binding profiles”. In: *Nucleic acids research* 32.Database issue (2004), pp. D91–4. ISSN: 1362-4962. DOI: 10.1093/nar/gkh012.

[77] Kenzie D. MacIsaac et al. “An improved map of conserved regulatory sites for Saccharomyces cerevisiae”. In: *BMC bioinformatics* 7 (2006), p. 113. ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-113.

[78] Roger Pique-Regi et al. “Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data”. In: *Genome research* 21.3 (2011), pp. 447–455. ISSN: 1088-9051. DOI: 10.1101/gr.112623.110.

[79] Roja Babazadeh et al. “The yeast osmostress response is carbon source dependent”. In: *Scientific reports* 7.1 (2017), p. 990. ISSN: 2045-2322. DOI: 10.1038/s41598-017-01141-4.

[80] Thomas Bonfert et al. “ContextMap 2: fast and accurate context-based RNA-seq mapping”. In: *BMC bioinformatics* 16 (2015), p. 122. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0557-5.

[81] Daehwan Kim et al. “TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions”. In: *Genome biology* 14.4 (2013), R36. DOI: 10.1186/gb-2013-14-4-r36.

[82] Alexander Dobin et al. “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics (Oxford, England)* 29.1 (2013), pp. 15–21. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts635.

[83] Daehwan Kim, Ben Langmead, and Steven L. Salzberg. “HISAT: a fast spliced aligner with low memory requirements”. In: *Nature methods* 12.4 (2015), pp. 357–360. ISSN: 1548-7105. DOI: 10.1038/nmeth.3317.

[84] Li Ni et al. “Dynamic and complex transcription factor binding during an inducible response in yeast”. In: *Genes & development* 23.11 (2009), pp. 1351–1363. ISSN: 1549-5477. DOI: 10.1101/gad.1781909.

[85] Ben Langmead and Steven L. Salzberg. “Fast gapped-read alignment with Bowtie 2”. In: *Nature methods* 9.4 (2012), pp. 357–359. ISSN: 1548-7105. DOI: 10.1038/nmeth.1923.

[86] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics (Oxford, England)* 26.1 (2010), pp. 139–140. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp616.

[87] Simon Anders and Wolfgang Huber. “Differential expression analysis for sequence count data”. In: *Genome biology* 11.10 (2010), R106. DOI: 10.1186/gb-2010-11-10-r106.

[88] Matthew E. Ritchie et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic acids research* 43.7 (2015), e47. ISSN: 1362-4962. DOI: 10.1093/nar/gkv007.

[89] Gabriel Cuellar-Partida et al. “Epigenetic priors for identifying active transcription factor binding sites”. In: *Bioinformatics (Oxford, England)* 28.1 (2012), pp. 56–62. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btr614.

[90] Shobhit Gupta et al. “Quantifying similarity between motifs”. In: *Genome biology* 8.2 (2007), R24. DOI: 10.1186/gb-2007-8-2-r24.

5 Alternative Splicing

by Dominik Müller, Alexandra Morscher and Markus Joppich

The molecular mechanism of pre-mRNA splicing is an essential part of the protein biosynthesis in most higher eukaryots. It describes the splitting and reassembling process of the pre-mRNA to its coding sequence containing successor, the mature mRNA. One pre-mRNA transcript can be transformed into various products to generate different protein isoforms. This shows the variability of splicing wherefor the mechanism is fully called *Alternative Splicing*. The assembly patterns (exon order) of mRNA vary highly between different organism and tissues, which shows the complexity of the mechanism but leads to difficulties in the determination of the examined pre-mRNAs, and in general terms also of the way of operation of the system [91].

During mRNA editing certain pre-mRNA fragments are cut out. This process is called splicing and is one of the key steps in the processing of the initial messenger RNA (mRNA), which also includes the 5'-end capping and the 3'-end capping, which takes place during the transcription from the DNA (cotranscriptional) and before its translation [**2**].

6 The Concept of (Alternative) Splicing

6.1 Definition

During mRNA editing certain pre-mRNA fragments are cut out. This process is called splicing and is one of the key steps in the processing of the initial messenger RNA (mRNA), which also includes the 5'-end capping and the 3'-end capping, which takes place during the transcription from the DNA (cotranscriptional) and before its translation [91]. During the splicing process the introns of a pre-mRNA, which are defined as gene segments that are present in the pre-mRNA but absent from the final mRNA thus a consequence of splicing, will be removed or *spliced out* of the pre-mRNA [92] [93] [91]. The resulting gene segments, which aren't removed, are defined as exons [92] [93] [91] [94] [95]. After the removal, the exon parts are connected to each other and form the main sequence of the mature mRNA, which will be transported to its translation into a protein [92] [93] [91] [94] [95]. Splicing of the mRNA usually occurs in the nucleus of a cell and mostly in eukaryotic organisms [92] [93]. The complex behind splicing is for most cases the spliceosome which contains a variety of RNA and protein biomolecules [92] [93] [91] [94] [95]. The *borders* of an intron are represented by the Donor and Acceptor site [93] [91] [95] [96] [97]. The donor site is the front or start side of an intron and is located at the 5'-end of the intron whereas the acceptor site is the end side of an intron and is located at the 3'-end of the intron [93] [91] [95] [96] [97]. Through these two sites the spliceosome complex is able to identify and remove the intronic segment out of the pre-mRNA [93] [91] [94] [95] [96] [97]. Afterwards the spliceosome connects the two exons, which are located before and after the excluded intron, and closes the gap between them [92] [93] [91] [94] [95] [96] [97]. There will be a more detailed reflection of the molecular mechanism of the spliceosome later. Additionally there exist two more rare mechanisms how splicing can occur [92] [96]. The first one emerge if an intron forms a ribozyme, which performs the functions of the spliceosome by itself without requiring of the spliceosome [92] [96]. Therefore the intron splices itself from the pre-mRNA why this is called *self-splicing* [92] [96]. The other rare mechanism *tRNA splicing* occurs in the tRNA and performs the splicing reaction through a different biochemistry pathway as the self-splicing or spliceosome [92] [96]. For summary, in the *simple* splicing model a gene is only able to produce the identical mature mRNA and protein product by splicing all its introns out and including all its exons in the final mRNA. However in higher eukaryotic organism it is commonly observed that a single gene is able to produce multiple different proteins [92] [93] [91] [94] [95] [96] [97]. This is the result of the alternative splicing process. In the alternative splicing process, additional to the introns, particular exons of genes will be included or excluded from the pre-mRNA as well, which creates the possibility of a high number of different exon compositions or resulting isoforms by taking out one or multiple exons [92] [93] [91] [94] [95] [96] [97]. These diverse compositions of exons are the cause that a single gene may have the possible varying set of proteins instead of just coding for one protein. Through exon in- or exclusion, it is possible that key regions for the functionality of a protein can be removed or added, too [93] [94] [95] [97]. This scenario results that it is possible that two proteins from the same gene are able to have different functions. Alternative splicing is regulated by a variety of enhancers, which promote the usage of the corresponding splice site, and silencers, repressing the usage of the corresponding splice site, located in introns as well as in exons [92] [93] [91] [94] [95] [96] [97]. Due to alternative splicing multiple compositions of exons can be formed. In figure 1 five most common patterns of alternative splicing are shown and explained.

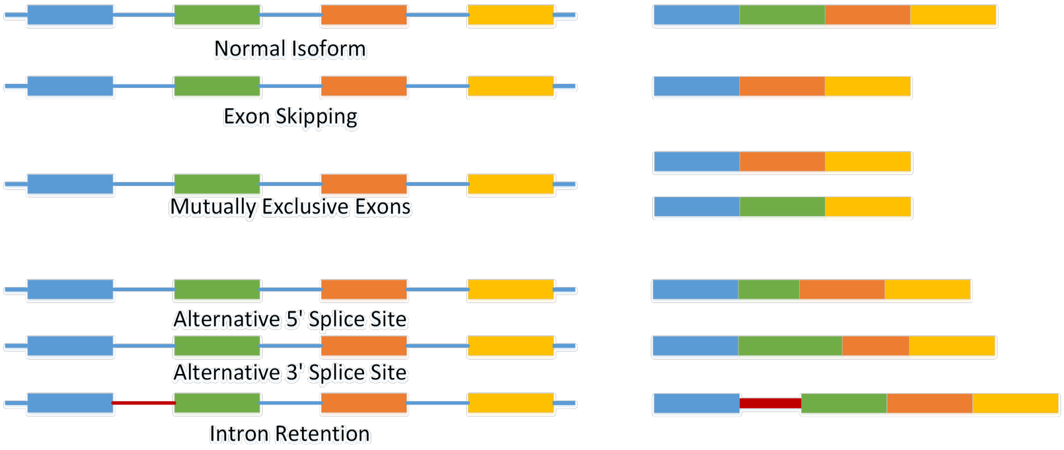


Figure 60 Figure 1: Illustrative schema of most common alternative splicing patterns. The colored boxes are representing exons and the blue lines are representing introns. The left side shows the exon/intron composition of the pre-mRNA and the right side the corresponding spliced mature mRNA. The *normal isoform* is the simple splicing mode in which all introns are spliced out and all exons are included in the final mRNA [93] [91] [94] [97] [98]. The *exon skipping* pattern indicate the case in which one or multiple exons are completely excluded and skipped in the final mRNA [93] [91] [94] [97] [98]. In this example the green exon is completely removed from the mRNA. The second pattern demonstrates *mutually exclusive exons* where one of two exons (e.g. green and orange) is included in the mRNA but not both [93] [91] [94] [97] [98]. The *alternative 5'* (donor) or 3' (acceptor) splice site indicate the usage of an alternative splice site and therefore changes the 3' (donor) or 5' (acceptor) border of the upstream (donor) or downstream (acceptor) exon [93] [91] [94] [97] [98]. This results in a reduction of the exon size or in an increase of exon size by including intronic sequence. The last common alternative splicing pattern is the *intron retention* in which a complete intron is included in the final mRNA [93] [91] [94] [97] [98].

6.2 Molecular Mechanism and Regulation

The *cutting* and *sewing* process of the spliceosome complex is possible due to the teamwork of a large variety of RNA's and proteins [93] [91] [96] [97] . The spliceosome consists of five different small nuclear ribonucleoproteins (snRNP?s) [93] [91] [96] [97]. Each snRNP is a complex of a single RNA and multiple proteins [93] [91] [96] [97]. In order to create the large spliceosome complex, four of the five snRNP compartments start merging after the snRNP U1, the auxiliary factor U2AF and the splicing factor SF1 bind to an intron by identifying conserved nucleotide tags [93] [91] [96] [97]. Therefore U1 identifies and binds to the GU tag of the donor site (5'-end) of the intron, U2AF to the AG tag of the acceptor site (3'-end) and to the PPT (polypyrimidine tract, upstream of the 3'-end which is a region high in pyrimidines) of the intron and SF1 to the A tag of the branch point site (upstream after the polypyrimidine tract near the 3'-end) of the intron [93] [91] [96] [97]. After binding of these starting factors to the intron the spliceosome complex starts to assemble which is illustrated in figure 2 [91]. The first step of the assembly is the exchange of SF1 to snRNP U2 [93] [91] [96] [97]. The resulting interactions between U1 and U2 conclude with the binding of U1 to U2 and creating the pre-spliceosome also called *complex-A* [93] [91] [96] [97]. After this step the U4/U5/U6 trimer complex binds to the U1/U2 complex and forms the pre-catalytic spliceosome [93] [91] [96] [97]. Then the re-catalytic spliceosome releases the U1 and U4 snRNP out of its complex, U6 moves to the 5'-splice site of the released U1 and the spliceosome become a catalytic spliceosome [93] [91] [96] [97]. This results into the first of two transesterification in which the 5'-splice site is cleaved and the intron forms to a lariat [93] [91] [96] [97]. Due to the following second transesterification the 3'-splice site of the intron is cleaved and the two exons are ligated [93] [91] [96] [97]. Then the spliced intron is released and the splicing process is done.

The process of skipping entire exons of mRNA are forms of alternative splicing [92]. There has to be a complex regulation for exon exclusion or inclusion to indicate which exon should be cleaved by the spliceosome and which exon should be included in the mRNA [93] [91] [96] [97]. Therefore splicing is regulated by trans-acting (regulation by another molecule) proteins, which act as activators or repressors, and their corresponding cis-acting (regulation by the same molecule) regulatory binding sites on the pre-mRNA, which increase or decrease the probability that the particular exon will be included or excluded in the mRNA [93] [91] [96] [97]. There are a large variety of different trans-acting proteins or splicing factors which bind to the corresponding cis-acting regulatory sites [93] [91] [96] [97]. These regulatory sites on the pre-mRNA can be classified into two categories. The first category is the splicing silencers which reduce the probability that the specific splice site will be used and therefore reduce the probability that the exon will be included [93] [91] [96] [97]. These splicing silencers can be located in the intron which are called ISS (intronic splicing silencers) or in the exon which are called ESS (exonic splicing silencers) [93] [91] [96] [97]. The other category are the splicing enhancers which increase the probability of using the specific splice site and therefore increase the probability that the exon will be included [93] [91] [96] [97]. Also the splicing enhancers can be located in the intron which are called ISE (intronic splicing enhancers) or in the exon which are called ESE (exonic splicing enhancers) [93] [91] [96] [97]. The corresponding trans-acting splicing factors could be classified into activators which binds to splicing enhancers and repressors which binds to splicing silencers. There are a variety of proteins which bind to splicing enhancers as well as splicing silencers and therefore would be classified as activators and repressors [93] [91] [96] [97].

Additionally the regulation of alternative splicing is a complex system of various splicing factors which can simultaneously increase as well as reduce the probability of including an exon or intron (intron retention) into the mRNA. Additionally these splicing factors are influenced by signaling pathways, development stage, cell type and sex [93] [97]. Therefore the understanding and ideally the prediction of alternative splicing outcomes are complex

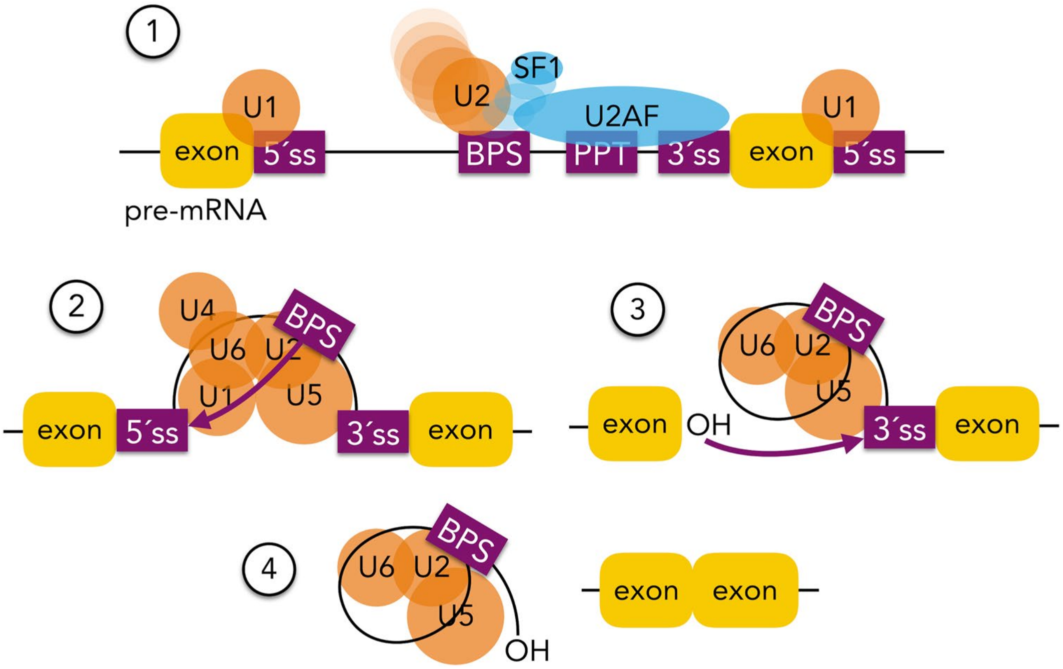


Figure 61 Figure 2: Illustration of the binding, assembling and cleaving mechanism of the spliceosome [91]. (1) Exchange of SF1 to the snRNA U2 at the branch point site (BPS) which results in the interaction of U1/U2 and the formation of the pre-spliceosome [91]. (2) Forming the pre-catalytic spliceosome after binding of U4/U5/U6 and starting the 5'-splice site cleavage by releasing U1/U4 and performing the first transesterification [91]. (3) Resulting intron lariat due to the 5'-splice site cleavage [91]. Performing the second transesterification and cleavage of the 3'-splice site [91]. (4) Finished splicing process in which the intron lariat is released and the exons are ligated [91].

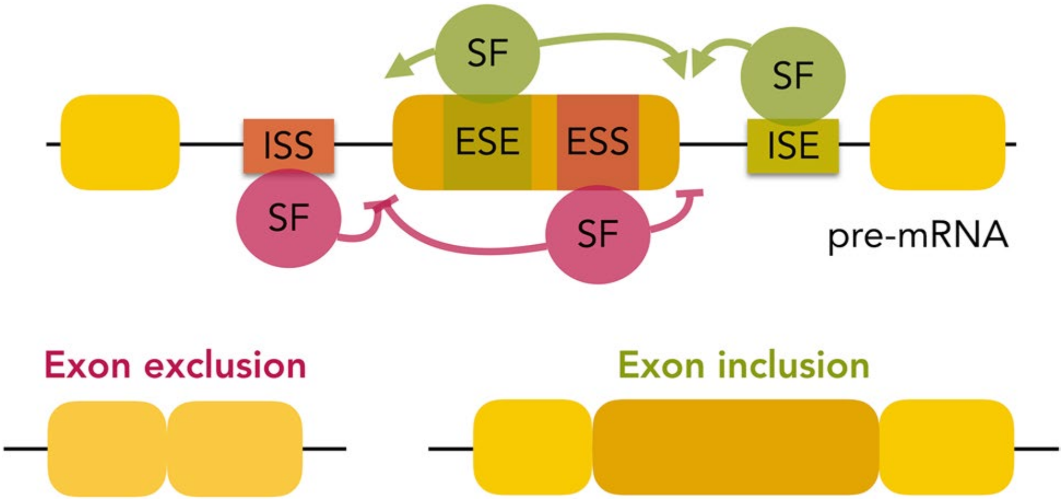


Figure 62 Figure 3: Representation of the alternative splicing regulation [91]. Multiple splicing factors (SF) can bind to exonic splicing silencers (ESS) or splicing enhancers (ESE) and to intronic splicing silencers (ISS) or splicing enhancers (ISE) due to reduce or increase the probability of exclusion or inclusion of the corresponding exon [91].

and challenging competition.

6.3 Alternative Splicing Prevalence

Due to alternative splicing, one gene can be the source of multiple different mRNAs and therefore proteins [92] [93] [97]. Thus one human cell with around 20.000 genes is able to code for 500.000 to 1.000.000 different proteins which is an increase of up to 50 times more proteins out of the same DNA [97]. An extreme example of the huge proteomic increase by alternative splicing is the Drosophila melanogaster gene DSCAN [93] [97]. This single gene is the source of 38.016 different isoforms [93] [97]. This explains that the number of genes in eukaryotic organisms doesn't indicate the number of different expressed proteins. In humans the actual prevalence of alternative splicing varies between 40-60% of all genes according to current research [93] [97] [98]. In a more detailed analysis one group discovered that more than 80% of the genes from the human chromosome 21 and chromosome 22 do alternative splicing [97] . These statistics indicate that alternative splicing is more the rule than an exception as first assumed [97]. Half of all cases of alternative splicing events, which effects the coding sequence, result into altering the reading frame. [97]. One third of alternative splicing events lead to nonsense-mediated decay (controll mechanism which identifies premature stop codons and stops the expression of these mRNAs) of the resulting mature mRNA [97].

Due to the fact, that the alternative splicing estimates strongly vary depending on publications, we decided to calculate an alternative splicing prevalence estimate by ourselves. Therefore we obtained the gene annotation data of the three organism human, mouse and rat from the two large databases NCBI and Ensembl. To estimate the alternative splicing prevalence we simply defined that a gene is processing alternative splicing if it has more than 1 annotated transcript. Also we obtained the gene *biotype* annotation of Ensembl by which it was possible to identify the only protein coding genes. Additionally we created another subset of genes which has at least one ortholog in one or both of the other studied organism. As seen in figure 4, the current alternative splicing processing human genes take around 37% in Ensembl and more than 50% in NCBI. The percentage of alternative splicing genes for all three species is higher in NCBI than in Ensembl. This could be explained by the fact, that Ensembl has nearly 66.000 human genes in its database whereas NCBI holds only 60.000 human genes. The same difference in gene numbers can be observed in the two other organism, too. Furthermore the protein coding genes and genes, with at least one ortholog in one of the other two organism, get closer to the current alternative splicing estimates. More than 80% of human protein coding genes and nearly 75% of mouse protein coding genes process alternative splicing which confirms the assumption that alternative splicing is more the rule than an exception. Only the rat genes seem to be processing normal splicing (under 25% in Ensembl). However this current situation in the data could be artificial. The two organism human and mouse belong to the most detailed experimental characterized organisms (e.g. GENCODE) whereby it is conceivable that rare alternative spliced isoforms in rat genes could have been missed in experiments so far.

7 Evolution of Splicing

There are vast differences between levels of alternative spliced genes in the kingdoms. For example plants show lover levels, where intron retention is present mostly with ˜30% [99]. By comparing alternative splicing patterns, the highest similarity consists between same tissue types of different organisms while different cell types in the same organism show lower

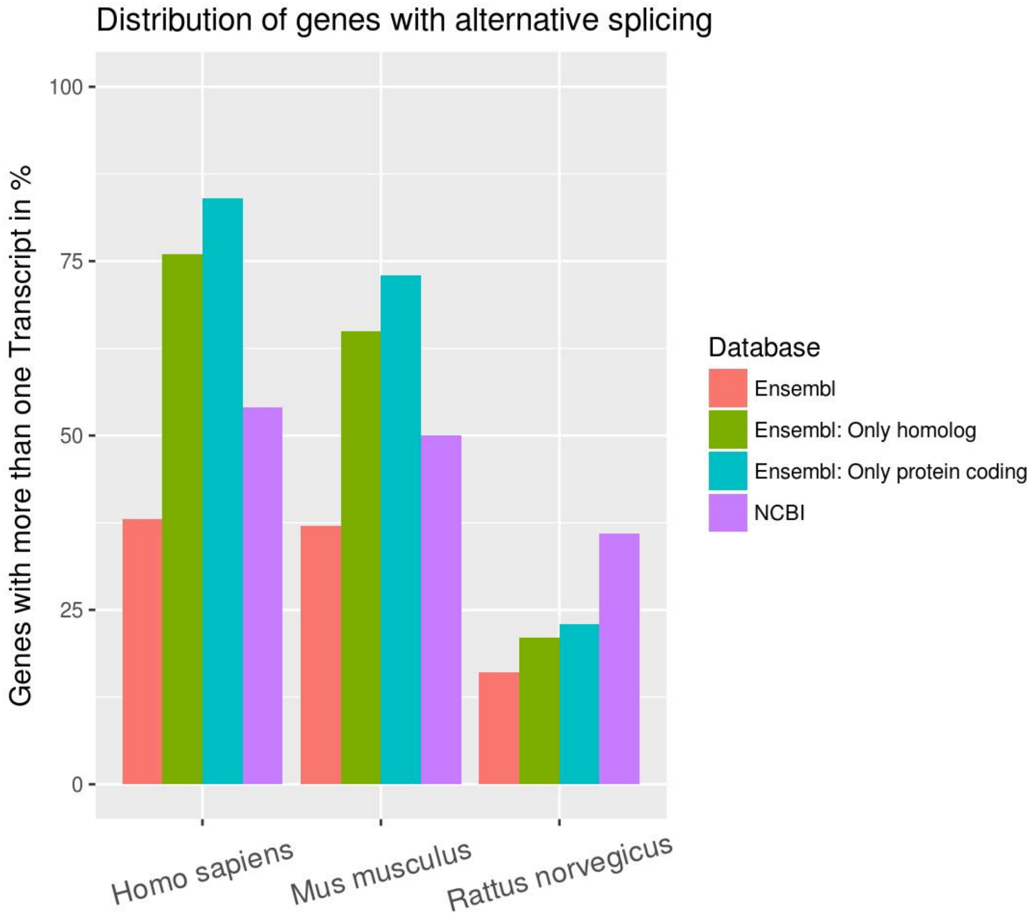


Figure 63 Figure 4: Alternative splicing prevalence of the three organism human (homo sapiens), mouse (mus musculus) and rat (rattus norvegicus). The gene annotation was obtained from the two databases Ensembl and NCBI. A gene was defined as alternative spliced when two or more transcripts were annotated.

similarity (Figure 5). Especially the brain, muscle and heart show strong conserved splicing signatures. Here we speak of *tissue-dominated clustering*, whereas other tissues show more of a *species-dominated clustering* [100]. Some patterns of splicing can be conserved for millions of years. This shows that also the splicing signature conservation varies greatly in tissue kinds. The comparison of the splicing mechanism over all kingdoms of the eucaryots suggests that alternative splicing might have existed in the early eucaryotic evolution and the *ancestor* most likely showed similarities to mammals in terms of splicing [99]. Relating to spliced exons, the evolution of spliced genes is subject to recent exon creation and loss [100]. The mechanism of exon shuffling occurs when a new exon is inserted or duplicated into a gene, whereas exonisation creates an exon *out of thin air*, when genomic sequences become exons. Transition describes the process where consecutive spliced exons become alternative spliced ones, but also the reverse may occur.

7.1 Methods and Databases

7.1.1 Prediction of Alternative Splicing Site:

There are many prediction tools for alternative splicie sites. Most of them work based on differential RNAseq analysis and statistic evaluations.

The open source software bioconductor offers packages for the statistical language R [101]: DEXseq focuses on finding alternative splice sites using RNAseq exon counts from different samples. It is based on the negative binomial distribution to estimate the variance between the samples. It also contains visualization and exploration functions [102]. JunctionSeq is another package which follows DEXseq to find differential usage of exon and splice junctions [**isar14**]. Another R package is Solas which works with RNAseq and anotations and predicts alternative exons and quantifies alternative splice forms in one or more cell samples [103].

Apart from R packages, there is a wide diversity of tools which stand on their own. Some of the statistical analysis of RNAseq based tools are Cufflinks and Cuffdiff from the Cufflinks tool collection [104], MISO [105], ALEXA-Seq [106] and MATS [107]. Cufflinks annotates exons from RNAseq with given annotations and therefor is able to detect new exons. Cuffdiff then computes a differential analysis for different samples on these annotations. It works solely with the statistical analysis of read counts [104] [108]. MISO (Mixture of Isoforms) detects isoforms out of RNAseq by using Bayesian interference for the probability that a read originated from a given isoform [105]. The Multivariate Analysis of Transcript Splicing (MATS) calculates the P-value and the false discovery rate so that a genes isoform ratio difference of two samples goes over a certain threshold. The user defines this threshold [107]. The paired-end RNAseq analyzing tool, ALEXA-Seq, uses the provided a database of sequence features and therefore supplies information on the sequence support, conservation and protein coding effect of each resulting feature. It also includes a visualization tool [106].

Slightly different from other tools, AltAnalyze uses microarray, single-cell and bulk RNAseq data. It detects novel and known alternative exon expression, domain compositions and miRNA targeting. It comes with different features like an alternative exon visualization and a graphical user interface or command-line execution [109].

Another method, the empirical Bayes change-point model (EBChangePoint) detects alternative 3' and 5' splice sites. It doesn't need prior information but such can be integrated through a systematic framework to improve the prediction. The empirical Bayes model pools information across genes and so detects the alternative splice Site [110].

SplAdder takes RNAseq alignments and annotation files as input to increase the annotation based on the RNAseq data and then identifies alternative splicing events trough the aug-

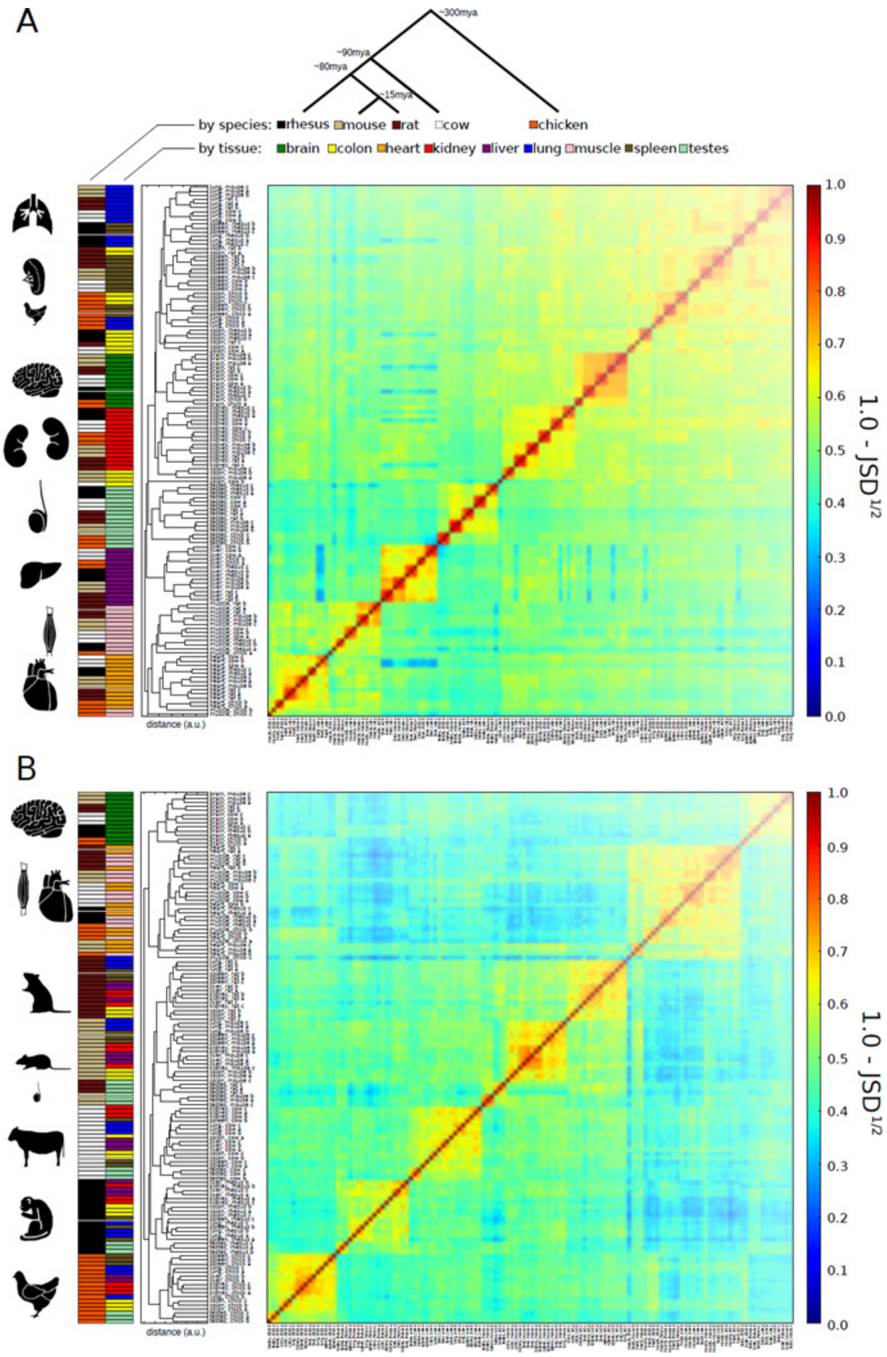


Figure 64 Figure 5: Conservation of expression signatures in all tissues and of alternative splicing signatures in some tissues [100] A) Clustering of FPKM of singleton orthologous genes. B) alternative splicing detected and clustering of samples based on PSI values of exons in genes conserved to chicken. (Figure from 'Evolutionary dynamics of gene and isoform regulation in Mammalian tissues' by Merkin J1, Russell C, Chen P and Burge CB)

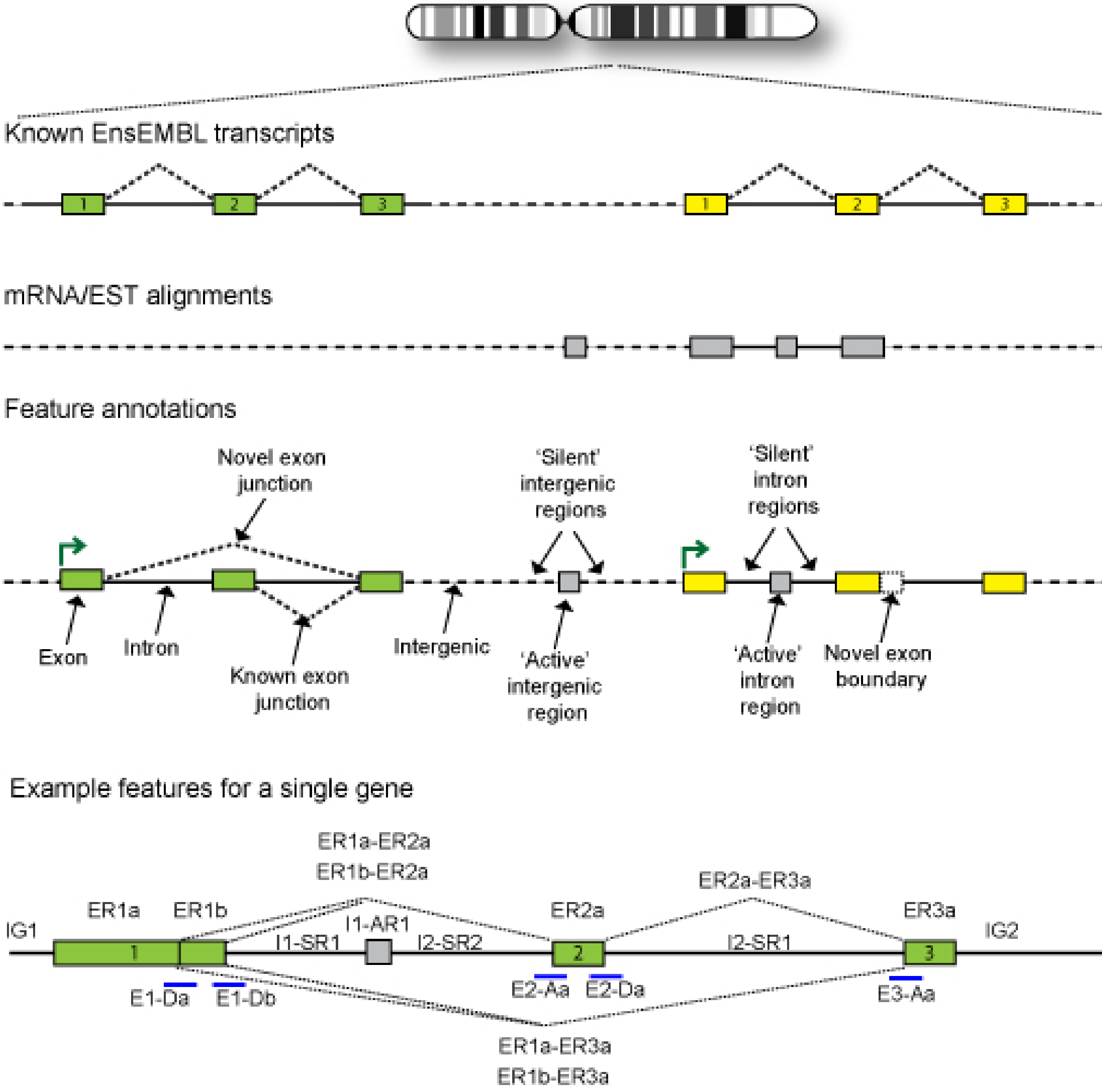


Figure 65 Figure 6: It shows the creation process of the ALEXAseq annotation database. Same colored rectangles represent exons of a single transcript/gene whereas grey ones are new detected annotations [106].

mented annotation graph. It quantifies these on the RNAseq data and searches for significant differences between samples [111].

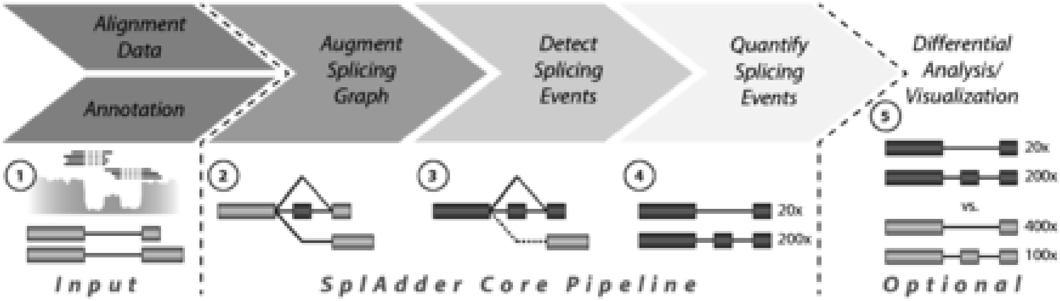


Figure 66 Figure 7: SplAdder analysis flowchart. The workflow consist of (1) integrating annotation and RNA-Seq data, (2) generating an augmented splicing graph, (3) extraction of splicing events, (4) quantifying the extracted events and optionally (5) the differential analysis and visualizations

7.1.2 Visualization Tools:

There are specialized tools for visualization of alternative splicing events. The Modeling Alternative Junction Inclusion Quantification (MAJIQ) together with Voila detects, quantifies, and visualizes local splicing variations from RNAseq. The MAJIQ Builder module determines splice graphs and Local Splice Variations (LSV) both known and novel. The second module MAJIQ Quantifier defines relative abundance (PSI) of LSVs and changes in relative LSV abundance (delta PSI) between two samples,conditions or replicates. Voila is the visualization package that takes the output of the other two modules and creates interactive summary files with gene splice graphs, LSVs, and their quantification, using interactive D3 components and HTML5 [112]. Astalavista extracts alternative splicing sites from exon/intron site annotations and compares given transcripts to find splicing structure variations and alternative splicing events. It is a visualization tool for transcriptional landscapes [113].

7.1.3 Annotation Databases:

There are a lot of databases which provide genomic annotations of all kinds. To get intron-exon annotations, the annotations of whole genome annotations can be filtered for the intron-exon tags. The geneids used are mostly from the Ensembl database or, if self-generated by the database, it usually provides a converting file or list for Ensembl ids.
Ensembl Genomes is the most frequently used database for genomic features. It uses and combines data from different databases, like the GENECODE [114] database for human and mouse genomes [115].
Another on alternative splicing database is ProSplicer. It contains protein, mRNA and Expressed Sequence Tag (EST) information. It takes its data from various databases like Ensembl, TrEMBL and UniGene and integrates the predicted alternative splicing data from these (see Figure 8) [116].

8 Project

The goal of the project *Alternative Splicing Orthology* was to create a splicing orthology and therefore analyze the isoforms of orthologous genes by visualizing a multiple sequence

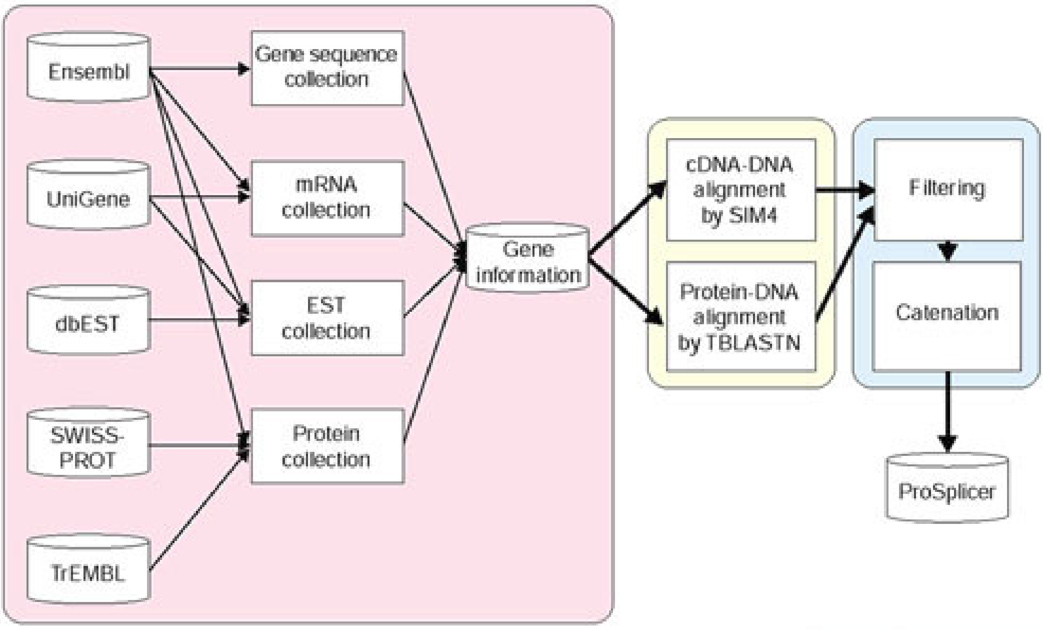


Figure 67 Figure 8: Predicting approach of alternative splicing. (From 'A Putative Alternative Splicing Database Based on Proteins, mRNAs and EST Clusters' by Jorng-Tzong Horng, Hsien-Da Huang, Chau-Chin Lee, and Baw-Jhiune Liu) [116]

alignment of these homologous isoforms. The idea behind this approach is to easily detect alternative splice events in multiple species. The information, gained by the splicing orthology, and other additional evidences can be used for an isoform reliability prediction.

8.1 Dataset

8.1.1 Homology database comparison

The Ensembl mentioned database also provides homology information for genes and additionally gives a reference on support levels of the annotation data with the the Transcript Support Level (TSL) flag [115]. It contains homologies for 87 genomes, for both coding and non coding genes.

The HomoloGene database of the NCBI also provides information on gene homology. Amino acid sequences can be given as input, which then are blasted against the database. The database gives a tree with all organisms which hold a homologue gene to the input, sorted by their homology degree [117]. It contains homologies only for 21 organisms, because these have to have a complete genome or >10,000 UniGene entries. Also the cluster contain only protein coding genes.

The Database of Complete Genome Homologous Genes Families, HOGENOM, contains information on homologous genes in fully sequences organisms from all three kingdoms. It also provides homology trees for the respective genes [118]. It contains about 7 million genes from 1470 organism of all kingdoms. They combine data from Ensembl, NCBI, and other additional genomes.

To decide upon a dataset of gene homology clusters, we took the available datasets from HomoloGene, HOGENOM and Ensembl. We converted the GeneIDs of the three datasets each to Entrez IDs to compare the content of the clusters. Hereby some identifiers couldn't be converted, so these cases were excluded from further analysis. Figure 9 shows how the three databases are related to each other. The overlap between Ensembl and HomoloGene is better than HOGENOME and the others. Therefore we excluded HOGENOME from further steps. Because HomoloGene conferred to only protein coding genes, we also discarded this database from our final dataset. Our final homology dataset was taken from Ensembl. Therefor we worked with all data provided by Ensembl, like homology clusters, gene annotations and sequences.

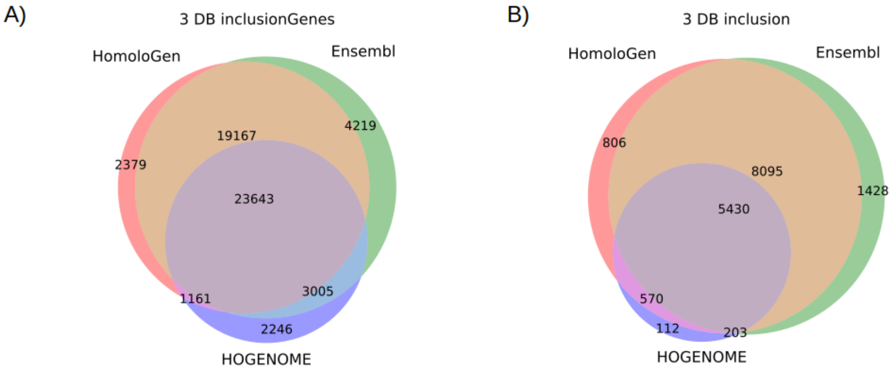


Figure 68 Figure 9:Only Clusters containing all 3 organisms were considered! A) shows the number of EntrezIDs, which can be found in the three databases, which also exist in the other two, one other or only in the single dataset. B) Here are the numbers of clusters displayed. The overlaps show which number of clusters containing all three organisms appear in the others as well.

8.1.2 Obtaining the data and clustering homologous genes

After deciding to choose the Enembl database, the first step, to be able to analyze the splicing structure of homologous gene clusters, was to create clusters of homologous genes. This was achieved by clustering the x:y:z homology relation (human, mouse, rat) annotation of genes in Ensembl. For this and the following data obtaining steps in Ensembl, we used the homo sapiens data set GRCh38.p10, the mus musculus data set GRCm38.p5 and the rattus norvegicus data set Rnor_5.0. After creating these homology clusters, we obtained all unspliced genes and cDNA (transcript) sequences of these homolog genes. As in figure 11 is shown, we received around one third of all human genes, slightly less than the half of mouse genes and more than two third of all rat genes from Ensembl. Also the majority of these genes were protein coding which could be explained through the homology prediction of Ensembl. The Ensembl homology prediction works mainly on phylogenetic protein comparisons whereby non protein coding genes weren't analyzed. Anyway after recent updates Ensembl offers through their new non coding gene homology prediction the homology relation of these genes. Through this circumstance it is understandable that not all non coding genes of these three organism were fully analyzed or just don't have any ortholog relation to the other two species.

Furthermore we identified the genomic structure annotation of the genes among it the genomic gene start and end position, the transcript start and end positions, the exon start and end positions, and the associated identifier which links associated exons, transcripts and genes to each other. Through these steps we received in total 21151 homology cluster in which mostly all three organism were present (17431 homology cluster). A more detailed

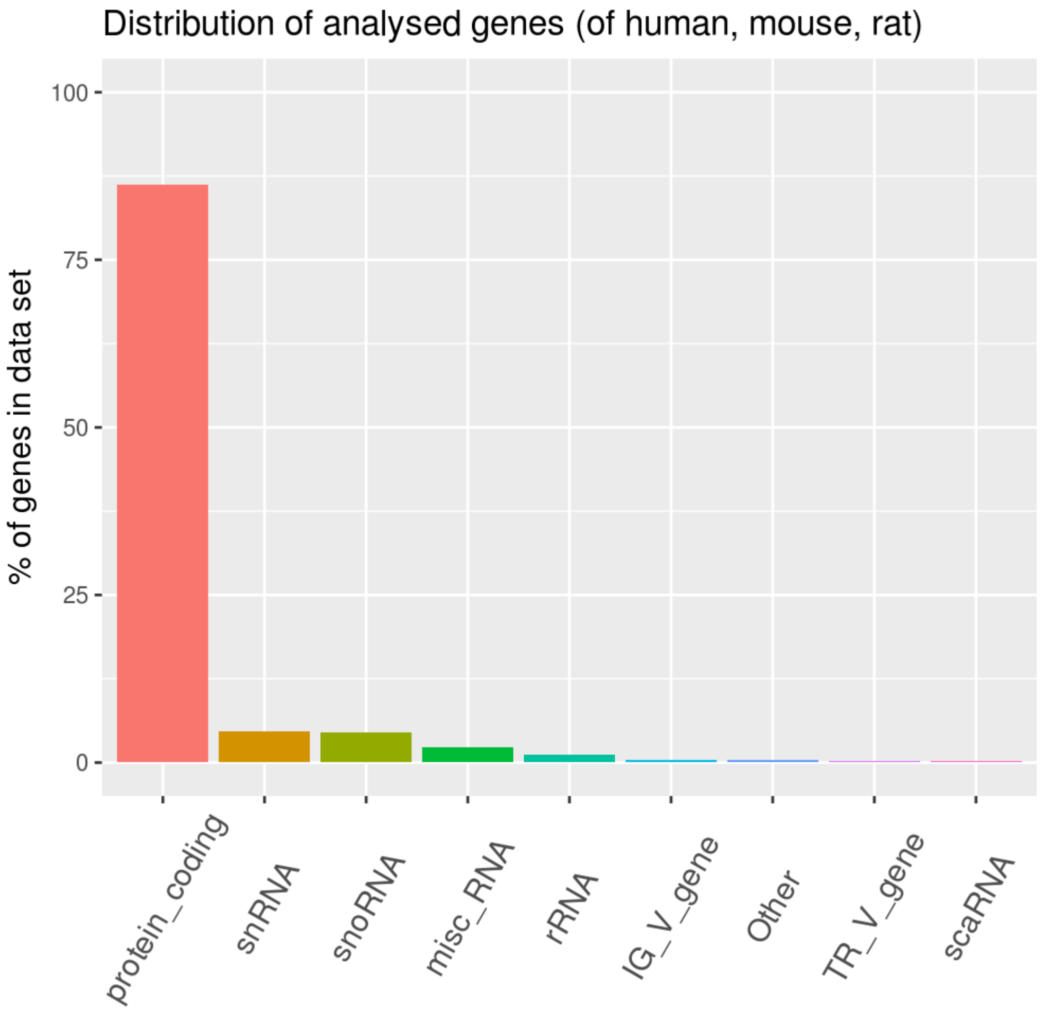


Figure 69 Figure 10: Overview of the annotated biotype (gene type) of the obtained complete gene data set which consists of human, mouse and rat genes with at least one homolog in each other. More than 80% of all genes were protein coding whereas less than 20% of genes in the data set are non coding genes.

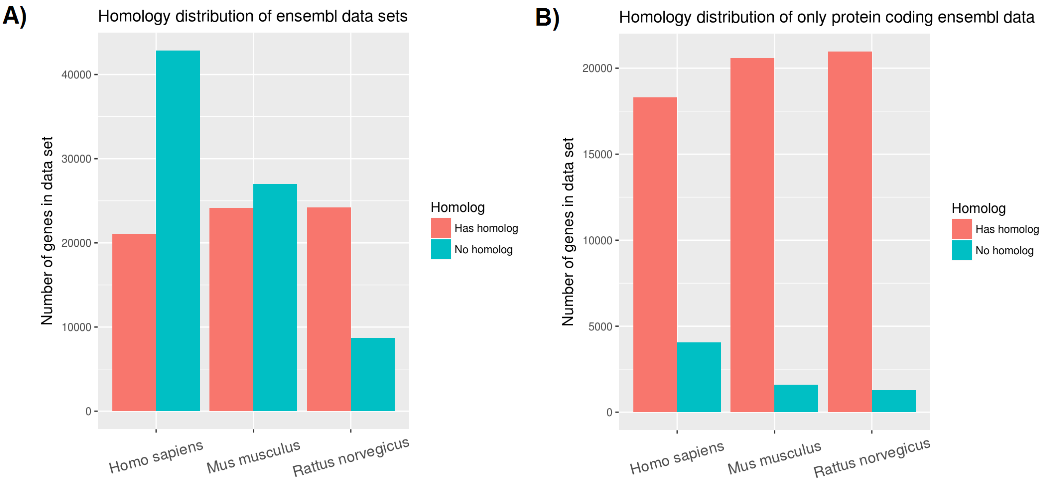


Figure 70 Figure 11: Overview of the distribution of genes with at least one ortholog in the two other species from the database Ensembl. The illustration displays the number of genes which have at least an ortholog (homolog) in contrast to the number of genes which haven't an ortholog (homolog) for the complete gene set (non coding genes + protein coding genes) in Ensembl (A) and only all protein coding genes of Ensembl (B). In plot (A) there is a wide variety of trends depending on the organism. In human there are more than double as much as genes with no homolog in one of the other two organism, in mouse the number of genes without a homolog are only slightly more than the mouse genes with a homolog and the number of rat genes with a homolog in human and mouse are more than double as high as the number of genes without a homolog. On the other hand the distribution of the protein coding genes in Ensembl with a homolog are quite similar to each other, except that there are slightly more human protein coding genes with no homolog.

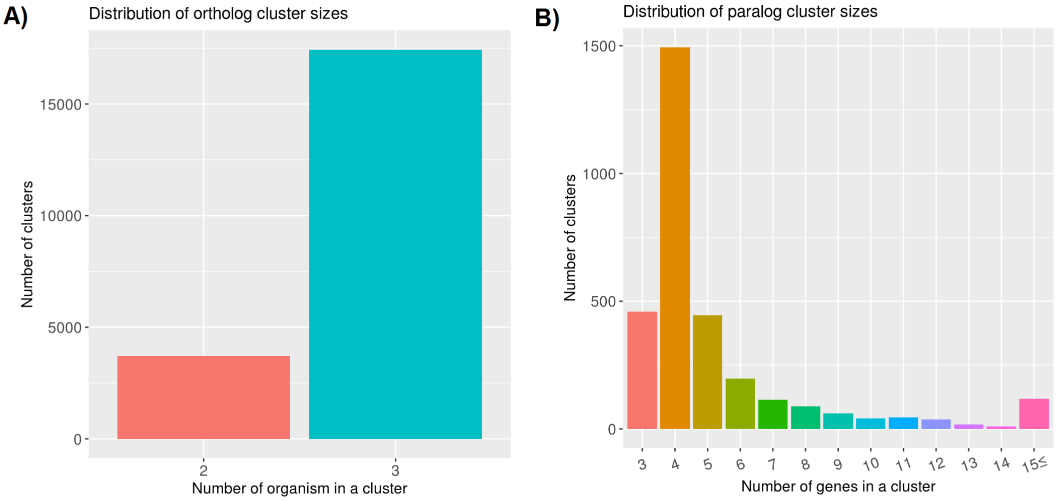


Figure 71 Figure 12: Distribution of the number of organism (A) and the number of genes (B) in a homology cluster. The number of organism in a cluster represents the number of ortholog relations in a cluster and indicates that the majority of homology clusters have all 3 selected species in it (human, mouse, rat). The distribution of genes (B) includes only clusters which contain more than one gene of the same species and therefore was defined as a paralog cluster. The comparison of the number of genes in these two plots, reveal that the number of ortholog clusters is considerably higher than the number of paralog clusters. Additionally the distribution of paralog clusters validate that most paralog cluster have 4 genes in it and that the frequency of clusters will get lower if more genes are in it.

distribution of the cluster content is shown in figure 11.

8.1.3 Isoform structured alignment calculation

In the next step the isoform structured multiple sequence alignments of the transcripts (isoforms) in the same homology cluster were with created the in-house software ISAR (Isoform Structure Alignment Representation). The reason, for not using a common multiple sequence alignment software like clustalW2 (clustal Omega) or MUSCLE, is that conventional multiple sequence alignment tools don't take the gene structure of aligned transcripts into account. This can result in alignments of different exons, because the alignment software can not distinguish between similar sequence and exons if the two aligned exons of two transcripts belongs to the same gene as illustrated in figure 13. These logical errors in multiple sequence alignments of isoforms have to be take into account, reason why ISAR was used.
To maintain the gene structure, ISAR requires the gene sequence and the transcript sequences with the annotated exon positions within the corresponding gene. The software identifies through pairwise alignments of the genes, the homologous exons between them. With this mapping of homolog exons it is possible to create a correct and consistent multiple sequence alignment of the corresponding transcripts with the correct gene structure using partial ordered graphs. ISAR was run for every homology cluster with the associated homolog genes as input. As output we received the pairwise alignments of homologous exon regions between the genes. This mapping was used to extract the additional features of ISAR. The first feature was the calculation of the isoform structured alignment. These transcript multiple sequence alignments of a homology cluster were the resulting core data of the project which we wanted to visualize and analyze. However these isoform structured

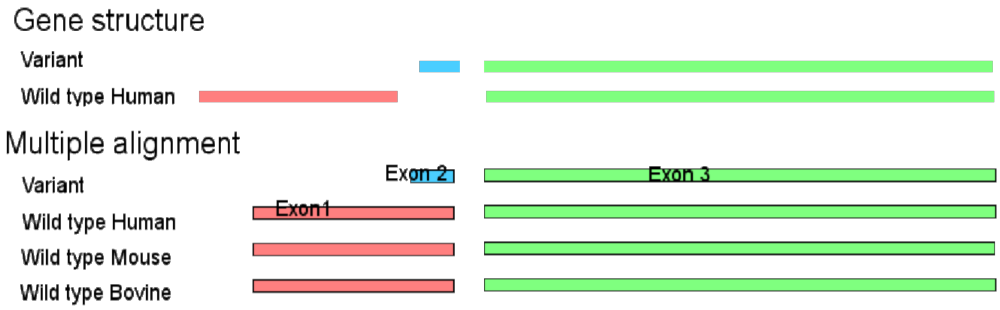


Figure 72 Figure 13: Illustration of the difference between the gene structure and a multiple sequence alignment from a conventional multiple sequence alignment software. The colored bars represent the exons of a gene. In contrast to the gene structure, exon 2 (blue middle bar) will be aligned with exon 1 (first red bar) instead of aligning with noone.

alignments remain dynamic. In the majority of homology clusters the sorting of specific exons determines the visual output. This can be explained by the following example: If there are two genes and there is one non-homologous and unaligned exon in each gene between two aligned and homolog exons, these two unaligned non-homologous exons can be sorted in any order as long as these two exons are between the aligned and homolog exons. Thus the gene structure after an unaligned exon sorting is correct, but can have a huge impact on the later visualization of the isoform structured alignment especially by a large number of unaligned exons in a homology cluster. Therefore it would be practical if the visualizer would be able to create the isoform structured alignment according to the provided ISAR mapping of aligned exon regions for a suitable visualization.
The second feature of ISAR is the output of the isoform structured alignment as an non-interactive image, allowing an easy validation for the correctness of the later interactive visualization. Also the interactive reproduction of these images was the first goal we wanted to archive with the visualizer.

8.1.4 Enhanced MSA analysis

8.1.4.1 Transcript support level

The Transcript Support Level (TSL) is a rating of transcripts, both provided by Ensembl. It is based on mRNA and Expressed Sequence Tag (EST) evidence for GENCODE transcripts. The TSL displays a rating for a trancript, which can be seen as the reliability that the transcript really exists (TSL1) or has yet to be confirmed (TSLNA). The following table of categories was taken from the Ensembl webpage: `http://www.ensembl.org/Help/Glossary?id=492`

- tsl1 – all splice junctions of the transcript are supported by at least one non-suspect mRNA
- tsl2 – the best supporting mRNA is flagged as suspect or the support is from multiple ESTs
- tsl3 – the only support is from a single EST
- tsl4 – the best supporting EST is flagged as suspect
- tsl5 – no single transcript supports the model structure
- **tslNA – the transcript was not analysed for one of the following reasons:**
    - pseudogene annotation, including transcribed pseudogenes

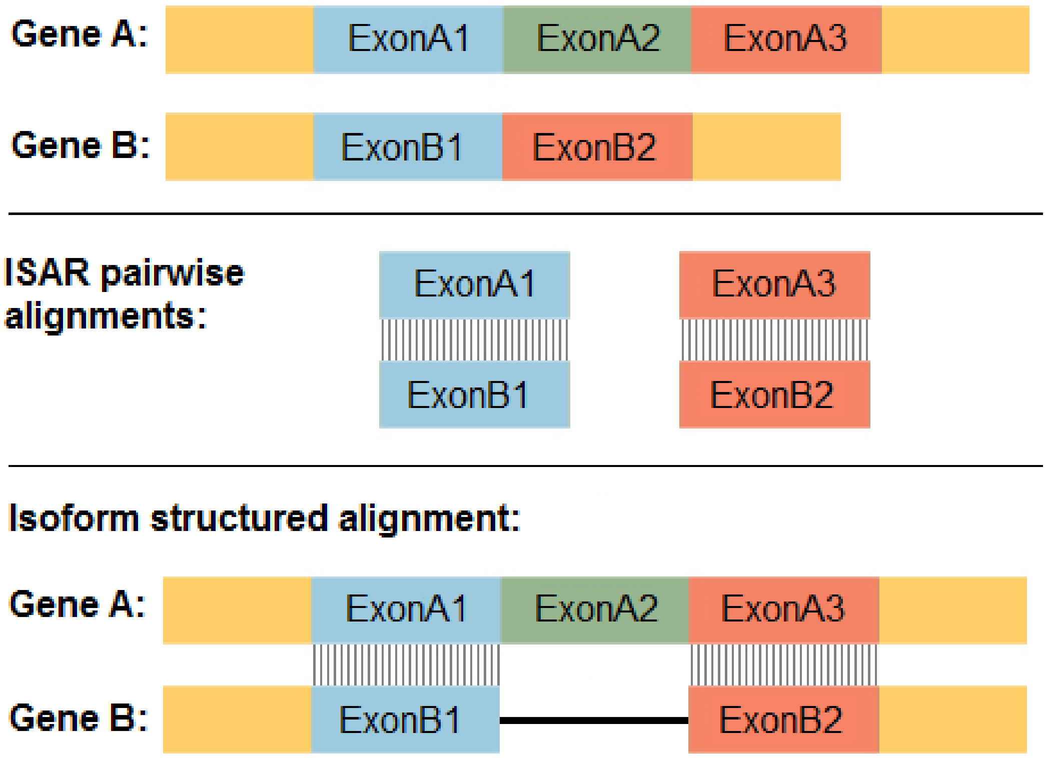


Figure 73 Figure 14: Illustration of the processing in ISAR. This example shows the correct identification of the two homolog blue and red exons in the two genes A and B. In the later created isoform structured alignment, the green exon A2 is correctly gapped whereas the two homolog exons blue and red are aligned.

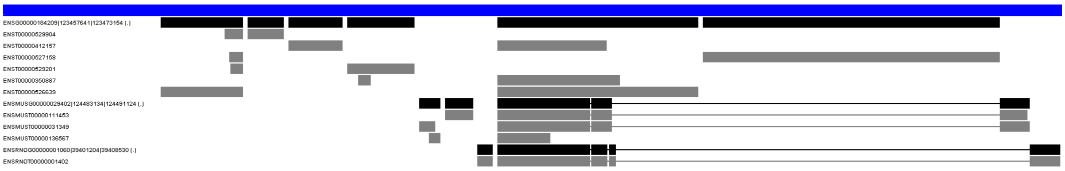


Figure 74 Figure 15: ISAR image output of the homology cluster 16386 which contains the genes: ENSG00000184209, ENSMUSG00000029402 and ENSRNOG00000001060.

- human leukocyte antigen (HLA) transcript
- immunoglobin gene transcript
- T-cell receptor transcript
- single-exon transcript (will be included in a future version)

We obtained the transcript support levels (TSL) for each gene from the two organism homo sapiens and mus musculus. The transcript support levels are a measurement for the existence reliability of a transcript. We implemented this value of reliability just as an external additional feature for the possible manual analysis by combining and summing up multiple evidences in the visualizer. We also used the TSL values later in the *Analyzer* for weighting our automatic calculated feature scores depending on which transcript these feature scores come from.

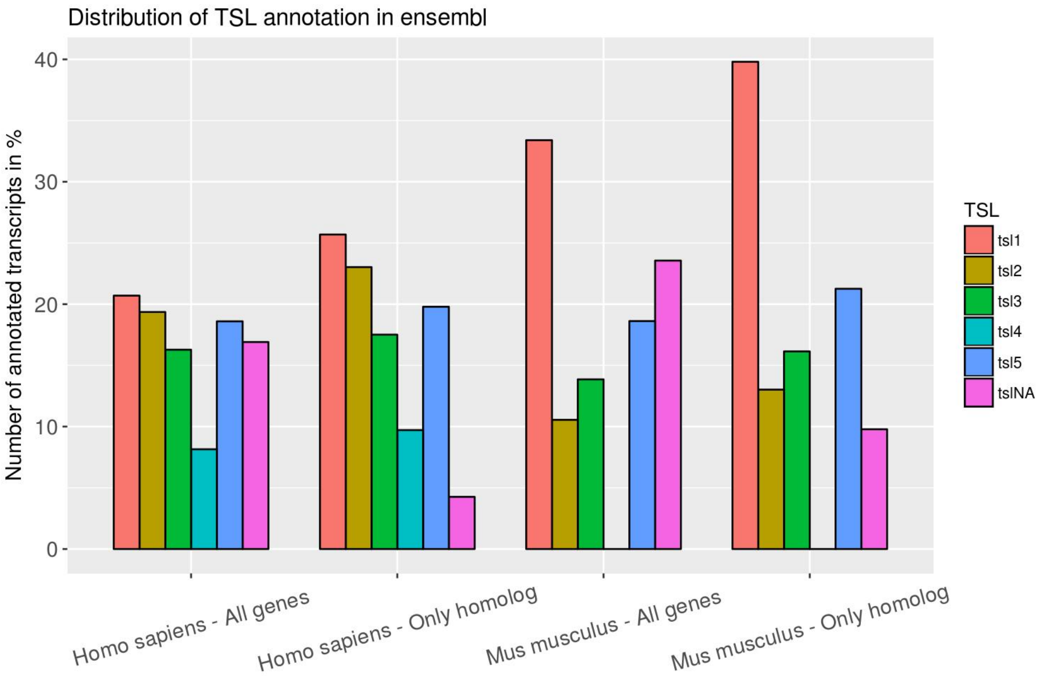


Figure 75 Figure 16: Distribution of the transcript support level of all transcripts in all genes and once only in the homolog gene subset (only genes with at least one ortholog in one or both of the two other organism) from the two organism human and rat. It is indicated that the homolog gene subset in both organism (which are mostly protein coding genes) have more tsl1 and tsl2 annotations than in comparison to all genes. Most of the genes of the homolog subset have a TSL value annotated and therefore not the value tslNA (no TSL value annotated yet) annotated. In addition it is unexpected that there are no tsl4 values annotated to mouse genes.

8.1.4.2 MiRNA

MicroRNA are short, about 22 nucleotides long, non-coding RNA molecules, which are found in most organisms [119]. They take part in RNA and posttranscriptional gene expression silencing. The transcription of miRNA is usually done by RNA polymerase II and the resulting pre-miRNA, which builds a hairpin loop and then is posttranscriptionally processed. The mature miRNA molecules are spliced out of its pre-miRNA. Some of the miRNA genes

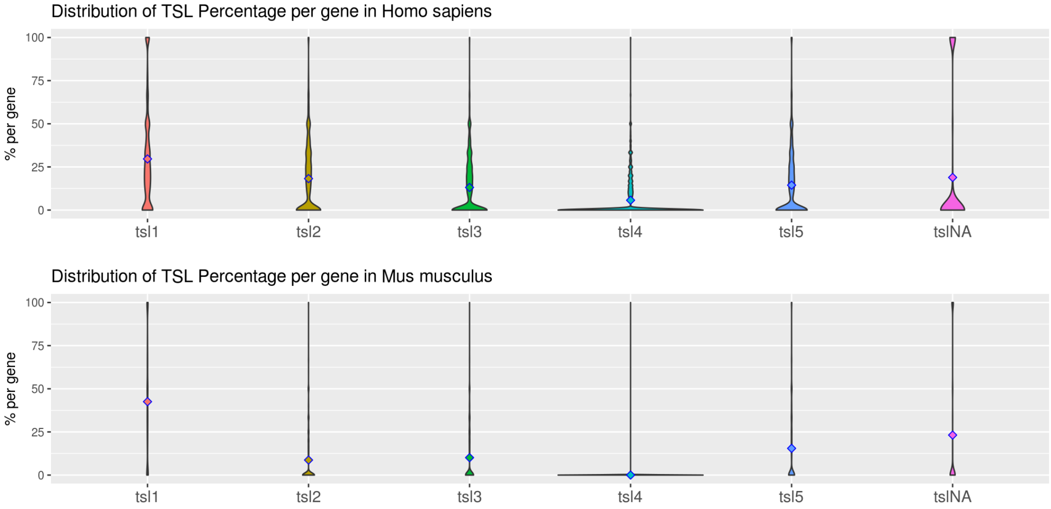


Figure 76 Figure 17: Distribution of the TSL percentage per gene for all genes in homo sapiens and mus musculus. The violin plots illustrate that the different TSL levels are nearly equally present in all genes, except of TSL4. The human genes have, in comparison to mouse genes, more likely an unequal TSL distribution.

are co-transcribed with a protein-coding gene. These kind of miRNA genes are located in the introns of their respective genes, which results to their designation as *mirtrons*. They are post-transcriptionally spliced out together with the pre-mRNAs introns. As miRNA are highly conserved sequences, we thought to use their existence in our cluster-genes as an evidence for transcripts, respectively their spliced out introns.
We used the dataset provided by miRBase with miRNA primary transcript annotations on the mouse, human and rat genome. With these we calculated the counts of miRNAs (primary transcripts without sequence variation, like *hsa-mir-4528*). In these datasets about 68% of the miRNAs in human, 75% in mouse and 38% in rat overlap genes in our clusters, as shown in the table of Figure 19. We also discovered that some genes have more than one kind of miRNA gene located inside, which can be observed in Figure 12. Our analysis showed that from our 21151 clusters 1590, about 7.5%, contain one or more miRNA genes. The distribution of miRNA of organisms per cluster is shown in Figure 21. Please note that no further identification of the miRNA genes as mirtrons was done.

8.1.4.3 Conservation

Because donor and acceptor sites of introns play an important role in the splicing mechanism, their sequence conservation is an evidence for splicing events. As described in the previous chapter certain sequences prevail as donor-acceptor patterns. Therefor we calculated the splice site conservation with sequence length of 2 for all transcripts in an cluster, respectively of the whole dataset to give an evidence of accuracy for the existence of a transcripts. For our conservation score we weighted the given transcripts with their TSL level, to give transcripts with a higher probability to be true transcripts a better rating. In our dataset the most common splice site sequence was 'GT,AG' with 'GT' as donor and 'AG' as acceptor. This splice site occurs in about 99% our splice sites of all clusters (see figure 22). Hereby is the difference between unweighted and weighted splice site counts minimal as both contain the 99% conservation of 'GT,AG'. In the unweighted counts some splice sites are filtered out,

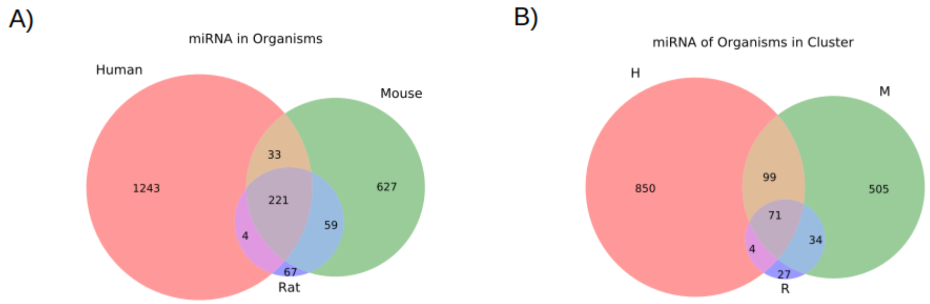
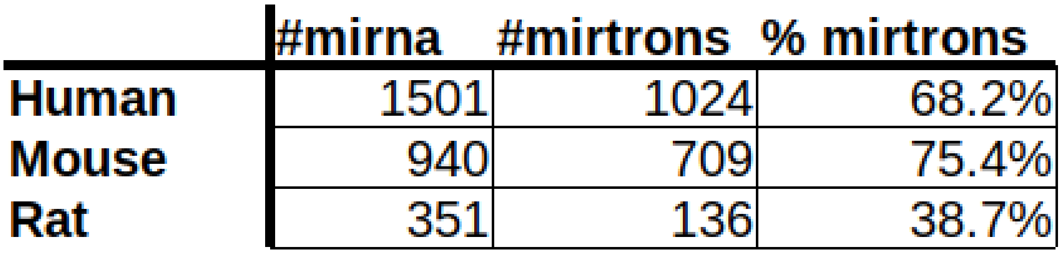


Figure 77 Figure 18: This Figure shows the numbers of miRNAs, which exist in the 3 organism. A) Here are the counts of all the miRNAs, without different versions of the miRNA, which occure in our data set. B) shows the number of miRNAs contained in genes from our Clusters (mirtrons our clusters). These are also the miRNAs which we used in our project. The plot is comparable with Figure 11.B).



	#mirna	#mirtrons	% mirtrons
Human	1501	1024	68.2%
Mouse	940	709	75.4%
Rat	351	136	38.7%

Figure 78 Figure 19: The table shows the number of miRNAs and mirtron miRNAs in the three organisms.

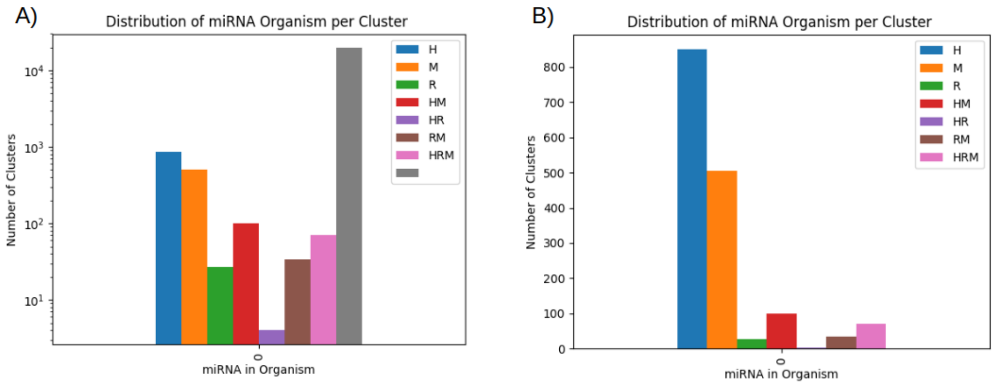


Figure 79 Figure 20: Number of clusters with miRNA originating from specific organism combinations (mouse 'M', human 'H', rat 'R'). A) Shows the distribution of miRNA containing clusters for all organism combinations of human, mouse and rat, additionally with the number of clusters without miRNAs. The empty clusters are labeld with "' in gray. Because of the low number of miRNA containing clusters, the y-Axis shows log values. B) shows the number of miRNA containing clusters for all Organisms without the miRNAless clusters.

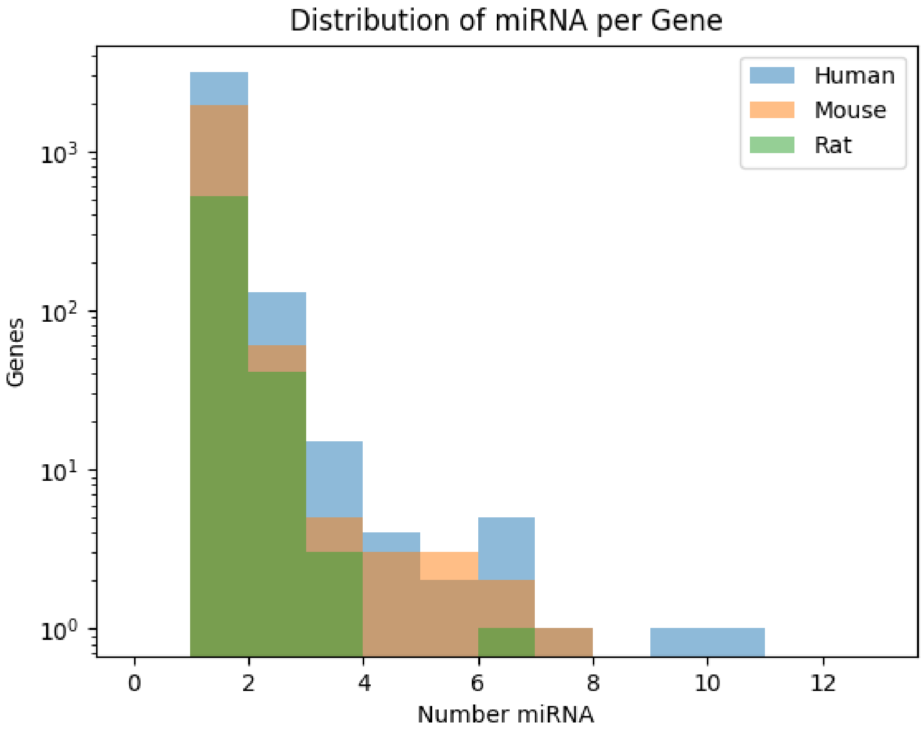


Figure 80 Figure 21: Number of miRNAs which are located in a gene. The plot shows the distribution of log number of genes to the number of miRNA they contain.

because of their unknown transcript evidence, but all in all the distribution of the high counts stays the same, as shown in figure 22.B). Also the relatively high conservation core of the intron start site 'GC' compared to the other possible sequences except 'GT', might be worth further research (figure 23). This phenomena might result from a nucleotide conversion from Thymine to Cytosine.

8.1.4.4 Expression Level

RNAseq is the main method to messure gene expression in cells. It is used to annotate genome regions which are transcribed and validate the level of their abundances. To get this, RNAs are extracted from cell samples, their sequence is read out and mapped to the according genome sequence of the samples organism. Therefor the resulting expression patterns of gene regions could be strong evidence for the existence of transcripts. Since the gene expression differs in different cell types, tissues, of Organisms, to get an general expression pattern for all genes, we took the expression from different tissues and calculated the mean expression for the genes. There should be a difference in level for introns and exons, because the data used only contained mature mRNA, so that all introns should have been spliced and not existent in our analysis. For our dataset we took RNAseq alignments from two different experiments: Mouse and rat from Merkin et al. [100] and human from Brawand et al. [120]. Of these, we took the tissue runs which appear in both studies: Brain, heart, kidney, liver and testis. For each tissue we have more than one replicate. For our expression analysis, only the uniquely mapped reads were used, by filtering for the samflag NH:i:1 (only once aligned) and the MAPQ value of over 10 (score of rightness of alignment), see figure 24, using the SAMtools package [121]. Using the Ensembl gene annotation, coverage was calculated for all 5 nucleotide (nt) bins in the cluster genes by using BEDtools [122] with the read alignments (bams) of the genes organism. The result are files for each gene with the counts of each sample run per 5 nt.

9 Analyzer

Using the isoform structured alignments from ISAR and the additional evidences mentioned above, the next step was to design a reliability score prediction which considers all evidences and calculate one single reliability score for a new isoform. Therefore our aim was to define rational scores which are able to differ between likely, plausible isoforms and unlikely isoforms.

Consequently we characterized three main scores. The mean of all splice junction reliability scores, the exon combination score and the transcript start and stop score. The core score is the mean of all splice junction reliability scores. In order to calculate this score, every splice junction (intronic gap between two exons) of a studied transcript is individually analyzed. The splice junction reliability score is put together by the isoform structured alignment score and the splice site pattern conservation score. The isoform structured alignment score evaluates whether the donor splice site and acceptor splice site exists in other isoforms in the observed homology cluster. Additionally the number of isoforms with such splice sites, the TSL value of these isoforms and whether the two splice sites occur together in an isoform or separated are all information which is used for weighting and result in the ISA scores with the interval from 0 (both splice sites don't exist in this homology cluster) to 1 (both splice sites exists at least more than three times in the homology cluster, within a tsl1 isoform and occur together as a splice junction in the homology cluster). The splice site pattern conservation scores consist of the splice site base conservation. All splice sites base patterns of sequence

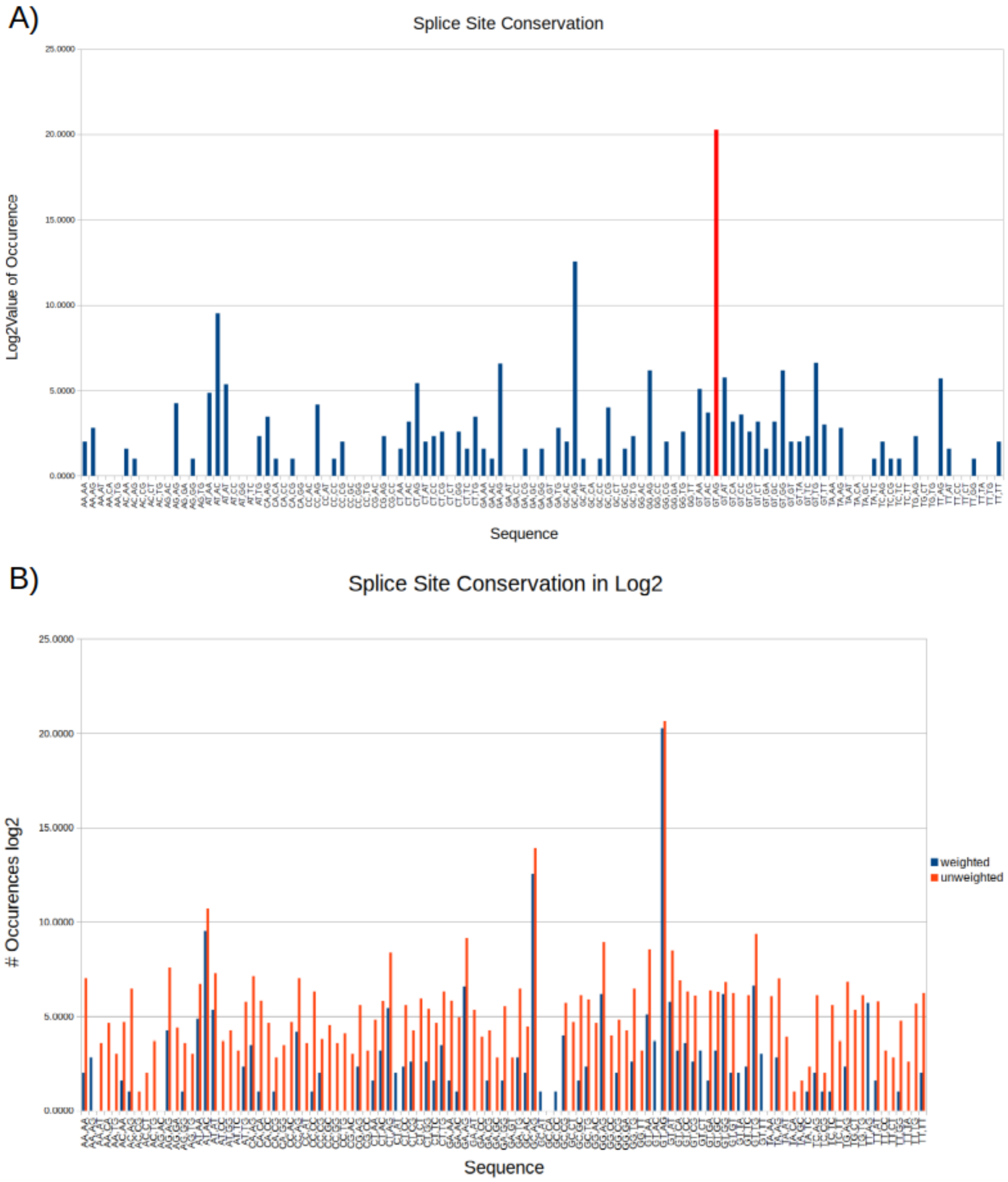


Figure 81 Figure 22: The plot shows the log2 numbers of splice site sequences in our dataset with A) the TSL weighted sequence counts and B) both unweighted and weighted splice site counts.

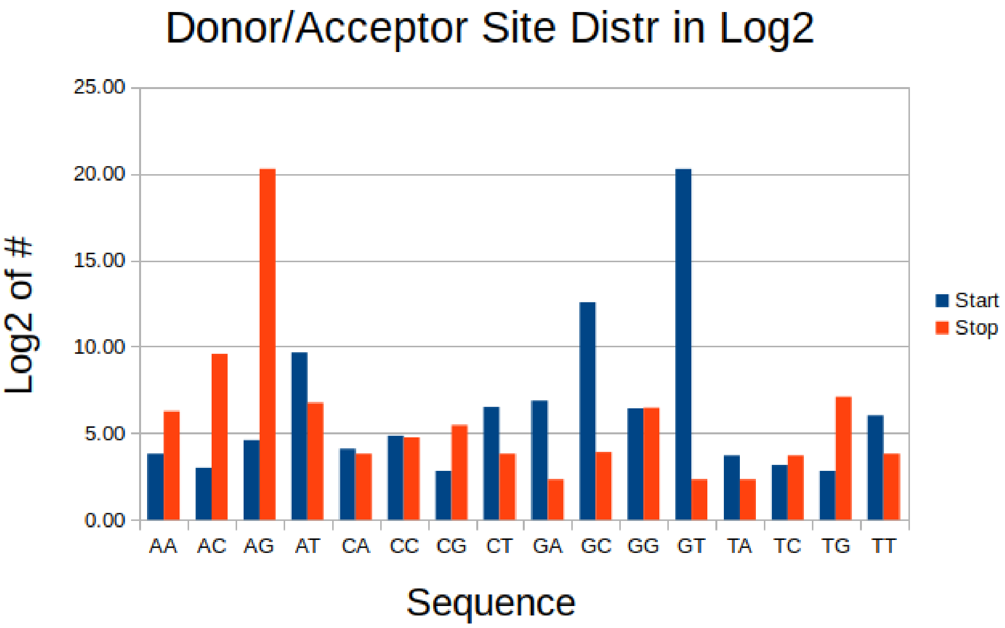
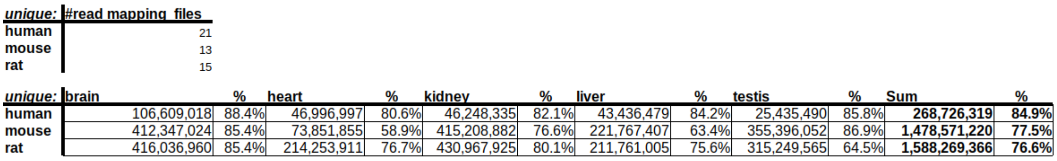


Figure 82 Figure 23: This is the log2 distribution of splice site sequence counts in our dataset (weighted).



unique:	#read mapping files
human	21
mouse	13
rat	15

unique:	brain	%	heart	%	kidney	%	liver	%	testis	%	Sum	%
human	106,609,018	88.4%	46,996,997	80.6%	46,248,335	82.1%	43,436,479	84.2%	25,435,490	85.8%	**268,726,319**	**84.9%**
mouse	412,347,024	85.4%	73,851,855	58.9%	415,208,882	76.6%	221,767,407	63.4%	355,396,052	86.9%	**1,478,571,220**	**77.5%**
rat	416,036,960	85.4%	214,253,911	76.7%	430,967,925	80.1%	211,761,005	75.6%	315,249,565	64.5%	**1,588,269,366**	**76.6%**

Figure 83 Figure 24: The tables show read mapping file numbers. The first is the number of read mapping files we use for our organism. The second shows the number of uniquely mapped reads for the 5 tissues (sum of all replicates) and the % compared to all mapped reads.

length two are observed whereas the percentage of occurrences of splice site base patterns in the homology cluster can be calculated. Therefore the splice site sequence patterns of an analyzed transcript can be compared with the splite site patterns of the homology cluster which results in a percentage score of how frequently the splice site pattern of the analyzed splice junction occurs in the homology cluster. After calculating the ISA score and the splice site pattern conservation score of a splice junction, an offset of these two scores to a single splice site reliability score is required. Due to the high scoring splice site conservation in nearly all splice junctions (˜99%), a 1:1 weighting and union of these two scores wouldn't be sensitive in order to differ between likely and unlikely splice junctions. Therefore we decided to implement a support vector machine (SVM) for a reasonable score weighting. We defined the SVM for predicting a two-class classification with a radial basis function kernel and default training parameter. To train the SVM we created a data set of splice junctions from all tsl1 transcripts as positives and all tsl5 transcripts as negatives and calculated the two scores for each splice junction. Following the SVM was able to analyze a splice junction and predict if it is likely or unlikely existing which results in the splice junction reliability score. The exon combination score for an analyzed transcript was calculated by viewing every exon of the transcript and seek through the homology cluster if the same exon (with same length and start/end positions) already exists in another isoform. This search results into a boolean outcome for every exon (yes, the exon exists in the homology cluster or no, the exon doesn't exist in the homology cluster) whereby the final exon combination score is defined as the number of already existing exons divided by the number of all exons of the analyzed transcript. The last score, the transcript start and stop score, highlights the existence of the transcription start and end position of an isoform. Similar to the ISA score, the transcript start and stop score evaluates if the transcription start and the transcript end position of the analyzed isoform already exists in the homology cluster and weights the existence based on the TSL value of the transcript. After calculating these three scores for the studied transcript, the scores are summed together and are divided by 3.

We tried to validate our defined scores by comparing existing transcripts of all homology clusters, sorted after its TSL value, with random created transcripts. We created random transcripts by choosing an existing transcript in a homology cluster as the template of our random transcript and introducing different alternative splicing patterns like exon skipping, intron retention, alternative splice sites or alternative transcript start or/and stop positions. Additionally we created completely random transcripts with random start and stop of the transcript and the exons in it.

As in figure 25 illustrated, we started to evaluate our isoform structured alignment score for the splice junction reliability prediction and therefore only analyzed the splice junctions and their scores at first. The majority of tsl1 and tsl2 splice junctions are mostly high scoring due to the high TSL weighting of these splice junctions. Nevertheless there are low scoring tsl1 and tsl2 splice junctions,too. These can be explained by small homology clusters with only a few transcripts, through which the exclusion of the current viewed transcript is also the complete exclusion of some splice junctions in the homology cluster which results in a low ISA score. The other existing transcripts in the homology cluster (tsl3-tslNA) have significant lower ISA scores due to the fact that these transcripts include splice junctions which don't exist in any other isoforms in the homology cluster. The better scoring splice junctions of the low TSL valued transcripts can be explained through the existence of these splice junctions in other isoforms (with a high TSL value). In a detailed reflection of the random created transcripts, we observed as expected high ISA scores in the altered exon skipping, intron retention and alternative transcript start and/or stop transcripts because

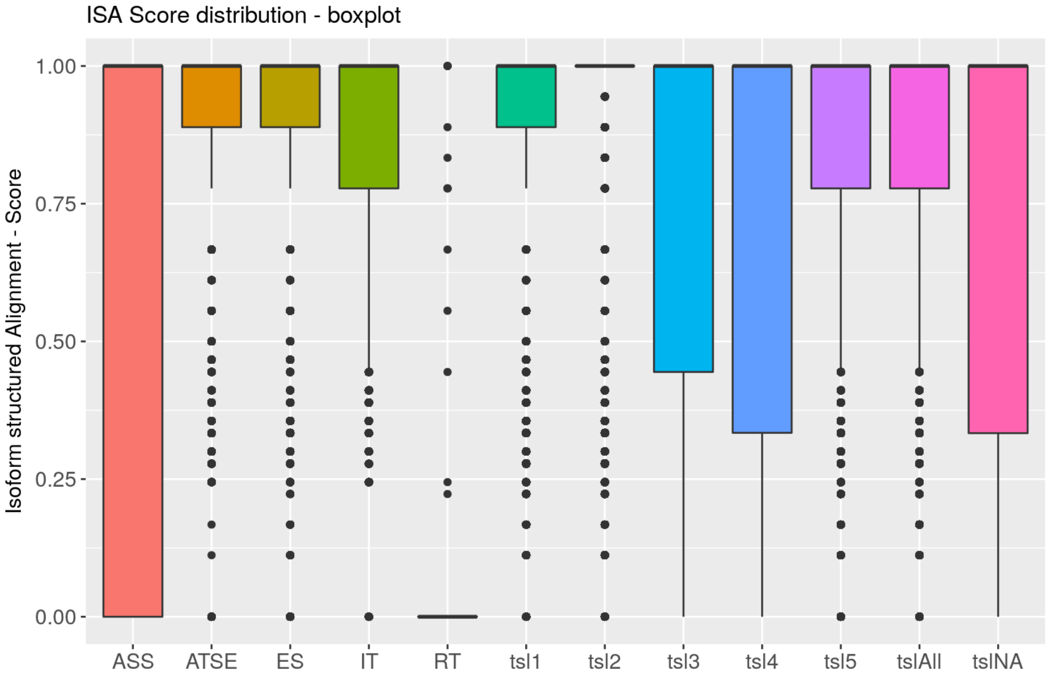


Figure 84 Figure 25: Distribution of the isoform structured alignment splice junction score for all splice junctions in every homology cluster with its associated TSL level (tsl1-tslNA, tslAll defined by union of all existing splice junctions in a homology cluster). Additional to splice junctions of existing transcripts, we created random transcripts for better score validation. The random transcripts (except the complete random transcript) were calculated by randomly choosing a existing template transcript and introducing alternative splicing patterns or different transcript start/stop positions. Random transcripts: Alternative splice site (ASS) – changing splice site positions of existing exons; alternative transcript start/stop (ATSE) – altering the start or/and stop position of the transcript; exon skipping (ES) – skip one or multiple exons of the template transcript; intron retention (IT) – include an intron of the template template transcript; complete random transcript (RT) – completely random transcript creation without a template transcript.

there are no changes of the splice sites themselves and therefore every analyzed splice site is maintained in the homology cluster. This results in the conclusion that the splice junction ISA score should be equal for the most cases (splice junction exist in template transcript) or lower (exon skipping results into a longer splice junction but with existing splice sites) than the template transcript splice junction score. As expected we observed that the alternative splice sites and completely random transcripts yielding the most low ISA scores. This can be explained due to the new created splice sites which don't exist in the homology cluster. Through this validation we confirmed that the ISA score is consistent to what we wanted to measure and express with this score. Finally the global isoform reliability score were evaluated with the same procedure as used for the splice junctions (random created transcripts by template).

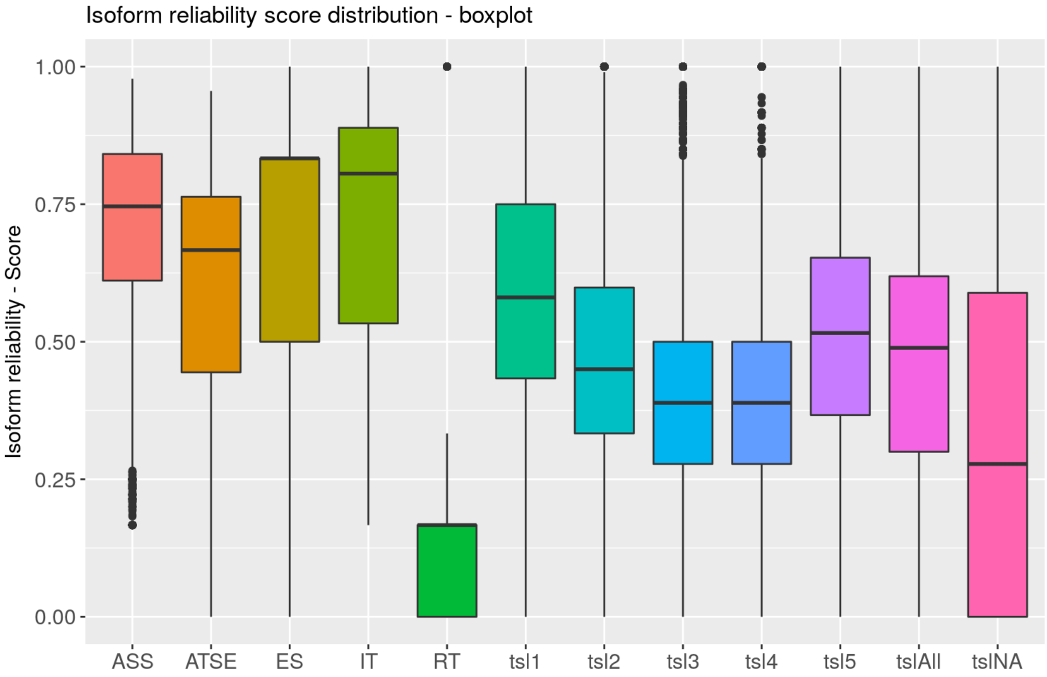


Figure 85 Figure 26: Distribution of the global isoform reliability score for all transcripts in every homology cluster with its associated TSL level (tsl1-tslNA, tslAll defined by union of all transcripts). Random transcripts: Alternative splice site (ASS) – changing splice site positions of existing exons; alternative transcript start/stop (ATSE) – altering the start or/and stop position of the transcript; exon skipping (ES) – skip one or multiple exons of the template transcript; intron retention (IT) – include an intron of the template template transcript; complete random transcript (RT) – completely random transcript creation without a template transcript.

The distribution of isoform reliability scores may be odd at the first view, because of the high scoring by random created transcripts in contrary to the existing TSL transcripts. However this can be easily explained. The scores for the existing transcripts are calculated by assuming it's a new transcript. Therefore the transcript is removed from the homology cluster and then scored with the information of the remaining isoforms. Especially in small clusters with only a few transcripts, it is likely that the splice junctions or exons of the analyzed transcript are unique and result into a lower scoring. On the other hand the scores of the random created transcripts by a template (ASS, ATSE, ES, IT) are calculated with all transcripts and therefore with more information which increases the probability that an exon or splice junction already exists in the homology cluster. Another important reason for the high scoring random transcripts are that the transcript alteration only change a specific property of an already existing isoform which results into altering one specific transcript score. For example the exon skipping random isoform only excludes some exons of an existing transcript and therefore only changes the splice junction reliability scores but has a high exon combination score and a high transcript start/stop score due to the obvious existence of the exon and transcript start/stop position (template transcript). The completely random created transcript (RT) had, as expected, a low global isoform reliability score due to the high probability that the random created transcript don't share any isoform properties with the other isoforms in the homology cluster. Therefore we are able to confirm our global isoform reliability score prediction, too.

10 Visualizer

Our final goal was to interactively visualize the before mentioned information and illustrate the isoform structured alignment from ISAR with the raw evidences as well as the calculated reliability scores. The user of the interactive visualizer should be able to browse our database of all homology clusters which we obtained from Ensembl. Therefore some kind of search or selection to browse a specific homology cluster of interest was required. We set ourselves the goal that the user could upload a new isoform to an existing homology cluster which should result into a new isoform structured alignment and an isoform reliability prediction for the new transcript based on the information of the selected homology cluster.

Our visualizer is capable of doing all the features we intended to implement and many more. The visualizer is based on the Zk framework and the in-house plug-in Zk − plotly. Additionally the visualizer calculates the isoform structured alignment by its own based on the transcript mapping from ISAR. The idea behind the self-calculation of the ISA is, that the user could change the transcript order and therefore change the overall appearance of the isoform structured alignment as it is mentioned before. Next to the ISA self-calcuation, the feature to drag transcripts in the transcript list overview tree and therefore change the transcript order is already implemented but deactivated in the current visualizer version due to the missing connection of these two features. The working procedure of the visualizer and its features are presented in the following screenshots.

10.1 TSL

As already been mentioned before, the transcript support level of isoforms is mainly used for weighting reliability scores. Nevertheless it is also possible to obtain the TSL level of an existing isoform by hovering over one of the exons of the isoform as it can be seen in figure 30.

10.2 MiRNA

We show miRNAs overlapping the union transcripts of our clusters, by marking their location with a pink diamonds slightly above the structure plot. Each displayed miRNA owns a hovertext, shown when the cursor is on the label, with information on its designation, length, miRNA-ID, Type of annotation, overlapping Ensembl geneID and a description of its location (intronic, exonic, overlap).

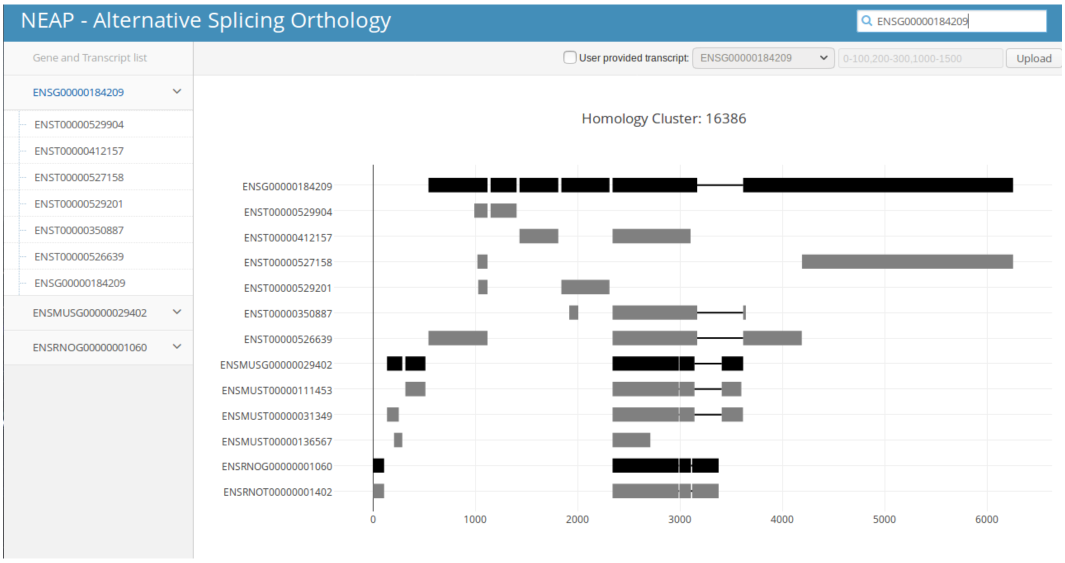
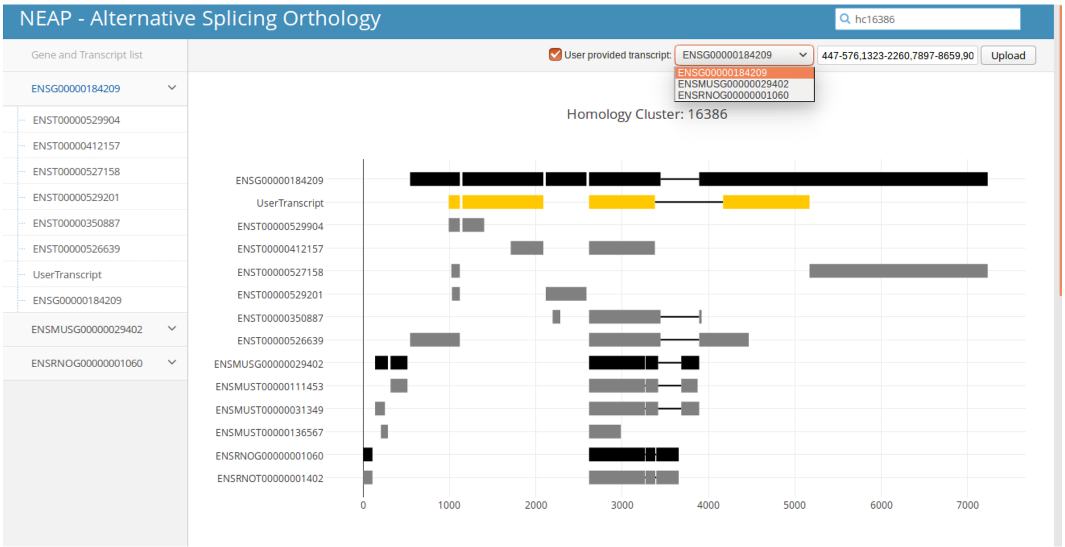


Figure 86 Figure 27: The main graphical user interface of the visualizer without a loaded homology cluster. On the top right a search bar is located which can be used to search by a homology cluster ID or by an Ensembl gene ID and therefore to load the corresponding homology cluster into the visualizer. Below the search bar the user input area is located which is explained in detail in the next screenshot. The main window in the center is the place where the isoform structured alignment will be visualized. In the bottom are the user transcript reliability score prediction and the cluster splice site conservation panel with the calculated evidence and reliability scores. Additionally it isn't possible to interact with the user input area as well as the conservation and reliability score prediction panel as long as no homology cluster is loaded. On the left side the gene and transcript tree is located which shows all genes with their transcripts if a homology cluster will be loaded.



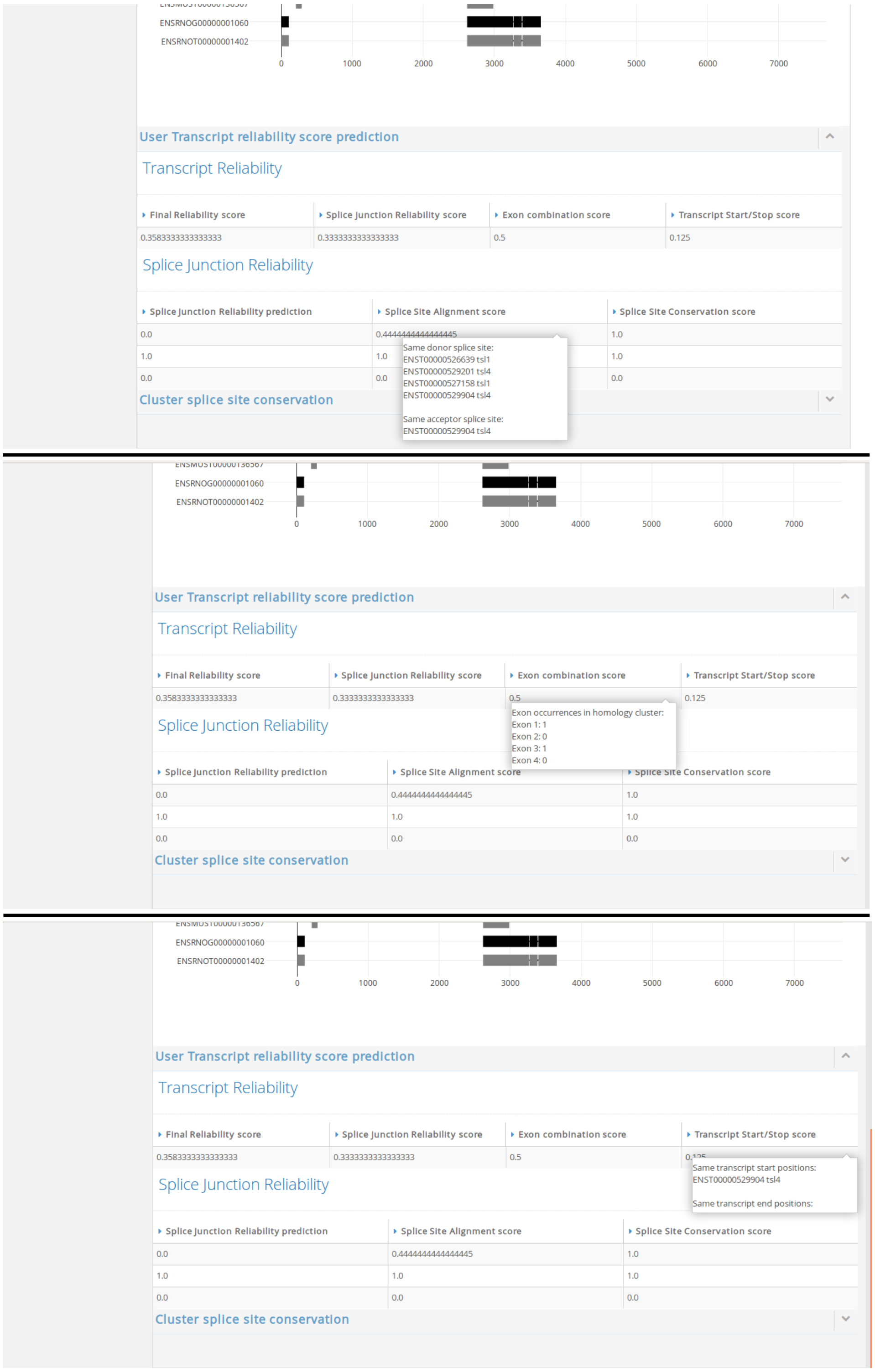
NEAP - Alternative Splicing Orthology

hc16386

☑ User provided transcript: ENSG00000184209

447-576,1323-2260,7897-8659,90

Figure 87 Figure 28: The loaded homology cluster 16386 with an uploaded user transcript. The homology cluster 16386 includes the human gene ENSG00000184209, the mouse gene ENSMUSG00000029402 and the rat gene ENSRNOG00000001060. On the left side the gene and transcript tree structure are visible. Every gene has its own panel which can be expanded to see all transcripts of the corresponding gene. The isoform structured alignment is presented in the center. The gray bars represent the exons of an isoform. The split exons which are connected with a small black line represent one exon which is just gapped in the isoform alignment. The black colored transcripts are the union transcripts of the genes. The gaps between exons indicate introns in the isoform structured alignment to establish a better overview (intron gaps have fixed sizes based on the homology cluster size and therefore don't represent the actual intron size). Due to loading a homology cluster, it is possible to interact with the user input area. It is possible to select a gene of the current loaded homology cluster and put an user isoform in text format in. For example the text input '0-100,200-230' defines a transcript with two exons. One exon of length 100, which starts at position 1 of the selected gene sequence, and another exon of length 30, which starts at position 201 of the selected gene sequence. The uploaded isoform is presented with the transcript name *UserTranscript* and is yellow colored.



User Transcript reliability score prediction

Transcript Reliability

Final Reliability score	Splice Junction Reliability score	Exon combination score	Transcript Start/Stop score
0.3583333333333333	0.3333333333333333	0.5	0.125

Splice Junction Reliability

Splice Junction Reliability prediction	Splice Site Alignment score	Splice Site Conservation score
0.0	0.4444444444444445	1.0
1.0	1.0	1.0
0.0	0.0	0.0

Same donor splice site:
ENST00000526639 tsl1
ENST00000529201 tsl4
ENST00000527158 tsl1
ENST00000529904 tsl4

Same acceptor splice site:
ENST00000529904 tsl4

Cluster splice site conservation

User Transcript reliability score prediction

Transcript Reliability

Final Reliability score	Splice Junction Reliability score	Exon combination score	Transcript Start/Stop score
0.3583333333333333	0.3333333333333333	0.5	0.125

Exon occurrences in homology cluster:
Exon 1: 1
Exon 2: 0
Exon 3: 1
Exon 4: 0

Splice Junction Reliability

Splice Junction Reliability prediction	Splice Site Alignment score	Splice Site Conservation score
0.0	0.4444444444444445	1.0
1.0	1.0	1.0
0.0	0.0	0.0

Cluster splice site conservation

User Transcript reliability score prediction

Transcript Reliability

Final Reliability score	Splice Junction Reliability score	Exon combination score	Transcript Start/Stop score
0.3583333333333333	0.3333333333333333	0.5	0.125

Same transcript start positions:
ENST00000529904 tsl4

Same transcript end positions:

Splice Junction Reliability

Splice Junction Reliability prediction	Splice Site Alignment score	Splice Site Conservation score
0.0	0.4444444444444445	1.0
1.0	1.0	1.0
0.0	0.0	0.0

Cluster splice site conservation

Figure 88 Figure 29: User transcript reliability score prediction panel zoomed in. After uploading an user transcript, it is possible to open the user transcript reliability score prediction panel. It contains two tables. The first one is the global transcript reliability table which shows the three transcript reliability scores (splice junction reliability score, exon combination score and transcript start/stop score) with the final reliability score. The second table shows the calculated splice site alignment scores (ISA score), splice pattern conservation scores and the predicted splice junction reliability score by the SVM (which takes the two scores as input) for each splice junction. It is also possible to hover over the calculated reliability scores as it is illustrated in the last three screen shots. Through this it is possible to get more detailed information about the specific score and it explains which features or transcripts were used. This was implemented to let the user be able to easily get an overview why his/her isoform scores are scored as they are.

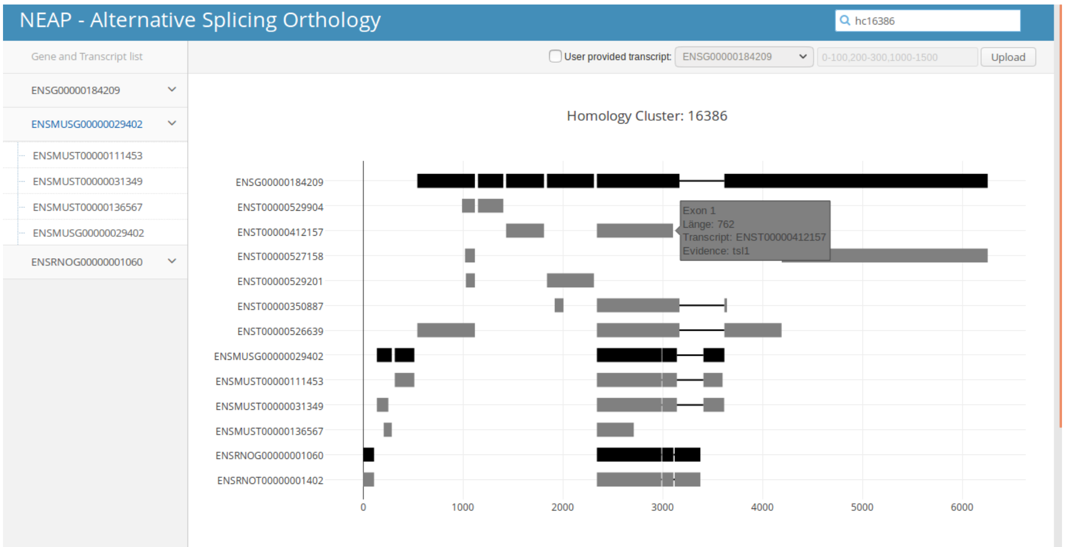


Figure 89 Figure 30: Hovering over an exon in the isoform structured alignment plot. If the user hovers over an exon a window with multiple features appears. The first feature indicates over what exon rank of the transcript the user is hovering and the second describes the corresponding length of that exon. The third feature points the corresponding transcript name of the exon out and the last feature shows the TSL level of that transcript.

10.3 Expression Level

The described tissue read-count per gene annotation files (bed) serve as input for the Visualizer, to show expression pattern for the union transcripts. When a cluster is loaded, the contained genes bed-files are read and for each intron and exon the expression level is build. For this, the different tissue counts are normalized by the minimal number of reads over the whole gene (minimal tissue) and assigned to the gene regions. The region values, divided by region length, are then displayed by green boxes under the union transcripts with the width as relation to their expression value (high value equals big box).

10.4 Conservation

For the visualization of the splice sites and their conservation, we extracted the donor and acceptor sides sequences of introns. The intron of the clusters union transcript start, ends is marked with a blue bar and their sequence as a hovertext, as shown in figure 33.A). We also calculated the sequence conservation of the start,stop sequence pairs in the cluster, under consideration of their origins TSL score. The splice site conservation scores are displayed in an panel with a list, as shown in figure 33.B). The Items in the list are checkable so that the selected splice site bars in the plot change their color to yellow.

11 Conclusion

This project gives a interactive visualization of splicing events in gene homology clustered transcripts and provides a reliability score for new user added transcripts. The Visualizer can easily be extended with further evidences in following projects. The exon start, stop sequence conservation is additionally implemented but not used and the expression could be visualized over small bins, instead of the intron/exon range, to ensure a more detailed

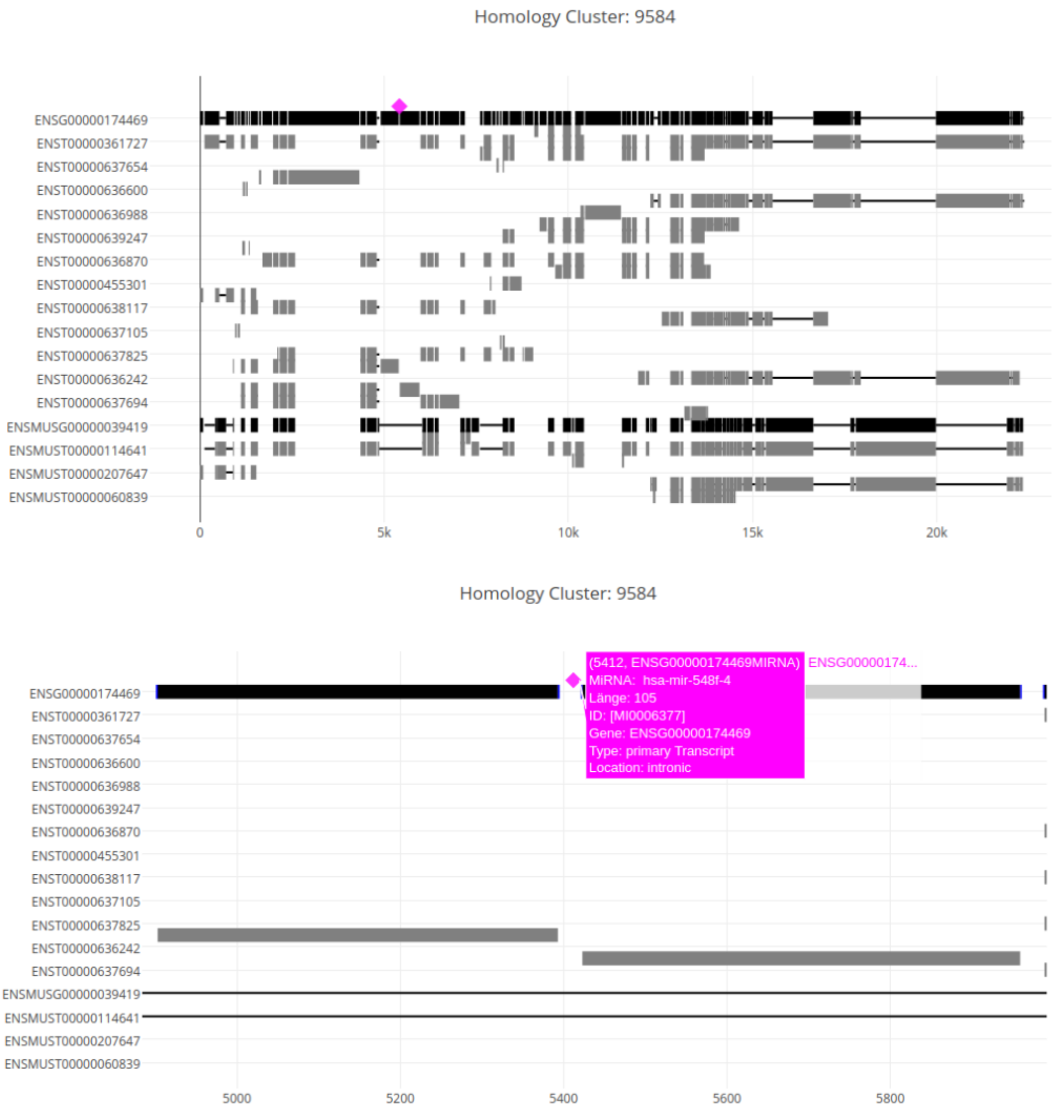


Figure 90 Figure 31: A view how miRNAs are visualized in our project. Overview and explicit zoomed in.

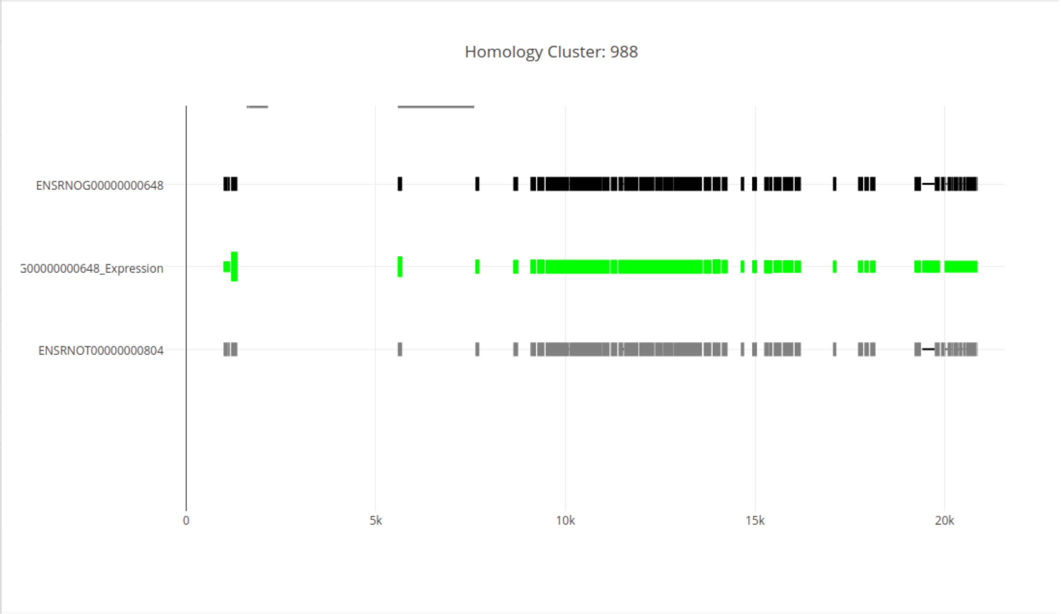
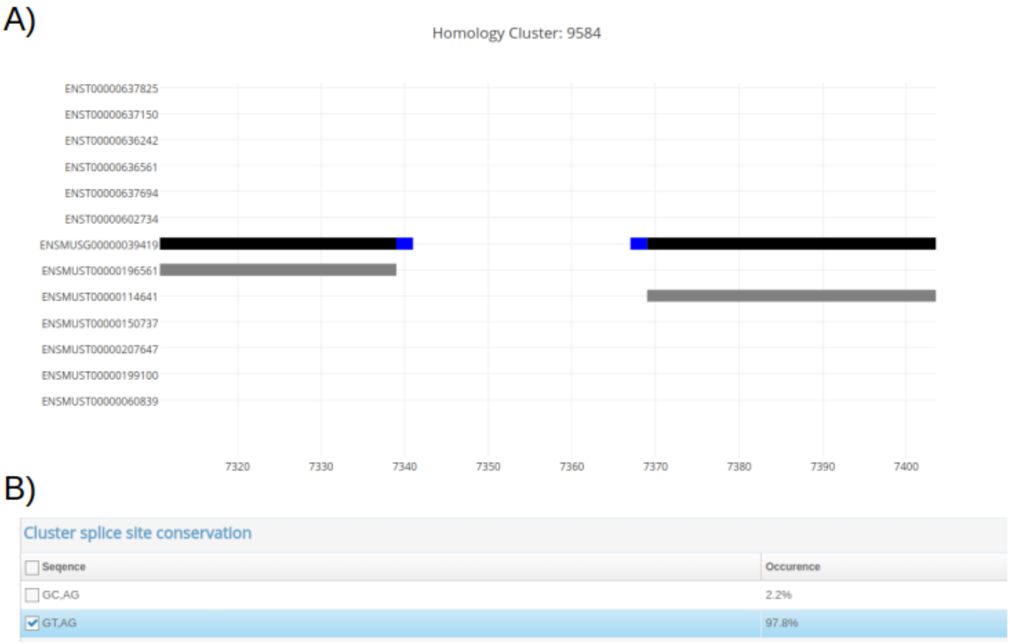


Figure 91 Figure 32: A view of expression visualization in the project. Read counts drawn in green for available union transcripts.



Cluster splice site conservation

□ Seqence	Occurence
□ GC,AG	2.2%
☑ GT,AG	97.8%

Figure 92 Figure 33: A view how splice sites are visualized in our project. A) the label of the intron start and stops as blue bars. B) A panel with the clusters splice site sequence conservation.

evaluation. These are just some points we thought of and partly translated into action, but couldn't fully realize.
Special thanks to our advisor Markus Joppich and supervisor Prof. Dr. Ralf Zimmer.

References

[91] L M Gallego-Paez et al. "Alternative splicing: the pledge, the turn, and the prestige : The key role of alternative splicing in human biological systems." In: *Human genetics* (2017). ISSN: 1432-1203. DOI: 10.1007/s00439-017-1790-y. URL: http://link.springer.com/10.1007/s00439-017-1790-yhttp://www.ncbi.nlm.nih.gov/pubmed/28374191.

12 Biological Networks

by Florian Hölzlwimmer, Daniel Schmitz and Evi Berchtold

12.1 Motivation

12.2 Basic network properties

When talking about biological networks, one first needs to understand the basics of the model used for their representation. We assume the reader is familiar with graph theory but for the sake of clarity we offer definitions of the terms used in this chapter for expressing important properties of graphs.
Let $G = (V, E)$ be a graph. We call V its *nodes* or *vertices* and $E \subseteq V \times V$ its *edges*. Let $n = |V|$ and $m = |E|$. n is called G's *order*, m is its *size*. An edge may be *directed*, i.e. one of the connected nodes is its *source* and the other its *target* (or *sink*), or *undirected*. Directed edges are usually represented as pairs of vertices while undirected edges can be represented as sets with two elements because the order of the connected nodes does not matter. An undirected graph contains only undirected edges whereas a directed graph contains directed ones. We call G connected if one can reach every node in V from every other node by walking along its edges. If it is not connected, it comprises at least two connected components which are the largest subgraphs that satisfy the condition for connectivity.
Each node $v \in V$ has a *degree* which is the number of edges it is connected to. If the graph is directed, each node has an *in-degree*, which is equal to the number of edges going to v, and an *out-degree* for the edges going from v. G's degree is the maximum degree of its nodes. The *diameter* of a graph is the longest path between two nodes in the entire graph.
A graph's clustering coefficient is a measure expressing how closely its nodes are connected. It is the ratio of existing edges and possible edges, i.e. $\frac{m}{n*(n-1)}$.
Each vertex v can be assigned a centrality. This is a measure for its importance in the modelled network. There are multiple approaches for determining it. These can be based on its degree, its probability of being visited on an arbitrary path or other more complicated algorithms. An example for a centrality measure is the betweenness centrality which is defined as the number of all shortest paths between each pair of nodes which pass through the selected node.

12.3 Network types

12.3.1 Gene Interaction Networks

Before applying information from biological networks to one's analyses, one must consider the different types they come in to make an educated decision which to use and how to use them. In the following paragraphs we will give a short overview over some of them.

12.3.1.1 Gene Co-expression Networks

Gene Co-expression Networks are networks modelling correlation of gene expression. The nodes of such a network are genes which are connected by an edge if there is a significant co-expression between them. Two genes are co-expressed if their expression correlates strongly between conditions. To assess genes' co-expressions one needs to generate expression data over many samples and conditions for all analysed genes. Because the relationships are derived from measurements of mRNA, one can only deduce co-regulation on the transcriptional level. Any regulation occuring after transcription cannot be captured.

Database	Evidence	№ of nodes	№ of edges
COEXPRESdb	E	43,617	*complete**
RegulonDB (*E. coli*)	E	4,653	3,261
RegNetwork	E, P	23,079	369,277
STRING	E, P, T, O	19,247	4,274,001
KEGG	E	812	2,645
Tarbase	P	799	831
ensembl	P	16,929	55,999
hprd	P	4,109	16,353
miRTarBase	P	1,949	2,727
miRecords	P	1,137	1,426
transmir	E	282	492
tred	E	2,953	6,726
UCSC	P	17,282	92,086

Table 5 Overview over the discussed databases (top section) as well as some which are note explicitly mentioned (bottom section). The numbers of edges and nodes are taken from the network for *Homo sapiens* unless otherwise noted. The abbreviations used in the column "Evidence" mean the following: *E*: Experimental, *P*: Predictions, *T*: Text-mining, *O*: Other.
*COEXPRESdb can be considered a complete graph: Since the graph contains co-expression data each pair of genes is "connected" by a correlation value.

A database providing access to co-expression networks is COEXPRESdb [123]. It contains 49 tissue-specific networks comprising between 50 and 1,001 genes. For the whole human organism, it provides co-expression data for 43,617 genes. The database can be queried by gene lists to obtain pairwise correlation values as well as genes which are co-regulated to all genes in the input. A query for a single gene returns a list of highly correlated genes and a network containing these genes as nodes and co-expression relationships as edges and additional genes (figure 93). These networks are annotated with correlation values and KEGG pathways containing multiple of their genes, providing evidence for the implied relationship.

12.3.1.2 Gene Regulatory Networks

Regulatory networks model effects of biomolecules on gene expression. Their nodes are proteins, RNA, DNA or complexes of those as well as genes. One database where one can obtain gene regulatory networks is RegulonDB [124] which is a comprehensive collection of operons, transcription factors and regulatory relationships of *E. coli*. The latest release contains 3,261 regulatory interactions of 4,653 genes. RegNetwork is another repository of regulatory interactions [125]. It contains curated data and predictions from TF-binding sites as well as interactions between transcription factors, genes and miRNAs. Its human regulatory network contains 23,079 nodes and 369,277 edges, can be searched by gene symbol or identifier from one of several other databases such as Ensembl or downlaoded as a whole.

12.3.2 Protein Interaction Networks

Proteins rarely perform their functions alone. Often they form complexes which can be permanent or temporary and become active together. Furthermore, proteins may modify each other, e.g. by phosphorylation. These protein-protein interactions are modelled by

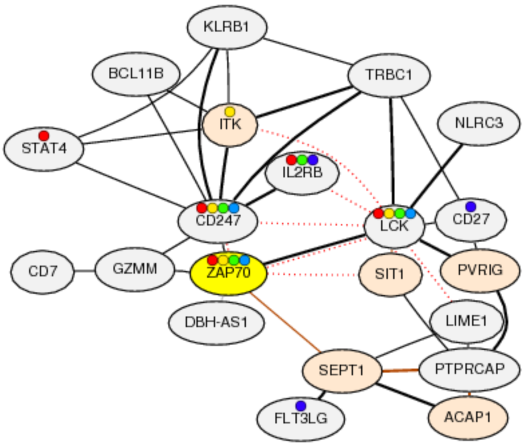


Figure 93 Co-expression network of gene `ZAP70`. The edge thickness encodes correlation, the color encodes evidence. The coloured rectangles in the nodes show common KEGG pathways.

protein interaction networks. There are many databases which provide access to them.

12.3.2.1 STRING

STRING is a database of known and predicted protein-protein interactions curated by EMBL [126]. While the access to FASTA and plain text files is free, the full database access using SQL requires a license. It contains networks from 2,031 organisms and over 1 billion interactions of around 9.6 million proteins. It can be searched by protein. The results are enriched using data fram GO, KEGG and PFAM as well as experimental data. One can browse every protein's neighbourhood, co-occurrences, and co-expression as well as information from text mining.

12.3.2.2 Phosphonetworks

Phosphonetworks is a database containing human phosphorylation substrates, sites and networks [127]. Each record has been experimentally verified. One can search for and download kinase-substrate relationships, phosphorylation motifs and high resolution networks by protein, phosphorylation site and network identifier.

12.3.3 Pathways

A type of biological network which can be particularly interesting and we focused on are pathways. The term "pathway" is vaguely defined and often used in a way which fits the current context. In general, a pathway can be considered a network of interactions and reactions of proteins, metabolites and other molecules related to a biological process. A change in activity of a pathway usually results in a change of the system's (i.e. cell's) state or a product.

12.3.3.1 KEGG

KEGG is probably the best-known pathway database in existence [128]. It contains 517 pathways including manually drawn maps (figure 94) represented in an XML based format with orders ranging from 14 to 4074. The pathways cover a multitude of biological processes

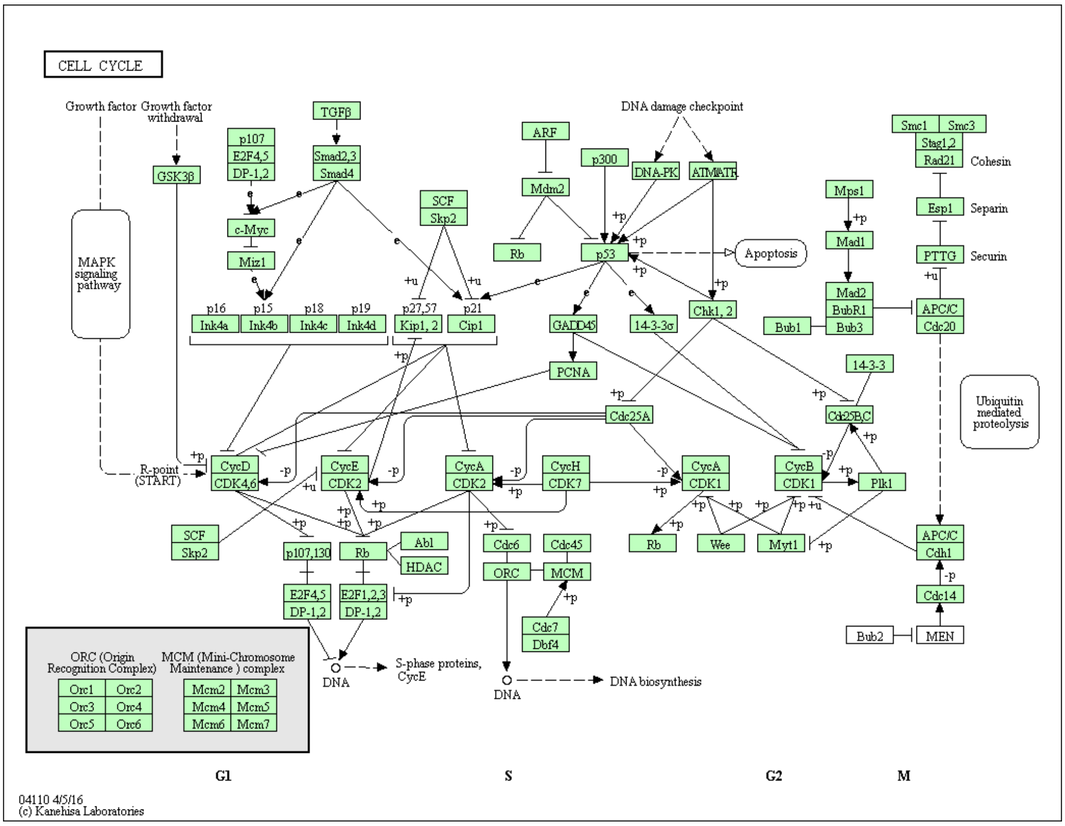


Figure 94 Pathway map of KEGG pathway `hsa04110`: *Cell cycle* showing the chemical processes relevant to the cell cycle in *Homo sapiens*. In addition to proteins it contains annotations for connected pathways, metabolites and general annotations providing information about the function of certain nodes. Edges are labelled according to their effect on their targets as well as the underlying process.

in 397 eukaryotes. They can be queried and analysed using multiple `R` packages which makes them especially useful for performing analyses.

12.3.4 Association networks

Association networks model relations between terms belonging to categories and can be considered the result of network- and set-based analyses.

12.3.4.1 Functional Networks

Functional networks are one of the results one can obtain from e.g. co-expression analyses. They contain information about cellular functions which are in relationship to one another—such as co-regulation. Here, these functions are represented as nodes and their relations as edges.

12.3.4.2 Disease Networks

Disease networks model the relationships of diseases either among each other or between them and e.g. genes or symptoms. They can be obtained from methods like GWAS and can act as an entry point for further studies. DNetDB is an example for a database providing a disease network associating diseases with each other based on similarity [129]. As this database was out of service at the time of writing, no statistics can be given.

12.3.5 Integrative Networks

All the types of networks mentioned above can be used to form integrative networks. These networks—as the name suggests—integrate information from multiple sources to enable discovery of underlying processes and relationships.

There are many more types of networks which can be integrated into a comprehensive analysis but these are not mentioned here as this inclusion would beyond the scope of this document due to their sheer number.

12.4 Analysis Methods

12.5 Network-based Analysis of Expression Data

Usually the first step in the investigation of expression data is to create a differential expression analysis. As this can result in thousands of significantly expressed genes, it is difficult to decide which expression changes are really important. Therefore, different types of methods to reduce this huge amount of values to easy-interpretable results. For example, “set-based enrichment” methods try to identify sets of genes (e.g. pathways), which are active in a certain experiment.

Many of these methods do not incorporate relations between genes. Though, considering expression values in context of a biological network delivers additional information which might be important to take into account. Due to this, “network-based enrichment” methods evaluate the consistency of certain expression patterns, rather than simply considering the amount of significant genes in a certain gene set. Other types of network-based expression analysis methods like “significant area search” use a biological network to identify sub-graphs of genes which are somewhat “interesting” in a certain experiment.

Another example of network-based analysis methods are “active transcription factor” predictors, which utilize a biological network to model the relations between genes and transcription factors.

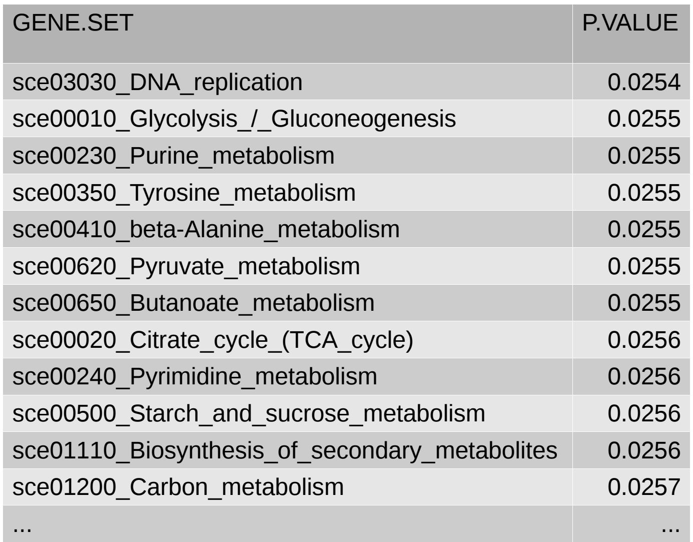
12.5.1 Enrichment Methods

Like already mentioned, “enrichment methods” try to find active biological processes by identifying certain groups of significantly enriched genes. These gene sets can be retrieved from various sources, e.g. biological annotation databases like “Gene Ontology”, which assign genes to certain functional classes, or biological pathways like KEGG and NCI. An example result for such an analysis can be seen in figure 95.

Many enrichment methods solely rely on gene set definitions and investigate the amount of significantly enriched genes in each set. A very common method of this type is called “Overrepresentation Analysis”:

12.5.1.1 Overrepresentation Analysis (ORA)

“Overrepresentation Analysis”[130] is a set-based enrichment method which aims at identifying significantly over-enriched gene sets. Given a certain gene set, ORA performs a



GENE.SET	P.VALUE
sce03030_DNA_replication	0.0254
sce00010_Glycolysis_/_Gluconeogenesis	0.0255
sce00230_Purine_metabolism	0.0255
sce00350_Tyrosine_metabolism	0.0255
sce00410_beta-Alanine_metabolism	0.0255
sce00620_Pyruvate_metabolism	0.0255
sce00650_Butanoate_metabolism	0.0255
sce00020_Citrate_cycle_(TCA_cycle)	0.0256
sce00240_Pyrimidine_metabolism	0.0256
sce00500_Starch_and_sucrose_metabolism	0.0256
sce01110_Biosynthesis_of_secondary_metabolites	0.0256
sce01200_Carbon_metabolism	0.0257
...	...

Figure 95 Example result of an enrichment analysis. This is an ordered list of gene sets, each one annotated with a p-value. Here, "DNA replication" is the most enriched result and therefore to be considered as the most clearly active process in the data set. Note that this example shows a hypothetical result.

hypergeometric test to check whether there is an unexpected number of differentially expressed genes in this set.

Lets assume a hypothetical expression anaysis, where a total of 5500 genes contain M=500 significant differentially expressed genes, while the remaining N=5000 genes are not significantly expressed. A hypergeometric test, also called Fisher's exact test, can be explained more easily by considering the significant genes as red balls and non-significant genes as black balls in an urn. When randomly drawing k=50 balls (respectively genes) without replacement from this urn, the corresponding hypergeometric probability distribution for obtaining exact *"x"* red balls (respectively significant genes) can be seen in figure 96. For example, if a gene set would contain k=50 genes from which x=10 genes are significantly expressed, the probability to observe this result *by chance* would be:

$$P(X = 10|k = 50, M = 500, N = 5000) = 0.85\% \qquad (14)$$

Therefore, this gene set would be considered as significantly enriched under a significance threshold of $\alpha = 5\%$, since $P \leq \alpha$.

To decide if some gene is differentially expressed ORA uses a fixed p-value threshold. Therefore, the analysis results of ORA strongly depend on the choice of this threshold.

12.5.1.2 Gene Set Enrichment Analysis (GSEA)

However, it is hard to reliably distinguish between differentially and non-differentially expressed genes. Using fixed p-value or fold change thresholds might causes unwanted artifacts when evaluating values close to this threshold. Also, such a threshold is always a somewhat "artificial" value.

Therefore, "Gene Set Enrichment Analysis"[131] tries to address this problem by using ranked lists of genes rather than fixed thresholds. GSEA ranks genes by their fold change or p-value and uses a Kolmogorov-Smirnov statistic to test whether the gene ranks in each gene set resemble a uniform distribution.

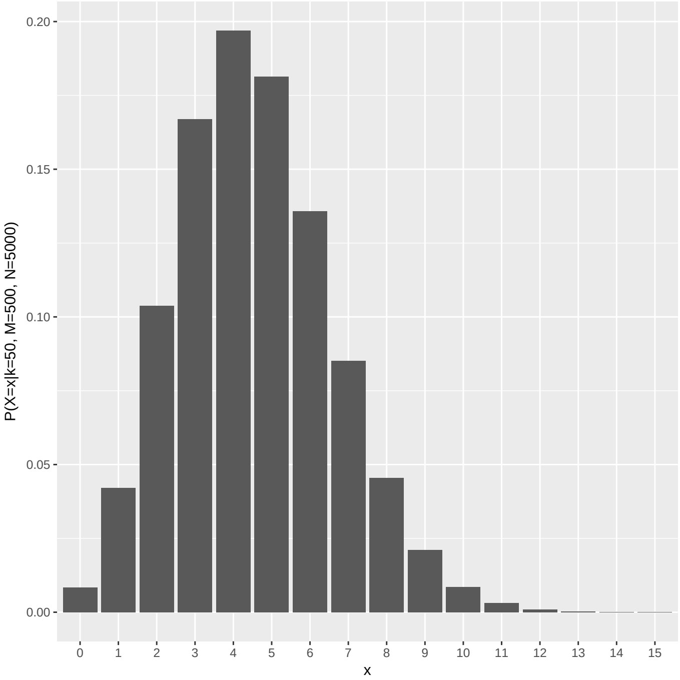


Figure 96 Hypergeometric distribution for k=50 draws, M=500 red and N=5000 black balls in an urn.

12.5.1.3 Gene Graph Enrichment Analysis (GGEA)

Since set-based enrichment methods like ORA and GSEA solely rely on comparing the amount of differentially expressed genes in each gene set, they ignore the relations between these genes. However, these relations might deliver important additional information. For example, when considering regulatory interactions between the genes in a certain set, some expression patterns can be more consistent than others (see figure 97).
In order to incorporate the relations between genes methods were developed which take into account information from gene interaction networks. An example for such network-based enrichment methods is “Gene Graph Enrichment Analysis”[132]. GGEA evaluates the consistency of the expression pattern in each gene set by assessing the meaningfulness of all relations inside the respective gene sets. These relations have to be provided in the form of a gene-regulatory network, in which each edge is known to be either activating or inhibiting. GGEA processes each gene set mainly in four steps:

1. Reduce the gene-regulatory network to the genes in the gene set. The result of this reduction is called the “induced subnetwork”.
2. Assign differential expression values (fold change, p-value) to each node in the induced subnetwork.
3. Decide, if the edges in the subnetwork are consistent or inconsistent.
4. Evaluate, if the observed consistencies are due to chance.

The first and the second step should be self-explanatory. In order to assess the consistency of an edge, GGEA performs a consistency check by comparing the expected expression change of the target gene with its observed change, similar to the example in figure 97. In detail, an

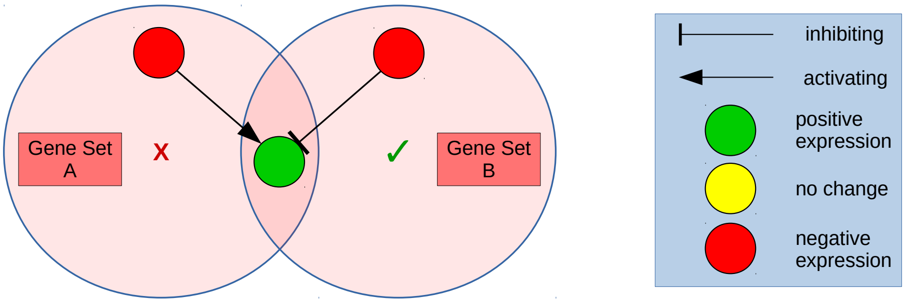


Figure 97 Consistency of expression patterns. In this example there are two gene sets, each containing one negatively expressed gene regulating one positively expressed gene. In gene set A this regulation is activating, while in gene set B this regulation is inhibiting. Logically seen, in case of an activating regulation the expression sign of the source gene should be equal to the expression sign of the target. Therefore, the expression pattern in gene set B is more consistent than in A.

approach called “fuzzification” is used to estimate the expected change of the edge’s target gene.

Estimating the expected change of some regulated target gene is not just as simple as in the previously shown example. Again, some thresholds have to be defined in order to decide if a gene is “postitively”, “unchanged” or “negatively” expressed. Also, the expression change of a gene might be not only “clearly positive/negative” or “not changed”, it could be also “intermediate positive/negative”. For each of these classes, the expected expression value of some target gene can be expressed, for example if the source gene is “intermediate negatively” expressed, any target gene inhibited by this source gene should be “intermediate positively” expressed.

Fuzzification aims at modelling these “fuzzy classes” by estimating the probability of each class that, given the expression of some gene, the gene belongs to this class. Therefore, instead of working with hard thresholds, expression values can be seen as “fuzzy” values. Then a so-called “fuzzy rule set” is applied to estimate the expected expression change of the target gene as “fuzzy” values. The resulting “fuzzy” values then can be “de-fuzzified” to real expression values. A schematic example of this procedure can be seen in figure 98.

The consistency of a whole gene set is calculated as the sum of the consistency scores of all edges inside the gene set, normalized by the number of edges. Significance then is tested by permutating the expression values and re-calculating consistency scores. The p-value then corresponds to the proportion of scores that are higher than the observed score.

12.5.1.4 Signalling Pathway Impact Analysis (SPIA)

SPIA[133] combines an “Overrepresentation Analysis” with a so-called “perturbation factor” to detect significantly enriched KEGG pathways. This “perturbation factor” of some gene $PF(g_i)$ describes the abnormal perturbation of that pathway, as measured by propagating measured expression changes across the pathway topology. It is defined as the sum of the gene’s expression change $\Delta E(g_i)$ and all upstream genes’ perturbation factors $PF(g_j)$, each of them normalized by their corresponding number of downstream genes $N_{down(g_j)}$:

$$PF(g_i) = \Delta E(g_i) + \sum_{j=1}^{N_{up(g_i)}} \frac{PF(g_j)}{N_{down(g_j)}} \tag{15}$$

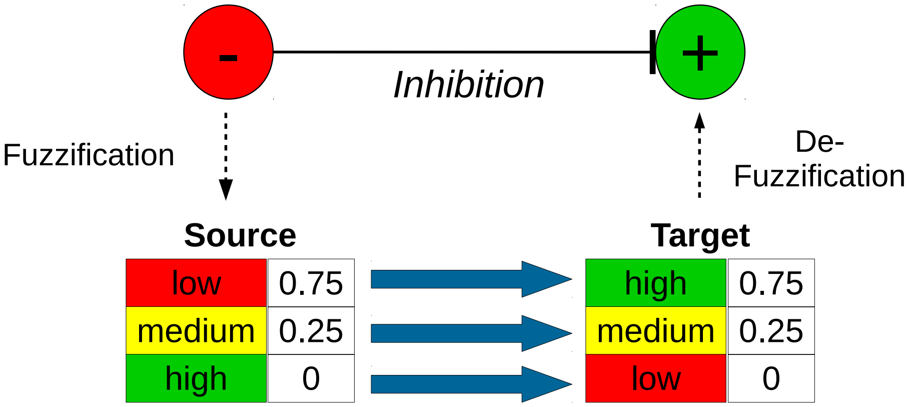


Figure 98 Schematic estimation of expression changes using fuzzification. In this example a negatively expressed source gene inhibits a positively expressed target gene. First, the expression value of the source gene is “fuzzified”, i.e. all probabilities that the expression value belongs to a certain “fuzzy class” are calculated. Next, a fuzzy rule set is used to transform the source’s fuzzy values to the target gene’s expected expression change. For example, in case of a low expressed source gene the target gene should be expressed high, therefore the probability of the “low” class is transformed into the probability of the “high” class. At last, the “fuzzy” values are being “de-fuzzified” to retrieve the expected change of the target gene as real expression value.

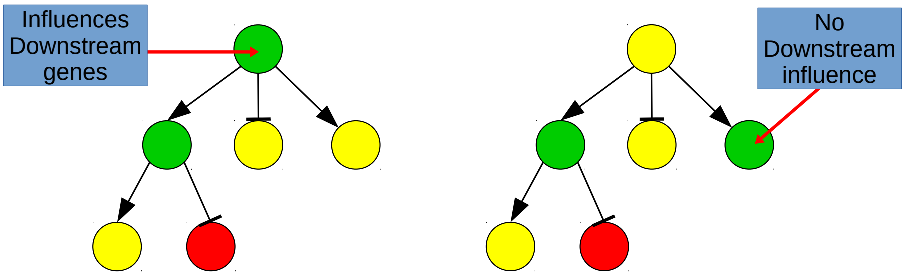


Figure 99 Genes regulating a lot of other (downstream) genes have a higher impact on the whole pathway than genes regulating less other genes

The idea behind that “perturbation factor” is that genes which regulate lots of other (downstream) genes have a higher impact on the whole pathway than genes regulating fewer other genes (see figure 99). Conversely, the impact of a gene on a target will be higher if the source gene exclusively regulates this single target. However, KEGG pathways are no acyclic trees, i.e. the “permutation factors” depend on the actual state of the system, as there can be cyclic dependencies between the “permutation factors”. Therefore, SPIA performs some type of *flux balance analysis* to identify the stable state of this perturbation system. Subsequently, the net perturbation accumulation at the level of each gene is calculated:

$$Acc(g_i) = PF(g_i) - \Delta E(g_i) \tag{16}$$

This accumulation value then is used for an “Overrepresentation Analysis” of all pathways.

12.5.1.5 Gene Association Network-based Pathway Analysis (GANPA)

GANPA[134] is a method which determines, given a network of gene interactions, the importance of a gene for different gene sets. For example, if a gene only interacts with other genes in the same gene set and therefore does not interact with genes outside that gene set,

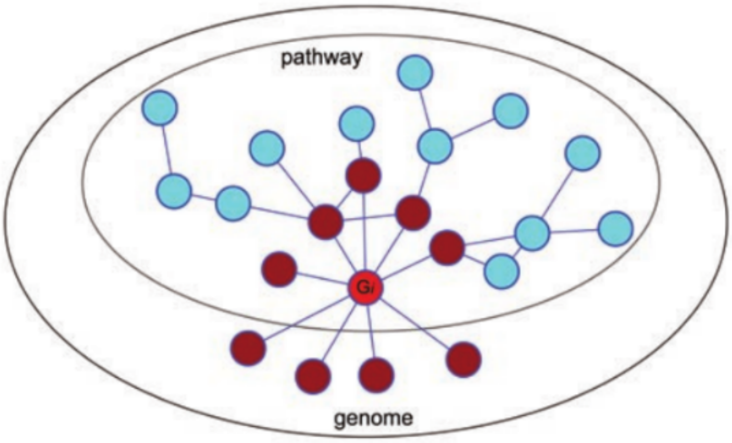


Figure 100 Importance assumption of GANPA[134]. Given a gene G_i in some gene set, the gene is more important to the gene set if the gene has more interaction partners (red genes) inside that gene set than outside.

that gene is very important to this particular gene set. More specifically, GANPA uses a hypergeometric distribution model to estimate the specific association between a gene and a gene set. The resulting gene weights then are used in a customized Gene Set Enrichment Analysis called "weighted GSEA" or "W-GSEA".

12.5.1.6 Functional Link Enrichment of Gene Ontology or Gene Sets (LEGO)

LEGO[135] is a very similar approach like GANPA. As well as GANPA, its target is to incorporate the contribution of a gene to some gene set, based on the interactions between the genes. In fact, GANPA was developed by the same team as LEGO.

However, LEGO adopts GANPA's network-based gene-weighting approach to decide between "interesting" and "uninteresting" genes in each set. This decision not only considers the centrality of some gene to some gene set but also the number of links to other central genes in the gene set. Then, LEGO uses an Overrepresentation Analysis using these interesting genes to detect significantly enriched gene sets.

12.6 Active TF prediction

Transcription factors play a major role in gene regulation. As proteins binding to specific DNA sequences, they influence gene expression greatly. Using techniques like ChIP-seq one can determine the binding sites of a given transcription factor pretty easily. A transcription factor's activity can be quantified as well using experimental strategies but an exhaustive analysis is not feasible because of their abundance. Performing measurements for each transcription factor would take a lot of resources. Therefore it is advisable to identify candidates whose activity might be linked to an observed condition such as an inherited or acquired disease *in silico* and confirm or reject the result experimentally. TF activity prediction tools try to estimate the change in activity of given transcription factors based on expression data and network models under certain conditions.

12.6.1 ISMARA

ISMARA offers an all-in-one pipeline for processing experimental gene expression data such as RNA-seq or microarray data [136]. It is available through a web interface and offers the ability to upload experimental data in multiple formats.

ISMARA tries to estimate transcription factor binding motif activities using a linear model called the MARA model which is formulated as follows:

$$E_{ps} = \bar{c}_s + c_p + \sum_m N_{pm} * A_{ms} + \text{noise}$$

where E is the expression matrix, N_{pm} the number of motif hits of the given feature, A_{ms} the activity of the transcription factor in the current sample and $\bar{c}_s, c_p$ are constants specific to sample and feature.

This model was not designed to explain much of the observed variance. To make a prediction, one instead performs an *in silico* knock-out of each transcription factor. This is done by choosing a motif m and setting N_{pm} for all promoters p having a binding site for m to 0, i.e. all these promoters are mutated so that the transcription factor cannot bind. After knocking out one TF, the model is fitted again. From the change of the variance explained by the new fit a score is calculated for each transcription factor, predicting which features are regulated by which transcription factors.

Although ISMARA provides a convenient way of analysing data, the necessity of uploading raw RNA-seq data made it unsuitable for our project because of privacy concerns.

12.6.2 RACER

RACER (Regression Analysis of Combined Expression Regulation) does not predict transcription factor activity *per se*. Instead, it makes a prediction of condition-specific transcription factor/miRNA-gene interactions, internally predicting their activities, as well [137]. The process of analysis is shown in figure 101.

In stage 1, RACER models the observed expression dependent on copy numbers, DNA methylation, transcription factor binding signals and miRNA expression as well as miRNA targets. This model is fitted for every sample over all genes. As a result, one obtains a matrix of activities per sample. The coefficients fitted for methylation and copy number variations are discarded, while the ones for miRNAs and transcriptions factors are considered measures for their activities. They are subsequently called α_{miR} and α_{TF}, respectively.

In stage 2, RACER fits a second linear model. The gene expression is modelled as dependent on copy number variations, DNA methylation and the activities obtained from stage 1—α_{miR} and α_{TF}. This model is fitted for each gene across all samples. The fit results in a matrix containing the condition specific association of genes and transcription factors/miRNAs inferred from their activities.

12.7 Subnetwork Search Methods

12.7.1 Significant Area Search

12.7.1.1 SteinerNet

SteinerNet is a web service using a model of the human interactome—namely a network of protein-protein interactions and transcription factor-gene interactions—in conjunctions with proteomics and transcriptional data to discover previously unknown associations [138].

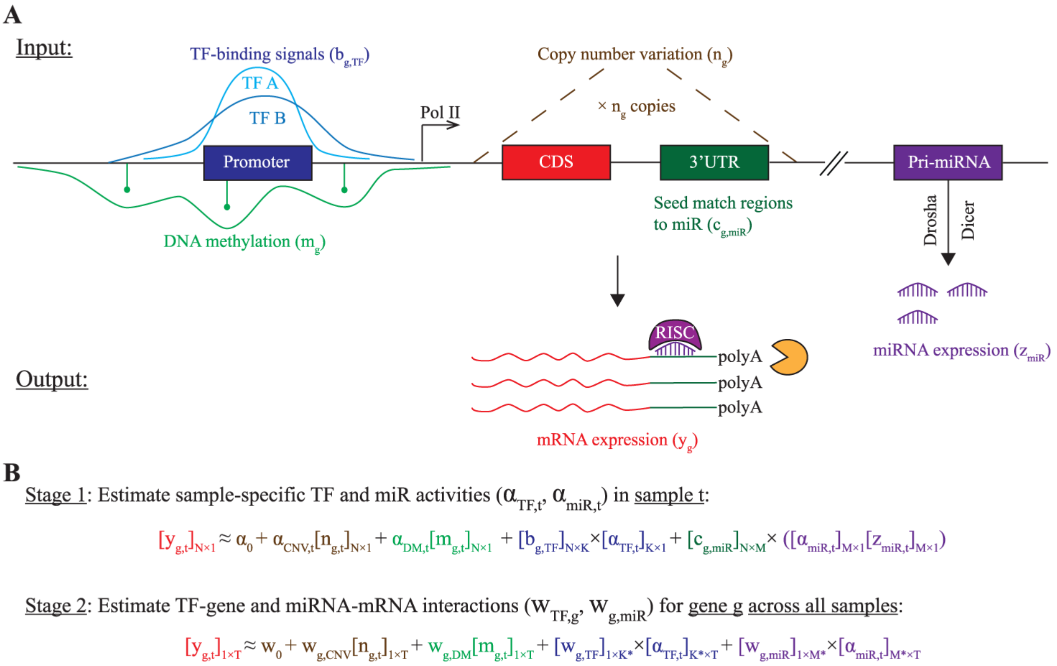

A
Input:
TF-binding signals ($b_{g,TF}$)
TF A
TF B
Promoter
Pol II
Copy number variation (n_g)
× n_g copies
CDS
3'UTR
Pri-miRNA
Seed match regions to miR ($c_{g,miR}$)
DNA methylation (m_g)
Drosha
Dicer
RISC
polyA
polyA
polyA
miRNA expression (z_{miR})
Output:
mRNA expression (y_g)
B
Stage 1: Estimate sample-specific TF and miR activities ($\alpha_{TF,t}$, $\alpha_{miR,t}$) in sample t:
$[y_{g,t}]_{N\times1} \approx \alpha_0 + \alpha_{CNV,t}[n_{g,t}]_{N\times1} + \alpha_{DM,t}[m_{g,t}]_{N\times1} + [b_{g,TF}]_{N\times K}\times[\alpha_{TF,t}]_{K\times1} + [c_{g,miR}]_{N\times M}\times([\alpha_{miR,t}]_{M\times1}[z_{miR,t}]_{M\times1})$
Stage 2: Estimate TF-gene and miRNA-mRNA interactions ($w_{TF,g}$, $w_{g,miR}$) for gene g across all samples:
$[y_{g,t}]_{1\times T} \approx w_0 + w_{g,CNV}[n_{g,t}]_{1\times T} + w_{g,DM}[m_{g,t}]_{1\times T} + [w_{g,TF}]_{1\times K*}\times[\alpha_{TF,t}]_{K*\times T} + [w_{g,miR}]_{1\times M*}\times[\alpha_{miR,t}]_{M*\times T}$

Figure 101 Visualization and formulations of the linear models used by RACER. Taken from [137].

It maps the interactome data to a network whose nodes represent genes, proteins and transcription factors and whose edges represent known interactions. It then tries to find the optimal tree connecting the nodes deemed significant.

SteinerNet models the problem of finding the association between significant genes and proteins using the prize-collecting Steiner tree problem (PCST). Let $G = (V, E)$ be a graph of interactions. Let $v' \subseteq V$ a set of nodes called *terminal nodes*. In the case of SteinerNet, the terminal nodes are significantyle differentially expressed genes and abundant proteins. The task is to find a tree in G connecting as many terminal nodes as possible. Each terminal node is assigned a penalty $p(v)$ and each edge $e \in E$ a cost $c(e)$. The penalties and costs are assigned according to the experimental measurements and the interaction probabilities of the connected nodes. The result of SteinerNet is a Tree $T \subset G$ with $T = (V_T, E_T)$ which minimizes the following objective function f:

$$f(T) = \beta * \sum_{v \in V \setminus V_T} p(v) + \sum_{e \in E_T} c(e)$$

where β is a user-supplied parameter representing the trade-off between including additional edges and leaving out terminal nodes. As the formulation of f shows, leaving out terminal nodes if their distance to the other ones is too big is a cromulent strategy. Finding a Steiner tree is a NP-hard problem, but there exists an exact solution for PCST which is used by SteinerNet [139]. This solution is based on a branch-and-cut approach and employs a preprocessing step to reduce the search space. When calculating the optimal solution, its complexity is not polynomial. SteinerTree's result can be viewed on-line and downloaded.

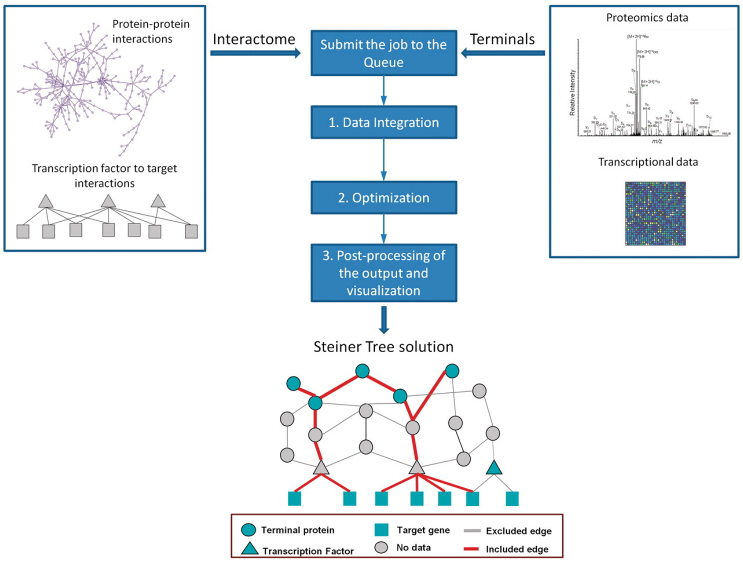


Figure 102 Overview of the SteinerNet workflow. Taken from [138].

12.7.1.2 RelExplain

RelExplain[140] is a method to analyse a single biological process in detail. It aims at identifying the most meaningful explanation for a given expression pattern inside that biological process.

Similar to SteinerNet, RelExplain tries to explain a set of "terminal nodes" by connecting them with a price-collecting Steiner tree. Therefore, RelExplain first maps differential expression data to a network of gene relations and identifies the terminal nodes as all significantly differentially expressed genes inside the target process.

RelExplain is designed to incorporate as much information as possible about the genes under investigation. To this end RelExplain uses a network which can contain different types of edges from sources like "Protein-Protein Interaction" databases, edges from gene regulatory networks or even gene relations gathered by text mining. Therefore, edges may be directed or undirected as well as activating or inhibiting. Subsequently, RelExplain scores all edges in this network based on the expression value of the adjacent genes, the membership in the target process as well as "reliability scores" which describe the reliability of the respective edge source (text mining edges will be less reliable than e.g. experimentally validated PPI's). At last, RelExplain uses a heuristics to connect all terminal nodes with a price-collecting Steiner tree. The subnetwork induced by this Steiner tree then is referred to as the "best" explanation for the given expression pattern.

A schematic visualization of the RelExplain algorithm [140] can be viewed in figure 103.

12.7.2 Identification of Driver Pathways

An obvious question one might ask if working with biological networks like pathways is whether there are pathways or parts of them which are significantly deregulated and if yes, which parts these are. Such subnetworks might be especially interesting because they can show which reactions or interactions are essential for a given process and what components are important.

One such tool is InFlo [141]. In general, InFlo tries to assess the relevance of a pathway and

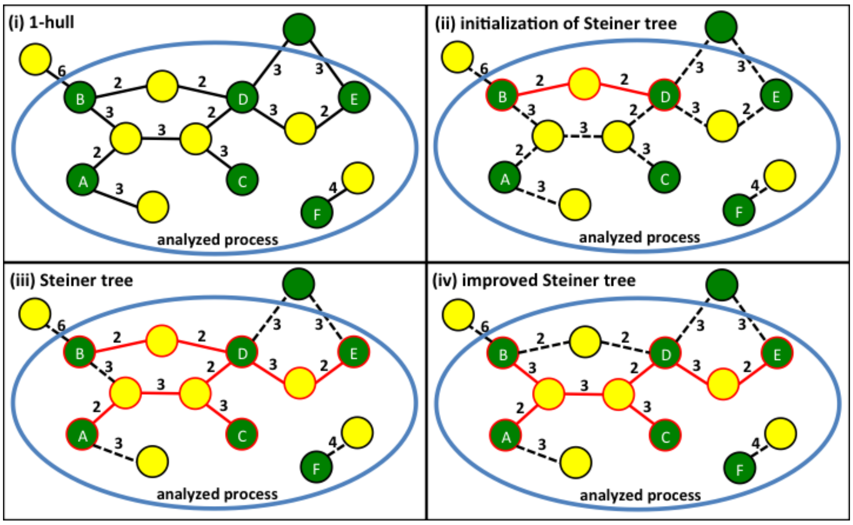

(i) 1-hull
(ii) initialization of Steiner tree
(iii) Steiner tree
(iv) improved Steiner tree
A
B
C
D
E
F
analyzed process

Figure 103 Schematic visualization of the RelExplain algorithm [140]**.** Green: terminal nodes (= differentially expressed genes inside analysed process), yellow: all other gene nodes. The current Steiner tree is marked red.
At first (i), the network is restricted to the d-hull of all terminal nodes (here: d=1). Next (ii), a new Steiner tree is being initialised with the shortest path between two terminal nodes. Then (iii), the remaining terminal nodes are added iteratively to the current Steiner tree, ordered by the shortest path between any node in the tree and any terminal node outside the tree. At last (iv), for each non-terminal node it is checked, if it can be removed, i.e. the tree would still be connected without it.

its interaction by modelling the flow of activation state through the network. It incorporates expression data (RNA-seq or microarray) methylation data and copy number variations into its analysis to account for their influence on gene expression.
For the analysis of a pathway for a given sample set it operates in the following way:

1. Model the pathway internally using the supplied representation and use the available annotation to map each node to its corresponding gene, if possible.
2. Calculate each gene's activity $A(i)$ for each sample.
   To derive the activity, a differential expression analysis is performed. The resulting value and the respective methylation β-values and copy numbers are incorporated into the activity value.
3. Predict each gene's state.
   Each gene i can be in one of three states: -1, meaning inhibited in disease samples, 1 meaning activated and 0 which indicates no difference between normal and disease samples. Because of the uncertainty coming from the measurements, no gene is assigned a single state. Instead, one computes a state vector which contains the probability that i is in the respective state: $S(i) = [P(-1 \mid A(i)), P(0 \mid A(i)), P(1 \mid A(i))]$. The most likely state is called $S_{MAP}(i)$. The probability function used for this calculation is learned from the normal samples provided with the data by fitting normal distributions to their measurements.
4. Deconstruct the pathway.
   Each pathway is split into interactions which consist of multiple sources and targets. An interaction is defined as a subgraph where each source is connected to every target and there is no larger subgraph containing the interaction for which this condition still holds.

5. Infer interaction state.
   Next, each interaction's state $S(I)$ is inferred from the calculated states of its sources. Just like genes, interactions are assigned a state vector. Let P be the set of sources, T the set of targets and p, t their respective numbers. For an arbitrary state a, its probability is calculated as follows:

   $$S_a(I) = \sum_{i \in P} \sum_{j \in T} r_{ij} * S_a(i)$$

   $$\text{with } r_{ij} = \begin{cases} \frac{1}{p} & \text{edge } (i,j) \text{ activating} \\ -\frac{1}{p} & \text{edge } (i,j) \text{ inhibiting} \end{cases}$$

   Like genes, interactions have a most likely state S_{MAP}, as well. The computed state vector can be propagated through the network to predict the activities of nodes which have no experimental data attached to them.
6. Consistency evaluation.
   After computing the interaction state, its consistency with the experimental data is assessed. An interaction is called consistent, if its S_{MAP} matches at least half of the targets' S_{MAP}. From this classification, the probability of this interaction modelling the observed data correctly is calculated using a Bayesian model. Each pathway is assigned a vector containing all probabilities for the scored interactions.

The resulting probability vectors can be used for further analysis, e.g. in a supervised setting. They also imply a simple way to define a distance between each sample.

In reality, Dimitrova *et al.* chose a somewhat different approach which they detail in their paper, as well, using the fundamental model as a guideline. Instead of propagating state vectors, state configurations for the whole pathway are sampled from the distributions. These configurations are checked for consistency and the probability of an interaction modelling the experimental data correctly is determined from the number of configurations it was consistent in. The output of InFlo is not a vector of probabilities but a vector of scores ranging from -4 to $+4$.

The default dataset consists of pathways from NCI-PID and BioCarta and experimental data from 23 patients with acute myeloid leukaemia.

12.8 TCGA

For our analysis we used experimental data from *The Cancer Genome Atlas* (TCGA) [142]. We took data from 99 breast cancer patients. These datasets include RNA-seq and microarray gene expression quantification, copy number variations, DNA methylation data and miRNA expression data from tumour as well as normal samples. Copy Number variations were supplied for arbitrary regions but our tools required them to be annotated per gene. Therefore, we modified our data by calculating the copy number of each gene as implied by the data provided. To do this, we intersected the annotated regions and all our known genes. If a gene overlaps with at least one copy number variation, the copy number for this gene is calculated by computing the average copy number of all regions overlapping with that gene weighted by the length of their overlap. Sequences in the given gene which don't overlap with any CNV are considered to have a copy number of 1 and are included in this average, as well.

As basis for our analyses we used pathway definitions and annotations from KEGG [128] (e.g. the pathway in figure 104), Gene Ontology [143] and NCI-PID (now included in NDEx) [144] maintained by the National Cancer Institute.

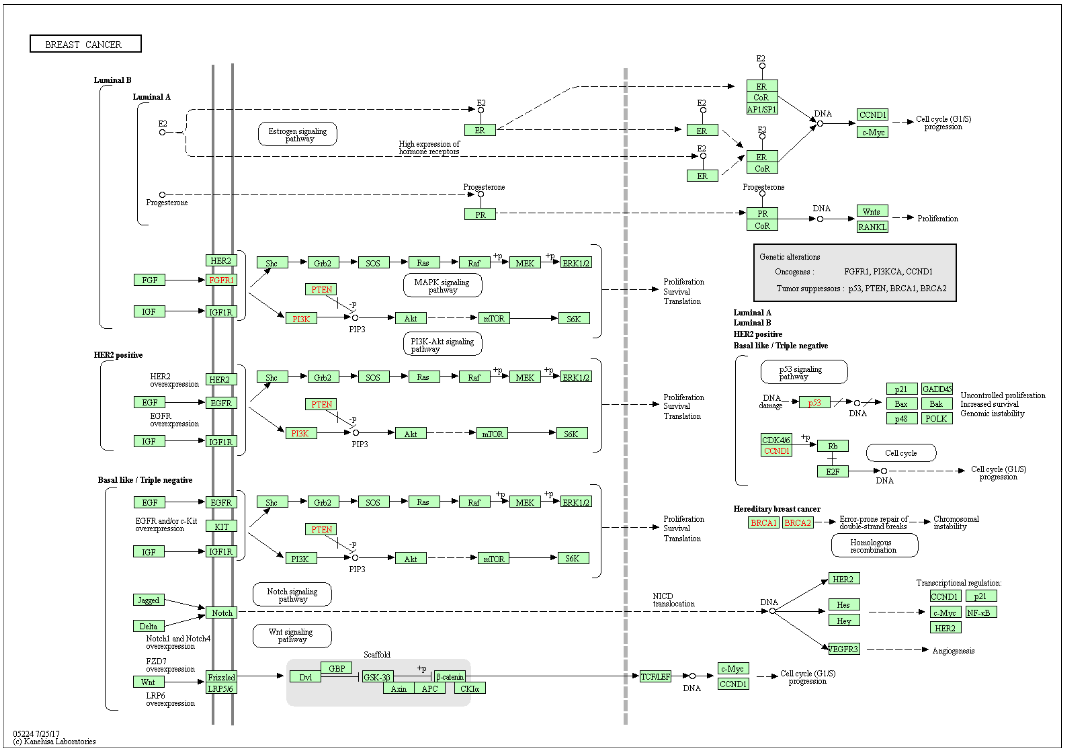


Figure 104 KEGG pathway "Breast Cancer" (`hsa:05224`). This pathway's order is 124 and its size is 98. It is one of the 22 sub-pathways of "Pathways in Cancer".

12.9 Enrichment Robustness

Here, we would like to compare different types of enrichment analysis methods. Our target is to check, whether the incorporation of gene relations helps network-based enrichment methods to create more sensible results.

To this end, we created an integrated workflow to systematically analyse the previously mentioned TCGA samples with different enrichment methods and compare the results in terms of robustness and meaningfulness.

All results can be viewed in detail by visiting our interactive web application at ■PLEASE ADD LINK HERE■. It allows to change different parameters like the selection of patients, the methods or gene sets and create custom analyses.

12.9.1 Workflow

A detailed overview of the workflow can be seen in Figure 105. In general, all 99 patient samples are examined for differential expression and subsequently analysed with different enrichment methods introduced in chapter 12.5.1. Then, it is possible to compare the enrichment methods or the 99 patient samples among themselves. The similarity of enrichment results can be measured by their "rank distance" or proportion of "gene set overlap" 12.9.1.3.

12.9.1.1 Differential Expression Analysis

Since our data from the TCGA patients consists of RNAseq reads, we used "edgeR"[145] to examine the samples for differential expression.

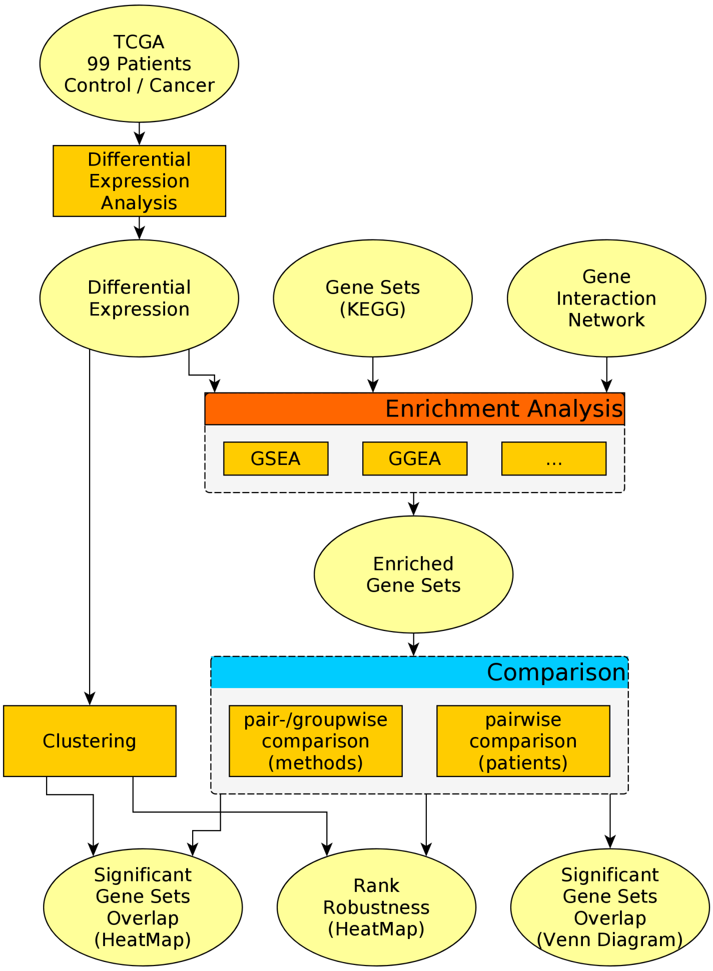

TCGA
99 Patients
Control / Cancer
Differential Expression Analysis
Differential Expression
Gene Sets (KEGG)
Gene Interaction Network
Enrichment Analysis
GSEA
GGEA
...
Enriched Gene Sets
Comparison
Clustering
pair-/groupwise comparison (methods)
pairwise comparison (patients)
Significant Gene Sets Overlap (HeatMap)
Rank Robustness (HeatMap)
Significant Gene Sets Overlap (Venn Diagram)

Figure 105 Workflow of enrichment analysis. At first, each TCGA patient sample has to be examined for differential expression. Next, the resulting differential expression analyses are further processed with different enrichment methods. At last, the enrichment results are compared by methods and patients. The similarity of two enrichment results can be measured by “rank distance” or “overlap” of the significantly enriched gene sets. These similarities can be visualized using heat maps and venn diagrams. In order to test the robustness of different enrichment methods, the patient heat maps can be clustered by fold change correlation of the different patient samples.

In order to analyse the differential expression of one single patient, we chose the tumor sample of this patient and compared it (using edgeR) to all reference samples. The reason why we chose this model is because of missing technical replicates: Usually TCGA patient samples are only sequenced once. As it is not possible to fit any meaningful distributions to a single value per group, we used the collection of reference samples as biological replicates instead. In this case, edgeR fits one distribution to all samples.

If there are more than three patients to be analysed for differential expression, we chose another model. In this case, we compare all tumor samples of the selected patients against all reference samples of these patients in a paired tumor versus control test.

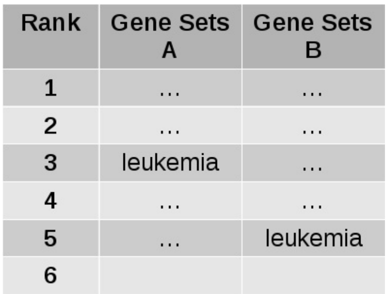
12.9.1.2 Clustering of Patient Samples

In order to cluster similar patients, we first created differential expression analyses for each single TCGA patient. Then, we used "pvclust"[146] to cluster the patients by their fold change correlation. "pvclust" is a tool which first performs a hierarchical cluster analysis and then assesses the uncertainty in the cluster analysis by calculating p-values using multiscale bootstrap resampling.

12.9.1.3 Statistics

There are two measures which we considered to compare two enrichment results, namely their "rank distance" and the proportion of "overlap" between their significantly enriched gene sets:

- **rank distance:**
  This measure describes the mean difference between the ranked enrichment results.



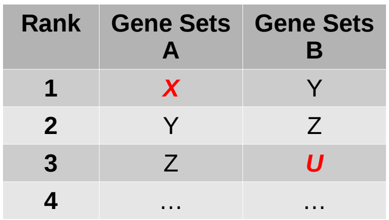
Rank	Gene Sets A	Gene Sets B
1	...	...
2	...	...
3	leukemia	...
4	...	...
5	...	leukemia
6		

In this example, the rank distance of leukemia would be $(5-3) = 2$. The overall rank distance corresponds to the mean of all rank distances.

- **overlap:**
  The overlap distance of two enrichment results is defined as the proportion of significantly enriched gene sets which are present in both results:

$$overlap = \frac{N(\text{“}significant\ in\ both\ results\text{”})}{N(\text{“}significant\ in\ A\text{”} \cup \text{“}significant\ in\ B\text{”})}$$

Therefore, in the following example the overlap proportion would equal 50%, as there are two significantly enriched gene sets which are present in both results (Y and Z) in a total of four gene sets (U, X, Y, Z) which are significant in at least one of both enrichment results.

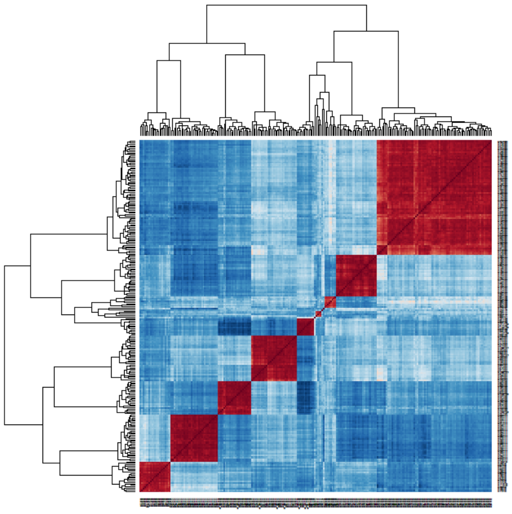


Rank	Gene Sets A	Gene Sets B
1	***X***	Y
2	Y	Z
3	Z	***U***
4	…	…

12.9.1.4 Visualization

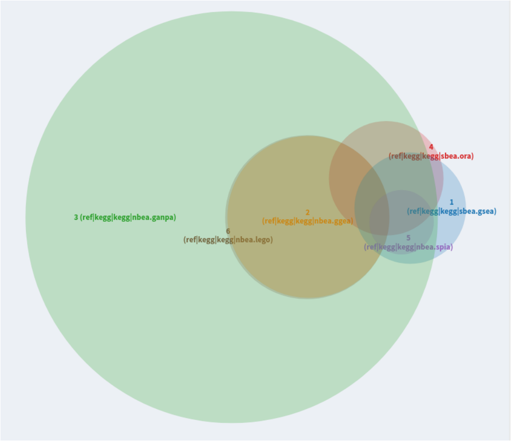
There are different types of visual representations used in our application. To display single enrichment results, the most obvious possibility is to print them as tables. However, it is also possible to represent them as a network of gene sets. By annotating nodes with different colors, this network representation can also be used for comparing multiple enrichment results. Other comparative visualizations include heat maps and Venn diagrams.

- **heat maps:**
  In order to display similarities between more than six entities, a heat map is often the best option:

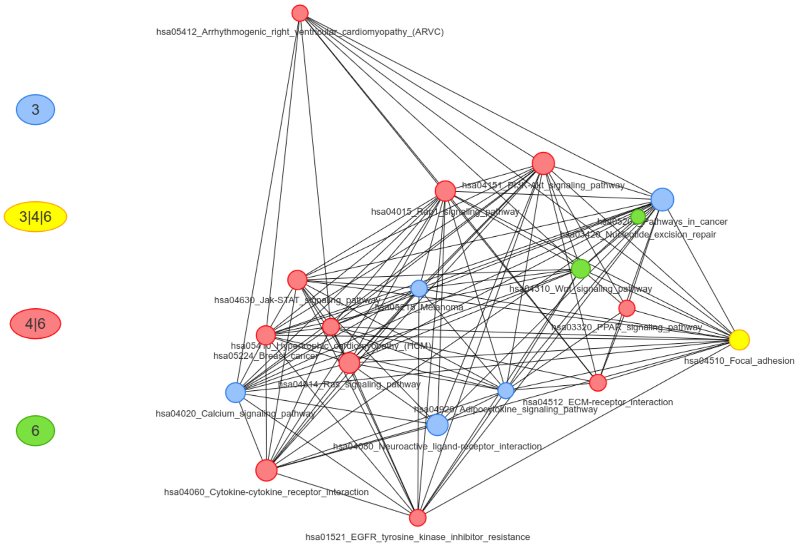
  Here, the rows and columns can also be ordered by different features, e.g. by clustering all patient samples by their expression correlation.

- **Venn diagrams:**
  To visualize the overlap of different enrichment results, Venn diagrams are a very intuitive representation:



Venn diagrams enable the identification of overlaps between more than two enrichment results. However, this type of visualization is only possible for a low number of enrichment results, since the possible number of overlaps grows exponentially with the number of enrichment results visualized.

- **gene set networks:**
  This type of network represents gene sets as nodes and the proportion of overlapping genes between two sets as weighted edges:


3
3|4|6
4|6
6
hsa05412_Arrhythmogenic_right_ventricular_cardiomyopathy_(ARVC)
hsa04151_PI3K-Akt_signaling_pathway
hsa04015_Rap1_signaling_pathway
hsa05200_Pathways_in_cancer
hsa03420_Nucleotide_excision_repair
hsa04310_Wnt_signaling_pathway
hsa04630_Jak-STAT_signaling_pathway
hsa05218_Melanoma
hsa03320_PPAR_signaling_pathway
hsa05410_Hypertrophic_cardiomyopathy_(HCM)
hsa05224_Breast_cancer
hsa04510_Focal_adhesion
hsa04014_Ras_signaling_pathway
hsa04512_ECM-receptor_interaction
hsa04020_Calcium_signaling_pathway
hsa04920_Adipocytokine_signaling_pathway
hsa04080_Neuroactive_ligand-receptor_interaction
hsa04060_Cytokine-cytokine_receptor_interaction
hsa01521_EGFR_tyrosine_kinase_inhibitor_resistance

Here, the size of the nodes corresponds to the logged number of genes in the respective gene set. Also, since this representation is using a force-directed graph layout, an edge will be shorter if the overlap between the two adjacent gene sets is higher. Nevertheless, the width of each edge directly shows the overlap weight of each edge.
The nodes in this network can be annotated with different features. In this case, the colors show the group of enrichment results which consider this gene set as significantly enriched.

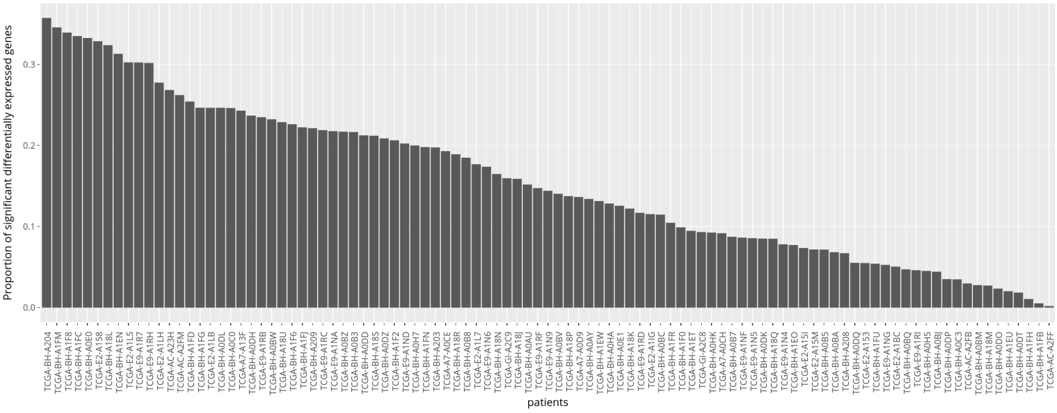


Figure 106 Proportion of significant differentially expressed genes in all TCGA patients. The patient with the highest value has a proportion of about 35% significant genes, but there are also patients where nearly no differential expression can be observed. The mean value is about 18% of significantly expressed genes per patient.

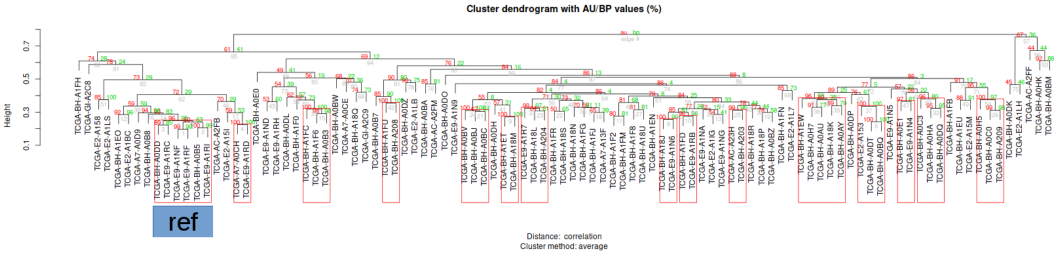


Figure 107 Cluster dendrogram of all TCGA patients. The patients were clustered by their fold change correlation. The red rectangles denote significantly clustered patients ($\alpha = 5\%$). The cluster labeled with "ref" shows the patient samples used as reference in later analyses. Using our interactive web application, the clusters can be selected by a custom α cutoff to choose groups of patients for further analyses.

12.9.2 Results

The first step in our workflow was to create differential expression analyses for all 99 TCGA patients using "edgeR" (see also chapter 12.9.1.1). The resulting proportion of significant differentially expressed genes per patient can be seen in Figure 106.

Next, we clustered all patients by their fold change correlation using "pvclust" (see also chapter 12.9.1.2). The resulting dendrogram can be seen in Figure 12.9.1.2.

Using our interactive web application, it is possible to choose different gene sets or patient groups for analysis. However, since changing the gene set produces similar results, in the following sections we will focus on "KEGG" gene sets together with a gene regulatory network compiled from "KEGG" pathways. In order to compare different enrichment methods we chose two patient selections as reference data sets. The first selection, abbreviated as "all", simply corresponds to all patients. For the second selection we chose six very similar patient samples as reference data set (labeled with "ref" in the dendrogram). For both selections we created differential expression analyses as explained in chapter 12.9.1.1. In Figure 108 the cumulative distribution of the p-values in both selections. It can be seen that the "all" selection contains a very high number of significant differentially expressed genes which indicates that the patient samples are quite different.

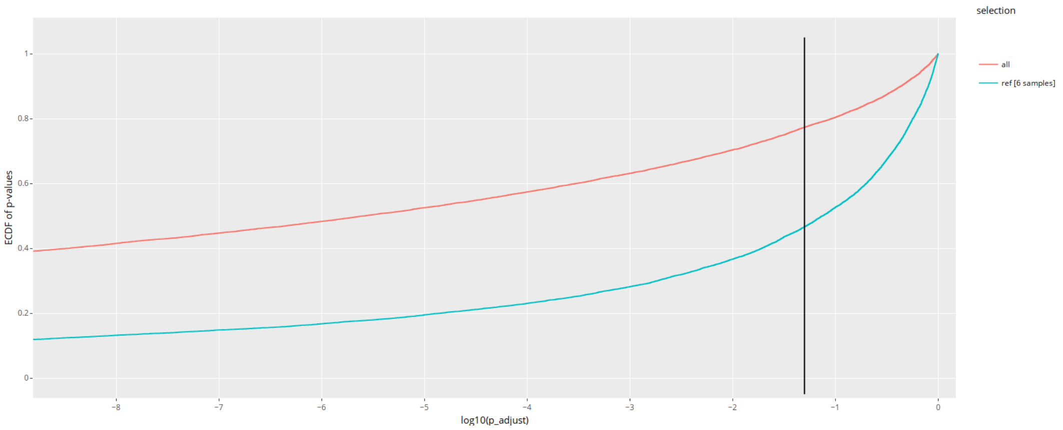


Figure 108 Cumulative distribution of p-values in both reference data sets. The black vertical line represents $alpha = 5\%$. It can be seen that the selection of all patients (“all”) contains nearly 78% significant differentially expressed, which is a very unusual high value. The “ref” selection with only six patient samples contains about 46% significant genes.

12.9.2.1 Single Enrichment Results

When looking at single enrichment results, it can be seen that both set-based enrichment methods “ORA” and “GSEA” are able to identify cancer pathways in the “ref” patients data set. In contrast, the network-based method “LEGO” and “GGEA” assess cancer pathways as not significantly enriched, even if they are able to detect some (breast) cancer-related pathways like the “PPAR Signalling Pathway”[147] between a high number of other gene sets (see Table 6). “SPIA” does not provide any indication for active cancer pathways in the samples (see Figure 111).

However, looking at the corresponding analyses for all patients (the “all” data set) changes everything. In this case, GGEA and LEGO directly identify “Breast Cancer” as best-ranked gene sets. Also, their number of significant results is lower and more precise (see Figure 112). SPIA now detects “Pathways in cancer” and “melanoma” as significantly enriched pathways, in addition to the gene sets found in the “ref” data set. GSEA and ORA instead are unable to detect any kind of significantly enriched gene sets, due to a limitation of this type of methods. As explained in section 12.5.1, ORA and GSEA both search for gene sets which are significantly enriched in comparison to other methods. Since the “all” data sets has a proportion of nearly 80% significant genes, they cannot find a gene set which is significantly more enriched than other gene sets, since all gene sets seem to be highly enriched. However, such a high number of differentially expressed genes is very unusual, such a high number of very different patient samples causes problems to estimate meaningful p-values for the differential expression of genes.

Surprisingly, “GANPA” does not create any sensitive results for both data sets, as it considers nearly all gene sets as significantly enriched (see Figure 113). We assumed that GANPA could have problems with our gene-regulatory network compiled from KEGG, since it was originally developed for bigger protein interaction networks. Therefore we examined GANPA’s behavior with different PPI networks like STRING. However, we were not able to achieve reasonable results under any circumstances.

	GENE.SET	P.VALUE
1	hsa05414 Dilated cardiomyopathy (DCM)	0.00132
2	hsa03320 PPAR signaling pathway	0.00132
3	hsa04512 ECM-receptor interaction	0.00132
4	hsa04971 Gastric acid secretion	0.00132
5	hsa04060 Cytokine-cytokine receptor interaction	0.00132
6	hsa04921 Oxytocin signaling pathway	0.00132
7	hsa04015 Rap1 signaling pathway	0.00132
8	hsa05032 Morphine addiction	0.00132
9	hsa04261 Adrenergic signaling in cardiomyocytes	0.00132
10	hsa04713 Circadian entrainment	0.00132
11	hsa04723 Retrograde endocannabinoid signaling	0.00132
12	hsa04510 Focal adhesion	0.00132
13	hsa04024 cAMP signaling pathway	0.00132
14	hsa04151 PI3K-Akt signaling pathway	0.00132
15	hsa04014 Ras signaling pathway	0.00132
16	hsa04918 Thyroid hormone synthesis	0.00198
17	hsa04970 Salivary secretion	0.00198
18	hsa04923 Regulation of lipolysis in adipocytes	0.00198
19	hsa04020 Calcium signaling pathway	0.00198
20	hsa04728 Dopaminergic synapse	0.00198
21	hsa04976 Bile secretion	0.0027
22	hsa04924 Renin secretion	0.0027
23	hsa04961 Endocrine and other factor-regulated calcium reabsorption	0.0043
24	hsa04727 GABAergic synapse	0.00634
25	hsa04724 Glutamatergic synapse	0.00634
26	hsa04540 Gap junction	0.0099
27	hsa05412 Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.01393
28	hsa05217 Basal cell carcinoma	0.01485
29	hsa05218 Melanoma	0.01639
30	hsa05410 Hypertrophic cardiomyopathy (HCM)	0.02096
31	hsa04913 Ovarian steroidogenesis	0.02096
32	hsa04972 Pancreatic secretion	0.02096
33	hsa04925 Aldosterone synthesis and secretion	0.02096
34	hsa04611 Platelet activation	0.02096
35	hsa04725 Cholinergic synapse	0.0215
36	hsa05020 Prion diseases	0.0297
37	hsa04630 Jak-STAT signaling pathway	0.03371
38	hsa04750 Inflammatory mediator regulation of TRP channels	0.03699

Table 6 Enrichment Result of "Gene Graph Enrichment Analysis". GGEA assesses cancer pathways as not significantly enriched, although it is able to detect (breast) cancer-related pathways like the "PPAR Signalling Pathway"[147]. Other significantly enriched gene sets like "Morphine addiction" or "Dilated cardiomyopathy" might be consequences of a therapy.

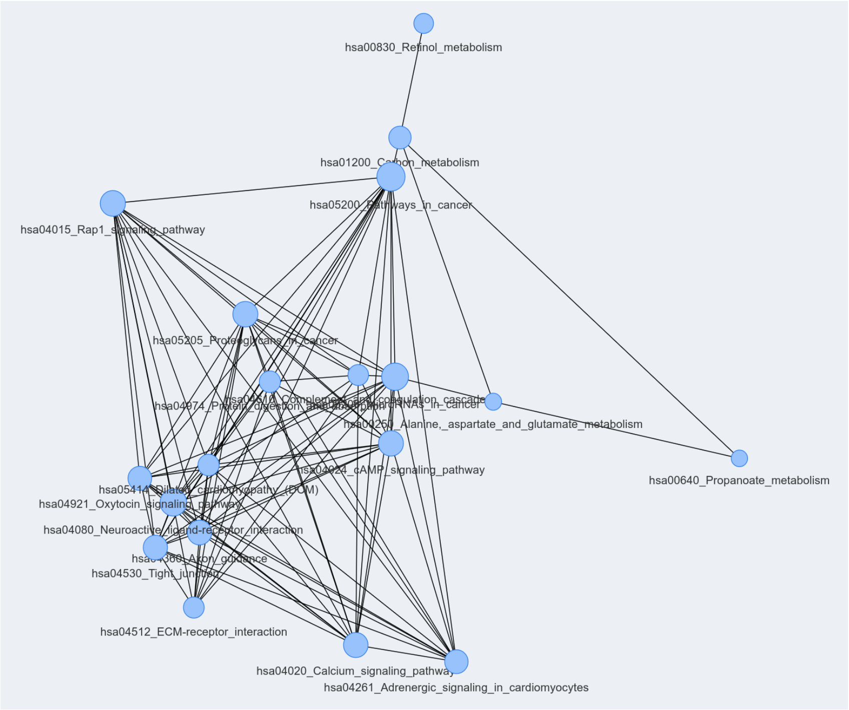

hsa00830_Retinol_metabolism
hsa01200_Carbon_metabolism
hsa05200_Pathways_in_cancer
hsa04015_Rap1_signaling_pathway
hsa05205_Proteoglycans_in_cancer
hsa04610_Complement_and_coagulation_cascades
hsa04974_Protein_digestion_and_absorption
hsa05206_MicroRNAs_in_cancer
hsa00250_Alanine,_aspartate_and_glutamate_metabolism
hsa04024_cAMP_signaling_pathway
hsa00640_Propanoate_metabolism
hsa05414_Dilated_cardiomyopathy_(DCM)
hsa04921_Oxytocin_signaling_pathway
hsa04080_Neuroactive_ligand-receptor_interaction
hsa04360_Axon_guidance
hsa04530_Tight_junction
hsa04512_ECM-receptor_interaction
hsa04020_Calcium_signaling_pathway
hsa04261_Adrenergic_signaling_in_cardiomyocytes

Figure 109 Network visualization of an "Overrepresentation Analysis" of the "ref" data set. It can be seen that ORA directly identifies "Pathways in cancer" and other cancer-related gene sets.

12.9.2.2 Comparison of Methods

Comparisons of the different enrichment methods can be seen in Figures 114, 115 and 116. It is noticeable that GGEA and LEGO always create very similar results. This is a rather surprising result, as LEGO could be seen as a successor of GANPA and therefore should create results more similar to GANPA. However, since GANPA considers nearly all gene sets as significant, it is not possible to make a statement about it.

As expected, both set-based enrichment methods (GSEA and ORA) produce similar results as well, even if not as similar as GGEA and LEGO. SPIA's result are rather independent from single methods; they rather are an intersection of other methods (see Figure 116).

12.9.2.3 Comparison of Patient Enrichments

Figures 117 and 118 show comparisons between all patients using SPIA and LEGO as enrichment methods (please view our interactive web application for more comparisons). Please note, that GANPA was not applicable to single patients, due to the chosen expression analysis model (section 12.9.1.1).

A method is called robust, if it produces similar enrichment results for similar input data. Both heat maps are ordered by the same dendrogram which resulted from fold change clustering as described in section 12.9.1.2. Therefore, a robust method should show high similarity values near the diagonal in these heat maps, since the regions near the diagonal denote comparisons between similar patient samples.

When comparing the enrichments of all single patients, it can be seen that ORA, GSEA and SPIA can produce robust enrichment results. For example, SPIA clearly shows some high-

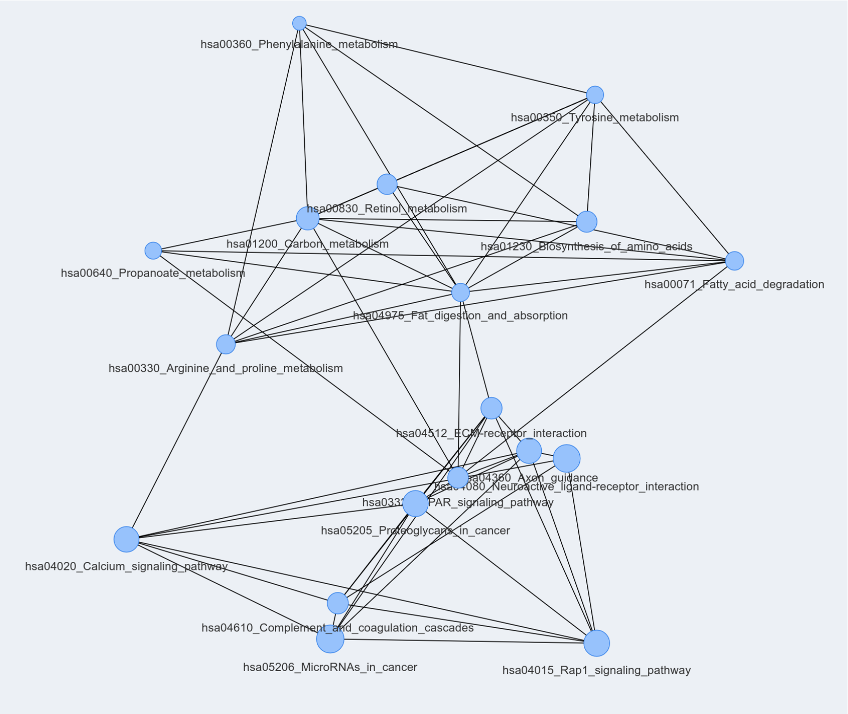

hsa00360_Phenylalanine_metabolism
hsa00350_Tyrosine_metabolism
hsa00830_Retinol_metabolism
hsa01200_Carbon_metabolism
hsa01230_Biosynthesis_of_amino_acids
hsa00640_Propanoate_metabolism
hsa00071_Fatty_acid_degradation
hsa04975_Fat_digestion_and_absorption
hsa00330_Arginine_and_proline_metabolism
hsa04512_ECM-receptor_interaction
hsa04360_Axon_guidance
hsa04080_Neuroactive_ligand-receptor_interaction
hsa03320_PPAR_signaling_pathway
hsa05205_Proteoglycans_in_cancer
hsa04020_Calcium_signaling_pathway
hsa04610_Complement_and_coagulation_cascades
hsa05206_MicroRNAs_in_cancer
hsa04015_Rap1_signaling_pathway

Figure 110 Network visualization of an "Gene Set Enrichment Analysis" of the "ref" data set. While GSEA does not directly dentify "Pathways in cancer" or "Breast cancer", it clearly finds some other cancer-related pathways like "MicroRNAs in cancer" (bottom half). In the top half of the network some other metabolism-related gene sets can be seen.

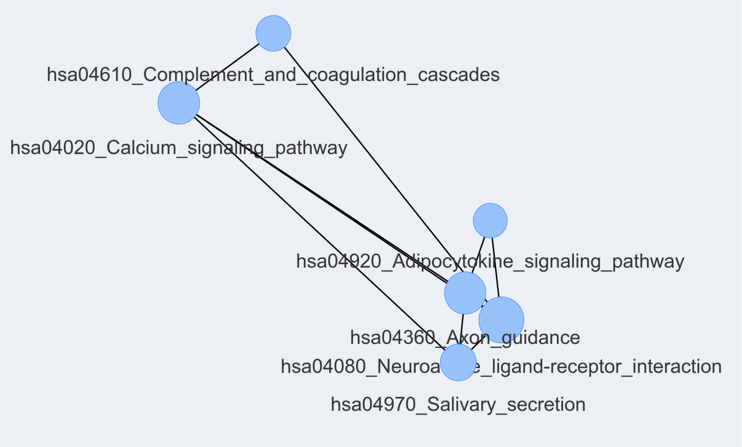

hsa04610_Complement_and_coagulation_cascades
hsa04020_Calcium_signaling_pathway
hsa04920_Adipocytokine_signaling_pathway
hsa04360_Axon_guidance
hsa04080_Neuroactive_ligand-receptor_interaction
hsa04970_Salivary_secretion

Figure 111 Network visualization of an "Signalling Pathway Impact Analysis" of the "ref" data set. None of the detected pathways are indicators for working with tumor samples.

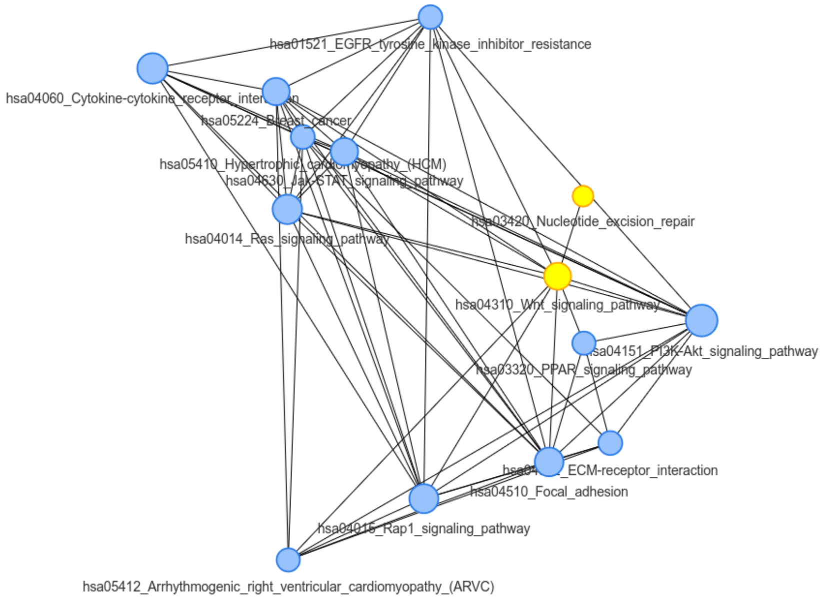

hsa01521_EGFR_tyrosine_kinase_inhibitor_resistance
hsa04060_Cytokine-cytokine_receptor_interaction
hsa05224_Breast_cancer
hsa05410_Hypertrophic_cardiomyopathy_(HCM)
hsa04630_Jak-STAT_signaling_pathway
hsa03420_Nucleotide_excision_repair
hsa04014_Ras_signaling_pathway
hsa04310_Wnt_signaling_pathway
hsa04151_PI3K-Akt_signaling_pathway
hsa03320_PPAR_signaling_pathway
hsa04512_ECM-receptor_interaction
hsa04510_Focal_adhesion
hsa04015_Rap1_signaling_pathway
hsa05412_Arrhythmogenic_right_ventricular_cardiomyopathy_(ARVC)

Figure 112 Network visualization of enrichment analyses of all patients with "GGEA" and "LEGO". Blue nodes are detected by both methods, yellow nodes are only identified by LEGO as significantly enriched. It can be seen that both methods create very similar enrichment results. Also, both methods are able to clearly identify "Breast cancer" as important pathway.

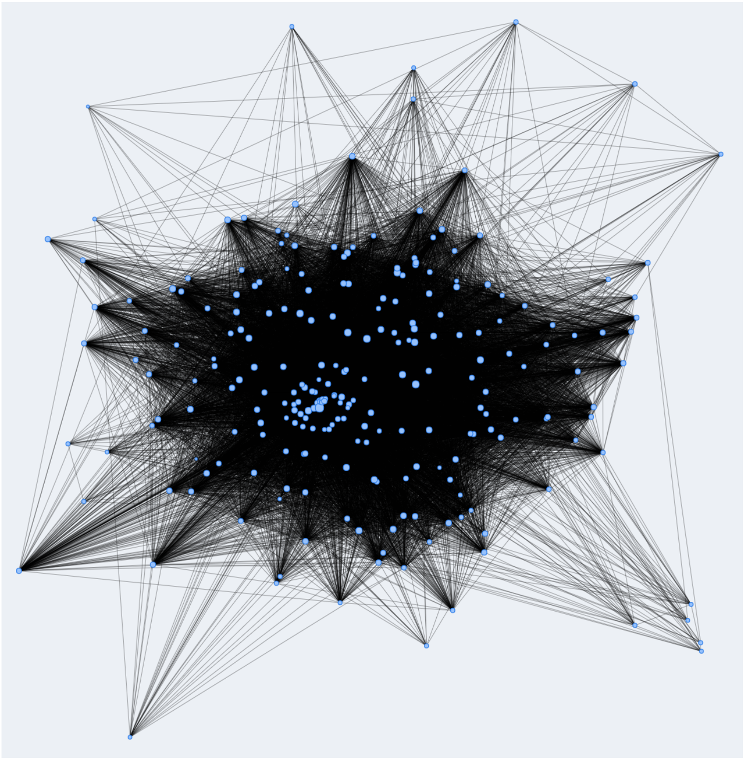


Figure 113 Network visualization of an "Gene Association Network-based Pathway Analysis" of the "ref" data set. GANPA considers 246 of 247 pathways as significantly enriched.

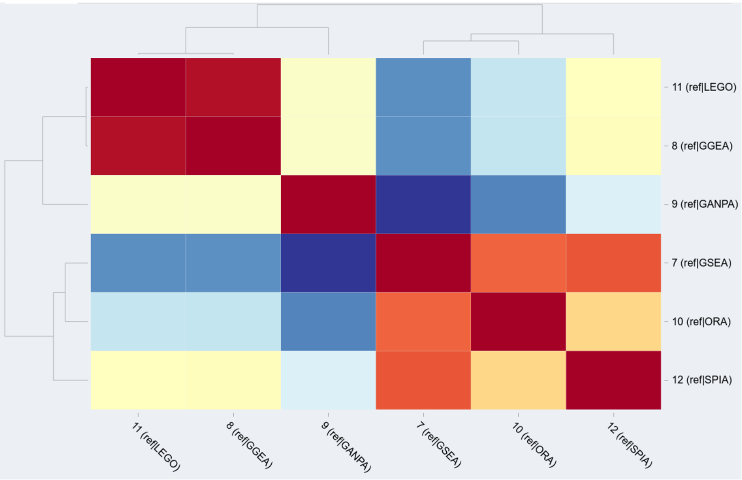


Figure 114 Heat map of rank distances between enrichment analyses of the "ref" data set. The best value of "0" is marked red. It can be seen that GGEA and LEGO are very similar, i.e. have a low rank distance. Also, ORA and GSEA produce similar results. However, in this case GSEA's results are more similar to SPIA than to ORA.

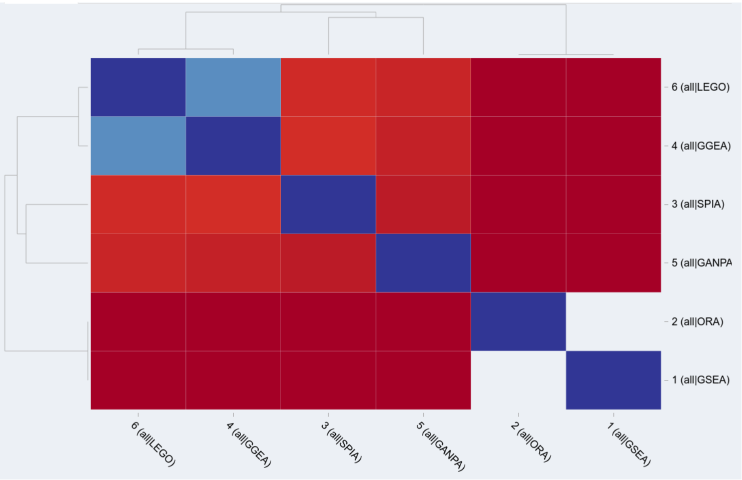


Figure 115 Heat map of overlap distances between enrichment analyses for the "all" data set. The best value of "100%" is marked blue. Since GSEA and ORA could not find significantly enriched gene sets in this data set, they got an overlap distance of "0%" or "NA" (division with 0) to other methods. Again, GGEA and LEGO produce very similar analyses.

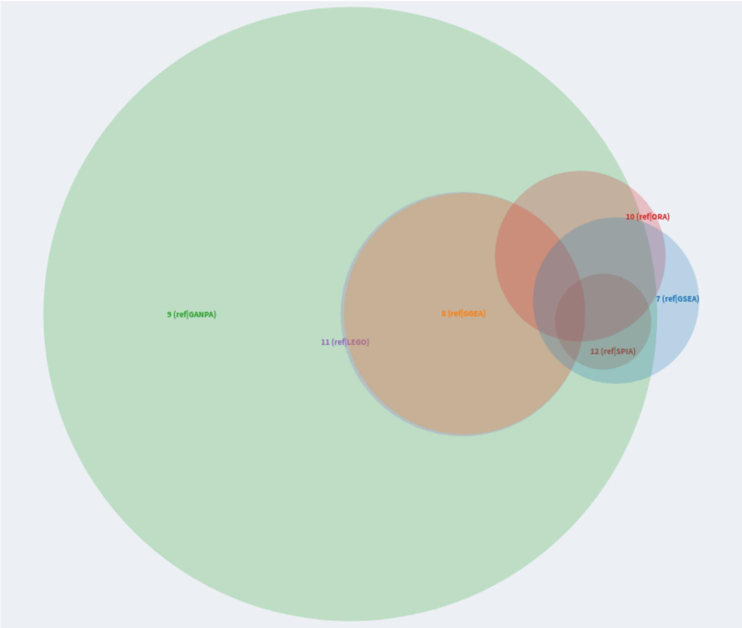


Figure 116 Venn diagram of overlaps between different enrichment analyses of the “ref” data set. It can be seen that GANPA contains the results of nearly all other methods. GGEA and LEGO produce pretty the same results, while ORA and GSEA have a big overlap. SPIA reports a subset of multiple other methods.

similarity clusters near the diagonal (Figure 117). Only LEGO and GGEA have less such hot spots. Their hot spots are near the hot spots of other methods like e.g. SPIA (figure 118). However, these hot spots cover patients of different sub trees, i.e. the similarity of the patients is less important to the similarity of the enrichment results of GGEA and LEGO.

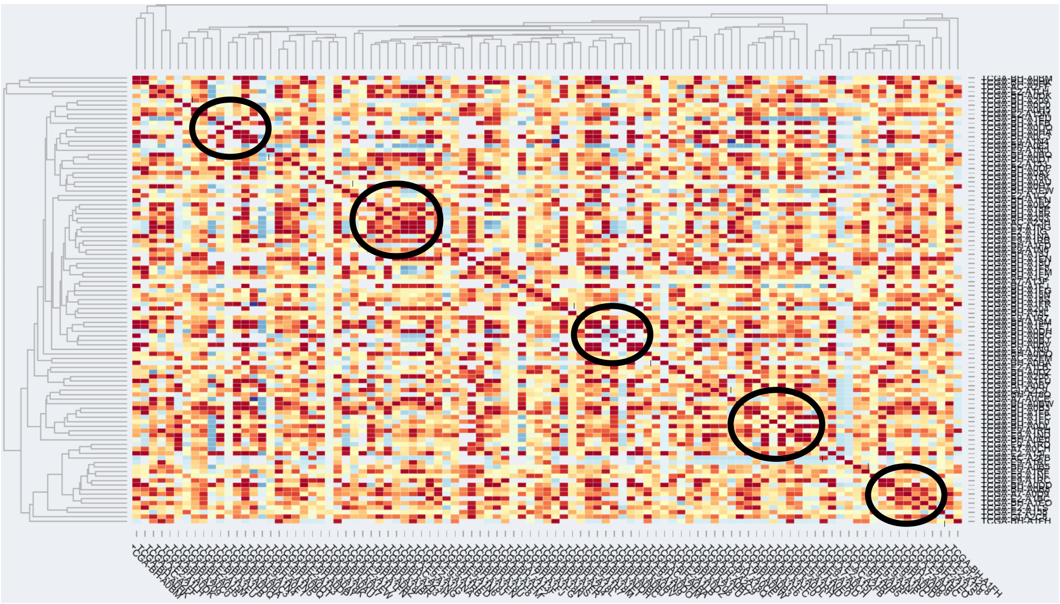


Figure 117 Heat map of rank distances between all patients, using SPIA as enrichment method. The rows and columns are ordered by the fold change cluster dendrogram (see section 12.9.1.2). Black circles mark some high-similarity clusters near the diagonal.

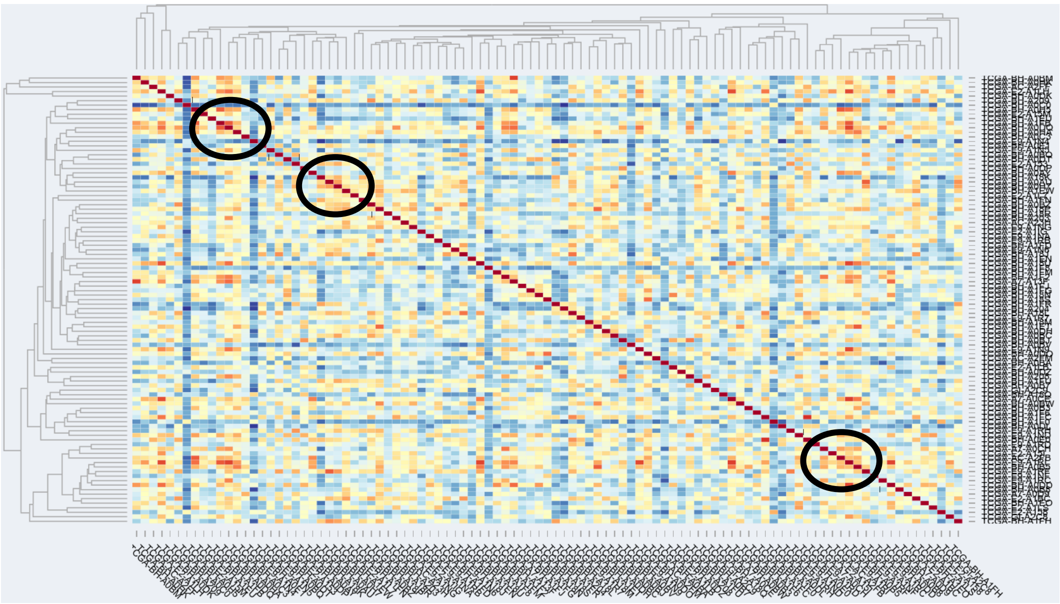


Figure 118 Heat map of rank distances between all patients, using LEGO as enrichment method. The rows and columns are ordered by the fold change cluster dendrogram (see section 12.9.1.2). Black circles mark some high-similarity clusters near the diagonal.

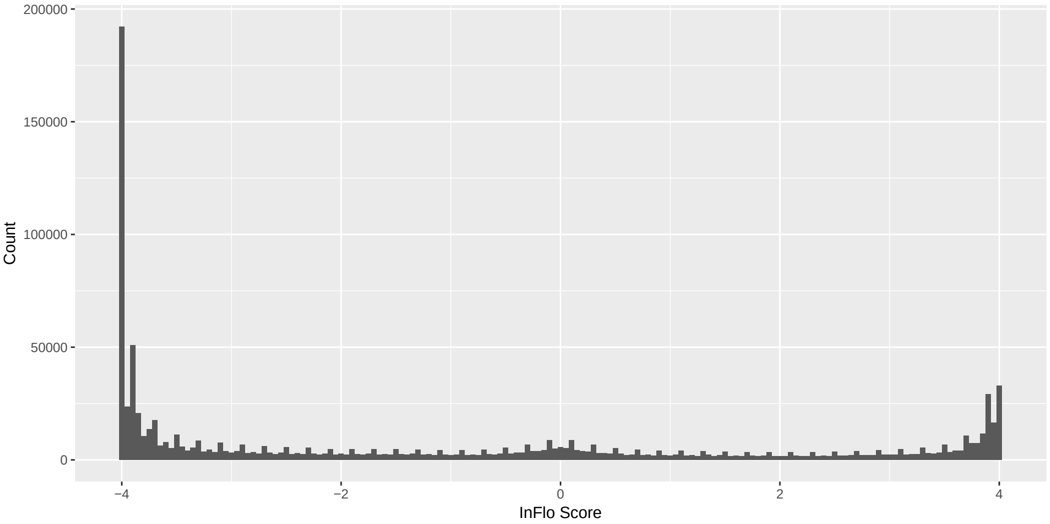


Figure 119 Distribution of scores assigned by InFlo. There is a clearly visible concentration of scores at the extrema of the spectrum with only a few values in-between. A minor accumulation can be observed around score 0. There are periodical spikes in score frequency which can be attributed to the algorithm InFlo uses to assign final scores.

12.9.3 Conclusion

Almost all investigated methods can create subjectively meaningful results. Only GANPA did not create any sensitive results as it classified nearly all gene sets as significantly enriched. Since running GANPA with different networks and gene sets did not influence this "overreaction" it seems to be caused by a methodic flaw or an error in GANPA's implementation.

Comparing the other methods the set-based enrichment methods GSEA and ORA could create more meaningful results for the six-patients data set than the network-based methods. However, GGEA and LEGO (both network-based enrichment methods) were the only methods which could detect "Breast Cancer" as significant pathway by analysing all patients, while GSEA and ORA could not produce any significant results for this data set. Therefore, incorporating additional information from biological networks can be beneficial.

Furthermore, all methods except GANPA (as it was not investigated here), GGEA and LEGO were able to produce robust results when provided with similar input data.

Please also try our interactive result visualizations at <PLEASE ADD LINK!!!>!

12.10 Integration of Subnetwork Search Methods

12.10.1 InFlo Analysis

To assess the applicability of InFlo, we performed an analysis on our chosen dataset. InFlo is available with a pre-compiled set of pathways from NCI-PID and Biocarta. Although N. Dimitrova *et al.* mention using KEGG pathways for cross-validation, these data were not included in the example data. Therefore, we had to reverse-engineer the format used for the representation of the pathways. As it turned out, InFlo expects networks to be supplied twice in two slightly differing formats. After converting all available KEGG pathways as faithfully as possible into the new format, we applied InFlo to our dataset.

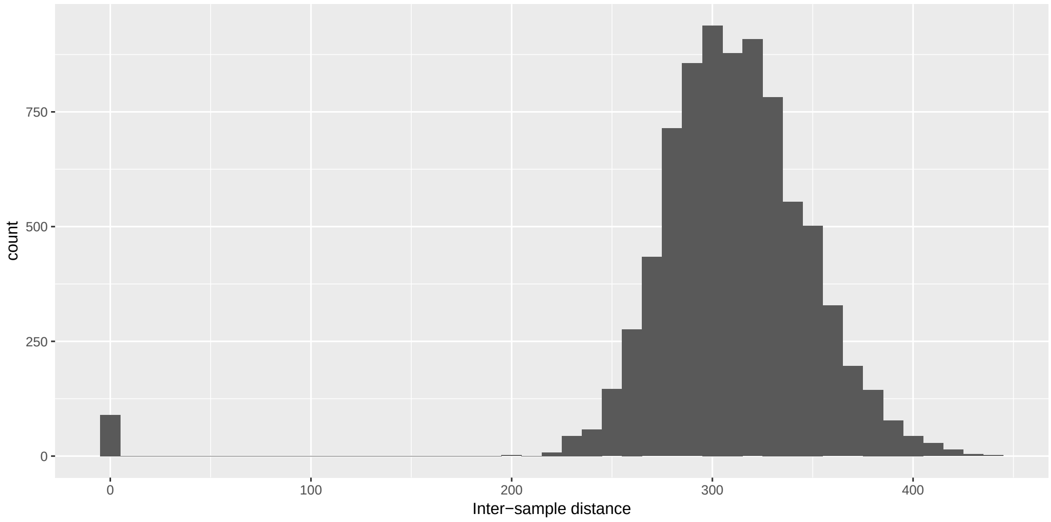


Figure 120 Distribution of inter-sample distances. An inter-sample distance is the euclidean distance between two score vectors assigned by InFlo, one vector representing one patient. One can clearly see there are no two samples with a distance less than 200. The accumulation of distances at point 0 stems from samples being compared to themselves.

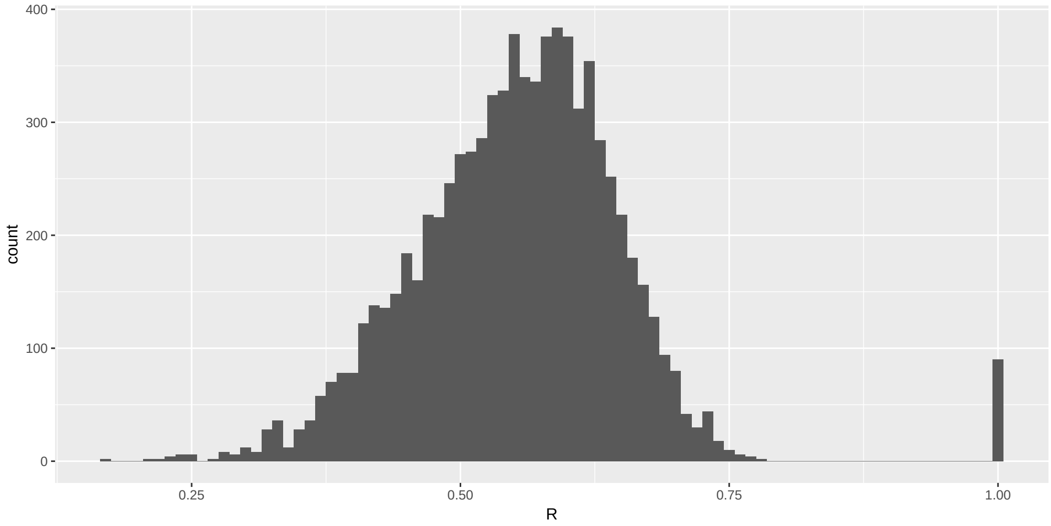


Figure 121 Distribution of sample correlations. There are no pairs of samples with a correlation above 0.8. The mean correlation coefficient is ca. 0.554. The high bar at value 1.00 stems from samples being correlated with themselves.

Although InFlo provides results on a continuous scale, most of the results are on either end (figure 119). That is, if the edges are scored at all. Out of 11103 supplied edges, only 4868 received scores in at least one sample. Additionally, InFlo supplied scores for gene expression, i.e. loop edges. The number of these newly generated edges is 7471. Therefore, we could evaluate only a fraction of the data we provided.

For a first, simple analysis, we had a look at the inter-sample distances as mentioned by N. Dimitrova *et al.* Inter-sample distances are calculated by treating the score vector returned by InFlo for each patient as a position vector. Consequently, one can compute the distance between two samples by calculating the distance between the two points representing them. We tried using the Euclidean distance as measure for our analysis but this resulted in all samples being seemingly equally spread apart in space (figure 120). Because of the high dimensionality of InFlo's results, the Euclidean distance and all other Minkowski distances become practically useless for further analysis. We examined the pairwise correlations of the score vectors but this analysis yielded no significant results with high correlations (figure 121). The large diversion of our results might stem from InFlo's tendency to score edges at the ends of the used scale, making a correlation analysis difficult. Additionally, InFlo's omission of the majority of edges even though there were experimental measurements for sources and targets of these edges leads to less expressive results. Why these edges were not scored is unknown to us as the scoring itself is implemented in C++. This module's source code was not available in the GitHub repository, only a pre-compiled binary was available for download.

12.10.2 Analysis

We applied RelExplain to our data by first calculating log fold changes for each sample and gene. The KEGG pathways were present in RelExplain's database. Afterwards, it was invoked for every patient.

The resulting Steiner Trees had orders ranging from 11 to 26 and sizes between 10 and 26. Because we called RelExplain using edges from all sources contained in its database, which may or may not be contained in KEGG, the overlap between the Steiner Tree and the pathways is at most five edges per sample. An analysis including only those edges contained in KEGG was started but did not complete in time.

For every sample, the Steiner Tree can be viewed in our Shiny app.

12.10.3 RACER Analysis

We applied RACER to our expression, methylation and CNV data. Additionally, we downloaded miRNA expression data from the GDC data portal for samples which had this data available to supply all values to the linear model employed by RACER. Consequently, we had to reduce the data set for RACER analysis to 81 patients.

Because RACER uses transcription factor binding signals measured in the K562 cell line for the estimation of transcription factor activity, it could only estimate the coefficients of the transcription factors whose signals were measured in this cell line of which there were 97. RACER estimates miRNA activities as well but the number of miRNAs which we had expression data and target motifs for was 6. Therefore we saw no value in analysing the results further.

The condition-specific transcription factor-gene interactions were more plentiful as the number of known transcription factor binding sites and transcription factors was high. Therefore, there were a lot of data to be examined. About half of the predicted interactions are in-

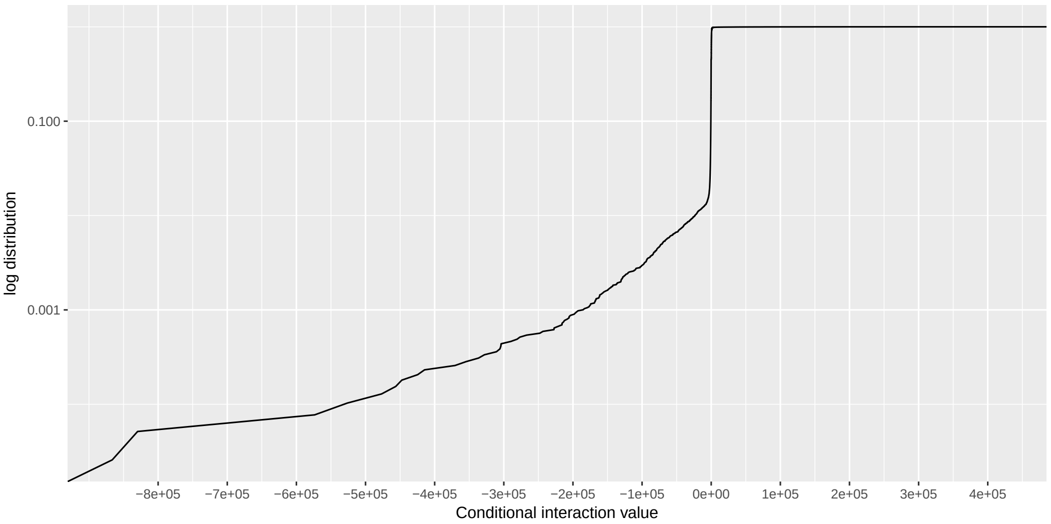


Figure 122 The logarithmized empirical cumulative distribution of the RACER values representing condition-specific interactions between transcription factors and genes. Their range is -866,375.3 to 420,659.9. Most of them lie in the neighbourhood of zero, meaning there is no interaction. About half of the observed values are negative which suggests many of the transcription factors show an inhibiting activity.

hibiting (figure 122). When examining edges representing a TF-gene interaction which were scored by InFlo, the results agree at least qualitatively. A quantitative comparison was deemed not appropriate because of the distribution of InFlo scores.

12.10.4 Comparison of InFlo and RelExplain

We wanted to compare InFlo's and RelExplain's results because of their different methodology. Because they use different models and therefore give differing outputs it was an interesting question to us whether the results would be largely the same or wildly different. Our first analysis was concerned with the distribution of InFlo scores of edges contained in RelExplain's Steiner Tree and those who were not included. We compared their empirical density functions and found that all edges in the Steiner Tree which were also scored by InFlo received a score at the lower end of the spectrum (figure 123).

12.10.5 Integrated Graphical View

We implemented a comprehensive graphical representation based on Cytoscape.js [148] of our results for the readers to browse and examine. It allows the selection of a patient and a KEGG pathway for exploration. It displays the selected pathway annotated with experimental data measured in their samples (figure 124). One datum is highlighted and therefore displayed as the nodes' colour. In the case of proteins this is the expression fold change observed for the respective coding gene, for transcription factors it is the fold change of activity observed in the associated samples.
Our graphical view allows simple actions one expects from a graph viewer such as zooming, panning and drag-and-drop functionality. Furthermore, it uses a layer system making it possible to show and hide the results from InFlo, RelExplain and RACER independently. These results can be displayed by clicking on a node or edge(figure 125). It also allows

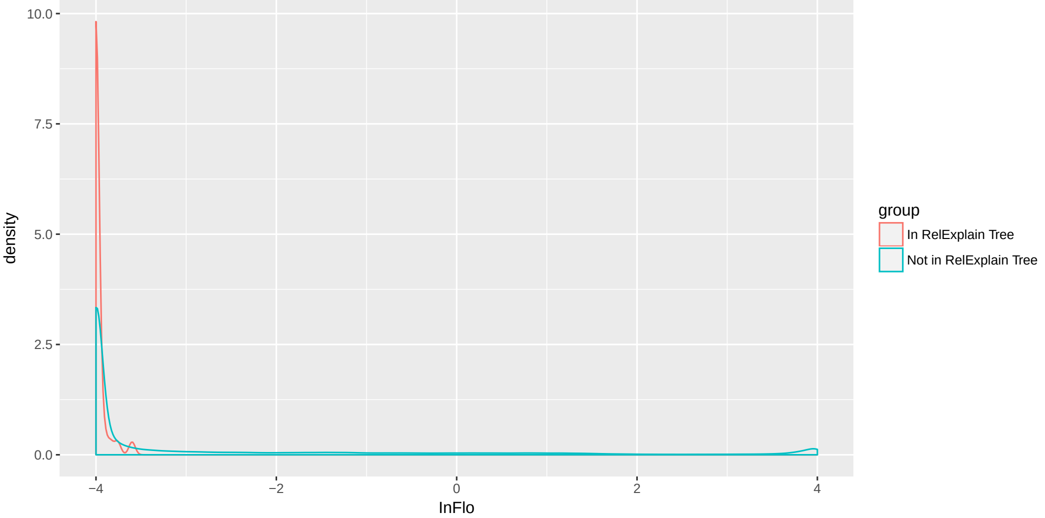
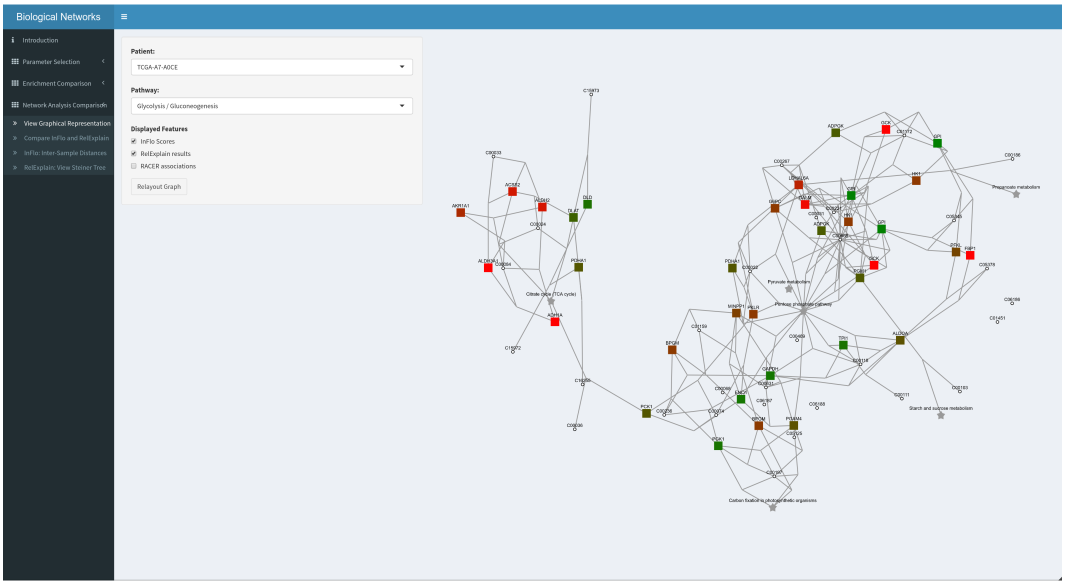


Figure 123 Comparison of InFlo score densities by classification from RelExplain. One can clearly see the spike on the left. Essentially, all edges in RelExplain's Steiner tree received a very low InFlo score if they were scored. An interactive version is available in our Shiny app.

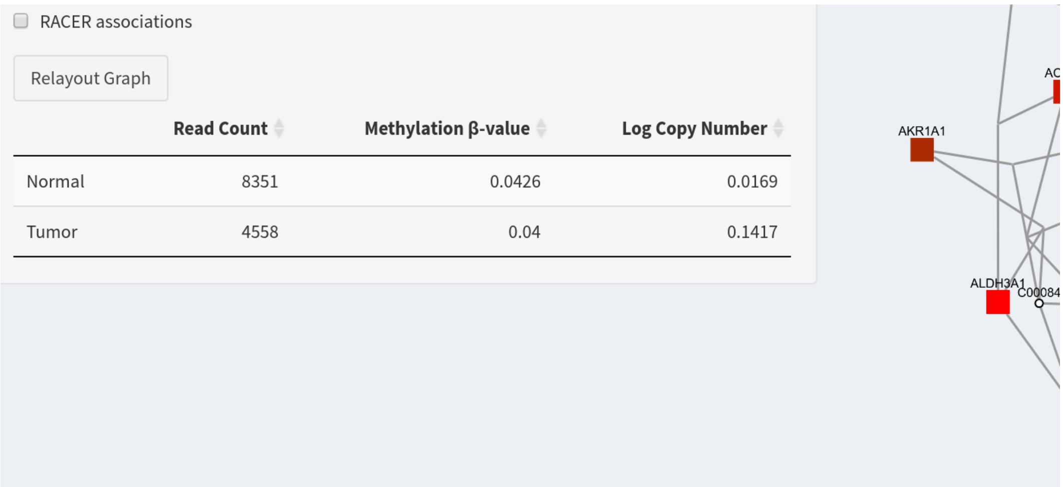


Patient: TCGA-A7-A0CE

Pathway: Glycolysis / Gluconeogenesis

Displayed Features
☑ InFlo Scores
☑ RelExplain results
☐ RACER associations

Figure 124 Our graph viewer showing the pathway "Glycolysis/Gluconeogenesis" with measurements from patient TCGA-A7-AOCE. On the left side one can select the patient and pathway as well as the layers to be shown. Currently, RACER results are hidden. One may also relayout the graph if the resulting layout does not satisfy the user.



☐ RACER associations

Relayout Graph

	Read Count	Methylation β-value	Log Copy Number
Normal	8351	0.0426	0.0169
Tumor	4558	0.04	0.1417

Figure 125 By clicking on a node or edge the user can display all measured values for that element.

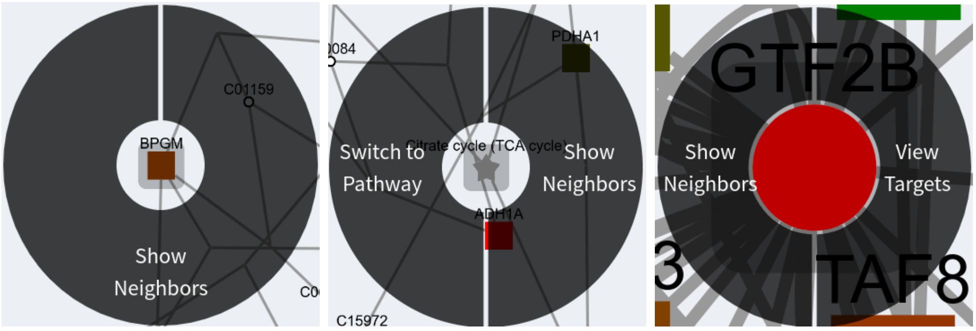
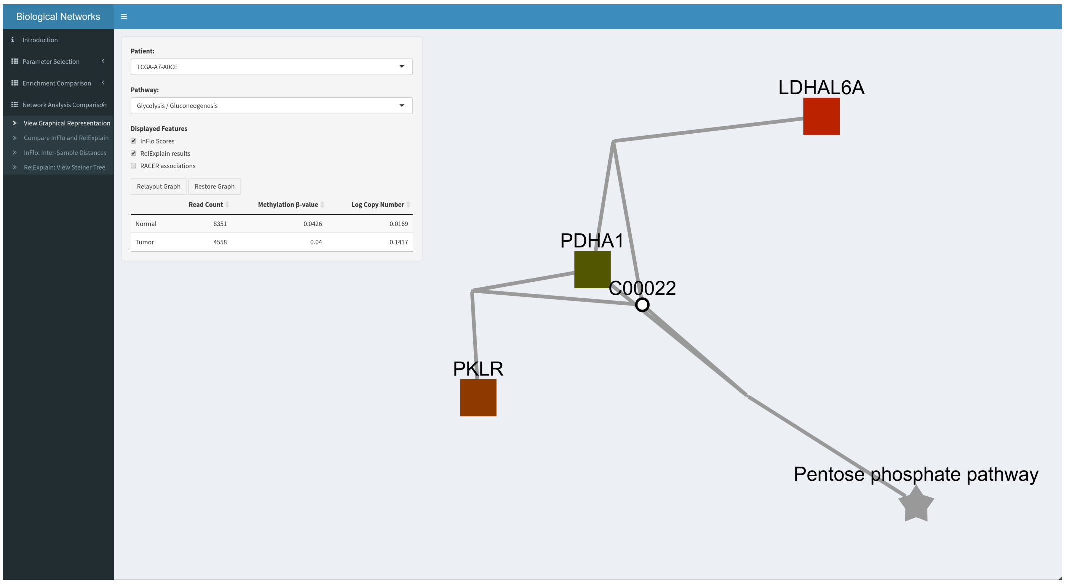


Figure 126 Available context actions in the graph view. When right-clicking on any node it is possible to reduce the shown nodes to its neighbourhood. When right-clicking on an associated pathway the user can switch to that pathway's view (middle). When right-clicking on a transcription factor one can view the top 5% targets of that transcription factor including their measurements (right).


Biological Networks
≡
Introduction
Parameter Selection
Enrichment Comparison
Network Analysis Comparison
View Graphical Representation
Compare InFlo and RelExplain
InFlo: Inter-Sample Distances
RelExplain: View Steiner Tree
Relayout Graph
Restore Graph
LDHAL6A
PDHA1
C00022
PKLR
Pentose phosphate pathway

Patient: TCGA-A7-A0CE

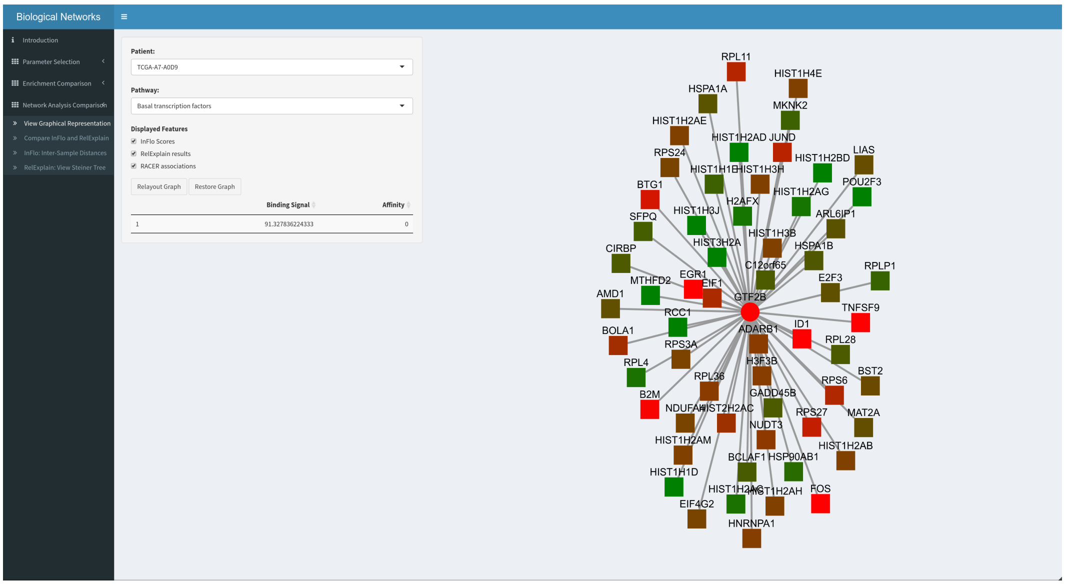
Pathway: Glycolysis / Gluconeogenesis

Displayed Features

☑ InFlo Scores
☑ RelExplain results
☐ RACER associations

	Read Count	Methylation β-value	Log Copy Number
Normal	8351	0.0426	0.0169
Tumor	4558	0.04	0.1417

Figure 127 Neighbourhood view of a gene. It is possible to restore the current graph by clicking the button labelled “Restore Graph”.



Patient: TCGA-A7-A0D9

Pathway: Basal transcription factors

Displayed Features

☑ InFlo Scores
☑ RelExplain results
☑ RACER associations

	Binding Signal	Affinity
1	91.327836224333	0

Figure 128 Target-view of a transcription factor. The edge to `TNSF9` is currently selected and the results from RACER are shown to the left. The user an return to the previous graph by clicking “Restore Graph”.

for a range of context-sensitive operations (figures 126 to 128) enabling a comprehensive view and allowing the user to explore possible alternative explanations. Additionally, the interface can easily be extended to use additional measurements or a new dataset. While it is not possible to upload data through the Shiny interface, the app comes with all functions needed to integrate experimental data supplied as Bioconductor's ExpressionSets.

12.10.6 Conclusion

In conclusion, the tools we used to analyse our dataset falling into the subnetwork search category—namely InFlo and RelExplain—were not able to provide an enhanced result over the enrichment methods. Especially InFlo did not meet our expectations as it only evaluated a small fraction of our data and did not provide a well-scaled result. When using RelExplain we encountered several problems during the execution because of external influence which would not be easy to fix even if we could find a solution at all. These problems were amplified by its closed nature which did not help in the process as it made the resolution process require more coordination. RACER provided some hints for further analysis, many of which were already verified by the KEGG pathway definitions—the interactions with especially extreme values were already present as edges in our KEGG pathways.

References

[91] L M Gallego-Paez et al. "Alternative splicing: the pledge, the turn, and the prestige : The key role of alternative splicing in human biological systems." In: *Human genetics* (2017). ISSN: 1432-1203. DOI: 10.1007/s00439-017-1790-y. URL: http://link.springer.com/10.1007/s00439-017-1790-yhttp://www.ncbi.nlm.nih.gov/pubmed/28374191.

[92] Suzanne Clancy. "RNA Splicing: Introns, Exons and Spliceosome". In: *Nature Education* (2008).

[93] Alberto R. Kornblihtt et al. "Alternative splicing: a pivotal step between eukaryotic transcription and translation". In: *Nature Reviews Molecular Cell Biology* 14.5 (2013), pp. 306–306. ISSN: 1471-0072. DOI: 10.1038/nrm3560. URL: http://www.nature.com/doifinder/10.1038/nrm3560.

[94] R E Breitbart, A Andreadis, and B Nadal-Ginard. "Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes." In: *Annual review of biochemistry* 56.1 (1987), pp. 467–95. ISSN: 0066-4154. DOI: 10.1146/annurev.bi.56.070187.002343. URL: http://www.ncbi.nlm.nih.gov/pubmed/3304142http://www.annualreviews.org/doi/10.1146/annurev.bi.56.070187.002343http://www.annualreviews.org/doi/abs/10.1146/annurev.bi.56.070187.002343.

[95] Douglas L. Black. "Mechanisms of Alternative Pre-Messenger RNA Splicing". In: *Annual Review of Biochemistry* 72.1 (2003), pp. 291–336. ISSN: 0066-4154. DOI: 10.1146/annurev.biochem.72.121801.161720. URL: http://www.ncbi.nlm.nih.gov/pubmed/12626338http://www.annualreviews.org/doi/10.1146/annurev.biochem.72.121801.161720.

[96] A. Gregory Matera and Zefeng Wang. "A day in the life of the spliceosome". In: *Nature Reviews Molecular Cell Biology* 15.2 (2014), pp. 108–121. ISSN: 1471-0072. DOI: 10.1038/nrm3742. URL: http://www.ncbi.nlm.nih.gov/pubmed/24452469http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4060434http://www.nature.com/doifinder/10.1038/nrm3742.

[97] Arianne J. Matlin, Francis Clark, and Christopher W. J. Smith. “Understanding alternative splicing: towards a cellular code”. In: *Nature Reviews Molecular Cell Biology* 6.5 (2005), pp. 386–398. ISSN: 1471-0072. DOI: 10.1038/nrm1645. URL: http://www.ncbi.nlm.nih.gov/pubmed/15956978http://www.nature.com/doifinder/10.1038/nrm1645.

[98] Barmak Modrek and Christopher Lee. “A genomic view of alternative splicing”. In: *Nature Genetics* 30.1 (2002), pp. 13–19. ISSN: 10614036. DOI: 10.1038/ng0102-13. URL: http://www.nature.com/doifinder/10.1038/ng0102-13.

13 (long) non-coding RNAs

by Johannes Rest and Alexander Grün

13.1 Introduction

Only about 2% of mammalian genomes are reported to be protein-coding regions. While those regions received great attention from the beginning, the majority of the genome was referred to as “junk DNA” for a long time. However, as we know today, more than 70% of the mammalian genome is transcribed. Those regions give rise to various classes of noncoding RNAs (ncRNA), which can be divided into two major groups: small ncRNAs with less than 200 nucleotides (such as microRNA or Piwi-interacting RNAs) and so called long ncRNA (lncRNA) [149, 150].

Although lncRNAs are loosely defined as RNA transcripts longer than 200 nucleotides that are not translated into proteins, the basic features of lncRNAs can be comparable to messenger RNAs (mRNAs). LncRNAs undergo alternative splicing, just as mRNAs, but are smaller and have fewer exons on average. The majority of them is also transcribed by RNA polymerase II und usually contain canonical polyadenylation signals, although some lncRNA are likely to be transcribed by polymerase III.

Unlike protein-coding genes, which are generally conserved across different species, most lncRNA sequences are poorly conserved. This lack of conservation makes predictions of lncRNA functions based on sequence analysis alone very difficult. However, the promoters of lncRNAs appear to be much better conserved, and even comparable to mRNA promoters.

LncRNAs form secondary and tertiary structures to serve their functions. This is a possible explanation for the weak sequence conservation of lncRNAs. While protein-coding genes are prone to frameshift mutations, lncRNAs lack an open reading frame (ORF). In consequence, lncRNAs have less pressure to retain their sequence in order to form a functional structure. This led to the suggestion, that the structure of lncRNAs is the main functional component and therefore highly conserved in comparison to their primary sequence [149, 151]. Furthermore lncRNAs are expressed at much lower rates than mRNAs. Additionally to their low expression rates their expression is highly tissue specific. Those two features combined make the discovery of lncRNAs challenging. However, todays sequencing technologies are capable of detecting lncRNA expression as well as discovering novel lncRNA transcripts, RNA-seq being the most widely used [152].

LncRNAs have received great attention in recent studies as they have been shown to carry out important biologial functions regarding gene regulation. Although plenty of lncRNA genes have been discovered, most of them lack functional annotation and the underlying mechanisms remain mostly unclear. Also, several lncRNAs have been proven to play a crucial role in different cancer types [153].

13.2 LncRNA classifcation

Although the importance of long non-coding RNAs (lncRNAs) in many biological processes is becoming increasingly clear and their number is steadily growing, their classification can be rather challenging. Up to a certain point this is due to a lack of a fundamental conceptual classifcation framework, which leads to a conflicting, confusing and overlapping terminology.

Here, we give an overview over different possibilities to characterize lncRNA and describe some conceptual guidelines that have emerged from the ever growing field of the non-coding transcriptome.

LncRNAs can be classified according to various features. One of the most used classification is based on their association with annotated protein-coding genes. This also serves as the foundation for the classification used by GENCODE, which defines five different biotypes: sense-overlapping, antisense, bidirectional, intronic and intergenic. Sense-overlapping represents a class of lncRNAs that overlap with the exons of a PCG on the same strand, while overlapping RNAs on the opposite strand are classified as anti-sense. An interesting class of lncRNAs is located head to head with a PCG on the opposite strand, so called bidirectional lncRNAs. Intronic lncRNAs are located either exclusively in the introns of a PCG, called totally intronic RNAs (TINs), or are only partially located in introns. Lastly, the term intergenic refers to lncRNAs that are not associated with PCGs regarding their genomic position, but lie in between coding regions [154, 155].

Another possible classification is simply based on the transcript length. We typically refer to a lncRNA only from 200 bp upwards, everything below this threshold being refered to as short ncRNA. However, some lncRNAs are much longer, especially in intergenic regions where they can span up to 1MB, so called very long intergenic non-coding RNAs (vlincRNAs) [154].

Enhancer- and promoter association represents yet another possibitlity to classify lncRNAs. This class represents a particular interesting group, since they are involved in linking the dynamics of nuclear architecture, chromatin signalling plasticity, and transcriptional regulation.

Regarding the fact that lncRNAs annotated by GENCODE are primarily spliced transcripts, one can classify lncRNA based on their mRNA resemblance. This helps to identify additional long RNAs that may play important roles in many biological processes.

Further possible classifcations are based on the association of lncRNAs with biochemical pathways, their stability, sequence or structure conservation, their subcellular location and lastly their function.

As shown, the list of possible lncRNA classifcations is huge. In order to provide a basis for further lncRNA classifications and to avoid problamatic overlapping terminologies, a consolidated conceptual framework is needed. St Laurent et al proposed a possible framework, which can be divided into four different tiers [154].

Tier 1 refers to the mapping of the longest unprocessed transcript. This would result in standalone ncRNA loci which allows for consolidation of disparate and often incomplete ncRNAs. Since the distances that separate those elements can be rather big, this will clearly help to get a bigger picture of the whole, e.g. for understanding the underlying regulating mechanism associated with the locus. It also would allow for experiments to focus on the locus instead of many distinct non-coding elements. Further the association with certain gemomic features would be much clearer. In the case of enhancers, for example, one could easily distinguish whether the transcript originates from that location or if it simply overlaps the region. Basically it would combine all distinct annotations of lncRNAs into gene-like structure with their own transcription regulatory regions. An exception to this are lncRNAs that are not produced from a promoter, as it is the case for circular intronic RNA.

In the next step (tier 2) processed transcripts have to be defined. Mapping sites of polyadenilation can provide additional information, as well as some other highly-sensitive methods

targeted to specific regions.
Since lncRNA expression is highly tissue specific, genomic coordinates alone are not sufficient. Thus, we need to add an additional dimension of expression to our annotation (tier 3). This allows for improved function prediction and may be especially important regarding the lncRNAs potentially big role in various cancers.
The last tier (tier 4) refers to a map of all RNA modifications (which may be way more than 100) and serves as a very informative source for classification as it allows for easy distinguishing between RNA molecules. However, a complete genome-wide mapping of those RNA modifications still faces some technical limitations [154].

13.3 LncRNAs in cancer

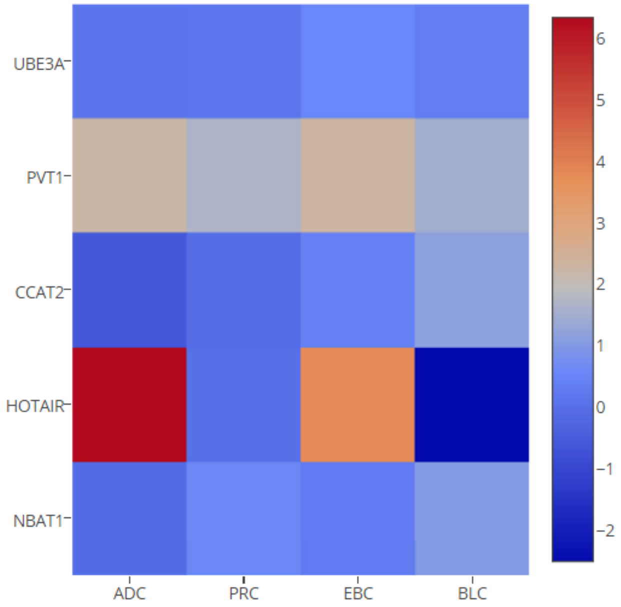
It is getting increasingly clear that lncRNAs play a significant role in the development and progression of various cancers. Expression analysis revealed that a lot of non-coding RNA is either highly up- or down-regulated in the cancer tissues compared to adjacent normal ones. Furthermore, many of the genomic mutations in cancer are located inside regions that do not encode proteins but are often transcribed into lncRNAs. Although only a small fraction of lncRNAs have been functionally characterized, several of them have been linked to the transformation of healthy cells into tumor cells. Possessing key roles in gene regulation, they affect various aspects of cellular homeostasis while exhibiting a huge diversity in strategies [153, 156, 157, 158].

In the next section, we highlight some lncRNAs that were recently shown to be associated with tumor progression and poor patient outcomes and give a quick impression of the underlying mechanisms. Figure 129 shows the differential expression of the mentioned lncRNAs in four different cancer types: lung adenocarcinoma (ADC), bladder cancer (BLC), ER+ breast cancer (EBC) and prostate cancer (PRC).

13.3.1 HOTAIR and CCAT2 are up-regulated in ovarian cancer

HOTAIR has been considered as a pro-oncogene in multiple cancers, including ovarian cancer. The PI3K/AKT/mTOR pathway is suggested to be frequently activated in ovarian cancer tissues and plays a crucial role in the malignant transformation of human tumors as well as tumor growth, proliferation and metastasis. Recently, the lncRNA HOTAIR was shown to be a driving factor in this pathway by regulating PIK3R3, which is regarded as potential therapeutic target of ovarian cancer.
The silencing of HOTAIR resulted in a down-regulation of the protein and mRNA level of PIK3R3, while silencing PIK3R3 down-regulated the expression of HOTAIR, suggesting an interaction between the two. Furthermore, when silencing HOTAIR or PIK3R3, the expression of miR-214 and miR-217 was increased, which are also known oncogenes. This suggests that the interaction of PIK3R3 and HOTAIR is mediated by miR-214 and miR-217. Interestingly, the proliferation, migration and invasion of the tumor cells was inhibited when HOTAIR or PIK3R3 was silenced. Thus, down-regulating HOTAIR can inhibit the malignant behavior of cancer, making it a potential therapeutic target [156].
CCAT2 is another example of an up-regulated lncRNA in ovarian cancer. While it was known to play a crucial role in promoting tumor metastasis, growth and chromosomal instability in colon, lung, breast and gastric cancers, until recently the role of CCAT2 in ovarian cancer was unclear. It could be shown, that its gene expression is elevated in ovarian cancer tissues and that high levels of CCAT2 are correlated with poor prognosis in patients. Consistently

Figure 129 Differential expression of the cancer associated lncRNAs in lung adenocarcinoma (ADC), bladder cancer (BLC), ER+ breast cancer (EBC) and prostate cancer (PRC). The reported fold change (log2(FC)) is based on the work presented in the 'Co-expression network analysis' chapter. Red denotes upregulation and blue downregulation of the lncRNA.



with these observations, the knockdown of CCAT2 inhibits cell proliferation, migration and invasion [157].

13.3.2 NBAT1 is down-regulated in various cancers

The lncRNA NBAT1 is involved in many different cancers. Loss of NBAT1 promotes proliferation leads to an impairment of differentiation of neuronal precursors. Its role in breast cancer and ovarian cancer was recently investigated [158, 159].

As observed in other types of cancer, NBAT1 was significantly down-regulated in breast cancer. Furthermore, low expression of NBAT1 was correlated with lymph node metastasis, migration, invasion and overal poorer patient outcomes. Higher expression of NBAT1 inhibits migration and invasion of breast cancer cells. Overexpressing NBAT1 leads to changes in the global gene expression profile. Amongst the genes with the biggest fold change with overexpressed NBAT1 is DKK1, which serves as an inhibitor of the WNT signaling pathway. Previous studies demonstrated that DKK1 could inhibit cell migration and invasion in breast cancer.

NBAT1 is also known to regulate gene expression by modulating the functions of PRC2. EZH2 is a catalytic subunit of PRC2 and is often overexpressed in breast cancer. The effect of NBAT1 on cell migration and invasion is considered to be mediated through EZH2, as the effect was reversed after adding concomitant EZH2 inhibitors. However, NBAT1 is not the only lncRNA that interacts with PRC2. Xist and HOTAIR have been proven to promote EZH2 and PRC2 functions, while NBAT1 represses them. Similar observations could be made in ovarian cancer. In conclusion, recent studies suggest that NBAT1 acts as an anti-oncogene in the development of various cancers.

13.3.3 Clinical implications

Since an increasing number of lncRNAs have been proven to be involved in the transformation and maintenance of cancer phenotypes, they represent an important new field for the diagnostics and treatment of cancer. Their high tissue- and cell type-specific expression patterns suggest lncRNAs as promising new biomarkers that could accurately classify cancer subtypes. However, not every lncRNA is suitable for this application, as they should be stable and easily detectable in body fluids. PCA3, for example, represents a very promising new biomarker for prostate cancer. It can be detected in patient urine samples and is a more specific and sensitive marker than the currently widely used prostate-specific antigen (PSA) [153].

Beyond their application as biomarkers, lncRNAs are promising new drug targets. Their nature as long RNA molecules allows for the development of highly specific oligonucleotide antagonists. This principle has been succesfully applied to the Angelman syndrome in mice. The syndrome is caused by a lack of expression of the imprinted ubiquitin protein ligase E3a gene (Ube3a), which is repressed by its lncRNA antisense transcript Ube3a-ats. By specifically targeting this lncRNA and thus inhibiting it, Angelman syndrome was successfully cured [153, 160].

Studies like these encourage further exploration of therapeutic strategies and highlight the urgent need to functionally categorize lncRNAs.

13.4 Coexpression network analysis

As mentioned earlier, functional characterization of lncRNAs remains a great challenge. While majority of the protein-coding genes have functional annotation, this is not the case for lncRNAs. While reliable experimental methods for exploring functions of lncRNA exist, e.g. through gene knockdown, those methods have only low throughput. Additionally, they require some prior knowledge about the potential mechanisms to explore. Another problem arises from the nature of lncRNAs itself. Because of their low sequence conservation, predicting functions of lncRNA through homology modeling is rather difficult. This approach is even more hindered by the lack of functional annotations among lncRNAs. With an increasing number of functionally characterized lncRNAs this may be an option in the future. However, it was suggested that the structure of lncRNAs, despite their low sequence conservation, may be the main functional unit and evolutionary constraint. Thus, exploring functions through predicting lncRNA structures is a promising approach. Unfortunately, current state of the art RNA structure predictions still remain a high false-positive rate [161, 149].

Co-expression networks are widely used to explore the system-level functionality of genes. Genes that are co-expressed are likely to be co-regulated and functionally related. Thus, finding co-expressed protein-coding genes can help us to assign functions to unexplored lncRNAs.

Here, we describe our approach to reproduce the publication on co-expression network analysis by Squing li et al [161]. They focused on the expression rates of lncRNAs amongst four different cancer types: lung adenocarcinoma (ADC), bladder cancer (BLC), ER+ breast cancer (EBC) and prostate cancer (PRC). After analyzing expression profiles and perform-

Table 7 RNA-Seq datasets used.

Accession	Cancer type	Samples (pairs)	Library layout	Reads length
SRP012656	lung adenocarcinoma (ADC)	12	paired	78
SRP042620	ER+ breast cancer (EBC)	30	paired	50
ERP000550	Prostate cancer (PRC)	14	paired	90
SRP018008	Bladder cancer (BLC)	16	paired	100

ing a differential expression analysis, they constructed a co-expression network of lncRNAs that were altered in at least two cancer types and their correlated protein-coding genes. Finally, they obtained 12 functionally characterized modules, which allow to draw conclusions on previously unknown lncRNA functions.

13.4.1 RNA-Seq Datasets

We used the same RNA-seq datasets as used by Suqing Li et al., which are publicly available at the European Nucleotide Archive (http://www.ebi.ac.uk/ena). They only used tumor samples with matched adjacent normal samples. Trying to obtain the same number of matched tumor and normal samples, we faced a number of difficulties.

The bladder cancer dataset (SRP018008) provided us with two matching normal samples for each tumor sample. Ultimately we only used one of those matching normal samples in our analysis. Comparing the differential expression rates between each of the two normal samples and the tumor sample may be done in further steps. Another striking thing was that the group reported to have used 11 sample pairs, although matching samples were available for 16 patients. After checking the supplementary material of the bladder cancer paper, we noticed that 5 of the 16 patients had a tumor of grade 1 (ranging from 1 to 3) [162]. These 5 patients may have been excluded from the analysis. However, this is not reported in the paper and remains pure speculation. In the end, we analyzed all of the 16 sample pairs.

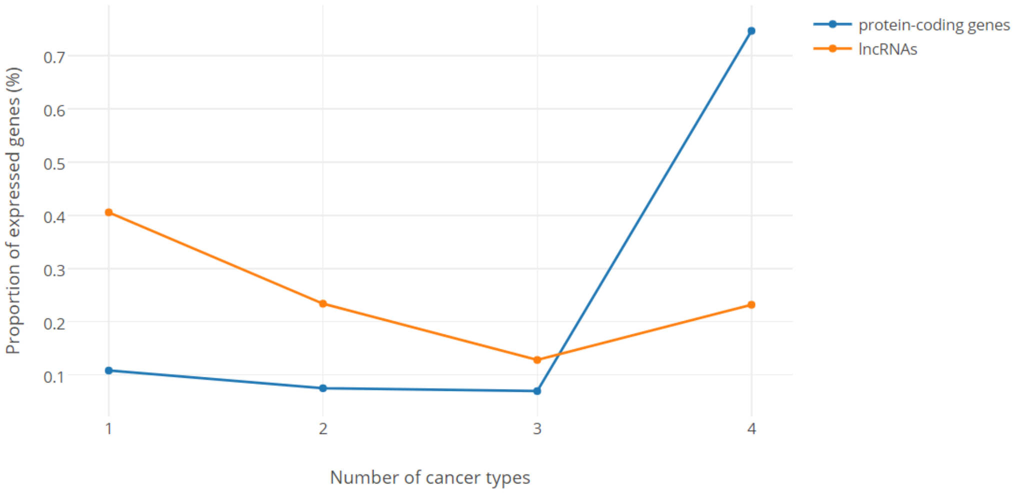
Prostate cancer RNA-seq data (ERP000550) provided normal and tumor samples for six patients. For each patient, two tumor samples and two matching normal tissue samples were available. Since Suqing Li et al. reported to have used 10 sample pairs for this dataset, we included both sample pairs per patient into our analysis, although each patient is accounted twice consequently.

The resulting number of useful sample pairs is reported in table 7.

13.4.2 Raw reads alignment and expression quantification

TopHat v2.0.13 [163] was used to map the raw reads to the reference genome with the GENCODE v23 gtf file [164] as annotation file. In order to distinguish between protein-coding genes and lncRNA genes, we downloaded the long non-coding RNA gene annotation file from the GENCODE website. This file contains 15.931 lncRNA genes, which correspond to the following biotypes: sense_intronic, sense_overlapping, antisense, lincRNA, macro_lncRNA, bidirectional_promoter_lncRNA, TEC, processed_transcript and 3prime overlapping_ncRNA (see http://www.gencodegenes.org/gencode_biotypes.html for detailed information). While most of these biotypes clearly refer to lncRNAs, some of them have rather shady definitions and should be looked at in detail before including them into the analysis. However, in order to achieve similar results as Suqing Li et al, we included all 15.931 lncRNA genes into the

Figure 130 The proportion of expressed protein-coding genes (blue) and lncRNA genes (red) appearing in different number of cancers. The x axis depicts the number of cancer types. The y-axis depicts the proportion of expressed genes, which is the ratio between the counts of expressed genes appearing in different number of cancers and the total counts of all expressed genes.



analysis. The FASTQ files were available as paired end reads, so we ran TopHat [163] accordingly. After all the FASTQ files were succesfully mapped we had 144 BAM files to be used for further processing.

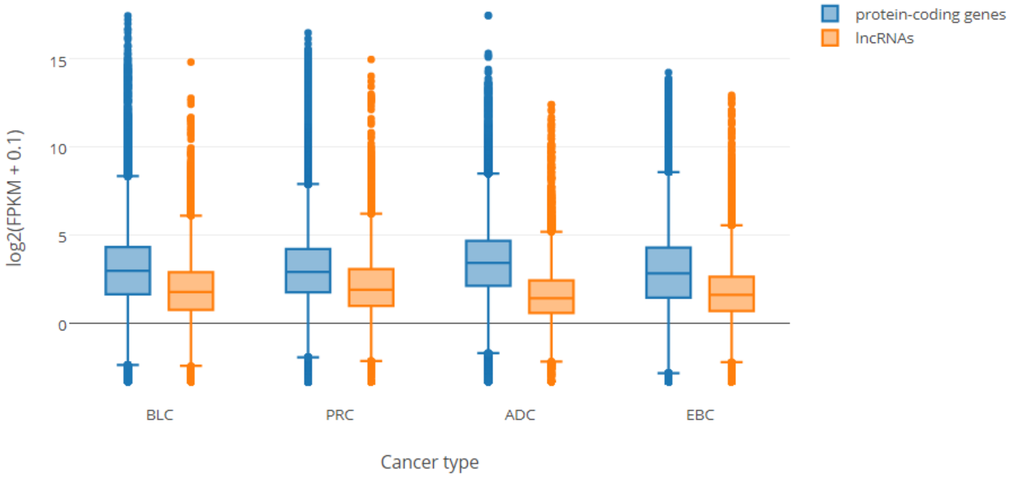
Cufflinks v2.2.1 [165] was used to perform gene assembly and quantification on the BAM files. The reported FPKM values (Fragment Per Kilobase per Million mapped reads) were used as measure of the gene expression levels. In order to minimize the number of false positives in the data, only significantly expressed genes were retained. Significantly expressed genes were required to have $FPKM \geq 1$ in more than 80% of the normal samples or 80% of the tumor samples for each cancer type. Filtering the gene set according to this criterion, we were left with a total of 14.290 expressed protein-coding genes and 2.004 expressed lncRNA genes in the four cancer types. These numbers are lower than the ones reported by Suqing Li et al, especially the number of expressed lncRNA genes (2.004 instead of 2.902). Due to a lack of detailed information we were not able to increase those numbers any further.

Figure 130 illustrates the number and proportion of expressed genes appearing in different numbers of cancers. We can see a clear difference between PCGs and lncRNA genes. While the vast majority of PCGs (74.7%), are found to be expressed in all of the four cancer types, the lncRNA genes display a more specific expression, with only 23.2% being expressed in all four cancer types.

Instead, 40.6% of the lncRNA genes are expressed in only one cancer type, compared to 10.8% of the PCGs. Furthermore, the distribution of FPKM values of lncRNA genes and PCGs for each cancer type is shown in figure 131 . Notably, lncRNA genes have overall lower expression levels than PCGs in every cancer type.

As it was covered in previous sections of this article, those observations go inline with previous studies that pointed out the low expression and high tissue-specificity of lncRNAs.

Figure 131 The expression levels (log2(FPKM + 0.1)) of expressed protein-coding genes (blue) and lncRNAs (orange) in each cancer type.



13.4.3 Differential expression analysis

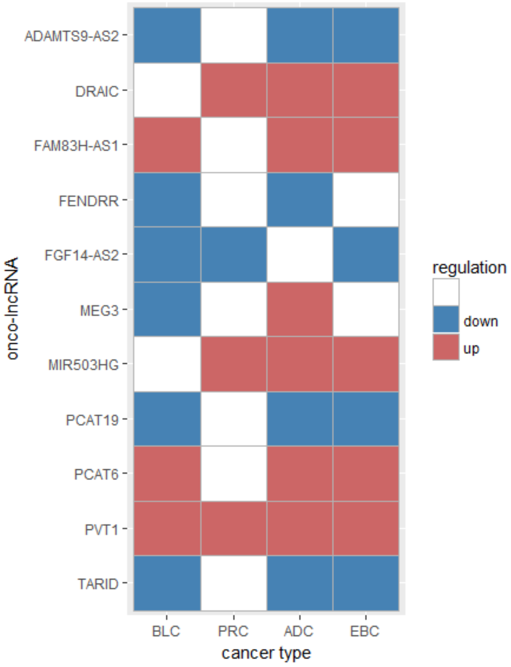
The next section covers a differential expression analysis on the remaining expressed genes. DESeq2 [166] was used to test for differential expression between the tumor and normal samples. Since Cufflinks [165] only reports normalized expression values (FPKM) and it is not recommended to run DESeq2 with already normalized input, we performed HTSeq [167] on the BAM files to obtain the raw counts for every sample. After running DESeq2 with the raw counts, we had the fold change (FC) and FDR (False Discovey Rate) corrected p value for every cancer type. A gene was defined to be differentially expressed if its FC is at least 2 times higher or lower ($|log2FC \geq 1|$) and its FDR corrected p value is less than 0.01 ($FDR \leq 0.01$).

This way we obtained 866 differentially expressed lncRNA genes (DELs) among all four cancer types with 247, 200, 243 and 445 DELs in BLC, PRC, ADC and EBC, respectively. We also identified 5686 differentially expressed protein-coding genes in all four cancer types. Although these numbers are not identical to the numbers reported by the group around Suqing Li, they are somewhat comparable [161].

LncRNAs that are differentially expressed in more than one cancer type represent a particularly interesting group, since they are likely to possess key roles in the development and maintenance of tumor cells. Therefore, we defined those lncRNAs as onco-lncRNAs, which were used to construct a co-expression network later on. However, as to be expected most lncRNAs (75.1%) were altered only in one cancer type. The remaining 216 were thus classified as onco-lncRNAs. Only five lncRNAs were altered in all of the four cancer types, amongst them PVT1, which is already known to be associated with different types of cancer.

Most of the onco-lncRNAs identified have not been reported to be associated with cancer before. Figure 132 shows the differential expression of the 11 genes among our onco-lncRNAs that are known to be associated with cancer. However, only two of those genes have been proven to be altered in more than one cancer type whereas the other nine were only studied

Figure 132 Differential expression of 11 functionally characterized cancer associated lncRNAs in lung adenocarcinoma (ADC), bladder cancer (BLC), ER+ breast cancer (EBC) and prostate cancer (PRC). The differential expression is measured as log2(FoldChange). Red denotes upregulation and blue downregulation of the lncRNA.



in one cancer type.

13.4.4 Co-expression network data prepariation

In the next step, a co-expression network was constructed based on the normalized expression profiles of the 216 onco-lncRNAs and their “closely correlated” PCGs. The normalized expression values for the network construction were provided by Cuffnorm [165], which gave us five different expression tables: one with all the samples for each corresponding cancer type and one with the expression values for all the cancer types. The advantage of running Cuffnorm is that it normalizes not only for one library, but among all the samples, which makes the ouput better comparable.

There are multiple ways to define the correlated PCGs. As proposed by Suqing Li et al, we defined a PCG to be “closely correlated” with the onco-lncRNAs when its absolute values of Pearson correlation coefficients with more than 5 onco-lncRNAs are equal or greater than 0.5. According to this definition, the PCGs were filtered using two different approaches. In the first approach we used the whole set of onco-lncRNAs to search for correlated PCGs within all the samples combined. This way we got 9000 correlated PCGs.

In the other approach, the onco-lncRNAs were divided into subsets, according to their association with a specific cancer type. This resulted in four sets of onco-lncRNAs, one for each cancer type. Correlations were then calculated individually for every cancer, using only the corresponding onco-lncRNAs and samples for this cancer type. We got 13931, 9661, 10309 and 9332 correlated PCGs for ADC, BLC, EBC and PRC, respectively. The intersection

of those gene sets was then used as the final list of “closely correlated” PCGs, which were 2557 genes in total. Although this approach is much stricter, every PCG contained in the network now is assured to be co-expressed with a lncRNA in every cancer type.
In the end, we calculated an individual co-expression network for both sets of correlated PCGs. The two networks can be compared against each other in order to assure that the predicted modules are correct.

13.4.5 Soft thresholding

In a co-expression network nodes represent genes and nodes are connected if the corresponding genes are significantly co-expressed across the tissue samples. When constructing such a co-expression network, one needs to define what a connection between two genes displays. A straighforward approach is to use binary information (connected=1, unconnected=0), so that two genes are connected if their correlation exceeds a certain threshold. We refer to this approach as hard thresholding or an unweighted co-expression network. This inevitable leads, however, to information loss, given the continuous nature of the underlying co-expression information [168].
Weighted Gene Correlation Network Analysis (WGCNA) was used for network construction. WGCNA is an R software package which uses a soft-thresholding approach to construct a co-expression network [169]. A co-expression network corresponds to an underlying adjacency matrix. In order to calculate the adjacency matrix we need to define a similarity matrix. In this case, our similarity matrix S contains the absolute values of parwise Pearson correlations, which we denote as $S = [s_{ij}]$, with $s_{ij} = |cor(i,j)|$. This similarity matrix is then transformed into an adjacency matrix A of connection strengths through soft thresholding. This was done with the adjacency function $a_{ij} = power(s_{ij}, \beta) \equiv |s_{ij}|^{\beta}$, with the single parameter β.
There are several ways to choose the parameter β, depending on the desired characteristics of our network. In this case, we want the network to display a scale free topology. Many real networks in fact resemble scale free networks with their degree distribution following a power law $p(k) \sim\ k^{-\gamma}$. As a result they exhibit a few highly connected nodes (hubs) with the majority of the nodes having only a small number of links. This characteristic makes the network very robust to random errors while being extremely vulnerable to attacks. In order to achieve such a scale free topology, we try to maximize the scale-free topology model fit R^2 while retaining high number of mean connections (Figure 133) [22,23]. In the presented work, there is a trade-off between maximizing R^2 and a high mean connectivity at $\beta = 9$.

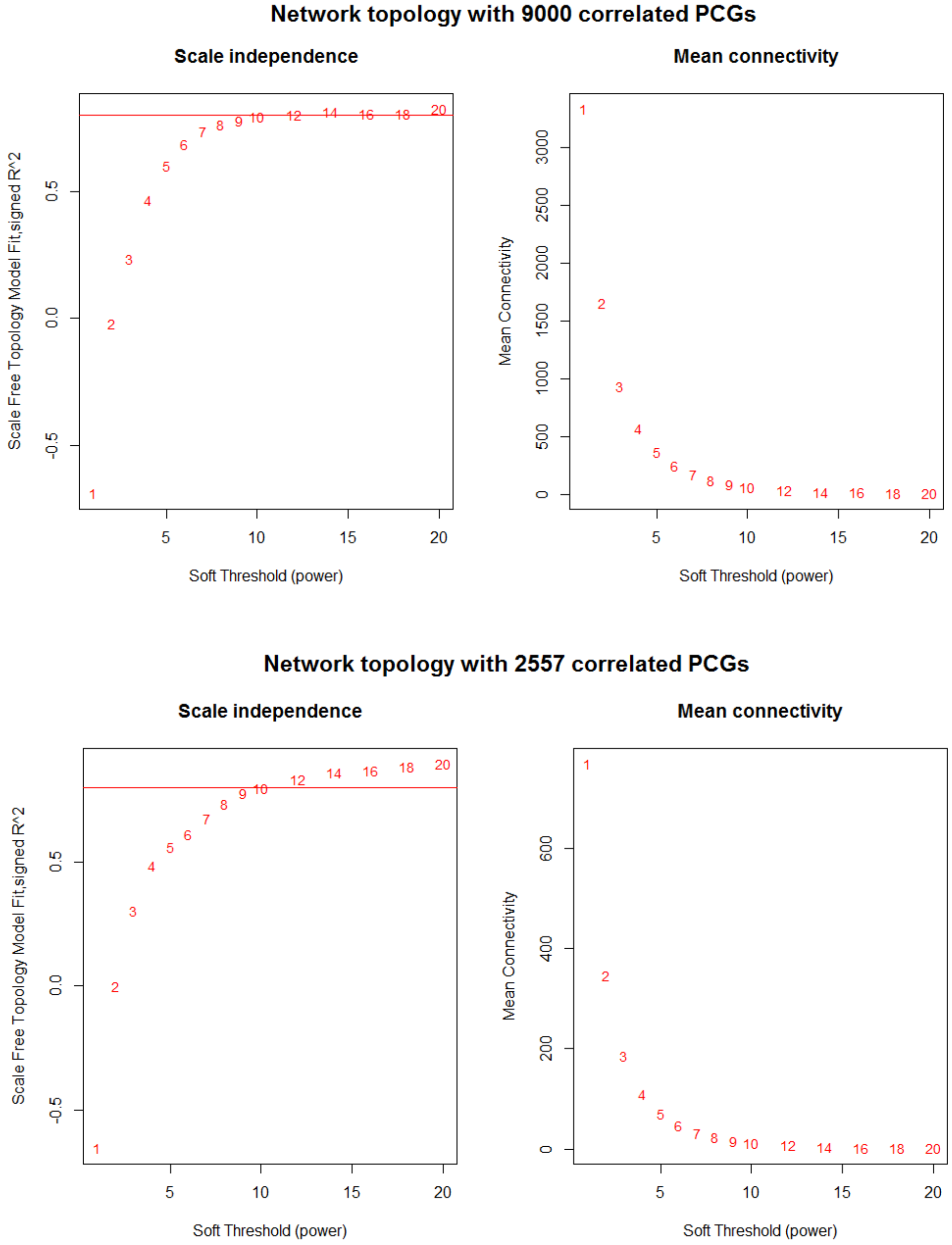
13.4.6 Module detection

In the next step, we want to subset the network into clusters of densely connected and therefore functional similar nodes (modules). This was performed through average linkage hierarchical clustering. For the clustering a dissimilarity matrix was used, based on the topological overlap matrix (TOM). The topological overlap of two nodes is calculated as follows:

$w_{ij} = \frac{l_{ij}+a_{ij}}{min\{k_i,k_j\}+1-a_{ij}}$ where $l_{ij} = \sum_u a_{iu}a_{ui}$ and $k_i = \sum_u a_{iu}$ is the node connectivity.

This measure can adapt values ranging from 0 to 1. Speaking of hard thresholding, a value of 0 would refer to both nodes being unconnected and not sharing a single neighbor, while a value of 1 would represent two nodes that are connected and share all of their neighbours.

Figure 133 Analysis of network topology for different soft thresholds. The left panel shows the scale-free topology fitting index (R^2, y-axis) as a function of the soft-thresholding power (x-axis). The right panel displays the mean connectivity (degree, y-axis) as a function of the soft-thresholding power (x-axis). Red numbers in the panels denote different soft-thresholds. The red line in left panel means $R^2 = 0.8$. The first plot corresponds to the network based on the 9000 correlated PCGs, the second to the network based on 2557 correlated PCGs.



But the formula can be applied to soft thresholing as well. To transform the TOM matrix into a dissimilarity matrix, we simply substract all of its values from 1 ($d_{ij} = 1 - wij$) [11,12].

As already mentioned, hierarchical clustering was applied to detect the modules. The resulting tree for both datasets is illustrated in fig. 6 as well as the corresponding modules, which were defined by a dynamic tree cutting algorithm [170]. Since in the first dataset (with more correlated PCGs) 58 modules were detected, a lot of them contain only very little genes. Therefore we used the merge function from the WGCNA package [169] to merge close modules together. This left us with 11 modules in the first dataset (figure 134).
In the other dataset 13 different modules were detected, two of which did not contain any lncRNAs and thus were excluded from further analysis.

13.4.7 Functional characterization of detected modules

With the modules being identified, the genes had to be mapped to external information in order explore the possible functions of the lncRNAs. The Database for Annotation, Visualization and Integrated Discovery (DAVID) was used to perform such a functional enrichment analysis.
The reported modules shown can be explored interactively. By clicking on a node in the network, all co-expressed PCGs are listed with their corresponding functions. Furthermore, the significance of enriched GO-terms as well as KEGG pathways is displayed for the selected module, which gives a good overview of the predominant functions.
In the following, some of the modules are looked at in detail and possible functions of previously uncharacterized lncRNAs are proposed. The second network was used for this purpose.

The purple module has a lot of significantly enriched GO-terms, the five most significant being angiogenesis, positive regulation of angiogenesis, response to hypoxia, leukocyte migration and vasculogenesis, in that order (table 8). Signifcant KEGG pathways are cell adhesion molecules, transcriptional misregulation in cancer, leukocyte transendothelial migration, Ras signaling pathway and pathways in cancer (table 9). Most of these terms are directly linked to angiogenesis, whose disregulation is crucial to the growth of tumors. Among the lncRNAs in this module is FENDRR, which is an essential regulator of heart and body wall development. It modulates the histone-modifying complex PRC2, which is important for cell differentiation. This function fits the functional enrichment of the module. Although FENDRR is not known to be associated with cancer, it is significantly down regulated in lung adenocarcinoma and bladder cancer. Interestingly, FENDRR as well as other lncRNAs show very high co-expression with the protein HSPA12B, which was recently discovered to stimulates lung tumor growth via a Cox-2-dependent mechanism [171]. They show another strong co-expression with RAMP2, which was found down-regulated in a majority of lung tumors, and RAMP2 down-regulation was correlated with high tumor grade [172] (figure 135). Summarizing the presented results, lncRNAs in the purple module are likely to be involved in tumor growth, especially in the formation of new blood vessels.
The magenta module is significantly enriched in funcions that refer to RNA splicing. Amongst the central nodes in this module is the known cancer associated lncRNA MALAT1, which regulates alternative splicing by modulating SR splicing factor phosphorylation and thus expression of metastasis-associated genes [173]. Some of the it's stronger connections are involved in processes affecting cell motility, which MALAT1 is known to enhance in lung adenocarcinoma cells [174]. Other lncRNAs in this module also show strong connections to genes involved in phosphorylation. In summary, we can conclude that genes in the magenta

Figure 134 Modules defined by the weighted correlation network analysis (WGCNA). Tree on the top is the clustering dendrogram of genes, and the colorful bands represent modules in this network. The first plot corresponds to the network based on the 9000 correlated PCGs, the second to the network based on 2557 correlated PCGs.

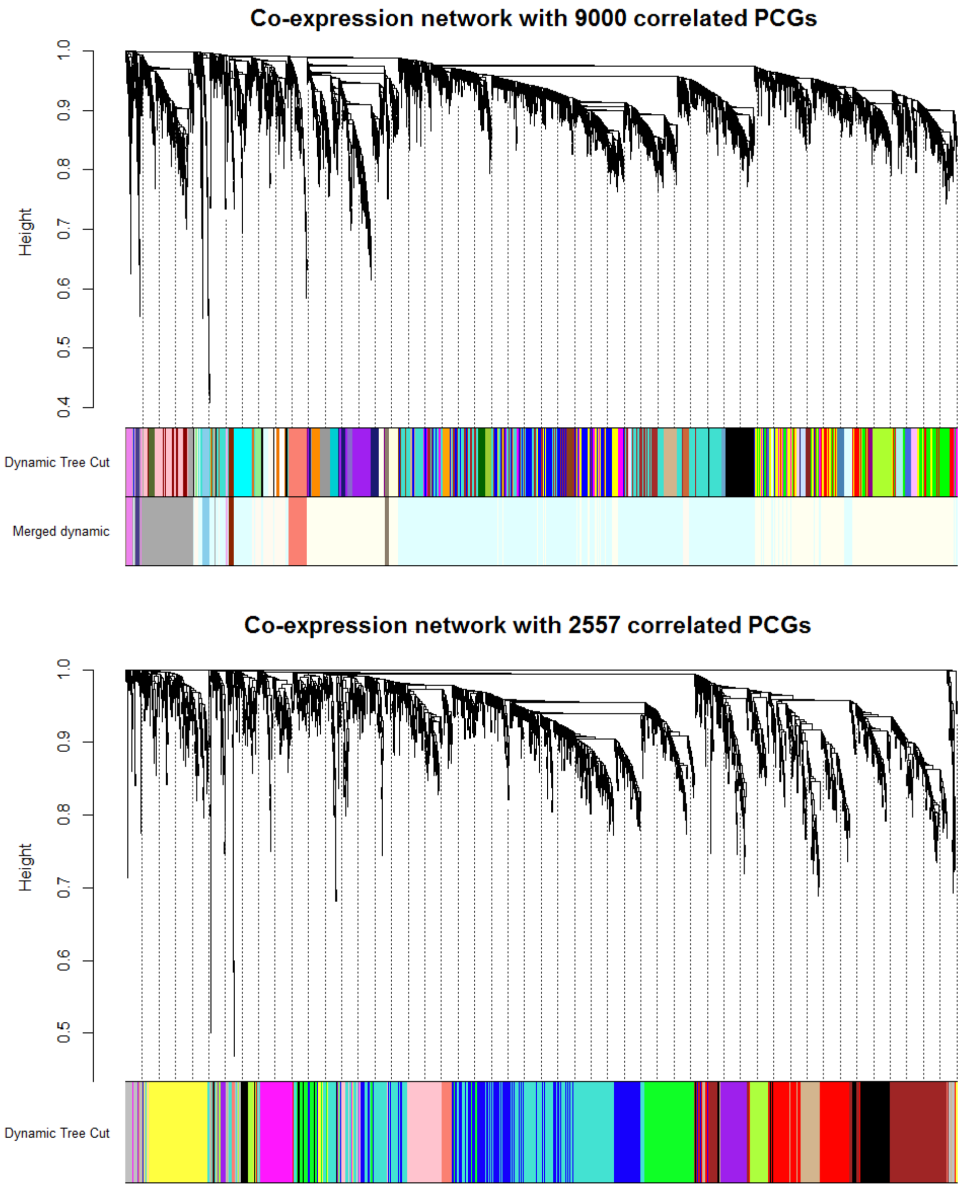


Table 8 Enrichment of GO terms in the purple module. Count refers to the number of genes in the module that exhibit the GO term.

ID	function	count	p value
GO:0001525	angiogenesis	12	7.2476404551477825E-9
GO:0045766	positive regulation of angiogenesis	9	6.58737640537967E-8
GO:0001666	response to hypoxia	9	1.446834327422244E-6
GO:0050900	leukocyte migration	7	2.3435315069869458E-5
GO:0001570	vasculogenesis	5	1.3839967589956192E-4
GO:0007165	signal transduction	15	9.482859188686749E-4

Table 9 Enrichment of KEGG pathways in the purple module. Count refers to the number of genes in the module that are associated with the KEGG pathway.

ID	pathway	count	p value
hsa04514	Cell adhesion molecules (CAMs)	7	1.9322267961281267E-4
hsa05202	Transcriptional misregulation in cancer	7	4.793944670031836E-4
hsa04670	Leukocyte transendothelial migration	4	0.03439455095176435
hsa04014	Ras signaling pathway	5	0.0473008140350735
hsa05200	Pathways in cancer	6	0.08789987217536796

module are likely to be involved in the regulation of alternative splicing.

The brown modules is mainly characterized by functions related to signal transduction. This includes the Wnt signaling pathway and the Hippo signaling pathway, which are involved into cell proliferation, apoptosis and cell migration. The module contains ADAMTS9-AS2, a lncRNA that is known to inhibit migration of glioma cells [175]. Therefore we propose lncRNAs contained in this module to be involved in the regulation of signal transduction.

Notably, these conclusions represent only a minor part of all possible observations that can be made by simply looking at the co-expression network. It may contain several more insightful reveals to the functional landscape of lncRNAs in the human genome.

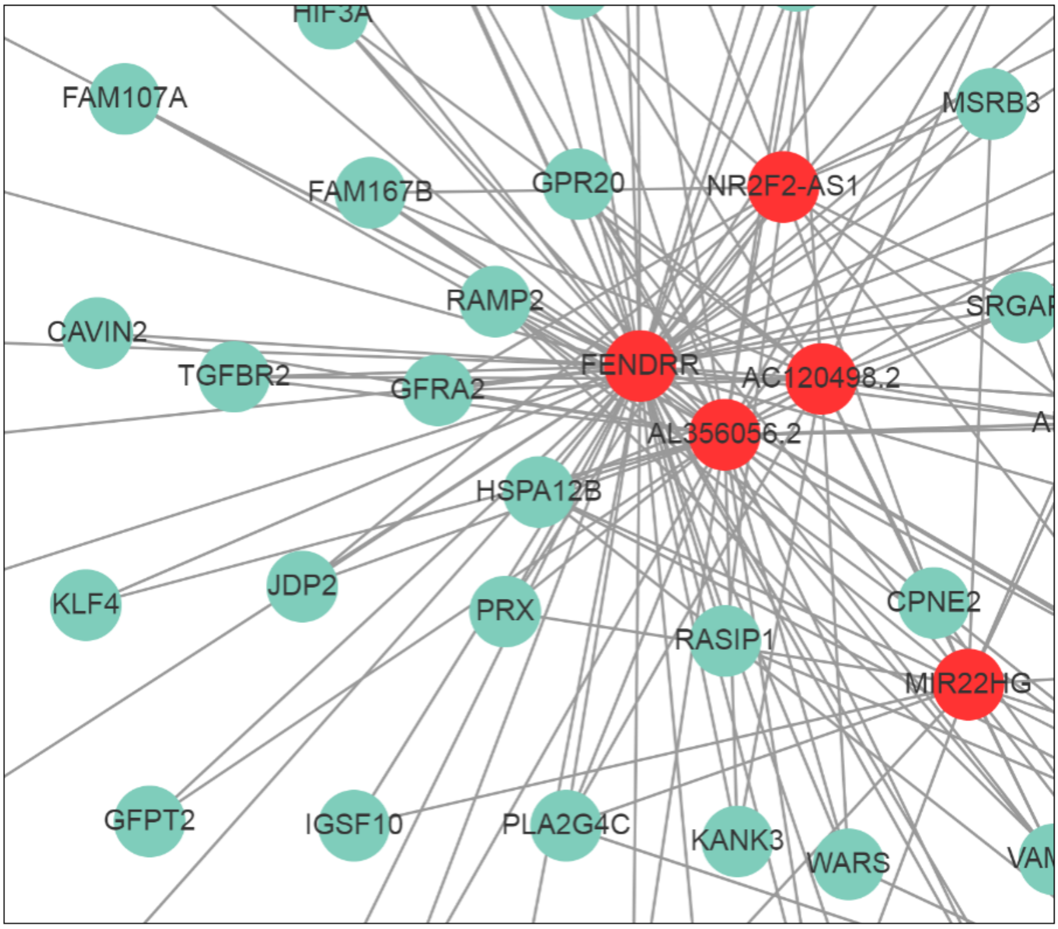
13.5 Conclusion

LncRNAs carry out important biological functions, their disregulation is associated with various forms of cancer in a lot of cases. However, the information on the function and mechanisms underneath most lncRNAs are unknown. In this work, we presented an approach to narrow down potential functions of previously uncharacterized lncRNAs that are likely to be associated with multiple types of cancer using co-expression networks. Our results reveal some lncRNAs that may indeed play crucial roles in the development and maintenance of cancer. Since experimental characterization in this field is still time consuming and requires some previous knowledge, this may serve as a potential guide for subsequent experimental studies.

References

[99] Hadas Keren, Galit Lev-Maor, and Gil Ast. “Alternative splicing and evolution: diversification, exon definition and function”. In: *Nature Reviews Genetics* 11.5 (2010), pp. 345–355. ISSN: 1471-0056. DOI: `10.1038/nrg2776`. URL: `http://www.ncbi.`

Figure 135 An extract of the purple module from the co-expression network build on 9000 correlated PCGs and the onco-lncRNAs. Red nodes denote long non-coding RNAs and green nodes denote protein coding genes.



nlm.nih.gov/pubmed/20376054http://www.nature.com/doifinder/10.1038/nrg2776.

[100] J. Merkin et al. “Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues”. In: *Science* 338.6114 (2012), pp. 1593–1599. ISSN: 0036-8075. DOI: 10.1126/science.1228186. arXiv: NIHMS150003. URL: http://www.ncbi.nlm.nih.gov/pubmed/23258891http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3568499http://www.sciencemag.org/cgi/doi/10.1126/science.1228186.

[101] Wolfgang Huber et al. “Orchestrating high-throughput genomic analysis with Bioconductor”. In: *Nature Methods* 12.2 (2015), pp. 115–121. ISSN: 1548-7091. DOI: 10.1038/nmeth.3252. arXiv: 9809069v1 [arXiv:gr-qc]. URL: http://www.nature.com/doifinder/10.1038/nmeth.3252http://www.ncbi.nlm.nih.gov/pubmed/25633503http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4509590.

[102] Simon Anders, Alejandro Reyes, and Wolfgang Huber. “Detecting differential usage of exons from RNA-seq data”. In: *Genome Research* 22.10 (2012), pp. 2008–2017. ISSN: 10889051. DOI: 10.1101/gr.133744.111. arXiv: arXiv:1011.1669v3. URL: http://www.ncbi.nlm.nih.gov/pubmed/22722343http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3460195http://genome.cshlp.org/cgi/doi/10.1101/gr.133744.111.

[103] Hugues Richard et al. “Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments”. In: *Nucleic Acids Research* 38.10 (2010), e112. ISSN: 03051048. DOI: 10.1093/nar/gkq041. URL: http://www.ncbi.nlm.nih.gov/pubmed/20150413http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2879520.

[104] Cole Trapnell et al. “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation”. In: *Nature Biotechnology* 28.5 (2010), pp. 511–515. ISSN: 1087-0156. DOI: 10.1038/nbt.1621. arXiv: 171. URL: http://www.ncbi.nlm.nih.gov/pubmed/20436464http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3146043http://www.nature.com/doifinder/10.1038/nbt.1621.

[105] Yarden Katz et al. “Analysis and design of RNA sequencing experiments for identifying isoform regulation”. In: *Nature Methods* 7.12 (2010), pp. 1009–1015. ISSN: 1548-7091. DOI: 10.1038/nmeth.1528. arXiv: 9605103 [cs]. URL: http://www.nature.com/doifinder/10.1038/nmeth.1528http://www.ncbi.nlm.nih.gov/pubmed/21057496http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3037023.

[106] Malachi Griffith et al. “Alternative expression analysis by RNA sequencing”. In: *Nature Methods* 7.10 (2010), pp. 843–847. ISSN: 1548-7091. DOI: 10.1038/nmeth.1503. URL: http://www.ncbi.nlm.nih.gov/pubmed/20835245http://www.nature.com/doifinder/10.1038/nmeth.1503.

[107] Shihao Shen et al. “rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data”. In: *Proceedings of the National Academy of Sciences* 111.51 (2014), E5593–E5601. ISSN: 0027-8424. DOI: 10.1073/pnas.1419161111. arXiv: arXiv:1408.1149. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1419161111http://www.ncbi.nlm.nih.gov/pubmed/25480548http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4280593.

[108] Cole Trapnell et al. “Differential analysis of gene regulation at transcript resolution with RNA-seq.” In: *Nature biotechnology* 31.1 (2013), pp. 46–53. ISSN: 1546-1696. DOI: 10.1038/nbt.2450. arXiv: NIHMS150003. URL: http://www.nature.com/doifinder/10.1038/nbt.2450http://www.ncbi.nlm.nih.gov/pubmed/23222703{\%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3869392.

[109] Dorothea Emig et al. “AltAnalyze and DomainGraph: Analyzing and visualizing exon expression data”. In: *Nucleic Acids Research* 38.SUPPL. 2 (2010), W755–62. ISSN: 03051048. DOI: 10.1093/nar/gkq405. URL: http://www.ncbi.nlm.nih.gov/pubmed/20513647http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2896198.

[110] Jie Zhang and Zhi Wei. “An empirical Bayes change-point model for identifying 3’ and 5’ alternative splicing by next-generation RNA sequencing”. In: *Bioinformatics* 32.12 (2016), pp. 1823–1831. ISSN: 14602059. DOI: 10.1093/bioinformatics/btw060. URL: http://www.ncbi.nlm.nih.gov/pubmed/26873932https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw060.

[111] André Kahles et al. “SplAdder: Identification, quantification and testing of alternative splicing events from RNA-Seq data”. In: *Bioinformatics* 32.12 (2016), pp. 1840–1847. ISSN: 14602059. DOI: 10.1093/bioinformatics/btw076. URL: http://www.ncbi.nlm.nih.gov/pubmed/26873928http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4908322https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw076.

[112] Scott Norton, Jorge Vaquero-Garcia, and Yoseph Barash. “Outlier detection for improved differential splicing quantification from RNA-Seq experiments with replicates”. In: *bioRxiv* (2017), pp. 1–15. DOI: 10.1101/104059. URL: https://www.biorxiv.org/content/early/2017/05/11/104059.

[113] Sylvain Foissac and Michael Sammeth. “ASTALAVISTA: Dynamic and flexible analysis of alternative splicing events in custom gene datasets”. In: *Nucleic Acids Research* 35.SUPPL.2 (2007), W297–W299. ISSN: 03051048. DOI: 10.1093/nar/gkm311. URL: http://www.ncbi.nlm.nih.gov/pubmed/17485470http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1933205https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkm311.

[114] Jennifer Harrow et al. “GENCODE: producing a reference annotation for ENCODE”. In: *Genome Biology* 7.Suppl 1 (2006), S4. ISSN: 14656906. DOI: 10.1186/gb-2006-7-s1-s4. URL: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2006-7-s1-s4.

[115] Andrew Yates et al. “Ensembl 2016.” In: *Nucleic acids research* 44.D1 (2016), pp. D710–6. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1157. arXiv: arXiv:1011.1669v3. URL: http://www.ncbi.nlm.nih.gov/pubmed/26687719http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4702834http://www.ncbi.nlm.nih.gov/pubmed/26687719{\%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4702834.

[116] Hsien-Da Huang et al. “ProSplicer: a database of putative alternative splicing information derived from protein, mRNA and expressed sequence tag sequence data.” In: *Genome biology* 4.4 (2003), R29. ISSN: 1465-6914. DOI: 10.1186/gb-2003-4-4-r29. URL: http://www.ncbi.nlm.nih.gov/pubmed/12702210http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC154580.

14 miRNAs

by Gregor Sturm, Sebastian Wilzbach and Markus Joppich

14.1 Introduction

MicroRNAs (miRNAs) are a class of small non-coding RNAs of about 22 nucleotides (nt) length that play an important role in gene regulation ([176]). While both plants and animals have miRNAs that are conserved in several aspects, the process of biogenesis differs to a certain extent between these two *regna* ([176]).

miRNAs are essential for animal development ([177]) and dysregulation of miRNA expression has been associated with the pathogenesis of various fatal diseases such as cancer ([178]), neurodegeneration ([179]) or atherosclerosis ([180]). A thorough understanding of the process of miRNA biogenesis and regulation will contribute to disease understanding and identifying novel drug targets.

In this chapter, we review the biosynthesis of miRNAs and miRNA-induced gene regulation. Moreover, we provide an overview over publicly available miRNA resources such as databases and *in silico* prediction algorithms and present a new tool wich combines several existing miRNA resources into one unified database with a user-friendly modern graph-based user interface.

14.1.1 Biosynthesis of miRNAs

The biogenesis of mature miRNAs is a multi-step-process (illustrated in figure 136). The primary miRNA precursors (pri-miRNA) are mainly transcribed by RNA Polymerase II ([181]), although some are transcribed by RNA Polymerase III ([182]). They are further processed by the Microprocessor complex into a ~ 70 nt hairpin precursors (pre-miRNA) with a ~ 2 nt 3' overhang ([183, 184]). Not all proteins forming this complex are known, but it has been shown that both Drosha, which is a member of the RNase III family ([185]) and its partner DGCR8/Pasha, which provides a dsRNA binding domain ([186]) are essential. After the cleavage, the Ran-GFP dependent Exportin 5 mediates the export of the pre-miRNAs to the cytoplasm ([187, 188]). There, they are cleaved by Dicer ([189]), another RNase III, that acts in a complex with TRBP ([190]). Dicer measures the distance from the 3' overhang generated by Drosha ([191]) and yields a short (~ 22 nt) double stranded RNA with a 2 nt 3' overhang.

Besides the canonical pathway, alternative ways of miRNA processing exist (reviewed in [192]). Yang and Lai [193] reported of the Dicer-independent *miR-451*, which is conserved among vertebrates. Instead of Dicer it is cleaved by the catalytic center of AGO2. Moreover, there exists the Drosha-independent *mirtron*-pathway, first discovered by Okamura et al. [194] in *D. Melanogaster*. Instead of being cleaved from a pri-miRNA, the pre-miRNA results directly from a spliced intron. Berezikov et al. [195] identified several human mirtrons.

14.1.2 miRNA-induced gene silencing

Usually, only one strand of the mature miRNA-duplex is loaded into the RNA-induced silencing complex (RISC), while the other strand is degraded ([197]). However, in some cases, both strands are loaded ([198]). Schwarz et al. [199] proposed a model in which the duplex is unwound by a helicase that starts at the thermodynamically less stable end and directs only one strand into the RISC. If both ends are equally stable, both strands might be selected.

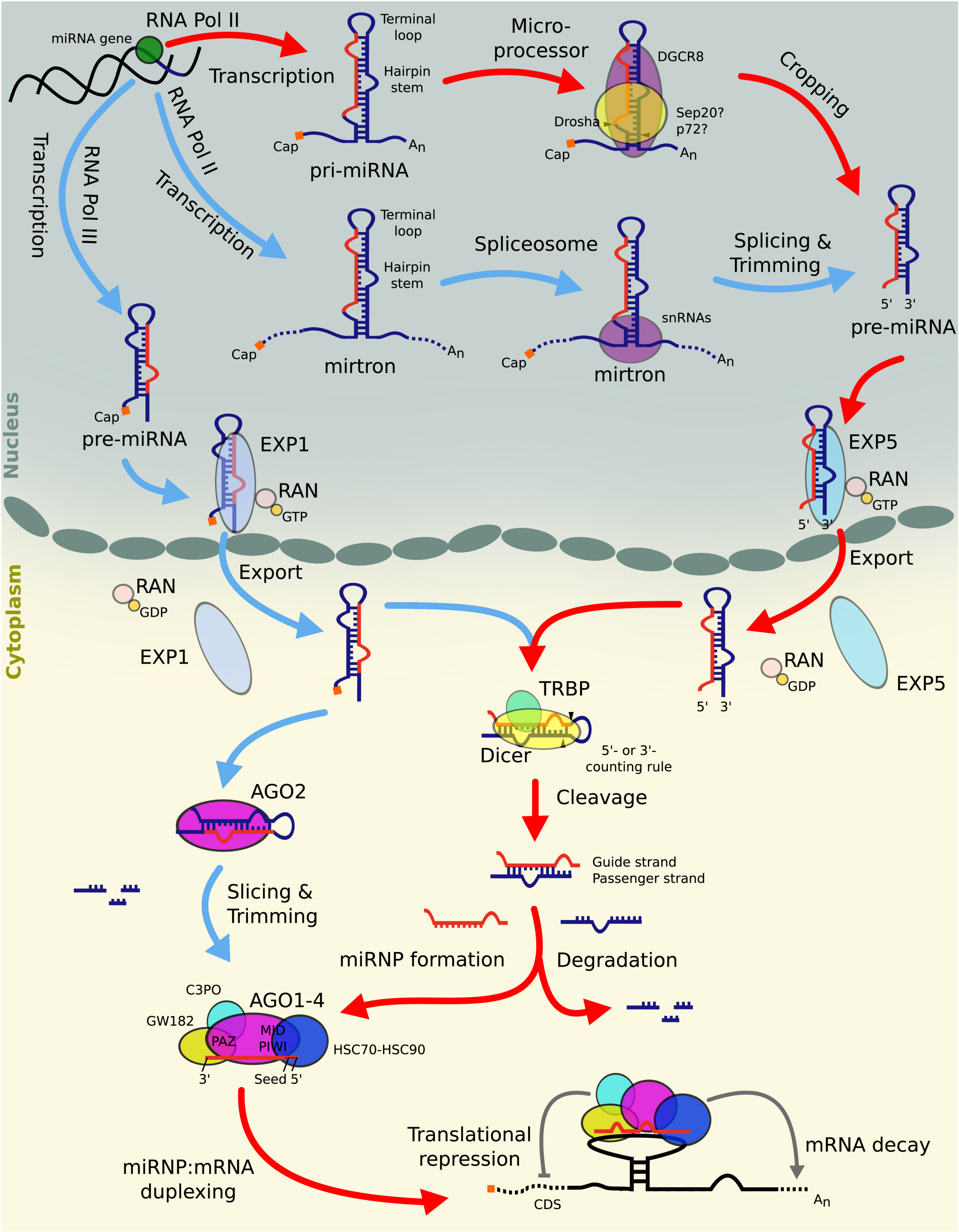


Figure 136 MicroRNA synthesis and regulation. Figure from Ellwanger [196] p. 7

Argonaute (AGO) proteins form the active component of the RISC that binds the miRNA and eventually directs the complex towards its target. AGO2 further provides the nuclease domain for nucleolytic cleavage of mRNA (reviewed in [200]).

In metazoa, RISC target selection is mainly guided by perfect base-pairing of a 6-8 nt *seed region* within the first 8 nt of the mature miRNA and the target ([201, 202]), although other determinants have also been identified (see section 14.2). Once bound, the RISC interacts with its target through different mechanisms: It can (1) inhibit translation, (2) deadenylate the poly(A) tail of mRNAs, leading to degeneration or (3) degrade mRNA by nucleolytic cleavage, given high sequence complementary between miRNA and target (reviewed in [203]).

14.2 Target Recognition and Target Prediction

Despite of considerable research effort, many challenges in understanding the RISC:target binding process remain ([204]). Recent findings that miRNA:target interactions are highly context-specific ([205]) make the landscape of miRNA targeting even more complex.

To identify factors licensing RISC binding several groups have analysed features of experimentally verified AGO binding sites. Alternatively, the predictive power of different features can be assessed *in silico*.

The adaption of CLIP-Seq methods (see section 14.3) has enabled high-throughput (HT) profiling of miRNA:target interactions leading to a growing set of experimental evidence. The downsides of this approach are experimental biases and limited accuracy. Moreover, binding sites detected with CLIP-Seq are not necessarily functional in terms of target regulation, even though the actual binding could be independently confirmed ([206]).

Alternatively, effective target sites can be predicted computationally based on the assumption that miRNA transfection leads to repression of its target. The goodness of a prediction can therefore be assessed by comparing the predicted strength of an interaction with the repression of the target measured by small RNA transfection assay, such that the predictive power of different features can be assessed individually without prior knowledge about the binding sites ([206]). Multiple features can be combined to improve accuracy of miRNA target prediction.

The most important feature is certainly the perfect Watson-Crick base pairing between the 6-8 nt 5' seed region of the miRNA and its target ([201, 202]). Multiple groups identified non-canonical binding events such as a bulge within the seed region. However, even though the binding could be experimentally verified, the regulative efficacy of such sites has been questioned ([206]).

Apart from the seed the following predictive features are commonly used (reviewed [207]):

1. **Site location within the 3' untranslated region (UTR)**. Most target sites reside within the 3' UTR of target genes. Although the RISC does bind 5' UTR and coding sequence (CDS) the 3' UTR is preferred, likely because there is less competition with other protein complexes.
2. **Conservation**. miRNA families are miRNAs with the same seed sequence and homologous targets that are well conserved among related species. While this feature has a strong predictive power in most cases, evidence suggest that about 30% of the binding sites are species-specific. Therefore it is important to run target prediction algorithms that do not only rely on this feature to increase sensitivity.
3. **Site accessibility**. An effective miRNA:target interaction needs an open structure to begin hybridization. Unfortunately, RNA secondary structure is hard to predict *in silico*.

Local A:U content as a surrogate heuristic has comparable predictive performance and is easy to compute.

4. **Multiple interacting sites**. Target sites within close distance ($\sim$ 14 nt–46 nt) act synergistically and enhance efficacy beyond the expected additive effect. If the distance is too close (<8 nt) the efficacy decreases, likely due to steric hindrance ([208, 209]),
5. **Additional 3' Watson-Crick pairing**. Additional Watson-Crick Pairing at Nucleotides 12-17 enhances downregulation of the target ([209]).

Many different target prediction algorithms are available, taking different features into account (reviewed in Saito and Sætrom [207] and Bartel [204]). As of July 2017, tools4mirs.org ([210]), an online platform gathering all tools and databases related to miRNAs, lists 45 different target prediction algorithms for human and mouse. Condition-specific prediction with computational methods yet needs to be addressed.

The arguably most commonly used target prediction algorithm is TargetScan v7.1 with the context++ model ([206]). It is based on four separate linear regression models for the seed types 6mer, 7mer-A1, 7mer-m8 and 8mer, taking 14 features into account, including conservation, target accessibility, sequence features and the position of the target site. The authors claim a predictive performance comparable to high-throughput experimental assays.

14.3 Experimental detection of miRNA:target interactions

In general, experiments to determine miRNA:target interactions can be divided into two categories: (a) classical, low-throughput methods and (b) high throughput methods. Traditionally, low-throughput experimental methods like Northern Blotting, qPCR, Western Blot, ELISA, and reporter assays have been used for determining miRNA:target interactions and are still used for high accuracy validations (reviewed in [196]).

Experimentally validated interactions are often seen as the 'gold standard' of miRNA interaction data. However, coverage is insufficient and the accuracy and reproducibility of recent high-throughput (HT) experiments is limited. In particular, HT experiments test for a binding event of a miRNA with a target, irrespective of whether the binding is functional in terms of genetic regulation or not.

14.3.1 CLIP-seq protocol

A more recent, state-of-the-art technology is ultraviolet (UV) crosslinking and immunoprecipitation (CLIP). Through UV crosslinking interactions between the *in vivo* RNA-binding proteins (RBP) and RNA from living cells or tissue samples are preserved. Subsequently, after cell lysis, partial digestion by Proteinase K into small fragments allows isolating the protein-RNA complex of interest with immunoprecipitation using a respective antibody. Afterwards, the extracted RNA is purified through amongst others radioactive labeling, ligated with 3' and 5' adaptors before being reverse-transcribed to cDNA, PCR amplified and sequenced ([211]). As this method is often combined with the power of high-throughput sequencing, it is commonly referred to as HITS-CLIP ([212, 196]).

Identification of binding sites relies on overlapping sequence clusters which limits the precision to a range of $\sim$ 30 nt. To identify the precise position, several variations of CLIP have been developed.

The first variation of CLIP is photoactivatable-ribonucleoside-enhanced CLIP (PAR-CLIP) where cells are fed with nucleoside analogue like 4-thiouridine or 6-thioguanosine, such that UV crosslinking at 365 nm instead of 256 nm is possible. The introduced nucleoside analogue leads to a base transition during the reverse transcription which can be used to track the

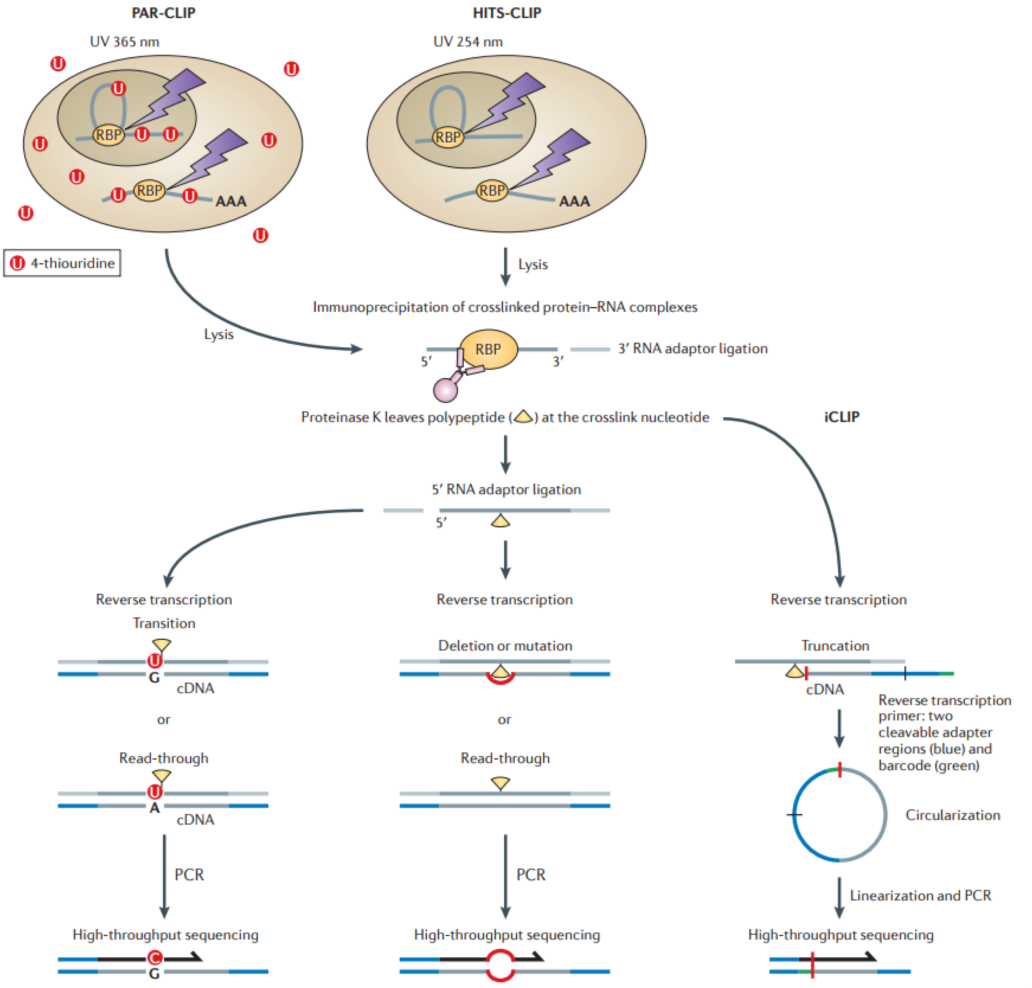


Figure 137 Comparison of HITS-CLIP and its latest variants, PAR-CLIP and iCLIP First, covalent bounds are induced between the RBP and RNA through UV cross-linking. PAR-CLIP cells are fed with a nucleoside analog, s.t. UV crosslinking can happen at 365 nm instead of 265 nm. The protein-RNA complex is isolated by immunoprecipitation and partially digested. Subsequently, the fragments are reverse transcribed and high-throughput sequenced. Specializations of CLIP like PAR-CLIP or iCLIP allow to determine the precise nucleotide position. PAR-CLIP utilizes the by the nucleoside analog induced base transition to track specific nucleotides, whereas iCLIP utilizes the truncation at the cross-linking site to infer the precise nucleotide position. For PCR-amplification, 5' adaptor ligation for iCLIP is necessary and happens via circularization and subsequent linearization. Figure from Konig et al. [211]

specific nucleotide. However, the need for the preincubation of the cells with the nucleoside analogs and their potential toxicity are severe downsides of this method ([213, 211]).

Another, alternative method is the individual nucleotide resolution CLIP (iCLIP) protocol which uses an alternative strategy for adaptor ligation and reverse transcription. While after purification of the protein-RNA complex for both methods the 3' adaptor is ligated, iCLIP does not ligate a 5' adaptor before the reverse transcription. Instead the RNA is reverse transcribed without a 5' adaptor and the key difference is that the cDNA gets truncated at the cross-link site, which allows the identification of the cDNA at the cross-link sites ([214, 215]). However, to allow for PCR amplification and subsequent HT sequencing, the 5' adaptor is appended to the cDNA sequences via circularization and linearization.

Recently, more variations of the CLIP method have been published. For example, eCLIP (enhanced CLIP) improves iCLIP with a better library preparation and replacing the circular adaptor ligation with two separate steps ([216]), or irCLIP (infrared-CLIP) which improves on iCLIP by replacing the radio-labeled adaptor with an infrared-dye-conjugated and biotinylated ligation adaptor ([217]). An overview of CLIP, PAR-CLIP and iCLIP is provided in figure 137 on p. 197.

While HITS-CLIP experiments (and its variants) play an important role in experimentally determining miRNA:target interactions and enabled studying thousands of interactions, one should never forget that the error rate of these experiments is vast compared to traditional low-throughput experiments.

14.3.2 miRNA:target interaction databases

Several groups have lanced projects to collect experimentally verified miRNA:target interactions. Commonly used collections include DIANA Tarbase ([218]), miRTarBase ([219]), starBase ([220]) and miRecords ([221]). DIANA Tarbase v7 is to our knowledge the largest publicly available resource of manually curated miRNA:target interactions. It unifies $\sim 500\,000$ manually curated interactions from experiments on 356 cell types and 24 species. A major advantage of this resource is, that it contains not only miRNA:target interactions, but also annotations of experimental conditions, tissue or cell type and both positively and negatively validated interactions. DIANA Tarbase is thus a valuable ressource to understand the condition-specific effects of miRNA regulation.

14.4 Textmining

A vast number of experimentally verified interactions is hidden in the haystack of biomedical literature. MEDLINE/PubMed contains more than 27 million article references and is rapidly increasing with over 3000 new articles published every day ([222]). In particular, the growth of miRNA:target interaction related publications has been exponential in the recent years (see figure 138). Manual extraction of information from this vast amount of knowledge is no longer feasible and thus automated procedures which can autonomously extract interactions from biomedical text such as PubMed abstracts or PubMed Central fulltexts have been developed. These automatic procedures are known as text-mining.

Various groups have presented approaches for mining miRNA:target interactions from biomedical literature. *BioContext* from Gerner et al. [223] uses their internal tool TextPipe for the identification of gene and protein mentioning and EventMine to detect related entities. Murray et al. [224] extracted miRNA:target relations using a hand-coded list of phrases. Their method identifies (miRNA, verb, gene) triplets, but relies on manual verification as evaluation. *mirSel* from Naeem et al. [225] uses a NER (Named Entity Recognition) approach

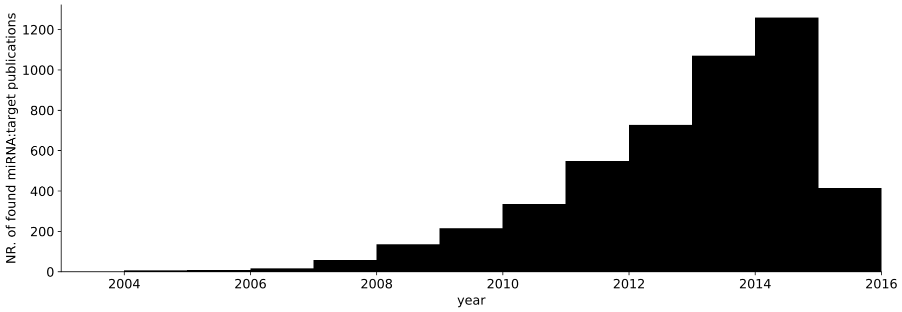


Figure 138 Found PubMed articles with miRNA:target interactions per year. Pubmed articles with miRNA:target interaction show almost an exponential growth since the discovery of miRNAs. The Pubmed abstract snapshot used by `syngrep` was taken in mid 2016.

and detects gene and microRNA synonyms in the same sentence. On a evaluation set of 89 sentences, the authors obtained a F-Score of 0.83. Bagewadi et al. [226] used a manually annotated corpus to a train a binary classifier. Their best F-Score on a manually curated corpus of 301 articles was 0.76. *mirTex* ([227]) uses custom, hand-crafted lexico-syntatic rules to extract miRNA-gene relations. For [226]'s corpus they reached an F-score of 0.87. More recently, Lamurias, Clarke, and Couto [228] developed *IBRel* which uses a sparse multi-instance learning algorithm to reduce the manual effort of creating an annotated corpus. They built a corpus automatically using cooccurrence search via a NER framework, grouped found miRNA:gene relations into (miRNA, gene) bags and labeled these bags binary with interaction existence using as knowledge base all non-human entries from *TransmiR* ([229]) - a database of known relations found in scientific literature. With this labeled set of bags of miRNA:gene relations, the sparse multi-instance learner, which is an optimized algorithm for bags with only a few positive instances, was trained.

Similarly to *mirSel* ([225]) we have implemented a text-mining pipeline approach using a Named Entity Recognition (NER) tool `syngrep` ([230]). A NER framework extracts and classifies entities from a text into pre-defined categories. `syngrep` uses Aho-Corasick trees for efficient string matching of synonyms.

To this end, we created synonym dictionaries for genes and miRNAs by compiling an alias database from various resources (HGNC, miRBase, Uniprot, Ensembl, MGI). These synonym dictionaries are detected in biomedical texts by string matching using `syngrep`.

This technique removes the problems of building hand-crafted rules or relying on a manually annotated corpus. However, there are a some problems that have to be taken into account: (1) human miRBase entries contain a "hsa-" prefix, which is rarely used in the literature, (2) a miRNA:target interaction may be n:m, e.g. "hsa-mir-42 and hsa-mir-45 regulate GEN-1, GEN-2, GEN-3", (3) a miRNA:target interaction may be negative "hsa-mir-42 does not regulate GEN-1", (4) a miRNA:target interaction may be found multiple times in a text, but a cooccurrence within the same sentence is more likely to be a true positive.

To tackle some of these problems, we use a score taking the proximity of a miRNA:gene pair into account. Individual matches within the same sentence receive a high score ($>$ 5), whereas matches in the same paragraph or even fulltext receive a low score ($<$ 1). Consequently, the overall score of a miRNA:gene relation is the sum over all found instance pairs and percentile-rank normalized (see section 14.6.4 for details).

Text-mining results are evaluated and compared to other resources in section 14.6.5.

14.5 miRNAs in atherosclerosis

Atherosclerosis is the underlying cause of heart disease and stroke which together represent the most common cause for death worldwide ([231]). It is a chronic inflammatory disease of the large arteries characterized by the accumulation of lipids and fibrous elements in the blood vessels.
The structure of a large artery is illustrated in figure 139A. During atherosclerosis, plaque is being formed inside the artery wall, eventually blocking the blood flow (see figure 139B). This atherosclerotic plaque forms gradually in multiple stages (reviewed in [232, 233]):

1. Lipoprotein particles circulating in peripheral blood aggregate at the tunica intima (figure 139A and 139B).
2. Monocytes adhere to the endothelial layer, subsequently migrating into the intima (figure 139C).
3. The monocytes proliferate and differentiate into macrophages. The macrophages take up the lipoproteins, forming so-called *foam-cells* (figure 139D).
4. Over time, the foam cells die, filling up the lesion. Additionally, smooth muscle cells migrate from the media into the lesion, secreting fibrous elements. The plaque starts increasing in size, as new mononuclear cells migrate from the blood into the lesion (figure 139E).
5. At a certain point, a thrombus can block the blood flow, and lead to rupture of the lesion. This can cause myocardial infarction or stroke (figure 139F).

Chemokines are small chemotactic cytokines playing an essential role in the recruitment of immune cells to sites of inflammation and in cell regulation and homeostasis. These processes are fundamentally involved during different stages of atherosclerosis ([235]). miRNAs have been shown to mediate the expression of chemokines and thereby modulating the function of endothelial cells, smooth muscle cells and macrophages in atherosclerosis ([236]).
The detailed mechanisms and pathways underlying these regulatory processes are still poorly understood. In the following section, we present a novel resource combining different sources of evidence for miRNA:target interactions. Based on the miRNA-chemokine interactome, we demonstrate, how our resource enables experimental scientists to perform more targeted *in vitro* and *in vivo* assays, leading to an improved understanding of the disease and eventually contributing to novel therapeutic options.

14.6 mirHAHA - A miRNA meta database

In the preceding chapters, we have introduced three dimensions of evidence for miRNA:target:

1. Experimentally verified interactions
2. *In silico* predicted interactions
3. Interactions automatically extracted from biomedical literature.

Unfortunately, this knowledge is scattered over many different, partly contradicting resources (see section 14.6.5). This makes it highly cumbersome for researchers to study miRNA:target interactions. We therefore believe that the scientific community would vastly benefit from a resource combining all evidence in one place, conveniently accessibly through a graphical user interface.
To our knowledge, there is only one resource so far addressing this issue to a certain extent. COGERE ([196]) is a meta-database for miRNA:target and transcription-factor:target

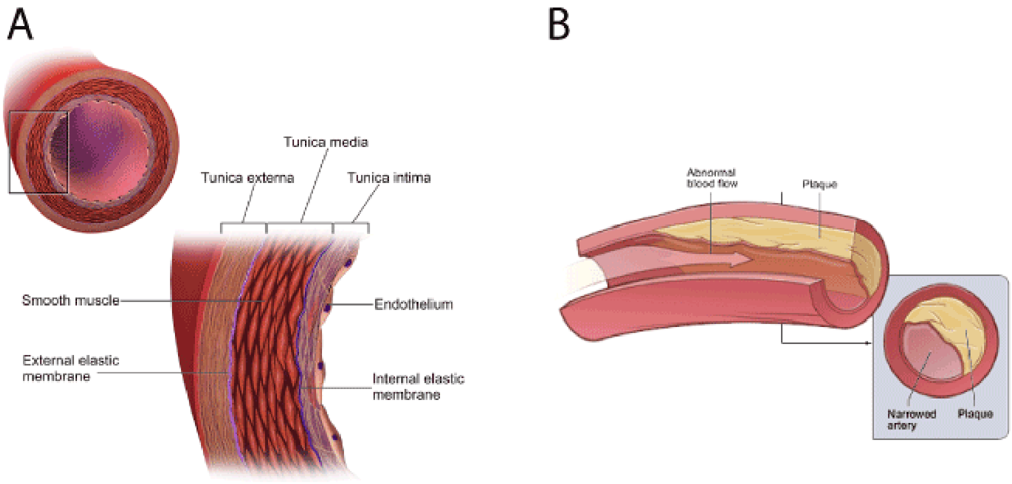


Figure 139 (A) **Structure of a normal large artery**. The innermost layer (*intima*) is a very thin layer made up of a sheet of elastic fibres bounded by a layer of endothelial cells on the luminal side. The next layer (*media*) consists of smooth muscle cells. The outermost layer (*adventia* or *externa*) is composed of connective tissues. (B) Atherosclerosis leads to plaque formation, disturbing the blood flow. Eventually a thrombus can lead to occlusion causing stroke or myocardial infarction. Figure adapted from Wikimedia figures[a,b]

[a] https://commons.wikimedia.org/wiki/File:Blausen_0055_ArteryWallStructure.png

[b] https://commons.wikimedia.org/wiki/File:Atherosclerosis_diagram.png

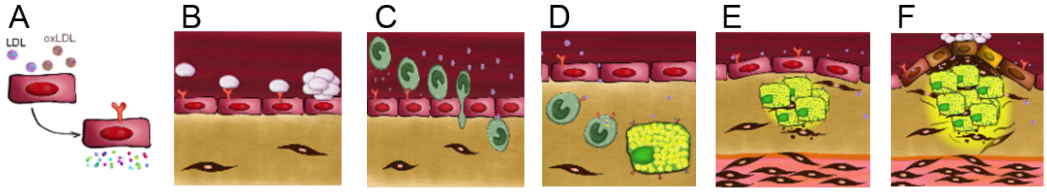


Figure 140 Atherosclerotic plaque formation. (A, B) Lipoprotein particles circulating in peripheral blood aggregate at the tunica intima. (C) Monocytes adhere to the endothelial layer, subsequently migrating into the intima. (D) The monocytes proliferate and differentiate into macrophages. The macrophages take up the lipoproteins, forming so-called *foam-cells*. (E) Over time, the foam cells die, filling up the lesion. Additionally, smooth muscle cells migrate from the media into the lesion, secreting fibrous elements. The plaque starts increasing in size, as new mononuclear cells migrate from the blood into the lesion. (F) At a certain point, a thrombus can block the blood flow, and lead to rupture of the lesion. This can cause myocardial infarction or stroke. Figure adapted from supplementary material of Talikka, Boue, and Schlage [234]

interactions. This is already an excellent resource integrating evidence from seven miRNA-target prediction algorithms, three text-mining tools and five databases for experimentally verified interactions into a 'prior network'. COGERE makes use of this network to perform condition-specific scoring of its interactions, based on user-provided experimental data.
Unfortunately, COGERE does not provide a graphical user interface allowing to explore the network interactively, nor is it possible to drill-down interactions to their origin for manual inspection. Instead, all sources are combined into a single 'prior-score' making it hard to assess the reliability of an interaction.
Additionally, there is GeneMANIA ([237, 238]), which implements a well-designed web-based user-interface for exploring biological networks. GeneMANIA allows to analyze gene-gene interaction networks from various data sets (predicted, physical interaction, co-expression, co-localization, pathway, shared protein domains, and genetic interactions). However, the main focus of GeneMANIA is to find functionally similar genes resource and it is thus limited to gene-gene interactions and therefore not suitable for examining miRNA:target interactions.

14.6.1 Web app

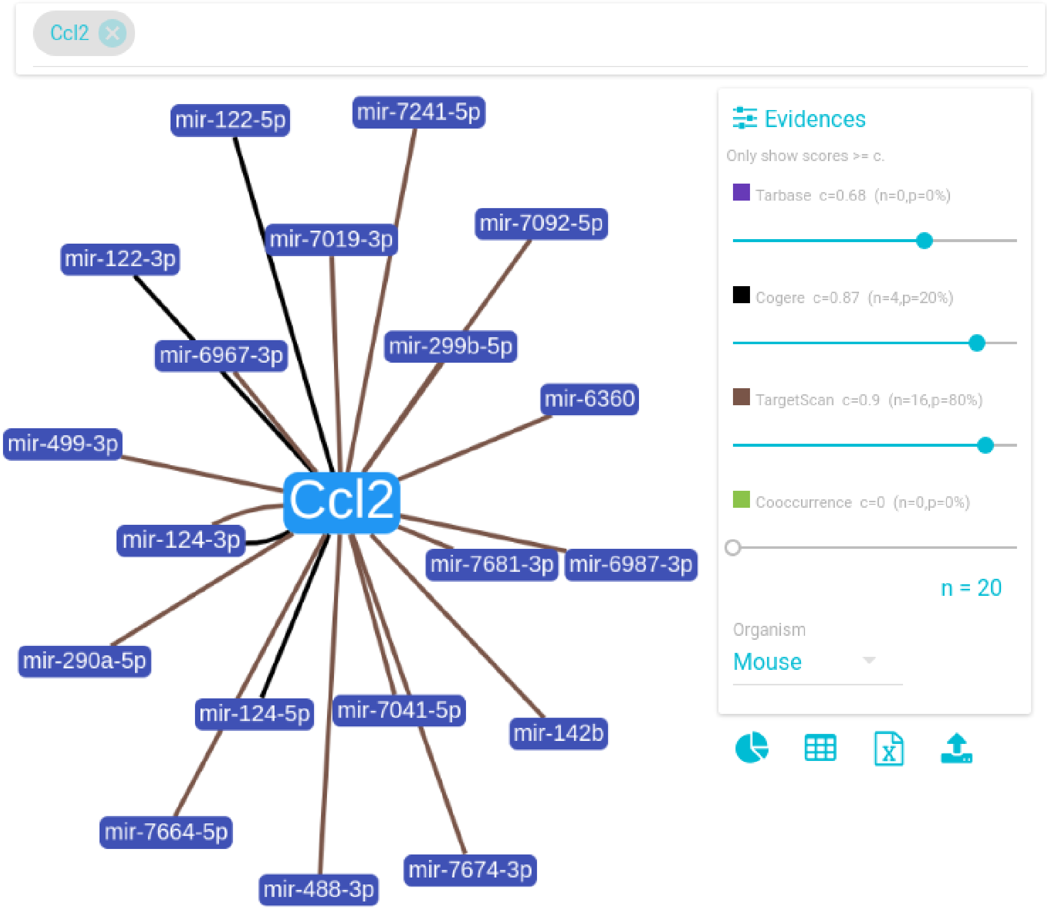
To overcome these limitations, we built mirHAHA (**miR**NA-specific **H**uge **A**ggregation of **H**eterogeneous **A**nnotations), a freely available web application. Our modern, user-friendly web interface allows to inspect miRNA:target interactions visually and interactively as a network of nodes (miRNAs or genes respectively) and edges (interactions) (see figure 141 on p. 203). Additional information for each node and interaction is available upon click on the corresponding item. Interactions can be filtered by separate scores for each evidence in real-time.
Additionally, we provide functionality to perform Gene Ontology (GO) term enrichment analysis and allow to visualize the results within the network graph (see figure 142 on p. 204) Addressing that miRNA:target interactions are highly context-specific ([205]), user-provided experimental data can be loaded into the web app and projected onto the network (see section 14.6.3 for more details). User-provided data is analyzed directly locally in the web app and not sent to our backend, such that no privacy issue arise and sensitive, confidential data, such as patient data, can be used without any concerns.
Lastly, to facilitate sharing of interactive networks, with each change of parameters, the browser URL is updated to encode the current state of the network. Opening the URL on a different device will reproduce exactly the same graph as seen by the current user.
We have performed several optimizations to provide a fast interface. For example, all identifiers are cached locally in IndexedDB to avoid repetitive downloads. As an additional benefit, this allows to use trie search trees ([239], chapter 8) for quickly finding related identifiers during auto-completion. The graph itself is rendered with Cytoscape.js ([240]) using a customized variant of the CoSE layout algorithm.

14.6.2 Availability

The web interface is freely accessible on `https://neap.bio.sh`. Both data and source code for the web server and all analyses performed are freely available from the GitHub repositories `wilzbach/neap-server`[1] and `wilzbach/neap-analyses`[2] respectively. Information

[1] `https://github.com/wilzbach/neap-server`

[2] `https://github.com/wilzbach/neap-analyses`



Evidences

Only show scores >= c.

🟪 Tarbase c=0.68 (n=0,p=0%)

■ Cogere c=0.87 (n=4,p=20%)

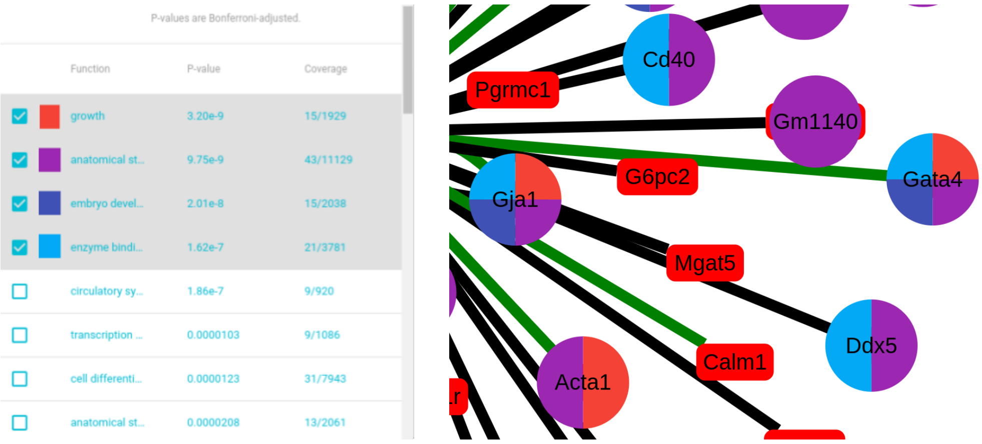
🟫 TargetScan c=0.9 (n=16,p=80%)

🟩 Cooccurrence c=0 (n=0,p=0%)

n = 20

Organism
Mouse

Figure 141 mirHAHA – feature overview. mirHAHA is a modern graph-based web app for systematically exploring miRNA:target interactions. The graph in the center represents the miRNA:target interactions, where dark blue nodes stand for mirRNAs and light blue nodes for targets genes. Edges connecting these nodes depict miRNA:target interactions from our database whereby the coloring is representative for the evidence source of the respective interaction. The size of a node corresponds to its degree. A user-friendly auto-completion assisted input box on top allows to conveniently add new identifiers (miRNA or target) to the graph. The graph is zoomable, draggable and for individual nodes a meta information box with more information (e.g. description text, aliases, links to external database, PMIDs, stem loop of the respective miRNA etc.) is available on click. At the evidences box on the right, cut-off score sliders for each evidences can be interactively controlled by the user to filter insignificant edges and transform a 'hairy ball' into a clear graph. With the score changing, for each evidence the number of edges displayed in the graph n, the relative size compared to all edges in the graph p and the total number of edges in the graph is updated as well. The organism select box can be used to switch between interactions from human and mouse respectively. The buttons below the evidence box allow the user to (1) show the GO coloring box (see figure 142 for details), (2) export the current interaction set as CSV file, (3) export the current interaction set as Excel worksheet, (4) upload custom, condition- or cell-type specific experimental data (see section 14.6.3 for more details). Data can also be conveniently uploaded using drag&drop.



P-values are Bonferroni-adjusted.

		Function	P-value	Coverage
☑	🟥	growth	3.20e-9	15/1929
☑	🟪	anatomical st...	9.75e-9	43/11129
☑	🟦	embryo devel...	2.01e-8	15/2038
☑	🟦	enzyme bindi...	1.62e-7	21/3781
□		circulatory sy...	1.86e-7	9/920
□		transcription ...	0.0000103	9/1086
□		cell differenti...	0.0000123	31/7943
□		anatomical st...	0.0000208	13/2061

Figure 142 mirHAHA - Gene Ontology (GO) coloring. An expandable GO box allows projecting biological information onto nodes in the mirHAHA miRNA:target interaction graph. In the GO box (left panel), GO terms are sorted after Bonferroni-adjusted p-values and can be selected by the user. P-values are derived from a 2x2 contingency matrix using Pearson's chi-squared test. For all selected GO terms, the respective nodes will be colored. For examples, all nodes in the right panel which are related to cell growth (Gja1, Acta1, Gata4) are partially colored in red and all nodes which are related to enzyme binding are colored in light blue (Gja1, Cd40, Gata4, Ddx5).

on how to build and execute our software are documented in the repositories' READMEs and Makefiles. For maximal reproducibility, all analyses are written in self-explanatory notebooks in *literate programming style* ([241]). Our analysis pipeline is outlined in figure 143.

14.6.3 Condition-specific scoring

Erhard et al. [205] have demonstrated, that miRNA:target interactions are highly context-specific and vary between tissue, cell-type and medical conditions. Since the interaction networks from the different sources do not provide context information (except, to a certain extent, DIANA Tarbase v7) these networks cannot be considered universal. Rather, they provide a *hypothesis space* of potential interactions.

Therefore, it is highly relevant to project condition-specific experimental data onto that network. We have shown in section 14.6.1 that our web app supports the upload of custom differential gene expression data. The main advantage of using the network as hypothesis space in comparison to a classical differential expression experiment is, that the number of tested hypotheses is significantly reduced, therefore gaining statistical power. After uploading the data, our web app will automatically determine conditionally active interactions and calculate FDR-adjusted p-values based on the reduced hypothesis space.

Under the assumption, that miRNAs downregulate their target genes, we have two options to confirm an interaction: either (1) the miRNA is differentially up-regulated and the gene down-regulated (shown in green in the web app) or (2) a miRNA is down-regulated and the gene up-regulated (shown in orange). We display interactions where both miRNA and gene are differentially expressed at a FDR < 0.2 (see figure 149). The full list of interactions and the corresponding p-values is available through the CSV or Excel worksheet export (see figure 141).

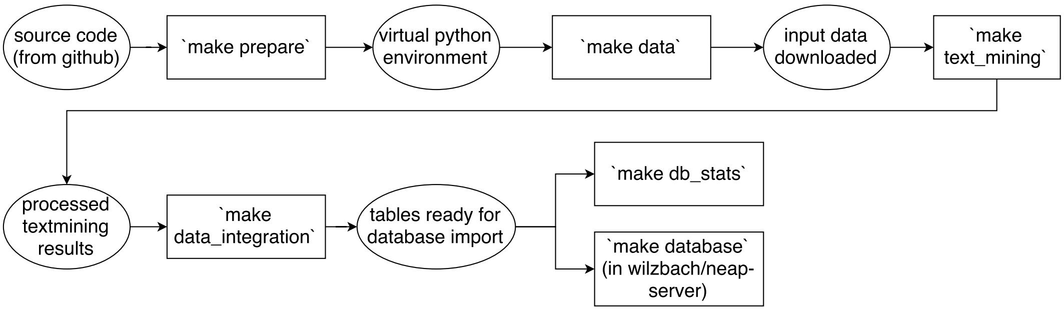

source code (from github)
`make prepare`
virtual python environment
`make data`
input data downloaded
`make text_mining`
processed textmining results
`make data_integration`
tables ready for database import
`make db_stats`
`make database` (in wilzbach/neap-server)

Figure 143 Simplified overview of our analysis pipeline. Invoking the `make` command from our source code will (1) setup a virtual environment with all dependencies, (2) download the input data from our web server, (3) process the text-mining results, (4) preprocess all sources of evidence to a common format and (5) generate statistics on these tables and insert the tables into the backend of the mirHAHA server. The high-level Makefile target consists of many subordinate Makefile targets which have been omitted for clarity.

14.6.4 Data integration

For the scope of this project, we have chosen one representative resource for each of the three dimensions of evidence. We integrated (1) experimentally verified interactions from DIANA Tarbase v7, (2) computationally predicted interactions from Targetscan 7.1 and (3) interactions automatically extracted from biomedical literature as described in in section 14.4. Despite of its limitations regarding usability, we consider COGERE a valuable source of information and integrated this meta database as a 'fourth dimension' into our resource. As mirHAHA is implemented in a highly generic fashion, adding more resources for each dimension of evidences is easily possible.

Except for DIANA Tarbase, we obtained the data from the databases' web portals. Unfortunately, DIANA Tarbase is not available for download. Since our request to access the full dataset has not been answered until completion of this project, we used a web crawler to extract the interactions from their web portal. Due to technical limitations, we could only obtain a subset ($\sim 15\,\%$) of all available interactions.

For each resource, we applied a percentile-rank normalization to its scores according to

$$s_{i_n} = \text{rank}(s_i)/N$$

where s_{i_n} is the normalized score for interaction i, s_i is the raw score of interaction i from the resource and N is the number of interactions contained in the resource.

The percentile rank normalization ensures comparability of the scores between the different resources and provides an intuitive interpretation of the scores: an interaction with score of *e.g.* $50\,\%$ has a higher confidence than half of all interactions.

14.6.5 Database statistics

Table 10 shows an overview of the number of unique miRNAs, genes and interactions contained in each database. Combining the four resources we have roughly 13M interactions in our database, with COGERE and Targetscan providing most ($\sim 92\,\%$) of the interactions. However, as both of these resources contain *in silico* predicted interactions, the vast majority of these interactions is of low reliability. Filtering by score or supporting these interactions with additional evidence is therefore indispensable. The number of experimentally verified interactions from DIANA Tarbase or text-mining is significantly lower and interactions are available only for a subset of miRNAs.

Table 10 Number of entities and interactions for each source of evidence. The numbers are given for human (hsa) and mouse (mmu) separately.

		miRNAs		genes		interactions	
	evidence	hsa	mmu	hsa	mmu	hsa	mmu
1	cogere	1312	1061	17981	17758	3868213	2415731
2	cooccurrence	978	737	2279	2359	46343	49152
3	diana_tarbase7	610	355	9555	4828	47834	18256
4	targetscan	2469	1799	18088	19971	7220438	6224074

Moreover, we were interested in the distribution of the number of genes regulated by a miRNA and vice versa. Without filtering by score, both COGERE and TargetScan predict on average an interaction of a single miRNA with multiple thousands of genes with a range from less than ten genes per miRNA to more than 10 000. Limiting the predicted interactions to high-confidence scores (> 0.95) only, the average number of interactions for a single miRNA is, comparable the text-mining and DIANA Tarbase around 100 (see figure 144). The median number of miRNAs predicted to regulate a single gene is in the range of hundreds for the two prediction algorithms with an arbitrary score and less than 10 for text-mining and DIANA Tarbase. Again, using high-confidence interactions only, the number of interactions predicted by COGERE and TargetScan converges to those of the two other methods. Interestingly, for all four resources, some outliers exist that do regulate hundreds of genes. One such example is the well-studied, highly-conserved miR-155 which is involved amongst others in hematopoiesis and immune cell regulation ([242]).
A comparison of the overlaps between each resource is illustrated in figure 145 as a venn-diagram. Curiously, the overlap of all four resources is surprisingly little: Only 148 of all interactions are contained in all four evidences. However, there is a substantial amount of interactions which is backed by two or three of the resources. For instance, we found $\sim 13\,500$ interactions with text-mining which are also backed by COGERE and ~ 7500 interactions which are also predicted by TargetScan. DIANA Tarbase contains $\sim 15\,000$ interactions that are also predicted by TargetScan. The large overlap of TargetScan and COGERE is not surprising, as an earlier version of TargetScan is integrated into that meta database. Nor is is surprising that COGERE contains more than 2.5 million distinct interactions, as it integrates amongst others six other prediction algorithms. Curiously, we find $\sim 25\,000$ interactions with text-mining which are not contained in any of the other resources. While a substantial amount of these edges are likely false positives that arise due to limitations of the NER-approach (see section 14.4), manual inspection on a subset of these edges could validate novel true-positive interactions.

14.7 Case Study: mirHAHA applied to the miRNA:Chemokine interactome in the scope of atherosclerosis

Hartmann, Schober, and Weber [236] published a figure of a miRNA:chemokine interactome in their paper about miRNAs in atherosclerosis. The network contains four Chemokine Receptors, 15 Chemokines and 110 miRNAs from mouse. The entities are connected with 146 edges. $\sim 25\,\%$ of all miRNAs have a degree ≥ 2. The median degree of a gene is 4, ranging up to 49 (Cxcl12). The network was constructed manually based on experimentally

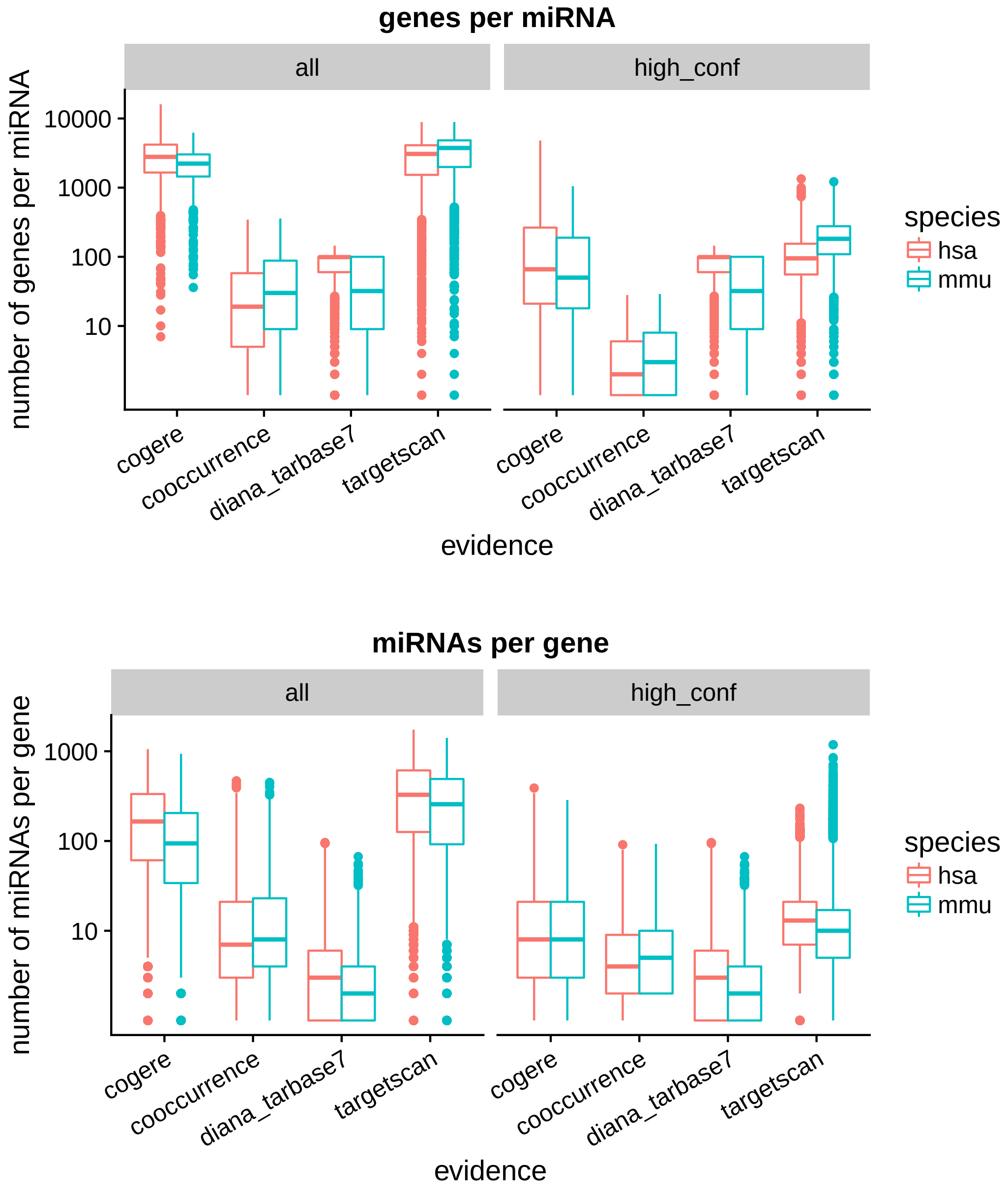


Figure 144 Upper panel: distribution of the number of genes per miRNA for each resource. Lower panel: distribution of the number of miRNAs per gene for each resource. High confidence only includes interactions with a score > 0.95.

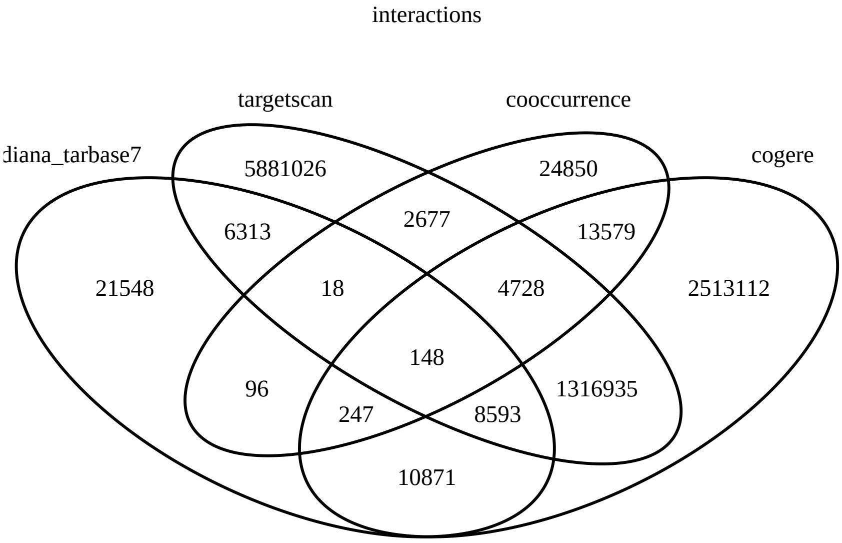


Figure 145 Venn diagram of the overlap of interactions contained in each of the four resources TargetScan (*in silico* predicted), DIANA Tarbase (experimentally verified), COGERE (meta database) and cooccurrence (text-mining).

verified interactions from DIANA Tarbase v7.0.

Using our tool, we reproduced an updated, interactive version of this figure (see figure 147 or the interactive online version[3]). To this end, we inserted all identifiers of the chemokine and chemokine receptor genes into the mirHAHA web portal. While the layout of both graphs is clearly different and many interactions only occur in one graph, some similarities can immediately be detected. For example, Cxcl12 is the largest hub in both graphs. However, a comparison of the network topology is not straight-forward as it highly depends on the cutoff-scores. For this reason, we made a more objective comparison taking different scores into account (see figure 148).

At a cutoff of 0.8, we find 68 (46 %) of the interactions in the Hartmann-graph in our network. But, with this cutoff we also find 2063 additional edges, not contained in the original network. As the vast majority of these edges is contributed by predicted interactions from COGERE and Targetscan, lowering the corresponding cutoffs is advisable. Lowering the cutoff to 0.99 our network has a comparable size of 154 interactions. At this cutoff, 25 edges (17 %) of the Hartmann-graph appear in our network.

Even when no cutoff is applied, our graph does not contain all interactions from the Hartmann-graph. This is due the fact that the Hartmann-graph was generated manually on the full version of DIANA, which we were not able to integrate in our resource (see section 14.6.5). The discrepancy is therefore due to interactions, which have been validated using (mostly high-throughput) experiments, but have not been found using text-mining nor been predicted using TargetScan or one of the prediction algorithms integrated in COGERE.

As both prediction algorithms and high-throughput experiments yield a high number of false

[3] an interactive version of this figure is available from `https://neap.bio.sh/organism/mmu/ids/Ccl2,Ccl22,Ccl3,Ccl4,Ccl7,Ccl9,Ccr5,Ccr7,Cx3cl1,Cxcl1,Cxcl10,Cxcl12,Cxcl13,Cxcl14,Cxcl5,Ppbp,Cxcl9,Cxcr2,Cxcr4?scores=diana_tarbase7:0,cogere:0.93,targetscan:1,cooccurrence:0.2`

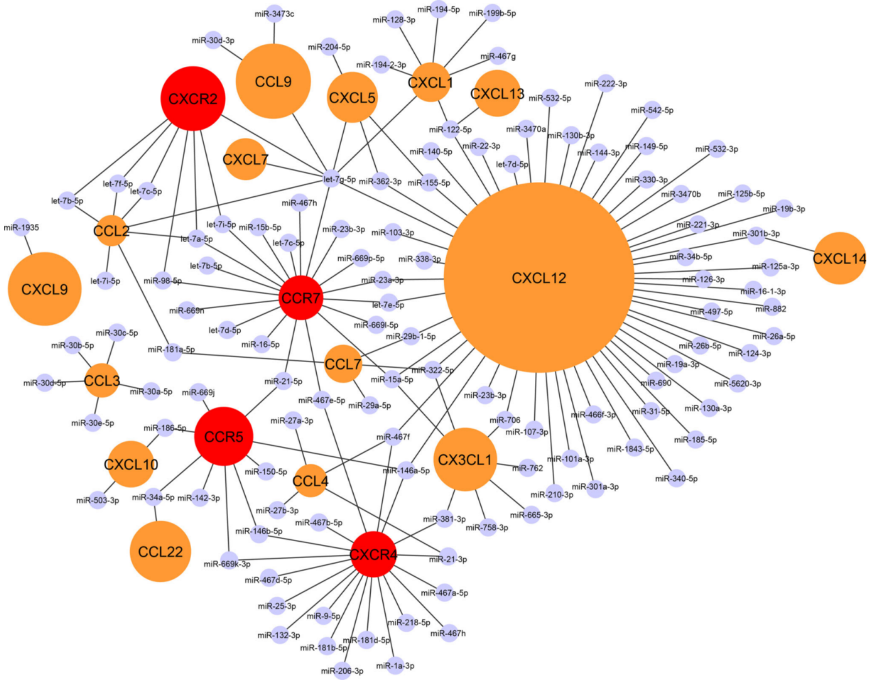


Figure 146 The chemokine:miRNA interactome based on experimental data from DIANA Tarbase v7.0 from [236]. Blue circles represent miRNAs, orange circles chemokines and red circles chemokine receptors. Black lines indicate an interaction between two nodes. The size of a gene-node corresponds to the degree of the node.

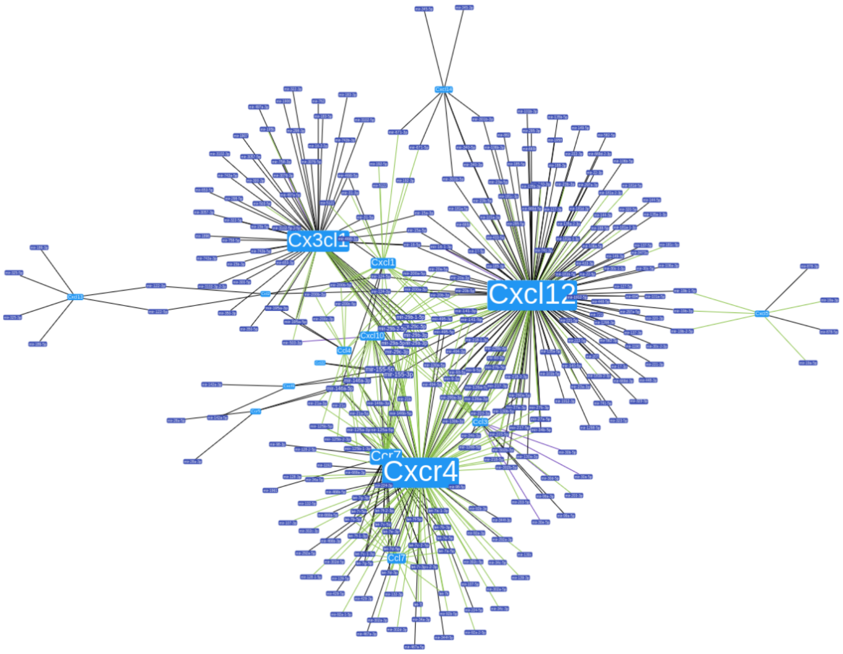


Figure 147 An extended version of the above miRNA-chemokine interactome generated with mirHAHA. Light blue boxes correspond to genes, dark blue boxes to miRNAs. Edges represent interactions and are colored according to the evidence (green: cooccurrence, black: COGERE, purple: DIANA Tarbase, brown: TargetScan). For this figure, edges were filtered with the following cutoff-scores: DIANA Tarbase: 0.0, COGERE: 0.93, TargetScan: 1.0, Textmining: 0.2. An interactive version of this figure is available online at `https://neap.bio.sh/organism/mmu/ids/Ccl2,Ccl22,Ccl3,Ccl4,Ccl7,Ccl9,Ccr5,Ccr7,Cx3cl1,Cxcl1,Cxcl10,Cxcl12,Cxcl13,Cxcl14,Cxcl5,Ppbp,Cxcl9,Cxcr2,Cxcr4?scores=diana_tarbase7:0,cogere:0.93,targetscan:1,cooccurrence:0.2`

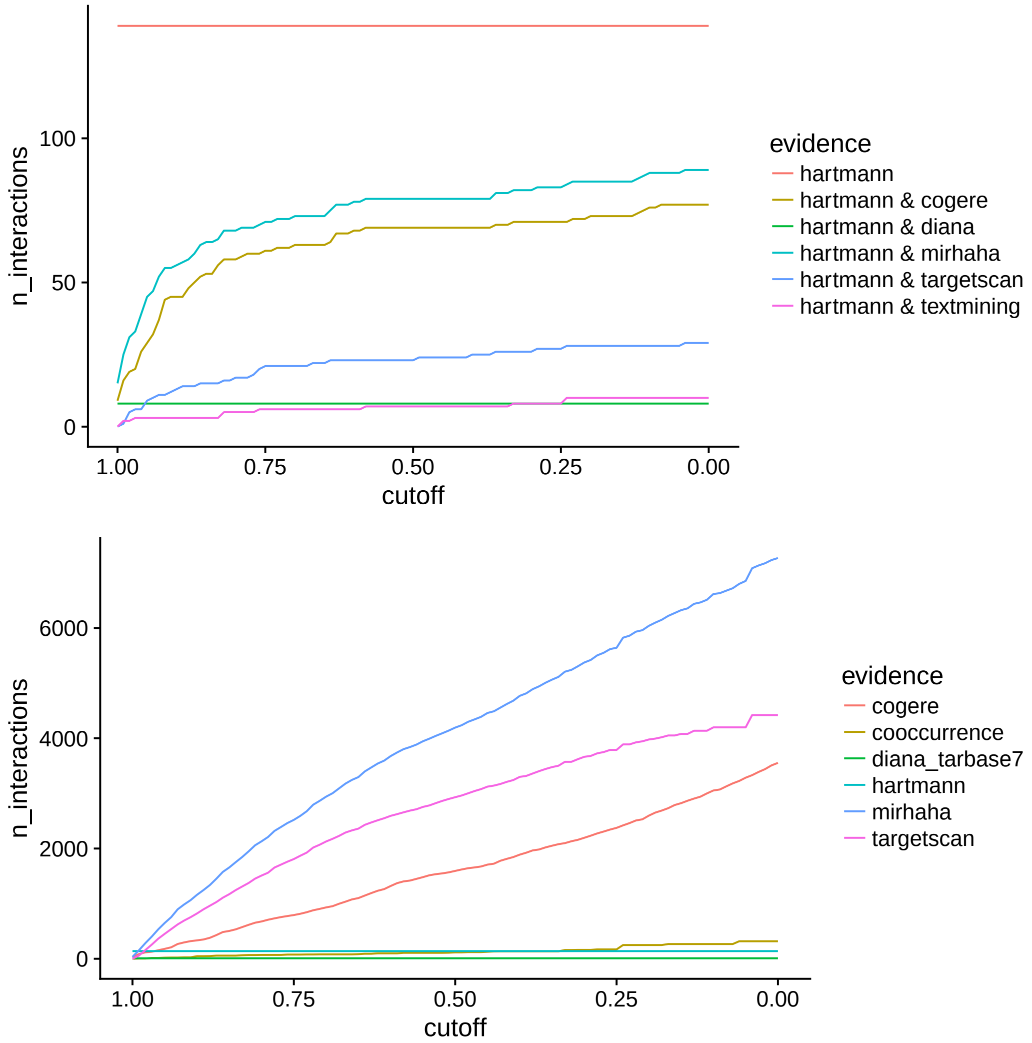
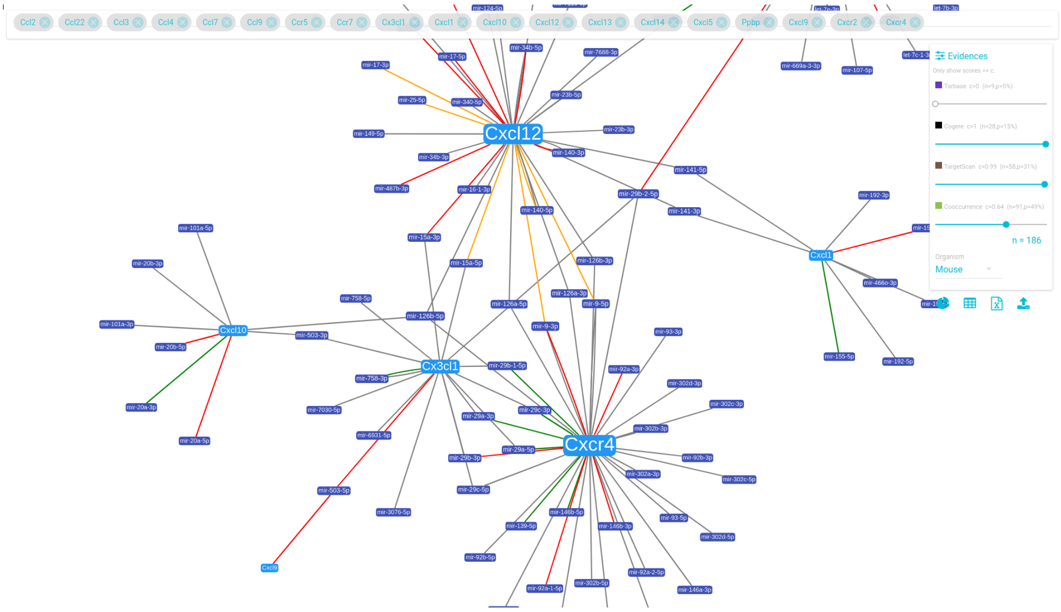


Figure 148 Comparison of the interactions in the chemokine interactome by Hartmann, Schober, and Weber [236] and the version generated with mirHAHA. Upper panel: Intersection of mirHAHA with the network by Hartmann, Schober, and Weber [236] for different cutoff scores. The orange line represents the number of interactions in the Hartmann-graph (n=144), the cyan line the aggregate interactions from mirHAHA. The other lines correspond to the individual evidences included in mirHAHA. Lower panel: The total number of interactions found in mirHAHA (blue line) and all individual evidences (other lines).



Evidences

Only show scores >= c.

Tarbase c=0 (n=9,p=5%)

Cogere c=1 (n=28,p=15%)

TargetScan c=0.99 (n=58,p=31%)

Cooccurrence c=0.64 (n=91,p=49%)

n = 186

Organism

Mouse

Figure 149 mirHAHA - Context specific scoring. The Chemokine interactome network enriched with differential expression data from a differential expression experiment. Native HDL has been transfected into human coronary artery cells and compared to control samples. The experiments are available from GEO accession numbers GSE53201 and GSE53315. If an edge is backed with a differential expression of both miRNA and target gene at a FDR of < 0.2 the edges are colored as follows: green if the miRNA has down-regulated, and the gene up-regulated, orange if the miRNA is up-regulated, and the gene down-regulated, red otherwise. The underlying network graph is available at `https://neap.bio.sh/organism/mmu/ids/Ccl2,Ccl22,Ccl3,Ccl4,Ccl7,Ccl9,Ccr5,Ccr7,Cx3cl1,Cxcl1,Cxcl10,Cxcl12,Cxcl13,Cxcl14,Cxcl5,Ppbp,Ccl9,Cxcr2,Cxcr4?scores=diana_tarbase7:0,cogere:1,targetscan:0.99,cooccurrence:0.64`

positives, particularly the edges backed by multiple evidences are the most interesting.

14.7.1 Enriching the graph with condition-specific experimental data

Given that (1) the resources tend to contain a high number of false positives and (2) miRNA regulation is context and cell-type specific ([205]), integrating condition-specific experimental data into the network is highly relevant. In section 14.6.3 we showed that our web app allows the user to upload custom experimental data from differential gene expression experiments and project it onto the network.

[243] transfected native high density lipoprotein (HDL) into human coronary artery cells and measured gene expression for both mRNA and microRNAs. We obtained the experimental data from GEO accession numbers GSE53315 and GSE53201 respectively. We performed a differential expression analysis of HDL-transfected samples versus control using the online-tool GEO2R ([244]). To project the data on the chemokine-interactome based on mouse-data, we mapped the human identifiers to the corresponding mouse orthologs using Ensemble Biomart. We loaded the results of the analysis into the mirHAHA app and obtained the edges the are supported with the experimental data (see figure 149). Performing an explorative analysis, we chose a relatively high FDR-cutoff of 0.1.

Given these cutoff, we found two miRNAs play an important role in the miRNA:chemokine interactome (see table 11): mmu-mir-223 regulates Cxcr4, Ccl3, Cx3cl1, Cxcl10 and mmu-mir-541 regulates Cxcl13, Cxcl5, Cx3cl1. We consider these interactions as novel candidates for treatment options and highly recommend to validate these interactions using a high-accuracy method such as qPCR.

Table 11 Condition-specific regulatory interactions

miRNA	gene	logFC miRNA	logFC gene	FDR miRNA	FDR gene
mmu-mir-223-3p	Cxcr4	4.045561	-0.09460	0.001972	0.004555
mmu-mir-223-3p	Ccl3	4.045561	-0.00792	0.001972	0.007766
mmu-mir-223-3p	Cx3cl1	4.045561	-0.47100	0.001972	0.001397
mmu-mir-223-3p	Cxcl10	4.045561	-1.98000	0.001972	0.000081
mmu-mir-541-3p	Cxcl13	5.096432	-0.03800	0.081188	0.006365
mmu-mir-541-3p	Cxcl5	5.096432	-0.19400	0.081188	0.002045
mmu-mir-541-3p	Cx3cl1	5.096432	-0.47100	0.081188	0.001397

14.8 Conclusions

The detailed mechanisms and pathways of miRNA-mediated gene regulation in disease is still poorly understood and is a major topic of ongoing research. Much effort has been made to provide miRNA:target interaction data, based on experiments, text-mining and *in silico* predictions. Unfortunately, this knowledge is scattered over many different resources and is partly contradicting, making it cumbersome for researchers to study. In this report, we presented mirHAHA, an interactive web-based application consolidating miRNA:target interactions at one place, conveniently and consistently accessible through a graphical user interface. Combining the evidence with user-provided condition-specific experimental data, addresses recent findings, that miRNA-mediated gene regulation is highly context-sensitive. At the example of atherosclerosis we demonstrate, how mirHAHA enables researchers to gain a systematic overview of the landscape of miRNA:target interactions. We firmly believe that our resource will enable scientists to perform more targeted *in vitro* and *in vivo* assays and contribute to an improved understanding of the disease, eventually contributing to novel therapeutic options.

Future improvements to our tool include the integration of more sources of evidence. In particular integrating the full version of DIANA Tarbase v7 would help assessing the confidence of interactions tremendously. Moreover, sensitivity would benefit from making use of the characteristics of different target-prediction algorithms. Finally, the user experience could benefit from implementing more filtering options, such as displaying only interactions that are contained in at least two resources.

References

[100] J. Merkin et al. “Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues”. In: *Science* 338.6114 (2012), pp. 1593–1599. ISSN: 0036-8075. DOI: 10.1126/science.1228186. arXiv: NIHMS150003. URL: http://www.ncbi.nlm.nih.gov/pubmed/23258891http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3568499http://www.sciencemag.org/cgi/doi/10.1126/science.1228186.

[115] Andrew Yates et al. “Ensembl 2016.” In: *Nucleic acids research* 44.D1 (2016), pp. D710–6. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1157. arXiv: arXiv:1011.1669v3. URL: http://www.ncbi.nlm.nih.gov/pubmed/26687719http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4702834http://www.ncbi.nlm.nih.gov/pubmed/26687719{\%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4702834.

[117] *Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information.* 2004. URL: https://www.ncbi.nlm.nih.gov/gene/ (visited on 06/04/2017).

[118] Simon Penel et al. “Databases of homologous gene families for comparative genomics”. In: *BMC Bioinformatics* 10.Suppl 6 (2009), S3. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-S6-S3. URL: http://www.biomedcentral.com/1471-2105/10/S6/S3.

[119] Jakub O Westholm and Eric C Lai. *Mirtrons: MicroRNA biogenesis via splicing.* 2011. DOI: 10.1016/j.biochi.2011.06.017. arXiv: NIHMS150003. URL: http://www.ncbi.nlm.nih.gov/pubmed/21712066http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3185189.

[120] David Brawand et al. “The evolution of gene expression levels in mammalian organs”. In: *Nature* 478.7369 (2011), pp. 343–348. ISSN: 0028-0836. DOI: 10.1038/nature10532. URL: http://www.nature.com/doifinder/10.1038/nature10532.

[121] Heng Li et al. “The Sequence Alignment/Map format and SAMtools”. In: *Bioinformatics* 25.16 (2009), pp. 2078–2079. ISSN: 13674803. DOI: 10.1093/bioinformatics/btp352. arXiv: 1006.1266v2. URL: http://www.ncbi.nlm.nih.gov/pubmed/19505943http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2723002.

[122] Aaron R Quinlan and Ira M Hall. “BEDTools: A flexible suite of utilities for comparing genomic features”. In: *Bioinformatics* 26.6 (2010), pp. 841–842. ISSN: 13674803. DOI: 10.1093/bioinformatics/btq033. URL: http://www.ncbi.nlm.nih.gov/pubmed/20110278http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2832824.

15 3D DNA Interactions

by Quirin Heiß, Carsten Uhlig and Alexander Grün

15.1 Introduction

In addition to the one-dimensional organization of nucleotides on the DNA strand, the human genome is organized in a three-dimensional way. This means that the linear distance of elements on the DNA strand does not necessarily correspond to the actual three-dimensional distance. Thus, regulatory elements (e.g. enhancers) can come into physical contact with their targets, which lie at very distant positions on the linear sequence. Understanding the influence of spatial chromosome organization on gene activity is important for the understanding of genomic processes such as gene expression.
Since microscopy is not sufficient to capture these processes at a high quality [245], research of genome organization relies on experimental approaches.
In recent years, several methods to detect chromosomal interactions have been developed. While chromosome conformation capture (3C) [246] is able to analyze interactions of two known loci, the recently developed Hi-C [247] method allows to capture all chromatin interactions on a genome-wide scale. The amount of data from these experiments is rapidly increasing, raising the need for efficient tools to process and analyze raw data automatically.

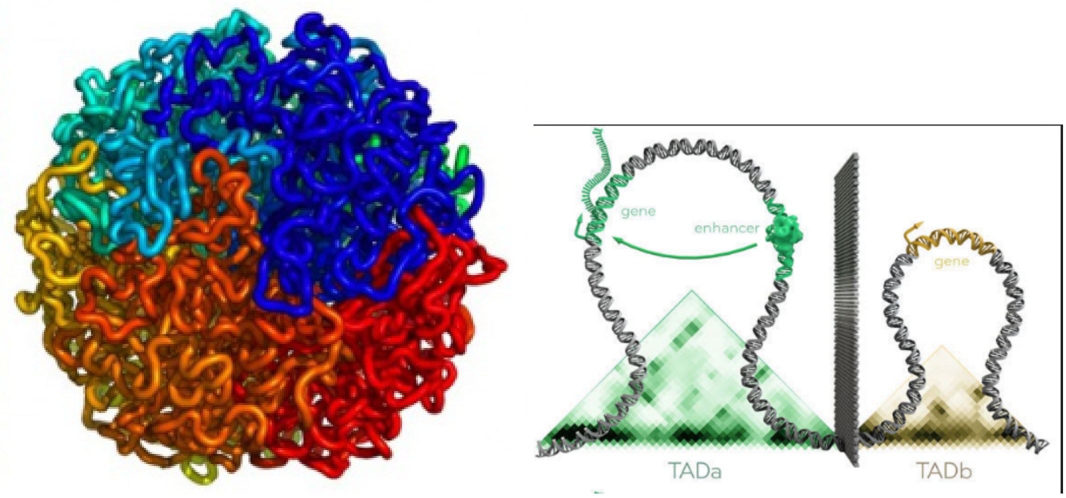


Figure 150 Schematic examples of three-dimensional chromatin organization - chromosome and TAD

Analysis of the contact data resulting from Hi-C experiments has shown the existence of chromosome territories where chromatin forms loops. This results in a compartmentalization of the genome into so-called Topologically Associated Domains (TAD). Contact of functional elements within a TAD is favored, and contacts between different compartments are suppressed [248]. This makes knowledge of genome compartmentalization useful for biomedicine and clinical research, since defective TAD boundaries can result in contact of elements (e.g. enhancers and oncogene) that allows pathogenic phenotypes to develop. Chromatin conformation capture has shown that deviations in chromatin organization play a role in limb malformations [249] and leukemia [250].

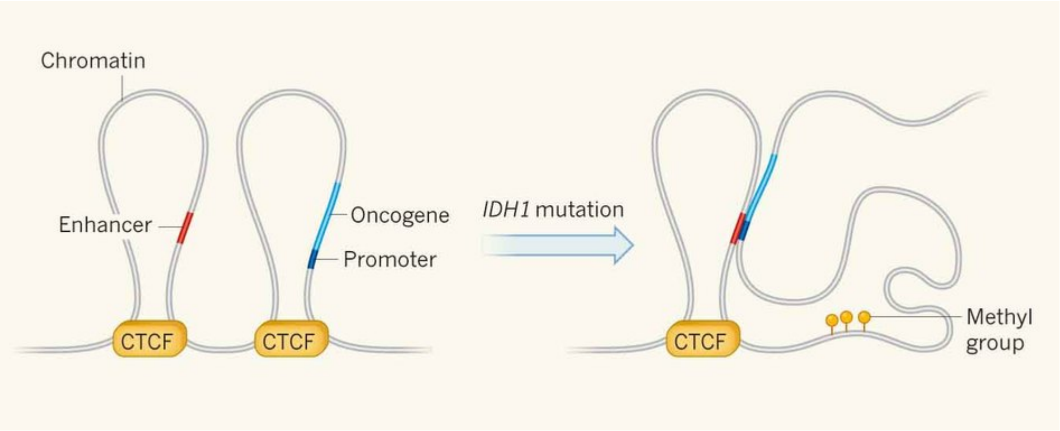


Figure 151 Comparison between normal (left) and mutated TAD boundaries. After boundary removal, the oncogene and enhancer come into contact, leading to cancerogenesis.

Analyses of contact matrices (that display interaction density between genomic regions, partitioned as a grid) have shown that the human genome is organized in a less dense compartment containing open, accessible chromatin and a densely packed compartment containing closed, inaccessible and inactive chromatin regions [251]. Furthermore, conserved CTCF sites are often enriched at strong TAD borders [252]. This raises interest for a combined analysis of chromatin binding and 3D contact data.

15.1.1 Hi-C and 3C

Different methods are used for this task; they are separated according to their scope: While 3C methods investigate only interactions between two pre-specified genomic loci, Hi-C (High throughput chromatin conformation capture) methods allow for genome-wide analysis of 3D interactions.
The procedure for Hi-C experiments, depicted in Figure 3, is as follows [251]:
First, cells are fixed with formaldehyde so that interacting loci are bound by covalent cross-links. Then, the DNA is fragmented with restriction enzymes and only the fixed loci stay linked. After the 5' ends are filled and marked with biotin, the fragments are ligated under conditions that favor the connection of cross-linked DNA fragments. The ligation products are marked with biotin at the junction site, so that marked junctions inside fragments can later be identified by streptavidin-coated magnetic beads. Then, the DNA fragments are purified and sheared and the junctions are isolated. The resulting fragments can be analyzed by a high-throughput sequencer, producing a catalog of interacting fragments.
Further analysis of the resulting data can give insight in the organization of the human genome: The data can be used to compute a contact matrix showing interactions on a genome-wide scale. Those matrices show the large-scale organization of the human genome into Topology Associated Domains, colloquially known as genome neighborhoods, wherein physical interactions occur frequently.

15.2 Hi-C Analysis Pipeline

To analyze raw paired-end reads generated with Hi-C experiments several steps are necessary. All Hi-C data processing methods need to do these steps before further analysis is possible. The four steps consist of [253]:

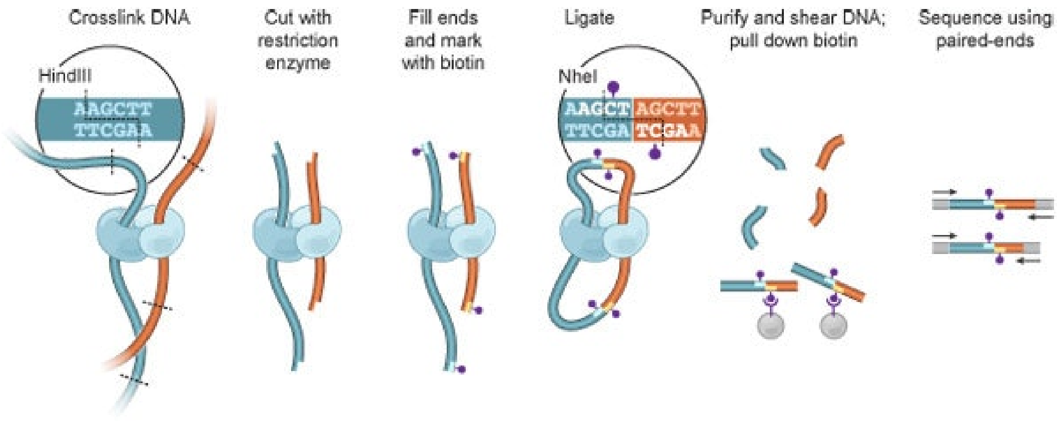

Crosslink DNA
HindIII
AAGCTT
TTCGAA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
NheI
AAGCT AGCTT
TTCGA TCGAA
Purify and shear DNA; pull down biotin
Sequence using paired-ends

Figure 152 Overview of Hi-C steps. Source: (1)

1. Alignments
2. Preprocessing
3. Binning
4. Normalization

In the end, the main output consists of a list of interactions between genomic regions. [254]

15.2.1 Alignments

The raw paired-end reads must be separated into two single-end read files, since the ends do not usually map onto the same region in the genome. If we would use a paired-end read mode for any aligner, we would receive almost no valid alignments, since the gap penalty would be too high [255].
So first the two-paired end reads have to be separated usually splitting the two-paired end read into half. 80 base pair long reads would be split into two 40 bp long reads by cutting the middle. In our pipeline we achieve this by using `fastq-dump` and the parameter `-split-files`.

Then we have two fastq files which for each we start the alignment-process. For this, we use BWA [256] and bowtie2 [257].

BWA - Burrows-Wheeler Aligner

This method is known to be memory-resource intensive. But it is well established and Juicer uses this in the first step [256].
There are three algorithms, which use an index of the reference genome. The index was created before and contains a suffix tree for quick exact search. To find sequences, that do not match with 100 % identity, each algorithm uses its own method - means, if the alignment would be of edit-distance one and higher.

- *BWA-backtrack*; Used for Illumina sequence reads up to 100 bp; This algorithm uses backtracking in its core. There are several additional heuristics to increase the speed (while losing accuracy). [256]
- *BWA-SW*; for read length: 70 bp to 1 Mbp; It is deprecated and only be used in cases, where the user might want to get higher sensitivity. [256]

- *BWA-MEM*; for read length: 70 bp to 1 Mbp; It is considered the latest algorithm, thus generally recommended for high-quality alignments, since it is faster and more accurate than *BWA-SW*; It performs better than *BWA-backtrack* for 70-100 bp. [256]

.

We only use *BWA-MEM* (bwa mem) and *BWA-backtrack* (bwa aln). The pipeline is set to use bwa mem for reads longer than 70 base pairs and bwa aln for reads lower than 70 base pairs.

bowtie2

Bowtie 2 is a fast and memory-efficient tool for the alignment of raw sequencing reads to long reference sequences. Its favourable runtime properties make bowtie a tool of choice for huge amounts of data [257]. Since we have many reads to process, for us it made a good choice. It is considered to be faster than BWA-MEM. This is the reason why we chose it. Additionally it is a common choice.

15.2.2 Preprocessing

Hi-C data is prone to several different types of bias that are targeted by normalization and preprocessing methods; the following four are most prominent:
1) Sequencing bias, i.e. bias that happens during the amplification step of sequencing; 2) Bias of read depth per region, e.g. due to reduced alignability stemming from repeats in certain regions; 3) Bias from linear distance, which is the main source of bias in Hi-C data. DNA fragments that are closer on the DNA strand tend to 'interact' because they are spatially close, even when no true 3D-interaction is given, during the ligation step of the Hi-C method ('spurious ligations'). Before normalization the diagonal of the contact matrix is the most visible feature, because of such spatially close interactions. Those ligations are, however, removed by some tools. 4) Bias from chromatin compaction: Heterochromatin is compacted closely, therefore there is a different interaction frequency in heterochromatin than in euchromatin.

To account for those various bias types, each method has their own algorithm. It should be noted that most algorithms remove exact duplicates (interactions, which have the same interaction partners) to remove duplicated regions which would otherwise show false signals in regions, that were only duplicated because of the NGS method.

15.2.3 Binning

In order to prepare contact data for visualization (e.g in a heatmap), interaction data is mapped onto a two-dimensional grid where x and y-axes correspond to the genomic region of interest and every cell corresponds to a genomic area of pre-defined size (e.g. 25000 bp). In this grid, for every cell (area A, area B) the number of interactions between fragments in area A and B are counted. The resulting matrix can be used for visualization in heatmaps that display the interaction density over large regions of the genome. [254]
The binning step is performed by the tools generating .hic files (Juicer, Homer) and is also performed by our web tool when entering .bed-type files listing interaction partners.

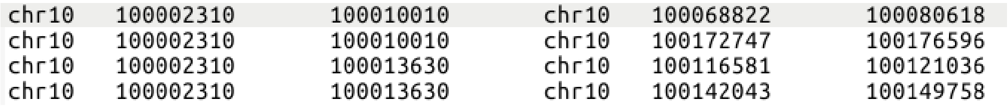
Juicer provides the **juicer-dump** tool for calculating and printing contact matrices in .tsv format, using .hic files as input. Since contact matrices with several output resolutions (bin-sizes) are stored in .hic files, this step is not very time-consuming and is integrated in our pipeline. Furthermore, the juicer-dump command allows the user to select a normalization method (e.g. Vanilla Coverage, see 9.4.3 for detailed explanation) to apply on the data. The user can upload either a contact file in .bed format or a .hic file and vizualise the input as a heatmap, after selecting a genomic range and bin-size to display.

Bin-Size

The right choice of the bin-size is important in all analyses that require contact matrices. Common bin-sizes range from 1kb to 1Mb; for analyses relying on contact matrices as input data, the optimal bin-size is usually in 5-digit area, since the matrix grid should be defined in a way that very few interacting fragments stretch over more than one bin (fragments being present in several different bins would give the faux impression of higher interaction density).

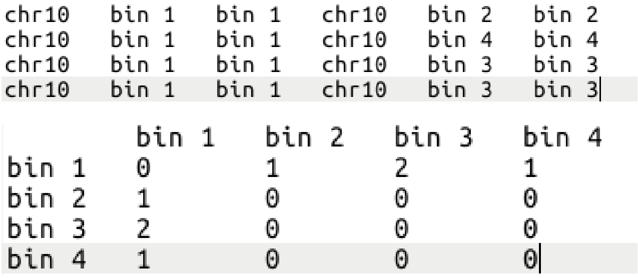
Binning Step

In the following, the binning process used in our pipeline will be described.
First, a function is defined that maps each genomic position to a certain bin, taking bin-size and a file listing genomic positions. For every entry, the bin positions of both fragments are calculated and stored in an array.



```
chr10   100002310       100010010       chr10   100068822       100080618
chr10   100002310       100010010       chr10   100172747       100176596
chr10   100002310       100013630       chr10   100116581       100121036
chr10   100002310       100013630       chr10   100142043       100149758
```

The next step is an iteration over the array containing contacts, where a pre-generated two-dimensional array is used to represent the contact matrix (both dimensions are genomic range of interactions). For every entry in the initial array consisting of an interaction between positions a and b, the entry in the contact matrix for positions a and b is increased by the contact number.



```
chr10   bin 1   bin 1   chr10   bin 2   bin 2
chr10   bin 1   bin 1   chr10   bin 4   bin 4
chr10   bin 1   bin 1   chr10   bin 3   bin 3
chr10   bin 1   bin 1   chr10   bin 3   bin 3

        bin 1   bin 2   bin 3   bin 4
bin 1   0       1       2       1
bin 2   1       0       0       0
bin 3   2       0       0       0
bin 4   1       0       0       0
```

Figure 153 Binning step, explained with simple example. The 'file' in the middle is only a schematic depiction of the internal step performed by our software. The bin-size in this example is 50 kb.

Problems With Binning

The major problem that can arise during binning stems from the bin size. If a too small bin size is selected, interactions spread over several bins, giving the faux impression of a higher interaction density. Also, many gaps will be present in the contact matrix, even in interaction-rich regions. This means that the memory needed to store the matrix bears no relation to the increase in information. If the bin-size is selected too big (e.g. 1 Mb), important information is lost, and detection of TADs (size usually within 100kb - 1 Mb range) will be impossible.

15.2.4 Normalization

The task of normalization is to make the data comparable with means to e.g. sequencing depth and other factors, that influence the data significantly [255]. There are several matrix balancing schemes and various easy, computationally fast algorithms, that let e.g. parts of regions that appear as cluster intensify and sometimes over-correct significant signals to a lower state, so that interesting regions might not be seen anymore. [258]

Main biases that are usually corrected for are the sequencing depth and the linear distance [253]. The methods are usually using the sum of one row or column in combination with the value in the cell itself. There is also the possibility to calculate a background model for which each element is later combined depending on the background model [255]. Most commonly represented is the expected background model based on the function of probability to see an interaction given a distance [259].

15.3 HOMER

HOMER (Hypergeometric Optimization of Motif EnRichment) [255] is a suite that provides several programmatic tools for Motif Discovery and next-generation sequencing analysis, including several command line programs supporting the analysis of Hi-C data. The Hi-C package provided by HOMER includes tools for multiple steps in our pipeline that aid in customizing and checking parameters, as well as tools for generating statistics to understand the underlying data. These steps are to be executed one-by-one in comparison with Juicer, where all the available steps are already included in one script. The advantage of HOMER over Juicer would be, that the user is able to customize HOMER to a fine-scale.

Initial steps

Hi-C and variant technologies use paired-end sequencing to find interactions between genomic regions. After sequencing, two files are required: read forward and read reverse from paired-end sequencing. The `fastq-dump` command can be used to convert SRA files to .fastq format which is used as a common basis for all further steps.

Due to the biological nature of fragments reads resulting from Hi-C experiments, alignments have to be generated in single-end and not paired-end read mode (explanation can be found in the Hi-C data processing section.

Before the HOMER pipeline is executed, there need to be at least two sam/bam-files representing forward and reverse reads that were mapped to the reference genome.

15.3.1 Pairing, Filtering and Quality control

The pipeline for the Hi-C analysis consists of the tools: *makeTagDirectory*, *analyzeHiC* and some additional Perl scripts, that were not used in this pipeline. All explained steps were done with these two tools.
makeTagDirectory is involved in pairing and filtering, whereas generation of the contact matrices including the normalization is done by *analyzeHiC*.
The following sections explain each step, with each explaining the corresponding tool of HOMER.

15.3.2 Pairing

makeTagDirectory: This tool, does pair two single-end read alignments (in various formats, e.g. sam, bam). The result is a directory that contains the paired reads as so called "tags" or fragments. They are saved for each chromosome in homer-specific .tsv format separately.

The directory created without any options can be called unprocessed, since no filtering options e.g. self-ligations have been removed. Furthermore, the user is prompted to experiment with following procedures to verify and improve the quality of the data.

15.3.3 Filtering

1. if needed, trim for restrictions sites to only contain read until restriction site
2. *makeTagDirectory -update*; This option is used, after *makeTagDirectory* was already executed once without the option *-update* which created a new paired-end tag directory. It is recommended to copy that directory and update the copied directory with different parameters to find the best parameter combination suited for the data.
   These parameters are defined through the options and those options can either be called on its own, or in combination, whereas some other options need to be used in conjunction with e.g. the option *-genome* to properly find the restriction sites (if they are not contained on the two-paired end tag).

   a. *restrictionSite XXXXXX*; This must be in conjunction with -genome genome to let homer search for all restriction sites; With the parameter -mis the user can specify how many mismatches for the restriction site are allowed. That would mean if the restriction site might have been mutated, but the enzyme to cut, would still split this section up, the user can search for these sites that do not have an exact match with the referenced restriction site.
   b. *removeSelfLigation*; Removes all fragments, that are adjacent to each other.
   c. *genome*; This is needed for the restriction site to be found. As explained before, this is mandatory if one wants to map the restriction sites to the reference genome.
   d. *both/none/one/onlyOne*; This is applied to keep only paired-end fragments if both/at-least-one/only-one/no-site-found-in-proximity single-end reads have a restriction site.
   e. *removePEbg*; This parameter is used to remove all paired-end fragments that are within 1.5 times the estimated fragment size range. This option is used to remove near duplicates, meaning fragments that lie within 1.5 times genomic range (on the reference genome), are removed since they are probably a (not exact) duplication that occurred from sequencing (according to HOMER Website).

15.3.4 Quality Control

Here we have various options to choose from: First, with every tag-directory there are statistics, that are generated by HOMER to check:

- Length distribution of fragments: This aspect of Hi-C data is important, because the distribution can give insights of high long the re-ligations are and then in the course of filtering provide a good hint where to set the threshold for too small or too big fragments, that probably did not show a real interaction with biologically relevant meaning.
- Fraction of interactions at given interaction distance: Many important interactions are of very long range. The profile of the distribution are distinct in that they have an increase in interactions at a distance of ten million and more. If this is not the case, maybe the data is not correct. Additionally too small interactions can show up in this diagram, which then helps to reset the threshold for too small interactions.
- Distance between 5'-Ends of the fragments: This particular aspect is different from the length distribution of fragments. Since Hi-C data is provided in paired-end fragments they get split up into single-end fragments which have to be paired. The distance between 5'-Ends provide the insight between opposite and same strand. This is further discussed below with an example provided.
- Distance between the single-end reads of the fragment and their found restrictions site: Since the restriction site defines the position, where the junction between the two single-end fragments resides, this quality control provides good insight for the common valid length of single-end fragment.
- Tag/read distribution per bp: This distribution is the coverage of the genome.

Second, the user can plot these statistics in any preferred form. Some examples are shown in the following section.

Fragment Size Distribution

To find the average length of the paired-end fragments, a histogram of the distances between the two single-end 5' ends of each two-end pair fragment, separated by same and opposite strand, can be used. In Figure 154, there is a peak for the opposite strand at around 350 bp, which is the average size of the fragments. At position zero for same strands self-ligations are shown whereas at the same position for the opposite strand represents dangling-ends (Figure 154).

Interaction Distance vs Fraction of Paired-end reads Log-Log-Scale

A key statistic is the overall distribution of the distances for the interacting genomic regions. It can be simply calculated by noting the difference of the genomic position on the reference genome. It is important to understand, that this can be only done for regions which lie on the same chromosome. The interactions that are between different chromosomes are noted as “interchromosomal” whereas the interactions inside of a chromosome are called “intrachromosomal”. The difference is measured in base pairs. The presented Figure fig:interaction-distance shows that at 100,000,000 base pairs distance the number of interactions does not decrease much. This can mean, that there are biologically meaningful interaction that appear such long distances.

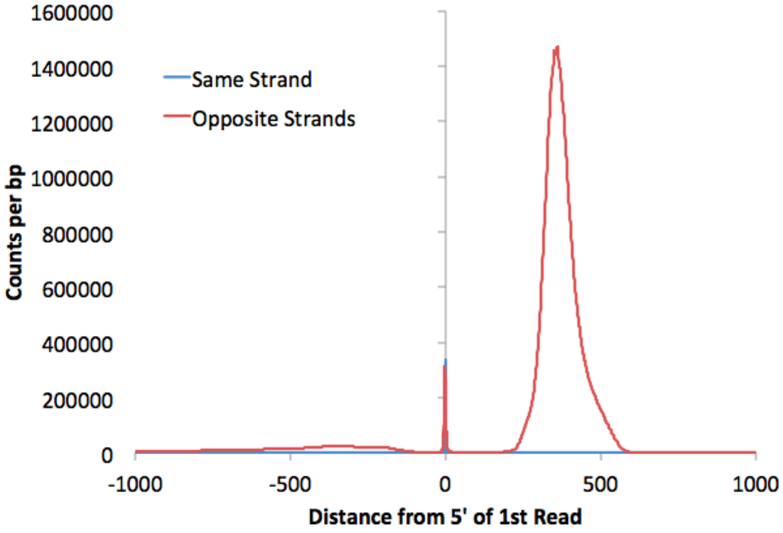


Figure 154 Histogram of the distances from 5' of 1st read to the 5' end of the second fragment (single-end read fragment from the other side of the paired-end read) after these fragments where mapped to the reference genome.

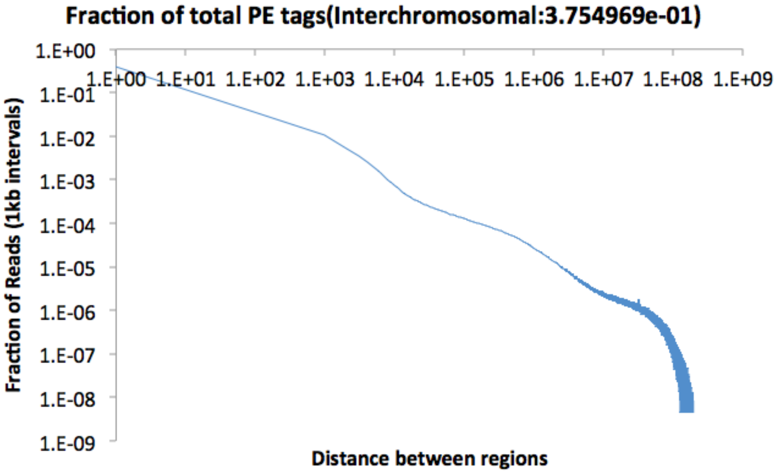


Figure 155 Histogram (fraction-wise/density) of distances of paired-end reads, that reside on one chromosome. The fraction of interchromosomal interactions are shown additionally, which in this case is 0.3754%. Both axes are shown as log10-scale.

15.3.5 Normalization and Generation of Contact Matrices

analyzeHiC: Primary analysis program, which generates interaction matrices, normalization, identification of significant interactions, clustering of domains, as well as generating the data for Circos plots (most of the following programs use this one internally)

There are various sources of bias which can be divided into 4 groups: Sequencing bias, read depth per region, linear distance and the chromatin compaction. The sequencing bias is common to all projects involving sequencing. Read depth per region can vary because of e.g. the align-ability of repeat regions. The main bias in Hi-C data is the linear distance between loci on a chromosome, which is also the reason why heatmaps without normalization or filtering have a diagonal, that represent close regions on the chromosome. And the last bias is the chromatin compaction, which is also valuable in finding Open Chromatin.

In order to remove the effect of those bias', normalization is applied to the (raw) data. In the following section, the algorithm developed and used by HOMER for normalization will be discussed:

Normalizing for sequencing depth

To normalize for sequencing depth HOMER simply uses the formula:

$$e_{ij} = \frac{(n_i)(n_j)}{N}$$

N = total number of reads

n_i = total number of reads at region i

n_j = total number of reads at region j

e_{ij} = expected number of reads at region i,j accounting for sequencing depth

Normalizing for linear distance

To also account for the linear distance HOMER introduces the function $f(i-j)$ as a function of distance. To normalize for regions that e.g. in reality would include interactions to unmappable regions such as repeat regions, HOMER tries to estimate the number of interactions, which should be higher in reality than the observed/measured interactions of a region adjacent to such repeat region. In the following formula n^{*i}, n^{*j} and N^* represent estimated numbers of interactions for regions i and j and total for N. These number can not be estimated by using the observed counts trivially, instead we iteratively aim to minimize the difference between current estimated number of reads per region to the observed number of reads using:

$$e_{ij} = f(i-j)\frac{(n^{*i})(n^{*j})}{N^*}$$

15.4 Juicer

Several tasks must be performed in order to process and analyze Hi-C data: alignment of raw reads, preprocessing of interaction data, binning and normalization of contact maps. As we see in Figure 7, Juicer [254] offers a command line tool that automates all these steps, allowing users with little experience in data management to transform raw sequencing data into complex contact maps. Combined with its flexibility (see following paragraph), this makes it a prime candidate for a Hi-C analysis pipeline.

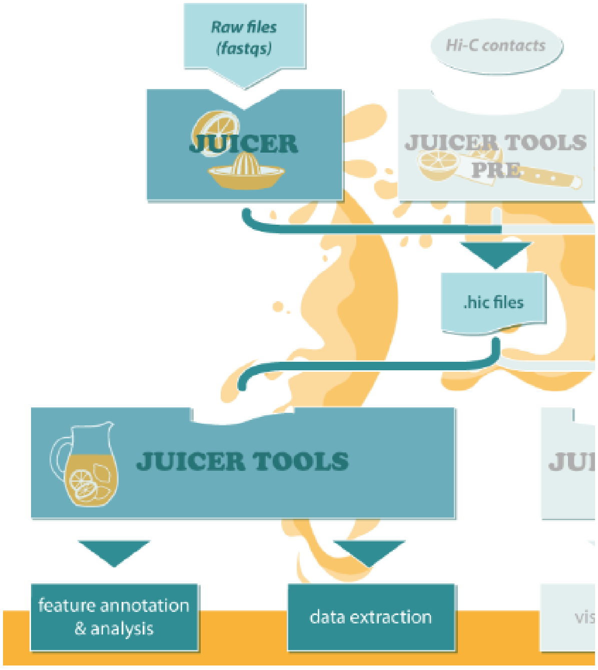


Figure 156 Overview of Hi-C steps: Juicer performs preprocessing steps, including binning and normalization, on .fastq files to generate .hic files. These files can be used to generate heatmaps. Source: (1)

15.4.1 Alignment Of Raw Reads

The alignment of raw reads, as a first step in every Hi-C data analysis, is performed by Juicer using BWA (Burrows-Wheeler Alignment) [256]. BWA represents the reference sequence as prefix trie and builds a word graph from the query sequence. Dynamic programming is then used to traverse both reference and query sequence to find all matches. This procedure is quite memory-consuming but well established.

The first step after sequencing is the transformation of raw sequence data into lists of contacts by aligning the read fragments to a reference genome. To accomplish this, the paired-end reads generated by Hi-C procedure must be read as single-end reads (one read per interaction partner). To avoid this, each read-pair is given an identifier and aligned as single-end reads.

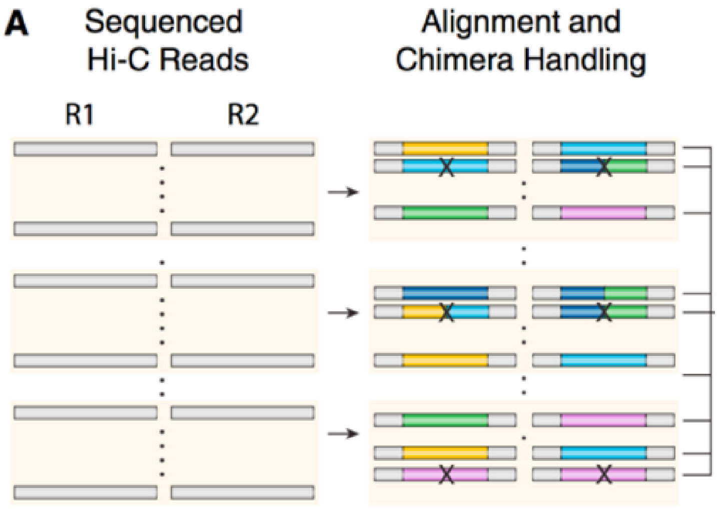


Figure 157 Schematic depiction of alignment step in Juicer. Chimeric reads, i.e. sequencing reads that align to two distinct portions of the genome with no overlap, consist of two different colors.

Given the appropriate hardware, this step can be accelerated by splitting up data in several chunks and running the process on several CPUs, or by using an FPGA.

Burrows-Wheeler Alignment:

This alignment method works by seeding alignment with maximal exact matches, and then extending seeds with the affine-gap Smith Waterman Algorithm.

15.4.2 Contact Map Calculation

After alignment of the raw reads, the files containing pairs of interacting fragments can be used to calculate contact matrices that portray contact frequency between sections of the genome.

15.4.3 Normalization

Juicer allows several different normalization types to be applied to the data, which provides us with an easy option to compare the output data of these methods. The available normalization techniques are: 1) Vanilla Coverage (VC) normalization [259, 258], 2) Vanilla Coverage normalization with modified formula (VC_SQRT), and 3) Knight and Ruiz (KR) normalization [260]. In the following paragraphs, we will describe these methods in more detail.

Vanilla Coverage When performing Vanilla Coverage normalization, for each position (i,j) in the matrix, the number of contacts is divided by both the total number of contacts in the respective column, and in the respective row. The value after normalization (VC) is defined as:

$$VC = C_i \cdot M_{ij} \cdot Rj$$

$$c_i = \frac{1}{Contact\ \ frequency\ \ in\ \ row\ \ i}$$

$$r_j = \frac{1}{Contact\ \ frequency\ \ in\ \ column\ \ j}$$

$$M_{ij} = \ Contact\ \ frequency\ \ between\ \ bin\ \ i\ \ and\ \ j$$

The resulting contact matrices can be visualized as heatmaps. The following example shows a heatmap of all interactions between 25 - 35 MB on chromosome 10, before and after Vanilla Coverage normalization:

VC_SQRT: In this method, a normalization similar to Vanilla Coverage is performed, with the difference that instead of dividing by the contact frequency, one divides by the square-root of the contact frequency. This is done because regular Vanilla Coverage someties 'overcorrects' the data.

KR: The other normalization method performed by Juicer is Knight and Ruiz normalization [260]. This method converges two magnitudes faster than the Sinkhorn-Knopp algorithm (which rescales all rows and columns of the matrix to sum to 1). Knight and Ruiz normalization applies a combination of the inexact Newton's method and inner-outer iteration with conjugate gradients to a system of linear equations to quickly find the next matrix in the iteration process.

15.5 TADs (Topologically Associating Domain)

In the three-dimensional architecture of the human genome, there are compartments of DNA, called Topologically Associating Domains. Within these compartments, interactions (e.g. between enhancers and promoters) occur more frequently, while interactions across TAD boundaries rarely occur, meaning that these compartments work as ‘barrier’ interactions. [248] Usually, the size of TADs is between 100 kb and 1 Mb.

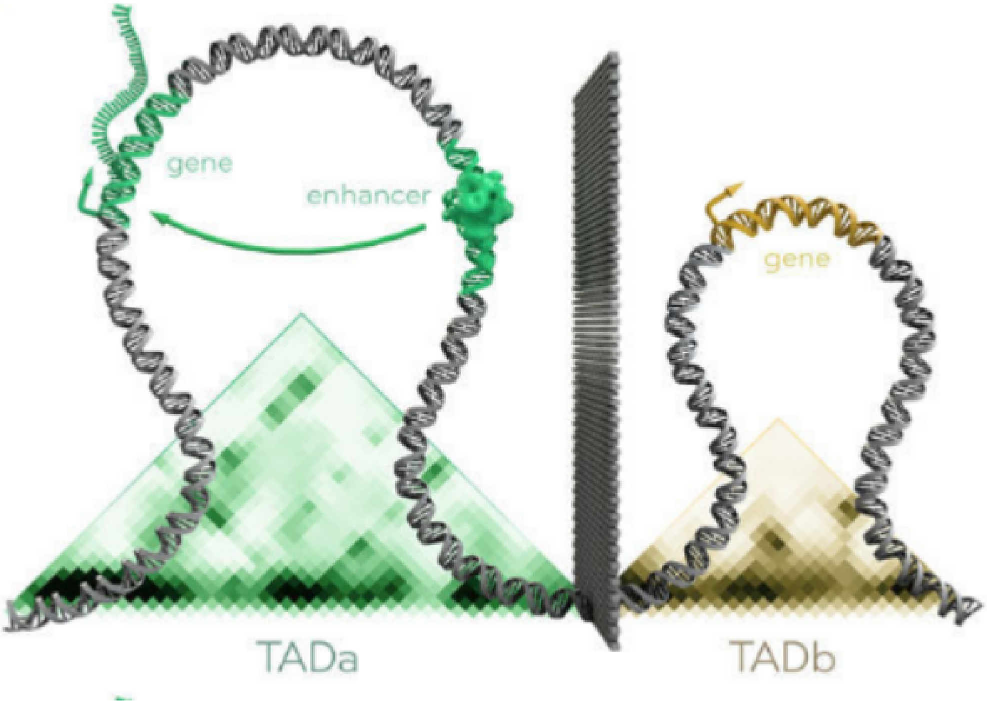


Figure 158 View of two TADs with interaction between gene and enhancer. The background plot shows a matrix representation of interaction density inside the TADs.

The majority of regulatory interactions in the genome does not ‘cross’ TAD boundaries. If a TAD boundary is removed or disrupted, new contacts between regulatory elements can form. These contacts can have massive influence on phenotype development and cause pathogenic phenotypes (e.g. limb malformation, cancerogenesis) [250]. This is one of the properties that make TADs interesting for medical Research.

15.5.1 TADs - Detection

Most interactions in the whole genome occur between close partners (< 25kb). In TADs, the distance between interaction partners is different, with a higher number of interactions between more distant fragments.
Figure 10 shows the distribution of interactor distances inside a TAD compared to the whole chromosome. While the distribution in the whole chromosome appears to roughly follow a negative exponential function, the distribution inside the TAD shows a much slower decrease in density, as can be seen in figures 11 and 12.
This observation provides a starting point for computational prediction of TADs, which will be explained in the following chapter.

15.5.2 TADTree

TADTree [261] is a free tool that predicts location and hierarchy of TADs based on contact matrix data. The model used by TADTree is based on the assumption that in TADs, enrichment of contacts over the background increases with growing distance (see Figure 11,

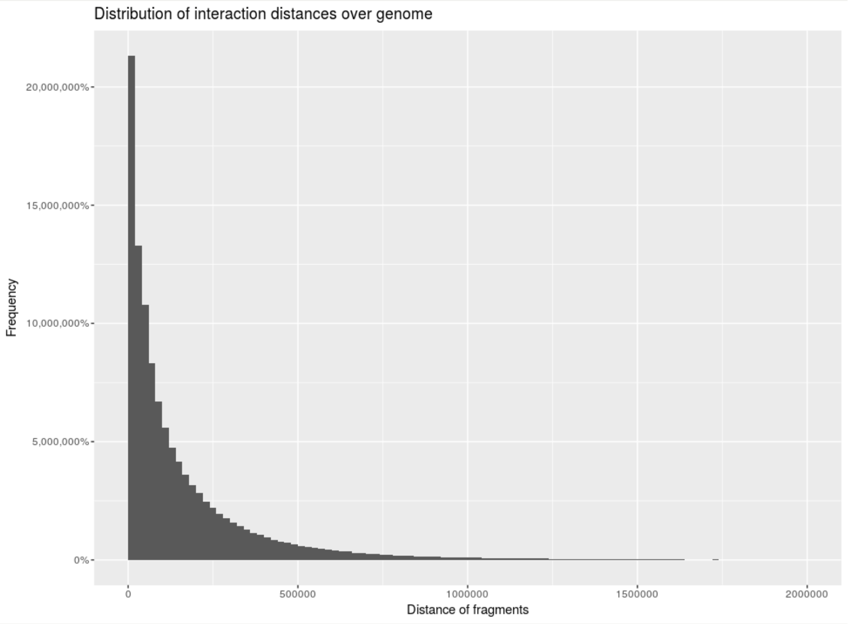


Figure 159 Distances of interacting fragments in whole genome.

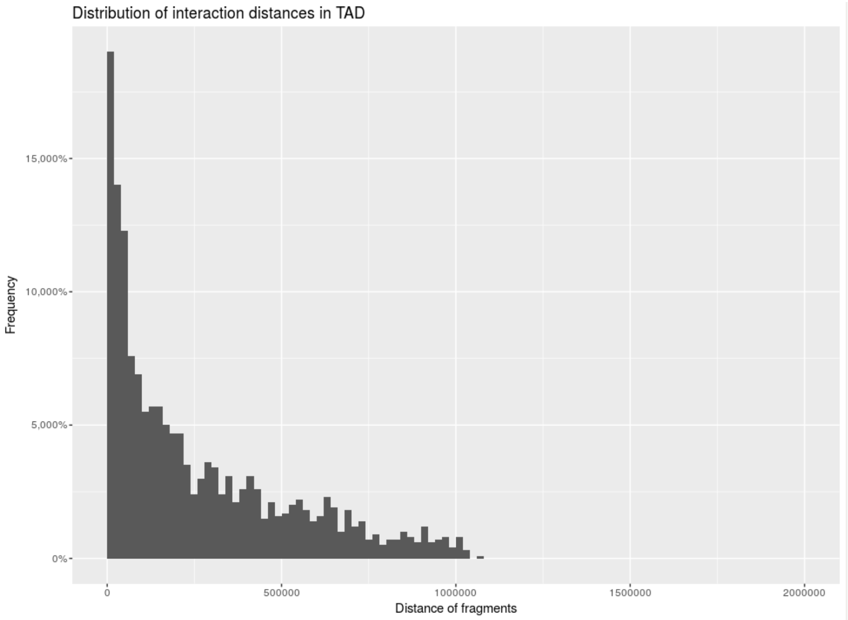


Figure 160 Distances of interacting fragments in TAD (Chromosome 10, 30.95 - 32.05 Mb). A high percentage interactions are between fragments that are spatially very close (< 25 Kb),

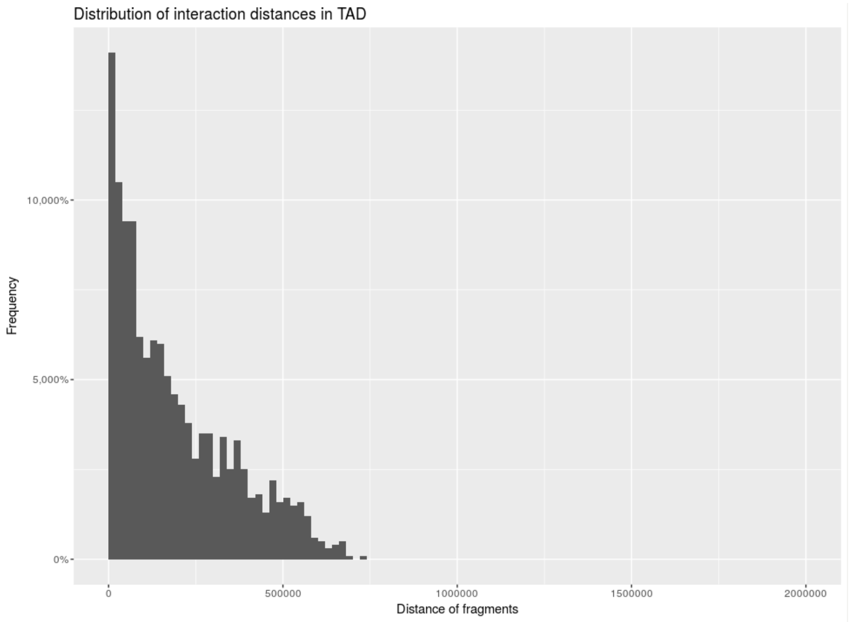


Figure 161 Distances of interacting fragments in TAD (Chromosome 10, 28.95 - 29.63 Mb)

12 above), and if a TAD lies inside another TAD, this increase is faster. In a heatmap, this shows as ‘borders’ of high-interaction-density areas sharpening towards the edge. Unlike other models, TADTree also calculates a hierarchy of ‘nested’ TADs (TADs inside another TAD).

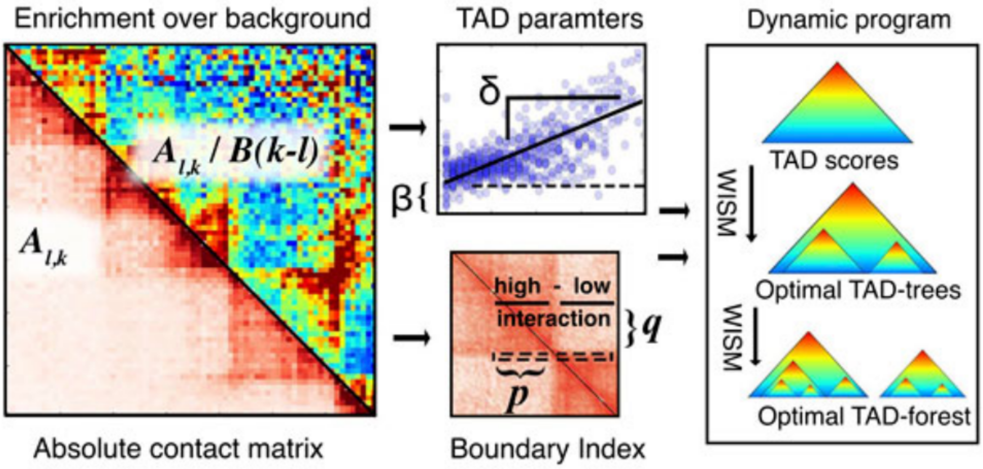


Figure 162 Schematic depiction of the model used by TADtree.

15.5.3 TADTree - Output

In our pipeline, we integrated TADTree in our contact matrix visualization, so that for a given heatmap, the user can calculate and display TADs. A maximal TAD size of 1 Mb was selected, with a standard bin size of 25 Kb. As we see in figure 14, the TADs calculated by TADTree overlap with the ‘dense’ regions, meaning that a high number of interactions are inside TADs, and few outside. The corners of the TADs appear sharp, especially for the large TAD around 14 Mb. Inside three out of four TADs, ’nested’ TADs were found.

As mentioned above, we use a maximum TAD size of 1 Mb. This was since the average size of TADs is between 100 Kb and 1 Mb; also, this parameter has proved best in ’preliminary testing’. Bigger TAD sizes would result in very large TADs containing many gaps with few interactions.

The TADs were generally calculated on a matrix with 25 kb resolution, and then mapped back on the original visualisation in order to keep input data quality as consistent as possible. The standard input parameters of TADTree for boundary threshold, i.e. how sharp boundaries of areas must be in order to qualify as TAD, have been used in our analyses.

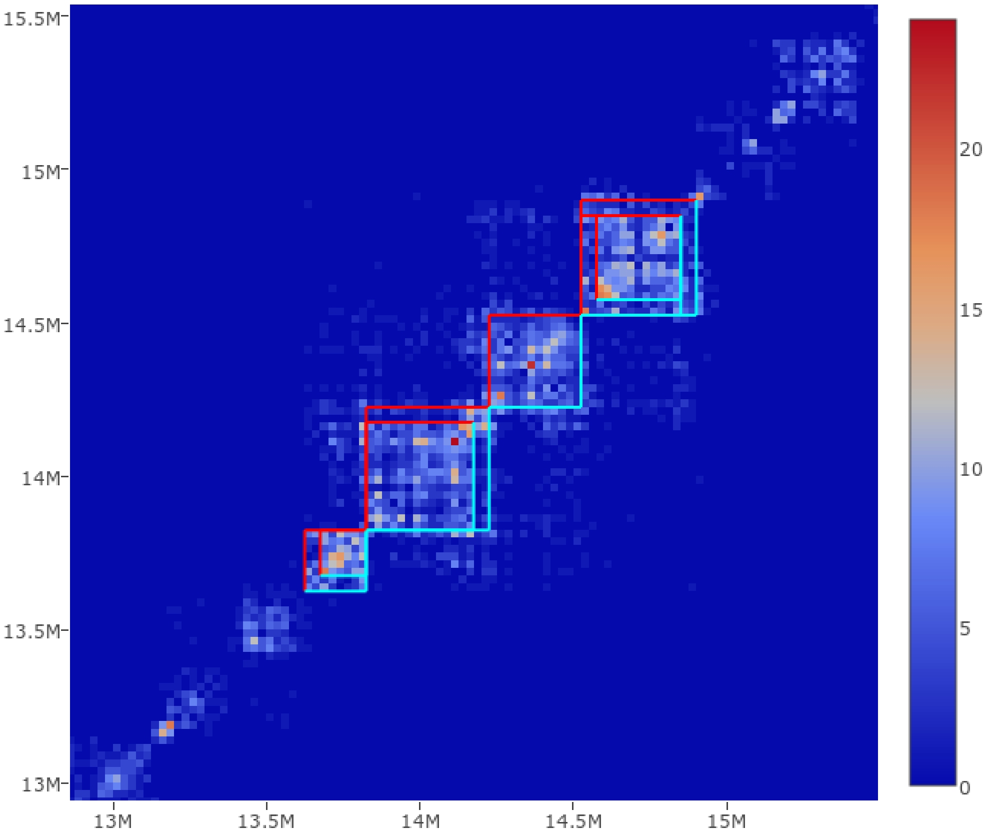


Figure 163 Heatmap of contact frequency in genomic range (chr10, 12 Mb - 17 Mb) with TADs calculated by TADtree.

15.6 Results and Discussion

15.6.1 Normalization

A major application of Hi-C data is the creation of heatmaps and use for domain recognition and peak calling [258]. To support the finding of these features in heatmaps, its recalculation of the matrices is usually necessary [259]. The use of logarithmic scales for biological data is well-established. However, the methods presented here are different: Figure 164 shows normalization algorithms for Juicer whereas figure 165 compares the normalization algorithms for Homer. At last, in figure 166 we compare the raw data from Juicer with Homer to show the importance of filtering reads before generating matrices.

Juicer Normalization Algorithms

It is possible to compare heatmaps by defining scores and measuring nearby specific parts - for instance - a peak[258]. Though in this case, the images are simply compared manually. This technique is more often used than systematic inspection by predefined scores, as the scores must be defined first. It should also be mentioned, that in this chapter the use for biologists and the like is more in the field of explorative studying: finding regions that can easily be detected by viewing the heatmaps. Here, all heatmaps (Figure 164) were generated by using the same data sets but different methods.

First, the raw data from the Juicer (Figure 164a) pipeline will be compared to the Vanilla Coverage (VC) algorithm (Figure 164b): in the raw data set the diagonal is well visible whereas the VC algorithm almost removes the diagonal. This phenomenon can be explained by the concept of the VC algorithm, which corrects too strong signals based on the overall signal of a region compared to all other regions. In this data set, unfortunately, the VC algorithm is not working that well.

One solution for the over-correction of the VC algorithm is the Vanilla Coverage Square Root algorithm (VC-sqrt). Figure 164c shows the VC-sqrt and demonstrates (in comparison to Figure 164b) the improvement of the over-correction effect.

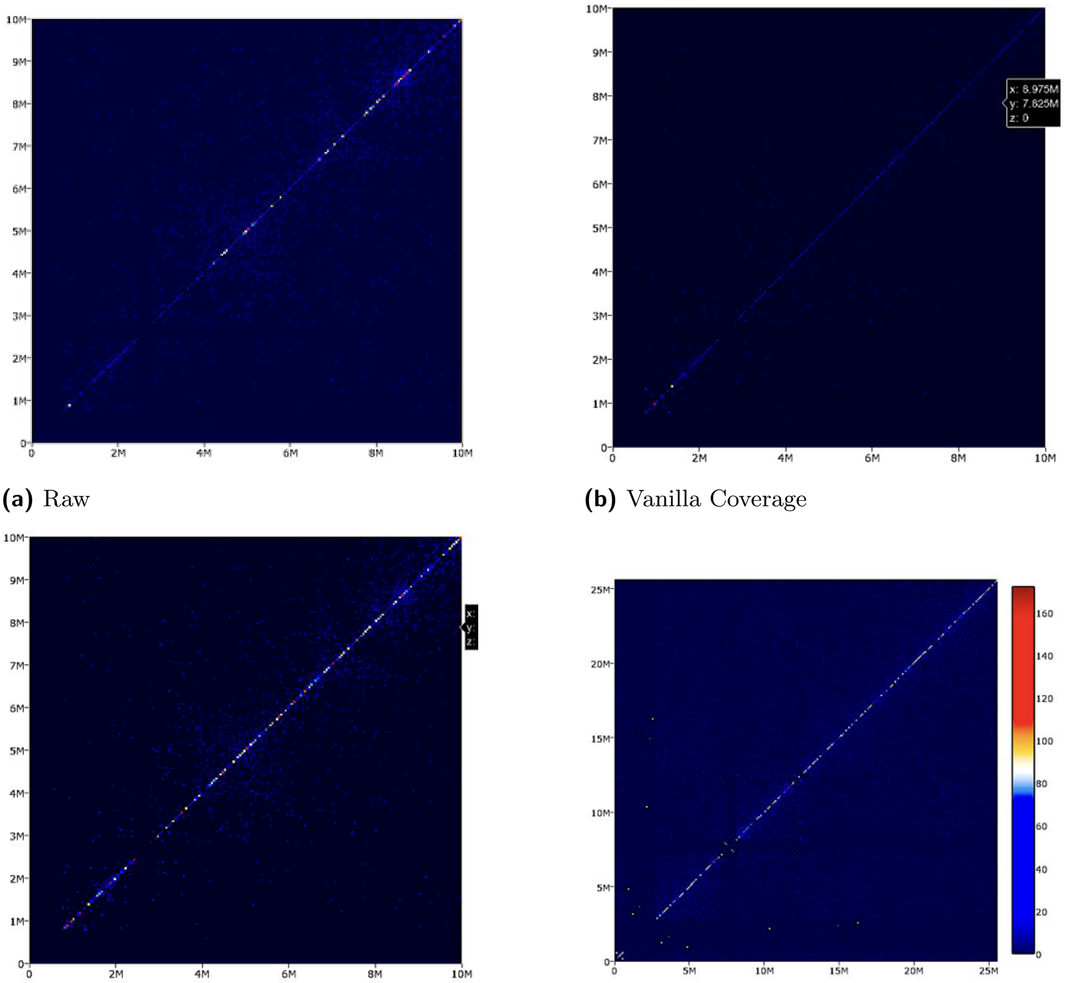
The newest default algorithm, used by the Aiden lab and their software package Juicer, is the Knight and Ruiz algorithm (KR) visible in Figure 164d. This heatmap demonstrates that now in certain regions the cluster can be seen. It represents a major advancement, since the KR algorithm imitates the repetition of the VC algorithm, which would lead to a similar behaviour if improved properly [258].

HOMER Normalization Algorithms

The raw data from HOMER comes already filtered (see Figure 165a). The diagonal is not continuously visible as in Juicer, while outer regions appear and clusters are slightly viewable.

In Figure 165b, the simple norm algorithm of HOMER highlights many single interactions and some narrow clusters appear. For example, at region eight to nine million base pairs a red cluster shows high interactions in this area. This region was already visible in Figure 165a, but now there is a difference between this red cluster and other regions on the diagonal, which had a similar appearance before.

A different background color is shown in Figure 165c. This can be explained as followed: under-represented regions become blue and over-represented regions become red, which implicates that a new color needs to be defined for neutral interaction, which in this case is beige. Since this algorithm also normalizes based on the interaction distances, clusters become visible now. Thereby, certain areas within the cluster become clusters themself. This



(a) Raw

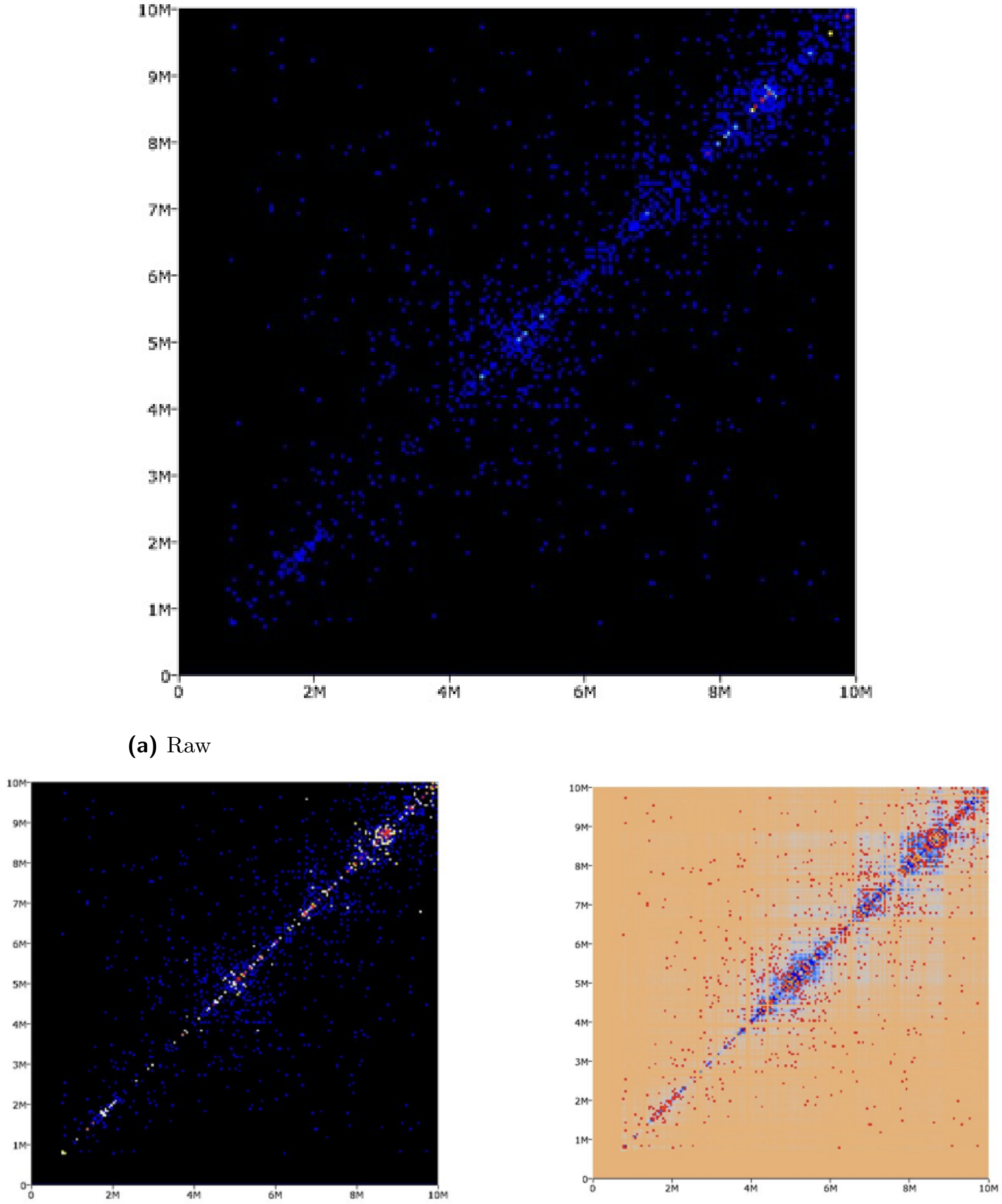
(b) Vanilla Coverage

(c) Vanilla Coverage Sqrt

(d) Knight and Ruiz

Figure 164 Comparison of the raw data and the three normalization algorithms of Juicer. The (a) Raw data shows a nice diagonal, whereas the normalization using (b) the Vanilla Coverage (VC) algorithm over-corrects the data diminishing the diagonal, (c) VC-Sqrt algorithm reduces the over-correction effect a little, showing a nice diagonal again, while (d) the Knight and Ruiz (KR) algorithm greatly balances the matrix and makes also clusters observable. The KR algorithm, although, iterates the VC normalization through an intelligent algorithm quickly and is the recommended choice.

is the advantage of the full norm algorithm compared to the simple version in HOMER.



(a) Raw

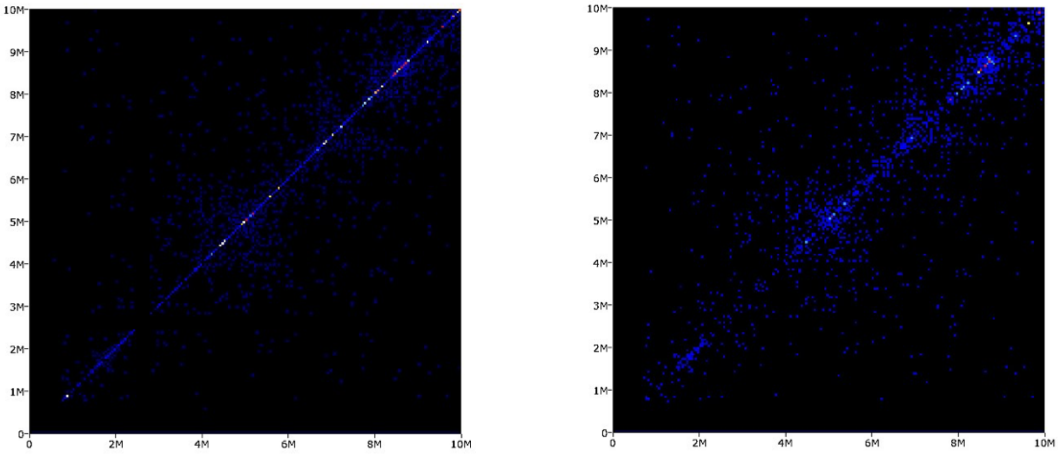
(b) Simple norm

(c) Full norm

Figure 165 Comparison of the raw data to the two normalization algorithms from Homer. The heatmaps using the (a) raw data, (b) the simple norm, and (c) the full norm normalization are shown. The simple norm normalized the data for sequencing depth and the full norm additionally normalized for linear distances.

HOMER vs. Juicer comparison

As explained in the previous section, before using a normalization algorithm HOMER filters the data more than Juicer. Figure 166 shows the two methods in comparison. Even without normalizing, HOMER already acquires well visible results, which can be used for manual examination. The main difference is the diagonal: it is visible in Figure 166a, but not really present in Figure 166b. One advantage of eliminating the diagonal is the change of scale to view worse signals, which come from the other regions. These would normally be overshadowed by the supposed interactions on the diagonal.



(a) Juicer

(b) Homer

Figure 166 Comparison of the raw heatmaps produced by (a) Juicer and (b) Homer. Juicer keeps the diagonal while Homer removes it by the filtering process.

15.6.2 EpiGenomeBrowser

In order to associate the Hi-C data with other regulatory features like RNA-Seq data, we needed an interface providing comparison of Hi-C heatmaps with these datasets. EpiGenomeBrowser [262] proved to be a favorable solution for this, which will be discussed now in more detail.

The EpiGenomeBrowser is a web interface showing available public datasets as well as private datasets hosted online together as tracks. It is easy to visually explore the data either on genome-wide scale or on local scale, when someone wants to explore specific regions e.g. a region around a custom gene set.

We chose this software package for various reasons. First, it has the capability to include heatmaps from Hi-C data; second, most datasets that are published and available online, are already integrated as tracks in JSON file format. The user is provided with the ability to specify its own dataset by creating a JSON file and hosting it on a server. Additionally we can easily customize the tracks to increase the visibility of hypotheses that we want to show. Another interesting feature is one of the various tools available to analyze the surrounding region of defined gene sets.

Since this web-interface is showing big datasets, the Javascript engine, does get slow with increasing view range and increasing number of tracks shown simultaneously [253]. Also, most additional data is loaded from other servers, so that every time the user adjusts the view range, the data is reloaded from the other server. One way to circumvent this behavior is to provide the datasets used for the analysis on the web server locally. For that reason, we installed the EpiGenomeBrowser on a server and loaded the heatmaps and JSON files onto the server for faster loading times.

Association with CTCF Binding Sites

CTCF is a transcription factor encoded by the CTCF gene [263]. It can bind together the strands of DNA. Because of this binding of two different loci, DNA forms into loops [258]. Visualization of Hi-C data in heatmaps makes regional clusters of interactions visible that usually resemble DNA loops.
In Figure 18, a juxtaposition by EpigenomeBrowser of CTCF ChipSeq data (as a barplot) and Hi-C data (contact frequencies displayed as heatmap at 5 kb resolution) can be seen. The genomic regions highlighted in pale blue and pale orange that show the highest signal in the CTCF data, coincide with the locations with the highest local interaction density (score of 170) in this region.

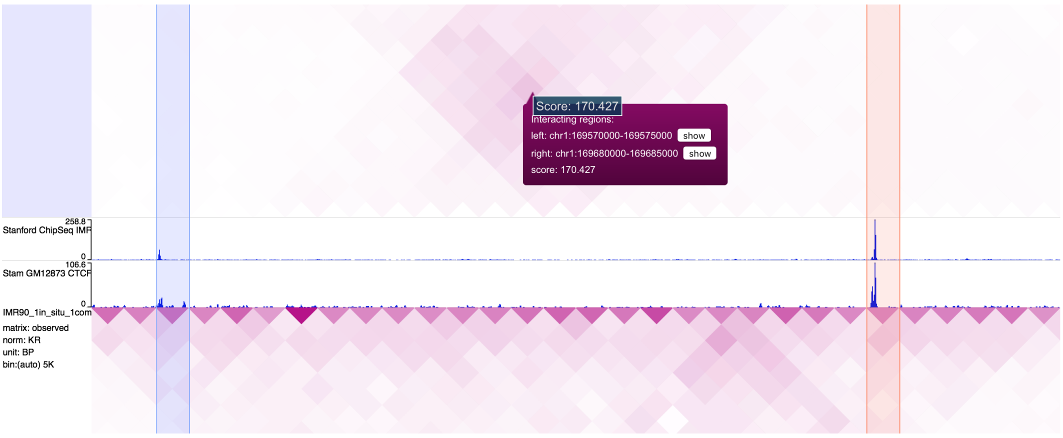

Score: 170.427
Interacting regions:
left: chr1:169570000-169575000 show
right: chr1:169680000-169685000 show
score: 170.427

Figure 167 Stanford ChipSeq for fetal lung tissue and lymphocytes and Hi-C data from Rao et. al, 2014 at 5k resolution from fetal lung and lymphocytes, show a correlation at highlighted region blue and red, which maps to the highest local measured interaction in the lymphocyte cell line. In the fetal lung tissue, the value is not as high and does show a difference in Hi-C data.

Association with RNA-Seq data

We can even integrate RNA-Seq data to show different expressions in different cell tissue types in combination with CTCF binding sites. The user can include various features to support their hypothesis or just explore local regions around e.g. a domain shown in Hi-C data. Figure 168 shows an example of Lymphocytes including CTCF binding sites, RNA-Seq data and finally the Hi-C data shown as either heatmaps or arcs.
In Figure 19, additional tracks of Genecode were added to include the ncRNA SELP and the protein SELP. In this figure, the difference between expression levels in the bottom track, corresponding to fetal ling tissue, and the top track, corresponding to blood cells (GM12878) is visible. The ncRNA SELP and the protein SELP need to be in spatial proximity in order to work; they are brought together in the CTCF binding site. As before in figure 18, the corresponding CTCF regions are highlighted pale red and pale blue; the interaction density score of 170 in the Hi-C data is visible as a local high. Summarily, it can be concluded that the association of Hi-C results with data generated by different experiments such as ChipSeq greatly supports understanding of the biological functions associated with functional elements in a genetic region.

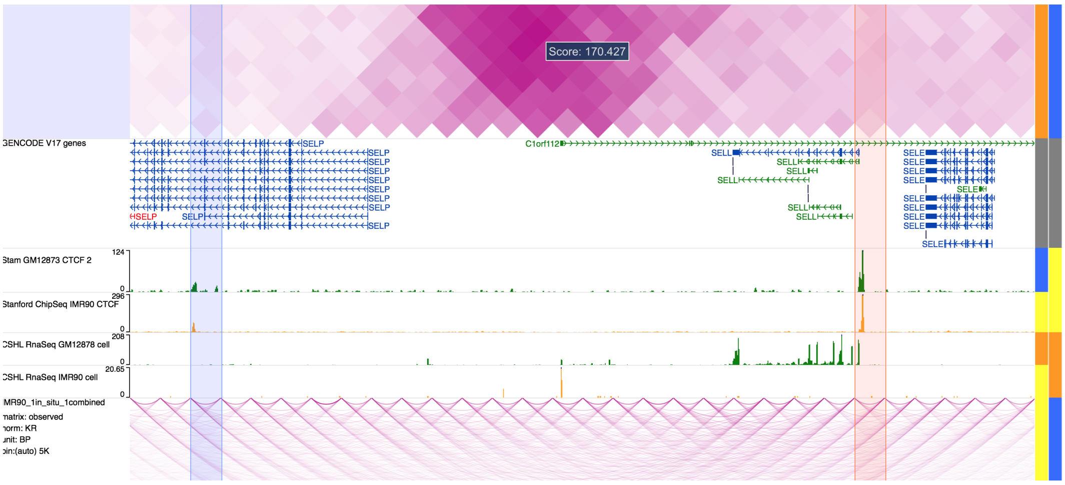


Figure 168 Combination of cell tissue lung and blood with features Hi-C data, CTCF ChIP-Seq data and RNA-Seq data. Additionally the annotation track of GENCODE is shown to visually correlate the position of the ncRNA SELL [264] and the protein SELP [265]. They are both involved within CTCF binding sites. Here we show that blood does express SELL ncRNA whereas lung does not express SELL. The Hi-C data shows a peak between regions highlighted in blue and red, where the two protein lie in proximity. It is clearly visible that IMR90 (fetal lung tissue) does not express the protein SELL. There are even no significant difference in the arc track of IMR90 as compared to the heatmap track of the GM12878 (blood cells/lymphocytes).

Finally, we can map the heatmap as a circular diagram as shown in Figure 169 to distinguish between regions. The user is able to apply additional features (annotation tracks) to associate e.g. CTCF binding sites to Hi-C data. In this particular case both tracks show blood cells with the outer track being CTCF binding sites (ChipSeq data) and the inner track being RNA-Seq data associating with the CTCF binding sites. The Hi-C data is presented as arcs and shows higher interaction in the region between the two CTCF binding sites that are correlated with the RNA-Seq data, which means that the proteins around the CTCF binding sites are expressed.

Association with Open Chromatin

Open Chromatin, can be mapped to Hi-C data. Regions that are dense and closed (closed chromatin) are shown as dense clusters in Hi-C data. Figure 170 shows two tracks of Hi-C data being at the top: IMR90 and below: K562. K562 shows less variation in interactions whereas IMR90 has open and closed regions of chromatin. Finally a Gencode track and the repeats are shown at the bottom which further strengthens the conclusion that at the region with few long range interactions (IMR90) the Open Chromatin has genes.

15.7 Conclusion (and future directions)

Since it is a new technology, the results of Hi-C experiments are currently primer used in addition to other experimental data and as helping data that support analysis of regulatory elements. The goal of Hi-C experiments is to capture chromatin conformation, which is crucial for the discovery of regulatory features such as TADs. Knowledge about these features is valuable as support of other analyses and to complement data generated by other experiments.

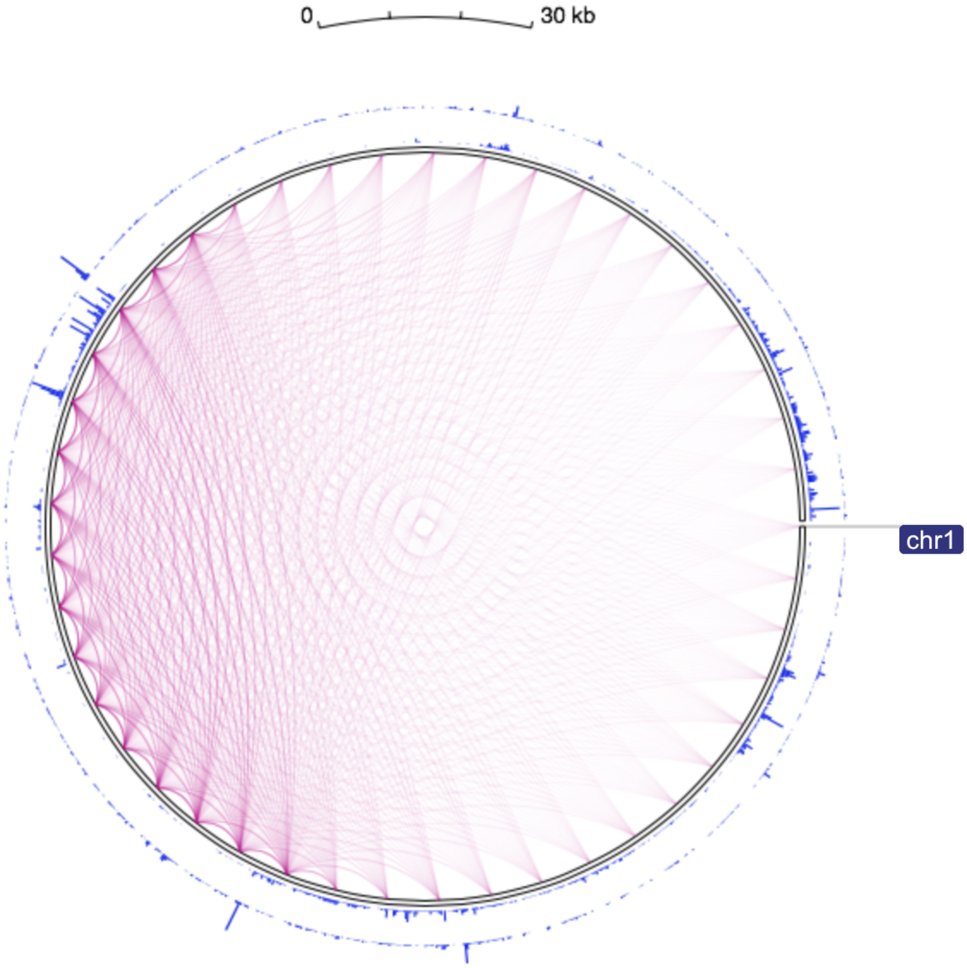


Figure 169 Circlet View: Outer track showing the CTCF binding sites of blood cells (lymphocytes, GM12873) and the inner track showing RNA-Seq data of blood cells (lymphocytes, GM12878). We can see higher probability of interactions where the CTCF binding sites reside.

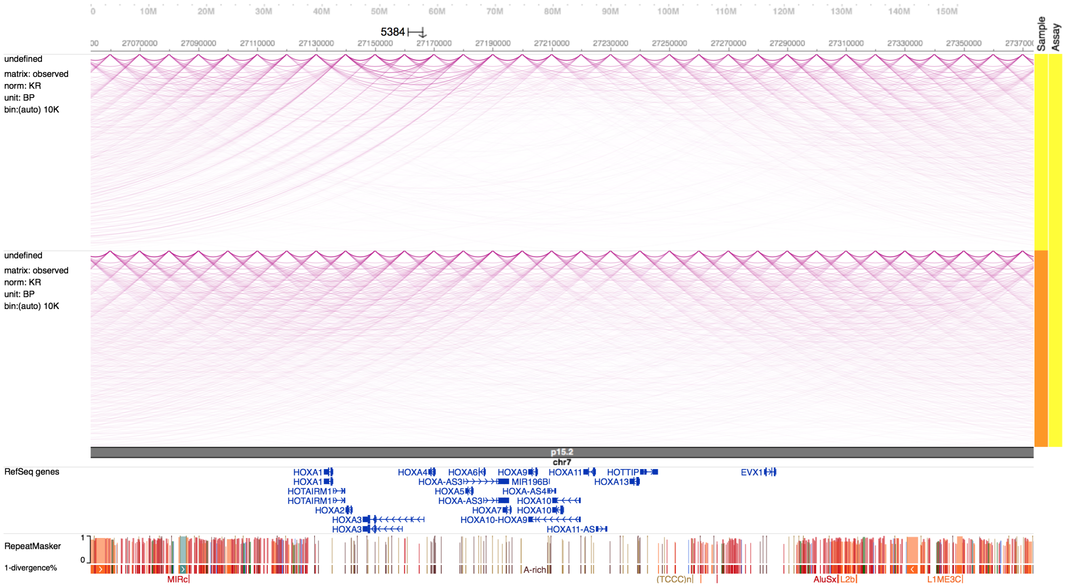


Figure 170 Arc-Top: IMR90, Arc-Bottom: K562; There are differences between cell-types: The gene-rich region shows less long-range interactions in IMR90. K562 does show less variation in open and closed chromatin. Additionally the repeats do map to dense regions including no genes. So Hi-C data associates well with Open/Closed Chromatin.

In our analysis, the choice of normalization method plays a major role in the output quality, and normalization allows better visibility of long-range interactions. For discovery of TADs, however, the use of raw data is the best choice, and over-normalization by common method is an obstacle. Also, different resolutions should be used to vizualise different regulatory features: megadomains, compartments and sub-compartments are best visible at >=25kb resolution; loops, however, are only visible at 1 kb resolution [258].
In many cases, the removal of spurious interactions has a bigger positive impact than normalization. Besides this, a sufficient sequencing depth is crucial for data usability.
[259]

16 Mass-Spectrometry based Proteomics

by Nick Lehner, Anne Hartebrodt and Constantin Ammar

With cells being able to splice genes and thereby creating alternative proteins, depending on their current specific needs, their own health status or the orgamism's health, a great interest lies in studying the abundance of these alternative spliced gene products. In this book's chapter we try to find evidence for alternative splicing, not on the level of mRNA expression patterns, but a level deeper, looking at the actual proteins present in a cell by using proteomics data from mass spectrometry experiments. To begin with, we will give an overview over the technical and computational aspects of mass spectrometry. Furthermore, we will briefly see some of the current applications of MS in proteomics, and more in detail the aforementioned alternative splicing.

16.1 Technical aspects of mass spectrometry

Mass spectrometry (MS) is an experimental technique to measure the mass(es) of a sample substance. In a common proteomics setup, a sample is first transfered to gas phase, ionized, and accelerated by an electric field. These accelerated ions are subsequently examined by a mass analyzer which separates them by their mass-to-charge ratio (m/z). This information is captured by a detector and used to create a mass spectrum, a histogram representing the detected ion flow on the y-axis and m/z on the x-axis (see Figure 171 bottom left). By comparing this measured spectrum with known masses it is possible to characterize the sample or fractions of the sample. There are many types of MS arrangements, for each section of the described work-flow there exist numerous different exchangeable alternatives each having strengths and weaknesses. For example the ionization technique can be specialized in low or high sample fragmentation, depending on the desired outcome. Furthermore, it is possible to arrange two mass analyzers in a row (MS/MS), to further dissect and analyze a sample. This arrangement is particularly interesting for this project, because it allows for direct peptide identification in a heterogeneous sample. A more in-depth description of a systematic MS arrangement and its execution, similar to the ones used to generate the data for this project, can be found in Figure 171 and its caption.

16.1.1 Technical limitations

Ideally, mass spectrometry experiments as described in Figure 171 should detect all proteins present in a cell/tissue sample. In spite of technological improvements, this is unfortunately not the case yet [268]. Several limiting factors impede the complete detection, processing and/or identification of all sample proteins. The peptide length after digestion with the protease in the process of sample preparation has a strong influence on detection probability. Peptides which are longer than about 30 to 40 amino acids can not be ionized and accelerated, due to an MS device's upper m/z limit, and are therefore not detectable. On the other extreme, peptides which are shorter than around 5 to 6 amino acids could be ionized, but they are not sequenced, because they pass through the liquid chromatography assay too fast and get discarded with salts [269]. Other peptide inherent factors may influence the ability to be ionized, such as hydrophobicity and basicity of the peptide chain [270]. Additionally, the dynamic spectrum of protein abundance in live cells decreases the chance of finding all proteins. Low abundance peptides are less likely to be selected for fragmentation in an MS/MS setup which implicates a decreased chanced of being sequenced. In human, protein concentrations can fluctuate 10-fold [268] which leads to redundant detection and identi-

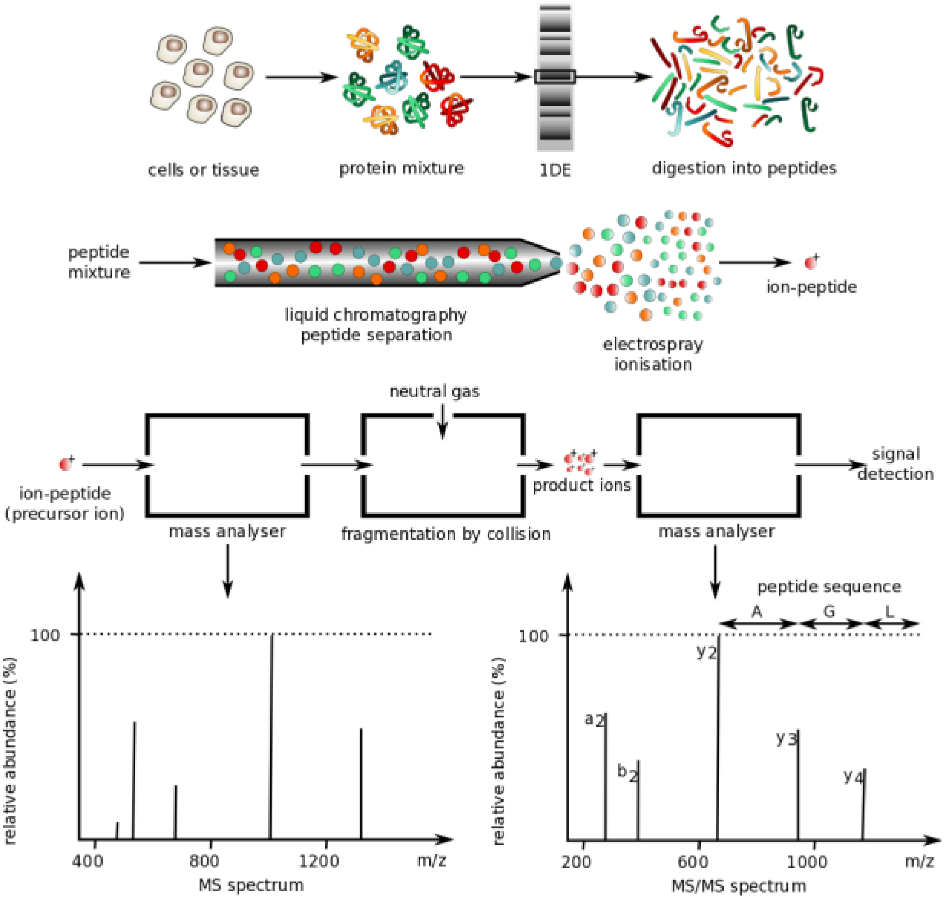


Figure 171 Setup of an MS/MS assay via liquid chromatography and electro-spray ionization [266]: First, the sample cells are fractionated, the protein part of the lysate is pulled out and run through a polyacrylamide gel to roughly separate the proteins by size. Each fraction is then digested into peptides using a protease, typically trypsin. The resulting peptide mix of different lengths is subsequently separated by affinity, size, or charge using liquid chromatography. Peptides reaching the needle at the end of the column are exposed high voltages to produce charged ions in the gas phase without further fragmentation. This precess is referred to as electrospray ionization and known as a "soft" ionization technique. The peptide ions have different mass/charge (m/z) ratios and in consequence different travel times, thus creating a distinct pattern upon arrival in the ion trap which is called an MS spectrum. Each spectrum of resolved m/z peaks is then automatically processed and the peptides corresponding to the highest quality peaks are selected for sequencing. These peptides are further fragmented via collision with a neutral gas, causing them to randomly break at their amino-bonds creating different ions. A second mass analyzer is then used to detect the resulting subsequences of the peptides, called b- and y-ions, resulting in an MS/MS spectrum. Those MS/MS spectra can then be used computationally infer the full peptide sequence [267].

fication of the most prevalent gene products [271]. Especially the detection of alternative isoforms, which are likely to be very sparsely or specifically translated, is rendered difficult by the instruments' bias towards higher abundance peptides [272].

16.2 Computational aspects of mass spectrometry

In the previous subsection we have seen, how an MS/MS-Spectrum is acquired. This is only the first step in peptide identification, the peptide sequences must still be determined from the spectra. Furthermore scores which indicate how reliable an identification is, must be calculated, and the identified peptide must be assigned to a protein. In the following, several approaches for peptide identification and their scoring schemes will be discussed. Furthermore, we will hint to some known limitations concerning the computational part of mass spectrometry.

16.2.1 Database searching – MaxQuant and Andromeda

MaxQuant in combination with the search engine Andromeda [273] uses the so-called database searching approach for peptide identification, which is a probabilistic approach. Two inputs are required, the MS/MS spectrum and a list of all protein sequences expected to be present in the sample. In a first step, all possible peptides for the masses in the spectrum are calculated based on the provided transcriptome. When calculating the theoretical masses, possible post-translational modifications of the amino acid chain are considered. The number of matches k of the actual masses to the theoretical masses is determined. Then the probability of observing at least k out of n matches by chance is computed. The Andromeda Score is the the negative decadic logarithm of this probability. (For further detail see the formula below which has been adapted from [273]). Note, that the formula that is actually applied in the software also considers side chain specific losses in the score calculation, and that the score is optimized over the number of peaks q allowed for the window of 100Th.

$$AndromedaScore = -log_{10} \sum_{k=j}^{n} \left[\binom{n}{k} \left(\frac{q}{100}\right)^{j} \left(1 - \frac{q}{100}\right)^{n-j} \right] \tag{17}$$

The Andromeda Score can be interpreted as a p-value with the null hypothesis that there is no similarity between the observed spectrum and the theoretical fragment masses. With $\frac{q}{100}$ being the approximation of the probability for a single random match, it is not an actual p-value, because the true probability of a random match is actually smaller since more that one nominal match per theoretical mass can exist. In addition to the Andromeda Score, a second score is provided, the Δ-score. This score is the difference in scores of the best and the second best peptide, and provides therefore an additional measure of reliability for the identification.

Beside the peptide identification and the calculation of the scores, MaxQuant also determines the protein an identified peptide belongs to, a process also referred to as protein identification. Peptides that cannot be unambiguously assigned to one protein, are associated with the protein with the most assigned peptides, also called razor protein, however the information of the other candidate proteins is kept [274].

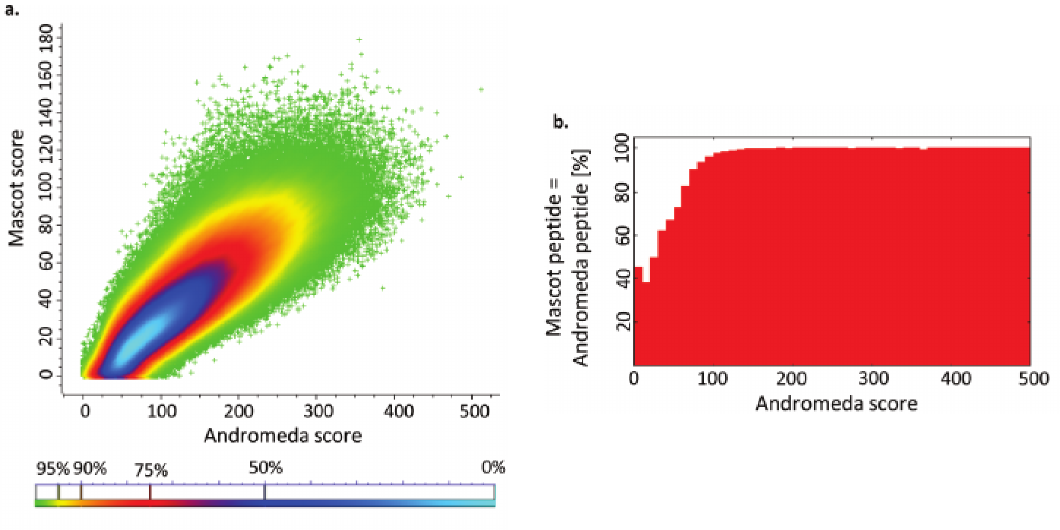


Figure 172 a) Comparison of Mascot and Andromeda Scores for the best scoring peptide. b) The set of peptide identifications obtained using MaxQuant and Mascot where the best scoring peptide is the same using both engines has been determined at varying Andromeda Scores. Above an Andromeda Score of 100 most of the highest scoring peptide identifications are the same using both methods. Figures taken from [273]

16.2.2 Other database search engines

Although the MaxQuant environment is now widely used in the proteomics community, a variety of popular alternatives exist. One of those tools is Mascot, which uses a similar method to determine the peptide sequence. Mascot is the older search engine and was part of MaxQuant before Andromeda was developed. In [273] the authors compare the Andromeda Score to the Mascot Score. They conclude that an overall good correlation between the scores for the best identification has been achieved, with the Andromeda Score being about three fold higher than the Mascot score. Furthermore, they assess how consistently the same peptides are identified as the best in both engines. They find that if the best scoring peptides have an Andromeda score higher than 100, they are consistent with the highest scoring peptides in Mascot (see figure 172). In addition to Mascot, other peptide search engines exist, like for example X!Tandem [275] and SEQUEST [276]. X!Tandem uses a probabilistic approach as well, however, the algorithm includes the distribution of the peptide masses in the search database in order to calculate the probability of a random match which avoids that larger proteins are favored in the identification process [277]. SEQUEST follows a procedure which is conceptually similar to Andromeda. The individual steps, such as preprocessing and score calculation however are different. Being one of the oldest algorithms, SEQUEST has been implemented several times, with newer versions like TIDE performing significantly faster than the original one [276].

16.2.3 Spectral Searching

An alternative concept for peptide identification is the direct comparison of curated, high quality experimental spectra with the spectrum of interest from a mass spectrometry run. This approach is referred to as spectral searching and implemented in software like SpectraST or X!Hunter [278][279]. In the case of SpectraST, the scoring metrics is a “spectral dot product" of the normalized intensity vectors. The preprocessing includes several steps which are supposed to assure that only relevant peaks (no contaminants or noise) are in the spectrum and that slightly shifted peaks can be detected. It is then followed by a normalization over the whole intensity vector and the calculation of the dot product. In addition to that score, two other values are used, the Δ-Score, like in Andromeda, and a score called dot bias which evaluates whether the score is dominated by a few peaks or by several. Those scores are combined to an overall score F using a simple equation [278]. The advantages of spectral searching over the traditional sequence based approach are an increased speed of identification, because the computationally expensive generation of theoretical fragments is foregone; higher precision due to exploitation of the characteristic of actual spectra, like the peak intensity or non canonical ions; and the possibility to include the results of multiple searches without increasing the runtime, since the reference spectra are precomputed. Obviously, one can only reliably identify peptides for which high-quality spectra are available. This means that the approach is suitable for routine identification, but might not be the method of choice for more specific tasks [280].

16.2.4 De-novo Sequencing – Overcome another sensitivity problem?

The third and last approach we would like to mention briefly is *de-novo* sequencing which is used in software like PEAKS, PepNovo, Novor and many more which have been designed for special purposes. Several ideas have been implemented in order to solve the problem. Novor for example uses decision trees which are trained on 300,000 experimental MS/MS spectra [281]. In contrast to database searching, it is not necessary to provide a transcriptome in order to determine the peptide sequence. This means, *de-novo* sequencing could in theory overcome a major disadvantage of the database searching methods – the dependency on a provided transcript database and thus the inability to identify peptides other than the ones previously known. However, as has been shown in a recent study [282], the identification rate is not yet comparable to the traditional methods. While 75% of the proteins can be identified, the number of correctly identified residues is below 50%. In the context of detection of alternative splicing which we will discuss later in this chapter, we will see that a high sequence coverage which means a high number of correctly identifies peptides, not only correctly identified proteins, is crucial. Therefore, *de-novo* sequencing is not yet a suitable method for the task at our hand [282].

16.3 Applications of Mass-spectrometry in proteomics

In the global scope of *The Regulatory Genome*, mass spectrometry can be used to study the proteomics component of the regulome. This layer of regulation is important since proteins are directly involved in metabolic processes but at the same time tight regulators of DNA and RNA related activities closing the loop to the upper levels of regulation. In this chapter we will explore some applications of mass spectrometry that can be applied to gain knowledge the structure or function of single proteins or the proteome as a whole. The study of post-translational protein modifications or alternative splicing are prominent examples for the application of MS in proteomics. Very briefly, we will see some techniques that can be used in combination with mass spectrometry and how those techniques can be used to infer protein function. This chapter aims to set the technique into the context of the Regulatory Genome before discussing the use of mass spectrometry to determine the presence of alternative splicing in more detail in the following chapter.

16.3.1 Post-translational modifications

Post-translational protein modifications (PTMs) are believed to be key players in the regulation of cellular processes. They play an important role in enzyme regulation, protein interactions and subcellular localization [283]. One of the best studied examples are the modifications of histones and their interaction with the genomic sequence. While there exists a whole variety of different post-translational modifications, only a few have been studied in a more detailed manner. Those include well known modifications like methylation, acetylation, phosphorylation and ubiquitination, but there are many more. Olsen and Mann estimate that about 70% of the proteins are phosphorylated at some point in their lives [284] which has been confirmed by a phosphoproteomics study [285].

Mass spectrometry offers the possibility to study common and more exotic modifications with an increasing speed and sensitivity. Tremendous advances have been made and are still on the way, including the study of PTMs in large scale experiments. One of the advantages of the application of MS is the possibility of unbiased study of PTMs. MS creates the opportunity to study PTM at varying cellular conditions and compare the evolutionary conservation of modified sites which could highlight functional sites from noise [283] and give insight in the speed of cellular signaling [285].

There are still challenges to overcome and unsurprisingly, sensitivity is one of them. In theory, for the identification of a protein, a single peptide matching unambiguously is sufficient. This is not case for post-translational modifications since the specific peptide carrying the modification needs to be identified. This means that the the identification success rate is decreased because not all peptide fragments will be detectable as we have seen in the technical part. Also it has become clear that a great amount of the proteome is modified post-translationally and the function of the modification remains yet to uncover.

16.3.2 Protein-Protein Interactions, Protein-Ligand Interactions

Protein-protein interaction (PPIs) and interactions of proteins with other ligands such as small molecules (e.g. pharmaceuticals) are of great interest for regulatory processes. There are several techniques that can be used in combination with mass spectrometry in order to investigate those interactions, as for example Affinity-Purification Mass Spectrometry (AP-MS), Crosslink-MS (XL-MS) and Hydrogen-Deuterium Exchange MS (HDX-MS). For schematic workflows of the experimental techniques see figure 173.

In *Affinity Purification MS*, a bait protein or molecule is immobilized on a matrix and a protein mixture is eluted over the column, such that proteins interacting with the bait proteins are captured. Those protein complexes are then purified and processed as usually in MS experiments. Using this technique, not only the interaction partners can be identified, but also the stoichiometry of the interactions can be probed. A possible application is the study of protein complexes in differing conditions which could for example hint on different subunit preferences in the different settings. From interaction partner stoichiometry one might also deduce chaperone activity of proteins, as well as classify PPIs into weak, transient and stable interactions [285][286].

Crosslink mass spectrometry uses, as the name indicates, chemical cross-links, to connect amino acids in or between peptide chains. The complexes are then digested and analyzed using mass spectrometry. Links may be made intra-molecular, between residues of the same subunit; or inter-molecular, between residues of different subunit chains. Various chemical cross-links can be made which means one cannot only link protein chains, but also proteins to RNA. XL-MS can be used to study protein-protein interactions, more specifically the interfaces of protein complexes, since links are more likely to be introduced at proximal residues. Likewise, the interaction of proteins with small molecules can be probed. Furthermore, extensive use of different intra-molecular crosslinks in combination with Cryo-EM can be utilized for investigation of the structure of proteins [287].

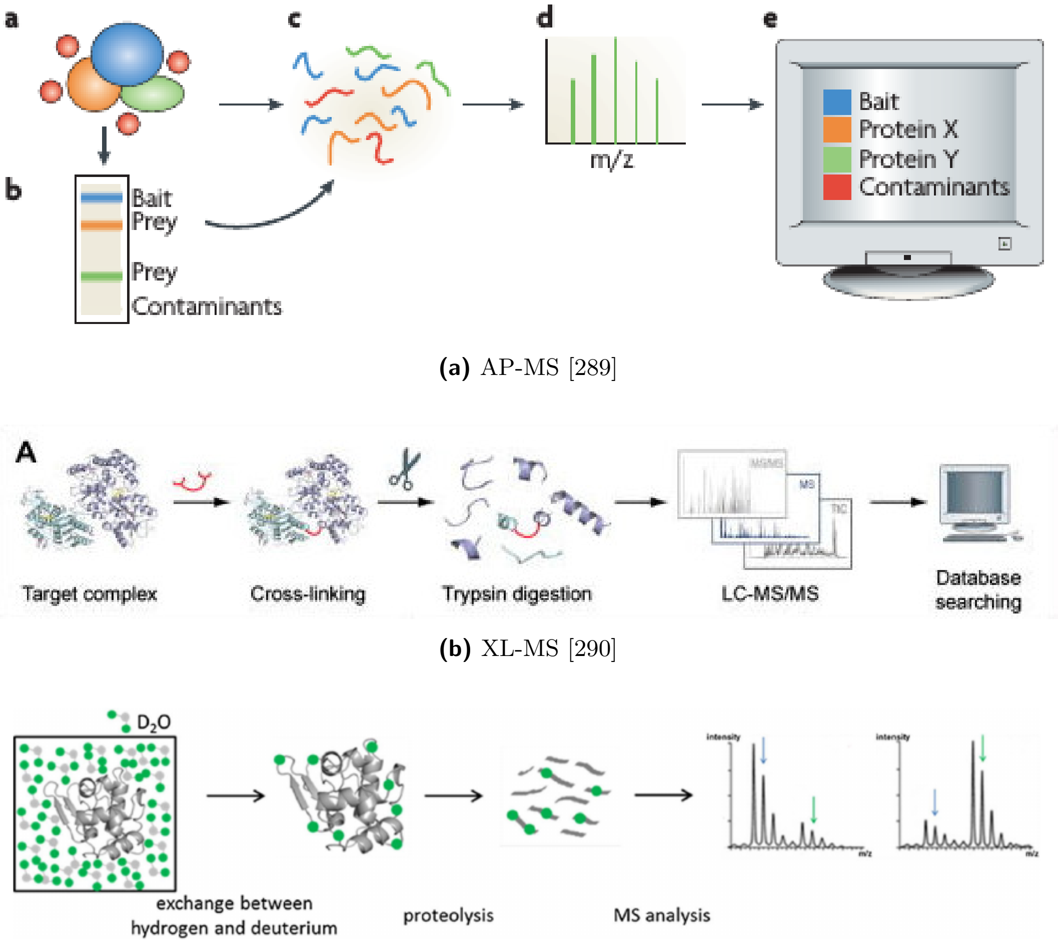
Another technique that can be used to study PPIs is *HDX-MS*. Deuterium is a stable isotope of hydrogen, carrying an additional neutron. Water containing deuterium ions instead of the lighter hydrogen ions, is referred to as heavy water. When incubating proteins in aqueous solution containing heavy water, solvent accessible side chains will exchange hydrogen ions with deuterium ions much faster than the buried residues, thus leading to slightly heavier peptide chains. In mass spectrometry, this slight difference can be detected which means the solvent accessible parts can be probed with MS. Interfaces of PPIs can be investigated, by labeling protein-complexes and the individual subunits, and from there deducing the interface of the interaction. Likewise, folding reactions or conformational changes can be studied (often using "pulse-labeling", where the heavy medium is only introduced for a very short timespan which provides a "snapshot" of the accessible surface). Of course, HDX can also be used to simply study the structure of proteins, with the advantage over X-ray crystallography that the proteins do not need to by crystallized [288].

16.3.3 Proteomics in Health and Disease

When studying the proteome, one goal could be to establish a "reference" proteome which reflects the healthy organism and consecutively can serve to identify the processes which are deregulated in the disease status. The proteome has been studied to a far lesser extent than the genome, also due to the technical difficulties we have mentioned in previous sections. Still, with the technical advances in MS based proteomics that are made today, increasing the speed and sensitivity of the procedure, the determination of the proteotype could become a standard application in clinical settings. Given that mass spectrometry runs are fast and reproducible, proteins could be used as biomarkers, for example when expressed at different levels in health and disease [285]. Another example of proteins being under the investigation as potential biomarkers is the common type II diabetes and the related cardiovascular diseases which often coincide. In order to optimize treatment, the study aimed to find a set of protein biomarkers that is able to cluster individuals into different groups according to their proteotype. Specific PTMs could be assigned to different groups, however due to the complex nature of the disease, the proteins that were selected are not yet sufficient, even when combined, to completely characterize the subtypes [291].

16.4 Case study: Proteomics and alternative splicing

The latest GENCODE release (v26 October 2016) lists 80,531 protein-coding transcripts for 19,817 human protein coding genes - more than four transcripts per gene on average [292]. Experimental assays like RNA-seq or microarrays are able to prove these isoforms' existence on the transcriptome level, but this level of proof is often not deemed sufficient to conclude a gene's isoform presence in a cell. Transcriptome to proteome relationship, although connected by the translation process may sometimes not be simplified to: more



(a) AP-MS [289]

(b) XL-MS [290]

(c) HDX-MS [288]

Figure 173 Schematic workflows of the different mass spectrometry coupled techniques mentioned in the text. Figure reference can be found in the subcaptions.

transcript equals more protein. In the steady state a cell's differences in protein levels can be sufficiently explained by transcript concentration, but once the cell diverges into another phase, the levels are much harder to correlate [293]. Another example for this can be translation on demand, where the mRNA level stays the same after a stimulus, but the corresponding protein's level rises independently [285]. The aforementioned mass spectrometric peptide identification allows for a more direct view into the proteome and allows a paradigmatic shift away from RNA expression [294][295].

16.4.1 Alternative splicing

Splicing messenger RNA in different ways can lead to a multitude of differing exon sets and their respective protein products as shown schematically in Figure 174. Thus, there is not only one, but several transcripts (isoforms) per gene to investigate. These isoforms can differ from each other by sequence and resulting protein structure in varying degree, from mostly the same to gravely modified. Altered isoforms can display different binding affinities to ligands, change in enzymatic activity or influence protein localization. For example a transcript with an exon which impairs the binding site in the folded protein can have severe consequences for the functionality of the resulting protein, if not spliced out (Figure 174, "Cassette exon"). The occurring isoforms of a gene can play a role in disease, tissue specific expression and regulation [295]. Additionally, alternative splicing of transcripts is often claimed to be a major factor in eucaryotic complexity [296].

The molecular mechanism of splicing pre-mRNA in eukaryotes is an extensively covered subject. Splice locations in a transcript are marked by two specific di-nucleotides, one at the 5' exon-intron junction (GU) and one on the 3' intron-exon junction (AG). Near the 3' junction, there is the so called branch point and a poly-pyrimidine sequence. First, this branch point conducts a nucleophilic attack on the 5' exon end di-nucleotides GU and detaches the 5' exon from the intron. In this intermediate step, the 5' intron end, now connected to the branch point forms a loop which itself is still attached to the 3' junction. The final step to exclude the intron from the mRNA is another nucleophilic attack carried out by the loose 5' exon end on the first nucleotide after the 3' intron-exon junction (AG) [296] which results in a mature mRNA and a detached intron. This mature mRNA can be further processed in the translational pathway.

16.4.2 Alternative splicing - a controversial topic

A recent publication by Tress et al. addressed the question, whether alternative splicing documented at the transcript level is also present at the protein level, and came to the conclusion that there is only one single main isoform for the vast majority of human genes. They conducted a systematic analysis of eight proteomic data sets generated by MS/MS experiments and compared the evidence for alternatively spliced isoforms between the proteomic and available transcriptomic data [295].
Ezkurdia et al. observe that the quality control and the control of the false discovery rate has been neglected in the early proteomics projects and call for more rigorous false discovery rate control. They suggest researchers using those data sets should proceed with extra caution [298]. Consequently, to reliably find splicing events within MS/MS experiments, Tress et al. implemented several filtering steps to lower incidence of false-positive identification. This filtering was carried out in the following way:

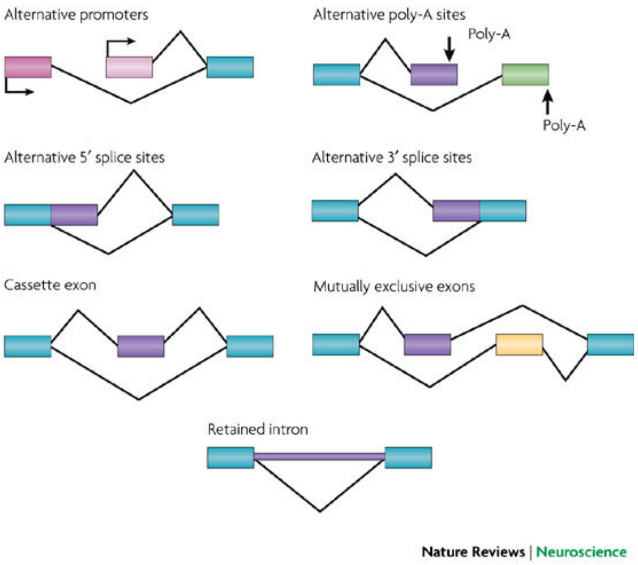


Figure 174 Different splicing patterns: Splicing can occur at different positions in the unspliced pre-mRNA and can alter the series of exons. Here, blue colored boxes are exons which are not affected by splicing, while other exons' color indicate an occurring splice event. The lines connecting two exons represent splice paths which depending on which exon is spliced out are on top or below the exons. These splice paths can be at the start and end of the transcript or in-between two unaffected exons [297]

.

1. Exclusion non-tryptic and semi-tryptic peptides (Peptides that have no or only one trypsin-specific terminus).
2. Missed cleavages that occur when a possible cleavage site is missed by the protease resulting in longer peptides are only allowed if a fully cleaved peptide is present.
3. Leucine and Isoleucine are considered to be the same amino acid.
4. Peptides are only included if they were identified by at least two search engines. Exception from this rule were peptides which had an Andromeda score above 100, because it can be assumed that the would also have been identified by Mascot (see Figure 172: b).
5. Removal of all peptides which are only present in one dataset.

After this filtering was applied to all eight datasets, they still found peptide evidence for 12,716 protein-coding genes, but more than 98% of these genes had evidence for only one single main isoform. Only 246 genes were found to have detectable peptides proving alternative splicing at the protein level. Following this surprising result, they deducted their conclusion that there is a single main cellular isoform for almost every gene. Subsequently, to evaluate whether there was a systematic mistake in their approach, they estimated the expected number of alternative splicing events on in-silico digested GENCODE20 data. The results of this simulation indicated that there should be roughly 5- to 15-fold more genes than they detected in the experiments, but still not to an extent transcriptomic profiling suggest.

Shortly after Tress et al. published their findings, their approach was commented by Benjamin Blencowe, who pointed out several systematic flaws and referenced contrasting findings on the same subject. His main arguments for the reason for Tress et al. finding so few alternative isoform evidence were the lack of peptide detection by MS/MS (undersampling) and the discarding of biologically relevant tissue-specific constraints. Blencowe also pointed out further evidence of alternative splicing sites detected by ribosomal profiling [299].

16.5 Our project

In the previous subsection, we have seen that alternative splicing is indeed a topic of controversial discussions. As detailed above, MS data can be used to make assumptions on alternative splicing on the protein level, rather than predict it from RNA. In the following, we will investigate data from publicly available resources and try to make our own conclusions on alternative splicing on proteome level. At the same time, we will draw parallels to the claims of Tress et al. that we have mentioned previously [295].

16.5.1 Data

There are two publicly available data resources that have been generated in an effort to characterize the human proteome, created by two groups that have simultaneously published their efforts [300][301]. We will call the datasets "Kim" and "Wilhelm" in the remainder of this book, after the first authors of the initial publications. We have an additional dataset containing sequence data from human myeloid leukemia cells [302] which we will call "Lamond". All those datasets have been analyzed using the MaxQuant Software, but using different proteomes. The Kim and Wilhelm datasets were processed using the UniProtKB reference proteome of 2015 [303] and have been provided with [304], while for the Lamond data RefSeq [305] has been used. The data can be downloaded from the PRIDE archive. Furthermore, we obtained data for four tissues from the Kim lab which were processed with the software Mascot and did an additional analysis of one of the Kim tissues with MaxQuant, but using the Ensembl proteome (version GRCh38) [306] instead of UniProt. The results of the execution of those programs are "evidence" files which contain the peptides, scores and other measures generated by the respective program.

16.5.2 In-silico digest and candidate peptide identification

In order to identify which peptides in a mass spectrometry run are proof for alternative splicing, in a first step we determine a list of all the candidate peptides that indicate the existence of more than one isoform. To this end, we perform an *in-silico* digestion of the proteome, by applying the appropriate cleavage rule [307]. Since the protease does not always work perfectly, and sometimes it does not cleave at a possible cleavage site, we include the possibility of *missed cleavages*. In the case study, we allow two missed cleavages in our digest, since the peptide calling programs generally use this setting. The list of peptides that is generated in this process needs then to be processed further, to find those peptides which can discriminate an isoform from another. Such a peptide needs to fulfill two requirements. First, for a given gene and two given isoforms, the peptide can only be present in one of the isoforms. Second, the peptide cannot appear in any other gene in the proteome. This alone is not enough to prove the existence of alternative splicing if said peptide is found, but for the approach we chose to implement, these candidates are sufficient. The list we obtain from this process are the peptides that can in theory discriminate two isoforms, so the next step is to check, whether we do actually find those peptides in our mass-spec experiments.

16.5.2.1 Excursus: used proteases

As mentioned in the technical part the protease mostly used for the digestion of the protein samples is trypsin, but as we have seen before the peptide lengths produced by this enzyme are not always optimal. Since for the identification of splice events, the identification of a protein is not sufficient, we need the specific peptides, as we have described before. This means the yield of eligible peptides in the experiment would need to be increased. In order to investigate whether the use of proteases other than trypsin would increase the number of theoretical peptides and thus whether a different protease would be better suited for our proceedings, we compiled the candidate peptide list for different proteases (ArgC (R), AspN (N), GluC (D), LysN (K)) which were already under investigation in [269]. Trypsin is known for producing small peptides of under six amino acids length which elute too quickly over the chromatography column and can not be detected. The other proteases mentioned above generate longer peptides which are detectable in mass spectrometry. In order to assess whether different proteases could potentially be better suited for the detection of alternative splicing we conducted a small *in-silico* test: We defined the longest isoform as the major isoform, which is arbitrary but justifiable, because the focus is on the fraction of peptides that map to different isoforms. We then calculated and plotted the cumulative distributions of the fractions of peptides mapping to shorter ("minor") isoforms for each protease. The fraction of peptides that map to minor isoform is the smallest for trypsin (see figure 175). This indicates that the use of alternative proteases could indeed increase the probability of finding a transcript discriminating peptide simply because the fraction of eligible peptides is greater.

16.5.3 Mapping of in-silico candidate peptides to MS/MS evidence

In the previous section we identified candidate peptides which are able to distinguish alternative isoforms for each gene. The next step was to match all these theoretically possible peptides to the real peptides detected by Kim and Wilhelm. Each in-silico peptide detected via MS/MS experiments was saved, all remaining peptides that had no MS/MS evidence, were discarded. In other words, we intersected all possible peptides with measures MS/MS peptides. In case an in-silico peptide was present multiple times in a MS/MS dataset, every instance of that peptide was saved.

16.5.3.1 Mapping files: Statistical evaluation

The resulting overlap between in-silico candidate peptides and MS/MS experimental peptides can be seen in Table 12. Wilhelm et al. had more total and unique candidate peptides matching an experimental peptide than Kim et al. lab. When comparing the difference of total peptides and unique peptides from Wilhelm et al. to Kim et al., we found that although Wilhelm et al. had almost one third more peptides mapped in total, the number of unique peptides only differs around 13%, revealing more redundant peptide evidence for candidate peptides in Wilhelm et al. These candidate peptides belong to 5,775 genes in Kim et al. and 5,740 genes in Wilhelm et al., with an average of 1132.2 and 1780.5 genes per dataset (tissue) on average.

16.5.4 Splice identification: Major/minor isoform

As outlined in our introductory section, detection of alternative splicing is of great interest. Tress et al. claim, that most human genes have one main protein isoform [295]. They

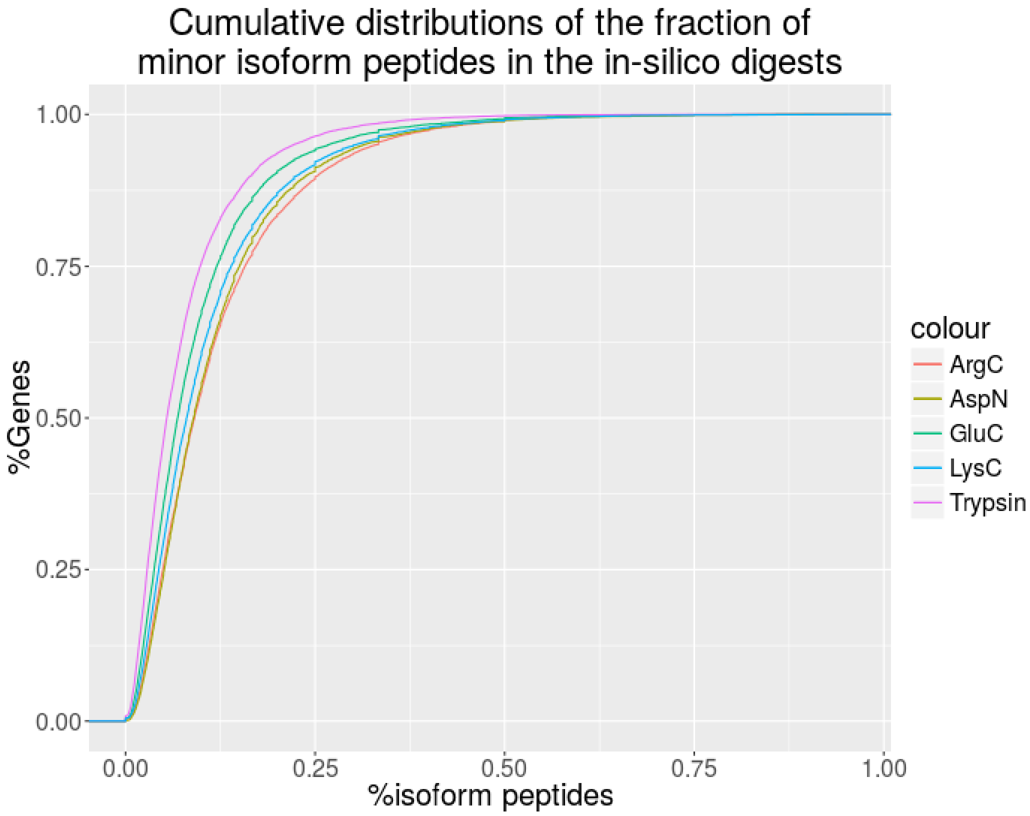


Figure 175 Cumulative distribution of the fraction of minor isoform peptides in the in-silico digests using different proteases. The x-axis shows the percentage of minor isoform peptides that map to a gene, and the y-axis shows the percentage of genes, for which more than one transcript are available in the Ensembl proteome (15,000). We observe that the trypsin curve has the steepest curve, which corresponds to the smallest fraction of minor isoform peptides.

	Peptides mapped	Unique Peptides	Unique Genes	Avg. # of Genes
Kim et al.	6,538,479	69,821	5,775	1132.2
Wilhelm et al.	10,220,472	79,752	5,740	1780.5

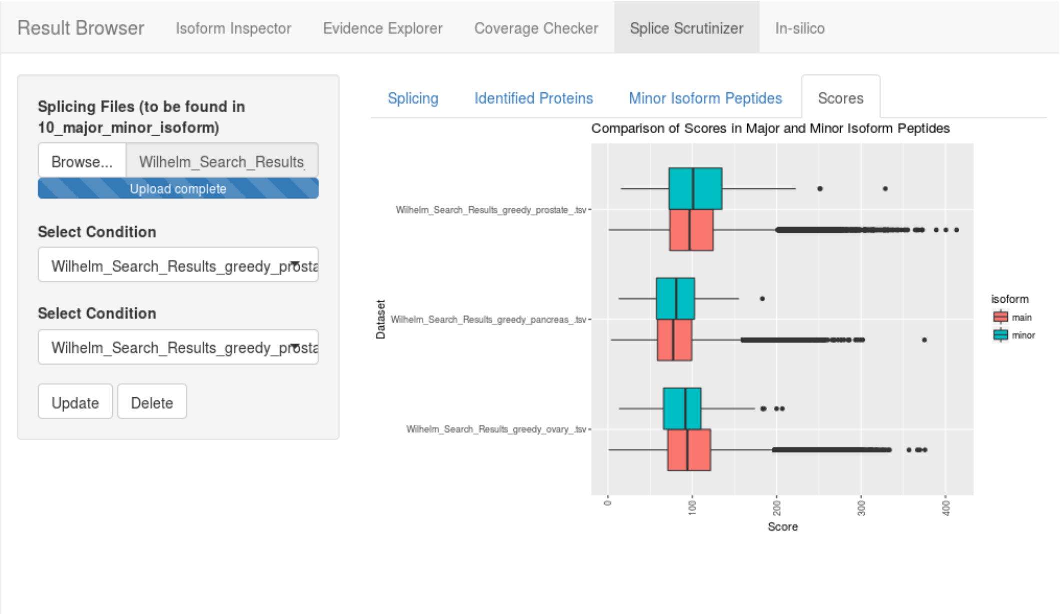
Table 12 Basic statistical analysis of in-silico digested candidate peptides mapped to MS/MS evidences: In-silico digested peptides from the uniprot proteome [303] were mapped to peptides found in MS/MS experiments from two different labs - Kim et al. [300] and Wilhelm et al. [301]. The columns show: 1. Total number of in-silico digested peptides mapped to MS/MS peptides; 2. Number of unique peptides from (1.); 3. Unique genes which had MS/MS peptide evidence; 4. Average number of those genes per tissue.

acknowledge the experimental proven existence of alternative transcripts, but do not see sufficient mass spectrometry-based evidence for translation of these transcripts. We investigate MS/MS data (1.5.1) to draw our own conclusions about the presence or absence of alternative gene products. To this end we define the term major isoform as the isoform we assume to be the most abundant one in the given cellular context. Consequently, any isoform that is less abundant is termed minor isoform. If evidence for peptides from major as well as minor isoform(s) are found, we identify a potential alternative splice event. A candidate peptide (1.5.3.1) will be either mapped to the major isoform if possible or to a minor isoform if no match with the major isoform can be found. To define which isoform is considered major/minor, two different approaches were implemented. The first approach was to choose the longest isoform of a gene as the main isoform. Peptides which did not match to a subsequence of this isoform, but matched another isoform were labeled “minor peptides”. If two or more transcripts had the same length, we picked in alphabetical order to maintain consistency. Although the decision to take the longest transcript of a gene as the major isoform has no biological background, Tress et al. found their main cellular isoforms to be the longest in 89.6% of the cases [295]. This approach will from now on be called “longest” approach.

The second approach was to count, for each gene, which isoform had the most mapped peptides. Whichever isoform had the most peptides, was labeled the major isoform. This approach is a lower estimation of major peptides under the assumption, that there are no unknown isoforms. Just as for the “longest” approach, if two or more isoforms happened to have an equal number of peptides, the isoform which came first alphabetically was chosen. Peptides which were associated with that gene, but did not match to the major isoform, were labeled “minor peptides”. For later comparison and analysis, the Andromeda score associated with each peptide was saved as well. In the following we will call this approach “greedy”. Both approaches yielded a list of genes, which contains: the number of peptides on the major/minor isoform, the major/minor isoform(s) and the Andromeda scores of each peptide.

16.5.5 User interface

Our project includes the development of a graphical user interface (GUI) which can be used to browse the data we create in an interactive manner and can be seen as a complement to this article. There are several tabs containing different modules. **Isoform Inspector** allows the user to specify a gene and two transcripts which are visualized according to their genomic exon structure. If there are peptides in the evidence file, those get aligned to the corresponding region in order to highlight the eventual presence of peptides that prove the existence of alternative isoforms. **Evidence Explorer** can be used to obtain basic information on one or more mass spectrometry runs, such as peptide length and compare the experimental peptide length to the in silico generated ones. **Coverage Checker** provides the possibility to compare the proteome coverage of different evidences, in order to get a feeling for the fraction of sequence that is typically covered by peptides in an MS experiment. **Splice scrutinizer** is the core module of our result browser. Here we provide basic statistics on the identified splice events in an interactive manner. The user can choose different results to compare. For two datasets, the user can compare the number of proteins that have been identified in the mass spectrometry runs as well as the number of genes, for which alternative isoforms have been detected. The fraction of peptides mapping to minor isoforms is highlighted and the scores of the peptides can be compared. Figure 176 shows a screenshot



Splicing Files (to be found in 10_major_minor_isoform)

Wilhelm_Search_Results

Upload complete

Select Condition

Wilhelm_Search_Results_greedy_prosta

Select Condition

Wilhelm_Search_Results_greedy_prosta

Figure 176 Screenshot of the result browser with example data loaded. The tab that is currently open visualizes the scores for the peptides mapping to major (orange) and minor (green) isoforms. The modules are briefly decribed in the main text.

of one of the tabs.

16.5.6 Pipeline settings

Our goal is not only to detect splice events using proteomics data, we also want to compare different parameter settings that have been applied during the procedure and which are likely to influence the results, or have already been shown to influence the results. For instance, we have seen in the section about the computational part of mass spectrometry that MaxQuant and Mascot do not always identify the same peptides from the raw spectra. Also different sequence databases are being used by different research groups during the peptide identification. Questions one might also ask are whether the identified splice events are consistent in the same tissue, but with the data coming from different labs, which implies slight differences during the experimental procedures, even when taking the uttermost care to follow the protocols. In the following we will have two types of tissue comparisons. The first one is to only compare equivalent tissues, for example two liver samples from different labs. We will call this comparison “equal”. Secondly a comparison of all tissues against all others can be made (comparison=“all”). In table 13 you will find an overview over the different parameter settings that we chose.

16.5.7 Results and Discussion

16.5.7.1 Identified Splice events – Total numbers

With the two approaches mentioned above, we identify a varying number of alternative isoforms in the different datasets. For the greedy approach the numbers range from 6 (approach=longest: 50) potential splice events in the adult liver dataset using Mascot to 384 (approach=longest: 732) in the Lamond leukemia dataset. See table 14 for the complete

fixed parameters		variable parameter
database	UniProtKB (2015)	Dataset: Kim vs. Wilhelm
program	MaxQuant	
approach	greedy	
comparison	equal	
dataset	Kim	Program: Mascot vs. MaxQuant
database	UniProtKB(2015)	
approach	longest	
comparison	equal	
dataset	Kim	Proteome: UniProtKB(2015) vs. Ensembl
program	MaxQuant	
approach	longest	
comparison	equal	
dataset	Kim	tissue: Kim all vs. all
program	MaxQuant	
approach	longest	
database	UniProtKB(2015)	

■ **Table 13** Parameter settings: Overview over the different comparisons made, with all parameters but one fixed. Proteomes: UniProtKB (2015) and Ensembl; Programs for peptide calling: Mascot and MaxQuant; Approaches for splice identification: longest, with the longest isoform fixed as the main isoform and greedy, with the isoform with the most mapping peptides as the main isoform; Comparisons: equal, where only matching tissues are compared, and all, where each tissue is compared against each.

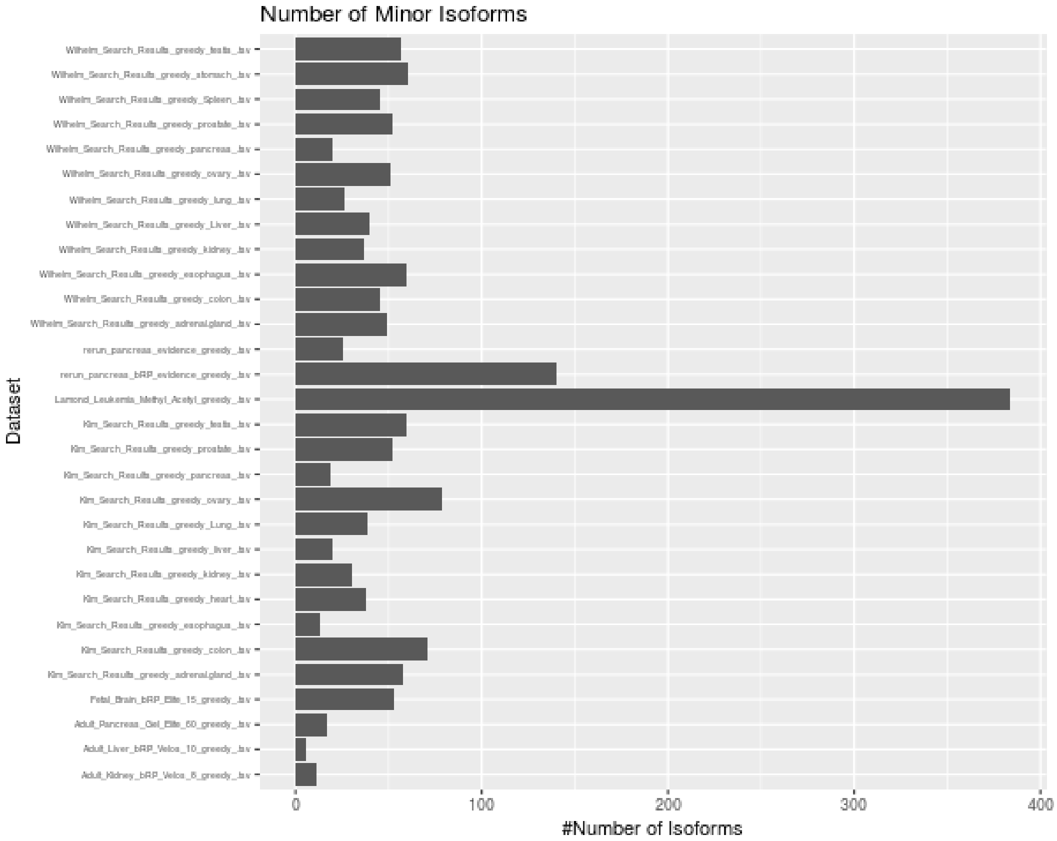


Figure 177 Total number of proteins identified in the different datasets and the number of proteins having an alternative isoform when using the greedy approach. Using the longest approach leads to considerably more minor isoform identifications (not shown here). This is not surprising because there is no biological indication that the longest isoform is the major isoform and using this assumption artificially increases the number of minor isoforms compared to the greedy approach.

list and figure 177 for a visual overview of the number of alternative splice events using the greedy approach. When combining the results of all the datasets we processed with UniProt we obtain 566 genes with alternative isoforms when using the greedy approach. For the combined Kim and Wilhelm datasets we only obtained 264 alternatively spliced genes, which is only slightly more than the 246 alternatively spliced genes that Tress et al. identified in their approach. However, in our analysis we only use a fraction of the mass spectrometry runs from the human proteome project, in contrast to Tress et al. who, in addition to the higher number of runs issued by Kim et al. and Wilhelm et al., use complementary data sources. We can, of course, not easily interpolate the number of alternative splice events in the whole dataset from this small number of MS runs, but it is to expect that we would find a greater number of potential splice events when using the whole dataset. By including the additional 300 mass spectrometry runs of the Lamond dataset, we already drastically increased the number of events.

16.5.7.2 Mapping peptides and scores

Each major and minor isoform is identified by a number of peptides that map to the respective isoform. We have computed the distribution of the number of peptides that map to the

Dataset	#Proteins	greedy	longest
Wilhelm_Search_Results_greedy_adrenal.gland_.tsv	4486	49	186
Wilhelm_Search_Results_greedy_colon_.tsv	4266	45	171
Wilhelm_Search_Results_greedy_esophagus_.tsv	4586	60	215
Wilhelm_Search_Results_greedy_kidney_.tsv	3429	37	150
Wilhelm_Search_Results_greedy_Liver_.tsv	4085	40	165
Wilhelm_Search_Results_greedy_lung_.tsv	3185	26	138
Wilhelm_Search_Results_greedy_ovary_.tsv	4188	51	200
Wilhelm_Search_Results_greedy_pancreas_.tsv	3465	20	116
Wilhelm_Search_Results_greedy_prostate_.tsv	3896	52	189
Wilhelm_Search_Results_greedy_Spleen_.tsv	3696	45	143
Wilhelm_Search_Results_greedy_stomach_.tsv	4297	61	176
Wilhelm_Search_Results_greedy_testis_.tsv	4597	57	215
Kim_Search_Results_greedy_adrenal.gland_.tsv	4278	58	192
Kim_Search_Results_greedy_colon_.tsv	4902	71	174
Kim_Search_Results_greedy_esophagus_.tsv	1995	13	60
Kim_Search_Results_greedy_heart_.tsv	2665	38	132
Kim_Search_Results_greedy_kidney_.tsv	3247	30	141
Kim_Search_Results_greedy_liver_.tsv	2660	20	105
Kim_Search_Results_greedy_Lung_.tsv	3354	39	141
Kim_Search_Results_greedy_ovary_.tsv	4689	79	223
Kim_Search_Results_greedy_pancreas_.tsv	2640	19	86
Kim_Search_Results_greedy_prostate_.tsv	3609	52	165
Kim_Search_Results_greedy_testis_.tsv	4156	60	198
Lamond_Leukemia_Methyl_Acetyl_greedy_.tsv	8945	384	723
rerun_pancreas_bRP_evidence_greedy_.tsv	8939	140	611
rerun_pancreas_evidence_greedy_.tsv	3978	25	198
Adult_Kidney_bRP_Velos_8_greedy_.tsv	3344	11	71
Adult_Liver_bRP_Velos_10_greedy_.tsv	2964	6	50
Adult_Pancreas_Gel_Elite_60_greedy_.tsv	3193	17	108
Fetal_Brain_bRP_Elite_15_greedy_.tsv	5636	53	181

Table 14 Overview over the number of identified proteins, and the number of proteins that have a minor isoform according to both approaches “greedy” and “longest”. We observe that the “longest” approach leads to the identification of significantly more alternative isoforms, which is likely due to the artificial determination of the major isoform.

major isoform and the one of the number of peptides that only map to minor isoforms for each gene for in the different datasets. While the medians of the distributions for the major isoform peptides range around 5 (with some exceptions) the majority of the minor isoforms are only identified by one or two peptides (see figure 178). In contrast to this clear difference between major and major isoforms when looking at the number of peptides, the distributions of the scores of minor peptides do not differ in such a way. The Andromeda Scores range between 50 and 150 in both datasets.. The datasets that contain Mascot Scores instead of Andromeda Scores have a median of about 50, which corresponds to an Andromeda score of roughly 150.

Two factors play into the small number of minor isoform peptides that we observe in the plots. First of all, we select only peptides that exclusively map to the minor isoform, which means we omit peptides that also map to the potentially large sections of the isoform that are equal to the main isoform. Second, the minor isoform is possibly present in lower abundance, which means that the respective peptides are harder to capture in the mass spectrometry run due to the technical limitations we have discussed earlier. When we look at the score distribution we do not observe a considerable difference in the scores of the major and minor isoforms. This means that the minor peptides are as reliable as the major peptides, which we esteem noteworthy, because while the number of identifications is small they are still eligible to identify an event and do not need to be omitted due to bad quality.

16.5.7.3 Fraction of minor isoform

In our *in-silico* digest (see section 16.5.2) we have calculated the theoretical fraction of peptides mapping to the minor isoforms for the different transcript databases. For instance, the Ensembl proteome digest yielded 8 969 770 unique peptides of which 494 798 mapped only to a minor isoform (UniProtKK, 6 385 09 and 419 200 respectively). We have about 5,5% and 6,5% of minor isoform peptides in-silico. Comparing those numbers to the actually observed fraction of minor isoform peptides we see that we observe a much smaller percentage in our experimental evidence. For the longest approach, we observe about 1% of minor isoform peptides in all datasets except for the ones that were processed using the Ensembl transcripts. The fractions are twice as high for those two datasets (more than 2%). The greedy approach leads to much lower fractions which resolve around 0.5% or below for all the datasets including the ones run with Ensembl. Figure 180 shows the fractions for each dataset and both approaches.

The fact, that the longest approach yields much higher fractions of minor isoform peptides is not surprising. Since we predetermine the major isoform as the longest one from the beginning, we do not include any experimental evidence in this decision process. This means the longest isoform is not necessarily the one the most peptides map to, which in return artificially increases the number of minor isoforms. This effect is especially strong when using our pipeline with the Ensembl database. Since the number of isoforms is only much higher (compared to the use of other databases) when using the longest approach and not when applying the greedy settings, we conclude that the high number of identifications must be an artifact. Also, the choice of the longest isoform has no biological foundation and is therefore to be handled with care and is even less justified when using databases containing predicted proteins. In our report, we include this evaluation for reference, since Tress et al. claim that the major isoform in most cases coincides with the longest isoform.When looking at the the number of minor peptides that are produced with the greedy approach, which is the most conservative estimation, based on the assumption that no other isoforms than

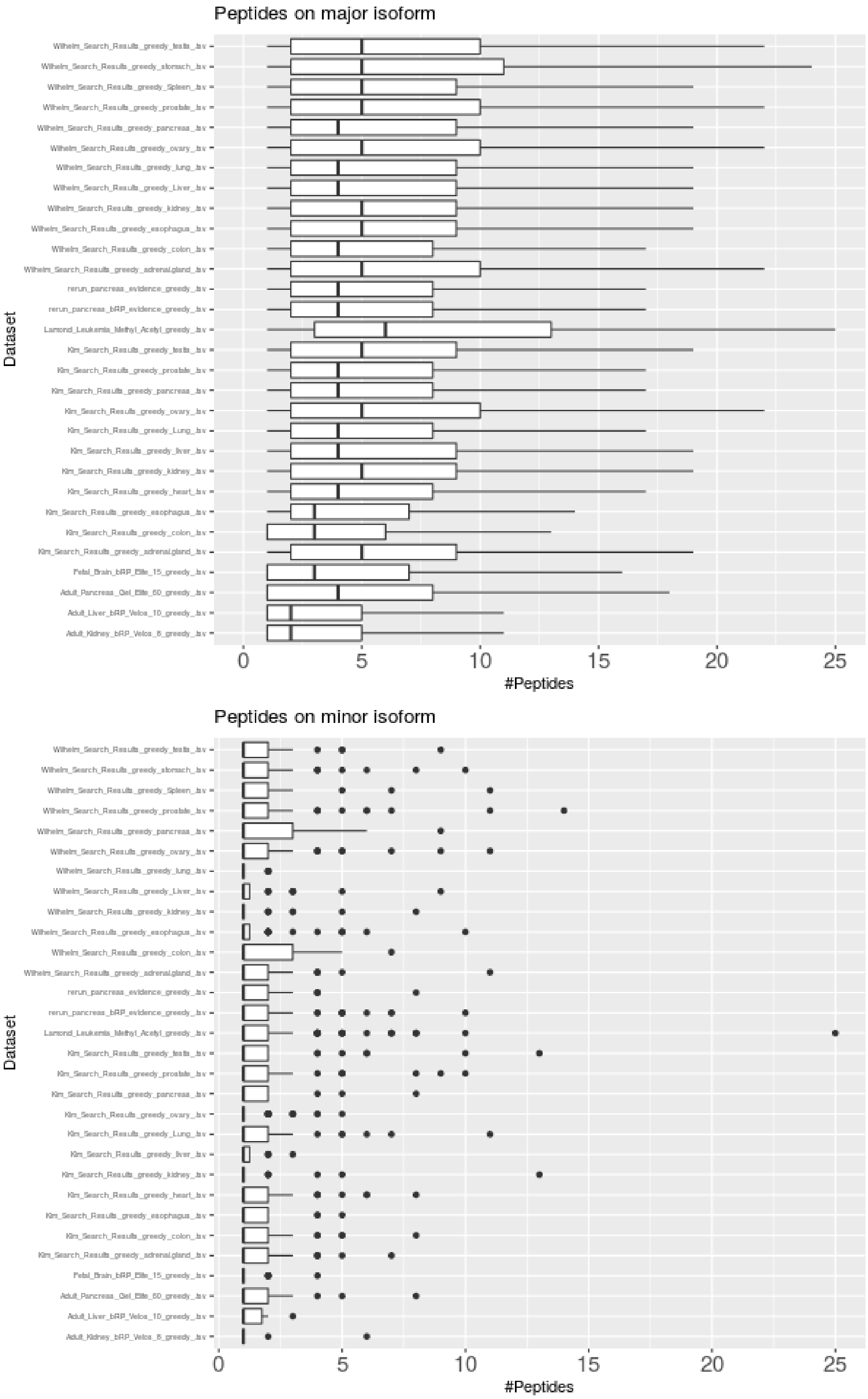
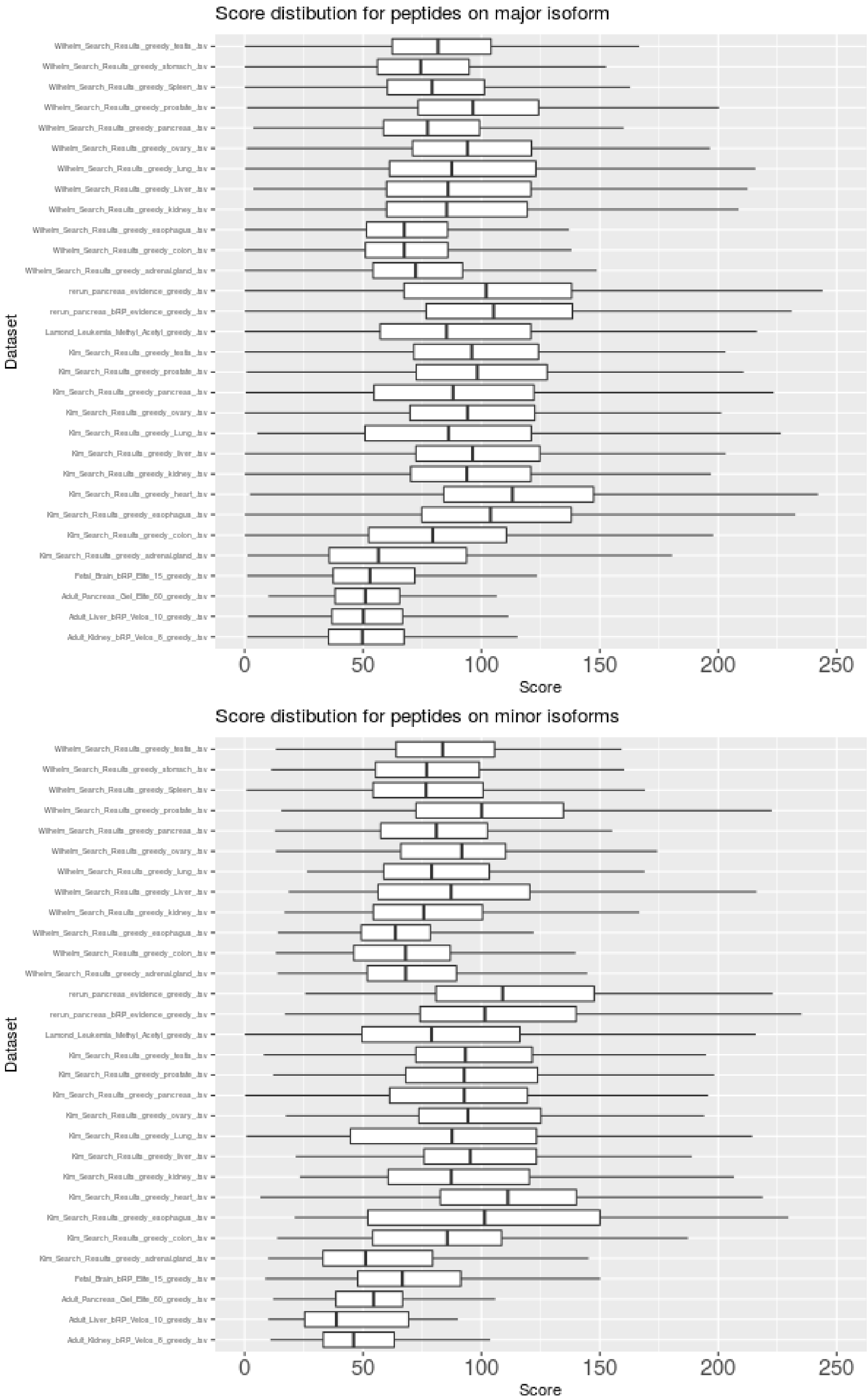


Figure 178 Number of peptides mapping to major and minor isoforms. Note the different x-axis scales. The number of minor isoform peptides seems quite small, a side effect of omitting the peptides that also map to the main isoform.


Score distibution for peptides on major isoform

Score distibution for peptides on minor isoforms

Figure 179 Distribution of scores for the peptides on the major isoform and for the peptides on the minor isoform using the greedy approach for all datasets (Kim, Wilhelm, Lamond, Kim_Ensembl and Mascot). Note, that the four lowest box-plots visualize data generated by Mascot and have therefore about three fold lower scores, than Andromeda.

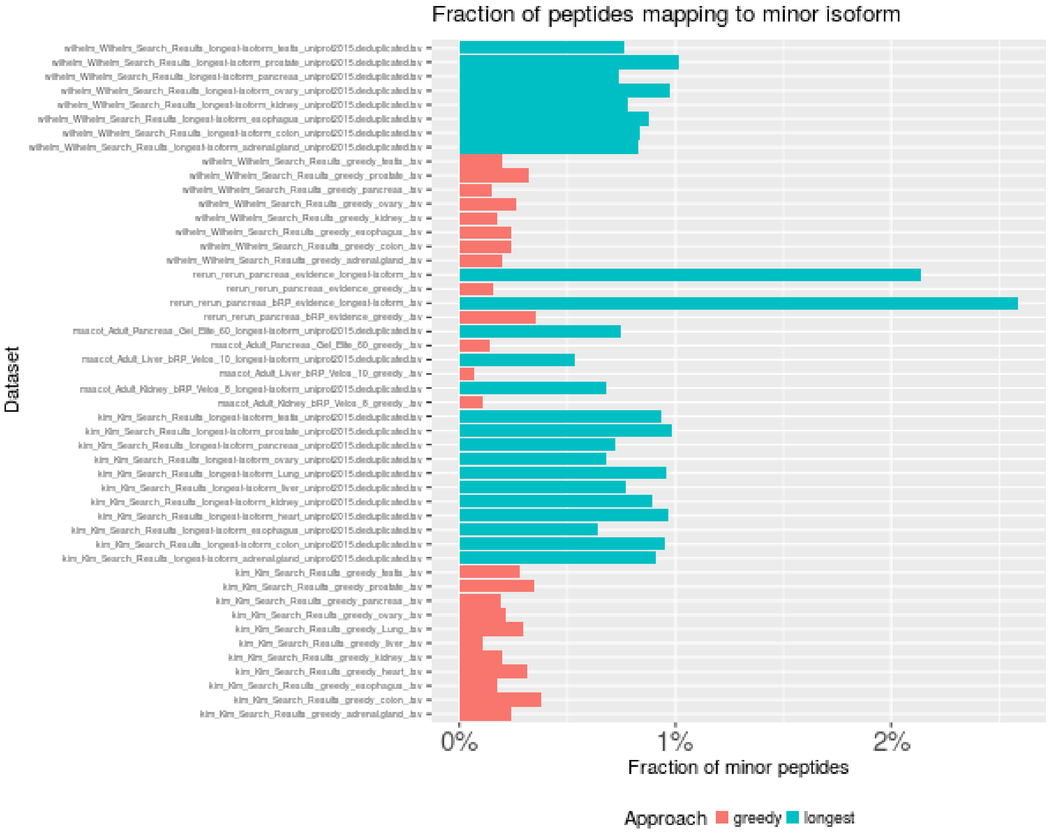


Figure 180 The figure shows an overview over the fraction of peptides that map to minor isoforms for the different datasets for both approaches. The "longest" approach is shown in blue, the "greedy" approach in red. For the greedy approach, we observe fractions of below 0.5% while for the "longest" approach, the number are about 1%. Interestingly, the fraction of minor isoform peptides is almost twice as large when running the pipeline using Ensembl and choosing the longest approach, which is likely an artifact resulting from the high number of transcripts in the database used.

the ones reported exist, we see this number decreased in contrast to the longest isoform. Furthermore, we compare the percentage of minor isoform peptides in the experiment to the fraction that we can theoretically observe. We see that the the fraction of theoretical is between 10 and 20 fold higher. However, the transcript databases we use contain a variety of transcripts that have been measured in different conditions. We do not expect to find all transcript versions of a gene in one cellular condition. This means that the estimation of the fraction of minor isoform peptides that we make based on the transcript databases is much higher as the fraction we expect to actually see in an experiment, even if several tissues are combined. Additionally, the technical constraints of mass spectrometry lead to an decreased probability of a minor isoform peptide to be detected.

16.5.7.4 Tissue-wise comparison of the greedy approach for different labs

The Kim and Wilhelm datasets contain peptide evidences for the same tissues that we can compare in order to evaluate how consistently we can measure alternative splicing in different runs from different labs. As the peptide search engine, transcript database, splice identification approach and tissue are the same, and only the lab where the data was generated

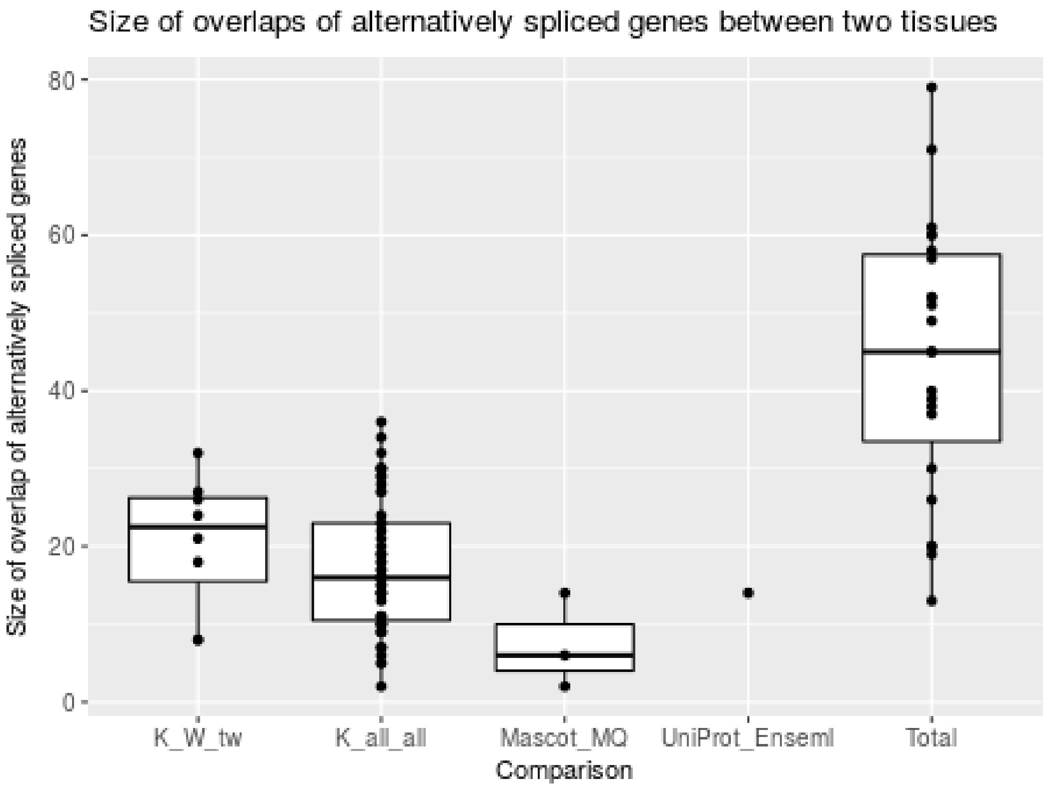
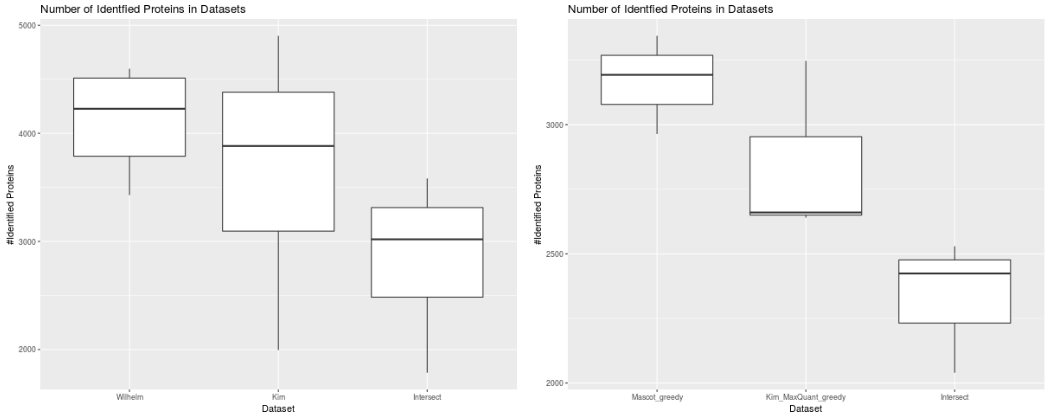


Figure 181 The boxplots show the distribution of size the overlap of the identified alternatively spliced genes (using the greedy identification approach), when doing pairwise comparisons of the result sets. K_W_tw shows the distribution of the number of potential alternative splicing events when comparing the same tissues from different labs, but with otherwise identical pipeline setting. The second boxplot, K_all_all, is an all against all comparison of the overlap of splice events between different tissues from the same lab (Kim). Mascot_MQ contains the overlap when using the same original data (3 tissues from the Kim lab), but using different peptide identification programs. Otherwise the pipeline settings are the same. UniProt_Ensembl is an exemplary comparison of the splice events when running the pipeline with identical data but executing the peptide calling program using different transcriptome databases. The last boxplot (Total) shows the distribution of the number of identified alternative splice events in all Kim and Wilhelm tissues, to set the other distributions into context.

differs, we can estimate the overlap of identified splice events we obtain from different data origins. In figure 182ba) the boxplots shows the overall number of proteins that have been identified in the mass spectrometry experiments for the different tissues split by lab, as well as the overlap of the identifications. We see that the overlap of the identified proteins in the two datasets is only a subset of the two individual ones, in both labs there are proteins that are unique to the respective dataset. We then evaluated the number of genes having minor isoforms that can be identified in both labs, which ranges between 15 and 25 depending on the tissue. In figure 177 the total number of identified splice events for the different labs are shown. As mentioned before number of identified events varies among the different tissues. We also compared the overlaps between different tissues of the Kim dataset against each other. In this scenario, we are sure that the parameters we set for the pipelines are identical up to the lab specific experimental procedures. We then proceeded in a similar fashion and calculated the overlap of the genes, having alternative splice events for each pair of tissues. We observe that the overlap of potential events is larger between the same tissues coming from different labs than the overlap between different tissues from the same lab (see figure 181, "K_W_tw" and "K_all_all").
There are several considerations, we need to make in order to evaluate the numbers for the comparison of the labs. Even when in theory following the same experimental and computational steps, there are various aspects that can lead to differing results. First of all, the technical appliances might not be exactly the same or calibrated in a slightly different way. As we have seen in the technical part, the identification of the peptides in the mass spectrometry run is far from being deterministic, so whether a peptide is identified or not is partly due to chance and this has an especially strong influence, when we are talking about low abundance peptides. The variability of the process is clearly visible when looking at the number and the overlap of protein identifications alone, where we see that there are proteins that have only been identified in one of the labs. Protein identification is the more reliable and more reproducible task in comparison to splice detection, because in theory one peptide is sufficient to prove the existence of a protein in a sample. Since we assume that alternative isoforms generally have a lower abundance than the main isoforms and we have the additional constraint of a limited number of peptides that unambiguously identify an isoform, the reproducibility of the identification of minor isoforms is obstructed by the statistical nature of the experimental process. In this light, the relatively small overlap of potential alternative isoforms between the two labs is not surprising. As mentioned before the overlap of two different tissues of the same lab is on average still smaller than the overlap of the same tissue from different labs. This could mean that some alternative isoforms are more consistently found in certain tissues and could hint towards tissue specific alternative splicing which has been previously reported on RNA level [308]. This however,would need to be confirmed by determining the exact isoforms that can be observed in these datasets.

16.5.7.5 Comparison of the peptide calling programs

As mentioned earlier we do not only compare different labs and different tissues, but we also investigate the effect of using different peptide identification programs. The median overlap of the number of protein identifications is below both of the medians of the individual datasets (see 182b). The size of the overlap of the proteins having alternative isoforms is also quite small being below 15 and thereby smaller than the overlap that we observed between different labs (using the same tissue, proteome and program) (see figure 181, second and third boxplot). The fact, that the overlap of alternatively spliced genes identified form identical raw data using two programs is smaller than the overlap of matched tissues coming from



(a) Same tissue, different lab

(b) Mascot and MaxQuant

Figure 182 a) Comparison of the number of identified proteins in the Wilhelm dataset (left) and the Kim dataset (right) using MaxQuant, the UniProtKB transcript database and the greedy approach, such that matching tissues are compared to each other. Intersect visualizes the number of proteins that have been identified in both pipelines. b) Comparison of the number of identified proteins when using different peptide identification software, in this case MaxQuant and Mascot, but otherwise identical pipeline settings (tissue, transcript database, greedy approach).

different labs is indeed surprising and indicates that not only the choice of technical setups determines the result, but also the computational part is crucial for the outcome. It has been shown in the literature (see figure 172) that the consistency of the peptide identifications using different programs is a matter of debate and the issue has been addressed by Tress et al. by only accepting peptides identified by two different engines, which on the other hand seems quite strict. There is a minor inconsistency since a different proteome (RefSeq) has been used in the Mascot runs while for the MaxQuant runs UniProtKB was used. This could lead to a decreased size of the overlaps, because the transcript databases are not identical and the search space is therefore reduced to the intersection of both.

16.5.7.6 Using different proteomes

We have made an attempt to investigate the influence of the use of an different transcript database by rerunning MaxQuant for one of our tissues using Ensembl instead of UniProtKB for the same raw data (figure 181, “UniProt_Ensembl”). However, this analysis can only serve as an example since we could not process enough data to make a representative comparison in the time given. The size of the overlap seems to be in the same order of magnitude as the overlaps when making tissue wise comparisons or comparing different search engines. We had difficulties mapping the UniProt identifier to the Ensembl transcript identifiers due to incompleteness of the matching tables we obtained from the UniProt mapping service. Some events might have been identified using both pipeline settings, but had to be discarded because the correct mapping of the identifiers is missing. In order to give a conclusive answer which is less dependent on the input we would need to add more datasets to the comparison and improve the mapping of the identifiers.

16.6 Conclusion

In the previous sections we have seen the technical and computational aspects of mass spectrometry and its limitations. We have seen that these limitations are especially important if we want to prove the existence of alternative splicing on proteome level, since we need to identify specific peptides that we additionally expect to be present in a lower abundance than the peptides belonging to the main isoform. Still in our project, we could show that we can use peptide evidence to identify potential alternative splicing events and we have shown that there are actually quite a few alternative isoforms depending on the tissue and condition under investigation. Also we established that the number of minor isoform peptides we found differs by an order of magnitude of the theoretical estimation for the fraction of eligible peptides in the transcript databases. Given the technical limitations of mass spectrometry and the dynamic range of the protein concentration in the cellular context itself, we can conclude that the measurements are in accordance to the theoretical estimates. As our project was partly motivated by the opinionated article by Tress et al. [295], we would like to comment their claims in the light of our own results. In their article they calculate an estimate for the number of alternative splice events and claim that this number is still an overestimate of the number of detectable isoforms. Our own *in-silico* study could not confirm that this theory. As we have elaborated above, the number of potential splice events we obtained using our pipeline is in accordance with the estimate we calculated in our own *in-silico* study.

Mass spectrometry based detection of alternative splicing on proteome level is dependent on multiple factors and seemingly, at the moment, it is still difficult to unambiguously determine whether alternative splicing on proteome level exists on a larger scale. In the light of the current technical limitations, we need to keep in mind is that the absence of evidence is not the evidence of absence.

17 Breast Cancer Subtype Classification

by Ines Scheller, Elena Mankina and Evi Berchtold

17.1 Introduction

In general, cancer is a complex disease that is characterized by cancer cells growing out of control and their possible invasion of different tissues. Cancer cells accumulate various somatic mutations that enable this abnormal growth. In the year 2000, Hanahan & Weinberg [309] proposed that most cancers develop in a multi-step process and that during their development cancer cells need to acquire characteristic hallmark capabilities that eventually enable their uncontrolled growth. Each of the characteristic capabilities can be acquired in several different ways via mutations in different genes. These hallmark capabilities of cancer refer to alterations in physiological processes and lead to the following properties of cancer cells: they are able to grow even in the absence of external growth factors (self-sufficiency in growth signals), they no longer listen to external anti-growth signals (insensitivity to anti-growth signals), they have found ways to escape their destruction via the pathway of programmed cell death (evading apoptosis), they can divide further even after reaching the normal replicative limit of generations (limitless replicative potential), they can stimulate surrounding tissue to form blood vessels that supply them with oxygen and nutrients (sustained angiogenesis) and eventually form metastases and invade other tissues (tissue invasion & metastasis).

In 2011, two further hallmarks were proposed by those authors [310]: the evasion of immune destruction and the reprogramming of the energy metabolism.

17.1.1 Breast cancer

As breast cancer is a very common cancer, especially in females, there are more than 1,300,000 cases and 450,000 deaths recorded in the world each year [311]. There are two types of invasive breast cancer: ductal and lobular carcinomas. Ductal tumors develop in the milk ducts of the breast and are the most common type. Lobular carcinomas develop in the milk-producing lobules or glands in the breast and account for only about 10% of all cases [312]. We will focus on ductal breast cancer in the following sections, since lobular breast cancer is rare.

17.1.2 Breast Cancer Subtypes

Breast cancer is a complex disease and breast tumors are very heterogeneous, so it is important to define breast cancer subtypes, for which different therapeutic strategies can be followed. Traditionally, clinicopathological variables like tumor size, tumor grade and lymph nodal involvement were used together with classical immunohistochemistry markers like estrogen receptor (ER), progesterone receptor (PR) and epidermal growth factor receptor 2 (HER2) for the prognosis of the patient and the assignment of a treatment [313].

If the tumor cells of a patient express ER or PR, they can be treated with endocrine therapy, while cells with a high expression of HER2 can respond to anti-HER2 therapies, e.g. with the antibodies trastuzumab and pertuzumab that inhibit the dimerization of HER2 with other members of its receptor family [314]. For patients with tumors without ER, PR or HER2 expression, chemotherapy is the only option [311].

With the availability of gene expression profiles, intrinsic molecular features have been used

Intrinsic subtype	IHC status	Grade	Outcome	Prevalence
Luminal A	[ER+|PR+] HER2-KI67-	1|2	Good	23.7%
Luminal B	[ER+|PR+] HER2-KI67+	2|3	Intermediate	38.8%
	[ER+|PR+] HER2+KI67+		Poor	14%
HER2 over-expression	[ER-PR-] HER2+	2|3	Poor	11.2%
Basal	[ER-PR-] HER2-, basal marker+	3	Poor	12.3%
Normal-like	[ER+|PR+] HER2-KI67-	1|2|3	Intermediate	7.8%

Table 15 Summary of the characteristics of breast tumor molecular subtypes, taken from [313]

to group tumors into subtypes. The standard of intrinsic breast cancer subtyping has been defined by Sørlie et al. [315]. According to them, five distinct subtypes can be identified: luminal A, luminal B, HER2 over-expression, basal and normal-like tumors (see Table 15). Other studies have proposed different classifications, e.g. with four or six subtypes and sometimes different names. But in general, there are three major classes that are identified: luminal (expressing ER), HER2 over-expression (expressing HER2) and triple negative phenotypic tumors (TNP or basal, no expression of either ER, PR or HER2). Within these three classes, the triple negative tumors are the most heterogeneous ones [313]. The Cancer Genome Atlas Network analysed tumor and germ line DNA samples from 825 patients and found that numerous genes were mutated frequently, including many genes previously reported in the context of breast cancer, but also many novel ones [311]. Several of these frequently mutated genes showed subtype-specific mutation patterns (see Figure 183) with regard to mRNA-expression-based subtypes as describes above (Luminal A/B, Her2, Basal). In this study, the overall mutation rate was highest in basal-like tumors and lowest in luminal A tumors.

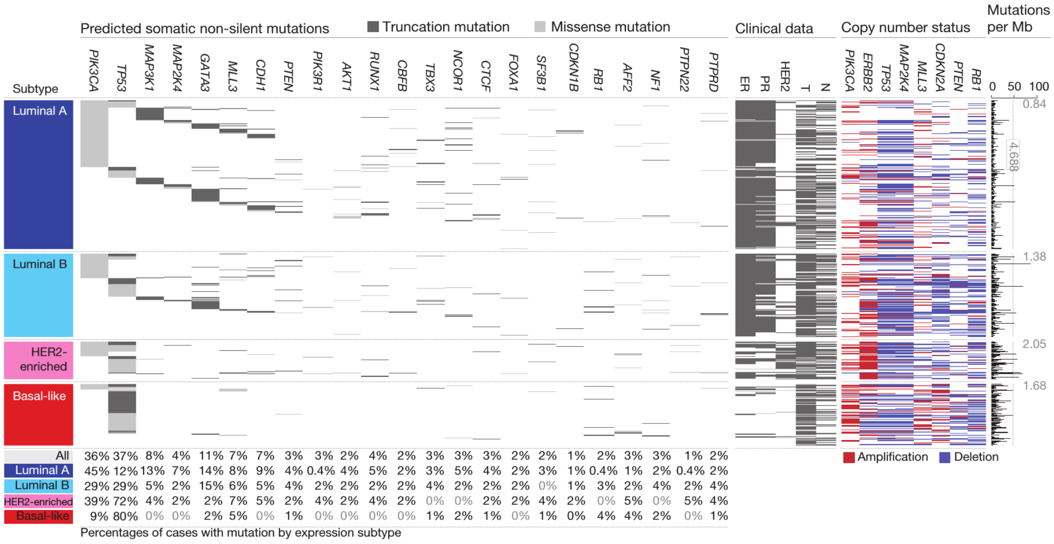


Figure 1 | Significantly mutated genes and correlations with genomic and clinical features. Tumour samples are grouped by mRNA subtype: luminal A ($n = 225$), luminal B ($n = 126$), HER2E ($n = 57$) and basal-like ($n = 93$). The left panel shows non-silent somatic mutation patterns and frequencies for significantly mutated genes. The middle panel shows clinical features: dark grey, positive or T2–4; white, negative or T1; light grey, N/A or equivocal. N, node status; T, tumour size. The right panel shows significantly mutated genes with frequent copy number amplifications (red) or deletions (blue). The far-right panel shows non-silent mutation rate per tumour (mutations per megabase, adjusted for coverage). The average mutation rate for each expression subtype is indicated. Hypermutated: mutation rates >3 s.d. above the mean (>4.688, indicated by grey line).

Figure 183 Frequencies of significantly mutated genes in the different mRNA-based breast cancer subtypes, taken from [311].

17.2 Materials & Methods

17.2.1 Datasets

In our analysis, we used 15 breast cancer datasets. These datasets were downloaded from Bioconductor (breastCancerVDX [316], breastCancerTRANSBIG [317], breastCancerUNT [318], breastCancerUPP [319], breastCancerMAINZ [320], breastCancerNKI [321]), ExperimentHub [322] (TCGA BRCA [323]), curatedBreastData [324](GSE12071 [325], GSE12093 [326], GSE16391 [327], GSE17705 [328], GSE22226 [329], GSE25055 [330], GSE25065 [330]) or Haibe-Kains et al. [331] (EXPO). These datasets contain 46-1.082 samples (TCGA (BRCA) is the largest dataset, containing 1.082 samples) and 22.293-54.675 genes. All sets except of EXPO contain survival data: overall survival (OS), distant metastasis-free survival (DMFS) or relapse-free survival (RFS). The datasets NKI and TRANSBIG contain information on all three survival types, UNT contains information on DMFS and RFS and GSE22226 - OS and RFS. All remaining datasets contain informations on only one survival type.

There are several datasets containing only untreated patients: VDX, TRANSBIG, UNT, MAINZ, GSE12093. In the dataset GSE17705 all patients are treated with hormonal therapy, while in GSE25055 all patients were treated with chemotherapy. The remaining studies contain patients, who received different types of treatment or combined treatment (including chemotherapy, hormonal therapy and radiotherapy). The treatment has an influence on the survival of the patients, but not on the expression data (as the samples are taken before treatment starts). For the dataset EXPO the treatment of the patients is not available.

Table 17 contains the breast cancer datasets with additional informations.

17.2.2 Survival analysis

17.2.2.1 Survival studies

In breast cancer studies, an outcome of interest can be the time interval to an event (survival data). In such survival studies, an event may be the recurrence of a tumor or the death of a patient. There are three survival types, which are important in breast cancer studies: overall survival (OS), distant metastasis-free survival (DMFS) and recurrence-free survival (RFS). The patients are usually recruited and followed up to a fixed date over several years, in order to assess whether the event of interest happened or not. After a follow up, some patients will not have experienced an event for different reasons: 1) the event just has not yet occurred and one does not know if and when it will occur; 2) the patient could not be followed up because he exited the study; 3) a competing event occurred, which makes a follow up impossible [332]. For these patients, the survival time is called 'censored'. Due to the censored data, specific survival analysis techniques are necessary, as such data still contributes important information and should not be excluded from the study. Hence, the aim of a survival study is to compare the survival times (including the censored times) between pairs of patient groups.

17.2.2.2 The Kaplan Meier method

The Kaplan Meier method is one of the techniques applicable to survival data. This method calculates conditional probabilities for each time interval: the probability that those patients who are alive at the beginning of the interval, will survive to the end of the interval [333]. The survival at a specific time point k is then calculated as the product of these conditional probabilities of surviving each time interval until this time point:

$$S(k) = p_1 \cdot p_2 \cdot p_3 \cdot ... \cdot p_k, \tag{18}$$

where p_1 is the proportion surviving the first period/interval, p_2 is the proportion surviving at the end of period 2 conditional on having survived up to the second period, and so on until the k-th time interval. The proportion p for a time interval i is given by:

$$p_i = \frac{r_i - d_i}{r_i}, \tag{19}$$

where r_i is the number of individuals alive at the beginning of i and d_i is the number of occurred events within period i [334]. The periods usually relate to days. Censored observations at a specific time point affect the number at risk at the start of the next time interval.

The Kaplan Meier plot represents the survival probabilities as a survival curve. A survival curve is a step function with sudden changes corresponding to the times when an event was observed. The times, where censored data occurs, are usually marked by a tick. Figures 184 and 185 show examples of Kaplan Meier plots showing one survival curve and comparing two survival curves, respectively.

There are several important assumptions which are made here: 1) patients whose data was censored have the same survival prognosis as patients who are still followed up; 2) the survival probabilities are always the same, no matter whether the patient was recruited at the beginning of the study or later; 3) the event occurred at the specified time (when it was observed).

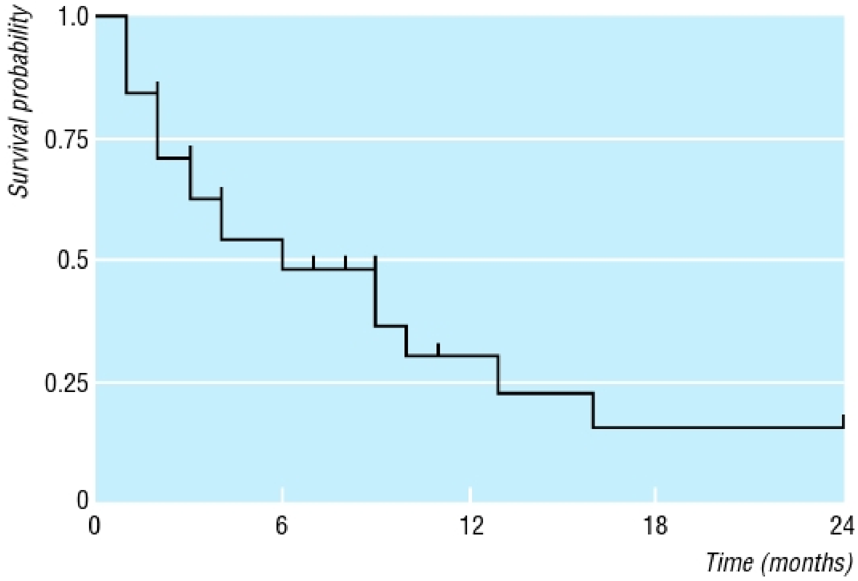


Figure 184 A Kaplan Meier Plot showing one survival curve with censored data. Censored data is marked by a horizontal line. Figure taken from [333].

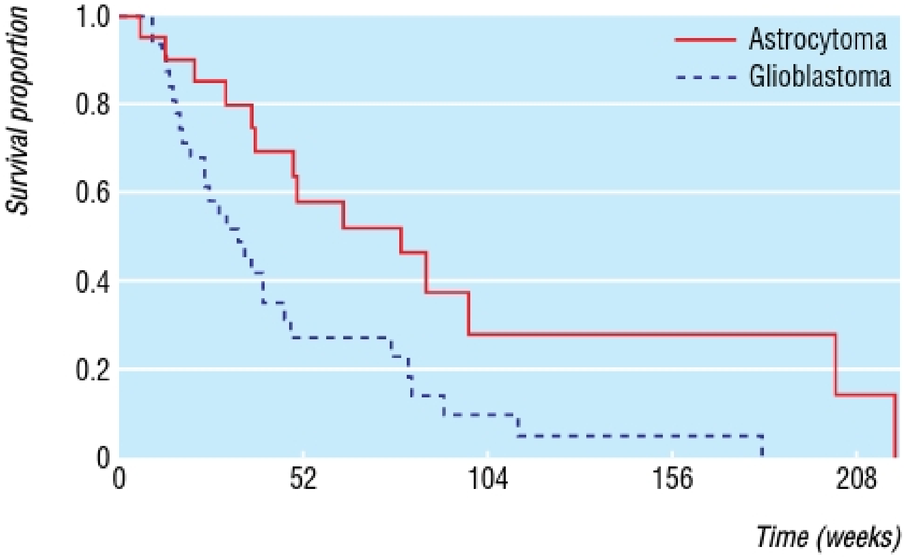


Figure 185 A Kaplan Meier Plot showing two survival curves comparing patients with glioblastoma and astrocytoma. The figure shows, that patients with astrocytoma have a higher survival rate than patients with glioblastoma, as after 180 weeks 0% of patients with glioblastoma are alive compared to about 30% of patients with astrocytoma. Figure taken from [335].

17.2.2.3 The Hazard rate

The hazard rate *h(t)* is the probability of experiencing the event of interest in the following time interval t divided by the length of this interval (if the event has not already occurred) [336]. The hazard ratio is the ratio of two hazard rates of two different groups:

$$Hazard\,ratio = \frac{h_2(t)}{h_1(t)}, \tag{20}$$

where $h_i(t)$ is the hazard rate of group i [337]. It measures how high the risk of experiencing an event is in group 2 compared to group 1. A hazard ratio of 1 indicates, that the risk is approximately equal in both groups, hazard ratio > 1 indicates, that the risk of event is higher in group 2 than group 1 and a hazard ratio < 1 indicates, that the risk of event is lower in group 2 than in group 1. The hazard ratio is a descriptive measure to compare survival times of two groups, which provides the estimate of the relative risk of events.
For example, if one would compare the time to event between two groups of patients: untreated and treated with

$$Hazard\,ratio = \frac{h_{untreated}(t)}{h_{treated}(t)} = 2, \tag{21}$$

it would mean, that a patient from the group with the higher hazard rate (untreated) will experience the event 'faster' than a patient from the other group. Whereby 'faster' means, that the patient has twice the chance of getting the event at the next time point compared to the patient from the treated group.

17.2.2.4 Cox proportional hazards

The Cox proportional hazards model is another technique to analyze survival data. It is a regression model which estimates the hazard ratio:

$$ln\frac{h(t)}{h_0(t)} = b_1x_1 + b_2x_2 + ... + b_px_p, \tag{22}$$

where h(t) is the hazard rate at time t, x_1, x_2, ..., x_p are the explanatory variables; b_1, b_2, ..., b_p are estimated from the data and $h_0(t)$ is the baseline hazard (when all explanatory variables x are zero) [334]. The Cox's model uses the assumption, that the hazard ratio does not depend on time [336].

17.2.2.5 The log rank test

Figure 185 shows survival curves of two groups of patients: patients with glioblastoma and patients with astrocytoma. In order to compare the survival between these two groups, one could compare the probability of survival a specific time point. This approach gives only a restricted comparison as it does not compare the total survival of the two groups.
A method for comparison of survival curves, which takes into account the total follow-up period, is the log rank test. The log rank test is a test of significance which tests a null hypothesis
H_0 = *'There is no difference between the groups in the probability of the occurrence of an event at any time point'* [335]. It calculates the number of observed and expected events in

each group for each time of event (i.e. the log rank test compares estimates of hazard rates from both groups).

The log rank test handles the censored data in the same way as the Kaplan Meier method and is based on the same assumptions. To calculate the log rank test, one should first plot the survival curves and make sure, that the curves of the two groups of interest do not intersect, as the log rank test uses the same assumption as the Cox's model, that the hazard ratio is constant over time. Hence, it is less likely to detect a difference between survival curves if the risk of an event is not consistently greater in one group.

17.2.3 Classifiers

A classifier assigns one of the "intrinsic" breast cancer subtypes to a tumor sample. There are two classes of classification models: single sample predictors (SSPs) and subtype classification models (SCMs). SSPs use hierarchical clustering on large "intrinsic" gene sets and classify a tumor sample using nearest centroid methods into "intrinsic" breast cancer subtypes: luminal A, luminal B, HER2-enriched, basal-like or normal-like; SCMs fit a mixture of Gaussian distributions on genes, whose expression is correlated with ER, HER2 and AURKA, the three key breast cancer genes. These Gaussian distributions represent three molecular subtypes of breast cancer: basal-like, HER2-enriched and luminal tumors. A tumor sample is then classified based on its maximum posterior probability to belong to one of the three molecular subtypes. Figure 186 shows conceptual designs of single sample predictors and subtype classification models [331].

17.2.3.1 PAM50

PAM50 [338] is a 50-gene subtype predictor and it belongs to the single sample predictors. PAM50 measures the expression of 50 genes, which were selected to characterize the "intrinsic" subtypes: luminal A, luminal B, basal-like, HER2-enriched and normal-like (i.e. with no matching clinico-pathological type).

To develop this classifier, an "intrinsic" gene set (1.906 genes) found in four microarray studies was analyzed by hierarchical clustering to identify the "intrinsic" subtypes. A minimized gene set containing 50 genes distinguishing between different subtypes was derived using the top "N" t test statistics. PAM50 uses a centroid-based prediction method (Prediction Analysis of Microarray (PAM)) to assign the subtype prediction to a tumor sample. PAM is an approach to predict the cancer class from gene expression profiles using the 'nearest shrunken centroids' method. This method identifies subsets of genes, which contribute most to the single classes, using cross validation to determine the amount of the 'shrinkage' [339]. After deriving the 50 best distinguishing genes, the PAM method is used here to get the best distinguishing genes for each cluster (out of 50 genes). Hence, the expression profile of a tumor sample have to be compared to shrunken centroids of each intrinsic subtype, each containing expression profiles of less than 50 genes.

A PAM50-based risk of recurrence score 'Prosigna' received a FDA (Food and Drug Administration) approval for clinical use [340].

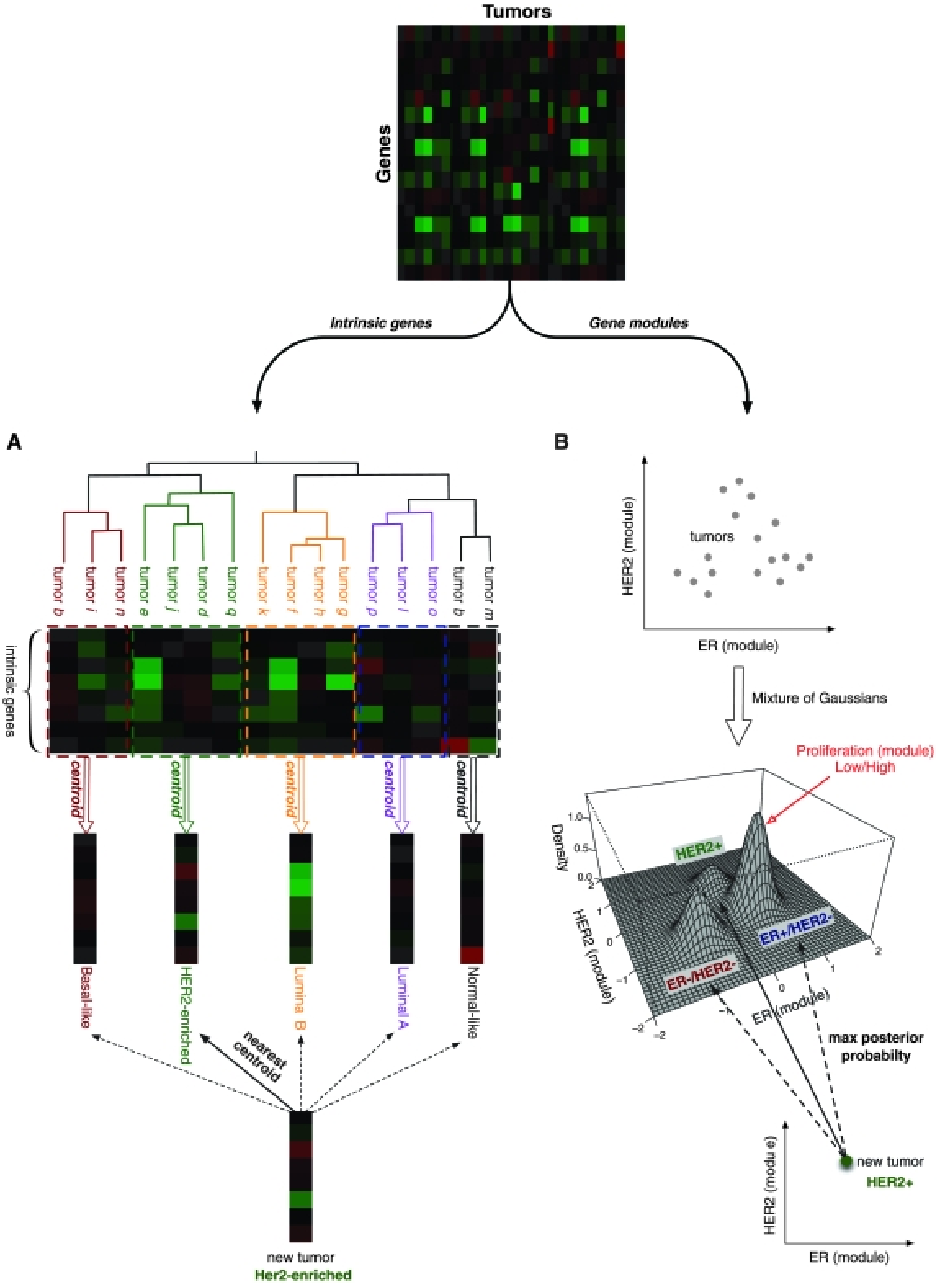
17.2.3.2 SCMOD1 & SCMOD2

SCMOD1 [341] and SCMOD2 [342] are subtype classification models, which are based on the concept of co-expression modules. The only difference between the two models is that, SCMOD1 is based on 726 genes and SCMOD2 is based on 663 genes.

Genes which are associated with the three key biological processes in breast cancer (estrogen receptor, HER2-signaling and proliferation) are considered as 'prototypes': ESR1, ERBB2 and AURKA. The aim of SCMOD1 is to identify gene sets (co-expression modules) specifically co-expressed with such 'prototypes'. The t statistic is used to test the association of a gene and a 'prototype'. After identifying the modules, SCMOD1 and SCMOD2 compute module scores and apply Gaussian mixture models to these scores in order to define three two-dimensional clusters (defined by ESR1 and ERBB2 module scores) corresponding to three molecular subgroups (ER-/HER2- (basal-like), HER2+ (HER2 enriched) and ER+/HER2- (luminal-like)). In order to discriminate between high (luminal B) and low (luminal A) proliferation luminal-like subtypes, SCMs incorporate proliferation module scores (referred by AURKA). A tumor sample is then automatically classified by computing the maximum posterior probability to belong to one of the computed clusters.

17.2.3.3 SCMGENE

SCMGENE [331] is a three-gene subtype classification model. It is the simplest form of a subtype classification model, which uses the expression of only three genes: ER, HER2 and AURKA, instead of co-expression modules like SCMOD1 and SCMOD2. Like other subtype classification models, a two dimensional clustering is performed using ESR1 and ERBB2 module scores to define the dimensions, where ESR1 is the gene encoding ER and ERBB2 - the gene encoding HER2. SCMGENE also uses the Gaussian mixture models to define three molecular subgroups as well as AURKA module score in order to differ between LumA and LumB, like described in the previous section for SCMOD1 and SCMOD2.



C

Model	Type	Number of genes	Reference
SSP2003	SSP	500	(6)
SSP2006	SSP	306	(2)
PAM50	SSP	50	(3)

D

Model	Type	Number of genes	Reference
SCMOD1	SCM	726	(1)
SCMOD2	SCM	663	(8)
SCMGENE	SCM	3	Present work

Figure 186 Conceptual design of the two breast cancer molecular subtyping methods: A) the Single Sample Predictor (SSP) and B) the Subtype Classification Model (SCM). Figure taken from [331].

17.2.4 Risc score predictors

17.2.4.1 EndoPredict

Endopredict (EP) [343] is a RNA-based multigene score predicting the likelihood of distant recurrence for patiens with early-stage ER+ and HER2- breast cancer, who are treated with adjuvant hormonal therapy. The EP risk score is a linear combination consisting of 8 cancer-associated genes: BIRC5, RBBP8, UBE2C, IL6ST, AZGP1, DHCR7, MGP and STC2; and 3 normalization/reference genes: CALM2, OAZ1, RPL37A. The EP score is a linear combination of the qPCR measurements (C_t) of these genes:

$$\begin{aligned} s_u =& 0.41 \cdot \Delta C_t(BIRC5) - 0.35 \cdot \Delta C_t(RBBP8) \\ &+ 0.39 \cdot \Delta C_t(UBE2C) - 0.31 \cdot C_t(IL6ST) \\ &- 0.26 \cdot \Delta C_t(AZGP1) + 0.39 \cdot C_t(DHCR7) \\ &- 0.18 \cdot \Delta C_t(MPG) - 0.15 \cdot \Delta C_t(STC2) - 2.63 \end{aligned} \tag{23}$$

The EP score ranges from 0 to 15, where a higher EP score indicates a higher risk of recurrence and a lower EP score indicates a lower risk of recurrence. The EP score can be combined with clinical parameters (tumor size and nodal status) to EPclin:

$$s_{clin} = 0.35 \cdot t + 0.64 \cdot n + 0.28 \cdot s, \tag{24}$$

where t is the tumor size and n - the lymph node status. In the training set, the cut-offs for both scores were defined, to discriminate between high and low risk of distant recurrence: a cut-off of 5 for EP and 3.3 for EPclin. Filipits et al. also validated both risk scores, concluding that the combined EPclin score outperformed the EP score.

17.2.4.2 OncotypeDx

OncotypeDx [344] is a multigene assay to predict the risk of recurrence of tamoxifen-treated, node-negative breast cancer. To develop OncotypeDx, 250 candidate genes were selected based on literature, other studies and genomic databases. Then, the association of the expression of these candidate genes and breast cancer recurrence was studied in three independent clinical studies involving 447 patients with node-negative, estrogen-receptor positive, tamoxifen-treated breast cancer. 16 genes, whose expression profiles correlated with breast cancer recurrence were selected as well as 5 additional reference genes (for normalization). The recurrence score of OncotypeDx is calculated by an algorithm based on the expression profiles of these selected genes and lies between 0 and 100. Figure 187 shows the OncotypeDx algorithm and the selected genes. OncotypeDx classifies patients into one of the following risk groups: low (score < 18), intermediate ($18 <=$ score < 31) or high (score > 31) risk of recurrence.

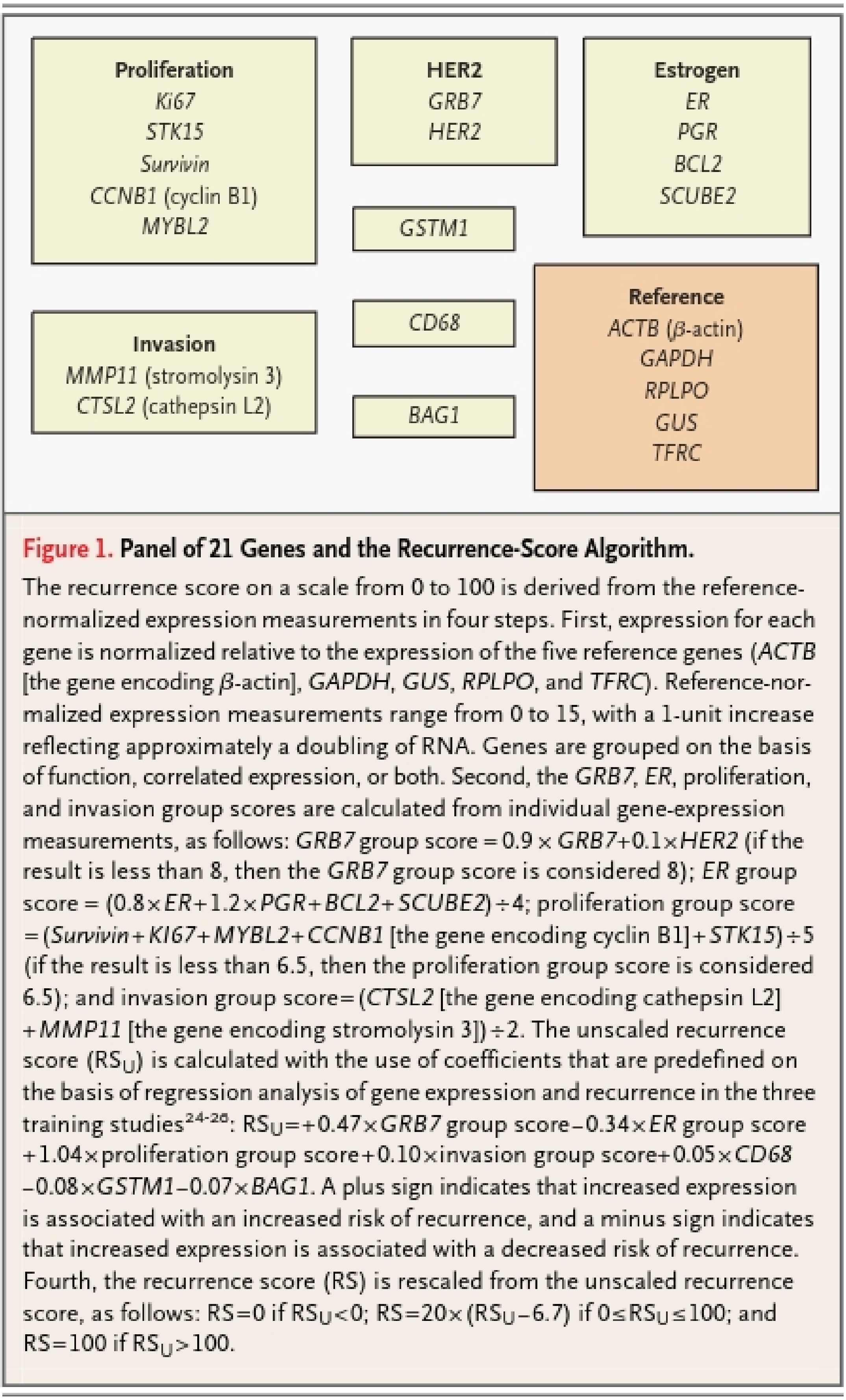
17.2.4.3 GGI

The gene expression grade index (GGI) [345] is a score, which summarizes the similarity between the expression profile and the tumor grade:

$$GGI = scale\left(\sum_{j\in G_3} x_j - \sum_{j\in G_1} x_j - offset\right), \tag{25}$$

with scale and offset - transformation parameters to standardize GGI, G$_i$ - set of genes with increased expression in histological grade i tumors and x - the logarithmic gene expression measure. A negative GGI score corresponds to tumor grade 1 (low grade), a positive GGI score corresponds to grade 3 (high grade). This score does not assign a tumor grade 2, which corresponds to intermediate risk of recurrence, as this information is not helpful for clinical decision making. The aim of Sotiriou et al. was to assign the histological grade 2 status to low (= 1) or high (= 3) risk categories. After analyzing expression profiles of tumor grade 2 tumors, the authors concluded, that there were no independent gene expression profiles distinguishing between tumor grade 2 and grades 1 and 3.

To develop GGI, datasets from breast carcinomas were analyzed to identify differentially expressed genes (by comparing the expression between tumors of histological grade 1 and 3). 97 genes, which were associated with tumor grade could be identified. Most of these were involved in proliferation and cell cycle regulation. GGI was then defined using the expression of these genes.



Proliferation
Ki67
STK15
Survivin
CCNB1 (cyclin B1)
MYBL2

HER2
GRB7
HER2

Estrogen
ER
PGR
BCL2
SCUBE2

GSTM1

CD68

BAG1

Invasion
MMP11 (stromolysin 3)
CTSL2 (cathepsin L2)

Reference
ACTB (β-actin)
GAPDH
RPLPO
GUS
TFRC

Figure 1. Panel of 21 Genes and the Recurrence-Score Algorithm.

The recurrence score on a scale from 0 to 100 is derived from the reference-normalized expression measurements in four steps. First, expression for each gene is normalized relative to the expression of the five reference genes (*ACTB* [the gene encoding β-actin], *GAPDH*, *GUS*, *RPLPO*, and *TFRC*). Reference-normalized expression measurements range from 0 to 15, with a 1-unit increase reflecting approximately a doubling of RNA. Genes are grouped on the basis of function, correlated expression, or both. Second, the *GRB7*, *ER*, proliferation, and invasion group scores are calculated from individual gene-expression measurements, as follows: *GRB7* group score = 0.9 × *GRB7*+0.1×*HER2* (if the result is less than 8, then the *GRB7* group score is considered 8); *ER* group score = (0.8×*ER*+1.2×*PGR*+*BCL2*+*SCUBE2*)÷4; proliferation group score =(*Survivin*+*KI67*+*MYBL2*+*CCNB1* [the gene encoding cyclin B1]+*STK15*)÷5 (if the result is less than 6.5, then the proliferation group score is considered 6.5); and invasion group score=(*CTSL2* [the gene encoding cathepsin L2] +*MMP11* [the gene encoding stromolysin 3])÷2. The unscaled recurrence score (RS_U) is calculated with the use of coefficients that are predefined on the basis of regression analysis of gene expression and recurrence in the three training studies[24-26]: RS_U=+0.47×*GRB7* group score−0.34×*ER* group score +1.04×proliferation group score+0.10×invasion group score+0.05×*CD68* −0.08×*GSTM1*−0.07×*BAG1*. A plus sign indicates that increased expression is associated with an increased risk of recurrence, and a minus sign indicates that increased expression is associated with a decreased risk of recurrence. Fourth, the recurrence score (RS) is rescaled from the unscaled recurrence score, as follows: $RS=0$ if $RS_U<0$; $RS=20\times(RS_U-6.7)$ if $0\leq RS_U\leq 100$; and $RS=100$ if $RS_U>100$.

Figure 187 OncotypeDx: 21 selected genes and the recurrence-score algorithm. Figure taken from [344].

17.2.5 Comparison

In order to assess the concordance between different classifiers and risk predictors, Wirapati et al. applied them on 36 publicly available breast cancer datasets containing 5.715 breast tumors. The subtype classification model SCMGENE was used as reference. Figure 188 shows the result of this analysis. Colored bars represent the breast cancer subtypes Basal-like (dark red), HER2-enriched (green), Luminal B (orange), Luminal A (purple) and Normal-like (black).
The basal-like subtype seems to be most consistently assigned by all classifiers, while luminal B and normal-like subtypes are assigned differently. Luminal A and B subtypes seem to be often mixed up by the classifiers (according to SCMGENE reference prediction). Overall, the coloured bars are roughly noticeable across all classifiers (Figure 188, A).

The results of risk predictions were also compared to the subtype classifications (Figure 188, B). Three risk predictors (including GGI and OncotypeDx) classified almost all Luminal A tumors a low risk of recurrence and most of the basal-like, HER2-enriched and Luminal B tumors - a high risk of recurrence.

Lastly, the concordance between “intrinsic” subtypes and several clinical parameters was assessed (Figure 188, C). The estrogen receptor (ER) status, which is defined by immunohistochemistry (IHC) matches with the classification of basal-like and luminal tumors (basal-like: mostly ER-, luminal: ER+). Similarly, the HER2-status (also defined by IHC or fluorescent in situ hybridization (FISH)) matches with HER2-enriched subtype. According to this analysis, other clinical parameters like progesterone (PGR) receptor status, the tumor size or age at diagnosis seem not to be associated with the “intrinsic” subtypes [331].

Buus et al. compared EndoPredict with OncotypeDX using TransATAC samples from the ATAC trial, which evaluated the “efficacy and safety of anastrozole vs. tamoxifen given for five years in postmenopausal women with localized primary breast cancer” [346]. The endpoint of interest was the relapse-free survival (RFS). This study was the first direct comparison of the clinical performance of two methods. They found that EndoPredict and OncotypeDx are similar in the years 0 to 5, but EndoPredict outperformed OncotypeDx in years 5 to 10. The authors pointed out, that the difference in performance might result from different training populations: EndoPredict algorithm was trained on HER2-negative and mixed node-negative and node-positive population, while OncotypeDx was trained on a mixed HER2-positive and negative population with node-negative status. The ATAC trial contained only HER2-negative samples.

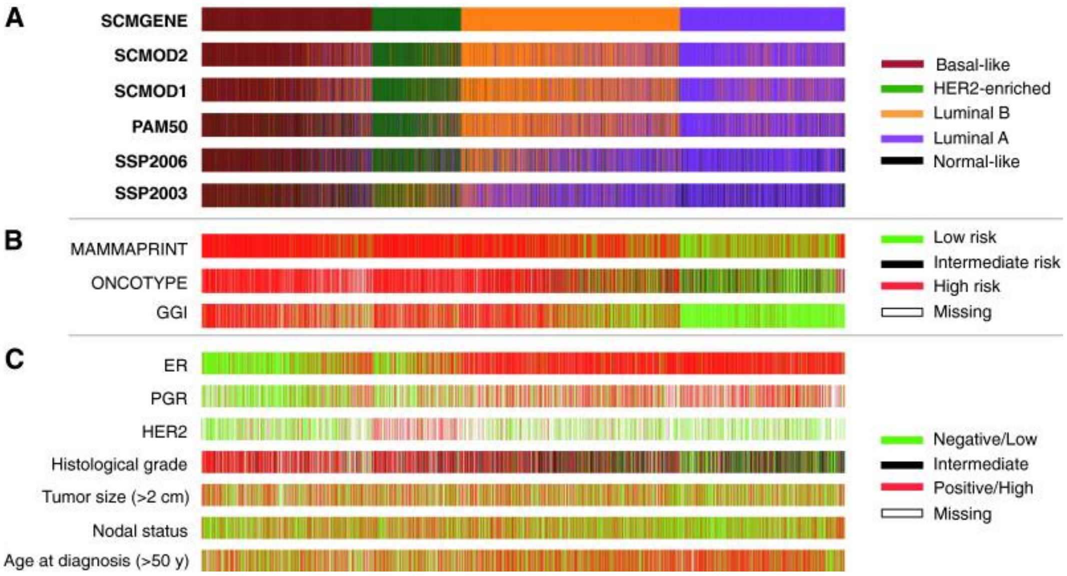


Figure 188 Concordance of classifiers for breast cancer molecular subtyping. Colored bars illustrate the molecular subtypes as computed by each of the six classifiers applied to the compendium of 5715 breast tumors. SCMGENE, the three-gene subtype classification model, was used as the reference [331].

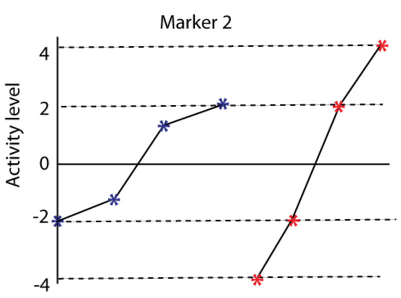
17.2.6 Network-based outcome or subtype prediction methods

17.2.6.1 A greedy procedure based on PPI networks (Chuang et al.)

Chuang et al. have implemented an protein-network-based approach to classify expression profiles as 'metastatic' or 'non-metastatic' [347]. They identify markers as sub-networks in the protein-protein interaction (PPI) network rather than as individual genes (see Figure 191). The PPI network they used contained 57 235 interactions among 11 203 proteins and was created by integrating data from yeast-two-hybrid experiments, predictions via co-citation and orthology, and manual curation of literature. They applied their approach to two data sets of breast cancer patients, which have been reported by van de Vijver et al. [348] and Wang et al. [349], respectively.
To find significant sub-networks in the whole PPI network, they first normalize gene expression values to z-transformed scores z_{ij} (for gene i and sample j) with mean $\mu = 0$ and standard deviation $\sigma = 1$ and overlay the normalized gene expression value of each gene on its corresponding protein in the PPI. The activity a_j of a given sub-network M in sample j is then assessed by averaging the normalized gene expression values z_{ij} of all contained genes i. This step gives one activity score per patient in the dataset for each sub-network. a denotes the vector of activity scores over all patient samples for a given sub-network M and a' denotes a discretized form of a, derived by discretizing the activity levels into $\lfloor \log_2(\#of samples) + 1 \rfloor$ bins.
In a second step, they calculated the mutual information between a sub-network's activity scores and the vector of class labels (metastatic or non-metastatic , denoted by c) over all patients to assess the discriminative potential of the sub-network. The discriminative potential is given by the discriminative score function, which is calculated via the mutual information between a' and c as

$$S(M) = MI(a', c) = \sum_{x \in a'} \sum_{y \in c} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \tag{26}$$

Here, "x and y enumerate values of a and c, respectively, $p(x, y)$ is the joint probability density function (pdf) of a' and c, and $p(x)$ and $p(y)$ are the marginal pdf's of a' and c" [347] (see Figure 189).



Original activity a = [-2 -1 1 2 -4 -2 2 4]

Discretized activity x = [-1.15 -1.15 1.15 1.15 -3.45 -1.15 1.15 3.45]

Class labels c = [1 1 1 1 2 2 2 2]

Joint probability

c \ x	-3.45	-1.15	1.15	3.45	
1	0	1/4	1/4	0	1/2
2	1/8	1/8	1/8	1/8	1/2
	1/8	3/8	3/8	1/8	

$$MI(\mathrm{a}',\mathrm{c}) = \sum_{x\in \mathrm{a}'} \sum_{y \in \mathrm{c}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = (\frac{1}{4}\times\log(\frac{1}{4}/(\frac{3}{8}\times\frac{1}{2})))\times 2 + (\frac{1}{8}\times\log(\frac{1}{8}/(\frac{1}{8}\times\frac{1}{2})))\times 2 + (\frac{1}{8}\times\log(\frac{1}{8}/(\frac{3}{8}\times\frac{1}{2})))\times 2 = 0.2157$$

Figure 189 Illustration of the calcutation of the mutual information between sub-network activity and class labels, taken from the supplement of Chuang et al. [347].

Sub-networks are derived by starting at each protein in the PPI network and seeding a candidate sub-network with this protein. This sub-network is then iteratively expanded by a greedy procedure. In each iteration step, the protein within distance $d = 2$ from the seed that maximizes the discriminative score function $S(M)$ is added to the network until no addition increases the score over the improvement rate $r = 0.05$.
Significant sub-networks are identified by comparing the network's discriminative potential to those of random networks. Three tests of significance were used: For the first test, the expression vectors of individual genes were permuted 100 times to break the correlation between expression and interaction. The scores of the obtained random sub-networks give a 'global' null distribution, on which real sub-network scores are indexed. For the second test, 100 random sub-networks are generated from the same seed gene and the real sub-network scores are indexed on this 'local' null distribution which is assumed to be gamma distributed. For the last test, class labels were randomly assigned to patients 20000 times, resulting in a null distribution of mutual information for each trial. Again, the real score is indexed on this null distribution. Significant sub-networks are those that satisfy all three tests with $P_1 < 0.05, P_2 < 0.05$ and $P_3 < 0.00005$.
Based on the identified significant sub-networks, a classifier was then trained using the activity scores of the sub-networks, which gives the sub-network activity matrix of significant sub-networks versus patient samples, as feature values for a logistic regression. The classification performance of the logistic regression model on the sub-network activity matrix in comparison to other markers can be seen in Figure 190.

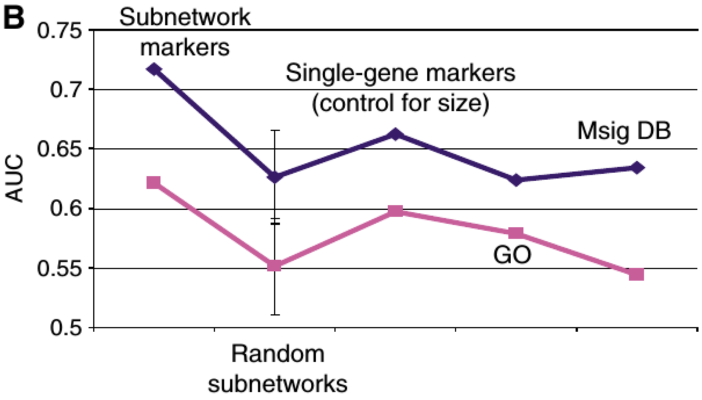
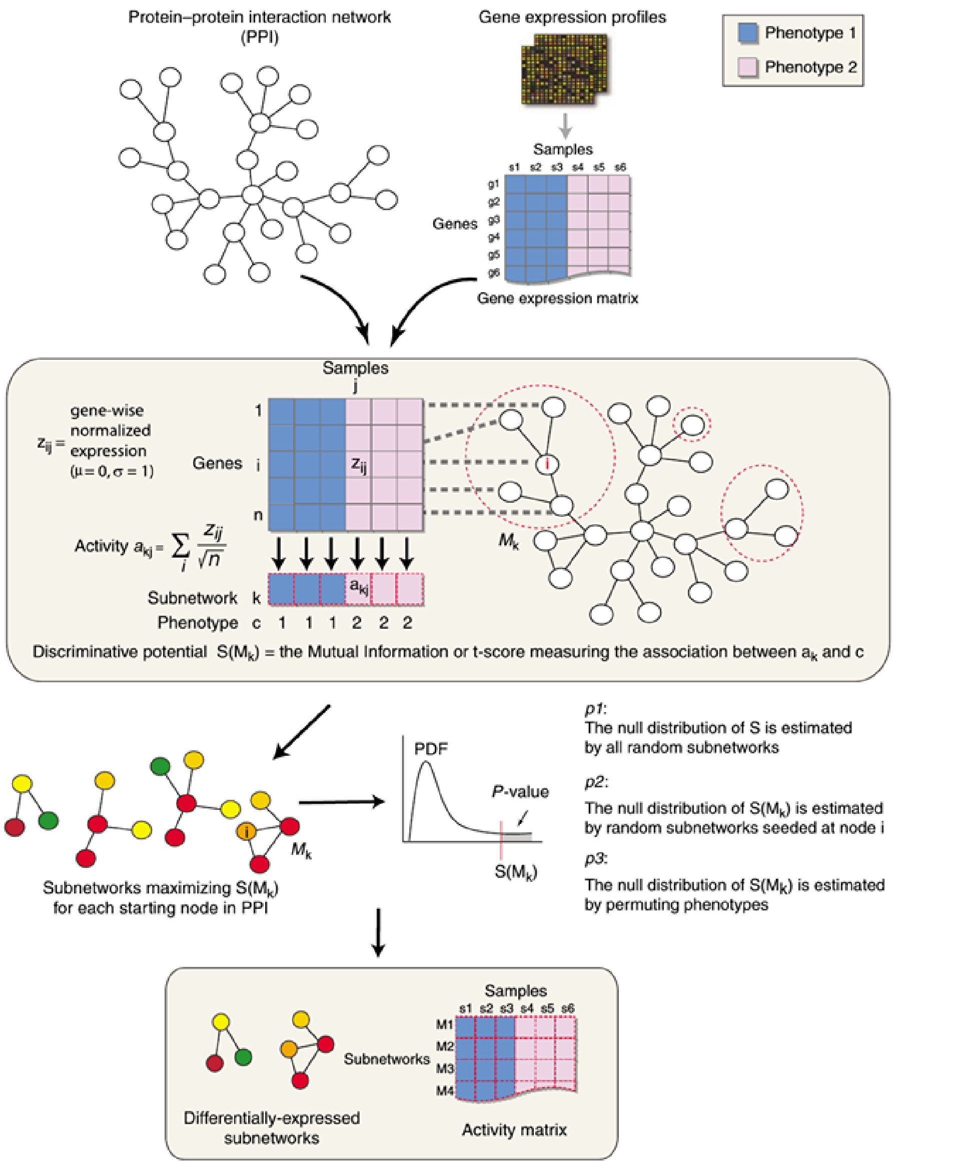


Figure 190 AUC classification performance of sub-network markers in comparison to other markers. Blue line: markers selected based on the dataset of Wang et al. [349] and tested on the van de Vijver et al. [348] dataset; pink line: training on the van de Vijver et al. dataset and testing on the dataset of Wang et al. Taken from the Chuang et al. [347].

Box 1 Schematic overview of subnetwork identification


Protein–protein interaction network
(PPI)
Gene expression profiles
Phenotype 1
Phenotype 2
Samples
s1 s2 s3 s4 s5 s6
Genes
g1 g2 g3 g4 g5 g6
Gene expression matrix
Samples
j
z_{ij} = gene-wise normalized expression ($\mu = 0, \sigma = 1$)
Genes
1
i
n
z_{ij}
Activity $a_{kj} = \sum_i \frac{z_{ij}}{\sqrt{n}}$
Subnetwork k
a_{kj}
Phenotype c 1 1 1 2 2 2
M_k
i
Discriminative potential $S(M_k)$ = the Mutual Information or t-score measuring the association between a_k and c
Subnetworks maximizing $S(M_k)$
for each starting node in PPI
i
M_k
PDF
P-value
$S(M_k)$
p1:
The null distribution of S is estimated by all random subnetworks
p2:
The null distribution of $S(M_k)$ is estimated by random subnetworks seeded at node i
p3:
The null distribution of $S(M_k)$ is estimated by permuting phenotypes
Samples
s1 s2 s3 s4 s5 s6
M1 M2 M3 M4
Subnetworks
Differentially-expressed subnetworks
Activity matrix

Protein–protein interaction networks are used to assign sets of genes to discrete subnetworks. Gene expression profiles of tissue samples drawn from each type of cancer (i.e., metastatic or non-metastatic) are transformed into a 'subnetwork activity matrix'. For a given subnetwork M_k in the interaction network, the activity is a combined *z*-score derived from the expression of its individual genes. After overlaying the expression vector of each gene on its corresponding protein in the interaction network, subnetworks with discriminative activities are found via a greedy search. Significant subnetworks are selected based on null distributions estimated from permuted subnetworks (see Materials and methods). Subnetworks are then used to identify disease genes, and the subnetwork activity matrix is also used to train a classifier.

Figure 191 Figure from Chuang et al. [347] describing their method.

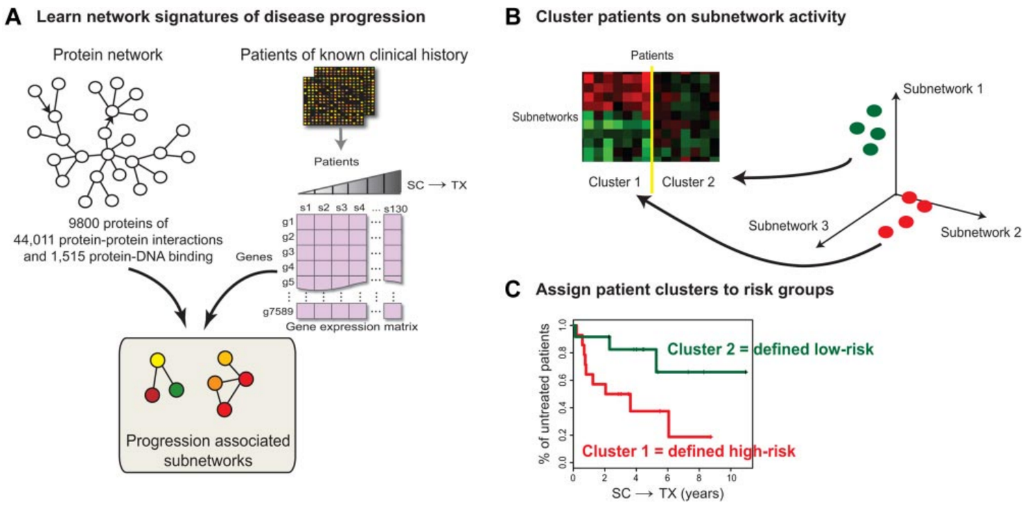

A Learn network signatures of disease progression
Protein network
Patients of known clinical history
9800 proteins of
44,011 protein-protein interactions
and 1,515 protein-DNA binding
Patients
SC → TX
s1 s2 s3 s4 ... s130
Genes
g1
g2
g3
g4
g5
g7589
Gene expression matrix
Progression associated subnetworks
B Cluster patients on subnetwork activity
Patients
Subnetworks
Cluster 1
Cluster 2
Subnetwork 1
Subnetwork 2
Subnetwork 3
C Assign patient clusters to risk groups
% of untreated patients
0.0
0.2
0.4
0.6
0.8
1.0
Cluster 2 = defined low-risk
Cluster 1 = defined high-risk
0
2
4
6
8
10
SC → TX (years)

Figure 1. Schematic overview of subnetwork identification and definition of risk groups. (A) The expression profile of each gene is projected onto its corresponding protein in a protein-protein interaction subnetwork. A greedy search is performed to find subnetworks for which the activities are associated with the time from sample collection to treatment (SC→TX). Significant subnetworks are selected based on null distributions estimated from permuted data. Subnetworks are used to identify disease genes, and the subnetwork activity is used to characterize the signatures of different risk groups. (B) K-means clustering segregates patients by their distinct subnetwork activity patterns. (C) Patient clusters are assigned high versus low risk based on median treatment-free probabilities in a Kaplan-Meier analysis.

Figure 192 Figure from Chuang et al. [350] describing their method.

17.2.6.2 The greedy method of Chuang et al. with survival data

Chuang et al. [350] also presented a way to extend their method to work with survival times (or time until disease progression or any other time variable) instead of binary class assignments for patients (see Figure 192).

This method is very similar to the one presented above [347] with the main difference being the function used to score the different sub-networks. Here they fit a Cox proportional hazard model $T : \frac{H(t)}{H_0(t)} = e^{ka}$, where $H(t)$ is the hazard function at time t and $H_0(t)$ is the baseline hazard when all a_j (sub-network activity in patient j) are set to zero. The score $S(M)$ is calculated as $\log$ p-value, where the p-value is obtained from a χ^2-test of the fitted hazard model against the baseline hazard as the null model. Therefore, this score estimates the statistical significance of the sub-network activities over all patients as the sole predictor of the patients survival time (or treatment need, etc.).

Furthermore, they adapted the greedy procedure that iterative expands the seed module. Now, after every addition of a new gene to the sub-network, the deletion of every gene that is not essential for sub-network connectivity is considered and those deletions that improve the score $S(M)$ over the given improvement rate r are adopted.

After significant sub-networks have been identified, they build a sub-network-activity versus patients matrix and use 2-means clustering to separate the patients into a low risk and a high risk group. Cluster labels are inferred from the visual analysis of the Kaplan-Meier-Plot of the two clusters. Then, a nearest shrunken centroid classifier is trained on the two risk groups to predict the risk of new patient samples (see Figure 192).

17.2.6.3 FERAL - a method based on Sparse Group Lasso

FERAL (DelFt nEtwoRk-bAsed cLassifier) is a recent network-based classification method that tries to assign patients to a good or poor outcome class that are defined with respect to the recurrence free survival time at a 5-year threshold [351]. For the prediction of new samples, FERAL uses the Sparse Group Lasso (least absolute shrinkage and selection operator, SGL), which performs simultaneous selection of marker genes and training of the prediction model. Another feature of FERAL is that is uses multiple unsupervised or supervised operators to summarize meta-genes, not only the average-operator that is typically applied in most previous methods, which enables the classifier to choose the best meta-gene for every gene set. In the notation of the authors, a meta-gene describes the aggregation of several functionally related genes (sub-networks in the method of Chuang et al.). The unsupervised operators are the average operator, the median operator, the variance operator and minimum and maximum operator. The supervised operators are the linear integration (implicitly provided by the SGL) and the direction-aware-average (DA2).

In FERAL, gene set size is kept constant. Therefore, a gene set consists of a seed gene and its $k-1$ closest neighbours. If the seed gene has more than $k-1$ neighbours, genes are randomly removed from the gene set until it contains only k genes. If there are less than $k-1$ neighbours, then the neighbours of the neighbours are taken into account. The authors set $k = 10$ to provide a balance between performance and the relevance of the marker genes to cancer. All genes were considered as seed genes to make sure that every gene is included in at least one gene set.

After for each gene a gene set of its $k-1$ closest neighbours has been selected, expression values are z-score normalized and meta-genes are computed in a next step. Then, the SGL classifier is trained on the meta-genes.

The authors used a previously collected dataset of 1606 breast cancer samples (ACES cohort, [352, 353]) that aggregates samples from 12 different studies for the evaluation of their method. They evaluate FERAL on three different networks: a PPI network (the I2D network [354], also used by [352]), a co-expression network (defined on the training data at a correlation threshold of 0.6) and a random network obtained by shuffling the nodes in the PPI network. According to their evaluation, FERAL performs superior to previously published network-based methods on all three network types.

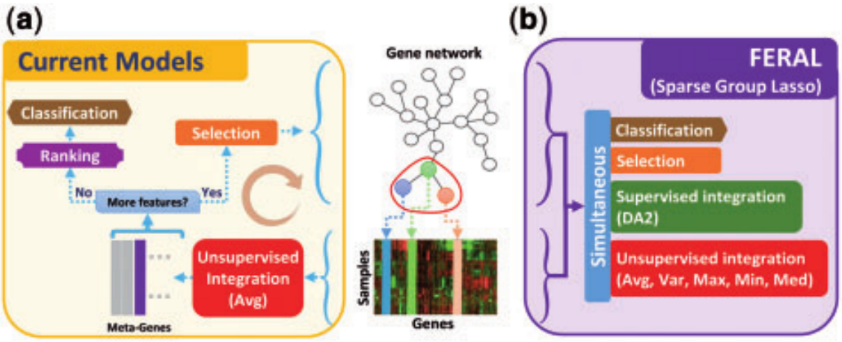


Figure 193 Comparison of FERAL to older methods, taken from [351].

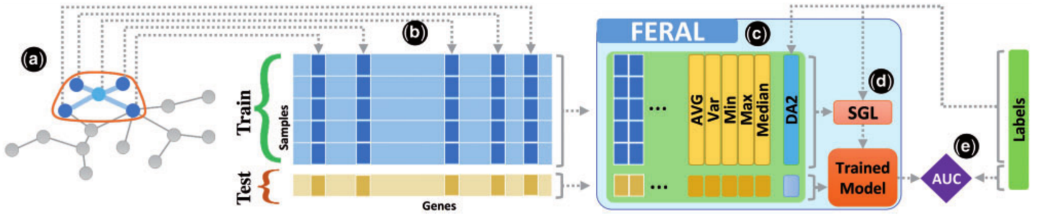


Fig. 3. Schematic of the training and testing procedures of FERAL. (**a**) In the first step, 10 genes are selected using given network. (**b**) Corresponding genes in expression dataset are selected and normalized using *z*-score. (**c**) Meta-genes are computed using the expression profiles of the gene set and target label (in case of a supervised integration). The expression of the individual genes is retained within the gene set. (**d**) The SGL is trained using training samples. (**e**) Test samples are used to assess the prediction performance (in terms of AUC) in the current fold

Figure 194 Schematic overview of the FERAL procedure, taken from [351].

17.2.6.4 NBS - Network-based stratification of tumor mutations

Hofree et al. [355] have developed a method called network-based stratification (NBS) that works on genome-wide somatic tumor mutations and integrates them with gene networks to infer subtypes predictive of clinical outcome [355].

The authors applied their method on three different cancer datasets of somatic mutations from the TCGA project (for ovarian, lung and uterine cancer). They also tested three different databases as sources of gene interaction network information: search tool for the retrieval of interacting genes (STRING) [356], HumanNet [357] or PathwayCommons [358].

In a first step, they represent each patient's mutation profile as a binary vector, where a 1 at position k means that the patient's tumor sample has acquired a mutation in gene k relative to the germ line. Then, this mutation profile is projected on a gene interaction network and network propagation is used, which spreads the influence of each mutation to its neighbourhood in the network. Network propagation simulates a random walk on the network according to the function

$$F_{t+1} = \alpha F_t A + (1 - \alpha) F_0 \tag{27}$$

where F_0 is the patient-versus-gene mutation matrix, A is a degree-normalized adjacency matrix and α is a network-dependent tuning parameter that regulates the distance that the signal is propagated (here between 0.5 and 0.7). This function is run iteratively until the matrix norm of $F_{t+1} - F_t < 1 \times 10^{-6}$ and therefore F_{t+1} converges.

The result of this step is a matrix of patients versus genes where each row contains one patients smoothed mutation profile. This matrix is quantile normalized and then clustered into a number of fixed, predefined subtypes via network-regularized non-negative matrix factorization (netNMF). This means that the function

$$min_{W,H>0} \parallel F - WH \parallel^2 + trace(W^t K W) \tag{28}$$

is minimized. Here, "W and H form a decomposition of the patient x gene matrix F (resulting from network smoothing as described above) such that W is a collection of basis vectors, or 'metagenes', and H is the basis vector loadings. The $trace(W^t K W)$ function constrains the basis vectors in W to respect local network neighborhoods. The term K is and adjacency matrix of a nearest neighbors network derived from the graph Laplacian of an influence distance matrix that is derived from the original network. The degree to which local network topology versus global network topology constrains W is determined by the number of nearest neighbors" [355].

To provide more robust cluster assignments, the authors use consensus clustering to aggregate the results of 1000 different subsamples (each containing 80% of patients and genes) into a single clustering result (see Figure 195).

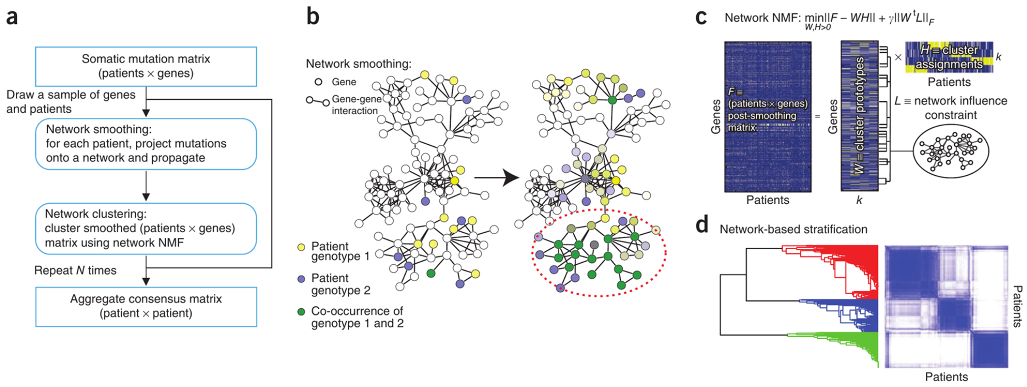


Figure 1 | Overview of network-based stratification (NBS). (**a**) Flowchart of the approach. (**b**) Example illustrating smoothing of patient somatic mutation profiles over a molecular interaction network. Mutated genes are shown in yellow (patient 1) and blue (patient 2) in the context of a gene interaction network. Following smoothing, the mutational activity of a gene is a continuous value reflected in the intensity of yellow or blue; genes with high scores in both patients appear in green (dashed oval). (**c**) Clustering mutation profiles using non-negative matrix factorization (NMF) regularized by a network. The input data matrix (*F*) is decomposed into the product of two matrices: one of subtype prototypes (*W*) and the other of assignments of each mutation profile to the prototypes (*H*). The decomposition attempts to minimize the objective function shown, which includes a network influence constraint *L* on the subtype prototypes. *k*, predefined number of subtypes. (**d**) The final tumor subtypes are obtained from the consensus (majority) assignments of each tumor after 1,000 applications of the procedures in **b** and **c** to samples of the original data set. A darker blue color in the matrix coincides with higher co-clustering for pairs of patients.

Figure 195 Schematic overview of the NBS procedure, taken from Hofree et al. [355].

The authors validated their method on a simulated dataset with known subtype assignments of patients and showed that by integrating network information and by using non-negative matrix factorization (NMF) instead of hierarchical clustering, their NBS-method performed much better on the simulated data than methods without network information or NMF, especially for larger network modules that are responsible for the assignment of patients to subtypes.

To train a classifier that can assign new patients to the computed subtypes based on his network-smoothed mutation profile, they used the nearest shrunken centroid approach, which calculates centroids for each subtype and then assigns a patient to the subtype with the closest centroid. Using tenfold cross-validation on their simulated dataset, this classifier showed a accuracy of over 95%. It it also possible to train a classifier on mRNA expression data with subtype assignments obtained from the NBS procedure using mutation profiles. Again, a nearest shrunken centroid classifier is used to learn expression signatures for each NBS subtype. Although the classifier trained on expression data showed a reduced accuracy of about 60%, it was still able to recover subtypes predictive of survival times [355].

17.2.7 Random signatures

Signatures are sets of genes, whose expression correlates with a specific outcome (for example, a disease). Venet et al. described how such signatures are typically derived: “1) characterize the mechanism in a model system, 2) derive from the model system a marker whose expression changes when the mechanism is altered, and 3) show that marker expression correlates with disease outcome in patients” [359]. The underlying assumption is, that these markers/signatures are more strongly associated with the outcome, than signatures which are not related to the mechanism of interest. Venet et al. proved this assumption wrong for breast cancer.
In order to do so, they compared 47 published breast cancer gene signatures with 1.000 random signatures of the same size focusing on the NKI dataset and the overall survival data (Figure 196). For each signature, the first principal component was calculated. According to the median of the first principal component, the cohort was splitted into two groups. After that, the log rank test was calculated for each splitted cohort and the resulting p-values were used to compare the real signatures with random signatures in respect to their association with breast cancer.

It could be shown, that 23% of published signatures have a weaker association with breast cancer than the median of random signatures. Furthermore, they defined a signature as 'biologically relevant' if it's association with breast cancer is stronger than that of the best 5% of random signatures. Only 40% of real signatures fulfilled this requirement.
Moreover, they found that larger signatures are more significant: “More than 90% of the signatures >100 genes we generated were significant at $p < 0.05$”. In Figure 196, for the two largest signatures (REUTER and HUA containing 714 and 1335 genes, respectively) all of the random generated signatures were significant at p-value < 0.05.
It was pointed out, that this study questioned the biological interpretation of the prognostic value of the published signatures, not their usefulness in the clinic, as a signature “may be accurate without yielding interesting biological insight regarding the mechanism of disease progression”.
Hence, Venet et al. concluded, that one should keep in mind, that “a random signature is more likely to be correlated with breast cancer outcome than not” and a study should not only investigate, whether their proposed signature is related to the outcome, but also if it is more related to the outcome than random signatures. In order to demonstrate this, they proposed to use much tighter p-values.

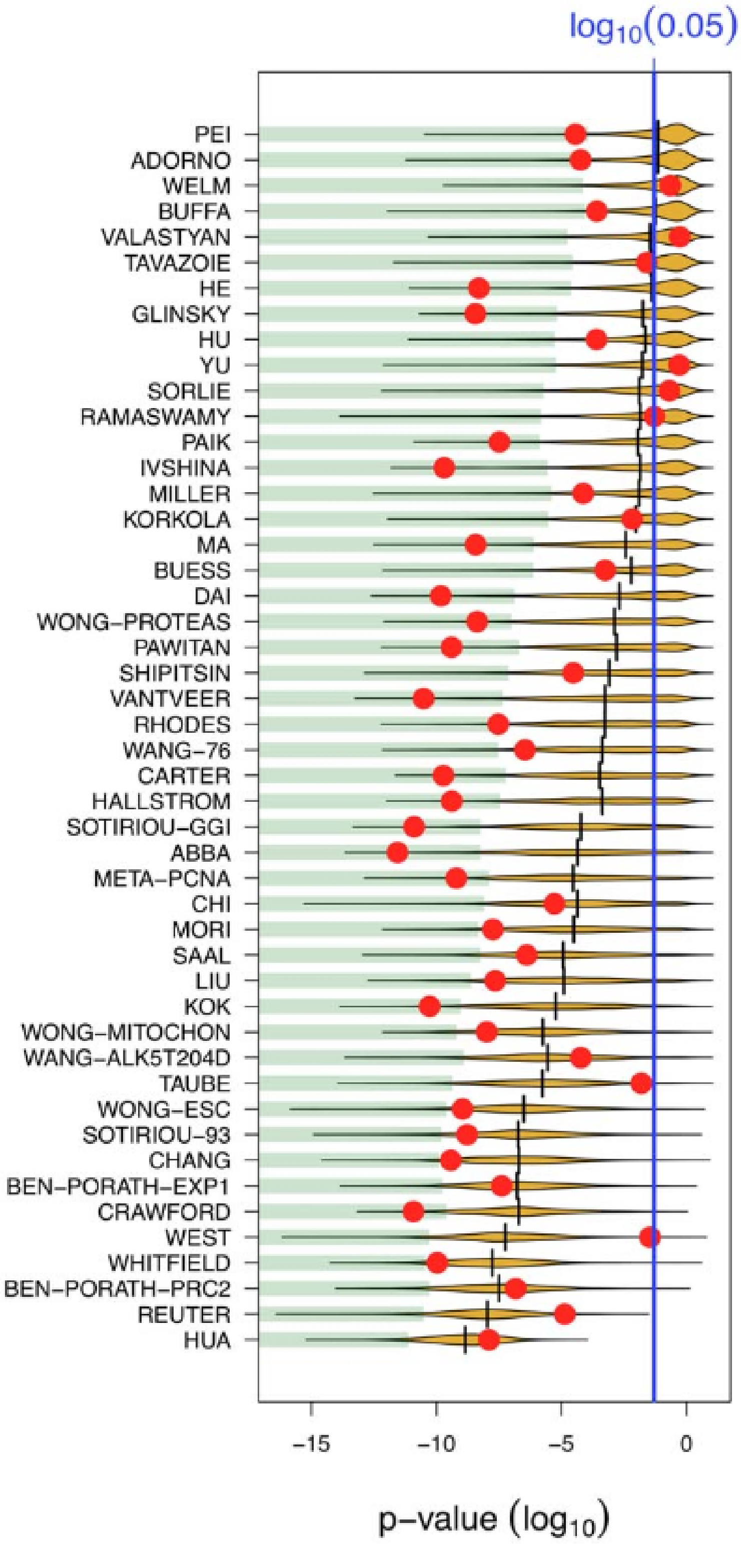


Figure 196 Comparison of random and published signatures. The x-axis denotes the p-value of association with overall survival. Red dots stand for published signatures, orange shapes depict the distribution of p-values for 1000 random signatures of identical size, with the lower 5% quantiles shaded in green and the median shown as black line. Signatures are ordered by increasing sizes. Figure taken from [359]

17.3 Project

17.3.1 Motivation

Many of the methods described above have been developed and tested on different datasets and only very few studies have compared the predictions of the different methods. But as the assignment of a patient's tumor to one of the subtypes is important for the selection of the appropriate therapy, it is vital to assess how much the different approaches agree or disagree with each other.
Therefore, the goal of our project was to create a interactive application that integrates and visualizes the various subtype and risk score prediction methods on the one hand, and many of the available datasets of breast cancer cohorts on the other hand. As we also integrated some network-based prediction methods, we wanted to investigate if the inclusion of network information during the training of a method improves the prediction performance. Additionally, we wanted to assess how “good” the signatures of the prediction methods are on each of the available datasets compared to random signatures of the same size.

17.3.2 Interactive application to visualize subtype predictions

We developed a shiny app for the interactive visualization of subtype and risk score predictions. Shiny is a web application framework for R [360]. The GUI of our app is build up of different 'views', which cover different components of survival analysis techniques and provides several interactive graphics and tables. The GUI contains the following views:

1. Explore a Predictor (Main view)
2. Compare 2 Predictors
3. Concordance Plot
4. Statistics
5. Random signatures
6. Random network signatures
7. Additional Informations

In the following sections, each 'view' will be described in more detail.

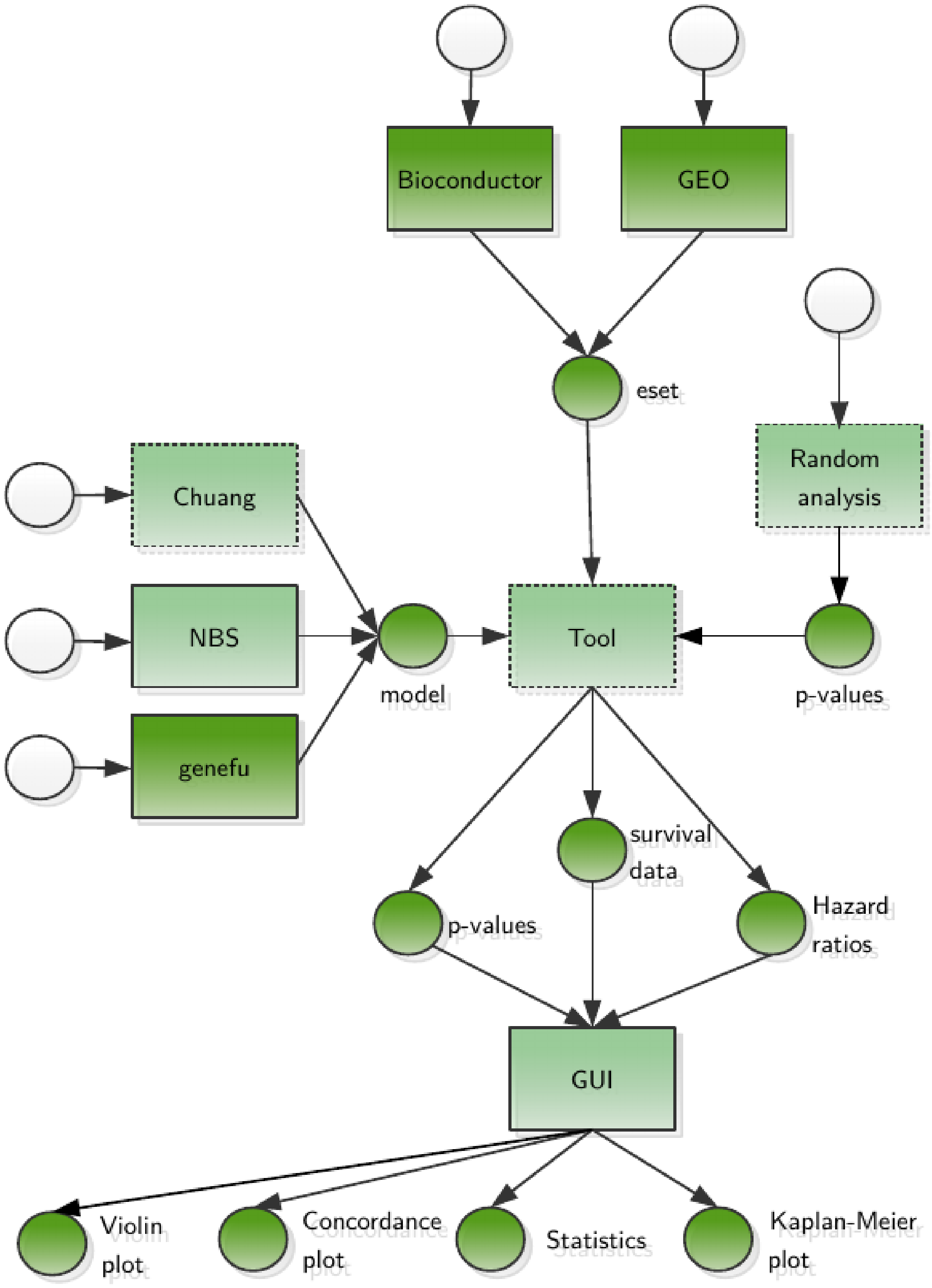


Figure 197 Our workflow as a petri net.

17.3.3 Main view: Explore a Predictor

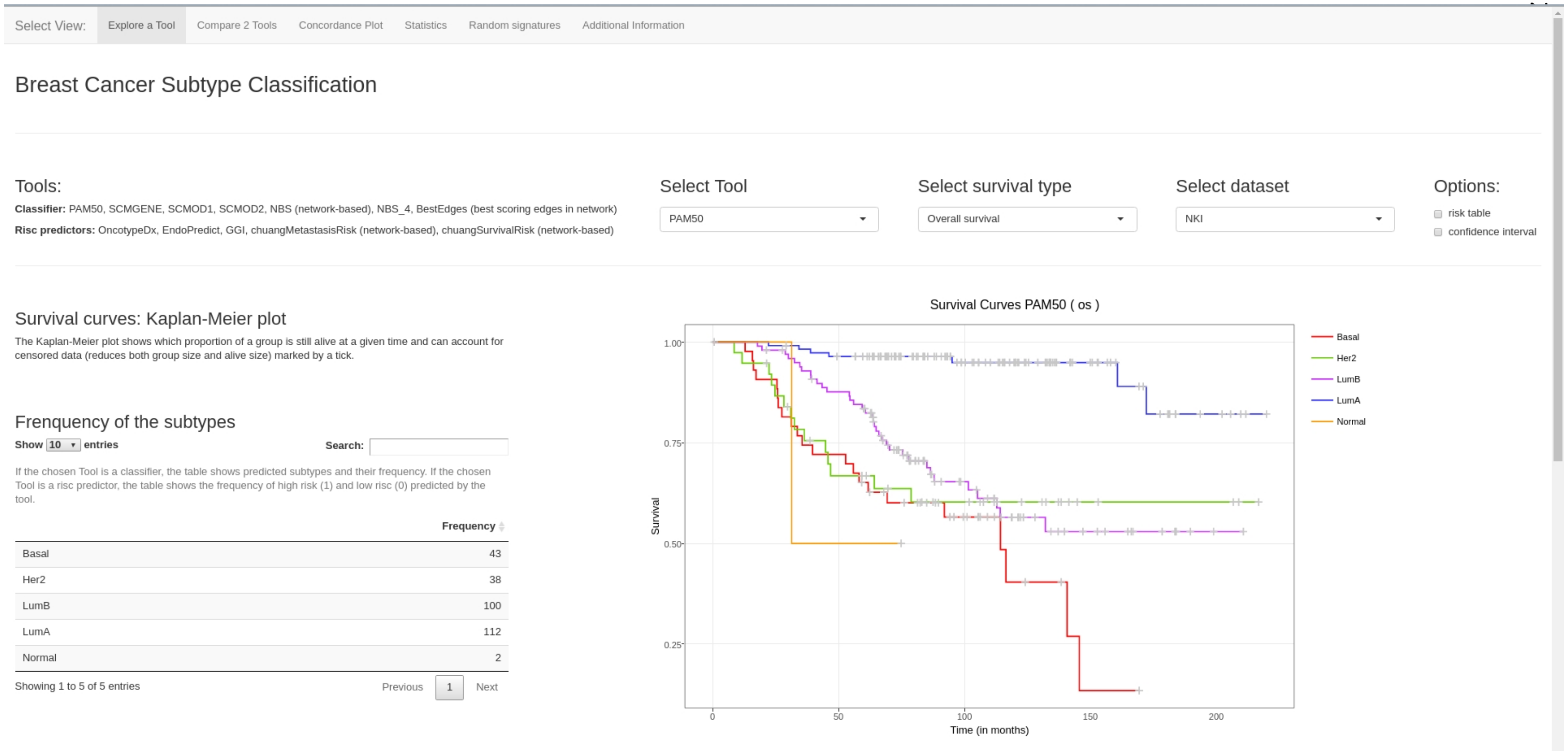
Figure 198 shows the upper 'Main view' of the application: “Explore a Predictor”. In the upper part, there are several tabs corresponding to different views.
In the upper left part of the 'Explore a Predictor' view, there is the description of available classifiers and risk predictors. On the right side, one can select a predictor, a dataset and survival type (only those, contained in the selected dataset). More informations on the available datasets and predictors are available in the view 'Additional Information' (see Section 17.3.9).
The following subtype classifiers are integrated into the application, all methods are described in previous sections:

- PAM50
- SCMGENE
- SCMOD1 and SCMOD2
- The NBS method, trained on mutation and gene expression data from the TCGA project to predict 3 and 4 subtypes, respectively.
- A classifier operating on edge scores (see section 17.3.8 'Random network signatures' for details)

Additionally, the following risk score predictors are integrated into the application:

- EndoPredict
- OncotypeDX
- GGI
- The method of Chuang et al. from 2007 [347] to predict risk of metastasis, trained on the network curated by Chuang et al. [347] and gene expression and phenotype information from van de Vijver et al. [348].
- The method of Chuang et al. from 2012 [350] to predict the risk of having a low overall survival rate, trained on the network curated by Chuang et al. [347] and gene expression and phenotype information from van de Vijver et al. [348].

Underneath the 3 selection boxes, the corresponding Kaplan-Meier plot is displayed. A short description of the Kaplan-Meier plot and a table containing breast cancer subtypes (predicted by the selected prediction method) and their frequencies are displayed to the left of the plot. There are two options, which can be selected to modify the Kaplan-Meier plot: 'risk table' and 'confidence interval'. If the 'risk table' option is selected, a 'number at risk' table appears beneath the Kaplan-Meier plot, indicating the number at risk for each breast cancer subtype at each time point (corresponding to x-axis ticks). If the option 'confidence interval' is selected, 95% confidence intervals are plotted as dashed lines for each breast cancer subtype. Figure 199 shows the Kaplan-Meier-Plot after both options were selected.
Figure 200 shows the lower section of the 'Main view' containing two tables. A table of hazard ratios for every pair of the predicted subtypes is shown on the left side. A table containing the log rank test p-values is shown on the right side. To verify, that the p-values are reliable, one should make sure, that the survival curves of the two subtypes of interest do not intersect on the Kaplan-Meier plot shown in the upper part of this view.



Select View: Explore a Tool | Compare 2 Tools | Concordance Plot | Statistics | Random signatures | Additional Information

Breast Cancer Subtype Classification

Tools:

Classifier: PAM50, SCMGENE, SCMOD1, SCMOD2, NBS (network-based), NBS_4, BestEdges (best scoring edges in network)

Risc predictors: OncotypeDx, EndoPredict, GGI, chuangMetastasisRisk (network-based), chuangSurvivalRisk (network-based)

Select Tool: PAM50

Select survival type: Overall survival

Select dataset: NKI

Options:

□ risk table

□ confidence interval

Survival curves: Kaplan-Meier plot

The Kaplan-Meier plot shows which proportion of a group is still alive at a given time and can account for censored data (reduces both group size and alive size) marked by a tick.

Frenquency of the subtypes

Show 10 **entries** **Search:**

If the chosen Tool is a classifier, the table shows predicted subtypes and their frequency. If the chosen Tool is a risc predictor, the table shows the frequency of high risk (1) and low risc (0) predicted by the tool.

	Frequency
Basal	43
Her2	38
LumB	100
LumA	112
Normal	2

Showing 1 to 5 of 5 entries

Figure 198 The main view of the Tool (upper section).

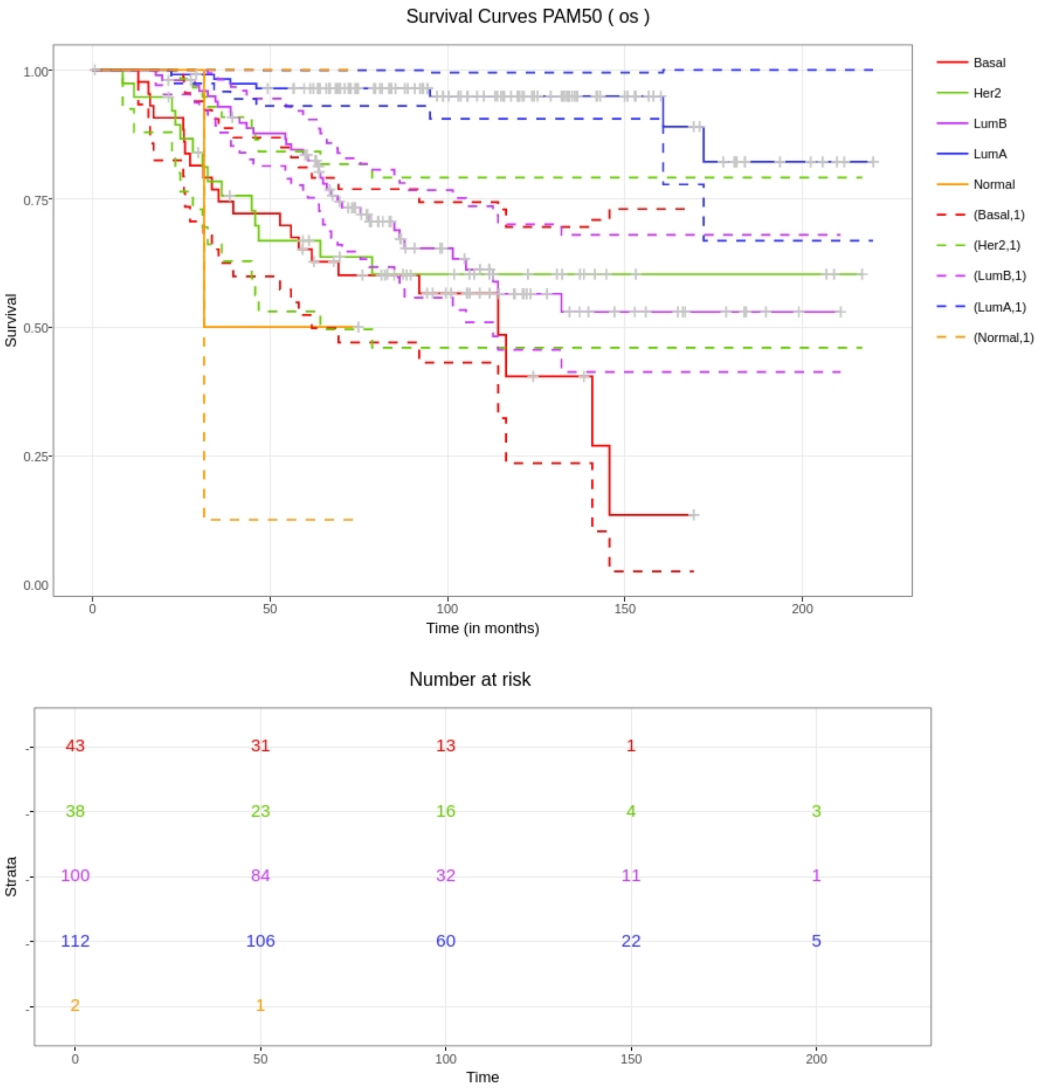
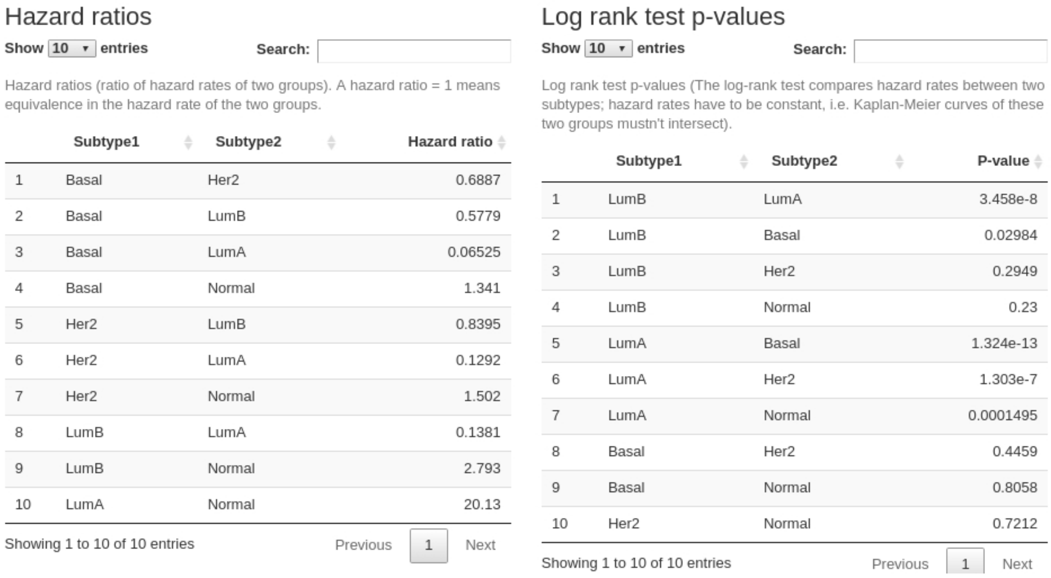


Figure 199 The Kaplan-Meier Plot from the Main view with selected options: confidence intervals and risk table.



Hazard ratios

Hazard ratios (ratio of hazard rates of two groups). A hazard ratio = 1 means equivalence in the hazard rate of the two groups.

	Subtype1	Subtype2	Hazard ratio
1	Basal	Her2	0.6887
2	Basal	LumB	0.5779
3	Basal	LumA	0.06525
4	Basal	Normal	1.341
5	Her2	LumB	0.8395
6	Her2	LumA	0.1292
7	Her2	Normal	1.502
8	LumB	LumA	0.1381
9	LumB	Normal	2.793
10	LumA	Normal	20.13

Showing 1 to 10 of 10 entries

Log rank test p-values

Log rank test p-values (The log-rank test compares hazard rates between two subtypes; hazard rates have to be constant, i.e. Kaplan-Meier curves of these two groups mustn't intersect).

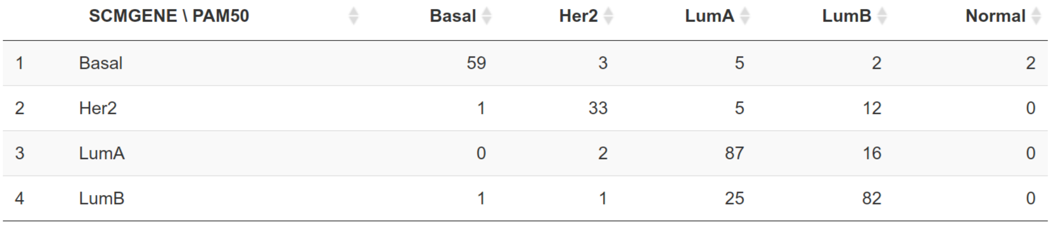
	Subtype1	Subtype2	P-value
1	LumB	LumA	3.458e-8
2	LumB	Basal	0.02984
3	LumB	Her2	0.2949
4	LumB	Normal	0.23
5	LumA	Basal	1.324e-13
6	LumA	Her2	1.303e-7
7	LumA	Normal	0.0001495
8	Basal	Her2	0.4459
9	Basal	Normal	0.8058
10	Her2	Normal	0.7212

Showing 1 to 10 of 10 entries

Figure 200 The main view of the Tool (lower section).

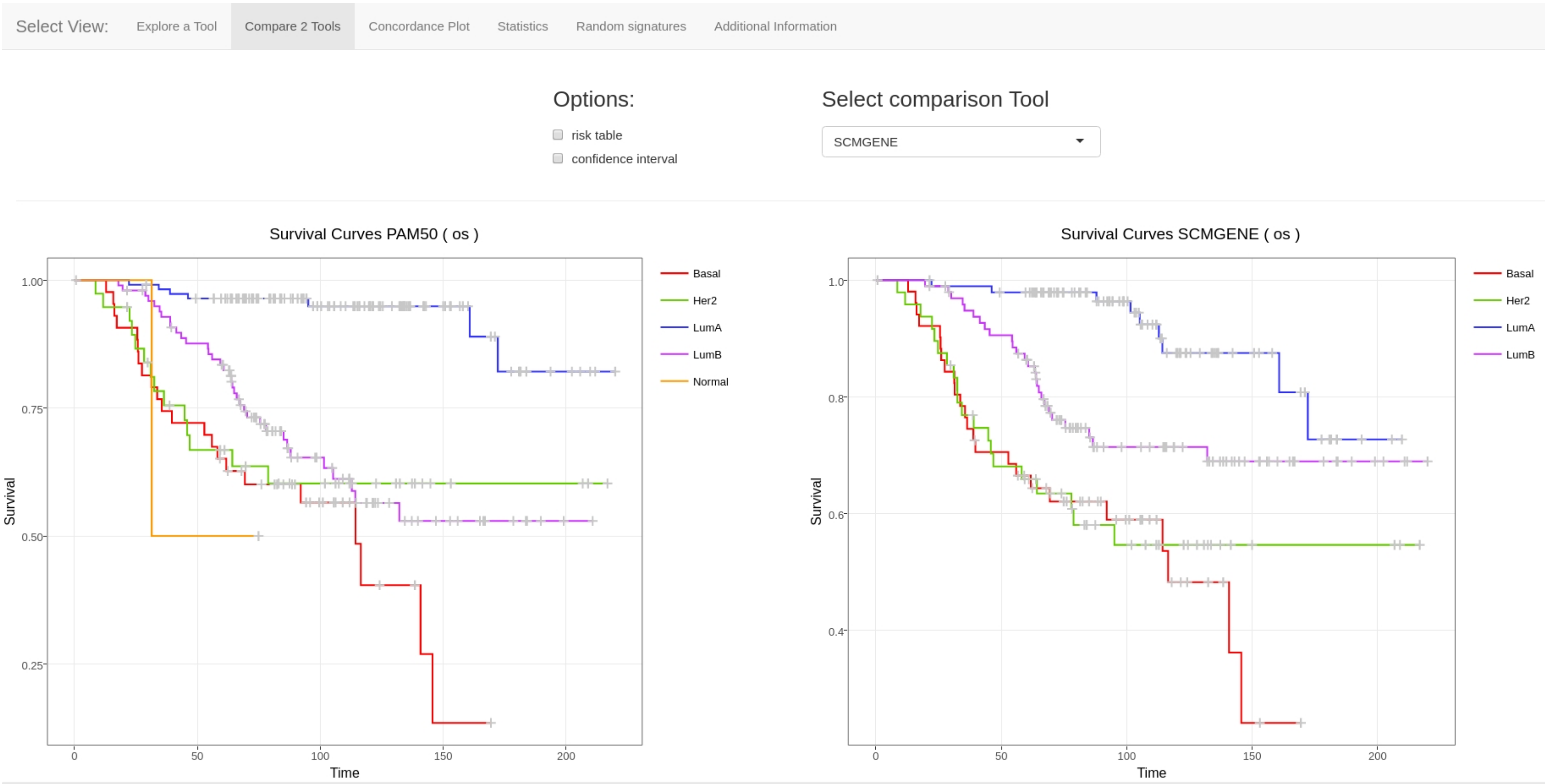
17.3.4 Compare 2 Predictors

The view 'Compare 2 Predictors' is shown in Figure 202. This view shows two Kaplan-Meier plots: one for the method selected in the view 'Explore a Predictor' and the other one for the method that can be selected in the current view. There are the same options available for the Kaplan-Meier-Plots as in the main view: one can add risk tables beneath the plots and the concordance intervals to both plots.
To assess the significance of the subtypes predicted by both methods, the hazard ratios and log-rank test p-values for each subtype combination of both methods are displayed next to each other below the Kaplan-Meier plots. Furthermore, a contingency matrix is displayed that enables the user to evaluate how much the subtype or risk score predictions of the two selected prediction methods disagree (see Figure 201).



	SCMGENE \ PAM50	Basal	Her2	LumA	LumB	Normal
1	Basal	59	3	5	2	2
2	Her2	1	33	5	12	0
3	LumA	0	2	87	16	0
4	LumB	1	1	25	82	0

Figure 201 The contingency table for PAM50 and SCMGENE on the dataset NKI as displayed in the Comparison view (lower section).



Options:

□ risk table

□ confidence interval

Select comparison Tool

SCMGENE ▾

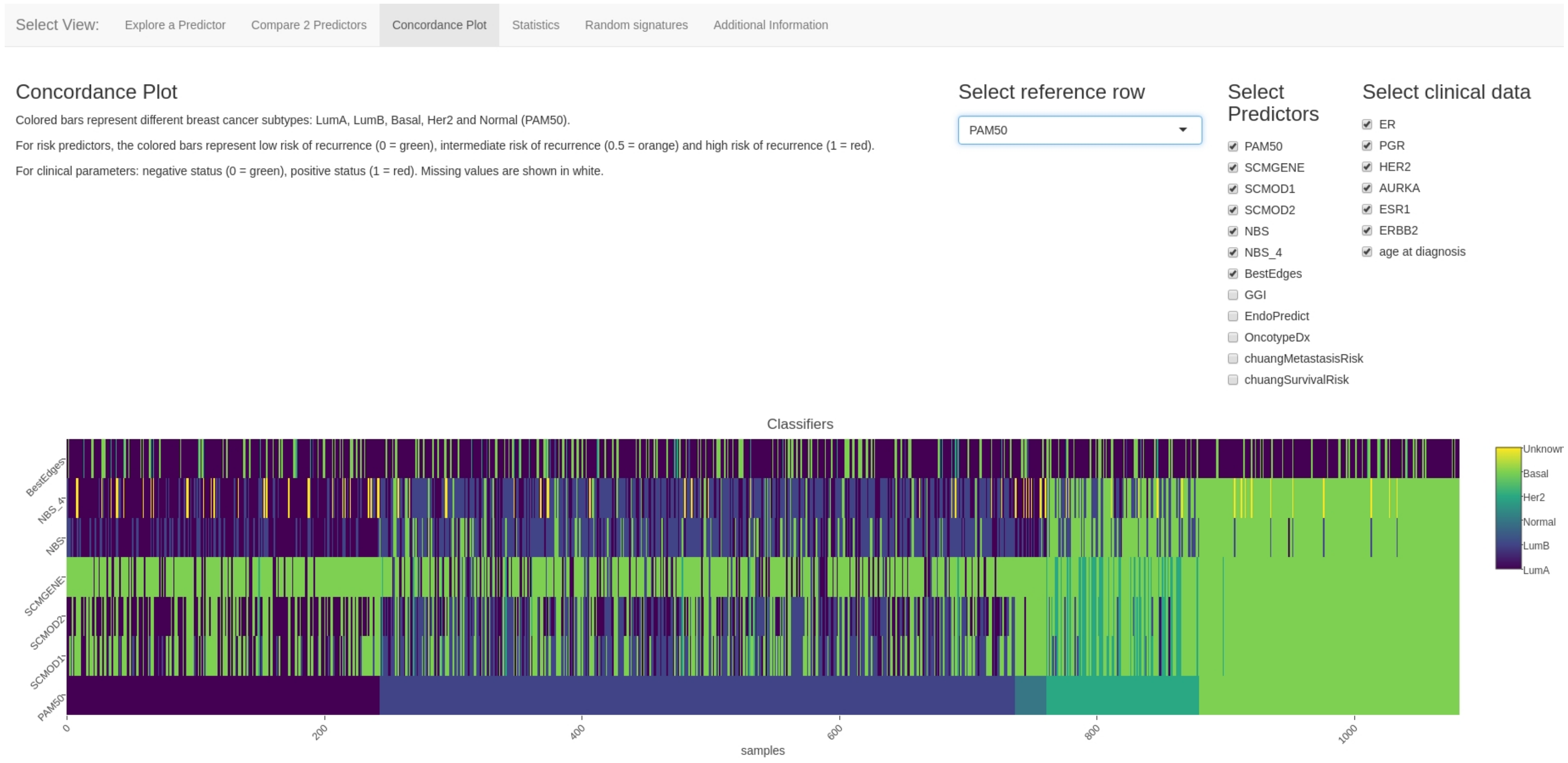
Figure 202 Compare 2 Tools - view showing PAM50 vs. SCMGENE on the NKI dataset.

17.3.5 Concordance Plot

The view 'Concordance Plot' shows a plot similar to the Concordance plot made by [331] shown in Figure 188. In the upper part, there is a short description of the plot. On the right side, several options can be selected: the checkboxes allow one to select or deselect different predictors and clinical data. A reference row, which is used to sort the samples, can also be selected. See Figure 203 for available options. It shows the upper part of the 'Concordance Plot' view with available options and the concordance plot for the classifiers on the TCGA dataset. The samples are sorted according to the prediction of PAM50.

Figure 204 is a reproduction of Figure 188 on our biggest dataset TCGA. The samples in the concordance plot are sorted by the prediction of SCMGENE, like in Figure 188. As in Figure 188, the subtypes LuminalA and LuminalB are often mixed up by the classifiers. Moreover, the right part of the classifier prediction, which is predicted to be Basal by SCMGENE shows noticeable disagreement between all other classifiers as well as GGI. EndoPredict predicts a low risk of recurrence for about half the LuminalA subtype, while OncotypeDx is not able to make a prediction at all.
The clinical parameter Her2 correlates with the Her2 subtype predicted by the classifiers (green color block in the upper part), as the clinical Her2 is positive (i.e. red = high) in this region. Other clinical parameters available for TCGA, as well as the 'age' seem not to correlate with any subtype or risk prediction.
Figure 205 shows all classifiers, risk score predictors and clinical parameters on the TCGA dataset. The samples are still sorted by SCMGENE prediction. There are several samples which are similarly classified by several predictors, but overall the predictors seem to disagree on the greater part of the samples.

Figure 206 shows the concordance plot containing all classifiers, clinical parameters and the gene expression of AURKA, ESR1 and ERBB2 (RNAseq data) on the TCGA dataset, sorted by SCMGENE prediction. The expression of the gene signature from SCMGENE correlates well with its prediction. The expression of ERBB2 (the Her2 gene) corresponds to the dark green color block from the classifiers (upper part), as well as to the clinical parameter “Her2”. The expression of ESR1 (estrogen receptor gene) is high only for the first ≈180 samples, which correspond to the luminal subtypes predicted by SCMGENE. The expression of AURKA (a gene associated with proliferation) seems not to clearly correlate with the subtypes over all samples, but for the first ≈180 samples, it divides clearly the luminal subtype (dark blue and blue) into luminal A and luminal B.



Concordance Plot

Colored bars represent different breast cancer subtypes: LumA, LumB, Basal, Her2 and Normal (PAM50).

For risk predictors, the colored bars represent low risk of recurrence (0 = green), intermediate risk of recurrence (0.5 = orange) and high risk of recurrence (1 = red).

For clinical parameters: negative status (0 = green), positive status (1 = red). Missing values are shown in white.

Select reference row

PAM50 ▾

Select Predictors

- ☑ PAM50
- ☑ SCMGENE
- ☑ SCMOD1
- ☑ SCMOD2
- ☑ NBS
- ☑ NBS_4
- ☑ BestEdges
- □ GGI
- □ EndoPredict
- □ OncotypeDx
- □ chuangMetastasisRisk
- □ chuangSurvivalRisk

Select clinical data

- ☑ ER
- ☑ PGR
- ☑ HER2
- ☑ AURKA
- ☑ ESR1
- ☑ ERBB2
- ☑ age at diagnosis

Figure 203 The 'Concordance Plot' view showing the available options and the upper part of the concordance plot.

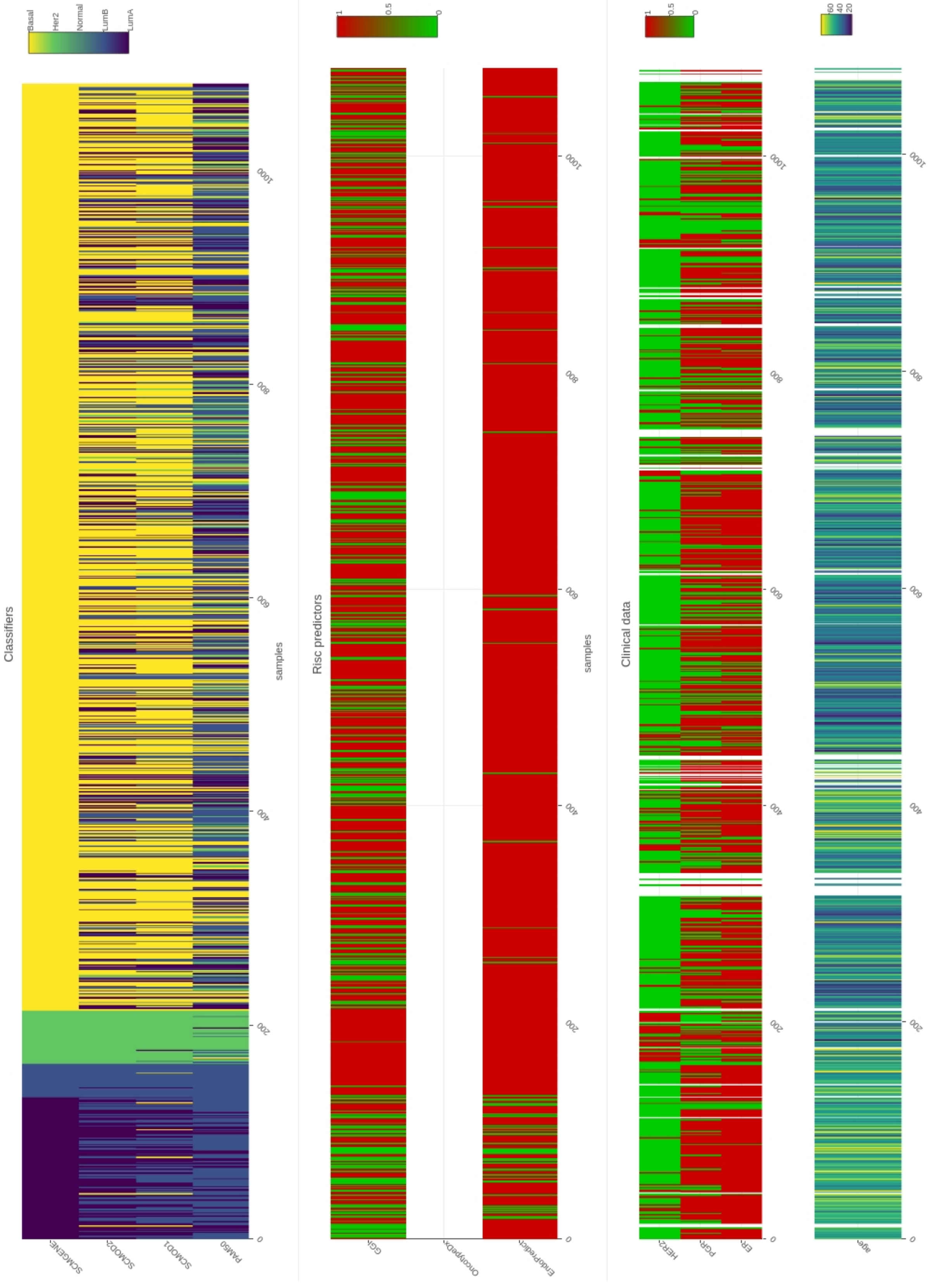


Figure 204 Concordance plot view. An attempt to reproduce the concordance plot made by [331] shown in Figure 188.

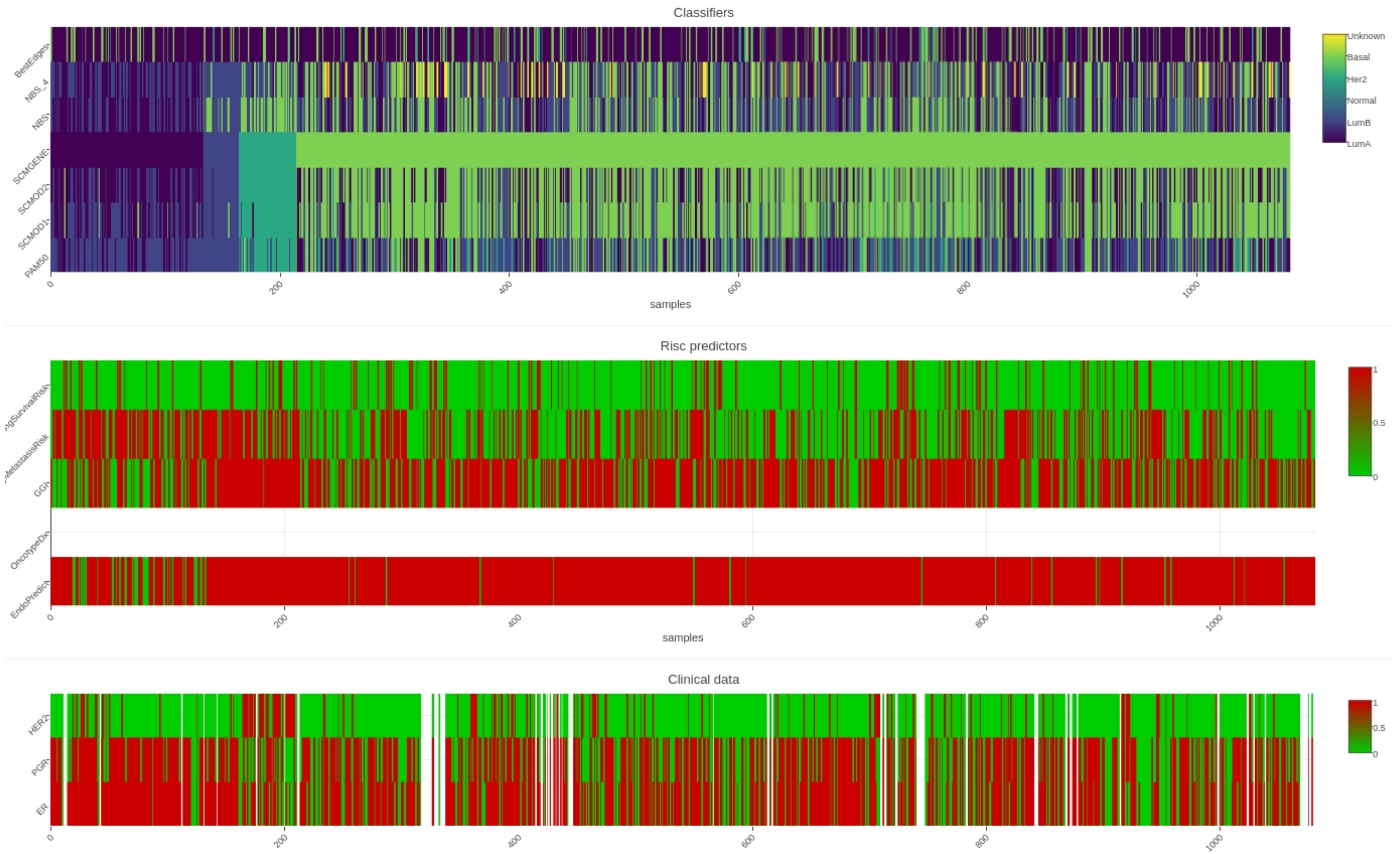


Figure 205 Concordance plot view showing the concordance plot containing all predictors and clinical parameters. Samples are sorted by SCMGENE prediction (TCGA dataset).

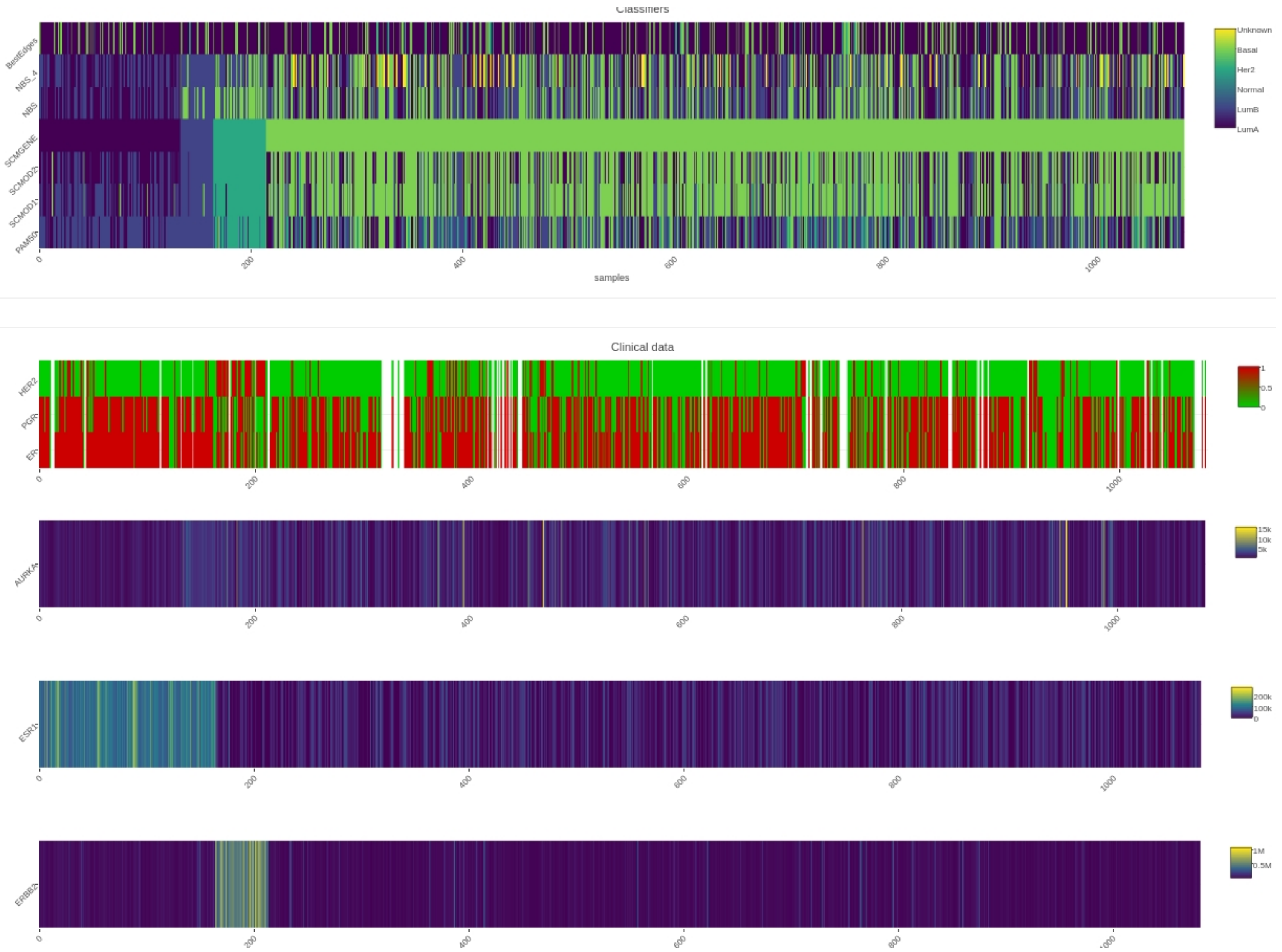
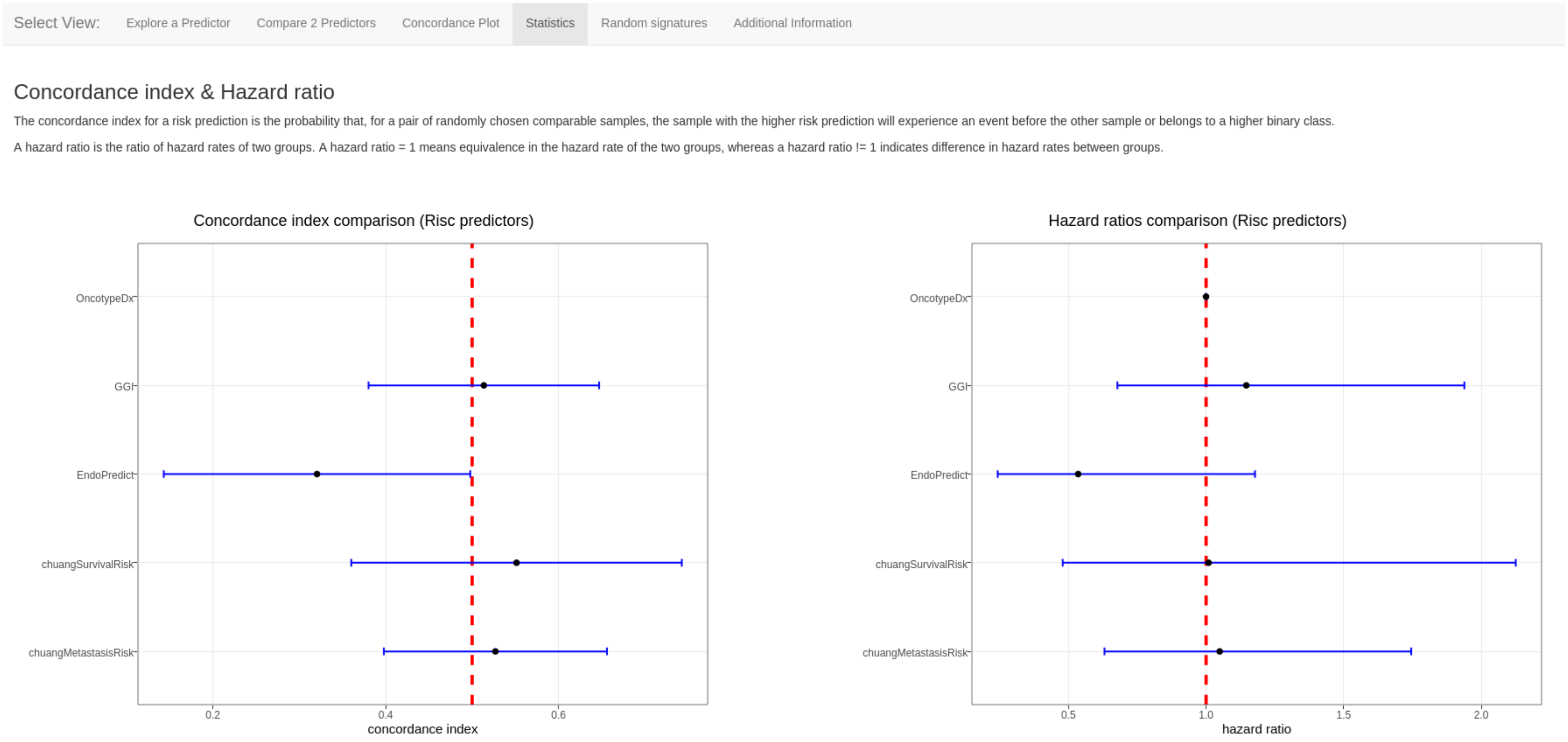


Figure 206 Concordance plot view showing the concordance plot containing all classifiers, the clinical parameters and the expression of AURKA, ESR1 and ERBB2 (RNAseq data), samples sorted by SCMGENE prediction (TCGA dataset).

17.3.6 Statistics

The Statistics view shows four forest plots: two plots visualize the concordance index and the hazard ratio for all risk score predictors (see Figure 207), two plots visualize the concordance index and the hazard ratio for all subtypes of the selected classifier (see Figure 208, selected classifier: PAM50). Both Figures show the Statistic view on the TCGA dataset.

The hazard rate and the hazard ratio are described in Section 17.2.2.3. A red dashed line on the Hazard ratio forest plot indicates a hazard rate of 1, corresponding to 'no difference between the compared groups'.
The concordance index for a risk prediction is the probability that, for a pair of randomly chosen comparable samples, the sample with the higher risk prediction will experience an event before the other sample or belongs to a higher binary class. A concordance index of 0.5 indicates that the prediction is not better than a random prediction. A red dashed line in the concordance index forest plot indicates a concordance index of 0.5.
The black circles correspond to the concordance index or hazard ratio and the blue lines correspond to the confidence intervals.
In the concordance index plot (left part of Figure 207), the only confidence interval not crossing the dashed line is that of EndoPredict. All other predictors seem to be not significantly better than random predictors. In the second concordance plot (left part of Figure 207) which compares the subtypes predicted by PAM50 on the TCGA dataset, only the confidence intervals of “Her2 vs. LumA” and “Her2 vs. LumB” seem to be better than a random prediction.
The hazard ratio plot (right side of the Figure 207) shows that for all risk score predictors, the difference between “high risk” and “low risk” is not significant, as all the confidence intervals intercept with the dashed line (hazard ratio = 1). Similar to the concordance plot in Figure 208, the hazard plot indicates a significant difference between the subtypes Her2 - LumA and Her2 - LumB. Moreover, the difference between Basal and LumA subtypes seems also to be significant. The confidence intervals of all other subtypes intercept with the dashed line.



Concordance index & Hazard ratio

The concordance index for a risk prediction is the probability that, for a pair of randomly chosen comparable samples, the sample with the higher risk prediction will experience an event before the other sample or belongs to a higher binary class.

A hazard ratio is the ratio of hazard rates of two groups. A hazard ratio = 1 means equivalence in the hazard rate of the two groups, whereas a hazard ratio != 1 indicates difference in hazard rates between groups.

Figure 207 Statistics view (upper part) showing a concordance index forest plot (left side) and a hazard rate forest plot (right side) for the risk score predictors on the TCGA dataset.

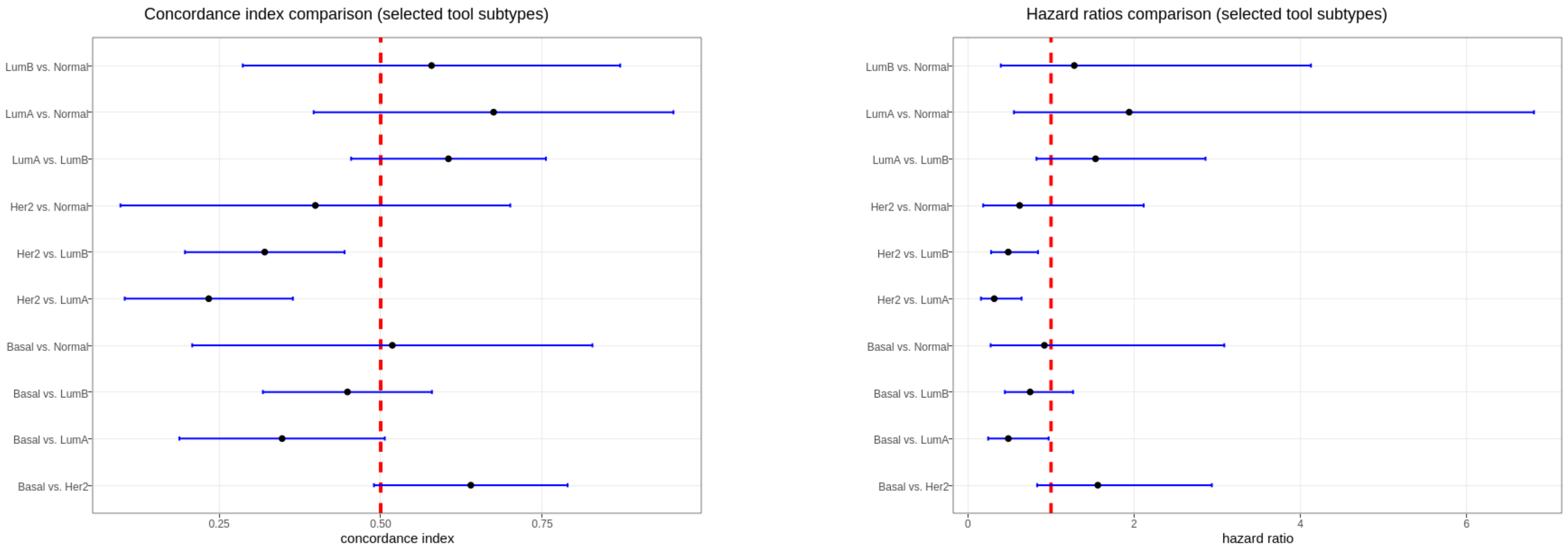


Figure 208 Statistics view (lower part) showing a concordance index forest plot (left side) and a hazard rate forest plot (right side) for the PAM50 classifier on the TCGA dataset.

17.3.7 Random signatures

An analysis of random signatures similar to the analysis done by Venet et al. (see section 17.2.7) was performed on the 14 breast cancer datasets. The differences to the quite simple approach used by Venet et al. (described in detail in section 17.2.7) are depicted below.

Figure 209 shows the workflow of the random signature analysis as a petri net. The random signature analysis was executed separately for SCMGENE and PAM50. First, a random gene sample was drawn from the EXPO dataset. In order to be able to make a prediction on every test dataset, only genes which were contained in EXPO and all other datsets were considered, resulting in 17.053 EXPO genes. The random gene set contained three random genes for SCMGENE and 50 random genes for PAM50. These gene set was then used to train SCMGENE/PAM50 in order to obtain a 'random SCMGENE'/'random PAM50' classifiers. All of the 14 remaining datasets were used to make predictions with the 'random classifiers'.

For every pair of predicted subtypes, a log rank test p-value was computed and the best p-value was chosen to represent the prediction quality of one 'random classifier'. The complete process was repeated 500 times for both classifiers.

The complete workflow was repeated two more times and the random genes were drawn from different subsets of the EXPO genes. First, the random genes were drawn from EXPO genes excluding the 51 'intrinsic' genes (50 genes from PAM50 and AURKA, as ESR1 and ERBB2 were already included in the PAM50 intrinsic genes). The second time, random genes were drawn from EXPO genes excluding the intrinsic genes and the 'proliferation' genes (taken from supplementary data of Venet et al. [359]). the 'intrinsic' genes were not included in the 131 'proliferation' genes of Venet et al.. 90 'proliferation' genes and 38 'intrinsic' genes were contained in the EXPO dataset and were removed.

Furthermore, the original SCMGENE and PAM50 tools were used to predict the subtypes from the 14 datasets (except EXPO) as well as SCMGENE and PAM50 trained on the EXPO dataset.

Figures 210, 211 and 212 show the results from the random signature analysis. Figures 210 and 210 compare random signatures with real signatures. The two plots on the left side correspond to SCMGENE, the two plots on the right side correspond to PAM50. The x-axis corresponds to the log10 p-values from the log rank tests, the y-axis contains the datasets. Both Figures show results for the Overall Survival (OS), but the other survival types (distant metastasis-free survival (DMFS) and relapse-free survival (RFS)) can also be selected in the interactive shiny app. The p-values computed on the random gene sets are shown as violin plots (upper part) and box plots (lower part). In both figures, a circle corresponds to a p-value computed on a dataset with the original PAM50/SCMGENE and a triangle corresponds to a p-value computed on a dataset using the classifiers trained on the EXPO dataset. A dashed blue line correspond to the log10(0.05) p-value.

Figure 210 show the result for the random genes, which where drawn from the complete EXPO gene set, Figure 211 show the results for the random genes drawn from EXPO genes excluding the 'intrinsic' and 'proliferation' gene sets. Both results look very similar: most of the random signatures have non-significant p-values, while several random signatures have a lower p-value than a real signature. Furthermore, some real signatures also have non-significant p-values. In most cases, there is only a small or no difference between the real classifiers and those trained on EXPO.

Figure 212 summarizes the results over all random gene sets and all survival types. The upper part shows the result for PAM50, the lower part shows the result for SCMGENE. PAM50, which uses 50 genes for the classification seems to outperform SCMGENE, which uses only 3 genes, as it has less p-values, which are weaker than the median of random signatures compared to SCMGENE. SCMGENE on the other side, have slightly more p-values, which are stronger than the lower whisker of the random p-values.

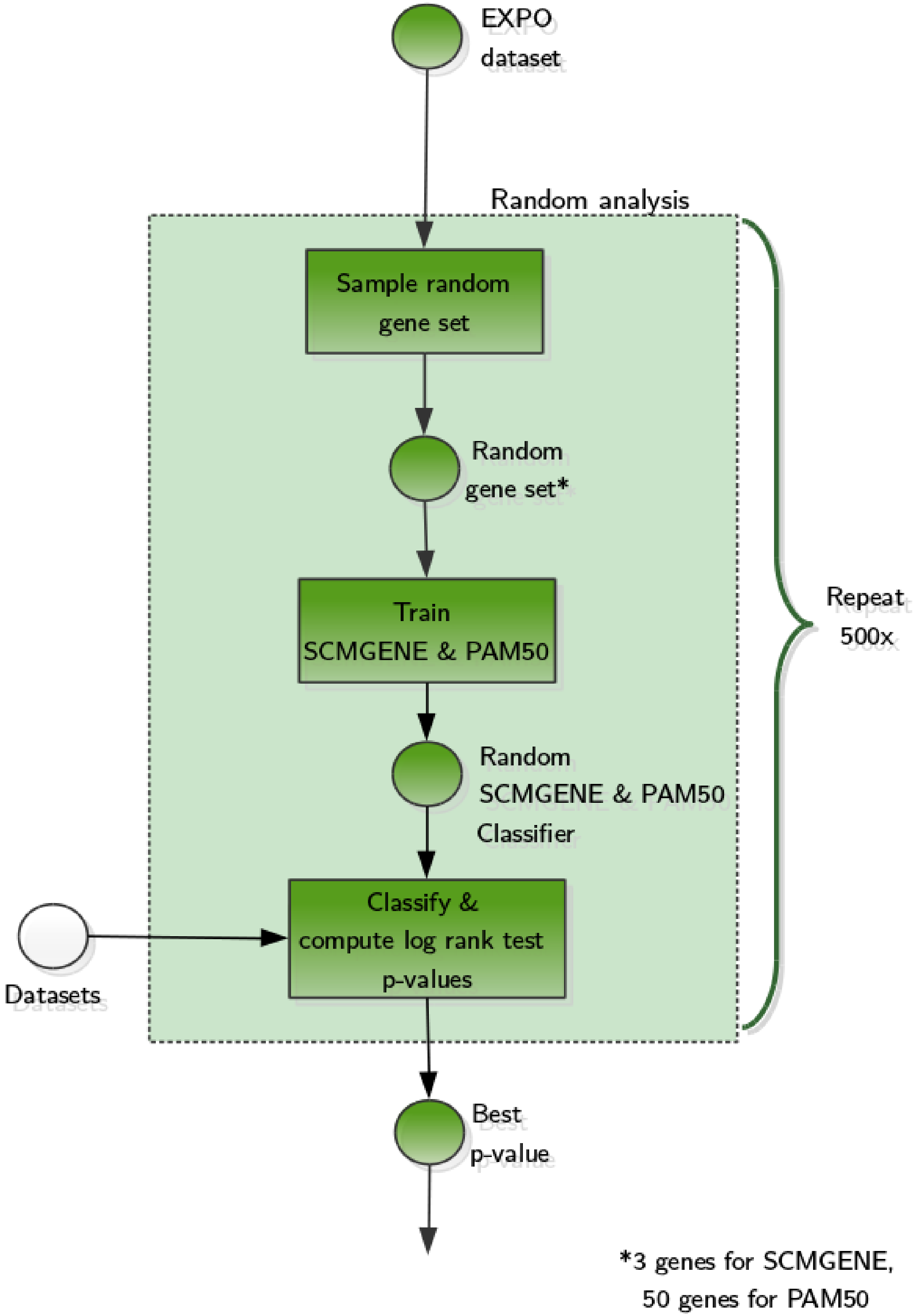


Figure 209 Workflow submodule: Random signatures analysis

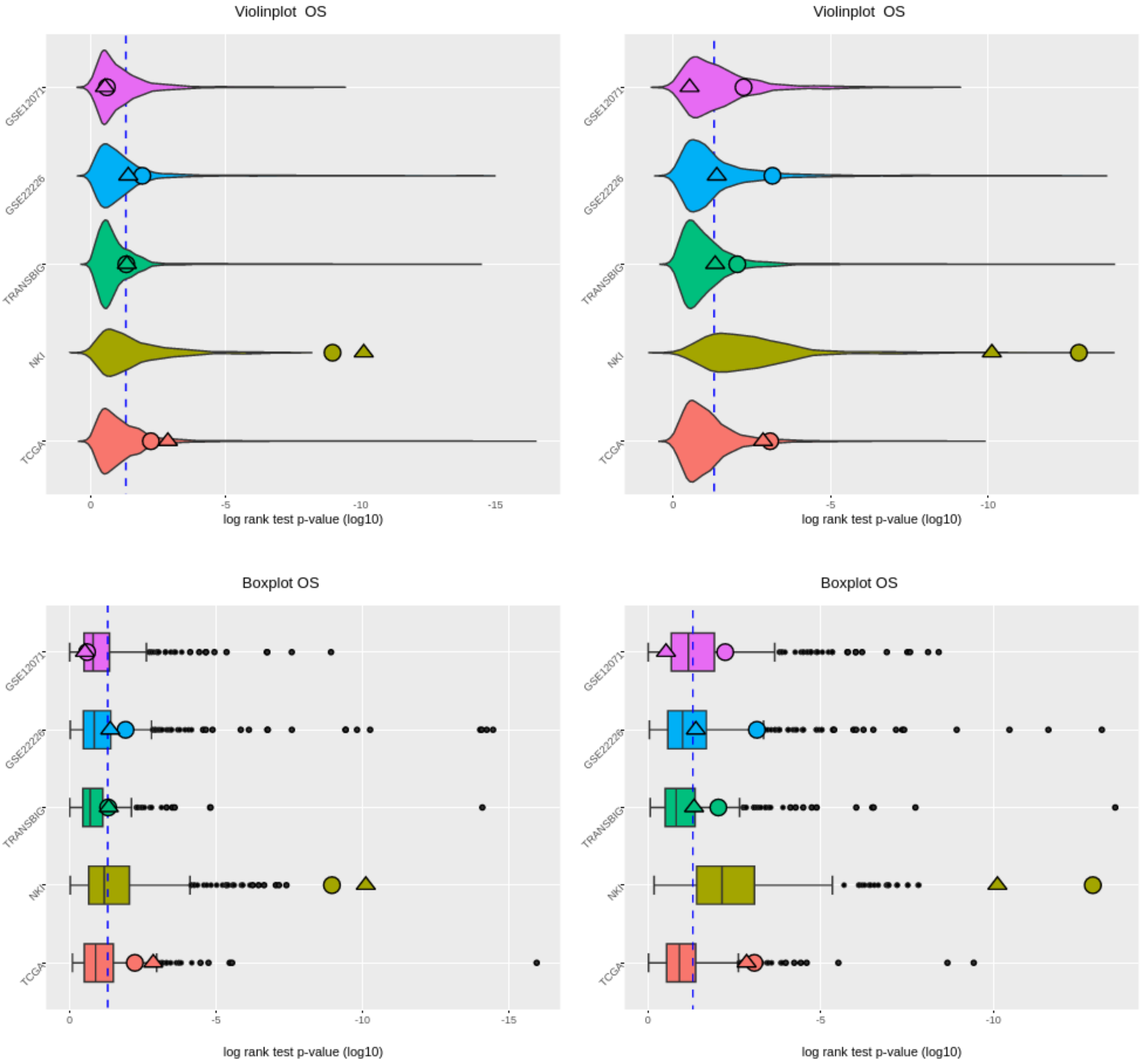
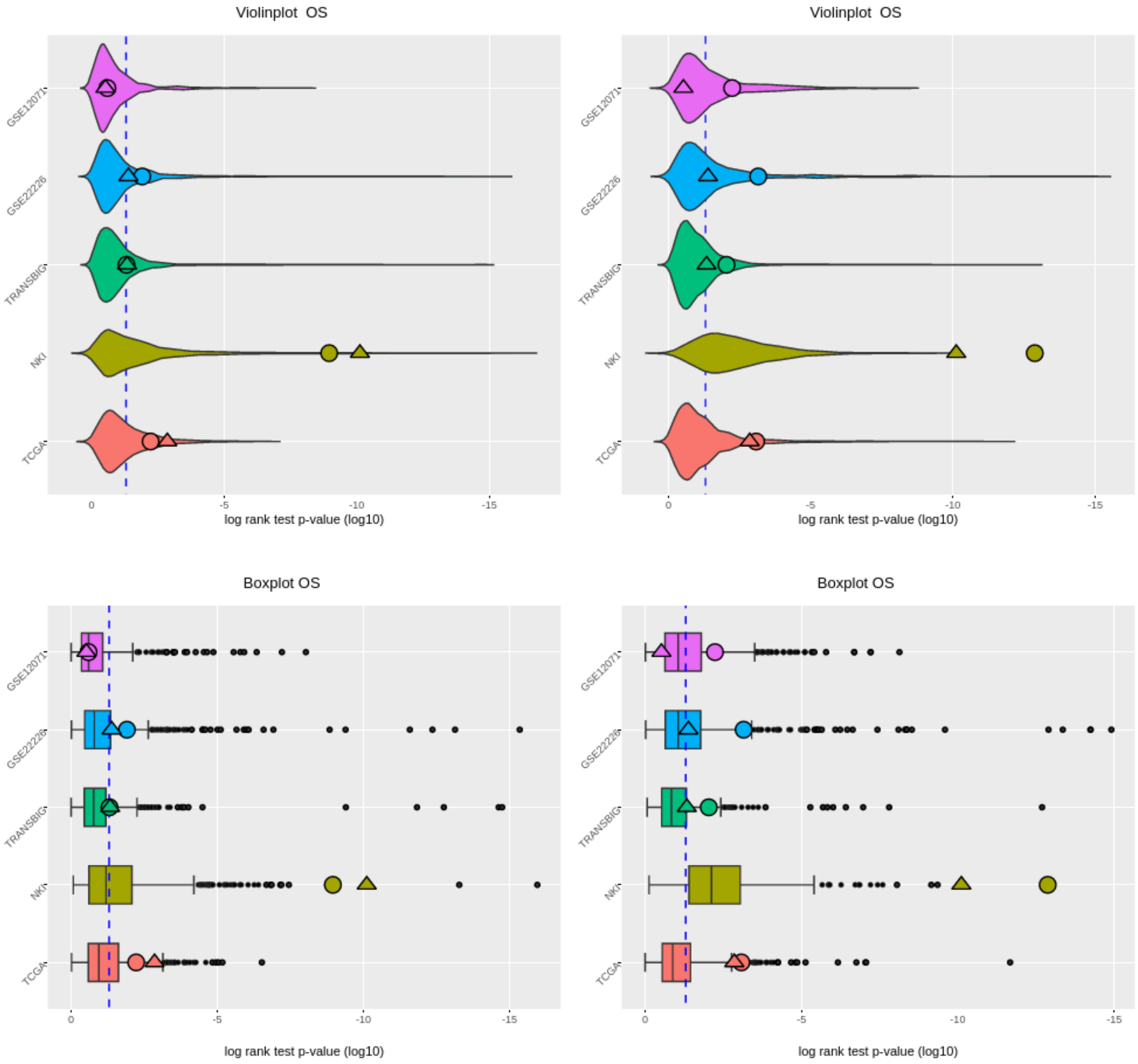
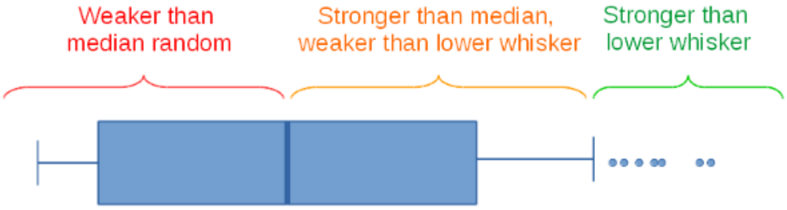
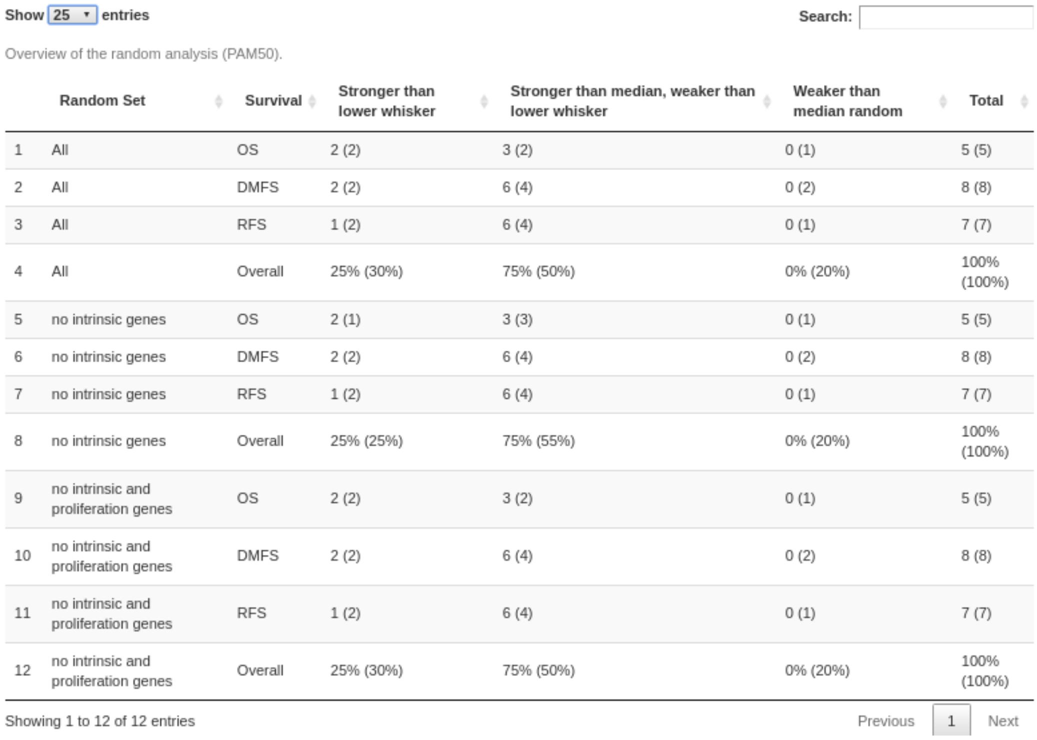


Figure 210 Comparison of real signatures (SCMGENE and PAM50) with random signatures of the same size. left side: SCMGENE, right side: PAM50. Random signatures were drawn from all EXPO genes.



■ **Figure 211** Comparison of real signatures (SCMGENE and PAM50) with random signatures of the same size. left side: SCMGENE, right side: PAM50. Random signatures were drawn from EXPO genes excluding the intrinsic and proliferation genes.


Weaker than median random
Stronger than median, weaker than lower whisker
Stronger than lower whisker

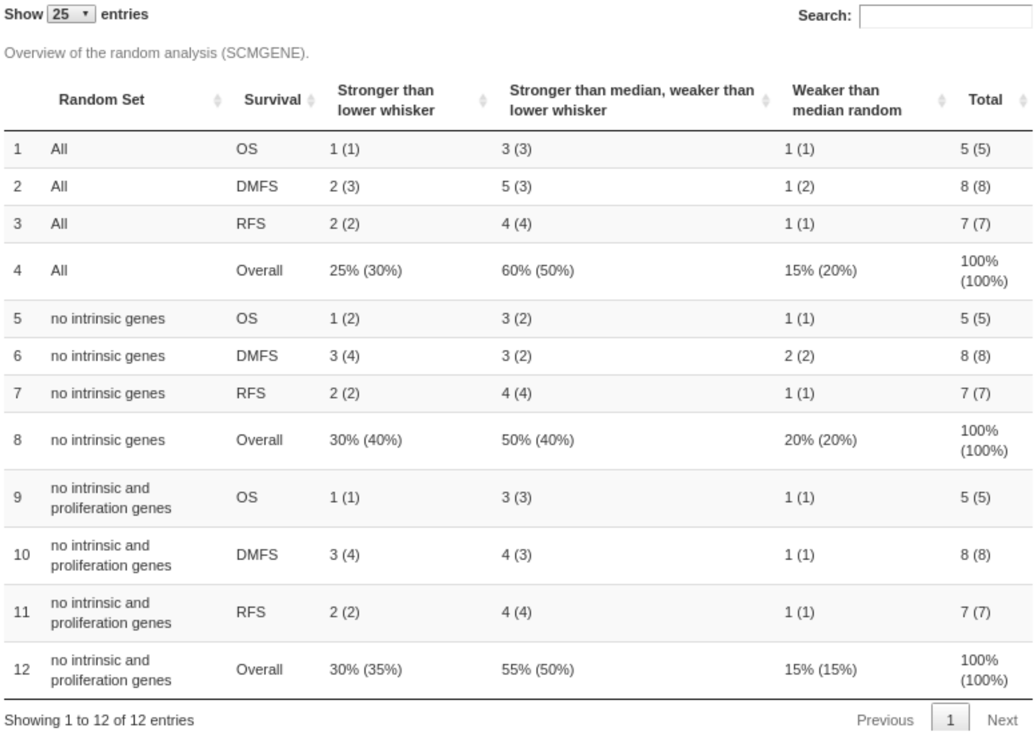


Show 25 entries Search:

Overview of the random analysis (PAM50).

	Random Set	Survival	Stronger than lower whisker	Stronger than median, weaker than lower whisker	Weaker than median random	Total
1	All	OS	2 (2)	3 (2)	0 (1)	5 (5)
2	All	DMFS	2 (2)	6 (4)	0 (2)	8 (8)
3	All	RFS	1 (2)	6 (4)	0 (1)	7 (7)
4	All	Overall	25% (30%)	75% (50%)	0% (20%)	100% (100%)
5	no intrinsic genes	OS	2 (1)	3 (3)	0 (1)	5 (5)
6	no intrinsic genes	DMFS	2 (2)	6 (4)	0 (2)	8 (8)
7	no intrinsic genes	RFS	1 (2)	6 (4)	0 (1)	7 (7)
8	no intrinsic genes	Overall	25% (25%)	75% (55%)	0% (20%)	100% (100%)
9	no intrinsic and proliferation genes	OS	2 (2)	3 (2)	0 (1)	5 (5)
10	no intrinsic and proliferation genes	DMFS	2 (2)	6 (4)	0 (2)	8 (8)
11	no intrinsic and proliferation genes	RFS	1 (2)	6 (4)	0 (1)	7 (7)
12	no intrinsic and proliferation genes	Overall	25% (30%)	75% (50%)	0% (20%)	100% (100%)

Showing 1 to 12 of 12 entries Previous 1 Next



Show 25 entries Search:

Overview of the random analysis (SCMGENE).

	Random Set	Survival	Stronger than lower whisker	Stronger than median, weaker than lower whisker	Weaker than median random	Total
1	All	OS	1 (1)	3 (3)	1 (1)	5 (5)
2	All	DMFS	2 (3)	5 (3)	1 (2)	8 (8)
3	All	RFS	2 (2)	4 (4)	1 (1)	7 (7)
4	All	Overall	25% (30%)	60% (50%)	15% (20%)	100% (100%)
5	no intrinsic genes	OS	1 (2)	3 (2)	1 (1)	5 (5)
6	no intrinsic genes	DMFS	3 (4)	3 (2)	2 (2)	8 (8)
7	no intrinsic genes	RFS	2 (2)	4 (4)	1 (1)	7 (7)
8	no intrinsic genes	Overall	30% (40%)	50% (40%)	20% (20%)	100% (100%)
9	no intrinsic and proliferation genes	OS	1 (1)	3 (3)	1 (1)	5 (5)
10	no intrinsic and proliferation genes	DMFS	3 (4)	4 (3)	1 (1)	8 (8)
11	no intrinsic and proliferation genes	RFS	2 (2)	4 (4)	1 (1)	7 (7)
12	no intrinsic and proliferation genes	Overall	30% (35%)	55% (50%)	15% (15%)	100% (100%)

Showing 1 to 12 of 12 entries Previous 1 Next

Figure 212 Comparison of real signatures (SCMGENE and PAM50) with random signatures of the same size. Summary of the result as tables.

17.3.8 Random network signatures

We also used a relatively simple approach to analyse network-based signatures that allows for randomization in order to evaluate if there is a difference to the non-network signatures. Figure 214 shows the workflow of the analysis of network-derived signatures as a petri net. Instead of using a gene expression matrix of patients versus genes as usual, we created a edge matrix of patients versus edges, where each edge in a network was assigned a score based on the gene expression values of the incident nodes. Two different ways to score the edges were examined: assigning each edge to a discrete score class (either $-1, -0.5, 0.5$ or 1) or assigning a continuous value to each edge defined as the average of the absolute z-scores of the two incident genes.

For both scoring schemas, the gene expression value of the two genes incident to the edge is first described a z-score for each patient based on the expression values of the gene of question over all patients. Z-scores for each patient i were calculated according to the following formula, where g_A and g_B are the vectors of expression values for genes A and B (incident to the edge) over all patients:

$$z_{A,i} = \frac{g_{A,i} - \mu(g_A)}{sd(g_A)} \tag{29}$$

To obtain a score for the edge from these z-scores, the following schema was used for the first approach to define the score of the edge:

	z_B < -1	-1 <= z_B <= 1	z_B > 1
z_A < -1	1	-0.5	0.5
-1 <= z_A <= 1	-0.5	-1	-0.5
z_A > 1	0.5	-0.5	1

The edge score of a patient is high (= 1) if both incident genes have high z-scores (z > 1) or if both have low z-scores (z < -1) for the given patient. In this case, the edge would represent an activating interaction between the genes. The score is relatively high (= 0.5) if one of the genes has a high z-score and the other has a low z-score, which might be indicative of an inhibitive interaction between the two genes. If the expression of both genes does not change compared to the other patients (z-score around 0), the the edge is assigned a low value (= -1) or relatively low value (= -0.5) if the expression of the other gene is high or low.

For the other scoring schema tested, the edge score is defined as $\frac{|z_A|+|z_B|}{2}$.

To create a matrix of edge scores for both approaches, gene expression values were taken from the EXPO dataset and the network curated by Chuang et al. [347] was used. The obtained edge matrices were then clustered into four clusters using the hierarchical clustering method of PAM (described in a previous section). The 353 patients in the EXPO cohort were assigned to the clusters as follows:

	discrete edge scores	continuous edge scores
cluster 1	104	249
cluster 2	27	61
cluster 3	148	25
cluster 4	74	18

Table 16 Sizes of clusters obtained from clustering the edge score matrices.

Next, t-test were performed to find the edges that are most significantly associated with the cluster assignment and the Bonferroni method was used to account for the multiple testing. Then the top 25 edges of each cluster were chosen for the discrete edge score approach, resulting in a set of 91 edges (there was some overlap) used to train the 'real' classifier. As the fourth cluster is very small for the approach using continuous edge scores, only the top 3 edges were used for this cluster, all other edges were not associated with this cluster. To account for this small set, only the top 20 edges were used for the other three clusters, resulting in a set of 59 edges. With each set of significantly associated edges, a nearest neighbour classifier was then trained.

In the last step, a set of 91 and 59 edges, respectively, was randomly sampled 100 times from the complete network. With each random edge set, a classifier was trained. The performance of the random classifier was assessed by predicting subtype assignments on all datasets and computing log-rank test p-values for each combination of subtypes. The smallest of those p-values was recorded for the random classifier on all datasets.

Figure 213 shows the result of this analysis. The 'real' classifier is indicated with a circle and can be compared to the distribution of all random classifiers. Most random edge sets do not predict subtypes with different survival rates at a significance level of 0.05, whereas a few random edge sets appear to predict subtypes with different survival rates, at least on some datasets. Interestingly, the 'real' classifiers also fail to define subtypes with different survival rates on most datasets. This is probably due to the simplicity of our approach and maybe also a result of the unequal cluster sizes (see Table 16).

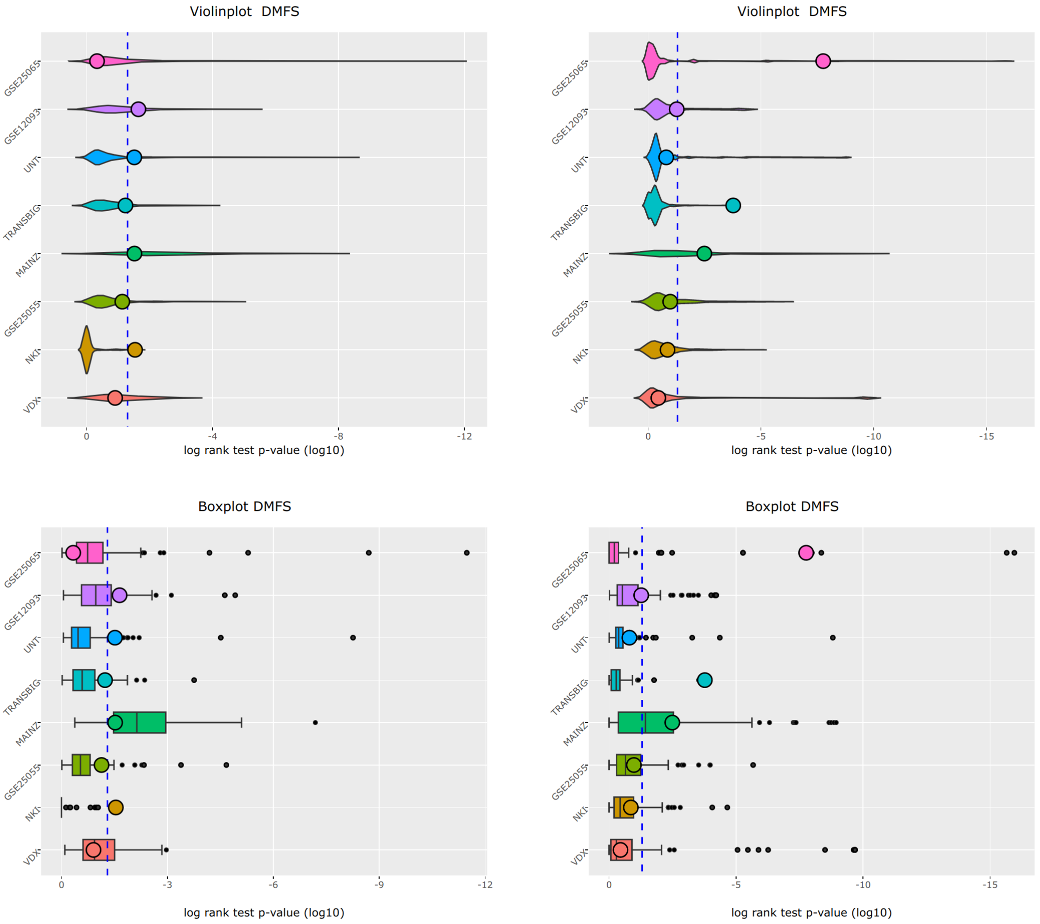


Figure 213 Comparison of log-rank test p-values between subtypes for 'real' network-based edge sets and random edge sets of the same size. For the two plots on the left, edges were assigned a discrete score, while for the plots on the right continuous values from averaging absolute z-scores were used. For this analysis the network curated by Chuang et al. [347] was used.

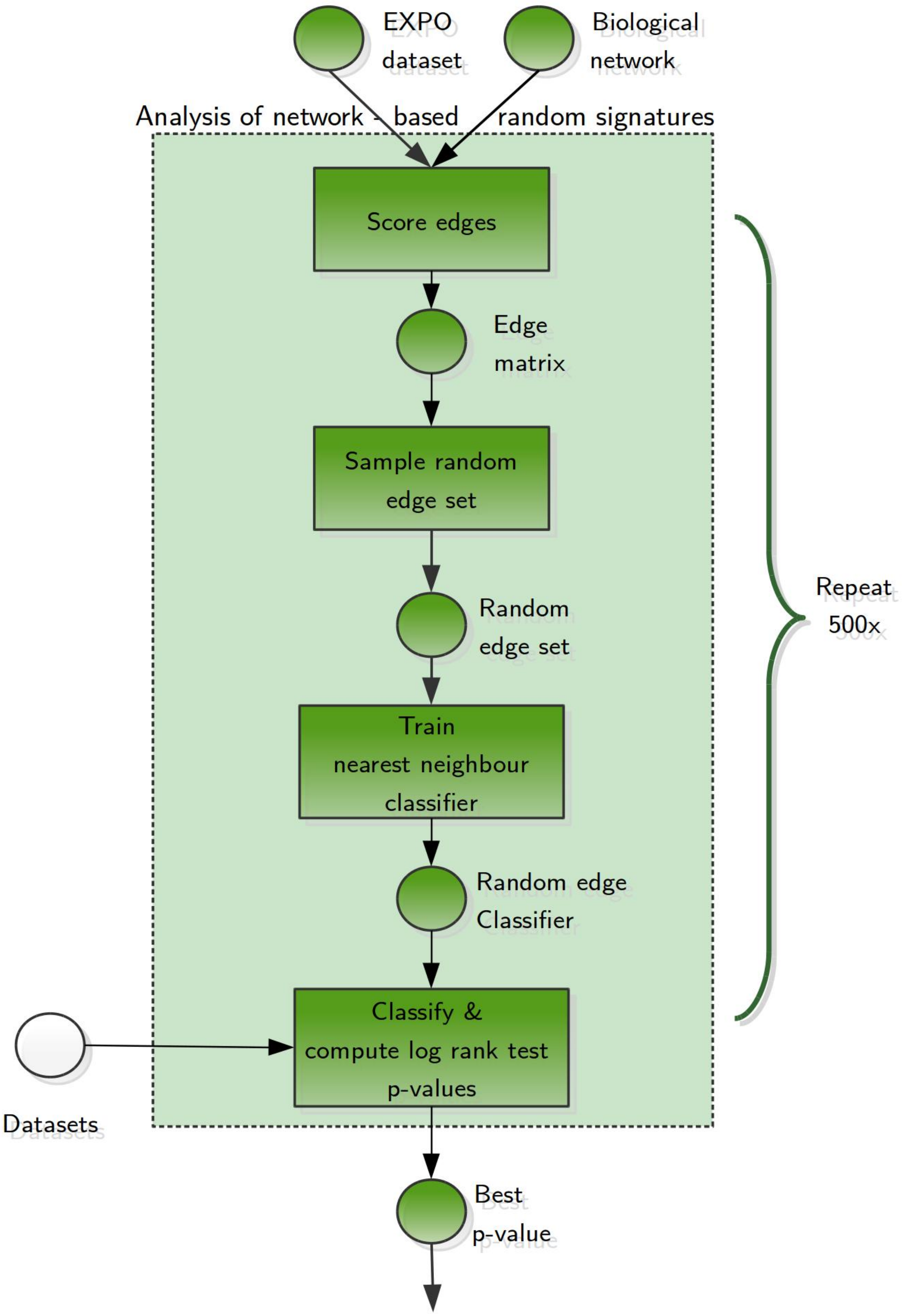
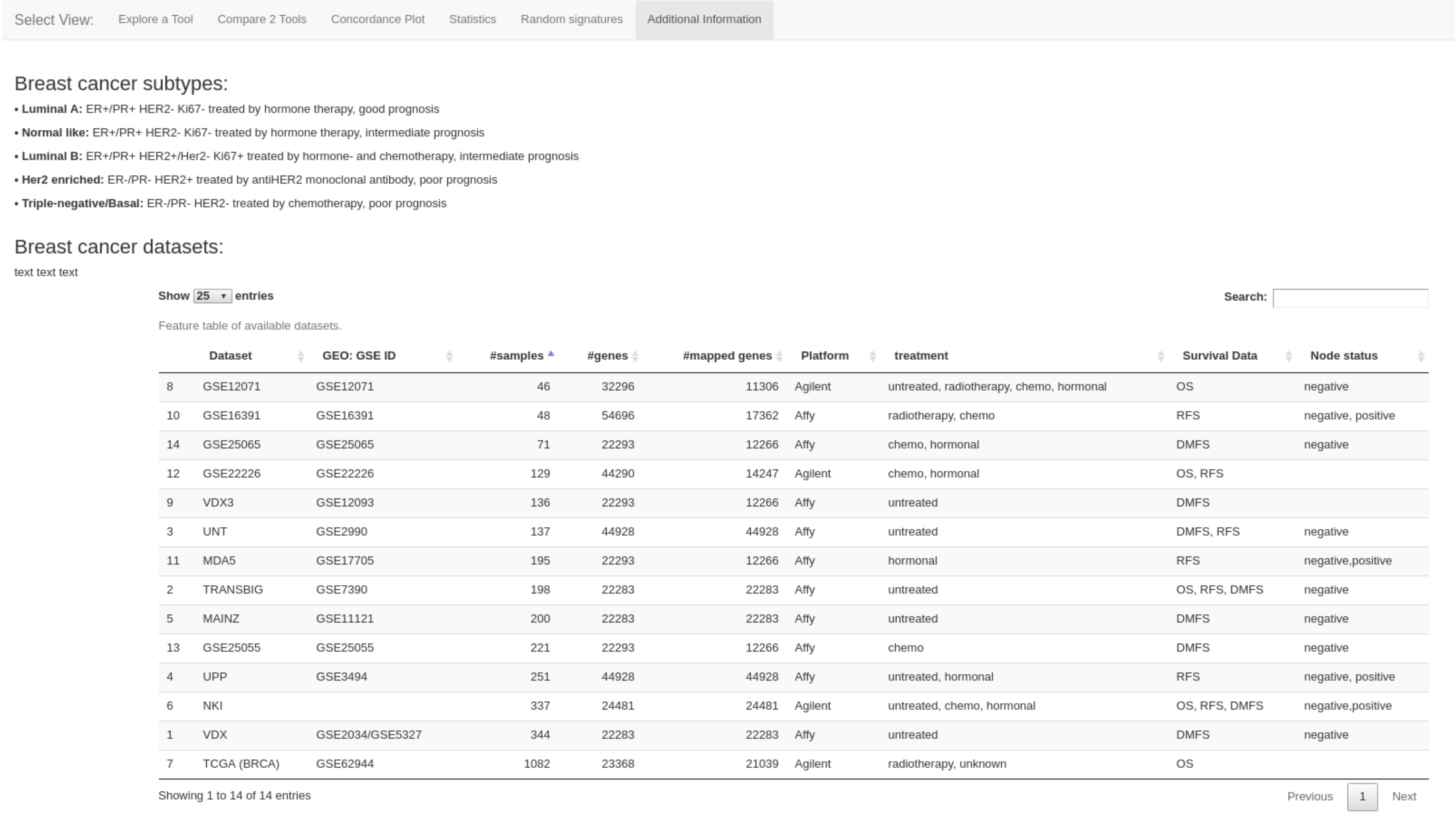


Figure 214 Workflow submodule: Random signatures analysis (Network-based)

17.3.9 Additional Information

The 'Additional Information view' contains various informations on the topic 'Breast Cancer Subtype Classification' like the definitions of the intrinsic subtypes and a table containing information about all datasets integrated into the application (a shorter version of Table 17). The upper section of this view is shown in Figure 215.



Breast cancer subtypes:

- **Luminal A:** ER+/PR+ HER2- Ki67- treated by hormone therapy, good prognosis
- **Normal like:** ER+/PR+ HER2- Ki67- treated by hormone therapy, intermediate prognosis
- **Luminal B:** ER+/PR+ HER2+/Her2- Ki67+ treated by hormone- and chemotherapy, intermediate prognosis
- **Her2 enriched:** ER-/PR- HER2+ treated by antiHER2 monoclonal antibody, poor prognosis
- **Triple-negative/Basal:** ER-/PR- HER2- treated by chemotherapy, poor prognosis

Breast cancer datasets:

text text text

Show 25 entries

Search:

Feature table of available datasets.

	Dataset	GEO: GSE ID	#samples	#genes	#mapped genes	Platform	treatment	Survival Data	Node status
8	GSE12071	GSE12071	46	32296	11306	Agilent	untreated, radiotherapy, chemo, hormonal	OS	negative
10	GSE16391	GSE16391	48	54696	17362	Affy	radiotherapy, chemo	RFS	negative, positive
14	GSE25065	GSE25065	71	22293	12266	Affy	chemo, hormonal	DMFS	negative
12	GSE22226	GSE22226	129	44290	14247	Agilent	chemo, hormonal	OS, RFS	
9	VDX3	GSE12093	136	22293	12266	Affy	untreated	DMFS	
3	UNT	GSE2990	137	44928	44928	Affy	untreated	DMFS, RFS	negative
11	MDA5	GSE17705	195	22293	12266	Affy	hormonal	RFS	negative,positive
2	TRANSBIG	GSE7390	198	22283	22283	Affy	untreated	OS, RFS, DMFS	negative
5	MAINZ	GSE11121	200	22283	22283	Affy	untreated	DMFS	negative
13	GSE25055	GSE25055	221	22293	12266	Affy	chemo	DMFS	negative
4	UPP	GSE3494	251	44928	44928	Affy	untreated, hormonal	RFS	negative, positive
6	NKI		337	24481	24481	Agilent	untreated, chemo, hormonal	OS, RFS, DMFS	negative,positive
1	VDX	GSE2034/GSE5327	344	22283	22283	Affy	untreated	DMFS	negative
7	TCGA (BRCA)	GSE62944	1082	23368	21039	Agilent	radiotherapy, unknown	OS	

Showing 1 to 14 of 14 entries

Figure 215 Additional Information - view, upper section showing the definitions of the breast cancer subtypes and an overview of the datasets.

17.4 Conclusion

When comparing the different subtype classifiers and risk score predictors on different datasets, we found that the agreement between the various methods is very dependent on the dataset of choice. In general, the predictors predict the same subtypes for many patients, but there are also many cases where the prediction are quite different (e.g Luminal A and Basal) for the same patient.
The analysis of survival rates obtained from random signatures of the same size as the real signatures of SCMGENE and PAM50 (3 and 50 genes), respectively, has shown that with most random signatures a smaller difference in survival between the subtypes is found. For the majority of the random signatures, there is no significant difference between the subtypes using the log-rank test. But on nearly all datasets, some random signatures perform better than the real signatures. For the network-derived random signatures, the results are similar: most random edge sets do not result in a significant difference in survival rates between the subtypes, but for these signatures also the 'real' edge set often fails to identify a significant difference in survival, although they are in most cases better than the random sets. A possible reason for this is the simplicity of the method to derive these with these edge sets and predict with them.

18 Appendix

Nr	Dataset	GEO:GSEID	# samples	# genes	# mapped	Platform	Treatment	Survival	Node status	Source	Ref.
1	VDX	GSE2034/ GSE5327	344	22283	22283	Affy	untreated	DMFS	negative	breastCancer VDX	[316]
2	TRANS BIG	GSE7390	198	22283	22283	Affy	untreated	OS, RFS, DMFS	negative	breastCancer TRANSBIG	[317]
3	UNT	GSE2990	137	44928	44928	Affy	untreated	DMFS, RFS	negative	breastCancer UNT	[318]
4	UPP	GSE3494	251	44928	44928	Affy	untreated, hormonal	RFS	negative, positive	breastCancer UPP	[319]
5	MAINZ	GSE11121	200	22283	22283	Affy	untreated	DMFS	negative	breastCancer MAINZ	[320]
6	NKI	NA	337	24481	24481	Agilent	untreated, chemo, hormonal	OS, RFS, DMFS	negative, positive	breastCancer NKI	[321]
7	TCGA (BRCA)	GSE62944	1082	23368	21039	Agilent	radiotherapy, unknown	OS	NA	Experiment Hub	[323], [322]
8	GSE 12071	GSE12071	46	32296	11306	Agilent	untreated, radiotherapy, chemo, hormonal	OS	negative	curated BreastData	[325], [324]
9	VDX3	GSE12093	136	22293	12266	Affy	untreated	DMFS	NA	curated BreastData	[326], [324]
10	GSE 16391	GSE16391	48	54696	17362	Affy	radiotherapy, chemo	RFS	negative, positive	curated BreastData	[327], [324]
11	MDA5	GSE17705	195	22293	12266	Affy	hormonal	RFS	negative, positive	curated BreastData	[328], [324]
12	GSE 22226	GSE22226	129	44290	14247	Agilent	chemo, hormonal	OS, RFS	NA	curated BreastData	[329], [324]
13	GSE 25055	GSE25055	221	22293	12266	Affy	chemo	DMFS	negative	curated BreastData	[330], [324]
14	GSE 25065	GSE25065	71	22293	12266	Affy	chemo, hormonal	DMFS	negative	curated BreastData	[330], [324]
15	EXPO*	GSE2109	353	54675	54675	Affy	NA	NA	NA	[331]	[331]

Table 17 Breast Cancer Datasets. Dataset marked with * was used only for the training of the random classifiers.

References

[1] Xianjun Dong and Zhiping Weng. "The correlation between histone modifications and gene expression". In: *Epigenomics* 5.2 (2013), pp. 113–116. DOI: `10.2217/epi.13.13`. arXiv: `NIHMS150003`. URL: `http://www.futuremedicine.com/doi/10.2217/epi.13.13`.

[2] Christoph M. Koch et al. "The landscape of histone modifications across 1% of the human genome in five human cell lines". In: *Genome Research* 17.6 (2007), pp. 691–707. DOI: `10.1101/gr.5704207`.

[3] Artem Barski et al. "High-Resolution Profiling of Histone Methylations in the Human Genome". In: *Cell* 129.4 (2007), pp. 823–837. DOI: `10.1016/j.cell.2007.05.009`. arXiv: `NIHMS150003`.

[4] S. Vinod Kumar and Philip A. Wigge. "H2A.Z-Containing Nucleosomes Mediate the Thermosensory Response in Arabidopsis". In: *Cell* 140.1 (2010), pp. 136–147. ISSN: 00928674. DOI: `10.1016/j.cell.2009.11.006`.

[5] T. Jenuwein. "Translating the Histone Code". In: *Science* 293.5532 (2001), pp. 1074–1080. DOI: `10.1126/science.1063127`. arXiv: `arXiv:1011.1669v3`. URL: `http://www.sciencemag.org/cgi/doi/10.1126/science.1063127`.

[6] `http://www.amsbio.com/images/featureareas/nucleosomes-and-histone-proteins/nucleosomes.jpg`.

[7] "Histone modification levels are predictive for gene expression". In: *Proceedings of the National Academy of Sciences* 107.7 (2010), pp. 2926–2931. DOI: `10.1073/pnas.0909344107`. URL: `http://www.pnas.org/cgi/doi/10.1073/pnas.0909344107`.

[8] Chao Cheng et al. "A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets". In: *Genome Biology* 12.2 (2011), R15. URL: `http://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-2-r15`.

[9] Xiaojiang Xu et al. "Application of machine learning methods to histone methylation ChIP-Seq data reveals H4R3me2 globally represses gene expression." In: *BMC bioinformatics* 11 (2010), p. 396.

[10] Ivan G Costa et al. "Predicting gene expression in T cell differentiation from histone modifications and transcription factor binding affinities by linear mixture models". In: *BMC Bioinformatics* 12.Suppl 1 (2011), S29. URL: `http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-S1-S29`.

[11] Xianjun Dong et al. "Modeling gene expression using chromatin features in various cellular contexts". In: *Genome Biology* 13.9 (2012), R53. URL: `http://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-9-r53`.

[12] Anirudh Natarajan et al. "Predicting cell-type - specific gene expression from regions of open chromatin the genome". In: *Genome Research* (2012), pp. 1711–1722. DOI: `10.1101/gr.135129.111`.

[13] Ritambhara Singh et al. "DeepChrome: Deep-learning for predicting gene expression from histone modifications". In: *Bioinformatics* 32.17 (2016), pp. i639–i648. ISSN: 14602059. DOI: `10.1093/bioinformatics/btw427`.

[14] Wouter Kundaje, Anshul Meuleman et al. "Integrative analysis of 111 reference human epigenomes". In: *Nature* 518.7539 (2015), pp. 317–330. ISSN: 0028-0836. DOI: `10.1038/nature14248`.

[15] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. “Rectifier Nonlinearities Improve Neural Network Acoustic Models”. In: *Proceedings of the 30 th International Conference on Machine Learning* 28 (2013), p. 6. URL: https://web.stanford.edu/{~}awni/papers/relu{_}hybrid{_}icml2013{_}final.pdf.

[16] Natalie de Souza. “The ENCODE project”. In: *Nature Methods* 9.11 (2012), pp. 1046–1046. ISSN: 1548-7091. DOI: 10.1038/nmeth.2238. URL: http://www.nature.com.ezp.lib.unimelb.edu.au/nmeth/journal/v9/n11/full/nmeth.2238.html{\%}5Cnhttp://www.nature.com.ezp.lib.unimelb.edu.au/nmeth/journal/v9/n11/pdf/nmeth.2238.pdf.

[17] Felipe Albrecht et al. “DeepBlue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets.” In: *Nucleic acids research* 44.i (2016), gkw211–. ISSN: 1362-4962. DOI: 10.1093/nar/gkw211. URL: http://nar.oxfordjournals.org/content/early/2016/04/15/nar.gkw211.full.

[18] *BLUEPRINT - A BLUEPRINT of Haematopoietic Epigenomes.* http://www.blueprint-epigenome.eu.

[19] *Canadian Epigenetics, Environment and Health Research Consortium (CEEHRC) initiative.* http://www.epigenomes.ca.

[20] *CREST program Development of Fundamental Technologies for Diagnosis and Therapy Based upon Epigenome Analysis (Disease Epigenome).* http://crest-ihec.jp/english/index.html.

[21] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[22] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.* Software available from tensorflow.org. 2015. URL: http://tensorflow.org/.

[23] J. M. Vaquerizas et al. “A census of human transcription factors: function, expression and evolution”. In: *Nat. Rev. Genet.* 10.4 (Apr. 2009), pp. 252–263.

[24] G. J. Narlikar, H. Y. Fan, and R. E. Kingston. “Cooperation between complexes that regulate chromatin structure and transcription”. In: *Cell* 108.4 (2002), pp. 475–487.

[25] L. Xu, C. K. Glass, and M. G. Rosenfeld. “Coactivator and corepressor complexes in nuclear receptor function”. In: *Curr. Opin. Genet. Dev.* 9.2 (1999), pp. 140–147.

[26] C. K. Osborne et al. “Estrogen receptor: current understanding of its activation and modulation”. In: *Clin. Cancer Res.* 7.12 Suppl (2001), 4338s–4342s; discussion 4411s–4412s.

[27] T. Pawson. “Signal transduction–a conserved pathway from the membrane to the nucleus”. In: *Dev. Genet.* 14.5 (1993), pp. 333–338.

[28] J. C. Bryne et al. “JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update”. In: *Nucleic Acids Research* 36.Database (2007), pp. D102–D106. DOI: 10.1093/nar/gkm955.

[29] J. Wang et al. “Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors”. In: *Genome Research* 22.9 (2012), pp. 1798–1812. DOI: 10.1101/gr.139105.112.

[30] S. G. Landt et al. “ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia”. In: *Genome Research* 22.9 (2012), pp. 1813–1831. DOI: 10.1101/gr.136184.111.

[31] B. Ren. “Genome-Wide Location and Function of DNA Binding Proteins”. In: *Science* 290.5500 (2000), pp. 2306–2309. DOI: 10.1126/science.290.5500.2306.

[32] A. Ozdemir et al. “High resolution mapping of Twist to DNA in Drosophila embryos: Efficient functional analysis and evolutionary conservation”. In: *Genome Research* 21.4 (2011), pp. 566–577. DOI: 10.1101/gr.104018.109.

[33] Qunhua Li et al. “Measuring reproducibility of high-throughput experiments”. In: *The Annals of Applied Statistics* 5.3 (2011), pp. 1752–1779. DOI: 10.1214/11-aoas466.

[34] Peter V Kharchenko, Michael Y Tolstorukov, and Peter J Park. “Design and analysis of ChIP-seq experiments for DNA-binding proteins”. In: *Nature Biotechnology* 26.12 (2008), pp. 1351–1359. DOI: 10.1038/nbt.1508.

[35] Joel Rozowsky et al. “PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls”. In: *Nature Biotechnology* 27.1 (2009), pp. 66–75. DOI: 10.1038/nbt.1518.

[36] Yong Zhang et al. “Model-based Analysis of ChIP-Seq (MACS)”. In: *Genome Biology* 9.9 (2008), R137. DOI: 10.1186/gb-2008-9-9-r137.

[37] B. C. Haynes et al. “Mapping functional transcription factor networks from gene expression data”. In: *Genome Res.* 23.8 (2013), pp. 1319–1328.

[38] J. J. Faith et al. “Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles”. In: *PLoS Biol.* 5.1 (2007), e8.

[39] G. STOLOVITZKY, D. MONROE, and A. CALIFANO. “Dialogue on Reverse-Engineering Assessment and Methods: The DREAM of High-Throughput Pathway Inference”. In: *Annals of the New York Academy of Sciences* 1115.1 (2007), pp. 1–22. DOI: 10.1196/annals.1407.021.

[40] Andrea Pinna, Nicola Soranzo, and Alberto de la Fuente. “From Knockouts to Networks: Establishing Direct Cause-Effect Relationships through Graph Analysis”. In: *PLoS ONE* 5.10 (2010). Ed. by Mark Isalan, e12912. DOI: 10.1371/journal.pone.0012912.

[41] S.-J. Dunn et al. “Defining an essential transcription factor program for naive pluripotency”. In: *Science* 344.6188 (2014), pp. 1156–1160. DOI: 10.1126/science.1248882.

[42] C. Angelini and V. Costa. “Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems”. In: *Front Cell Dev Biol* 2 (2014), p. 51.

[43] Z. Ouyang, Q. Zhou, and W. H. Wong. “ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells”. In: *Proceedings of the National Academy of Sciences* 106.51 (2009), pp. 21521–21526. DOI: 10.1073/pnas.0904863106.

[44] A. M. Tsankov et al. “Transcription factor binding dynamics during human ES cell differentiation”. In: *Nature* 518.7539 (2015), pp. 344–349.

[45] Denes Hnisz et al. “Super-Enhancers in the Control of Cell Identity and Disease”. In: *Cell* 155.4 (2013), pp. 934–947. DOI: 10.1016/j.cell.2013.09.053.

[46] Michael B. Stadler et al. “DNA-binding factors shape the mouse methylome at distal regulatory regions”. In: *Nature* (2011). DOI: 10.1038/nature10716.

[47] Matt Thomson et al. “Pluripotency Factors in Embryonic Stem Cells Regulate Differentiation into Germ Layers”. In: *Cell* 145.6 (2011), pp. 875–889. DOI: 10.1016/j.cell.2011.05.017.

[48] Joseph K. Pickrell et al. “False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions”. In: *Bioinformatics* 27.15 (2011), pp. 2144–2146. DOI: 10.1093/bioinformatics/btr354.

[49] R. Stark and G. D. Brown. “DiffBind: Differential Binding Analysis of ChIP-Seq Peak Data.” In: *Bioconductor* (2011). URL: http://bioconductor.org/packages/release/bioc/html/DiffBind.html.

[50] Anthony Mathelier et al. “JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles”. In: *Nucleic Acids Research* 42.D1 (2013), pp. D142–D147. DOI: 10.1093/nar/gkt997.

[51] T. L. Bailey et al. “MEME: discovering and analyzing DNA and protein sequence motifs”. In: *Nucleic Acids Research* 34.Web Server (2006), W369–W373. DOI: 10.1093/nar/gkl198.

[52] V. Matys. “TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes”. In: *Nucleic Acids Research* 34.90001 (2006), pp. D108–D110. DOI: 10.1093/nar/gkj143.

[53] K. Robasky and M. L. Bulyk. “UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions”. In: *Nucleic Acids Research* 39.Database (2010), pp. D124–D128. DOI: 10.1093/nar/gkq992.

[54] P. Kheradpour et al. “Reliable prediction of regulator targets using 12 Drosophila genomes”. In: *Genome Research* 17.12 (2007), pp. 1919–1931. DOI: 10.1101/gr.7090407.

[55] Gong-Hong Wei et al. “Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo”. In: *The EMBO Journal* 29.13 (2010), pp. 2147–2160. DOI: 10.1038/emboj.2010.106.

[56] Thomas Bonfert et al. “ContextMap 2: fast and accurate context-based RNA-seq mapping”. In: *BMC Bioinformatics* 16.1 (2015). DOI: 10.1186/s12859-015-0557-5.

[57] Daehwan Kim, Ben Langmead, and Steven L Salzberg. “HISAT: a fast spliced aligner with low memory requirements”. In: *Nature Methods* 12.4 (2015), pp. 357–360. DOI: 10.1038/nmeth.3317.

[58] Alexander Dobin et al. “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29.1 (2012), pp. 15–21. DOI: 10.1093/bioinformatics/bts635.

[59] Daehwan Kim et al. “TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions”. In: *Genome Biology* 14.4 (2013), R36. DOI: 10.1186/gb-2013-14-4-r36. URL: https://doi.org/10.1186/gb-2013-14-4-r36.

[60] Michael I Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biology* 15.12 (2014). DOI: 10.1186/s13059-014-0550-8.

[61] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics* 26.1 (2009), pp. 139–140. DOI: 10.1093/bioinformatics/btp616.

[62] M. E. Ritchie et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic Acids Research* 43.7 (2015), e47–e47. DOI: 10.1093/nar/gkv007.

[63] M. D. Bethesda. “National Library of Medicine (US), National Center for Biotechnology Information. Gene ID: 5460, Homo sapiens POU class 5 homeobox 1 (POU5F1)”. In: *Internet* (1988). [Online; accessed 19-July-2017].

[64] Florian Erhard and Ralf Zimmer. “Count ratio model reveals bias affecting NGS fold changes”. In: *Nucleic Acids Research* (2015), gkv696. DOI: 10.1093/nar/gkv696.

[65] Stanford University. *ENCODE: Encyclopedia of DNA Elements.* https://www.encodeproject.org/. Accessed: 2017-07-25.

[66] Robert E. Thurman et al. "The accessible chromatin landscape of the human genome". In: *Nature* 489.7414 (2012), pp. 75–82. ISSN: 0028-0836. DOI: 10.1038/nature11232.

[67] Peter J. Park. "ChIP-seq: advantages and challenges of a maturing technology". In: *Nature reviews. Genetics* 10.10 (2009), pp. 669–680. ISSN: 1471-0064. DOI: 10.1038/nrg2641.

[68] Anirudh Natarajan et al. "Predicting cell-type-specific gene expression from regions of open chromatin". In: *Genome research* 22.9 (2012), pp. 1711–1722. ISSN: 1088-9051. DOI: 10.1101/gr.135129.111.

[69] Jason D. Buenrostro et al. "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". In: 10.12 (2013), pp. 1213–1218. ISSN: 1548-7091. DOI: 10.1038/nmeth.2688.

[70] Maria Tsompana and Michael J. Buck. "Chromatin accessibility: a window into the genome". In: *Epigenetics & chromatin* 7.1 (2014), p. 33. DOI: 10.1186/1756-8935-7-33.

[71] P. G. Giresi et al. "FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin". In: *Genome research* 17.6 (2007), pp. 877–885. ISSN: 1088-9051. DOI: 10.1101/gr.5533506.

[72] Alan P. Boyle et al. "High-Resolution Mapping and Characterization of Open Chromatin across the Genome". In: *Cell* 132.2 (2008), pp. 311–322. ISSN: 00928674. DOI: 10.1016/j.cell.2007.12.014.

[73] Yong Zhang et al. "Model-based analysis of ChIP-Seq (MACS)". In: *Genome biology* 9.9 (2008), R137. DOI: 10.1186/gb-2008-9-9-r137.

[74] Kristin Brogaard et al. "A map of nucleosome positions in yeast at base-pair resolution". In: *Nature* 486.7404 (2012), pp. 496–501. ISSN: 0028-0836. DOI: 10.1038/nature11142.

[75] Alicia N. Schep et al. "Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions". In: *Genome research* 25.11 (2015), pp. 1757–1770. ISSN: 1088-9051. DOI: 10.1101/gr.192294.115.

[76] Albin Sandelin et al. "JASPAR: an open-access database for eukaryotic transcription factor binding profiles". In: *Nucleic acids research* 32.Database issue (2004), pp. D91–4. ISSN: 1362-4962. DOI: 10.1093/nar/gkh012.

[77] Kenzie D. MacIsaac et al. "An improved map of conserved regulatory sites for Saccharomyces cerevisiae". In: *BMC bioinformatics* 7 (2006), p. 113. ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-113.

[78] Roger Pique-Regi et al. "Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data". In: *Genome research* 21.3 (2011), pp. 447–455. ISSN: 1088-9051. DOI: 10.1101/gr.112623.110.

[79] Roja Babazadeh et al. "The yeast osmostress response is carbon source dependent". In: *Scientific reports* 7.1 (2017), p. 990. ISSN: 2045-2322. DOI: 10.1038/s41598-017-01141-4.

[80] Thomas Bonfert et al. "ContextMap 2: fast and accurate context-based RNA-seq mapping". In: *BMC bioinformatics* 16 (2015), p. 122. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0557-5.

[81] Daehwan Kim et al. "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions". In: *Genome biology* 14.4 (2013), R36. DOI: 10.1186/gb-2013-14-4-r36.

[82] Alexander Dobin et al. “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics (Oxford, England)* 29.1 (2013), pp. 15–21. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts635.

[83] Daehwan Kim, Ben Langmead, and Steven L. Salzberg. “HISAT: a fast spliced aligner with low memory requirements”. In: *Nature methods* 12.4 (2015), pp. 357–360. ISSN: 1548-7105. DOI: 10.1038/nmeth.3317.

[84] Li Ni et al. “Dynamic and complex transcription factor binding during an inducible response in yeast”. In: *Genes & development* 23.11 (2009), pp. 1351–1363. ISSN: 1549-5477. DOI: 10.1101/gad.1781909.

[85] Ben Langmead and Steven L. Salzberg. “Fast gapped-read alignment with Bowtie 2”. In: *Nature methods* 9.4 (2012), pp. 357–359. ISSN: 1548-7105. DOI: 10.1038/nmeth.1923.

[86] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics (Oxford, England)* 26.1 (2010), pp. 139–140. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp616.

[87] Simon Anders and Wolfgang Huber. “Differential expression analysis for sequence count data”. In: *Genome biology* 11.10 (2010), R106. DOI: 10.1186/gb-2010-11-10-r106.

[88] Matthew E. Ritchie et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic acids research* 43.7 (2015), e47. ISSN: 1362-4962. DOI: 10.1093/nar/gkv007.

[89] Gabriel Cuellar-Partida et al. “Epigenetic priors for identifying active transcription factor binding sites”. In: *Bioinformatics (Oxford, England)* 28.1 (2012), pp. 56–62. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btr614.

[90] Shobhit Gupta et al. “Quantifying similarity between motifs”. In: *Genome biology* 8.2 (2007), R24. DOI: 10.1186/gb-2007-8-2-r24.

[91] L M Gallego-Paez et al. “Alternative splicing: the pledge, the turn, and the prestige : The key role of alternative splicing in human biological systems.” In: *Human genetics* (2017). ISSN: 1432-1203. DOI: 10.1007/s00439-017-1790-y. URL: http://link.springer.com/10.1007/s00439-017-1790-yhttp://www.ncbi.nlm.nih.gov/pubmed/28374191.

[92] Suzanne Clancy. “RNA Splicing: Introns, Exons and Spliceosome”. In: *Nature Education* (2008).

[93] Alberto R. Kornblihtt et al. “Alternative splicing: a pivotal step between eukaryotic transcription and translation”. In: *Nature Reviews Molecular Cell Biology* 14.5 (2013), pp. 306–306. ISSN: 1471-0072. DOI: 10.1038/nrm3560. URL: http://www.nature.com/doifinder/10.1038/nrm3560.

[94] R E Breitbart, A Andreadis, and B Nadal-Ginard. “Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes.” In: *Annual review of biochemistry* 56.1 (1987), pp. 467–95. ISSN: 0066-4154. DOI: 10.1146/annurev.bi.56.070187.002343. URL: http://www.ncbi.nlm.nih.gov/pubmed/3304142http://www.annualreviews.org/doi/10.1146/annurev.bi.56.070187.002343http://www.annualreviews.org/doi/abs/10.1146/annurev.bi.56.070187.002343.

[95] Douglas L. Black. “Mechanisms of Alternative Pre-Messenger RNA Splicing”. In: *Annual Review of Biochemistry* 72.1 (2003), pp. 291–336. ISSN: 0066-4154. DOI: 10.1146/annurev.biochem.72.121801.161720. URL: http://www.ncbi.nlm.nih.

gov/pubmed/12626338http://www.annualreviews.org/doi/10.1146/annurev.biochem.72.121801.161720.

[96] A. Gregory Matera and Zefeng Wang. "A day in the life of the spliceosome". In: *Nature Reviews Molecular Cell Biology* 15.2 (2014), pp. 108–121. ISSN: 1471-0072. DOI: 10.1038/nrm3742. URL: http://www.ncbi.nlm.nih.gov/pubmed/24452469http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4060434http://www.nature.com/doifinder/10.1038/nrm3742.

[97] Arianne J. Matlin, Francis Clark, and Christopher W. J. Smith. "Understanding alternative splicing: towards a cellular code". In: *Nature Reviews Molecular Cell Biology* 6.5 (2005), pp. 386–398. ISSN: 1471-0072. DOI: 10.1038/nrm1645. URL: http://www.ncbi.nlm.nih.gov/pubmed/15956978http://www.nature.com/doifinder/10.1038/nrm1645.

[98] Barmak Modrek and Christopher Lee. "A genomic view of alternative splicing". In: *Nature Genetics* 30.1 (2002), pp. 13–19. ISSN: 10614036. DOI: 10.1038/ng0102-13. URL: http://www.nature.com/doifinder/10.1038/ng0102-13.

[99] Hadas Keren, Galit Lev-Maor, and Gil Ast. "Alternative splicing and evolution: diversification, exon definition and function". In: *Nature Reviews Genetics* 11.5 (2010), pp. 345–355. ISSN: 1471-0056. DOI: 10.1038/nrg2776. URL: http://www.ncbi.nlm.nih.gov/pubmed/20376054http://www.nature.com/doifinder/10.1038/nrg2776.

[100] J. Merkin et al. "Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues". In: *Science* 338.6114 (2012), pp. 1593–1599. ISSN: 0036-8075. DOI: 10.1126/science.1228186. arXiv: NIHMS150003. URL: http://www.ncbi.nlm.nih.gov/pubmed/23258891http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3568499http://www.sciencemag.org/cgi/doi/10.1126/science.1228186.

[101] Wolfgang Huber et al. "Orchestrating high-throughput genomic analysis with Bioconductor". In: *Nature Methods* 12.2 (2015), pp. 115–121. ISSN: 1548-7091. DOI: 10.1038/nmeth.3252. arXiv: 9809069v1 [arXiv:gr-qc]. URL: http://www.nature.com/doifinder/10.1038/nmeth.3252http://www.ncbi.nlm.nih.gov/pubmed/25633503http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4509590.

[102] Simon Anders, Alejandro Reyes, and Wolfgang Huber. "Detecting differential usage of exons from RNA-seq data". In: *Genome Research* 22.10 (2012), pp. 2008–2017. ISSN: 10889051. DOI: 10.1101/gr.133744.111. arXiv: arXiv:1011.1669v3. URL: http://www.ncbi.nlm.nih.gov/pubmed/22722343http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3460195http://genome.cshlp.org/cgi/doi/10.1101/gr.133744.111.

[103] Hugues Richard et al. "Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments". In: *Nucleic Acids Research* 38.10 (2010), e112. ISSN: 03051048. DOI: 10.1093/nar/gkq041. URL: http://www.ncbi.nlm.nih.gov/pubmed/20150413http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2879520.

[104] Cole Trapnell et al. "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation". In: *Nature Biotechnology* 28.5 (2010), pp. 511–515. ISSN: 1087-0156. DOI: 10.1038/nbt.1621. arXiv: 171. URL: http://www.ncbi.nlm.nih.gov/pubmed/20436464http://

www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3146043http://www.nature.com/doifinder/10.1038/nbt.1621.

[105] Yarden Katz et al. "Analysis and design of RNA sequencing experiments for identifying isoform regulation". In: *Nature Methods* 7.12 (2010), pp. 1009–1015. ISSN: 1548-7091. DOI: 10.1038/nmeth.1528. arXiv: 9605103 [cs]. URL: http://www.nature.com/doifinder/10.1038/nmeth.1528http://www.ncbi.nlm.nih.gov/pubmed/21057496http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3037023.

[106] Malachi Griffith et al. "Alternative expression analysis by RNA sequencing". In: *Nature Methods* 7.10 (2010), pp. 843–847. ISSN: 1548-7091. DOI: 10.1038/nmeth.1503. URL: http://www.ncbi.nlm.nih.gov/pubmed/20835245http://www.nature.com/doifinder/10.1038/nmeth.1503.

[107] Shihao Shen et al. "rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data". In: *Proceedings of the National Academy of Sciences* 111.51 (2014), E5593–E5601. ISSN: 0027-8424. DOI: 10.1073/pnas.1419161111. arXiv: arXiv:1408.1149. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1419161111http://www.ncbi.nlm.nih.gov/pubmed/25480548http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4280593.

[108] Cole Trapnell et al. "Differential analysis of gene regulation at transcript resolution with RNA-seq." In: *Nature biotechnology* 31.1 (2013), pp. 46–53. ISSN: 1546-1696. DOI: 10.1038/nbt.2450. arXiv: NIHMS150003. URL: http://www.nature.com/doifinder/10.1038/nbt.2450http://www.ncbi.nlm.nih.gov/pubmed/23222703{\%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3869392.

[109] Dorothea Emig et al. "AltAnalyze and DomainGraph: Analyzing and visualizing exon expression data". In: *Nucleic Acids Research* 38.SUPPL. 2 (2010), W755–62. ISSN: 03051048. DOI: 10.1093/nar/gkq405. URL: http://www.ncbi.nlm.nih.gov/pubmed/20513647http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2896198.

[110] Jie Zhang and Zhi Wei. "An empirical Bayes change-point model for identifying 3' and 5' alternative splicing by next-generation RNA sequencing". In: *Bioinformatics* 32.12 (2016), pp. 1823–1831. ISSN: 14602059. DOI: 10.1093/bioinformatics/btw060. URL: http://www.ncbi.nlm.nih.gov/pubmed/26873932https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw060.

[111] André Kahles et al. "SplAdder: Identification, quantification and testing of alternative splicing events from RNA-Seq data". In: *Bioinformatics* 32.12 (2016), pp. 1840–1847. ISSN: 14602059. DOI: 10.1093/bioinformatics/btw076. URL: http://www.ncbi.nlm.nih.gov/pubmed/26873928http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4908322https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw076.

[112] Scott Norton, Jorge Vaquero-Garcia, and Yoseph Barash. "Outlier detection for improved differential splicing quantification from RNA-Seq experiments with replicates". In: *bioRxiv* (2017), pp. 1–15. DOI: 10.1101/104059. URL: https://www.biorxiv.org/content/early/2017/05/11/104059.

[113] Sylvain Foissac and Michael Sammeth. "ASTALAVISTA: Dynamic and flexible analysis of alternative splicing events in custom gene datasets". In: *Nucleic Acids Research* 35.SUPPL.2 (2007), W297–W299. ISSN: 03051048. DOI: 10.1093/nar/gkm311. URL:

http://www.ncbi.nlm.nih.gov/pubmed/17485470http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1933205https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkm311.

[114] Jennifer Harrow et al. "GENCODE: producing a reference annotation for ENCODE". In: *Genome Biology* 7.Suppl 1 (2006), S4. ISSN: 14656906. DOI: 10.1186/gb-2006-7-s1-s4. URL: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2006-7-s1-s4.

[115] Andrew Yates et al. "Ensembl 2016." In: *Nucleic acids research* 44.D1 (2016), pp. D710–6. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1157. arXiv: arXiv:1011.1669v3. URL: http://www.ncbi.nlm.nih.gov/pubmed/26687719http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4702834http://www.ncbi.nlm.nih.gov/pubmed/26687719{\%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4702834.

[116] Hsien-Da Huang et al. "ProSplicer: a database of putative alternative splicing information derived from protein, mRNA and expressed sequence tag sequence data." In: *Genome biology* 4.4 (2003), R29. ISSN: 1465-6914. DOI: 10.1186/gb-2003-4-4-r29. URL: http://www.ncbi.nlm.nih.gov/pubmed/12702210http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC154580.

[117] *Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information.* 2004. URL: https://www.ncbi.nlm.nih.gov/gene/ (visited on 06/04/2017).

[118] Simon Penel et al. "Databases of homologous gene families for comparative genomics". In: *BMC Bioinformatics* 10.Suppl 6 (2009), S3. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-S6-S3. URL: http://www.biomedcentral.com/1471-2105/10/S6/S3.

[119] Jakub O Westholm and Eric C Lai. *Mirtrons: MicroRNA biogenesis via splicing.* 2011. DOI: 10.1016/j.biochi.2011.06.017. arXiv: NIHMS150003. URL: http://www.ncbi.nlm.nih.gov/pubmed/21712066http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3185189.

[120] David Brawand et al. "The evolution of gene expression levels in mammalian organs". In: *Nature* 478.7369 (2011), pp. 343–348. ISSN: 0028-0836. DOI: 10.1038/nature10532. URL: http://www.nature.com/doifinder/10.1038/nature10532.

[121] Heng Li et al. "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16 (2009), pp. 2078–2079. ISSN: 13674803. DOI: 10.1093/bioinformatics/btp352. arXiv: 1006.1266v2. URL: http://www.ncbi.nlm.nih.gov/pubmed/19505943http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2723002.

[122] Aaron R Quinlan and Ira M Hall. "BEDTools: A flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6 (2010), pp. 841–842. ISSN: 13674803. DOI: 10.1093/bioinformatics/btq033. URL: http://www.ncbi.nlm.nih.gov/pubmed/20110278http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2832824.

[123] Yasunobu Okamura et al. "COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems". In: *Nucleic acids research* 43.D1 (2014), pp. D82–D86.

[124] Socorro Gama-Castro et al. "RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation". In: *Nucleic acids research* 36.suppl_1 (2008), pp. D120–D124.

[125] Zhi-Ping Liu et al. "RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse". In: *Database* 2015 (2015), bav095.

[126] Damian Szklarczyk et al. "STRING v10: protein–protein interaction networks, integrated over the tree of life". In: *Nucleic acids research* 43.D1 (2014), pp. D447–D452.

[127] Jianfei Hu et al. "PhosphoNetworks: a database for human phosphorylation networks". In: *Bioinformatics* 30.1 (2013), pp. 141–142.

[128] Minoru Kanehisa and Susumu Goto. "KEGG: kyoto encyclopedia of genes and genomes". In: *Nucleic acids research* 28.1 (2000), pp. 27–30.

[129] Jing Yang et al. "DNetDB: The human disease network database based on dysfunctional regulation mechanism". In: *BMC systems biology* 10.1 (2016), p. 36.

[130] Rainer Breitling, Anna Amtmann, and Pawel Herzyk. "Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments". In: *BMC bioinformatics* 5.1 (2004), p. 1.

[131] Aravind Subramanian et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15545–15550.

[132] Ludwig Geistlinger et al. "From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems". In: *Bioinformatics* 27.13 (2011), pp. i366–i373.

[133] Adi Laurentiu Tarca et al. "A novel signaling pathway impact analysis". In: *Bioinformatics* 25.1 (2009), pp. 75–82. ISSN: 13674803. DOI: 10.1093/bioinformatics/btn577.

[134] Zhaoyuan Fang, Weidong Tian, and Hongbin Ji. "A network-based gene-weighting approach for pathway analysis". In: *Cell Research* 22.3 (2012), pp. 565–580. URL: http://dx.doi.org/10.1038/cr.2011.149.

[135] Xinran Dong et al. "LEGO : a novel method for gene set over-representation analysis by incorporating network-based gene weights". In: *Nature Publishing Group* (2016), pp. 1–17. URL: http://dx.doi.org/10.1038/srep18871.

[136] Piotr J Balwierz et al. "ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs". In: *Genome research* 24.5 (2014), pp. 869–884.

[137] Yue Li, Minggao Liang, and Zhaolei Zhang. "Regression analysis of combined gene expression regulation in acute myeloid leukemia". In: *PLoS computational biology* 10.10 (2014), e1003908.

[138] Nurcan Tuncbag et al. "SteinerNet: a web server for integrating ?omic?data to discover hidden components of response pathways". In: *Nucleic acids research* 40.W1 (2012), W505–W509.

[139] Ivana Ljubić et al. "An algorithmic framework for the exact solution of the prize-collecting Steiner tree problem". In: *Mathematical programming* 105.2 (2006), pp. 427–449.

[140] Evi Berchtold, Gergely Csaba, and Ralf Zimmer. "RelExplain?integrating data and networks to explain biological processes". In: *Bioinformatics* 33.12 (2017), pp. 1837–1844. DOI: 10.1093/bioinformatics/btx060. eprint: /oup/backfile/content_public/journal/bioinformatics/33/12/10.1093_bioinformatics_btx060/1/btx060.pdf. URL: +http://dx.doi.org/10.1093/bioinformatics/btx060.

[141] N Dimitrova et al. "InFlo: a novel systems biology framework identifies cAMP-CREB1 axis as a key modulator of platinum resistance in ovarian cancer". In: *Oncogene* 36.17 (2017), p. 2472.

[142] National Cancer Institute. *The Cancer Genome Atlas.* URL: http://cancergenome.nih.gov/.

[143] Gene Ontology Consortium et al. "The Gene Ontology (GO) database and informatics resource". In: *Nucleic acids research* 32.suppl 1 (2004), pp. D258–D261.

[144] Dexter Pratt et al. "NDEx, the network data exchange". In: *Cell systems* 1.4 (2015), pp. 302–305.

[145] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1 (2010), pp. 139–140. DOI: 10.1093/bioinformatics/btp616. eprint: /oup/backfile/content_public/journal/bioinformatics/26/1/10.1093/bioinformatics/btp616/2/btp616.pdf. URL: +http://dx.doi.org/10.1093/bioinformatics/btp616.

[146] Ryota Suzuki and Hidetoshi Shimodaira. "Pvclust: an R package for assessing the uncertainty in hierarchical clustering". In: *Bioinformatics* 22.12 (2006), pp. 1540–1542. DOI: 10.1093/bioinformatics/btl117. eprint: /oup/backfile/content_public/journal/bioinformatics/22/12/10.1093/bioinformatics/btl117/2/btl117.pdf. URL: +http://dx.doi.org/10.1093/bioinformatics/btl117.

[147] YZ Chen et al. "PPAR signaling pathway may be an important predictor of breast cancer response to neoadjuvant chemotherapy". In: *Cancer chemotherapy and pharmacology* 70.5 (2012), pp. 637–644.

[148] Max Franz et al. "Cytoscape. js: a graph theory library for visualisation and analysis". In: *Bioinformatics* 32.2 (2015), pp. 309–311.

[149] X. Guo et al. "Advances in long noncoding RNAs: identification, structure prediction and function annotation". In: *Briefings in Functional Genomics* 15.1 (2016), pp. 38–46. DOI: 10.1093/bfgp/elv022.

[150] J. T. Y. Kung, D. Colognori, and J. T. Lee. "Long Noncoding RNAs: Past, Present, and Future". In: *Genetics* 193.3 (2013), pp. 651–669. DOI: 10.1534/genetics.112.146704.

[151] A. Kapusta and C. Feschotte. "Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications". In: *Trends in Genetics* 30.10 (2014), pp. 439–452. DOI: 10.1016/j.tig.2014.08.004.

[152] K. Kashi et al. "Discovery and functional analysis of lncRNAs: Methodologies to investigate an uncharacterized transcriptome". In: *Biochimica et Biophysica Acta* 1895.1 (2016), pp. 3–15. DOI: 10.1016/j.bbagrm.2015.10.010.

[153] M. Huarte. "The emerging role of lncRNAs in cancer". In: *Nature Medicine* 21.11 (2015), pp. 1253–1261. DOI: 10.1038/nm.3981.

[154] G. St Laurent, C. Wahlestedt, and P. Kapranov. "The Landscape of long non-coding RNA classification". In: *Trends in Genetics* 31.5 (2015), pp. 239–251. DOI: 10.1016/j.tig.2015.03.007.

[155] L. Ma, V. B. Bajic, and Z. Zhang. "On the classification of long non-coding RNAs". In: *RNA Biology* 10.6 (2013), pp. 924–933. DOI: 10.4161/rna.24604.

[156] L. Dong and L. Hui. "HOTAIR Promotes Proliferation, Migration, and Invasion of Ovarian Cancer SKOV3 Cells Through Regulating PIK3R3". In: *Medical Science Monitor* 31.22 (2016), pp. 325–331. DOI: 10.12659/MSM.894913.

[157] S. Huang et al. "The long non-coding RNA CCAT2 is up-regulated in ovarian cancer and associated with poor prognosis". In: *Diagnostic Pathology* 11.49 (2016). DOI: 10.1186/s13000-016-0499-x.

[158] P. Hu et al. “NBAT1 suppresses breast cancer metastasis by regulating DKK1 via PRC2”. In: *Oncotarget* 6.32 (2015), pp. 32410–32425. DOI: 10.18632/oncotarget.5609.

[159] C. Yan et al. “Long noncoding RNA NBAT-1 suppresses tumorigenesis and predicts favorable prognosis in ovarian cancer”. In: *Onco Targets and Therapy* 10 (2017), pp. 1993–2002. DOI: 10.2147/OTT.S124645.

[160] L. Meng et al. “Towards a therapy for Angelman syndrome by targeting a long non-coding RNA”. In: *Nature* 518.7539 (2015), pp. 409–412. DOI: 10.1038/nature13975.

[161] S. Li et al. “Exploring functions of long noncoding RNAs across multiple cancers through co-expression network”. In: *Scientific Reports* 7.1 (2017). DOI: 10.1038/s41598-017-00856-8.

[162] G. Guo et al. “Whole-genome and whole-exome sequencing of bladder cancer identifies frequent alterations in genes involved in sister chromatid cohesion and segregation”. In: *Nature Genetics* 45.12 (2013), pp. 1459–1463. DOI: 10.1038/ng.2798.

[163] C. Trapnell, L. Pachter, and S. L. Salzberg. “TopHat: discovering splice junctions with RNA-Seq”. In: *Bioinformatics* 25.9 (2009), pp. 1105–1111. DOI: 10.1093/bioinformatics/btp120.

[164] J. Harrow et al. “GENCODE: the reference human genome annotation for The ENCODE Project”. In: *Genome Research* 22 (2012), pp. 1760–1774. DOI: 10.1101/gr.135350.111.

[165] C. Trapnell et al. “Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks”. In: *Nature Protocols* 7.3 (2012), pp. 562–578. DOI: 10.1038/nprot.2012.016.

[166] M. I. Love, W. Huber, and S. Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biology* 15.550 (2014). DOI: 10.1186/s13059-014-0550-8.

[167] S. Anders, P. T. Pyl, and W. Huber. “HTSeq - a Python framework to work with high-throughput sequencing data”. In: *Bioinformatics* 31.2 (2015), 166?169. DOI: 10.1093/bioinformatics/btu638.

[168] B. Zhang and S. Horvarth. “A general framework for weighted gene co-expression network analysis”. In: *Statistical Applications in Genetics and Molecular Biology* 4.1 (2005). DOI: 10.2202/1544-6115.1128.

[169] P. Langfelder and S. Horvarth. “WGCNA: an R package for weighted correlation network analysis”. In: *BMC Bioinformatics* 9.559 (2008). DOI: 10.1186/1471-2105-9-559.

[170] P. Langfelder, B. Zhang, and S. Horvath. “Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R”. In: *Bioinformatics* 24.5 (2008), pp. 719–720. DOI: 10.1093/bioinformatics/btm563.

[171] H. Ma et al. “HSPA12B: a novel facilitator of lung tumor growth”. In: *Oncotarget* 6.12 (2015), pp. 9924–9936. DOI: 10.18632/oncotarget.3533.

[172] W. Yue et al. “Frequent inactivation of RAMP2, EFEMP1 and Dutt1 in lung cancer by promoter hypermethylation”. In: *Clinical Cancer Research* 13.15 Pt. 1 (2007), pp. 4336–4344. DOI: 10.1158/1078-0432.CCR-07-0015.

[173] V. Tripathi et al. “The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation”. In: *Molecular Cell* 39.6 (2010), pp. 925–938. DOI: 10.1016/j.molcel.2010.08.011.

[174] D. Liu et al. “Knockdown of long non-coding RNA MALAT1 inhibits growth and motility of human hepatoma cells via modulation of miR-195”. In: *Journal of Cellular Biochemistry* (2017), pp. 925–938. DOI: 10.1002/jcb.26297.

[175] J. Yao et al. “A new tumor suppressor LncRNA ADAMTS9-AS2 is regulated by DNMT1 and inhibits migration of glioma cells”. In: *Tumor Biology* 35.8 (2014), pp. 7935–7944. DOI: 10.1007/s13277-014-1949-2.

[176] D. P. Bartel. “MicroRNAs: genomics, biogenesis, mechanism, and function”. In: *Cell* 116.2 (2004), pp. 281–297. DOI: 10.1016/s0092-8674(04)00045-5.

[177] Victor Ambros. “The functions of animal microRNAs”. In: *Nature* 431.7006 (2004), pp. 350–355. DOI: 10.1038/nature02871. URL: http://dx.doi.org/10.1038/nature02871.

[178] Thalia A Farazi et al. “miRNAs in human cancer”. In: *The Journal of Pathology* 223.2 (2011), pp. 102–115. DOI: 10.1002/path.2806. URL: http://dx.doi.org/10.1002/path.2806.

[179] Peter T. Nelson, Wang-Xia Wang, and Bernard W. Rajeev. “MicroRNAs (miRNAs) in Neurodegenerative Diseases”. In: *Brain Pathology* 18.1 (2008), pp. 130–138. DOI: 10.1111/j.1750-3639.2007.00120.x. URL: http://dx.doi.org/10.1111/j.1750-3639.2007.00120.x.

[180] Donato Santovito, Virginia Egea, and Christian Weber. “Small but smart: MicroRNAs orchestrate atherosclerosis development and progression”. In: *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids* 1861.12 (2016), pp. 2075–2086. ISSN: 18792618. DOI: 10.1016/j.bbalip.2015.12.013. URL: http://dx.doi.org/10.1016/j.bbalip.2015.12.013.

[181] Yoontae Lee et al. “MicroRNA genes are transcribed by RNA polymerase II”. In: *The EMBO Journal* 23.20 (2004), pp. 4051–4060. DOI: 10.1038/sj.emboj.7600385. URL: http://dx.doi.org/10.1038/sj.emboj.7600385.

[182] Xuefeng Zhou et al. “Characterization and Identification of MicroRNA Core Promoters in Four Model Species”. In: *PLoS Computational Biology* 3.3 (2007), e37. DOI: 10.1371/journal.pcbi.0030037. URL: http://dx.doi.org/10.1371/journal.pcbi.0030037.

[183] Richard I. Gregory et al. “The Microprocessor complex mediates the genesis of microRNAs”. In: *Nature* 432.7014 (2004), pp. 235–240. DOI: 10.1038/nature03120. URL: http://dx.doi.org/10.1038/nature03120.

[184] Ahmet M. Denli et al. “Processing of primary microRNAs by the Microprocessor complex”. In: *Nature* 432.7014 (2004), pp. 231–235. DOI: 10.1038/nature03049. URL: http://dx.doi.org/10.1038/nature03049.

[185] Y. Lee et al. “The nuclear RNase III Drosha initiates microRNA processing”. In: *Nature* 425.6956 (2003), pp. 415–419.

[186] Markus Landthaler, Abdullah Yalcin, and Thomas Tuschl. “The Human DiGeorge Syndrome Critical Region Gene 8 and Its D. melanogaster Homolog Are Required for miRNA Biogenesis”. In: *Current Biology* 14.23 (2004), pp. 2162–2167. DOI: 10.1016/j.cub.2004.11.001. URL: http://dx.doi.org/10.1016/j.cub.2004.11.001.

[187] E. Lund. “Nuclear Export of MicroRNA Precursors”. In: *Science* 303.5654 (2004), pp. 95–98. DOI: 10.1126/science.1090599. URL: http://dx.doi.org/10.1126/science.1090599.

[188] M. T. Bohnsack. “Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs”. In: *RNA* 10.2 (2004), pp. 185–191. DOI: 10.1261/rna.5167604. URL: http://dx.doi.org/10.1261/rna.5167604.

[189] Emily Bernstein et al. In: *Nature* 409.6818 (2001), pp. 363–366. DOI: 10.1038/35053110. URL: http://dx.doi.org/10.1038/35053110.

[190] Thimmaiah P. Chendrimada et al. "TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing". In: *Nature* 436.7051 (2005), pp. 740–744. DOI: 10.1038/nature03868. URL: http://dx.doi.org/10.1038/nature03868.

[191] Jong-Eun Park et al. "Dicer recognizes the 5′ end of RNA for efficient and accurate processing". In: *Nature* 475.7355 (2011), pp. 201–205. DOI: 10.1038/nature10198. URL: http://dx.doi.org/10.1038/nature10198.

[192] Julia Winter et al. "Many roads to maturity: microRNA biogenesis pathways and their regulation". In: *Nat Cell Biol* 11.3 (2009), pp. 228–234. DOI: 10.1038/ncb0309-228. URL: http://dx.doi.org/10.1038/ncb0309-228.

[193] Jr-Shiuan Yang and Eric C. Lai. "Dicer-independent, Ago2-mediated microRNA biogenesis in vertebrates". In: *Cell Cycle* 9.22 (2010), pp. 4455–4460. DOI: 10.4161/cc.9.22.13958. URL: http://dx.doi.org/10.4161/cc.9.22.13958.

[194] Katsutomo Okamura et al. "The Mirtron Pathway Generates microRNA-Class Regulatory RNAs in Drosophila". In: *Cell* 130.1 (2007), pp. 89–100. DOI: 10.1016/j.cell.2007.06.028. URL: http://dx.doi.org/10.1016/j.cell.2007.06.028.

[195] Eugene Berezikov et al. "Mammalian Mirtron Genes". In: *Molecular Cell* 28.2 (2007), pp. 328–336. DOI: 10.1016/j.molcel.2007.09.028. URL: http://dx.doi.org/10.1016/j.molcel.2007.09.028.

[196] Daniel Christian Ellwanger. "Computational modeling of miRNA-mediated gene regulation in consideration of miRNP binding information from AGO-bound CLIP-Seq data analysis". PhD thesis. Technische Universität München, 2015. URL: https://mediatum.ub.tum.de/doc/1244082/1244082.pdf.

[197] Richard I. Gregory et al. "Human RISC Couples MicroRNA Biogenesis and Posttranscriptional Gene Silencing". In: *Cell* 123.4 (2005), pp. 631–640. DOI: 10.1016/j.cell.2005.10.022. URL: http://dx.doi.org/10.1016/j.cell.2005.10.022.

[198] Benjamin Czech et al. "Hierarchical Rules for Argonaute Loading in Drosophila". In: *Molecular Cell* 36.3 (2009), pp. 445–456. DOI: 10.1016/j.molcel.2009.09.028. URL: http://dx.doi.org/10.1016/j.molcel.2009.09.028.

[199] Dianne S. Schwarz et al. "Asymmetry in the Assembly of the RNAi Enzyme Complex". In: *Cell* 115.2 (2003), pp. 199–208. DOI: 10.1016/s0092-8674(03)00759-1. URL: http://dx.doi.org/10.1016/S0092-8674(03)00759-1.

[200] Gyorgy Hutvagner and Martin J. Simard. "Argonaute proteins: key players in RNA silencing". In: *Nature Reviews Molecular Cell Biology* 9.1 (2008), pp. 22–32. DOI: 10.1038/nrm2321. URL: http://dx.doi.org/10.1038/nrm2321.

[201] Benjamin P. Lewis et al. "Prediction of Mammalian MicroRNA Targets". In: *Cell* 115.7 (2003), pp. 787–798. DOI: 10.1016/s0092-8674(03)01018-3. URL: http://dx.doi.org/10.1016/S0092-8674(03)01018-3.

[202] D. C. Ellwanger et al. "The sufficient minimal set of miRNA seed types". In: *Bioinformatics* 27.10 (2011), pp. 1346–1350. DOI: 10.1093/bioinformatics/btr149. URL: http://dx.doi.org/10.1093/bioinformatics/btr149.

[203] A van den Berg, J Mols, and J Han. "RISC-target interaction: Cleavage and translational suppression". In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1779.11 (2008), pp. 668–677. DOI: 10.1016/j.bbagrm.2008.07.005. URL: http://dx.doi.org/10.1016/j.bbagrm.2008.07.005.

[204] David P. Bartel. "MicroRNAs: Target Recognition and Regulatory Functions". In: *Cell* 136.2 (2009), pp. 215–233. ISSN: 00928674. DOI: 10.1016/j.cell.2009.01.002.

[205] Florian Erhard et al. “Widespread context dependency of microRNA-mediated regulation”. In: *Genome Research* 24.6 (2014), pp. 906–919. ISSN: 1088-9051. DOI: 10.1101/gr.166702.113. URL: http://genome.cshlp.org/cgi/doi/10.1101/gr.166702.113.

[206] Vikram Agarwal et al. “Predicting effective microRNA target sites in mammalian mRNAs”. In: *eLife* 4.AUGUST2015 (2015), pp. 1–38. ISSN: 2050084X. DOI: 10.7554/eLife.05005.

[207] Takaya Saito and Pal Sætrom. “MicroRNAs - targeting and target prediction”. In: *New Biotechnology* 27.3 (2010), pp. 243–249. ISSN: 18716784. DOI: 10.1016/j.nbt.2010.02.016.

[208] Pål Sætrom et al. “Distance constraints between microRNA target sites dictate efficacy and cooperativity”. In: *Nucleic Acids Research* 35.7 (2007), pp. 2333–2342. ISSN: 03051048. DOI: 10.1093/nar/gkm133.

[209] Andrew Grimson et al. “MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing”. In: *Molecular Cell* 27.1 (2007), pp. 91–105. ISSN: 10972765. DOI: 10.1016/j.molcel.2007.06.017.

[210] Anna Lukasik, Maciej Wójcikowski, and Piotr Zielenkiewicz. “Tools4miRs ? one place to gather all the tools for miRNA analysis”. In: *Bioinformatics* 32.17 (2016), pp. 2722–2724. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw189. URL: https://oup.silverchair-cdn.com/oup/backfile/Content{_}public/Journal/bioinformatics/32/17/10.1093{_}bioinformatics{_}btw189/2/btw189.pdf?Expires=1501270255{\&}Signature=GXHBXxwUd3SdRZoEWV9{~}yFFQDm-96LpW8jyASY8{~}0tH9RdsBK6JqOfuKPopyoL2lTCvdTfhMt9l{~}OeLTuS6ba88PiFVdxNYusghSgVtAPYL3z9DUvawSV3rr7sPHbNrSc
1ycRDF63VEGURrXsiSuGhVmZkzy2roLgme57hJvsGiPb1E{~}o1dMZqNwXRAofGDBhhGV3ASU03Hf7s-W2iKuwMny5U9lIFKcOzObMab48jEbTXqnGei{~}UnTWCzQ2VA6SRjeOpHlFxZqkUkhS{~}JfnwnNIy8JtOB7owmErx0tgmw7RA{_}{_}{\&}Key-Pair-Id=APKAIUCZBIA4LVPAVW3Qhttps://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw189.

[211] J. Konig et al. “Protein-RNA interactions: new genomic technologies and perspectives”. In: *Nat. Rev. Genet.* 13.2 (2012), pp. 77–83.

[212] M. Uhl et al. “Computational analysis of CLIP-seq data”. In: *Methods* 118-119 (2017), pp. 60–72.

[213] Markus Hafner et al. “Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP”. In: *Cell* 141.1 (2010), pp. 129–141. ISSN: 0092-8674. DOI: 10.1016/j.cell.2010.03.009. URL: http://dx.doi.org/10.1016/j.cell.2010.03.009.

[214] J. Konig et al. “iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution”. In: *Nat. Struct. Mol. Biol.* 17.7 (2010), pp. 909–915.

[215] Y. Sugimoto et al. “Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions”. In: *Genome Biol.* 13.8 (2012), R67.

[216] E. L. Van Nostrand et al. “Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP)”. In: *Nat. Methods* 13.6 (2016), pp. 508–514.

[217] B. J. Zarnegar et al. “irCLIP platform for efficient characterization of protein-RNA interactions”. In: *Nat. Methods* 13.6 (2016), pp. 489–492.

[218] Ioannis S. Vlachos et al. “DIANA-TarBase v7.0: Indexing more than half a million experimentally supported miRNA:mRNA interactions”. In: *Nucleic Acids Research* 43.D1 (2015), pp. D153–D159. ISSN: 13624962. DOI: 10.1093/nar/gku1215.

[219] Chih-Hung Chou et al. “miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database”. In: *Nucleic Acids Research* 44.D1 (2016), pp. D239–D247. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1258. URL: http://www.ncbi.nlm.nih.gov/pubmed/26590260http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4702890https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1258.

[220] Jun-Hao Li et al. “starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein?RNA interaction networks from large-scale CLIP-Seq data”. In: *Nucleic Acids Research* 42.D1 (2014), pp. D92–D97. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1248. URL: https://oup.silverchair-cdn.com/oup/backfile/Content{_}public/Journal/nar/42/D1/10.1093{_}nar{_}gkt1248/3/gkt1248.pdf?Expires=1501271164{\&}Signature=aGXxtpXzO-IX8IN0UGNLX2gdJBb8kR-LYk{~}sznMv86KXgpPNghSPpBxfLYfl1sR06HXaJ-3kXdQcsZqnXq98cuHkjF82qNdDFs3HK9L-R2nYRqx-pQvLsFc1hssL6e1p16t24pMIeR2f-LKXE3{~}THzc4Mt7-oI3z7eQ9m1HYQKcNdJ408f45SoxpNOrsUKtd5Y9J3VnFK-2tfLa{~}Hknuba{~}J3PsZxGNKclkLoOpgCRA2Ltr2Q2x-EgWz5OwEn11dwWH5ltDIbiPIfwhuI7bMApEfA34fBiiE{~}pf{~}tmgi1j8yKwvEUYvnmGJc1mGO-iYa-fRyroFGE4du4TgXLKg{_}{_}{\&}Key-Pair-Id=APKAIUCZBIA4LVPAVW3Qhttps://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1248.

[221] Feifei Xiao et al. “miRecords: an integrated resource for microRNA-target interactions”. In: *Nucleic Acids Research* 37.Database (2009), pp. D105–D110. ISSN: 0305-1048. DOI: 10.1093/nar/gkn851. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkn851.

[222] U.S. National Library of Medicine. *Detailed Indexing Statistics: 1965-2016*. 2017. URL: https://www.nlm.nih.gov/bsd/index_stats_comp.html (visited on 08/01/2017).

[223] Martin Gerner et al. “BioContext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events”. In: *Bioinformatics* 28.16 (2012), pp. 2154–2161. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts332. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3413385/pdf/bts332.pdfhttps://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts332.

[224] B. Stuart Murray et al. “An in silico analysis of microRNAs: Mining the miRNAome”. In: *Mol. BioSyst.* 6 (10 2010), pp. 1853–1862. DOI: 10.1039/C003961F. URL: http://dx.doi.org/10.1039/C003961F.

[225] Haroon Naeem et al. “miRSel: Automated extraction of associations between microRNAs and genes from the biomedical literature”. In: *BMC Bioinformatics* 11.1 (2010), p. 135. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-135. URL: http://services.bio.ifi.lmu.de/mirsel.http://www.biomedcentral.com/1471-2105/11/135.

[226] Shweta Bagewadi et al. “Detecting miRNA Mentions and Relations in Biomedical Literature”. In: *F1000Research* (2015). ISSN: 2046-1402. DOI: 10.12688/f1000research.4591.3. URL: https://f1000researchdata.s3.amazonaws.com/manuscripts/7643/76cac8ca-64f5-4e5b-84ed-d7c260309e6b{_}4591{_}-{_}shweta{_}bagewadi{_}v3.pdf?doi=10.12688/f1000research.4591.3http://f1000research.com/articles/3-205/v3.

[227] Gang Li et al. “miRTex: A Text Mining System for miRNA-Gene Relation Extraction”. In: *PLOS Computational Biology* 11.9 (2015). Ed. by Andrey Rzhetsky, e1004391. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004391. URL: https://

www.ncbi.nlm.nih.gov/pmc/articles/PMC4583433/pdf/pcbi.1004391.pdfhttp://dx.plos.org/10.1371/journal.pcbi.1004391.

[228] A. Lamurias, L. A. Clarke, and F. M. Couto. "Extracting microRNA-gene relations from biomedical literature using distant supervision". In: *PLoS ONE* 12.3 (2017), e0171929.

[229] Juan Wang et al. "TransmiR: a transcription factor–microRNA regulation database". In: *Nucleic acids research* 38.suppl_1 (2009), pp. D119–D122.

[230] Gergely Csaba. "Context based bioinformatics". PhD thesis. 2013. URL: http://nbn-resolving.de/urn:nbn:de:bvb:19-157252.

[231] Mohsen Naghavi et al. "Global, regional, and national age?sex specific all-cause and cause-specific mortality for 240 causes of death, 1990?2013: a systematic analysis for the Global Burden of Disease Study 2013". In: *The Lancet* 385.9963 (2015), pp. 117–171. ISSN: 01406736. DOI: 10.1016/S0140-6736(14)61682-2. arXiv: arXiv:1011.1669v3. URL: http://dx.doi.org/10.1016/S0140-6736(14)61682-2http://linkinghub.elsevier.com/retrieve/pii/S0140673614616822.

[232] Aldons J Lusis. "Atherosclerosis". In: *Nature* 407.September (2000), pp. 233–241. ISSN: 0028-0836. DOI: 10.1038/35025203. arXiv: 0102091 [cond-mat]. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2826222{\&}tool=pmcentrez{\&}rendertype=abstract.

[233] Ira Tabas, Guillermo García-Cardeña, and Gary K. Owens. "Recent insights into the cellular biology of atherosclerosis". In: *Journal of Cell Biology* 209.1 (2015), pp. 13–22. ISSN: 15408140. DOI: 10.1083/jcb.201412052.

[234] Marja Talikka, Stephanie Boue, and Walter K. Schlage. "Causal biological network database: A comprehensive platform of causal biological network models focused on the pulmonary and vascular systems". In: *Computational Systems Toxicology* (2015), pp. 65–93. ISSN: 19406053. DOI: 10.1007/978-1-4939-2778-4_3.

[235] Alma Zernecke and Christian Weber. "Chemokines in atherosclerosis: Proceedings resumed". In: *Arteriosclerosis, Thrombosis, and Vascular Biology* 34.4 (2014), pp. 742–750. ISSN: 15244636. DOI: 10.1161/ATVBAHA.113.301655.

[236] Petra Hartmann, Andreas Schober, and Christian Weber. "Chemokines and microRNAs in atherosclerosis". In: *Cellular and Molecular Life Sciences* 72.17 (2015), pp. 3253–3266. ISSN: 1420-682X. DOI: 10.1007/s00018-015-1925-z. URL: http://link.springer.com/10.1007/s00018-015-1925-z.

[237] David Warde-Farley et al. "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function". In: *Nucleic Acids Research* 38.suppl_2 (2010), W214–W220. ISSN: 1362-4962. DOI: 10.1093/nar/gkq537. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2896186/pdf/gkq537.pdfhttps://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq537.

[238] K. Zuberi et al. "GeneMANIA prediction server 2013 update". In: *Nucleic Acids Res.* 41.Web Server issue (2013), W115–122.

[239] Peter Brass. *Advanced data structures*. Vol. 1. Cambridge University Press Cambridge, 2008.

[240] M. Franz et al. "Cytoscape.js: a graph theory library for visualisation and analysis". In: *Bioinformatics* 32.2 (2016), pp. 309–311.

[241] Donald E Knuth. "Literate programming". In: *CSLI Lecture Notes, Stanford, CA: Center for the Study of Language and Information (CSLI), 1992* 1.2 (1984), pp. 97–111. ISSN: 0010-4620. DOI: 10.1093/comjnl/27.2.97.

[242] Isabella Faraoni et al. “miR-155 gene: A typical multifunctional microRNA”. In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1792.6 (2009), pp. 497–505. ISSN: 09254439. DOI: 10.1016/j.bbadis.2009.02.013. URL: http://ac.els-cdn.com/S0925443909000428/1-s2.0-S0925443909000428-main.pdf?{_}tid=29c1ea1c-7924-11e7-94f8-00000aacb35d{\&}acdnat=1501858368{_}5b9e7f398fc8ab0636ae02dff561d8bahttp://linkinghub.elsevier.com/retrieve/pii/S0925443909000428.

[243] Fatiha Tabet et al. “HDL-transferred microRNA-223 regulates ICAM-1 expression in endothelial cells”. In: *Nature Communications* 5 (2014). ISSN: 2041-1723. DOI: 10.1038/ncomms4292. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4189962/pdf/nihms576607.pdfhttp://www.nature.com/doifinder/10.1038/ncomms4292.

[244] T. Barrett et al. “NCBI GEO: archive for functional genomics data sets–update”. In: *Nucleic Acids Research* 41.D1 (2013), pp. D991–D995. ISSN: 0305-1048. DOI: 10.1093/nar/gks1193. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks1193.

[245] Marta Kubiak and Marzena Anna Lewandowska. “Can chromatin conformation technologies bring light into human molecular pathology?” In: *Acta Biochimica Polonica* 62.3 (2015), pp. 483–489. ISSN: 0001-527X. DOI: 10.18388/abp.2015_984. URL: http://www.ncbi.nlm.nih.gov/pubmed/26328275http://www.actabp.pl/{\#}File?./html/3{_}2015/2015{_}984.html.

[246] J. Dekker et al. “Capturing Chromosome Conformation”. In: *Science* 295.5558 (2002), pp. 1306–1311. ISSN: 00368075. DOI: 10.1126/science.1067799. URL: http://www.ncbi.nlm.nih.gov/pubmed/11847345http://www.sciencemag.org/cgi/doi/10.1126/science.1067799.

[247] Jon-Matthew Belton et al. “Hi?C: A comprehensive technique to capture the conformation of genomes”. In: *Methods* 58.3 (2012), pp. 268–276. ISSN: 10462023. DOI: 10.1016/j.ymeth.2012.05.001. URL: http://www.ncbi.nlm.nih.gov/pubmed/22652625http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3874846http://linkinghub.elsevier.com/retrieve/pii/S1046202312001168.

[248] Ana Pombo and Niall Dillon. “Three-dimensional genome architecture: players and mechanisms”. In: *Nature Reviews Molecular Cell Biology* 16.4 (2015), pp. 245–257. ISSN: 1471-0072. DOI: 10.1038/nrm3965. URL: http://www.ncbi.nlm.nih.gov/pubmed/25757416http://www.nature.com/doifinder/10.1038/nrm3965.

[249] Darío G. Lupiáñez et al. “Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions”. In: *Cell* 161.5 (2015), pp. 1012–1025. ISSN: 00928674. DOI: 10.1016/j.cell.2015.04.004. URL: http://www.ncbi.nlm.nih.gov/pubmed/25959774http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4791538http://linkinghub.elsevier.com/retrieve/pii/S0092867415003773.

[250] D. Hnisz et al. “Activation of proto-oncogenes by disruption of chromosome neighborhoods”. In: *Science* 351.6280 (2016), pp. 1454–1458. ISSN: 0036-8075. DOI: 10.1126/science.aad9024. URL: http://www.ncbi.nlm.nih.gov/pubmed/26940867http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4884612http://www.sciencemag.org/cgi/doi/10.1126/science.aad9024.

[251] Nynke L. van Berkum et al. “Hi-C: A Method to Study the Three-dimensional Architecture of Genomes.” In: *Journal of Visualized Experiments* 39 (2010). ISSN: 1940-087X. DOI: 10.3791/1869. URL: http://www.ncbi.nlm.nih.gov/pubmed/

20461051http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3149993http://www.jove.com/index/Details.stp?ID=1869.

[252] Matteo Vietri Rudan et al. "Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture". In: *Cell Reports* 10.8 (2015), pp. 1297–1309. ISSN: 22111247. DOI: 10.1016/j.celrep.2015.02.004. URL: http://www.ncbi.nlm.nih.gov/pubmed/25732821http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4542312http://linkinghub.elsevier.com/retrieve/pii/S2211124715001126.

[253] Galip Gürkan Yardimci and William Stafford Noble. "Software tools for visualizing Hi-C data". In: *Genome Biology* 18.1 (2017), p. 26. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1161-y. URL: http://www.ncbi.nlm.nih.gov/pubmed/28159004http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5290626http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1161-y.

[254] Neva C. Durand et al. "Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom". In: *Cell Systems* 3.1 (2016), pp. 99–101. ISSN: 24054712. DOI: 10.1016/j.cels.2015.07.012. URL: http://www.ncbi.nlm.nih.gov/pubmed/27467250http://linkinghub.elsevier.com/retrieve/pii/S240547121500054X.

[255] Sven Heinz et al. "Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities". In: *Molecular Cell* 38.4 (2010), pp. 576–589. ISSN: 10972765. DOI: 10.1016/j.molcel.2010.05.004. URL: http://www.ncbi.nlm.nih.gov/pubmed/20513432http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2898526http://linkinghub.elsevier.com/retrieve/pii/S1097276510003667.

[256] H. Li and R. Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform". In: *Bioinformatics* 25.14 (2009), pp. 1754–1760. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp324. URL: http://www.ncbi.nlm.nih.gov/pubmed/19451168http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2705234https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp324.

[257] Ben Langmead et al. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome". In: *Genome Biology* 10.3 (2009), R25. ISSN: 1465-6906. DOI: 10.1186/gb-2009-10-3-r25. URL: http://www.ncbi.nlm.nih.gov/pubmed/19261174http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2690996http://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-3-r25.

[258] Suhas S.P. Rao et al. "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping". In: *Cell* 159.7 (2014), pp. 1665–1680. ISSN: 00928674. DOI: 10.1016/j.cell.2014.11.021. URL: http://www.ncbi.nlm.nih.gov/pubmed/25497547http://linkinghub.elsevier.com/retrieve/pii/S0092867414014974.

[259] Erez Lieberman-Aiden et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." In: *Science (New York, N.Y.)* 326.5950 (2009), pp. 289–93. ISSN: 1095-9203. DOI: 10.1126/science.1181369. URL: http://www.ncbi.nlm.nih.gov/pubmed/19815776http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2858594.

[260] Philip A Knight and Daniel Ruiz. "A FAST ALGORITHM FOR MATRIX BALANCING". In: (2012). URL: http://drops.dagstuhl.de/volltexte/2007/1073/pdf/07071.KnightPhilip.Paper.1073.pdf.

[261] Caleb Weinreb and Benjamin J. Raphael. "Identification of hierarchical chromatin domains". In: *Bioinformatics* 32.11 (2016), pp. 1601–1609. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv485. URL: http://www.ncbi.nlm.nih.gov/pubmed/26315910http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4892410https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv485.

[262] Xin Zhou et al. "Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser." In: *Nature biotechnology* 33.4 (2015), pp. 345–6. ISSN: 1546-1696. DOI: 10.1038/nbt.3158. URL: http://www.ncbi.nlm.nih.gov/pubmed/25690851http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4467764.

[263] G N Filippova et al. "An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes." In: *Molecular and cellular biology* 16.6 (1996), pp. 2802–13. ISSN: 0270-7306. URL: http://www.ncbi.nlm.nih.gov/pubmed/8649389http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC231272.

[264] Shuangdi Li, Yanqiu Wang, and Jiarong Zhang. "L-Selectin Ligands Expression in Human Fallopian Tube Epithelia of Tubal Pregnancies". In: *BIOLOGY OF REPRODUCTION* 90133.6 (2014), pp. 1–6. DOI: 10.1095/biolreprod.113.113654. URL: https://oup.silverchair-cdn.com/oup/backfile/Content{_}public/Journal/biolreprod/90/6/10.1095{_}biolreprod.113.113654/1/biolreprod0133.pdf?Expires=1504117811{\&}Signature=DQt3hWG5aFGaTi9poJXeEeQ5rLFSLvywi3HmD4vHnnJiuwr-snW-Tff9RsHzrh2RCOXs6cwakq3WryTuxUCvkT1cGCEjigo1Nky73OFAtoTpjfaxmdwxrNTrSzwBr1ynA-lCBCzsjpLpUZJX5PeQfgj3a6Dc5saQQZKPJUVRHIqjQyThEPpWcGE3Nt5mhklaWnSx0dkQuHFLHcNo4zoqi538la
SQfBDdbBOpSHOcG4Pc6d79aKGQQWP-WjWXGyk2ucPBDvi1amtS9YZBfCYQeuFqEHLTPxxcRIZEzFfabgGuAyTbjs
}{_}{\&}Key-Pair-Id=APKAIUCZBIA4LVPAVW3Q.

[265] G I Johnston, R G Cook, and R P McEver. "Cloning of GMP-140, a granule membrane protein of platelets and endothelium: sequence similarity to proteins involved in cell adhesion and inflammation." In: *Cell* 56.6 (1989), pp. 1033–44. ISSN: 0092-8674. URL: http://www.ncbi.nlm.nih.gov/pubmed/2466574.

[266] Philippe Hupé et al. "Computational Systems Biology of Cancer Chapman & Hall/CRC Mathematical & Computational Biology". In: (2012). URL: https://en.wikipedia.org/wiki/Protein_mass_spectrometry#/media/File:Mass_spectrometry_protocol.png.

[267] Ruedi Aebersold and Matthias Mann. "Mass spectrometry-based proteomics". In: *Nature* 422.6928 (2003), pp. 198–207. ISSN: 0028-0836. DOI: 10.1038/nature01511. URL: http://dx.doi.org/10.1038/nature01511.

[268] Matthias Gstaiger and Ruedi Aebersold. "Applying mass spectrometry-based proteomics to genetics, genomics and network biology". In: *Nat Rev Genet* 10.9 (2009), pp. 617–627. ISSN: 1471-0056. DOI: 10.1038/nrg2633. URL: http://dx.doi.org/10.1038/nrg2633.

[269] Danielle L. Swaney, Craig D. Wenger, and Joshua J. Coon. "Value of Using Multiple Proteases for Large-Scale Mass Spectrometry-Based Proteomics". en. In: *Journal of Proteome Research* 9.3 (Mar. 2010), pp. 1323–1329. ISSN: 1535-3893, 1535-3907. DOI:

10.1021/pr900863u. URL: http://pubs.acs.org/doi/abs/10.1021/pr900863u (visited on 06/14/2017).

[270] Sa?a M. Miladinovi? Et al. "In-Spray Supercharging of Peptides and Proteins in Electrospray Ionization Mass Spectrometry". en. In: *Analytical Chemistry* 84.11 (June 2012), pp. 4647–4651. ISSN: 0003-2700, 1520-6882. DOI: 10.1021/ac300845n. URL: http://pubs.acs.org/doi/abs/10.1021/ac300845n (visited on 07/06/2017).

[271] N. Leigh Anderson et al. "The Human Plasma Proteome: A Nonredundant List Developed by Combination of Four Separate Sources". In: *Molecular & Cellular Proteomics* 3.4 (2004), pp. 311–326. DOI: 10.1074/mcp.M300127-MCP200. eprint: http://www.mcponline.org/content/3/4/311.full.pdf+html. URL: http://www.mcponline.org/content/3/4/311.abstract.

[272] Annette Michalski, Juergen Cox, and Matthias Mann. "More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority is Inaccessible to Data-Dependent LC?MS/MS". en. In: *Journal of Proteome Research* 10.4 (Apr. 2011), pp. 1785–1793. ISSN: 1535-3893, 1535-3907. DOI: 10.1021/pr101060v. URL: http://pubs.acs.org/doi/abs/10.1021/pr101060v (visited on 07/02/2017).

[273] Ju?rgen Cox et al. "Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment". en. In: *Journal of Proteome Research* 10.4 (Apr. 2011), pp. 1794–1805. ISSN: 1535-3893, 1535-3907. DOI: 10.1021/pr101065j. URL: http://pubs.acs.org/doi/abs/10.1021/pr101065j (visited on 07/03/2017).

[274] Jürgen Cox and Matthias Mann. "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification". en. In: *Nature Biotechnology* 26.12 (Dec. 2008), pp. 1367–1372. ISSN: 1087-0156. DOI: 10.1038/nbt.1511. URL: http://www.nature.com/nbt/journal/v26/n12/full/nbt.1511.html (visited on 07/02/2017).

[275] R. Craig and R. C. Beavis. "TANDEM: matching proteins with tandem mass spectra". en. In: *Bioinformatics* 20.9 (June 2004), pp. 1466–1467. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/bth092. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bth092 (visited on 07/08/2017).

[276] Benjamin J. Diament and William Stafford Noble. "Faster SEQUEST Searching for Peptide Identification from Tandem Mass Spectra". en. In: *Journal of Proteome Research* 10.9 (Sept. 2011), pp. 3871–3879. ISSN: 1535-3893, 1535-3907. DOI: 10.1021/pr101196n. URL: http://pubs.acs.org/doi/abs/10.1021/pr101196n (visited on 09/11/2017).

[277] Jan Eriksson and David Fenyö. "Probity: A Protein Identification Algorithm with Accurate Assignment of the Statistical Significance of the Results". en. In: *Journal of Proteome Research* 3.1 (Feb. 2004), pp. 32–36. ISSN: 1535-3893, 1535-3907. DOI: 10.1021/pr034048y. URL: http://pubs.acs.org/doi/abs/10.1021/pr034048y (visited on 09/11/2017).

[278] Henry Lam et al. "Development and validation of a spectral library searching method for peptide identification from MS/MS". en. In: *PROTEOMICS* 7.5 (Mar. 2007), pp. 655–667. ISSN: 1615-9861. DOI: 10.1002/pmic.200600625. URL: http://onlinelibrary.wiley.com/doi/10.1002/pmic.200600625/abstract (visited on 07/23/2017).

[279] R. Craig et al. "Using Annotated Peptide Mass Spectrum Libraries for Protein Identification". en. In: *Journal of Proteome Research* 5.8 (Aug. 2006), pp. 1843–1849. ISSN:

1535-3893, 1535-3907. DOI: 10.1021/pr0602085. URL: http://pubs.acs.org/doi/abs/10.1021/pr0602085 (visited on 07/08/2017).

[280] Henry Lam et al. "Building Consensus Spectral Libraries for Peptide Identification in Proteomics". In: *Nature methods* 5.10 (Oct. 2008), pp. 873–875. ISSN: 1548-7091. DOI: 10.1038/nmeth.1254. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2637392/ (visited on 07/08/2017).

[281] Bin Ma. "Novor: Real-Time Peptide de Novo Sequencing Software". In: *Journal of the American Society for Mass Spectrometry* 26.11 (2015), pp. 1885–1894. ISSN: 1044-0305. DOI: 10.1007/s13361-015-1204-0. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4604512/ (visited on 07/23/2017).

[282] Thilo Muth and Bernhard Y. Renard. "Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification?" en. In: *Briefings in Bioinformatics* (Mar. 2017). ISSN: 1467-5463, 1477-4054. DOI: 10.1093/bib/bbx033. URL: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbx033 (visited on 07/08/2017).

[283] Pedro Beltrao et al. "Evolution and functional cross-talk of protein post-translational modifications". en. In: *Molecular Systems Biology* 9.1 (2013), n/a–n/a. ISSN: 17444292. DOI: 10.1002/msb.201304521. URL: http://msb.embopress.org/cgi/doi/10.1002/msb.201304521 (visited on 07/24/2017).

[284] Jesper V. Olsen and Matthias Mann. "Status of Large-scale Analysis of Post-translational Modifications by Mass Spectrometry". en. In: *Molecular & Cellular Proteomics* 12.12 (Jan. 2013), pp. 3444–3452. ISSN: 1535-9476, 1535-9484. DOI: 10.1074/mcp.O113.034181. URL: http://www.mcponline.org/content/12/12/3444 (visited on 07/24/2017).

[285] Ruedi Aebersold and Matthias Mann. "Mass-spectrometric exploration of proteome structure and function". en. In: *Nature* 537.7620 (Sept. 2016), pp. 347–355. ISSN: 0028-0836. DOI: 10.1038/nature19949. URL: https://www.nature.com/nature/journal/v537/n7620/full/nature19949.html (visited on 06/28/2017).

[286] Arne H. Smits and Michiel Vermeulen. "Characterizing Protein?Protein Interactions Using Mass Spectrometry: Challenges and Opportunities". en. In: *Trends in Biotechnology* 34.10 (Oct. 2016), pp. 825–834. ISSN: 01677799. DOI: 10.1016/j.tibtech.2016.02.014. URL: http://linkinghub.elsevier.com/retrieve/pii/S0167779916000482 (visited on 07/07/2017).

[287] Andrew N. Holding. "XL-MS: Protein cross-linking coupled with mass spectrometry". en. In: *Methods* 89 (Nov. 2015), pp. 54–63. ISSN: 10462023. DOI: 10.1016/j.ymeth.2015.06.010. URL: http://linkinghub.elsevier.com/retrieve/pii/S1046202315002522 (visited on 07/06/2017).

[288] Antonio Artigues et al. "Protein Structural Analysis via Mass Spectrometry-Based Proteomics". In: *Advances in experimental medicine and biology* 919 (2016), pp. 397–431. ISSN: 0065-2598. DOI: 10.1007/978-3-319-41448-5_19. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5271599/ (visited on 07/07/2017).

[289] Anne-Claude Gingras et al. "Analysis of protein complexes using mass spectrometry". In: *Nature Reviews Molecular Cell Biology* 8.8 (Aug. 2007), pp. 645–654. ISSN: 1471-0072, 1471-0080. DOI: 10.1038/nrm2208. URL: http://www.nature.com/doifinder/10.1038/nrm2208 (visited on 07/31/2017).

[290] Juri Rappsilber. "The beginning of a beautiful friendship: Cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes". In: *Journal of Structural Biology* 173.3 (Mar. 2011), pp. 530–540. ISSN: 1047-8477. DOI: 10.1016/

j.jsb.2010.10.014. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3043253/ (visited on 07/31/2017).

[291] Chad R. Borges et al. "Building Multidimensional Biomarker Views of Type 2 Diabetes on the Basis of Protein Microheterogeneity". In: *Clinical chemistry* 57.5 (May 2011), pp. 719–728. ISSN: 0009-9147. DOI: 10.1373/clinchem.2010.156976. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3761388/ (visited on 07/31/2017).

[292] Jennifer Harrow et al. "GENCODE: The reference human genome annotation for The ENCODE Project". In: *Genome Research* 22.9 (2012), pp. 1760–1774. DOI: 10.1101/gr.135350.111. eprint: http://genome.cshlp.org/content/22/9/1760.full.pdf+html. URL: http://genome.cshlp.org/content/22/9/1760.abstract.

[293] Y. Liu, A. Beyer, and R. Aebersold. "On the Dependency of Cellular Protein Levels on mRNA Abundance". In: *Cell* 165.3 (2016), pp. 535–550.

[294] Mathias Uhlén et al. "Tissue-based map of the human proteome". In: *Science* 347.6220 (2015). ISSN: 0036-8075. DOI: 10.1126/science.1260419. eprint: http://science.sciencemag.org/content/347/6220/1260419.full.pdf. URL: http://science.sciencemag.org/content/347/6220/1260419.

[295] Michael L. Tress, Federico Abascal, and Alfonso Valencia. "Alternative Splicing May Not Be the Key to Proteome Complexity". en. In: *Trends in Biochemical Sciences* 42.2 (Feb. 2017), pp. 98–110. ISSN: 09680004. DOI: 10.1016/j.tibs.2016.08.008. URL: http://linkinghub.elsevier.com/retrieve/pii/S0968000416301189 (visited on 05/09/2017).

[296] Douglas L. Black. "Mechanisms of Alternative Pre-Messenger RNA Splicing". In: *Annual Review of Biochemistry* 72 (July 2003), pp. 291–336. DOI: 10.1146/annurev.biochem.72.121801.161720. URL: http://www.annualreviews.org/doi/abs/10.1146/annurev.biochem.72.121801.161720 (visited on 07/31/2017).

[297] Qin Li, Ji-Ann Lee, and Douglas L. Black. "Neuronal regulation of alternative pre-mRNA splicing". In: *Nat Rev Neurosci* 8.11 (2007), pp. 819–831. ISSN: 1471-003X. DOI: 10.1038/nrn2237. URL: http://dx.doi.org/10.1038/nrn2237.

[298] Iakes Ezkurdia et al. "The potential clinical impact of the release of two drafts of the human proteome". In: *Expert Review of Proteomics* 12.6 (2015), pp. 579–593. DOI: 10.1586/14789450.2015.1103186. URL: https://doi.org/10.1586/14789450.2015.1103186.

[299] Benjamin J. Blencowe. "The Relationship between Alternative Splicing and Proteomic Complexity". In: *Trends in Biochemical Sciences* 42.6 (2017), pp. 407–408. ISSN: 0968-0004. DOI: 10.1016/j.tibs.2017.04.001. URL: http://dx.doi.org/10.1016/j.tibs.2017.04.001.

[300] Min-Sik Kim et al. "A draft map of the human proteome". en. In: *Nature* 509.7502 (May 2014), pp. 575–581. ISSN: 0028-0836. DOI: 10.1038/nature13302. URL: http://www.nature.com/nature/journal/v509/n7502/full/nature13302.html (visited on 06/28/2017).

[301] Mathias Wilhelm et al. "Mass-spectrometry-based draft of the human proteome". en. In: *Nature* 509.7502 (May 2014), pp. 582–587. ISSN: 0028-0836. DOI: 10.1038/nature13319. URL: https://www.nature.com/nature/journal/v509/n7502/full/nature13319.html (visited on 06/28/2017).

[302] Tony Ly et al. "A proteomic chronology of gene expression through the cell cycle in human myeloid leukemia cells". In: *Elife* 3 (2014), e01630. URL: https://elifesciences.org/download/aHR0cHM6Ly9jZG4uZWxpZmVzY2llbmNlcy5vcmcvYXJ0aWNsZXMvMDE2MzAvZWxp

elife-01630-v1.pdf?_hash=sUQNgBcIYIbOnQkSyYCrVYbO9OUQtfkITOgv1gmIYwc%3D (visited on 08/01/2017).

[303] "UniProt: the universal protein knowledgebase". In: *Nucleic Acids Research* 45.D1 (Jan. 2017), pp. D158–D169. ISSN: 0305-1048. DOI: 10.1093/nar/gkw1099. URL: https://academic.oup.com/nar/article/45/D1/D158/2605721/UniProt-the-universal-protein-knowledgebase (visited on 08/01/2017).

[304] Alexander Franks, Edoardo Airoldi, and Nikolai Slavov. "Post-transcriptional regulation across human tissues". In: *PLOS Computational Biology* 13.5 (May 2017), e1005535. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005535. URL: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005535 (visited on 09/13/2017).

[305] Nuala A. O'Leary et al. "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". eng. In: *Nucleic Acids Research* 44.D1 (Jan. 2016), pp. D733–745. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1189.

[306] Bronwen L. Aken et al. "The Ensembl gene annotation system". In: *Database* 2016 (2016). DOI: 10.1093/database/baw093. URL: https://academic.oup.com/database/article/doi/10.1093/database/baw093/2630475/The-Ensembl-gene-annotation-system (visited on 08/01/2017).

[307] *ExPASy PeptideCutter tool: available enzymes.* URL: http://web.expasy.org/peptide_cutter/peptidecutter_enzymes.html (visited on 07/27/2017).

[308] Bushra Raj and BenjaminJ. Blencowe. "Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles". en. In: *Neuron* 87.1 (July 2015), pp. 14–27. ISSN: 08966273. DOI: 10.1016/j.neuron.2015.05.004. URL: http://linkinghub.elsevier.com/retrieve/pii/S0896627315004110 (visited on 09/14/2017).

[309] Douglas Hanahan and Robert A Weinberg. "The Hallmarks of Cancer Review evolve progressively from normalcy via a series of pre". In: *Cell* 100 (2000), pp. 57–70. URL: http://www.cell.com/cell/pdf/S0092-8674(00)81683-9.pdf.

[310] Douglas Hanahan and Robert A Weinberg. "Hallmarks of Cancer: The Next Generation". In: *Cell* 144 (2011), pp. 646–674. DOI: 10.1016/j.cell.2011.02.013. URL: http://www.cell.com/cell/pdf/S0092-8674(11)00127-9.pdf.

[311] The Cancer Genome Atlas Network. "Comprehensive molecular portraits of human breast tumours". In: (2012). DOI: 10.1038/nature11412.

[312] Wood WC et al. "Malignant Tumors of the Breast". In: *Cancer: Principles and Practice of Oncology.* Ed. by DeVita VT Jr, Hellman S, and Rosenberg SA. 7th ed. Philadelphia: Lippincott Williams & Wilkins, 2005, 1420ff.

[313] Xiaofeng Dai et al. "Breast cancer intrinsic subtype classification, clinical use and future trends". In: *Am J Cancer Res* 5.10 (2015), pp. 2929–2943. URL: www.ajcr.us.

[314] Dove Press. "Novel drugs that target the estrogen-related receptor alpha: their therapeutic potential in breast cancer". In: *Cancer Manag Res* (2014), pp. 225–252.

[315] T Sø rlie et al. "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications." In: *Proceedings of the National Academy of Sciences of the United States of America* 98.19 (2001), pp. 10869–74. ISSN: 0027-8424. DOI: 10.1073/pnas.191367098. URL: http://www.pnas.org/content/98/19/10869.full.pdfhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=58566{\&}tool=pmcentrez{\&}rendertype=abstract.

[316] Markus Schroeder et al. *breastCancerVDX: Gene expression datasets published by Wang et al. [2005] and Minn et al. [2007] (VDX).* R package version 1.14.0. 2011. URL: http://compbio.dfci.harvard.edu/.

[317] Markus Schroeder et al. *breastCancerTRANSBIG: Gene expression dataset published by Desmedt et al. [2007] (TRANSBIG).* R package version 1.14.0. 2011. URL: http://compbio.dfci.harvard.edu/.

[318] Markus Schroeder et al. *breastCancerUNT: Gene expression dataset published by Sotiriou et al. [2007] (UNT).* R package version 1.14.0. 2011. URL: http://compbio.dfci.harvard.edu/.

[319] Markus Schroeder et al. *breastCancerUPP: Gene expression dataset published by Miller et al. [2005] (UPP).* R package version 1.14.0. 2011. URL: http://compbio.dfci.harvard.edu/.

[320] Markus Schroeder et al. *breastCancerMAINZ: Gene expression dataset published by Schmidt et al. [2008] (MAINZ).* R package version 1.14.0. 2011. URL: http://compbio.dfci.harvard.edu/.

[321] Markus Schroeder et al. *breastCancerNKI: Genexpression dataset published by van't Veer et al. [2002] and van de Vijver et al. [2002] (NKI).* R package version 1.14.0. 2011. URL: http://compbio.dfci.harvard.edu/.

[322] Bioconductor Package Maintainer. *ExperimentHub: Client to access ExperimentHub resources.* R package version 1.2.0. 2017.

[323] Wilma Lingle et al. *Radiology Data from The Cancer Genome Atlas Breast Invasive Carcinoma [TCGA-BRCA] collection.* 2016. DOI: 10.7937/k9/tcia.2016.ab2nazrp.

[324] Katie Planey. *curatedBreastData: Curated breast cancer gene expression data with survival and treatment information.* R package version 2.4.0. 2016.

[325] E. Karlsson et al. "Gene expression variation to predict 10-year survival in lymph-node-negative breast cancer". In: *BMC Cancer* 8 (2008), p. 254.

[326] Y. Zhang et al. "The 76-gene signature defines high-risk patients that benefit from adjuvant tamoxifen therapy". In: *Breast Cancer Res. Treat.* 116.2 (2009), pp. 303–309.

[327] C. Desmedt et al. "The Gene expression Grade Index: a potential predictor of relapse for endocrine-treated breast cancer patients in the BIG 1-98 trial". In: *BMC Med Genomics* 2 (2009), p. 40.

[328] W. F. Symmans et al. "Genomic index of sensitivity to endocrine therapy for breast cancer". In: *J. Clin. Oncol.* 28.27 (2010), pp. 4111–4119.

[329] L. J. Esserman et al. "Chemotherapy response and recurrence-free survival in neoadjuvant breast cancer depends on biomarker profiles: results from the I-SPY 1 TRIAL (CALGB 150007/150012; ACRIN 6657)". In: *Breast Cancer Res. Treat.* 132.3 (2012), pp. 1049–1062.

[330] C. Hatzis et al. "A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer". In: *JAMA* 305.18 (2011), pp. 1873–1881.

[331] B. Haibe-Kains et al. "A three-gene model to robustly identify breast cancer molecular subtypes". In: *J. Natl. Cancer Inst.* 104.4 (2012), pp. 311–325.

[332] D. G. Altman and J. M. Bland. "Time to event (survival) data". In: *BMJ* 317.7156 (1998), pp. 468–469.

[333] J. M. Bland and D. G. Altman. "Survival probabilities (the Kaplan-Meier method)". In: *BMJ* 317.7172 (1998), p. 1572.

[334] V. Bewick, L. Cheek, and J. Ball. “Statistics review 12: survival analysis”. In: *Crit Care* 8.5 (2004), pp. 389–394.

[335] J. M. Bland and D. G. Altman. “The logrank test”. In: *BMJ* 328.7447 (2004), p. 1073.

[336] S. L. Spruance et al. “Hazard ratio in clinical trials”. In: *Antimicrob. Agents Chemother.* 48.8 (2004), pp. 2787–2792.

[337] I. Zwiener, M. Blettner, and G. Hommel. “Survival analysis: part 15 of a series on evaluation of scientific publications”. In: *Dtsch Arztebl Int* 108.10 (2011), pp. 163–169.

[338] Parker J Bernard P et al. “Supervised risk predictor of breast cancer based on intrinsic subtypes”. In: *Journal of Clinical Oncology* vol: 27.8 (2009), pp: 1160–1167.

[339] R. Tibshirani et al. “Diagnosis of multiple cancer types by shrunken centroids of gene expression”. In: *Proc. Natl. Acad. Sci. U.S.A.* 99.10 (2002), pp. 6567–6572.

[340] N. Harbeck et al. “Molecular and protein markers for clinical decision making in breast cancer: today and tomorrow”. In: *Cancer Treat. Rev.* 40.3 (2014), pp. 434–444.

[341] C. Desmedt et al. “Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes”. In: *Clin. Cancer Res.* 14.16 (2008), pp. 5158–5165.

[342] P. Wirapati et al. “Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures”. In: *Breast Cancer Res.* 10.4 (2008), R65.

[343] M. Filipits et al. “A new molecular predictor of distant recurrence in ER-positive, HER2-negative breast cancer adds independent information to conventional clinical risk factors”. In: *Clin. Cancer Res.* 17.18 (2011), pp. 6012–6020.

[344] S. Paik et al. “A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer”. In: *N. Engl. J. Med.* 351.27 (2004), pp. 2817–2826.

[345] C. Sotiriou and C. Desmedt. “Gene expression profiling in breast cancer”. In: *Ann. Oncol.* 17 Suppl 10 (2006), pp. x259–262.

[346] R. Buus et al. “Comparison of EndoPredict and EPclin With Oncotype DX Recurrence Score for Prediction of Risk of Distant Recurrence After Endocrine Therapy”. In: *J. Natl. Cancer Inst.* 108.11 (2016).

[347] Han-Yu Chuang et al. “Network-based classification of breast cancer metastasis.” In: *Molecular systems biology* 3.140 (2007), p. 140. ISSN: 1744-4292. DOI: 10.1038/msb4100180.

[348] Marc J. van de Vijver et al. “A Gene-Expression Signature as a Predictor of Survival in Breast Cancer”. In: *New England Journal of Medicine* 347.25 (2002), pp. 1999–2009. DOI: 10.1056/nejmoa021967. URL: https://doi.org/10.1056/nejmoa021967.

[349] Y WANG et al. “Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer”. In: *The Lancet* 365.9460 (2005), pp. 671–679. DOI: 10.1016/s0140-6736(05)70933-8. URL: https://doi.org/10.1016/s0140-6736(05)70933-8.

[350] H.-Y. Chuang et al. “Subnetwork-based analysis of chronic lymphocytic leukemia identifies pathways that associate with disease progression”. In: *Blood* 120.13 (2012), pp. 2639–2649. DOI: 10.1182/blood-2012-03-416461. URL: https://doi.org/10.1182/blood-2012-03-416461.

[351] Amin Allahyar and Jeroen De Ridder. “FERAL: Network-based classifier with application to breast cancer outcome prediction”. In: *Bioinformatics* (2015). ISSN: 14602059. DOI: 10.1093/bioinformatics/btv255.

[352] “A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer”. In: *PLoS ONE* 7.4 (2012). ISSN: 19326203.

[353] "Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis". In: *Frontiers in Genetics* 4.DEC (2013), pp. 1–15. ISSN: 16648021. DOI: 10.3389/fgene.2013.00289.

[354] K. R. Brown and I. Jurisica. "Online Predicted Human Interaction Database". In: *Bioinformatics* 21.9 (2005), pp. 2076–2082. DOI: 10.1093/bioinformatics/bti273. URL: https://doi.org/10.1093/bioinformatics/bti273.

[355] Matan Hofree et al. "Network-based stratification of tumor mutations". In: *Nature Methods* 10.11 (2013), pp. 1108–1115. DOI: 10.1038/nmeth.2651. URL: https://doi.org/10.1038/nmeth.2651.

[356] D. Szklarczyk et al. "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored". In: *Nucleic Acids Research* 39.Database (2010), pp. D561–D568. DOI: 10.1093/nar/gkq973. URL: https://doi.org/10.1093/nar/gkq973.

[357] I. Lee et al. "Prioritizing candidate disease genes by network-based boosting of genome-wide association data". In: *Genome Research* 21.7 (2011), pp. 1109–1121. DOI: 10.1101/gr.118992.110. URL: https://doi.org/10.1101/gr.118992.110.

[358] E. G. Cerami et al. "Pathway Commons, a web resource for biological pathway data". In: *Nucleic Acids Research* 39.Database (2010), pp. D685–D690. DOI: 10.1093/nar/gkq1039. URL: https://doi.org/10.1093/nar/gkq1039.

[359] D. Venet, J. E. Dumont, and V. Detours. "Most random gene expression signatures are significantly associated with breast cancer outcome". In: *PLoS Comput. Biol.* 7.10 (2011), e1002240.

[360] Winston Chang et al. *shiny: Web Application Framework for R*. R package version 1.0.3. 2017. URL: https://CRAN.R-project.org/package=shiny.

[361] Chao Cheng et al. "Understanding transcriptional regulation by integrative analysis of transcription factor binding data". In: *Genome research* 22.9 (2012), pp. 1658–1667. ISSN: 1088-9051. DOI: 10.1101/gr.136838.111.

[362] Thordur Oskarsson. "Extracellular matrix components in breast cancer progression and metastasis". In: *The Breast* 22 (2013), S66–S72.

[363] Michael EG Sauria et al. "HiFive: a tool suite for easy and efficient HiC and 5C data analysis". In: *Genome Biology* 16.1 (2015), p. 237. ISSN: 1474-760X. DOI: 10.1186/s13059-015-0806-y. URL: http://www.ncbi.nlm.nih.gov/pubmed/26498826http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5410870http://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0806-y.

[364] Cameron S. Osborne and Borbála Mifsud. "Capturing genomic relationships that matter". In: *Chromosome Research* 25.1 (2017), pp. 15–24. ISSN: 0967-3849. DOI: 10.1007/s10577-016-9546-4. URL: http://www.ncbi.nlm.nih.gov/pubmed/28078515http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5346121http://link.springer.com/10.1007/s10577-016-9546-4.

[365] Eitan Yaffe and Amos Tanay. "Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture". In: *Nature Genetics* 43.11 (2011), pp. 1059–1065. ISSN: 1061-4036. DOI: 10.1038/ng.947. URL: http://www.ncbi.nlm.nih.gov/pubmed/22001755http://www.nature.com/doifinder/10.1038/ng.947.

[366] Guoliang Li et al. "Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application". In: *BMC Genomics* 15.Suppl 12 (2014), S11. ISSN: 1471-2164. DOI: 10.1186/1471-2164-15-S12-S11. URL: http://www.

ncbi.nlm.nih.gov/pubmed/25563301http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4303937http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-S12-S11.

[367] Wendy Weijia Soon, Manoj Hariharan, and Michael P Snyder. "High-throughput sequencing for biology and medicine." In: *Molecular systems biology* 9 (2013), p. 640. ISSN: 1744-4292. DOI: 10.1038/msb.2012.61. URL: http://www.ncbi.nlm.nih.gov/pubmed/23340846http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3564260.

[368] Nicolas Servant et al. "HiC-Pro: an optimized and flexible pipeline for Hi-C data processing". In: *Genome Biology* 16.1 (2015), p. 259. ISSN: 1474-760X. DOI: 10.1186/s13059-015-0831-x. URL: http://www.ncbi.nlm.nih.gov/pubmed/26619908http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4665391http://genomebiology.com/2015/16/1/259.

[369] Harianto Tjong et al. "Population-based 3D genome structure analysis reveals driving forces in spatial genome organization". In: *Proceedings of the National Academy of Sciences* 113.12 (2016), E1663–E1672. ISSN: 0027-8424. DOI: 10.1073/pnas.1512577113. URL: http://www.ncbi.nlm.nih.gov/pubmed/26951677http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4812752http://www.pnas.org/lookup/doi/10.1073/pnas.1512577113.

[370] Suqing Li et al. "Exploring functions of long noncoding RNAs across multiple cancers through co-expression network". In: *Scientific Reports* 7.1 (2017), p. 754. ISSN: 2045-2322. DOI: 10.1038/s41598-017-00856-8. URL: http://www.ncbi.nlm.nih.gov/pubmed/28389669http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5429718http://www.nature.com/articles/s41598-017-00856-8.

[371] Elzo de Wit and Wouter de Laat. "A decade of 3C technologies: insights into nuclear organization." In: *Genes & development* 26.1 (2012), pp. 11–24. ISSN: 1549-5477. DOI: 10.1101/gad.179804.111. URL: http://www.ncbi.nlm.nih.gov/pubmed/22215806http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3258961.

[372] Marc W. Schmid, Stefan Grob, and Ueli Grossniklaus. "HiCdat: a fast and easy-to-use Hi-C data analysis tool". In: *BMC Bioinformatics* 16.1 (2015), p. 277. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0678-x. URL: http://www.ncbi.nlm.nih.gov/pubmed/26334796http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4559209http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0678-x.

[373] Neva C Durand et al. "Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments Tool Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments". In: *Cell Systems* 3 (2016), pp. 95–98. DOI: 10.1016/j.cels.2016.07.002. URL: http://dx.doi.org/10.1016/j.cels.2016.07.002.

[374] S. A. Helvik, O. Snøve, and P. Sætrom. "Reliable prediction of Drosha processing sites improves microRNA gene prediction". In: *Bioinformatics* 23.2 (2006), pp. 142–149. DOI: 10.1093/bioinformatics/btl570. URL: http://dx.doi.org/10.1093/bioinformatics/btl570.

[375] J. Han et al. "Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex". In: *Cell* 125.5 (2006), pp. 887–901.

[376] X. Zhang and Y. Zeng. "The terminal loop region controls microRNA processing by Drosha and Dicer". In: *Nucleic Acids Res.* 38.21 (2010), pp. 7689–7697.

[377] H. Ma et al. “Lower and upper stem-single-stranded RNA junctions together determine the Drosha cleavage site”. In: *Proc. Natl. Acad. Sci. U.S.A.* 110.51 (2013), pp. 20687–20692.

[378] X. Hu, C. Ma, and Y. Zhou. “A novel two-layer SVM model in miRNA Drosha processing site detection”. In: *BMC Syst Biol* 7 Suppl 4 (2013), S4.

[379] A. R. Forrest et al. “A promoter-level mammalian expression atlas”. In: *Nature* 507.7493 (2014), pp. 462–470.

[380] Y. Lee. “MicroRNA maturation: stepwise processing and subcellular localization”. In: *The EMBO Journal* 21.17 (2002), pp. 4663–4670. DOI: 10.1093/emboj/cdf476. URL: http://dx.doi.org/10.1093/emboj/cdf476.

[381] Katsutomo Okamura, Na Liu, and Eric C. Lai. “Distinct Mechanisms for MicroRNA Strand Selection by Drosophila Argonautes”. In: *Molecular Cell* 36.3 (2009), pp. 431–444. DOI: 10.1016/j.molcel.2009.09.027. URL: http://dx.doi.org/10.1016/j.molcel.2009.09.027.

[382] R. D. Morin et al. “Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells”. In: *Genome Research* 18.4 (2008), pp. 610–621. DOI: 10.1101/gr.7179508. URL: http://dx.doi.org/10.1101/gr.7179508.

[383] Nicole Cloonan et al. “MicroRNAs and their isomiRs function cooperatively to target common biological pathways”. In: *Genome Biol* 12.12 (2011), R126. DOI: 10.1186/gb-2011-12-12-r126. URL: http://dx.doi.org/10.1186/gb-2011-12-12-r126.

[384] G. C. Tan et al. “5' isomiR variation is of functional and evolutionary importance”. In: *Nucleic Acids Research* 42.14 (2014), pp. 9424–9435. DOI: 10.1093/nar/gku656. URL: http://dx.doi.org/10.1093/nar/gku656.

[385] Chenfeng He et al. “MiRmat: Mature microRNA Sequence Prediction”. In: *PLoS ONE* 7.12 (2012). Ed. by Lukasz Kurgan, e51673. DOI: 10.1371/journal.pone.0051673. URL: http://dx.doi.org/10.1371/journal.pone.0051673.

[386] Anastasia Khvorova, Angela Reynolds, and Sumedha D. Jayasena. “Functional siRNAs and miRNAs Exhibit Strand Bias”. In: *Cell* 115.2 (2003), pp. 209–216. DOI: 10.1016/s0092-8674(03)00801-8. URL: http://dx.doi.org/10.1016/S0092-8674(03)00801-8.

[387] M. Kanamori-Katayama et al. “Unamplified cap analysis of gene expression on a single-molecule sequencer”. In: *Genome Research* 21.7 (2011), pp. 1150–1159. DOI: 10.1101/gr.115469.110. URL: http://dx.doi.org/10.1101/gr.115469.110.

[388] S. Griffiths-Jones. “The microRNA Registry”. In: *Nucleic Acids Research* 32.90001 (2004), pp. 109D–111. DOI: 10.1093/nar/gkh023. URL: http://dx.doi.org/10.1093/nar/gkh023.

[389] A. Kozomara and S. Griffiths-Jones. “miRBase: annotating high confidence microRNAs using deep sequencing data”. In: *Nucleic Acids Research* 42.D1 (2013), pp. D68–D73. DOI: 10.1093/nar/gkt1181. URL: http://dx.doi.org/10.1093/nar/gkt1181.

[390] S. K. Wyman et al. “Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity”. In: *Genome Research* 21.9 (2011), pp. 1450–1461. DOI: 10.1101/gr.118059.110. URL: http://dx.doi.org/10.1101/gr.118059.110.

[391] Haoquan Wu et al. “Alternative Processing of Primary microRNA Transcripts by Drosha Generates 5′ End Variation of Mature microRNA”. In: *PLoS ONE* 4.10 (2009). Ed. by Dominik Hartl, e7566. DOI: 10.1371/journal.pone.0007566. URL: http://dx.doi.org/10.1371/journal.pone.0007566.

[392] B. Li et al. “RNA-Seq gene expression estimation with read mapping uncertainty”. In: *Bioinformatics* 26.4 (2009), pp. 493–500. DOI: 10.1093/bioinformatics/btp692. URL: http://dx.doi.org/10.1093/bioinformatics/btp692.

[393] Geoffrey J. Faulkner et al. “A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE”. In: *Genomics* 91.3 (2008), pp. 281–288. DOI: 10.1016/j.ygeno.2007.11.003. URL: http://dx.doi.org/10.1016/j.ygeno.2007.11.003.

[394] J. E. Babiarz et al. “Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs”. In: *Genes & Development* 22.20 (2008), pp. 2773–2785. DOI: 10.1101/gad.1705308. URL: http://dx.doi.org/10.1101/gad.1705308.

[395] Ronny Lorenz et al. “ViennaRNA Package 2.0”. In: *Algorithms Mol Biol* 6.1 (2011), p. 26. DOI: 10.1186/1748-7188-6-26. URL: http://dx.doi.org/10.1186/1748-7188-6-26.

[396] Peter Kerpedjiev. *RNA Secondary Stucture as Graph Using the forgi Library.* 2014. URL: http://www.tbi.univie.ac.at/~pkerp/forgi/ (visited on 02/10/2015).

[397] Yan Zeng, Rui Yi, and Bryan R Cullen. “Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha”. In: *The EMBO Journal* 24.1 (2004), pp. 138–148. DOI: 10.1038/sj.emboj.7600491. URL: http://dx.doi.org/10.1038/sj.emboj.7600491.

[398] E. Berezikov et al. “Deep annotation of Drosophila melanogaster microRNAs yields insights into their processing, modification, and emergence”. In: *Genome Research* 21.2 (2010), pp. 203–215. DOI: 10.1101/gr.116657.110. URL: http://dx.doi.org/10.1101/gr.116657.110.

[399] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *SciPy: Open source scientific tools for Python.* 2001. URL: http://www.scipy.org/ (visited on 02/16/2015).

[400] Fernando Pérez and Brian E. Granger. “IPython: a System for Interactive Scientific Computing”. In: *Computing in Science and Engineering* 9.3 (May 2007), pp. 21–29. ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.53. URL: http://ipython.org.

[401] J. D. Hunter. “Matplotlib: A 2D graphics environment”. In: *Computing In Science & Engineering* 9.3 (2007), pp. 90–95.

[402] M. Waskom. *Seaborn: statistical data visualization.* 2014. URL: http://stanford.edu/~mwaskom/software/seaborn/index.html (visited on 02/16/2015).

[403] Martijn Huynen, Robin Gutell, and Danielle Konings. “Assessing the reliability of RNA folding using statistical mechanics”. In: *Journal of Molecular Biology* 267.5 (1997), pp. 1104–1112. DOI: 10.1006/jmbi.1997.0889. URL: http://dx.doi.org/10.1006/jmbi.1997.0889.

[404] R. Das and D. Baker. “Automated de novo prediction of native-like RNA tertiary structures”. In: *Proceedings of the National Academy of Sciences* 104.37 (2007), pp. 14664–14669. DOI: 10.1073/pnas.0703836104. URL: http://dx.doi.org/10.1073/pnas.0703836104.

[405] T. Hamelryck and B. Manderick. “PDB file parser and structure class implemented in Python”. In: *Bioinformatics* 19.17 (2003), pp. 2308–2310. DOI: 10.1093/bioinformatics/btg299. URL: http://dx.doi.org/10.1093/bioinformatics/btg299.

[406] G. E. Crooks. “WebLogo: A Sequence Logo Generator”. In: *Genome Research* 14.6 (2004), pp. 1188–1190. DOI: 10.1101/gr.849004. URL: http://dx.doi.org/10.1101/gr.849004.

[407] Weiyang Shi et al. “A distinct class of small RNAs arises from pre-miRNA–proximal regions in a simple chordate”. In: *Nat Struct Mol Biol* 16.2 (2009), pp. 183–189. DOI: 10.1038/nsmb.1536. URL: http://dx.doi.org/10.1038/nsmb.1536.

[408] D. Langenberger et al. “Evidence for human microRNA-offset RNAs in small RNA sequencing data”. In: *Bioinformatics* 25.18 (2009), pp. 2298–2301. DOI: 10.1093/bioinformatics/btp419. URL: http://dx.doi.org/10.1093/bioinformatics/btp419.

[409] Leo Breiman. In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/a:1010933404324. URL: http://dx.doi.org/10.1023/A:1010933404324.

[410] Julia Starega-Roslan et al. “Sequence Features of Drosha and Dicer Cleavage Sites Affect the Complexity of IsomiRs”. In: *IJMS* 16.4 (2015), pp. 8110–8127. DOI: 10.3390/ijms16048110. URL: http://dx.doi.org/10.3390/ijms16048110.

[411] Kaycee A. Quarles et al. “Ensemble Analysis of Primary MicroRNA Structure Reveals an Extensive Capacity To Deform near the Drosha Cleavage Site”. In: *Biochemistry* 52.5 (2013), pp. 795–807. DOI: 10.1021/bi301452a. URL: http://dx.doi.org/10.1021/bi301452a.

[412] M. B. Warf, W. E. Johnson, and B. L. Bass. “Improved annotation of C. elegans microRNAs by deep sequencing reveals structures associated with processing by Drosha and Dicer”. In: *RNA* 17.4 (2011), pp. 563–577. DOI: 10.1261/rna.2432311. URL: http://dx.doi.org/10.1261/rna.2432311.

[413] J. Krol. “Structural Features of MicroRNA (miRNA) Precursors and Their Relevance to miRNA Biogenesis and Small Interfering RNA/Short Hairpin RNA Design”. In: *Journal of Biological Chemistry* 279.40 (2004), pp. 42230–42239. DOI: 10.1074/jbc.m404931200. URL: http://dx.doi.org/10.1074/jbc.M404931200.

[414] Ali A. Hosin et al. “MicroRNAs in atherosclerosis”. In: *Journal of Vascular Research* 51.5 (2014), pp. 338–349. ISSN: 14230135. DOI: 10.1159/000368193. URL: http://dx.doi.org/10.1016/j.kjms.2012.04.001.

[415] Donato Santovito, Virginia Egea, and Christian Weber. “Small but smart: MicroRNAs orchestrate atherosclerosis development and progression”. In: *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids* 1861.12 (2016), pp. 2075–2086. ISSN: 18792618. DOI: 10.1016/j.bbalip.2015.12.013. URL: http://dx.doi.org/10.1016/j.bbalip.2015.12.013.

[416] Cydney B Nielsen et al. “Determinants of targeting by endogenous and exogenous microRNAs and siRNAs”. In: (2007), pp. 1894–1910. DOI: 10.1261/rna.768207.

[417] Jonas Sicking et al. *Indexed Database API*. W3C Recommendation. http://www.w3.org/TR/2015/REC-IndexedDB-20150108/. W3C, Jan. 2015.