Detection of interacting transcription factors in human tissues using predicted DNA binding affinity

Alena Myšičková [1], Martin Vingron

Max Planck Institute for Molecular Biology, Ihnestr. 73, D–14195 Berlin, Germany

Abstract

Tissue-specific gene expression is generally regulated by combinatorial interactions among transcription factors (TFs) which bind to the DNA. Despite this known fact, previous discoveries of the mechanism that controls gene expressions usually consider only a single TF. Here, we provide a prediction of interacting TFs in 30 human tissues based on their DNA binding affinity in promoter regions. We analyzed all possible pairs of 130 vertebrate TFs from JASPAR database. First, all human promoter regions were scanned for single TF-DNA binding affinities with TRAP and for each TF, a rank list of all promoters ordered by the binding affinity was created. We then studied the similarity of the rank lists and detected candidates for TF-TF interaction by applying a conditional independence test for multiway contingency tables. Our candidates are validated by both known protein-protein interactions (PPIs) and known gene regulation mechanism in the selected tissue. We find that the known PPIs are significantly enriched in the group of our predicted TF-TF interactions (2-21 times of random expectation). In addition, the predicted interacting TFs for a specific tissue are supported in literature to be active regulators or expressed in the studied tissue.

Introduction

Transcriptional regulatory networks determine a spatial and temporal gene expression which enables the tissue-specificity of the cell (Naef and Huelsken, 2005). Regulatory networks include groups of control proteins, such as transcription factors (TFs) binding to short DNA motifs, called transcription factor binding sites (TFBS). Each TF can be connected to a set of its target genes - genes on whose promoters the TF binds in order to activate or repress them (Tan et al., 2008). In mammalian tissues, TFs do not usually act alone but form complexes with other TFs and cofactor proteins, which bind together to the DNA to affect synergistically the transcription of the target genes (Fedorova and Zink, 2008). This combinatorial regulation increase the specificity and flexibility of genes in controlling tissue development and differentiation. Therefore, detection of interacting TFs can significantly increase our understanding how tissue specificity is determined.

In previous years, a variety of experimental approaches was introduced to detect TF interactions controlling tissue gene expression, such as genomic microarrays (Rada-Iglesias et al., 2005) or chromatin immunoprecipitation followed by microarrays or high-throughput sequencing (Odom et al., 2004; Johnson et al., 2007). However, these studies are able to detect TF interactions on a limited scale since they basically treat each TF separately. A novel two-hybrid screening method which can detect physical protein-protein interactions was applied in mouse and human (Suzuki et al., 2001; Ravasi et al., 2010). Thus, such technology is able to detect just a part (25%) of all possible TF interactions.

To overcome the experimental limitation, several computational models were built to predict tissue-specific interacting TFs. Some of these models combine gene expression information with promoter sequence features (Klein and Vingron, 2007; Smith et al., 2007; Yu et al., 2006b) or integrate the evolutionary conservation of tissue-specific genes and TFs controlling their expression (Hu and Gallo, 2010). However, the results of these studies can be biased by pairs of cooperating TFs with similar motifs, as discussed in (Pape et al., 2009). While comparing all these methods just a small fraction of predicted TFs interactions can be found in more then one study. This fact demonstrates that different methods are able to identify interacting TFs from different perspective and that the mechanism regulating the tissue differentiation and development is still not completely understood.

With our study we try to create the next component in understanding the transcriptional networks in human tissues. To identify interacting

---

[1]Corresponding author, email: `mysickov@molgen.mpg.de`

TFs, we combine the information of predicted binding affinity of single TF on its target genes while investigating all possible pairs of studied TFs. Further, we include the information about a tissue specificity of the target genes and apply a 3-way contingency table test to determine the significance of the overlap of tissue-specific top-ranked target genes for 2 different TFs.

Methods

Similarity of ranked lists of target genes measured by the hypergeometric test

In our model, we use a simple assumption that two interacting TFs should share a significant number of identical target genes. In other words, if two different TFs bind on the same promoter regions they would very likely act together to direct the expression of their target genes. To evaluate the significance of the shared genes, we apply the hypergeometric test for ranked lists of TF's target genes.

To create the ranked list of target genes we first scan all human promoter regions (-500-0bp transcription start site (TSS), from genome.ucsc.edu, GRCh37/hg19 assembly) with TRAP (Roider et al., 2007), which predicts the binding strength of a given TF to the sequence based on a physical model. The binding affinity of all 130 TFs, represented by position weight matrices (PWMs), in the JASPAR CORE Vertebrata database (Bryne et al., 2008) to all human promoters is calculated. Separately for each TF, we rank the promoter regions by their binding affinity in a decreasing order, such that the genes with high binding affinity are placed on the top of the list. We measure the similarity of these rank lists for all possible ($130 * 129/2 = 8385$) pairs of TFs such that we calculate the significance of the shared target genes among the top-L_1 (for the first TF) and the top-L_2 (for the second TF) ranked genes with the hypergeometric test (Fisher, 1922). This problem corresponds to a simple 2-way contingency table with random variables X and Y indicating genes ranked among the top-L_1 in the list of the first TF and and genes ranked among the top-L_2 in the list of the second TF, respectively.

To ensure an independence of the hypergeometric test statistic from the thresholds L_1 and L_2 we repeat the testing procedure for varying values of both cutoff points: $L_1, L_2 = 10, 20, \ldots, 990, 1000$ (together 10^4 combinations). We assume that the smallest obtained p-value of the hypergeometric test corresponds to the highest similarity between the two rank lists of target genes. Similar technique was applied in (Roider et al., 2009) to identify significant association of tissue specific genes and target genes of transcription factors.

Confounding factor: motif similarity

When two TFs have very similar motifs (represented by PWMs), with high probability their rank lists of target genes will be very similar (Pape et al., 2009). To exclude the choice of false positive candidates which would share a significant number of the identical genes in the top of the lists due to their similar matrices, we include a confounding factor into the analysis. For all pairs of TFs, we calculate their motif similarity using MOSTA $S^{\max}$ similarity measure (Pape et al., 2007) which is based on the log-odds ratio of the overlap probability and the independent probability of 2 motifs on both strands of a DNA sequence.

The similarity measure for all TF pairs ranges from -1.12 to 8.58. To avoid the presence of false positives in our prediction, we concentrate on TF pairs with motif similarity < 4.

Similarity of ranked lists of target genes in a tissue measured by testing in 3-way contingency tables

By definition, a 2-way contingency table depicts the association of 2 variables. In our case, the 2 variables come from 2 TFs. In order to stratify by tissue, we need to introduce a third dimension, thus arriving at a 3-way contingency table. We introduce variable Z_t, an indicator function of genes specific in the tissue t:

$$Z_t(i) = \begin{cases} 1 & \text{gene } i \text{ in tissue } t \\ 0 & \text{otherwise} . \end{cases}$$

As in previous section, random variables X and Y indicate genes ranked among the top-L_1 and top-L_2 in the list of the first and second TF, respectively. A graphic illustration of this situation is shown in Figure 1. All human genes are shown as dots, blue ones indicate tissue specific

genes (where $Z(i) = 1$). The green set highlights the top-ranked target genes of the first TF ($X(i) = 1$), red set highlights the top-ranked target genes of the second TF ($Y(i) = 1$). To test whether the number of the genes in the intersect of all 3 variables, e.g. $\sum_i (X(i) = 1, Y(i) = 1, Z(i) = 1)$, is larger than expected by random, 3-way contingency table test is applied (Gokhale and Kullback, 1978). We assume that the number of tissue-specific genes is fixed and test the null hypothesis of conditional independence of the two variables X and Y given Z, denoted as: $H_0 : (XY|Z)$. The corresponding $2 \times 2 \times 2$ contingency table is shown in Table 1, color coding of the random variables is identical with the coding in Figure 1.

The test statistic for the null hypothesis of conditional independence $(XY|Z)$ is the log-likelihood ratio of observed (μ_{xyz}) and expected frequencies ($\widehat{\mu}_{xyz}$) in the groups of variables X, Y and Z (Gokhale and Kullback, 1978):

$$2I(\mu : \widehat{\mu}) = 2 \sum_{xyz=1}^{2} \mu_{xyz} \log \left(\frac{\mu_{xyz}}{\widehat{\mu}_{xyz}} \right) \sim \chi^2_{df} .$$

df denotes the degrees of freedom of the χ^2 distribution and equals 2 for this particular test (one degree of freedom for each variable which expected frequencies has to be estimated). Under H_0, the xz- and yz- two-way marginals are identical with the observed one:

$$\widehat{\mu}_{x+z} = \mu_{x+z}; \ \widehat{\mu}_{+yz} = \mu_{+yz} \ ; x, y, z = 1, 2 .$$

Here, μ_{x+z} denotes the sum over all groups of variable y, e.g.: $\mu_{x+z} = \mu_{x1z} + \mu_{x2z}$ for $x, z = \{1, 2\}$. The expected frequencies in the contingency table are calculated with the following formula:

$$\widehat{\mu}_{x,y,z} = \frac{\mu_{x+z} \mu_{+yz}}{\mu_{++z}}$$

In the same way, $\mu_{++z} = \mu_{11z} + \mu_{12z} + \mu_{21z} + \mu_{22z}$, $z = \{1, 2\}$; denotes one-way marginals in the $2 \times 2 \times 2$ contingency table. The test statistic can be simply calculated using the loglinear representation (Gokhale and Kullback, 1978).
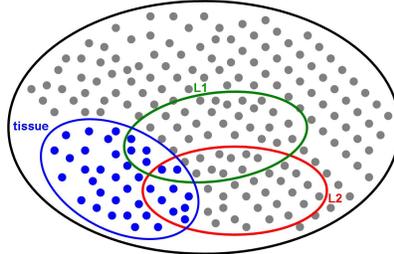

Figure 1: Venn diagram of the setting for independence test in 3-way contingency tables. Grey dots indicate all human genes, blue dots are genes known to be specific for a selected tissue. Green and red sets denote the top-ranked target genes of the first and second TF, respectively.

Table 1: $2 \times 2 \times 2$ contingency table for shared genes among the top-L_1 and top-L_2 ranked target genes of two different TFs and tissue-specific genes.

genes with …	tissue specific		not tissue specific		sum
	rank $\leq L_2$	rank $> L_2$	rank $\leq L_2$	rank $> L_2$	
rank $\leq L_1$	μ_{111}	μ_{121}	μ_{112}	μ_{122}	μ_{1++}
rank $> L_1$	μ_{211}	μ_{221}	μ_{212}	μ_{222}	μ_{2++}
sum	μ_{+11}	μ_{+21}	μ_{+12}	μ_{+22}	μ_{+++}

Results

Hypergeometric test

To assess the association between the similarity of rank lists and the similarity of PWMs, we study the relation between the smallest p-values obtained from the hypergeometric test and the similarity measure $S^{\max}$ (smoothed density scatterplot in Figure 2). As expected, TF pairs with very similar motifs ($S^{\max} \in [6; 8]$) correspond to highly significant p-values (data cloud in lower right corner). We identify already known protein-protein interactions (PPIs) and those TF pairs which have a joint already known co-factor (trios) found in Stark et al. (2011); Ravasi et al. (2010), (red dots and orange triangles in Figure 2). However, the majority of these known interactions corresponds to TF pairs with relative low significance ($p \in [10^{-3}; 10^0]$).

We define the candidates for TFs interactions as TF pairs with p-value $\leq 10^{-20}$, shown in Figure 4. The network consists of 76 interactions, 15.8% (7 fold enrichment, Figure 3) were found as already known PPIs (denoted as red edges). 22.4% are known trios, highlighted by orange color. For the prediction of TF interactions, we focused on 13 interactions between TF pairs with low motif similarity ($S^{\max} < 4$) which are represented by solid lines. 3 TF pairs have one or more already described common co-factors (EN1:TBP interacts with AP1 and PAX6; SP1:TFAP2A with TP53 and HOXA5:NR3C1 with PBX) which are indicated as grey nodes with corresponding grey edges. The evidence of a common co-factor increase the probability that these TFs can interact on the promoter. Manke et al. (2003) showed that the TFs build networks mostly with length of 2-4 molecules. Further, we find an experimental confirmation of our predictions in literature for these two interactions: SP1:TFAP2A (Pena et al., 1999; Perkins et al., 2001; Tellez and Bar-Eli, 2003) and GATA2:GATA3 (Minami et al., 2004).

Prediction of tissue-specific interactions

Before applying the new statistical test, tissue-specific genes have to be defined. For our analysis we use the data from Yu et al. (2006a) which are based on expression enrichment values calculated for 30 human tissues for ESTs clusters. The number of tissue-specific genes varies from 58 (uterus) to 860 (testis) which is a small number in comparison with the total number of promoters (42380, source: genome.ucsc.edu, GRCh37/hg19 assembly).

To achieve possible significance of the test statistic and to use the most relevant biological information, we fix the length of top-ranked target genes to 1000 for all TFs and do not repeat the testing procedures with various cut-off point as in the 2-way contingency tables. In this case, the expected number of shared top-ranked genes for 2 different TFs which are specific for a tissue is less or equal one. Thus, from now on, we concentrate only on those TF pairs which share at least one common tissue-specific top-ranked gene. The exact criterium for predicting the interactions for TF pairs is set as follows: take pairs with tissue-specific p-value $\leq 10^{-5}$ and the intersect of the 3 groups $\mu_{111} \geq 99\%$-quantile of the empiricial distribution of μ_{111} for all TF pairs. The 99%-quantile corresponds to 2 shared target genes (for bone, lung, mammary gland, small intestine, soft tissue, spleen and thymus) and 3 shared target genes (all other tissues) which guarantee a choice of highly significant TF pairs (since the number of shared genes expected by random chance is 0 or 1).

In total, we identify 352 significant TF pairs, 185 of them were between TFs with nonsimilar motifs ($S^{\max} < 4$). The most interactions are found in testis (108), the least in peripheral nervous system (2). 67 TF pairs are significant in more than 3 different tissues, just a minority (19.4%) of these multiple tissue pairs has low similar motifs. Among these belong: TFAP2A:MAFB, SP1:MAFB, ETS1:MAFB, SP1:ZFX, TFAP2A:ELK1, ELK1:MAFB, TFAP2A:PAX2, NFYA:TBP, ARNT:ARNT-AHR, TFAP2A:YY1, ETS1:ZFX and MAFB:ZFX. TF pair SP1:TFAP2A with low motif similarity is significant in 15 tissues and was identified also by the 2-dimensional hypergeometric test.

Evaluation by known protein-protein interactions

To further evaluate our predictions, we calculate the enrichment of known protein-protein interactions in the group of our candidates for each tissue. The percentage of known interactions is shown in the barplot in Figure 3. For all tissues,

the percentage is higher than what we expect by random chance (2.1%, dashed black line). The highest ratio is observed in peripheral nervous system (50%), the lowest ratio is recorded for bladder (4.2%).

In the next sections we present and validate our predictions of TF interactions in liver and skeletal muscle - 2 well-studied homogenous human tissues for which enough information is provided in the literature.

Prediction of interactions in liver

42 interactions are detected in liver by the criterium described above, network shown in Figure 5(a). Solid edges indicate 12 interactions between TFs with low motif similarity, remaining edges are between TF pairs with high motif similarity. 9 (20.5%) TFs in the network (HNF1A, HNF1B, HNF4A, NR2F1, NFKB1, NF-KappaB, RELA, PPARG-RORA, NR1H2-RORA) are supported in the literature to be transcriptional regulator in liver (Krivan and Wasserman (2001); Odom et al. (2004); Smith et al. (2007),TRANSFAC database, IPA Ingenuity Systems). 70% of the nodes (light green) was found to be expressed in liver tissue (source: TRANSFAC database, Gene Expression Atlas of EBI www.ebi.ac.uk/gxa/). However, known PPIs (red edges) are usually interactions between 2 TFs with very similar motifs and can be just partly used for validation of predicted interactions.

Next, we search for enriched GO annotation pathways for the predicted TFs. Among transcriptional regulation and DNA-binding, regulation of cholesterol transport (NR1H2,PPARG,NFKB1; Q-value=1.53×10^{-4}), regulation of lipid storage (NR1H2,PPARG,NFKB1; Q-value=3.29×10^{-4}), carbohydrate homeostasis, lipid homeostasis (HNF1A, STAT3, PPARG, USF1; Q-values=4.04×10^{-4}, 6.06×10^{-2}) and protein kinase binding (RELA, STAT3, USF1, FOXO3; Q-value=6.41×10^{-2}) are found.

Prediction of interactions in skeletal muscle

Figure 5(b) shows the network with 37 predicted interactions in muscle. Here, 6 TFs (MAFB, MEF2A, SP1, SRF, TBP and USF1) are known to regulate the gene expression in muscle (Smith et al., 2007). For 74% of factors evidence of expression in muscle is found (TRANSFAC database, Gene Expression Atlas of EBI www.ebi.ac.uk/gxa/). Again, known PPIs (red edges) are between factors with similar motifs. 7 of the predicted interactions are identified as known trios, which increases the validity of our predictions. TF pair SP1:TFAP2A discussed above was found in muscle too. MAFB:ETS1 is another TF pair with low motif similarity and with already described common co-factor (AP1). This interaction is found to be significant in other 6 tissues (bladder, blood, lymph node, placenta,thymus, spleen).

Further, several related GO pathways such as regulation of muscle contraction (SRF; Q-value=8.13×10^{-2}) and organ morphogenesis (PAX2, MYC, KLF4, SRF, STAT3; Q-value=1.68×10^{-3}) are found to be significantly enriched for the factors in the network (source: www.genemania.org/).

Discussion

Tissue-specific gene expression is in general regulated by interactions of multiple transcription factors. To better understand how cells in tissues and developmental states achieve their specificity the identification of interacting TFs regulating together the expression of their target genes is necessary. Previous computational studies were based either on common sequence features of promoters (Klein and Vingron, 2007; Smith et al., 2007; Yu et al., 2006b) or on function conservation of interacting TFs (Hu and Gallo, 2010). Although these studies make plausible predictions, they do not discriminate between factors with similar and different PWMs. Moreover, the mechanisms controlling tissue gene expression are still not fully understood.

In this study, we presented a new method to detect interacting TFs. We used the predicted binding affinity information for single TF and compared the ranked lists of the target genes for all pairs of studied TFs. To identify the interacting pairs in a tissue, tissue specificity information of the target genes was included. We applied the statistical testing in 3-way contingency tables to predict TF interactions. The number of false positives in our prediction was reduced by focusing on TF pairs with non similar motifs. In general, TFs with very similar motifs can jointly bind to the DNA sequence and regulate the transcription of the target gene.

However, our method is not able to distinguish between joint binding of both TFs and binding of a single TF for such similar TFBS.

We have shown that already known protein interactions are enriched (2-21 fold) in groups of selected candidates for all tissues and for candidates without tissue specification. In total, we have identified 352 significant TF pairs, 185 of them between TFs with nonsimilar motifs. Only a minority (19.4%) of TF pairs found in multiple tissues had nonsimilar motifs.

Networks made out of tissue-specific candidates include TFs which are known to regulate the gene expression in the studied tissue; such as HNF1A and HNF1B in liver and MAFB, USF, and TBP in muscle. Majority ($> 70\%$) of candidates in liver and skeletal muscle was found to be expressed in the studied tissue. These candidates factors were found just by the selection criterion from the statistical test, without any knowledge of their functions in human tissues. Further, we have identified 12 and 11 significant interactions among non-similar factors in liver and muscle, respectively. Some of our predicted interactions were supported by similar findings in literature (SP1:TFAP2A, GATA2:GATA4) or by already described common co-factors (MAFB:ETS1).

Despite of the successful predictions of novel pairs of interacting TFs, our method could be improved. Currently, we use a simple definition of promoter regions and tissue-specific genes. We could achieve much higher accuracy by using open chromatin regions for various cell types.

Further, we have used the groups of genes which are specific for a tissue. In general, many mammalian tissues are highly heterogenous and consist of different type of cells which could be regulated by different combination of TFs. Therefore, including cell-type specific genes would improve the accuracy of predicted interactions. On the other hand, if the groups include smaller number of specific genes the probability to have common specific genes in the top of the ranked lists will be even smaller. Next, to define the significant nonsimilar TF pairs, we use an arbitrary chosen threshold of 4 which excludes high similar PWMs from predicted candidates. This choice was made by a visual evaluation of binding motifs and could be improved for example by an implementation of weighting motif-similarity function or by a further analysis of binding site distances on the promoter.

In addition, a future experimental validation would provide a measure of the specificity and sensitivity of our predictions. In summary, our findings have shown that comparing the rank lists of target genes results in plausible predictions of interacting TFs in human tissues.

Acknowledgements

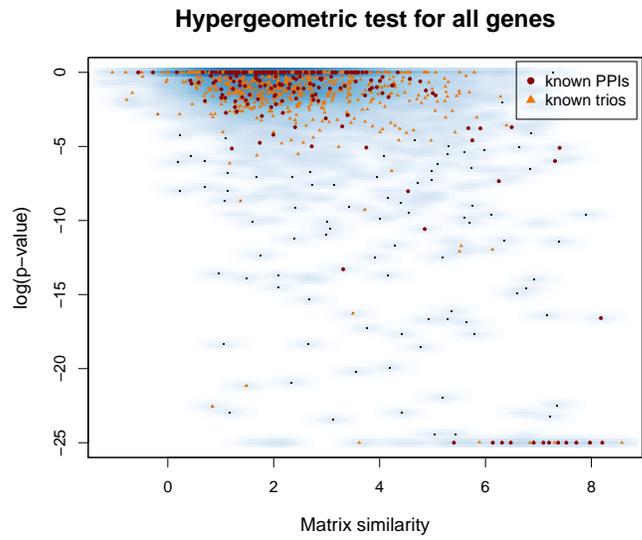**Hypergeometric test for all genes**

Figure 2: Logarithm of the smallest *p*-values of the hypergeometric test for all tested TF pairs (vertical axis) vs. motif similarity measure $S_{\max}$ (horizontal axis). Red points and orange triangles denote experimentally shown PPIs and trios with already known interacting co-factor, respectively.
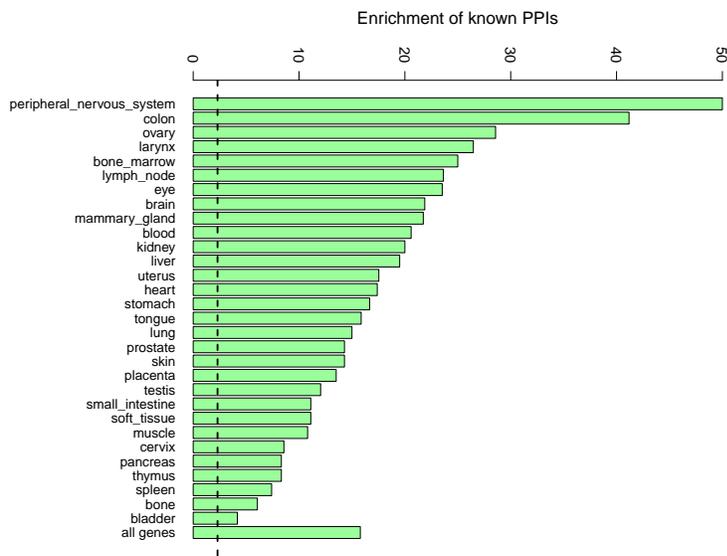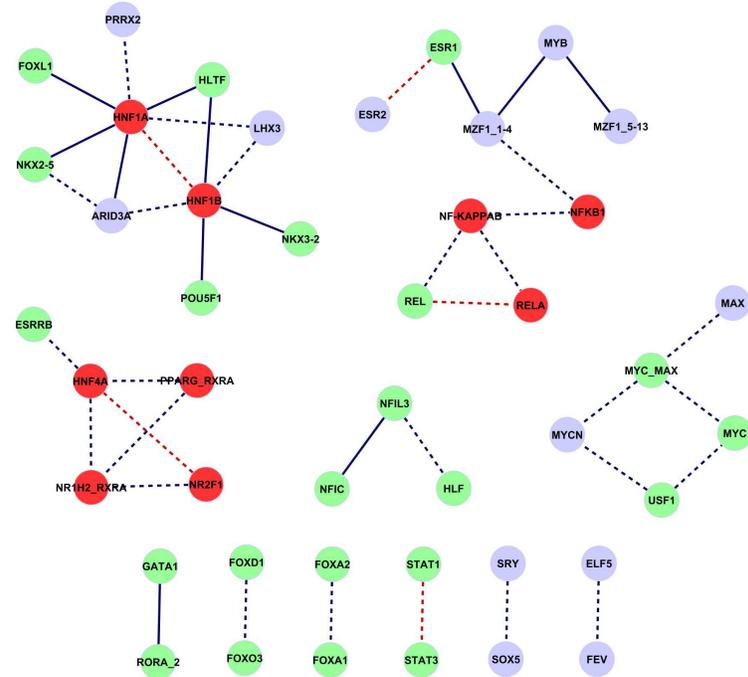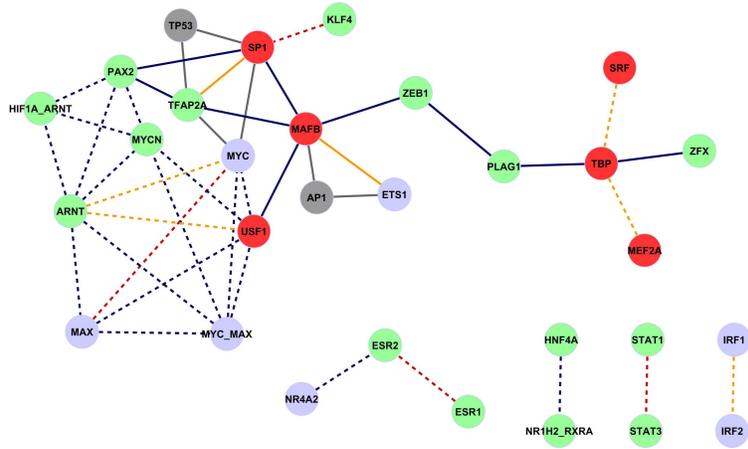

Enrichment of known PPIs

Figure 3: Percentage of known protein-protein interactions among TF interaction candidates in studied tissues with the expected percentage of known interactions by random chance (dashed line). The lowest bar corresponds to the 2-dimensional hypergeometric test.

Figure 4: Network of the predicted TF interactions based on the 2-dimensional hypergeometric test. Red and light green nodes denote TFs known to regulate gene expression in the corresponding tissue and TFs expressed in the corresponding tissue, respectively. Red and orange edges are known PPIs and known trios, respectively. Common co-factors which were included in the network but were not predicted are denoted by grey color.

(a) Liver

(b) Muscle

Figure 5: (a) Network of predicted TF interactions in liver based on testing in 3-way contingency tables. (b) Network of predicted TF interactions in muscle based on testing in 3-way contingency tables. Red and light green nodes denote TFs known to regulate gene expression in the corresponding tissue and TFs expressed in the corresponding tissue, respectively. Red and orange edges are known PPIs and known trios, respectively. Common co-factors which were included in the network but were not predicted are denoted by grey color.

References

J.C. Bryne, E. Valen, M.H. Tang, T. Marstrand, O. Winther, I. da Piedade, A. Krogh, B. Lenhard, and A. Sandelin. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, 36(Database issue):D102–6., 2008.

E Fedorova and D Zink. Nuclear architecture and gene regulation. *Biochimica et Biophysica Acta*, 1783(11):2174–2184, 2008.

R. A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

Dattaprabhakar V. Gokhale and Solomon Kullback. *The Information in Contingency Tables.* Marcel Dekker, Inc. New York and Basel, 1978.

Zihua Hu and Steven Gallo. Identification of interacting transcription factors regulating tissue gene expression in human. *BMC Genomics*, 11 (49)(1), 2010. ISSN 1471-2164.

David S. Johnson, Ali Mortazavi, Richard M. Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 8:316:1441–2, 2007.

H. Klein and M. Vingron. Using transcription factor binding site co-occurrence to predict regulatory regions. *Genome Informatics*, 18:109–18, 2007.

W. Krivan and W. W. Wasserman. A Predictive Model for Regulatory Sequences Directing Liver-Specific Transcription. *Genome Research*, 11(9):1559–1566, September 2001.

Thomas Manke, Ricardo Bringas, and Martin Vingron. Correlating Protein-DNA and Protein-Protein Interaction Networks. *Journal of Molecular Biology*, 333(1):75–85, October 2003. ISSN 00222836.

T. Minami, K. Horiuchi, M. Miura, M.R. Abid, W. Takabe, N. Noguchi, T. Kohro, X. Ge, H. Aburatani, T. Hamakubo, T. Kodama, and W.C. Aird. Vascular Endothelial Growth Factor- and Thrombin-induced Termination Factor, Down Syndrome Critical Region-1, Attenuates Endothelial Cell Proliferation and Angiogenesis. *J Biol Chem*, 279(48):50537–54, Nov 2004.

Felix Naef and Joerg Huelsken. Cell-type-specific transcriptomics in chimeric models using transcriptome-based masks. *Nucleic Acids Research*, 33, 2005.

Duncan T. Odom, Nora Zizlsperger, D. Benjamin Gordon, George W. Bell, Nicola J. Rinaldi, Heather L. Murray, Tom L. Volkert, Jörg Schreiber, P. Alexander Rolfe, David K. Gifford, Ernest Fraenkel, Graeme I. Bell, and Richard A. Young. Control of Pancreas and Liver Gene Expression by HNF Transcription Factors. *Science*, 303(5662):1378–1381, February 2004. ISSN 1095-9203.

Utz Pape, Sven Rahmann, and Martin Vingron. Natural similarity measures between position frequency matrices with an application to clustering. *Bioinformatics*, 24:350–357, 2007.

Utz Pape, Holger Klein, and Martin Vingron. Statistical detection of cooperative transcription factors with similarity adjustment. *Bioinformatics*, 25 (16):2103–2109, 2009.

P. Pena, AT Reutens, C Albanese, M D'Amico, G Watanabe, A Donner, IW Shu, T Williams, and RG Pestell. Activator protein-2 mediates transcriptional activation of the CYP11A1 gene by interaction with Sp1 rather than binding to DNA. *Mol Endocrinol*, 13(8):1402–16, Aug 1999.

KJ Perkins, EA Burton, and KE Davies. The role of basal and myogenic factors in the transcriptional activation of utrophin promoter A: implications for therapeutic up-regulation in Duchenne muscular dystrophy. *Nucleic Acids Res*, 29(23):4843–50, Dec 2001.

A. Rada-Iglesias, O. Wallerman, C. Koch, A. Ameur, S. Enroth, G. Clelland, K. Wester, S. Wilcox, O.M. Dovey, and P.D. Ellis. Binding sites for metabolic disease related transcription factors inferred at base pair resolution by chromatin immunoprecipitation and genomic microarrays. *Human molecular genetics*, 14(22):3435–3447, 2005.

Timothy Ravasi, Harukazu Suzuki, Carlo V. Cannistraci, Shintaro Katayama, Vladimir B. Bajic, Kai Tan, Altuna Akalin, Sebastian Schmeier, Mutsumi Kanamori-Katayama, and Nicolas Bertin. An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man. *Cell*, 140(5): 744–752, March 2010. ISSN 00928674.

Helge G. Roider, Aditi Kanhere, Thomas Manke, and Martin Vingron. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23:134–141, 2007.

Helge G. Roider, Thomas Manke, Sean O' Keeffe, Martin Vingron, and Stefan A. Haas. PASTAA: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics*, 25: 435–442, 2009.

Andrew D. Smith, Pavel Sumazin, and Michael Q. Zhang. Tissue-specific regulatory elements in mammalian promoters. *Mol Syst Biol*, 3:73, January 2007. ISSN 1744-4292.

C. Stark, B.J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M.S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, T. Reguly, J.M. Rust, A. Winter, K. Dolinski, and M. Tyers. The biogrid interaction database: 2011 update. *Nucleic Acids Res.*, 39 (Database issue):D698–704, 2011.

H. Suzuki, Y. Fukunishi, I. Kagawa, R. Saito, H. Oda, T. Endo, S. Kondo, H. Bono, Y. Okazaki, and Y. Hayashizaki. Protein-protein interaction panel using mouse full-length cdnas. *Genome Research*, 11:1758âĂŞ1765, 2001.

K. Tan, J. Tegner, and T. Ravasi. Integrated approaches to uncovering transcription regulatory networks in mammalian cells. *Genomics*, 91:219–31, 2008.

C. Tellez and M. Bar-Eli. Role and regulation of the thrombin receptor (PAR-1) in human melanoma. *Oncogene*, 22(20):3130–7, Jun 2003.

Xueping Yu, Jimmy Lin, Tomohiro Masuda, Noriko Esumi, Donald J. Zack, and Jiang Qian. Genome-wide prediction and characterization of interactions between transcription factors in Saccharomyces cerevisiae. *Nucleic Acids Research*, 34(3):917–927, 2006a. ISSN 0305-1048.

Xueping Yu, Jimmy Lin, Donald J. Zack, and Jiang Qian. Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic acids research*, 34(17):4925–4936, September 2006b. ISSN 1362-4962.