Rigorous assessment of gene set enrichment tests

Haroon Naeem, Ralf Zimmer, Robert Küffner
haroon.naeem@campus.lmu.de, ralf.zimmer@bio.ifi.lmu.de, robert.kueffner@bio.ifi.lmu.de
Department of Informatics, Ludwig-Maximilians Universität, Amalienstr. 17, 80333 München, Germany

Abstract

Background: Several enrichment tests are available to detect the enrichment of differential expression in gene sets. Such tests were originally proposed for analyzing gene sets associated with biological processes. The objective evaluation of tests on real measurements has not been possible as it is difficult to decide *a priori*, which processes will be affected in given experiments.
Results: For the rigorous assessment of enrichment tests we define gene sets based on the known targets of given regulators such as transcription factors (TFs) and microRNAs (miRNAs). In contrast to processes, TFs and miRNAs are amenable to direct manipulations, e.g. regulator over-expression or deletion. We assessed the ability of 12 different tests to predict the manipulations from expression measurements in *E. coli*, *S. cerevisiae* and human. We also analyzed how performance depends on the quality and comprehensiveness of the known regulator targets via an additional permutation approach. Overall, we find that ANOVA and Wilcoxons test consistently performed better than for instance Kolmogorov-Smirnov and hypergeometric tests. For novel scenarios where the optimal test is not known we suggest to combine all evaluated tests into an unweighted consensus.
Conclusion: We present a first large study to rigorously assess and compare the performance of gene set enrichment tests. Our results provide a guide for the selection of existing tests as well as a basis for the development and assessment of novel tests.

Introduction

The interpretation of gene expression studies reporting mRNA levels for a high number of genes or other expressed sequences is difficult. Instead of individual genes, it has been proposed to analyze gene sets corresponding to biological processes. The Gene Ontology (GO, *Harris et al.,* 2004) is an example source for biological process definitions and process associated gene sets. The analysis of expression data in the context of such gene sets can be performed by many different enrichment or over-representation tests (see section related work). These tests aim to detect gene sets exhibiting significant levels of differential expression. However, it is difficult to decide *a priori* which biological processes will be affected in a given gene expression experiment. This lack of a dependable standard of truth has prevented an objective selection and evaluation of enrichment tests on real data.
Targets of gene expression regulators such as transcription factors (TFs) and microRNAs (miRNAs) can also be treated as gene sets. TFs are regulatory proteins that bind to the promoter regions of target genes to regulate their levels of expression (*Chen et al.,* 2007; *Hobert et al.,* 2008; *Martinez et al.,* 2009). miRNAs are small (~22-nucleotides) non-coding RNAs that are incorporated into the RNA-induced silencing complex (RISC) to regulate the stability and translation of messenger RNA (mRNA) transcripts (*Bartel*, 2009; *Naeem et al.,* 2010).
The activity of such regulators is not directly visible on the mRNA level: TFs are frequently modulated at the post-transcriptional level (*Boorsma et al.,* 2008) and miRNAs are usually not profiled. It is thus important to indirectly determine the activity of regulators by analyzing their target genes (*Cheng et al.,* 2007; *Hu et al.,* 2010). Here, the same tests are employed that were devised for analyzing biological processes (*Cheng et al.,* 2009; *Naeem et al.,* 2011).
We propose TFs and miRNAs and their associated sets of target genes for the rigorous evaluation of gene set enrichment tests since their experimental manipulation offers the required standard of truth that is not available for biological processes. Given regulator deletion or over-expression experiments, we considered the experimentally manipulated and the remaining regulators (along with their corresponding sets of target genes) as positives and negatives, respectively. We thus evaluated the ability of statistical tests to infer the manipulated regulator from the expression of its target genes.
The present study thereby conducts the first large comparison and rigorous assessment of 12 statistical enrichment tests for analyzing gene sets. We applied start-of-the-art statistical methods such as

ANOVA, Wilcoxon's test, Kolmogorov-Smirnov test as well as the hypergeometric test to test the null hypothesis, i.e. that expression changes in regulator target sets might be due to noise.
The following section briefly reviews the field of gene set enrichment tests. Subsequently, we describe our approach to rank enrichment tests.

Related work: Gene set analyses. A microarray experiment typically results in a long list of differentially expressed genes (DEGs) that is the starting point to gain insights into biological mechanisms (*Gatti et al.,* 2010). Several statistical methods for the analysis of sets of DEGs have been proposed (*Goeman et al.,* 2007, *Rivals et al.,* 2007). Most test for the over-representation of predefined sets of genes (e.g., Gene ontology (GO), KEGG pathways) in the list of DEGs (*Hosack et al.,* 2003; *Zeeberg et al.,* 2003; *Zhang et al.,* 2004; *Martin et al.,* 2004; *Al-Shahrour et al.,* 2004, *Beissbarth et al.,* 2004; *Lee et al.,* 2005; *Pehkonen et al.,* 2005; *Khatri et al.,* 2005; *Yi et al.,* 2006).
Pavlidis *et al.* (2004) use the geometric mean to calculate the significance of the genes in the gene set. Gene Set Enrichment (GSE) analysis, proposed by *Mootha et al.,* (2003) and improved by *Subramanian et al.,* (2005) uses an enrichment score based on a Kolmogorov-Smirnov test statistic. GSEA has been extended (*Barry et al.,* 2005; *Huang et al.,* 2009) to cover multiclass, continuous and survival phenotypes, and more test statistics.
More recently, GSE tests have also been applied to gene sets representing TF or miRNA target genes. *Sohler et al.,* (2005), *Liu et al.,* (2010) and *Essaghir et al.,* (2010) identified the activity of TFs by analyzing whether the TF target gene sets are enriched among a list of DEGs using a hypergeometric test. GSE tests were also applied to detect expression changes of miRNAs based on the expression of their target gene set (*Farh et al.,* 2005; *Sood et al.,* 2006; *Arora et al.,* 2008; *Tu et al.,* 2009; *Volinia et al.,* 2010; *Ott et al.,* 2011). Recently *Cheng et al.,* 2009 proposed a test statistic based on difference of average ranks between the miRNA's non-targets and targets. None of these studies provide a comprehensive comparative analysis of the tests evaluated against real data.

Methods

Assessment of TF and miRNA activity

To determine activity changes of TFs and miRNAs we apply several gene set enrichment approaches to test the null hypothesis (H_o) whether the expression levels of regulator downstream targets could be sampled from the background distribution of the remaining (i.e. non-target) genes. Our approach to assess gene set enrichment tests is depicted in Figure 1. In the following, we describe how the standard of truth is derived and how sign annotations are used to treat the up- and down-regulation of target genes. The applied enrichment approaches are described in Appendix 2.

Standard of truth. In the proposed assessment scenario, we evaluate the ability of statistical tests to infer the experimentally manipulated (i.e. deleted or over-expressed) regulators from the expression of its target genes (see Appendix 1). Thus, the identity of the manipulated regulators represents the standard of truth. It is compiled into a label matrix that assigns 1 if the given regulator is manipulated (deleted or over-expressed) in the given measurement or 0 otherwise (Figure 1).
Some TFs are excluded from the assessment. We exclude manipulated TFs that exhibit fold-changes smaller than a predefined threshold: here, it is unclear whether the manipulation was effective. We also exclude TFs that exhibit large fold-changes but have not been directly manipulated: they could be direct or indirect targets of a manipulated TF. By varying fold-change thresholds, the performance dependency on the definition of positives can be explored. Since the expression levels of miRNAs have not been measured, all miRNAs are used to determine the performance of tests.

Pre-processing of the data matrix

Before applying enrichment tests, the given gene expression measurements need to be pre-processed, either utilizing or neglecting interaction signs, i.e. that a TF activates (+) or inhibits (-) a given target**.**

Absolute-one-sided (H_o^{abs}). In this scenario, interaction signs are ignored. Enrichment tests are applied to absolute log fold changes, i.e. we evaluate the degree of differential expression in the target genes regardless of up- and down-regulation (Figure 2, left).

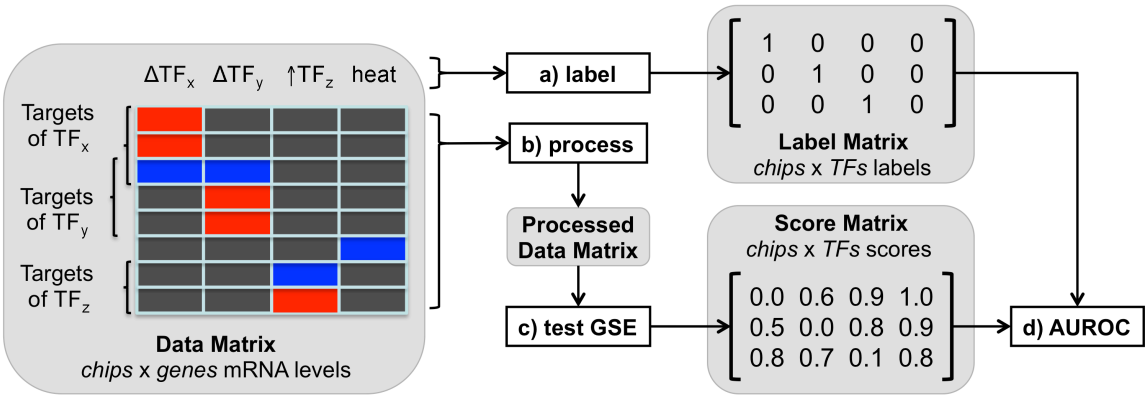


Figure 1: Overview. The data matrix consists of |*chips*| columns and |*genes*| rows where cells in the matrix represent mRNA levels. Chips are annotated by the treatment, e.g. the manipulation (Δ=deletion or ↑=over-expression) of expression regulators, here exemplified by TFs. This annotation is compiled (a) into a label matrix to represent the standard of truth. Manipulation of a regulator is expected to result in up- (red) or down-regulation (blue) of its target genes. After processing (b) the data matrix as shown in Figure 2, gene set enrichment (GSE) tests are applied (c) to determine the activity of regulators based on the differential expression of their target genes (see Appendix 1). This results in a score matrix containing the test results. For evaluation, the label matrix is compared to the score matrix to compute (d) an area under the receiver-operator characteristic (AUROC) curve.

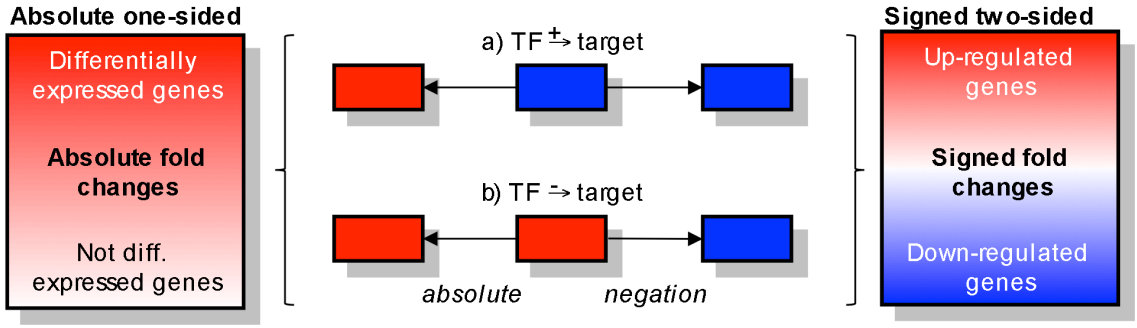


Figure 2: Preprocessing of the data matrix. Two null hypotheses, the absolute (H_0^{abs}, left) and the signed two-sided (H_0^{sign}, right) null hypotheses, can be tested after pre-processing the data matrix accordingly. For $\boldsymbol{H_0^{abs}}$, expression profiles are transformed into absolute log fold changes. For *E.coli*, where interactions are annotated as '+' for activation and '-' for inhibition, we also test $\boldsymbol{H_0^{sign}}$. Here, we negate fold changes for target genes that are inhibited by the regulator. Two-sided tests detect positive or negative fold-changes corresponding to an increase or decrease, respectively, in regulator activity.

Signed-two-sided (H_o^{sign}). This scenario can only be applied to *E. coli* since only RegulonDB provides sign annotations for gene regulatory interactions. We negate fold changes for target genes that are inhibited by the given regulator. Thus, all target genes of a regulator should either exhibit enrichment of positive or negative fold changes in case of increased or decreased, respectively, regulator activity. Enrichment at either tail of the distribution is then determined by two-sided tests (Figure 2, right).

Performance assessment

Statistical tests as described in Appendix 2 are applied to the processed data matrix (Fig. 2). Test predictions are then evaluated against the standard of truth via the area under the receiver-operating characteristic (AUROC) as discussed in (*Prill et al.,* 2010). The AUROC compares continuous test scores (Fig. 1: score matrix) against discrete regulator states (1=active, 0=inactive, compare Fig. 1: label matrix). Thus, AUROC is a summary measure of the test's ability to consistently assign higher scores to active regulators and lower scores to non-active regulators based on given chip measurements. AUROC's of 1 or 0.5 represent a perfect or random test, respectively.

Randomized testing. In addition to applying the tests to the data matrix, we also progressively randomized the set of regulator target genes to evaluate how much the performance of statistical

methods depends on the quality of gold standards. We generate new target sets that are randomized by x% (where x=25, 50, 75…), i.e. by randomly selecting x% of the interactions in the gold standard and exchanging the true target gene in such an interaction by a random non-target gene. An average AUROC is determined by applying GSE tests on 100 partially randomized networks for each x.

Results

Detection of TF activity without sign annotations. We first evaluated the ability of the applied enrichment tests to predict TFs that have been deleted or over-expressed. At this point, sign annotations are ignored, i.e. we test H_0^{abs}. Manipulations were only considered effective if the TFs exhibit a fold change of at least two or less than 0.5. Conversely, substantial fold-changes in non-manipulated (secondary) TFs could be due to a direct or indirect effect from the manipulated (primary TFs). Such cases are also excluded from the evaluation. In case of negative examples, we varied the fold-change cutoff to explore its influence on the performance of the enrichment tests (Figure 3). At a higher cutoff, more negative examples are included in the analysis, which leads to a slightly decreased performance but hardly influences the ranking of enrichment tests. The resulting AUROC values at a cutoff of 0.5 are shown in Table 1.
In addition, we also combined all individual tests into a consensus. The scores in the individual score matrices (Figure 1) are transformed into ranks and averaged. Although some of the constituent tests hardly perform better than random, the consensus shows consistently good results arcoss all scenarios.

Detection of TF activity with sign annotations. This section evaluates if test performance can be improved by exploiting the annotation provided by RegulonDB. This annotation distinguishes whether the TF activates or inhibits a given target gene. H_0^{abs} as applied in the previous section tested only for differential expression. By using H_0^{sign} instead, we additionally test whether the fold changes observed in TF targets are consistent with the given interaction sign annotations. Neglecting signs slightly improves the enrichment tests without significantly changing their ranks (Table 1).

Test performance on E. coli vs. S. cerevisiae. In addition, we also applied the enrichment tests to expression compendia in *S. cerevisiae.* The overall ranking of tests is very consistent between prokaryotic and eukaryotic datasets. The performance for *S. cerevisiae* is somewhat lower than that for *E. coli*. These results might be due to the better quality of gene regulatory networks in *E. coli* as well as the simpler gene regulation in prokaryotes (*Narendra et al.,* 2010).

Detection of miRNA activity. In addition to TF-target relationships we also evaluated miRNA target relationships based on miRNA transfection experiments in human cell lines.. Here, a range of miRNA-target set definitions has been employed: databases only (AUROC ANOVA 0.63), DBs+PICTAR+TargetScan (high precision prediction tools, AUROC ANOVA 0.83) and DBs+PITA (high recall prediction tool, AUROC ANOVA 0.84). Although the quality of computational miRNA target predictions has been discussed controversially (*Ritchie et al.* 2009), they are required to complement the currently available manual repositories, which appear to be not sufficiently comprehensive for such an analysis. Although this setting deviates considerably from the previously discussed ones, the overall ranking of methods is again very consistent (see Appendix 3). An exception is the hypergeometric test (HG-0.5) with the second best performance after ANOVA.

Randomized testing. To determine how the test performance depends on the quality of the available gene regulatory networks, we progressively randomized the regulator target sets. The ability of the different tests to infer the activity of TFs is surprisingly stable even if, on average, about 50% of the gene regulatory network is randomized (Figure 4, Appendix 3 for miRNAs).

Overall ranking of methods. Average ranks for the examined tests were computed based on their performance across different partially randomized expression compendia (*E. coli*, *S. cerevisiae* and human) and different scenarios (H_0^{abs} vs. H_0^{sign}). Thereby, we derived the following ordering of methods: ANOVA > CON > WR > HG-0.5 >FCR > KS > BT > FC > MED > HG-1.0 > HG-1.5 > FCRW. ANOVA, CON (consensus) and WR (Wilcoxons rank) performed consistently well across all

scenarios (average ranks between 1 and 3). While FCR (fold change rank), HG-0.5 (hypergeometric, with threshold 0.5), KS (Kolmogorov-Smirnov) and BT (bootstrapping) also deliver usable results (average ranks between 5 and 7), the remaining tests performed substantially below average.

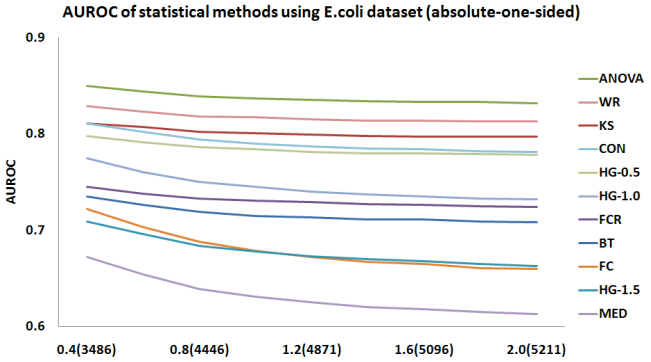


Figure 3: Dependency of AUROC on the set of negatives (*E.coli*). TFs are only considered as negatives in the AUROC analysis if they exhibit fold-changes of less than a pre-defined cutoff. The x-axis shows the sizes of different negative sets (in brackets) compiled based on different fold-change cutoffs ($|log_2$ (*fold change*)$|$). The size of the negative sets has only little influence on the AUROC (y-axis) or on the relative rank of the different enrichment tests.

Table 1: AUROC for enrichment tests across *E.coli* and yeast expression compendia.

Enrichment tests		*E.coli*		Yeast	
		(H_o^{abs})	(H_o^{sig})	Y1-(H_o^{abs})	Y2-(H_o^{abs})
ANOVA	Two-sample ANOVA≡t-test	0.85	0.80	0.71	0.71
CON	Consensus of all tests	0.81	0.77	0.74	0.68
WR	Wilcoxon's rank sum	0.83	0.72	0.72	0.68
KS	Kolmogorov-Smirnov	0.81	0.83	0.64	0.63
BT	Bootstrapping	0.72	0.68	0.73	0.67
HG-0.5	Hypergeometric, cut=0.5	0.79	0.60	0.69	0.60
FC	Average fold change	0.72	0.72	0.75	0.67
HG-1.0	Hypergeometric, cut=1.0	0.77	0.67	0.68	0.54
HG-1.5	Hypergeometric, cut=1.5	0.71	0.69	0.72	0.56
FCR	Average gene rank	0.75	0.65	0.71	0.68
MED	Median	0.67	0.67	0.70	0.66
FCRW	Avg. fold change rank weight	0.56	0.54	0.56	0.70

Appendix 2 describes the used tests. The table is sorted based on the average of the four AUROC's.

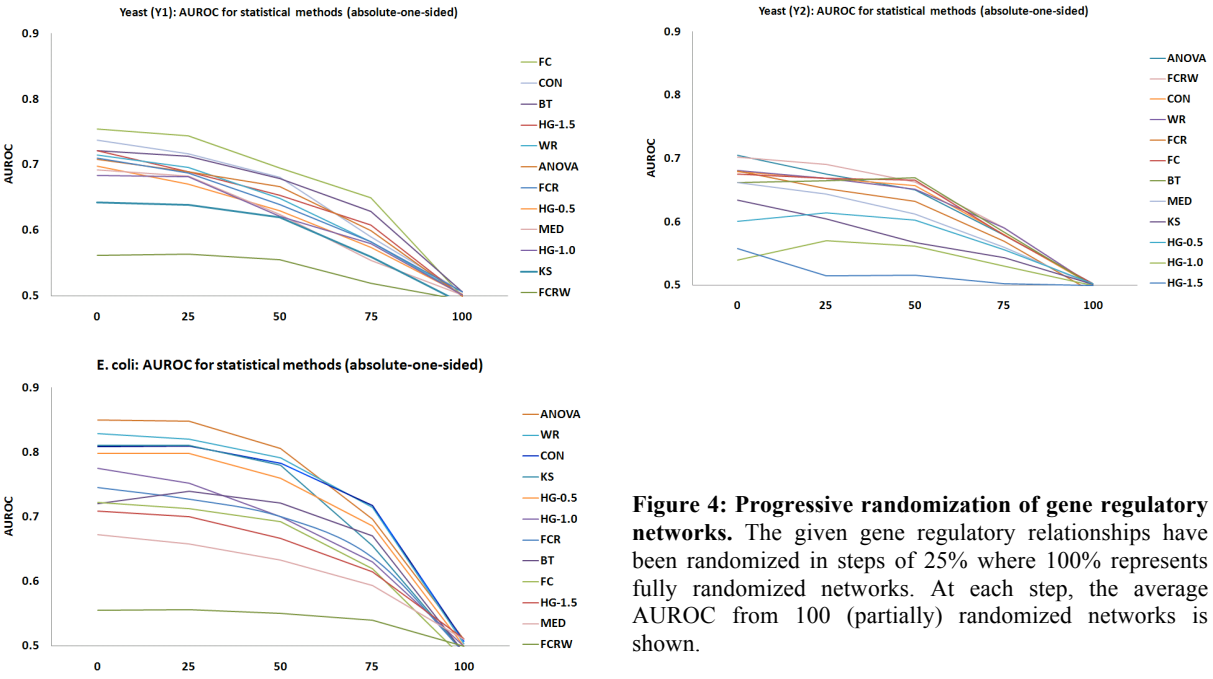


Figure 4: Progressive randomization of gene regulatory networks. The given gene regulatory relationships have been randomized in steps of 25% where 100% represents fully randomized networks. At each step, the average AUROC from 100 (partially) randomized networks is shown.

Discussion and Conclusion

Gene set enrichment tests have been devised to detect an over-representation of differentially expressed genes in pre-defined gene sets that correspond to biological processes. A dependable standard of truth is not available since it is difficult to decide *a priori*, which biological processes will be affected on the mRNA level. This has previously prevented the objective selection and evaluation of enrichment tests on real measurements. Instead, we derived gene sets from the targets of gene expression regulators (TFs and miRNAs) whose experimental manipulation directly offers the required standard of truth. In this setting, we evaluated the ability of 12 frequently used statistical tests (*Huang et al.,* 2009) to detect regulator manipulations.

The detection of regulator activities is difficult: simple tests based on the rank difference between regulator targets and non-targets are not appropriate. We observe that the hypergeometric and Kolmogorov-Smirnov tests, which are most frequently used in practice, are significantly outperformed by ANOVA and Wilcoxons test. The HG test yielded mixed results depending on threshold parameter and setting (TF vs. miRNA). Although the performance of the used tests was diverse (AUC between 0.5 and 0.85 for *E. coli*), an unweighted consensus integrating all approaches consistently showed good results.

Surprisingly, test performance did not improve by utilizing interaction signs (activate vs. inhibit). This might be due incomplete sign annotations, e.g. with respect to toggle switches (Morel et al., 2000) where a TF can activate or inhibit a target gene depending on the molecular context.

To ensure the wide applicability of our results, we employed a variety of settings. In terms of microarray data, we used TF manipulations in *E. coli* (one expression compendium) and *S. cerevisiae* (two compendia) to compare results between a prokaryote and a eukaryote model organism. We also analyzed a third setting, the transfection of human cell lines with miRNAs. Performance on *S. cerevisiae* and human is lower than that for *E. coli*, which might be due to the lower quality of the available gold standards of TF/miRNA target networks and the more complex regulation in eukaryotes (*Michoel et al.,* 2009; *Hu et al.,* 2007; *Narendra et al.,* 2010).

The performance ranking of the tests is very consistent between each of the examined scenarios, with methods such as ANOVA or Wilcoxon's test always performing substantially better than random guessing. We thus expect that the ranking of the 12 tests will be meaningful in novel settings that deviate from the ones described here. An example is the application of enrichment tests to biological processes, where we expect the consensus approach to yield the most reliable results.

Via an additional permutation approach, we analyzed how enrichment tests depend on the quality and comprehensiveness of the known regulator-target relationships. Most methods show only a moderate decrease in performance even after randomizing 50% of the gene regulatory network. We therefore conclude that the gene set definitions derived from the known gene regulatory interactions are sufficient to enable the comparative assessment of enrichment tests as well as the detection of regulator activities in real mRNA expression compendia.

References

1. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 2004, 20(4):578-80.
2. Arora A, Simpson DA (2008) Individual mRNA expression profiles reveal the effects of specific microRNAs. *Genome Biol* 9(5):R82.
3. Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* 2005, 21:1943-1949.
4. Barry WT, Nobel AB, Wright FA: A statistical framework for testing functional categories in microarray data. *Annals of Applied Statistics* 2008, 2:286-315.
5. Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell* 2009, 136(2):215-233.
6. Beissbarth T, Speed TP. GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics* 2004, 20:1464-1465
7. Boorsma A, Lu XJ, Zakrzewska A, Klis FM, Bussemaker HJ. Inferring condition-specific modulation of transcription factor activity in yeast through regulon-based analysis of genomewide expression. *PLoS One* 2008, 3(9):e3112.

8. Chen K, Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet* 2007, 8(2):93-103.
9. Cheng C, Fu X, Alves P, Gerstein M. mRNA expression profiles show differential regulatory effects of microRNAs between estrogen receptor-positive and estrogen receptor-negative breast cancer. *Genome Biol.* 2009, 10(9):R90.
10. Davis J, Goadrich M. The Relationship Between Precision-Recall and ROC Curves. In ICML '06: Proceedings of the 23rd international conference on Machine learning 2006, p.240.
11. Essaghir A, Toffalini F, Knoops L, Kallin A, van Helden J, et al. Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. *Nucleic Acids Res* 2010, 38:e120.
12. Farh KK, Grimson A, Jan C, Lewis BP, Johnston WK, et al. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* 2005, 10(5755):1817-21.
13. Gatti DM, Barry WT, Nobel AB, Rusyn I, Wright FA. Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics* 2010, 11:574.
14. Goeman JJ, Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 2007, 23(8): 980-7.
15. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, et al. Gene Ontology Consortium: The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004, 32:D258-61.
16. Hobert O: Gene regulation by transcription factors and microRNAs. *Science* 2008, 319:1785-1786.
17. Hosack DA,et al. Identifying biological themes within lists of genes with EASE. *Genome Biol.* 2003, 4:R70.
18. Hu H. An efficient algorithm to identify coordinately activated transcription factors. *Genomics* 2010, 95(3):143-50.
19. Hu Z, Killion PJ, Iyer VR. Genetic reconstruction of a functional transcriptional regulatory network. *Nat Genet.* 2007, 39(5):683-7.
20. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009, 37(1):1-13.
21. Khatri P, Drăghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 2005, 21:3587-3595.
22. Lee HK, Braynen W, Keshav K, Pavlidis P: ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics* 2005, 6:269.
23. Liu Q, Tan Y, Huang T, Ding G, Tu Z, Liu L, Li Y, Dai H, Xie L. TF-centered downstream gene set enrichment analysis: Inference of causal regulators by integrating TF-DNA interactions and protein post-translational modifications information. *BMC Bioinformatics* 2010, 11 Suppl 11:S5.
24. Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B. GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol.* 2004, 5(12):R101.
25. Martinez NJ, Walhout AJ. The interplay between transcription factors and microRNAs in genome-scale regulatory networks. *Bioessays* 2009, 31(4):435-45.
26. Michoel T, De Smet R, Joshi A, Van de Peer Y, Marchal K. Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. *BMC Syst. Biol.* 2009, 3:49.
27. Mootha VK,et al. PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet* 2003, 34:267-273.
28. Morel V, Schweisguth F. Repression by suppressor of hairless and activation by Notch are required to define a single row of single-minded expressing cells in the Drosophila embryo. *Genes Dev*. 2000, 14(3):377-88.
29. Naeem H, Küffner R, Csaba G, Zimmer R. (2010). miRSel: automated extraction of associations between microRNAs and genes from the biomedical literature. *BMC Bioinformatics*, 11:1-8.
30. Naeem H, Küffner R, Zimmer R. MIRTFnet: Analysis of miRNA regulated transcription factors. *PLoS One* 6(8): e22519. doi:10.1371/journal.pone.0022519.
31. Narendra V, Lytkin NI, Aliferis CF, Statnikov A. A comprehensive assessment of methods for de-novo reverse-engineering of genome-scale regulatory networks. *Genomics* 2011, 97(1):7-18.
32. Ott CE, Grünhagen J, Jäger M, Horbelt D, Schwill S, et al. MicroRNAs differentially expressed in postnatal aortic development downregulate elastin via 3' UTR and coding-sequence binding sites. *PLoS One* 2011, 6(1):e16250.
33. Pavlidis P, Qin J, Arango V, Mann JJ, Sibille E. Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem Res.* 2004, 29(6):1213-22.
34. Pehkonen P, et al. Theme discovery from gene lists for identification and viewing of multiple functional groups. *BMC Bioinformatics* 2005, 6:162.

35. Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, Clarke ND, Altan-Bonnet G, Stolovitzky G. Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS One* 2010, 5, e9202.
36. Ritchie W, Flamant S, Rasko JE: Predicting microRNA targets and functions: traps for the unwary. *Nat Methods* 2009, 6(6):397-8.
37. Rivals I, Personnaz L, Taing L, Potier MC. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* 2007, 23(4):401-7.
38. Sohler F, Zimmer R. Identifying active transcription factors and kinases from expression data using pathway queries. *Bioinformatics* 2005, 21:115-122.
39. Sood P, Krek A, Zavolan M, Macino G, Rajewsky N. Cell-type-specific signatures of microRNAs on target mRNA expression. *Proc Natl Acad Sci USA* 2006, 103(8):2746-51.
40. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005, 102:15545-15550.
41. Tu K, Yu H, Hua YJ, Li YY, Liu L, Xie L, Li YX: Combinatorial network of primary and secondary microRNA-driven regulatory mechanisms. *Nucleic Acids Res* 2009, 37:5969-5980.
42. Volinia S, Visone R, Galasso M, Rossi E, Croce CM. Identification of microRNA activity by Targets' Reverse EXpression. *Bioinformatics* 2010, 26(1):91-7.
43. Yi M,et al. Wholepathwayscope: a comprehensive pathway-based analysis tool for high-throughput data. *BMC Bioinformatics* 2006, 7:30.
44. Zeeberg BR, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* 2003. 4:R28.
45. Zhang B, et al. GO Tree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics* 2004, 5:16.
46. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem. 1993 Apr;39(4):561-77.

Appendix 1: Datasets

Gene expression compendia

TF deletion and over-expression studies. To investigate the influence of TFs on downstream target genes, we analyze large-scale experiments that in part consist of TF perturbations (knock-out=KO and over-expression=OE, Appendix table 1). A compendium of 907 *E.coli* microarray samples was taken from the M3D Database (*Faith et al.,* 2008). Another compendium of 263 *S. cerevisiae* microarrays was obtained from the study of *Hu et al.,* 2007. *Hu et al.* systematically deleted 263 TFs in yeast, and compared each deletion strain with the wild type for genome-wide expression. We also used the dataset by *Chua et al.,* (2006) providing microarray expression data resulting from over-expression and/or deletion of 55 *S. cerevisiae* TFs.

All analyses are based on comparing gene expression levels between deletion/over-expression and control via $\log_2$ fold-changes. The microarray datasets contain basal gene levels that can be quite different between experiments. To compensate for this, we transformed the absolute expression values into expression fold changes. Fold changes are computed by mapping each measured condition to one or more control conditions from the same experiment (see *Küffner et al.,* (2011) "Inferring gene regulatory networks by ANOVA", submitted to Bioinformatics.

Appendix Table 1: *E. coli* and yeast expression compendia used in this study.

Dataset	TFs	Targets	KO/OE TFs	Targets	Chips	References
***E.coli* (M3D)**	167	1377	17	949	907	*Faith et al.,* 2008
S. cerevisiae (Y1)	114	1934	102	1527	263	*Hu et al.,* 2007
S. cerevisiae (Y2)	114	1934	48	1094	270	*Chua et al.,* 2006

miRNAs transfection studies. Several microarray experiments with over-expression of miRNAs have been performed to measure the global changes in the transcriptome or proteome. We obtained 43 gene expression profiles of 18 different miRNA transfection studies in different human cell lines. Selbach *et al.* (2008) measured gene expression data in HeLa cells at 8h and 32h after miRNA over-expression of miR-155, miR-16 and let-7b. Expression profiles by He *et al.* (2007) include gene expression changes at 24h after miRNA over-expression of miR-34 family (i.e., miR-34a and miR-34b), in six

different cell lines (e.g., HeLa, A549 H1-term and TOV21G H1-term). Georges *et al.* (2008) measured p53-inducible miRNAs, miR-192 and miR-215, at 10h and 24h after miRNA transfection in a human cell line (i.e., HCT116 Dicer -/- #2). Baek *et al.* (2006) measured the gene expression data in HeLa cells at 24h after miR-124, miR-1 and miR-181a transfection. We also use the dataset by Grimson *et al.* (2007) that measured gene expression data in HeLa cells at 12h and 24h after miRNA over-expression of miR-7, miR-9, miR-122, miR-128, miR-132, miR-133, miR-142 and miR-181a.

Gene regulatory networks

TF-gene regulatory interactions. *E.coli* TF-gene regulatory interactions were obtained from RegulonDB (*Gama-Castro et al.,* 2008). RegulonDB contains 2066 experimentally validated and manually curated interactions. Recently, DREAM5 used RegulonDB to validate the predicted *E.coli* interactions (wiki.c2b2.columbia.edu/dream/index.php/D5c4). The Saccharomyces cerevisiae (*S. cerevisiae*) gold-standard network of 3940 interactions was obtained from the study of *MacIsaac et al.,* 2006 who re-analyzed the *Harbison et al.,* 2004 ChIP-chip data to determine the binding locations of TFs. The *E.coli* gold standard is considered more reliable than *S. cerevisiae* as suggested the analysis of *Narendra et al.* (2010).

miRNA-target gene associations. Several computational algorithms have been developed to predict miRNA-target genes. We obtained putative human miRNA-target pairs predicted by PITA (*Kertesz et al.,* 2007), PICTAR (*Krek et al.,* 2005) and TargetScan (Friedman et al., 2009). In addition, several databases collect target genes of the miRNAs in different organisms. Human miRNA-gene associations were obtained from the curated databases TarBase (*Papadopoulos et al.,* 2009), miRecords (*Xiao et al.,* 2009) and miR2Disease (*Jiang et al.,* 2009). From miRSel (*Naeem et al.,* 2010) we obtained putative miRNA-gene associations and relations extracted from biomedical abstracts by text mining (Appendix table 2).

Appendix Table 2 – miRNA-target associations from databases (DB) and predictions (PR).

Source	miRNAs	Target genes	Pairs
DB: miRSel	486	1969	7604
DB: TarBase	110	837	1023
DB: MiRecords	93	614	772
DB: miR2Disease	176	364	596
PR: PITA	640	14065	307465
PR: PICTAR	163	5975	44403
PR: TargetScan	249	9446	110172

Appendix 2: Enrichment tests

Wilcoxons test. The Wilcoxon nonparametric rank-sum (WR, *Mann and Whitney,* 1947; *Lehmann,* 1975) method is applied to test the null hypothesis, i.e. regulator targets exhibit no significant rank differences in comparison to other (non-targets) genes. The ranks were derived by sorting the genes based on either their absolute or signed log fold changes (Figure 2). If the rank distributions of targets and non-targets of the tested regulator are significantly different the null hypothesis will be rejected. We refer to such a TF/miRNA as active regulator for the tested experiment. The results of WR test statistic are p-values as a measure of significance of the observed change in means.

Hypergeometric test. Since hypergeometric (HG, *Spiegel,* 1992) test requires a threshold parameter to select regulated genes, we applied the HG test to test the null hypothesis given regulated gene sets of different sizes compiled based on absolute log fold changes, e.g. greater than 0.5, 1.0 or 1.5. For a given regulator *i* the *p*-value is computed according to the cumulative hypergeometric formula:

$$p_{HG} = 1 - \sum_{i=0}^{x} \binom{m}{i} \binom{N-m}{k-i} \Big/ \binom{N}{k}$$

where N is the population size or number of DEGs in a given chip measurement; m is the number of success in population or a set of DEGs filtered based on a given regulated gene threshold value; k is the the number of regulator target genes, and x is the number of common DEGs between m and k.

Kolmogorov-Smirnov test. The results of Kolmogorov-Smirnov test (KS, *Nikiforov,* 1994; *Siegel,* 1956) statistic are p-values as measure of differences between the empirical distribution (cumulative distribution) functions of a regulator targets and non-targets.

ANOVA. The results for ANOVA test are p-values that are calculated using F statistic/distribution. In a given setting two sample-ANOVA is equivalent to the two-sample t-test (*Miller,* 1997).

Bootstrap sampling. The significance probability of the bootstrap test (BT, *Efron et al.,* 1993) calculates the statistic for bootstrap two samples (such as regulator targets and non-targets gene set) drawn in some way (randomly) from the original data, and then calculates the proportion of these that are less than or equal to lower tail, greater than or equal to upper tail, or either (two tail). Bootstrap results in p-values as a measure of significance to the difference in means using ANOVA/t-test.

Average Fold Change. The Average Fold change (FC-score) of a regulator activity is defined as the difference of the average mean expression levels between its targets and non-targets. A positive FC score indicates that the target genes of a regulator are expressed at higher levels than non-target genes. The higher the FC score, the stronger the effect of a regulator on its targets (*Cheng et al.,* 2009).

Average gene rank. The average gene rank (FCR-score) of a regulator activity is defined as the difference of the average rank between its targets (T_{ravg}) and non-targets (nT_{ravg}). The genes ranks (FCR) were derived by sorting them based on their absolute or signed fold changes.

$$FCR = \frac{1}{n}\sum_{i=1}^{n} t_i - \frac{1}{j}\sum_{i=1}^{j} t_j$$

where n and j represent the number of a given regulator targets and non-targets. And t_i and t_j represent the ranks of regulator targets and non-targets.

Average fold change rank weight. The average fold change rank weight (FCRW-score) of a regulator activity is defined as the difference of the combined average rank and expression levels between its targets and non targets. The ranks of genes are derived by sorting them based on their absolute or signed fold changes (Figure 2).

$$FCRW = \frac{\sum_{i=1}^{n} w_i t_i}{\sum_{i=1}^{n} w_i} - \frac{\sum_{i=1}^{j} w_j t_j}{\sum_{i=1}^{j} w_j}$$

where w_i and w_j are the ranks and t_i and t_j are the fold changes of targets and non-targets, respectively.

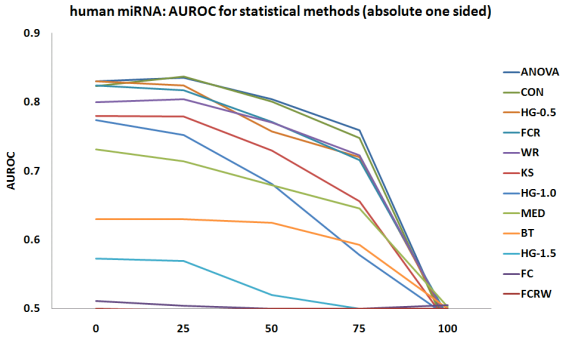
Median. The median (MED-score) of a regulator activity is defined as the difference of the median expression levels between its targets and non-targets.

Consensus prediction. A number of tests have been applied to a TF in a given experiment to test for over-representation of its targets among the DEGs. For each test, ranks of the regulators are determined by sorting them based on their scores. We define a consensus score (CON) to measure the regulator activity changes: the unweighted average of the ranks of a regulator determined by other statistical methods/tests as described above. This approach is called Borda count voting (Borda, 1781). For a given regulator j, the consensus score is calculated as:

$$CON = \frac{1}{n}\sum_{i=1}^{n} R_{ji}$$

where n represents the number of tests applied to calculate the significance of a regulator j in a given experiment. Thus, R_{ji} represents the rank of a regulator j for a given statistical test i.

Appendix 3: Additional results for miRNAs



Appendix Figure 1: Progressive randomization of miRNA-target gene regulatory networks. The regulatory relationships from databases and computational predictions (TargetScan and PicTar) have been randomized in steps of 25% where 100% represents fully randomized networks. At each step, the average AUROC from 100 (partially) randomized networks is shown. 43 single miRNA transfection experiments have been performed (43 positives), based 18 unique miRNAs. Thus, the negative set contains (18-1)*43=731 samples.

Appendix Table 3: AUROC for enrichment tests for human miRNA-target gene sets from databases and predictions.

Statistical Methods/ Tests	Human miRNAs			
	(H_o^{abs}) – Databases (P1)	(H_o^{abs}) – P1 + PICTAR+ TargetScan (P2)	(H_o^{abs}) P2+PITA(-20) (P3)	Average AUROC
ANOVA	0.63	0.83	0.84	0.77
HG-0.5	0.61	0.83	0.81	0.75
CON	0.61	0.82	0.80	0.74
FCR	0.61	0.82	0.75	0.73
WR	0.60	0.80	0.77	0.72
KS	0.60	0.78	0.76	0.71
HG-1.0	0.61	0.77	0.72	0.70
MED	0.61	0.73	0.68	0.67
BT	0.62	0.62	0.66	0.63
HG-1.5	0.59	0.57	0.50	0.55
FC	0.51	0.51	0.51	0.51
FCRW	0.50	0.51	0.50	0.50

Statistical methods are applied to test the null hypothesis (H_o^{abs}) given miRNA-target gene set derived from databases (miRSel, TarBase, miRecords) and computational prediction programs (PICTAR, TargetScan and PITA) following different test settings as described in method section. In total 50 (P1), 260 (P2) and 649 (P3) miRNAs have been evaluated in all miRNA transfection experiments. For AUROC analysis, the positive set includes those miRNAs that are used for transfection in a given experiment and contain more than 20 targets (which are 26 and 43 in case of databases and computational prediction methods).

Appendix Table 4: Performance-based ranking of enrichment tests.

	Ranking based on ≥10% permutation results Avg. of 4 test cases (*E.coli* H_o^{abs}+ 2x*Yeast* H_o^{abs}+*E. coli* H_o^{sign})	Ranking without permutation Avg. of 4 test cases (*E.coli* H_o^{abs}+ 2x*Yeast* H_o^{abs}+*E. coli* H_o^{sign})	Ranking without permutation Avg. of 3 test cases (human miRNAs: P1, P2, P3)	Ranking based on 25% permutation Avg. of 3 test cases (miRNAs: P1, P2, P3)
ANOVA	2	2	1	1
CON	1	1	3	2
WR	3	3	4	3
HG-0.5	8	9	2	4
FCR	6	7	6	5
KS	9	4	5	6
BT	4	6	9	9
FC	5	5	11	11
MED	10	8	8	8
HG-1.0	11	11	7	7
HG-1.5	7	10	10	10
FCRW	12	12	12	12

Appendix References

1. Baek D, Villén J, Shin C, Camargo FD, Gygi SP, et al. (2006) The impact of microRNAs on protein output. Nature 455:64-71.
2. Borda, J. Memoire sur les elections au scrutin. Histoire de l'Academie des Sciences, Paris (1781).
3. Cheng C, Fu X, Alves P, Gerstein M. mRNA expression profiles show differential regulatory effects of microRNAs between estrogen receptor-positive and estrogen receptor-negative breast cancer. Genome Biol. 2009;10(9):R90. Epub 2009 Sep 1.
4. Chua G, Morris QD, Sopko R, Robinson MD, Ryan O, Chan ET, Frey BJ, Andrews BJ, Boone C, Hughes TR.Source. Identifying transcription factor functions and targets by phenotypic activation.Proc Natl Acad Sci U S A. 2006 Aug 8;103(32):12045-50. Epub 2006 Jul 31.
5. Efron,B. and Tibshirani,R.J.(1993). An Introduction to the Bootstrap. Chapman & Hall.
6. Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, Juhn FS, Schneider SJ, Gardner TS.Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata.Nucleic Acids Res. 2008 Jan;36(Database issue):D866-70. Epub 2007 Oct 11.
7. Friedman RC, Farh KK, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 19:92-105.
8. Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muñiz-Rascado L, et al. RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units).Nucleic Acids Res. 2011 Jan;39(Database issue):D98-105. Epub 2010 Nov 4.
9. Georges SA, Biery MC, Kim SY, Schelter JM, Guo J, et al. (2008) Coordinated regulation of cell cycle transcripts by p53-Inducible microRNAs, miR-192 and miR-215. Cancer Res 68:10105-12.
10. Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, et al. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell 27:91-105.
11. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA. Transcriptional regulatory code of a eukaryotic genome. Nature. 2004 Sep 2;431(7004):99-104.
12. He L, He X, Lim LP, de Stanchina E, Xuan Z, et al. (2007) A microRNA component of the p53 tumour suppressor network. Nature 447:1130-1134.
13. Hu Z, Killion PJ, Iyer VR. Genetic reconstruction of a functional transcriptional regulatory network.Nat Genet. 2007 May;39(5):683-7. Epub 2007 Apr 8.
14. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, et al. (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res 37:D98-D104.
15. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E (2007) The role of site accessibility in microRNA target recognition. Nat Genet 39:1278-1284.
16. Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, et al. (2005) Combinatorial microRNA target predictions. Nat Genet 37:495-500.
17. Lehmann EL. (1975). Nonparametric Statistical Methods Based on Ranks. McGraw-Hill.
18. MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics. 2006 Mar 7;7:113.
19. Mann HB, Whitney DR. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics*, **18**(1):50-60.
20. Miller RG. (1997). Beyond ANOVA: Basics of Applied Statistics. Boca Raton, FL: Chapman & Hall.
21. Naeem H, Küffner R, Csaba G, Zimmer R. (2010). miRSel: automated extraction of associations between microRNAs and genes from the biomedical literature. BMC Bioinformatics, 11:1-8.
22. Narendra V, Lytkin NI, Aliferis CF, Statnikov A. A comprehensive assessment of methods for de-novo reverse-engineering of genome-scale regulatory networks. Genomics. 2011 Jan; 97(1):7-18. Epub 2010 Oct 14.
23. Nikiforov, A.M.(1994). Algorithm AS 288: Exact Smirnov two-sample tests for arbitrary distributions. Applied Statistics, Vol. 43, 265-284.
24. Papadopoulos GL, Reczko M, Simossis VA, Sethupathy P, Hatzigeorgiou AG (2009) The database of experimentally supported targets: a functional update of TarBase. Nucleic Acids Res 37:D155-D158.
25. R. Küffner et al. (2011) Inferring Gene Regulatory Networks by ANOVA, submitted to ISMB2011.
26. Selbach M, Schwanhäusser B, Thierfelder N, Fang Z, Khanin R, et al. (2008) Widespread changes in protein synthesis induced by microRNAs. Nature 455:58-63.
27. Siegel, S.(1956). Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill.
28. Spiegel MR. (1992). Theory and Problems of Probability and Statistics. McGraw-Hill, pp. 113-114.
29. Xiao F, Zuo Z, Cai G, Kang S, Gao X, et al. (2009) miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res 37:D105-D110.