ABSTRACT

Motivation: Several statistical 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: We present a first large study to rigorously assess and compare the performance of gene set enrichment tests on real expression measurements. Gene sets are defined based on the targets of given regulators such as transcription factors (TFs) and microRNAs (miRNAs). In contrast to processes, TFs and miRNAs are amenable to direct perturbations, e.g., regulator over-expression or deletion. We assess the ability of 14 different statistical tests to predict the perturbations from expression measurements in Escherichia coli, Saccharomyces cerevisiae and human. We also analyze how performance depends on the quality and comprehensiveness of the regulator targets via a permutation approach. We find that ANOVA and Wilcoxon's test consistently perform better than for instance Kolmogorov–Smirnov and hypergeometric tests. For scenarios where the optimal test is not known, we suggest to combine all evaluated tests into an unweighted consensus, which also performs well in our assessment. Our results provide a guide for the selection of existing tests as well as a basis for the development and assessment of novel tests.

Contact: robert.kueffner@bio.ifi.lmu.de

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on November 4, 2011; revised on March 3, 2012; accepted on April 2 2012

1 INTRODUCTION

The interpretation of gene expression studies is difficult because a large number of genes or other expressed sequences are profiled. Instead of individual genes, it has been proposed to analyze the expression of gene sets that correspond 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 gene sets can be performed by many different enrichment tests (Gatti et al., 2010; related work). These tests aim to detect gene sets exhibiting significant differential expression. However, it is not known a priori, which processes will be affected in a given expression experiment. This lack of a dependable standard-of-truth has prevented an objective selection and evaluation of enrichment tests on real data.

1.1 Related work: gene set analyses

Long lists of differentially expressed genes (DEGs) derived from microarray experiments are used as a starting point to gain biological insights (Gatti et al., 2010). Several statistical methods for the analysis of sets of DEGs have been proposed (reviewed by Ackermann and Strimmer, 2009; Goeman and Bühlmann, 2007; Nam and Kim, 2008; Rivals et al., 2007). Most test for the over-representation of predefined sets of genes (e.g., Gene Ontology) in the DEGs (Al-Shahrour et al., 2004; Khatri and Drăghici, 2005; Martin et al., 2004; Pavlidis et al., 2004; Pekkonen et al., 2005; Yi et al., 2006; Zeeberg et al., 2003).

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. GSE analysis has been extended (Barry et al., 2005; Barry et al., 2008;
Huang et al., 2009) to cover multiclass, continuous phenotypes and more test statistics.

More recently, GSE tests have also been applied to gene sets representing TF or miRNA TGs. Sohler and Zimmer, (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 (HG) test. GSE tests were also applied to detect expression changes of miRNAs based on the expression of their target gene sets (Farh et al., 2005; Ott et al., 2011; Sood et al., 2006; Tu et al., 2009). Recently, Cheng et al. (2009) proposed a test based on difference of ranks between the miRNA’s targets and the remaining genes. Levine et al. (2006), Efren and Tishbharani (2006), Nam and Kim (2008) as well as Ackermann and Strimmer (2009) rigorously and thoroughly evaluated the performance of different tests on simulated data. Only limited supporting evidence on real data was provided here as this would require manually curated gold standards. None of the mentioned studies provide a comprehensive and rigorous comparative evaluation of tests based on real data.

2 DATASETS AND METHODS

2.1 TF-gene regulatory interactions

To investigate the influence of TFs on downstream TGs, we use TF-gene regulatory interactions to analyze large microarray compendia (Table 1). To facilitate the reproducibility of results, we only used publicly available datasets and sources in this study. From RegulonDB (Gama-Castro et al., 2007) and 270 (Y2; Chua et al., 2006) microarrays perturbing 263 TFs and 1377 TGs. The Saccharomyces cerevisiae network of 3940 interactions between 114 TFs and 1934 TGs were obtained from Maichsaac et al. (2006). It is considered less reliable than the E.coli network as suggested by the analysis of Narendra et al. (2011).

2.2 TF deletion and over-expression compendia

As summarized in Table 1, a compendium of 907 E.coli microarrays were taken from the M3D Database (Faith et al., 2008). It included knock-out (KO) and over-expression (OE) experiments for 17 TFs targeting 949 genes. In case of S. cerevisiae, we analyzed two compendia of 263 (Y1; Hu et al., 2007) and 270 (Y2; Chu et al., 2006) microarrays perturbing 263 TFs and 55 TFs, respectively. In total, 102 TFs (targeting 1527 genes in Y1) and 48 TFs (targeting 1094 genes in Y2) were mapped to the known interactions. Basal gene levels in the datasets can be quite different between experiments. To compensate for this, we transform the absolute expression values into log2 fold-changes between deletion/over-expression and control. Fold-changes are computed by mapping a condition measuring a perturbed TF to one or more control conditions without the perturbation.

2.3 miRNA-target gene associations

Several computational algorithms have been developed to predict TGs of miRNAs. We obtained putative human miRNA-target pairs predicted by PITA (Kertesz et al., 2007), PICTAR (Krek et al., 2005) and TargetScan (Table 2). miRNAs were taken from the M3D Database (Faith et al., 2008). It included knock-out (KO) and over-expression (OE) experiments for 17 TFs targeting 949 genes. In case of S. cerevisiae, we analyzed two compendia of 263 (Y1; Hu et al., 2007) and 270 (Y2; Chu et al., 2006) microarrays perturbing 263 TFs and 55 TFs, respectively. In total, 102 TFs (targeting 1527 genes in Y1) and 48 TFs (targeting 1094 genes in Y2) were mapped to the known interactions. Basal gene levels in the datasets can be quite different between experiments. To compensate for this, we transform the absolute expression values into log2 fold-changes between deletion/over-expression and control. Fold-changes are computed by mapping a condition measuring a perturbed TF to one or more control conditions without the perturbation.

Table 1. E.coli and yeast compendia and interactions used in this study

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Interactions</th>
<th>mRNA compendia</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TFs</td>
<td>Targets</td>
</tr>
<tr>
<td>E.coli</td>
<td>167</td>
<td>1377</td>
</tr>
<tr>
<td>S.cerevisiae Y1</td>
<td>114</td>
<td>1934</td>
</tr>
<tr>
<td>S.cerevisiae Y2</td>
<td>114</td>
<td>1934</td>
</tr>
</tbody>
</table>

Table 2. miRNA-target pairs from databases (DB) and predictions (PR)

<table>
<thead>
<tr>
<th>Source</th>
<th>miRNAs</th>
<th>TGs</th>
<th>Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB: miRScI</td>
<td>486</td>
<td>1969</td>
<td>7604</td>
</tr>
<tr>
<td>DB: TarBase</td>
<td>110</td>
<td>837</td>
<td>1023</td>
</tr>
<tr>
<td>DB: MiRecords</td>
<td>93</td>
<td>614</td>
<td>772</td>
</tr>
<tr>
<td>DB: miR2Disease</td>
<td>476</td>
<td>364</td>
<td>596</td>
</tr>
<tr>
<td>PR: PITA</td>
<td>640</td>
<td>14065</td>
<td>30746</td>
</tr>
<tr>
<td>PR: PICTAR</td>
<td>163</td>
<td>5975</td>
<td>44403</td>
</tr>
<tr>
<td>PR: TargetScan</td>
<td>249</td>
<td>9446</td>
<td>10172</td>
</tr>
</tbody>
</table>

(Friedman et al., 2009). In addition, several databases collect TGs 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 (Nacem et al., 2010), we obtained putative miRNA–gene associations and relations extracted from biomedical abstracts by text mining (Table 2 and Supplementary Material 1).

2.4 miRNA transfection compendia

We obtained 43 gene expression profiles of 18 different miRNA transfection (i.e. over-expression) 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-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 -/-). Baek et al. (2006) measured the gene expression data in HeLa cells at 24h after miR-124, miR-1 and miR-181a transfection. Expression data were also measured by Grimsson et al. (2007) in HeLa cells at 12h and 24h after over-expression of miR-7, −9, −122, −128, −132, −133, −134 and −181a.

2.5 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 (H₀) 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 sections, we describe how the standard-of-truth is derived and how sign annotations are used to treat the up- and down-regulation of TGs.

In our assessment scenario, we evaluate the ability of statistical tests to infer an experimentally perturbed (i.e. deleted or over-expressed) regulator from the expression of its TG (see Section 2.9). Thus, the identities of the perturbed regulators represent the standard-of-truth. It is compiled into a label matrix that assigns 1 if the given regulator is perturbed in a given measurement or 0 otherwise (Fig. 1).

Some TFs are excluded from the assessment. We exclude TFs that exhibit fold-changes smaller than a predefined threshold: here, it is unclear whether the perturbation was effective. We also exclude TFs that exhibit large fold-changes but have not been directly perturbed: they could be direct or indirect targets of a perturbed TF. By varying fold-change thresholds (see Section 3), 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.

2.6 Preprocessing of the data matrix

Before applying enrichment tests, the given gene expression measurements can be preprocessed, so that the interaction signs are either utilized or neglected, i.e. that a TF activates (+) or inhibits (−) a given target.
2.8 Performance assessment and randomization

Enrichment tests are applied to the processed data matrix (Fig. 2). This results in one test score for each combination of a TF and a mRNA measurement. Scores are then evaluated against the standard-of-truth via the area under the receiver-operating characteristic (AUROC) as discussed in Phill et al. (2010). The AUROC compares continuous test scores (Fig. 1: Score Matrix) against discrete regulator states (1 = active, 0 = inactive; Fig. 1: Label Matrix). Thus, AUROC is a summary measure of the tests ability to consistently assign higher scores to active regulators and lower scores to non-active regulators based on given measurements. AUROCs of 1 or 0.5 represent a perfect or random test performance, respectively.

In addition to applying the tests to the data matrix, we also progressively randomize the set of regulator TGs to evaluate how much the performance of statistical methods depends on the quality of the known interactions. We generate new target sets that are randomized by x% (where x = 25, 50, 75 and 100), i.e. by randomly selecting x% of the available interactions and exchanging the true TG in such an interaction by a random non-TG. An average AUROC is determined by applying GSE tests on 100 randomized networks for each x.

2.9 Enrichment tests

The statistical tests below evaluate the null hypothesis that the two distributions of log2 fold-changes of targets and non-targets of a given regulator differ. The first category are over-representation tests [Kolmogorov-Smirnov (KS) and HG] that do not take the expression values or their ranks into account. The KS test (Nikitovor, 1994; Stiegel, 1956) estimates the maximum possible enrichment by varying the length of the list of genes. In contrast, the HG test (Stiegel, 1992) requires a threshold parameter to select DEGs. We select genes exhibiting fold-changes >0.5, 1.0 or 1.5 (see Supplementary Material 2 for additional thresholds). For a given regulator, i.e. P-value is computed via:

\[ P_{HG} = 1 - \sum_{i=0}^{N-1} \binom{m}{i} \binom{N-m}{N-i} \binom{N}{i} \]

where N is the population size or number of DEGs in a given chip measurement, m is the number of successes in a population or a set of DEGs filtered based on a given regulated gene threshold value; i is the number of regulator TGs, and x is the number of common DEGs in m and i.

ANOVA and the bootstrap test (BT) are evaluated on the gene expression values. In the given setting, the two sample-ANOVA are equivalent to the t-test (Miller, 1997). BT (Efron and Tibshirani, 1993) calculates the statistic for two bootstrap samples drawn randomly from the original data of regulator targets and non-targets, and then calculates the proportion of these that are less than or equal to the lower tail, greater than or equal to the upper tail, or either (two tails). Bootstrap is based on the difference in means measured by the t-test. In contrast to ANOVA and BT, the null hypothesis of the Wilcoxon non-parametric rank-sum method [WR; Mann and Whitney, 1947; Lehmann, 1975] tests whether regulator targets exhibit no significant rank differences in comparison to other (non-targets) genes. The ranks are derived by sorting the genes based on either their absolute or signed log fold-changes (Fig. 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. In contrast to KS and HG, expression values or their ranks are used by ANOVA, BT and WR. The latter three tests can thus effectively take the noise levels of genes into account. The test statistics of all tests discussed so far result in P-values, i.e. the probability that the observed differences between distributions are due to chance.

We further analyze several tests calculated based on fold-changes. The average fold-change (FC-score) of a regulator is defined as the difference between the average mean expression levels between its targets and non-targets. A positive FC-score indicates that the TGs of a regulator are expressed at higher levels than non-TGs. The higher is the FC-score, the stronger is the effect of a regulator on its targets (Cheng et al., 2009).
The average gene rank (FCR-score) of a regulator is defined as the difference of the average rank between its targets ($R_{ti}$) and non-targets ($R_{j}$): 

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

where for a given regulator, $n$ and $j$ represent the number and $t_i$ and $t_j$ represent the ranks of regulator targets and non-targets, respectively.

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

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

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) 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 TF 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) based on 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:

$$\text{CON} = \sum_{i=1}^{n} R_i$$

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

We also analyzed two recently developed tests, the Gene Set Z-score (GSZ, Törönen et al., 2009) and the model-based gene set analysis (MGSA; Bauer et al., 2010). GSZ combines features of the KS test (estimating the maximum enrichment) and the fold-change based methods that compare mean expression of targets and non-targets. MGSA models gene responses via Bayesian networks. In contrast to the tests described above, both GSZ and MGSA analyze all gene sets at once. Similar to the HG test, MGSA requires a threshold (used 0.5, 1.0 and 1.5 as for the HG test) to determine differentially expressed genes. Due to performance issues, the more involved 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:

$$\text{CON} = \sum_{i=1}^{n} R_i$$

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

We also analyzed two recently developed tests, the Gene Set Z-score (GSZ, Törönen et al., 2009) and the model-based gene set analysis (MGSA; Bauer et al., 2010). GSZ combines features of the KS test (estimating the maximum enrichment) and the fold-change based methods that compare mean expression of targets and non-targets. MGSA models gene responses via Bayesian networks. In contrast to the tests described above, both GSZ and MGSA analyze all gene sets at once. Similar to the HG test, MGSA requires a threshold (used 0.5, 1.0 and 1.5 as for the HG test) to determine differentially expressed genes. Due to performance issues, the more involved 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:

$$\text{CON} = \sum_{i=1}^{n} R_i$$

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

We also analyzed two recently developed tests, the Gene Set Z-score (GSZ, Törönen et al., 2009) and the model-based gene set analysis (MGSA; Bauer et al., 2010). GSZ combines features of the KS test (estimating the maximum enrichment) and the fold-change based methods that compare mean expression of targets and non-targets. MGSA models gene responses via Bayesian networks. In contrast to the tests described above, both GSZ and MGSA analyze all gene sets at once. Similar to the HG test, MGSA requires a threshold (used 0.5, 1.0 and 1.5 as for the HG test) to determine differentially expressed genes. Due to performance issues, the more involved 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:

$$\text{CON} = \sum_{i=1}^{n} R_i$$

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

We also analyzed two recently developed tests, the Gene Set Z-score (GSZ, Törönen et al., 2009) and the model-based gene set analysis (MGSA; Bauer et al., 2010). GSZ combines features of the KS test (estimating the maximum enrichment) and the fold-change based methods that compare mean expression of targets and non-targets. MGSA models gene responses via Bayesian networks. In contrast to the tests described above, both GSZ and MGSA analyze all gene sets at once. Similar to the HG test, MGSA requires a threshold (used 0.5, 1.0 and 1.5 as for the HG test) to determine differentially expressed genes. Due to performance issues, the more involved 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:

$$\text{CON} = \sum_{i=1}^{n} R_i$$

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

We also analyzed two recently developed tests, the Gene Set Z-score (GSZ, Törönen et al., 2009) and the model-based gene set analysis (MGSA; Bauer et al., 2010). GSZ combines features of the KS test (estimating the maximum enrichment) and the fold-change based methods that compare mean expression of targets and non-targets. MGSA models gene responses via Bayesian networks. In contrast to the tests described above, both GSZ and MGSA analyze all gene sets at once. Similar to the HG test, MGSA requires a threshold (used 0.5, 1.0 and 1.5 as for the HG test) to determine differentially expressed genes. Due to performance issues, the more involved 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:

$$\text{CON} = \sum_{i=1}^{n} R_i$$

where $n$ represents the number of tests applied to calculate the significance of a regulator in a given experiment. Thus, $R_i$ represents the rank of a regulator $j$ for a given statistical test $i$.
3.3 Test performance on E.coli versus S.cerevisiae
In addition, we also apply 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 (Narendra et al., 2011).

3.4 Detection of miRNA activity
In addition to TF-target relationships, we also evaluate miRNA target relationships based on miRNA transfection experiments in human cell lines. Here, a range of miRNA-target set definitions are 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 (e.g. Ritchie et al., 2009), they are required to complement manual repositories, which appear to be not sufficiently comprehensive for this analysis. Although this setting deviates considerably from the previously discussed ones, the overall ranking of methods is again very consistent (Fig. 5). An exception is HG-0.5 showing the third best performance after ANOVA and GSZ.

3.5 Randomized testing
To determine how the test performance depends on the quality of the available gene regulatory networks, we progressively randomize the regulator target sets. The ability of the different tests to infer the activity of TFs is surprisingly stable even if, on average, 50% of the gene regulatory network is randomized (Fig. 5).

3.6 Overall ranking of methods
Average ranks for the examined tests are computed based on their performance across different partially randomized expression compendia (E.coli, S.cerevisiae and human) and different scenarios ($H^{20}$ versus $H^{30}$). Thereby, we derive the following ordering of methods: ANOVA > GSZ > CON > WR > HG-0.5 > FCR > KS > HG-1.0 > BT > HG-1.5 > FD > MED > FCRW > MGSA. ANOVA, GSZ, CON and WR perform consistently well across all scenarios. While HG-0.5, FCR and KS also deliver usable results but fail in individual scenarios, the remaining tests (HG-1.0, BT, HG-1.5, FC, MED, FCRW and MGSA) performed below average across several scenarios (Table 3). We note that predictions by several methods are quite similar (see Supplementary Material 3) so a consensus of just ANOVA, WR and HG-0.5 is sufficient to improve the AUROC in E.coli from 0.80 to 0.84.

4 DISCUSSION AND CONCLUSION
Gene set enrichment tests have been devised to detect an over-representation of differentially expressed genes in predefined 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 (miRNAs and TFs) whose experimental perturbation directly offers the required standard-of-truth. In this setting, we evaluated the ability of 14 frequently used statistical tests (compare Huang et al., 2009) to detect regulator perturbations. Regulator activities cannot be derived directly from RNA expression levels: TFs are frequently regulated on the protein level and miRNAs are often not profiled. The indirect inference of regulator activities from the expression of their TGs is thus an important task.

We observed that, e.g. ANOVA and Wilcoxon test outperformed the HG and KS tests, which are most frequently used in practice. Tests depending on a threshold (HG, MGSA, compare Supplementary Material 2) yielded mixed results but improved for small thresholds below fold-changes of 2.

Despite the diverse test performance (AUC between 0.5 and 0.86 for E.coli), an unweighted consensus of all approaches consistently showed good results. Analogously, the GSZ score

Table 3. AUROC (±SD) for enrichment tests across expression compendia

<table>
<thead>
<tr>
<th>Enrichment tests</th>
<th>E.coli TFs ($H^{30}$)</th>
<th>E.coli TFs ($H^{30}$)</th>
<th>S.cerevisiae TFs ($H^{20}$)</th>
<th>Human miRNAs ($H^{30}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>0.86 (±0.03)</td>
<td>0.66 (±0.05)</td>
<td>0.71 (±0.04)</td>
<td>0.71 (±0.04)</td>
</tr>
<tr>
<td>GSZ</td>
<td>0.82 (±0.05)</td>
<td>0.67 (±0.05)</td>
<td>0.69 (±0.04)</td>
<td>0.73 (±0.03)</td>
</tr>
<tr>
<td>CON</td>
<td>0.80 (±0.05)</td>
<td>0.60 (±0.05)</td>
<td>0.73 (±0.03)</td>
<td>0.67 (±0.04)</td>
</tr>
<tr>
<td>WR</td>
<td>0.83 (±0.05)</td>
<td>0.64 (±0.05)</td>
<td>0.71 (±0.03)</td>
<td>0.68 (±0.04)</td>
</tr>
<tr>
<td>HG-0.5</td>
<td>0.80 (±0.04)</td>
<td>0.72 (±0.06)</td>
<td>0.70 (±0.03)</td>
<td>0.58 (±0.02)</td>
</tr>
<tr>
<td>FCR</td>
<td>0.74 (±0.05)</td>
<td>0.53 (±0.04)</td>
<td>0.71 (±0.03)</td>
<td>0.68 (±0.04)</td>
</tr>
<tr>
<td>KS</td>
<td>0.81 (±0.06)</td>
<td>0.69 (±0.04)</td>
<td>0.64 (±0.04)</td>
<td>0.63 (±0.04)</td>
</tr>
<tr>
<td>BT</td>
<td>0.72 (±0.03)</td>
<td>0.51 (±0.03)</td>
<td>0.72 (±0.03)</td>
<td>0.67 (±0.04)</td>
</tr>
<tr>
<td>FC</td>
<td>0.72 (±0.04)</td>
<td>0.51 (±0.03)</td>
<td>0.75 (±0.03)</td>
<td>0.68 (±0.04)</td>
</tr>
<tr>
<td>MED</td>
<td>0.67 (±0.05)</td>
<td>0.50 (±0.03)</td>
<td>0.69 (±0.03)</td>
<td>0.66 (±0.03)</td>
</tr>
<tr>
<td>HG-1.0</td>
<td>0.78 (±0.04)</td>
<td>0.71 (±0.04)</td>
<td>0.68 (±0.04)</td>
<td>0.54 (±0.03)</td>
</tr>
<tr>
<td>HG-1.5</td>
<td>0.73 (±0.04)</td>
<td>0.67 (±0.05)</td>
<td>0.72 (±0.03)</td>
<td>0.56 (±0.02)</td>
</tr>
<tr>
<td>FCRW</td>
<td>0.56 (±0.05)</td>
<td>0.50 (±0.002)</td>
<td>0.56 (±0.04)</td>
<td>0.71 (±0.03)</td>
</tr>
<tr>
<td>MGSA-0.5</td>
<td>0.63 (±0.08)</td>
<td>0.61 (±0.04)</td>
<td>0.53 (±0.04)</td>
<td>0.53 (±0.04)</td>
</tr>
</tbody>
</table>

The order reflects the overall ranking of methods (see section 3.6).
Assessment of gene set enrichment tests

Fig. 5. Progressive randomization of gene regulatory networks. Relationships are randomized in steps of 25% (100% = full randomization). In each case, the average AUROC from 100 randomized networks is shown. The order of curves at 0% corresponds to the order of methods in the legend.

(Törönen et al., 2009) combining features from enrichment and fold-change based methods also showed very good results. Parametric tests expecting normally distributed data such as ANOVA performed well as the data used in our study are indeed approximately normal (see Supplementary Material 3 or Zien et al., 2000).

Surprisingly, test performance did not improve by using interaction signs (activation versus inhibition). Here, we tested whether the fold-changes observed in TF targets are consistent with the given interaction sign annotations. Fold-changes and signs were clearly consistent in case of activation where activators caused target expression changes of the same sign. We did not find such coherent relationships between the fold-changes of repressors and their targets. According to Herrgård et al. (2003), this low correlation can be explained by the observation that either inhibitors or their targets exhibit low expression levels that are not profiled reliably.

To ensure the wide applicability of our results, we employed a variety of settings. In terms of microarray data, we used TF perturbations 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 TF-/miRNA-target networks and the more complex regulation in eukaryotes (Hu et al., 2007; Michoel et al., 2009; Narendra et al., 2011).

The performance ranking of the tests was very consistent between each of the examined scenarios, with several methods always performing substantially better than random guessing. We thus expect that the ranking of the 14 tests will be meaningful in novel settings that deviate from the ones described here. An example is the application of the tests to biological processes, where we expect the consensus 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 showed 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.

ACKNOWLEDGEMENTS
We thank the reviewers for help and comments.

Funding: BMBF (FKZ 01GS0801), DAAD (Referat 442: A0796865 to H. N.) and DFG (GRK 1563 to H. N.).

Conflict of Interest: none declared

REFERENCES


H.Naeem et al. (2008), e3112.


H.Naeem et al. (2010) miRSelect: automated extraction of associations between microRNAs and genes from the biomedical literature. BMC Bioinformatics, 11, 1–8.


