BIOINFORMATICS ORIGINAL PAPER
Vol. 28 no. 11 2012, pages 1480–1486
doi:10.1093/bioinformatics/bts164

Gene expression

Advance Access publication April 5, 2012

Rigorous assessment of gene set enrichment tests

Haroon Naeem, Ralf Zimmer, Pegah Tavakkolkhah and Robert Küffner*

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

Associate Editor: Trey Ideker

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 Wilcoxons 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.

*To whom correspondence should be addressed.

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 (TGs) to regulate their levels of expression (Chen *et al.,* 2007; Hobert, 2008; Martinez and Walhout, 2009). miRNAs are short (∼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 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 TGs (Cheng and Rajewsky 2007; Hu, 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 TGs for the rigorous evaluation of gene set enrichment tests. The experimental perturbation of regulators offers the required standard-of-truth that is not available for biological processes. Given regulator deletion or over-expression experiments, we consider the experimentally perturbed and the remaining regulators (with their corresponding sets of TGs) as positives and negatives, respectively. We thus evaluate the ability of statistical tests to infer the perturbed regulator from the expression of its TGs.

The present study, thereby conducts the first large comparison and rigorous assessment of 14 gene set enrichment tests on real data. We applied start-of-the-art statistical methods to test whether expression changes in regulator target sets might be due to chance. In the following sections, we review the field of gene set enrichment tests and describe our approach to rank enrichment tests.

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; Pehkonen *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), Efron and Tibshirani (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.,* 2011), we obtained 3425 *Escherichia coli* interactions between 167 TFs and 1377 TGs. The *Saccharomyces cerevisiae* network of 3940 interactions between 114 TFs and 1934 TGs were obtained from MacIsaac *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; Chua *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 $\log_2$ 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 (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 (Naeem *et al.*, 2010), we obtained putative miRNA–gene associations and relations extracted from biomedical abstracts by text mining (Table 2 and Supplementary Material 1).

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

Dataset	Interactions		mRNA compendia		
	TFs	Targets	Chips	KO/OE	Targets
E.coli	167	1377	907	17	949
S.cerevisiae Y1	114	1934	263	102	1527
S.cerevisiae Y2	114	1934	270	48	1094

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

Source	miRNAs	TGs	Pairs
DB: miRSel	486	1969	7604
DB: TarBase	110	837	1023
DB: MiRecords	93	614	772
DB: miR2Disease	176	364	596
PR: PITA	640	14 065	3 07 465
PR: PICTAR	163	5975	44 403
PR: TargetScan	249	9446	1 10 172

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 8 h and 32 h after miRNA over-expression of miR-155, miR-16 and let-7b. Expression profiles by He *et al.* (2007) include gene expression changes at 24 h 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 10 h and 24 h 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 24 h after miR-124, miR-1 and miR-181a transfection. Expression data were also measured by Grimson *et al.* (2007) in HeLa cells at 12 h and 24 h after over-expression of miR-7, −9, −122, −128, −132, −133, −142 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 (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 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.

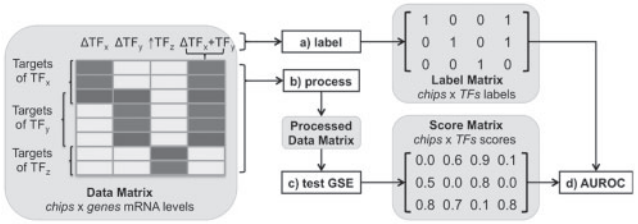

Fig. 1. Overview. The matrix of mRNA fold-changes consists of |*genes*| rows and |*chips*| columns. Chips are annotated by the treatment, e.g. triangle = deletion or up arrow = over-expression of TFs. This annotation is (a) compiled into a label matrix to represent the standard-of-truth. Perturbation of a regulator results in up- (red) or down-regulation (blue) of its TGs. After (b) processing the data matrix (Fig. 2), GSE tests are (c) applied to determine the activity of regulators based on the expression of TGs. The resulting score matrix (e.g. *P*-values) is compared with the label matrix to (d) compute the AUROC.

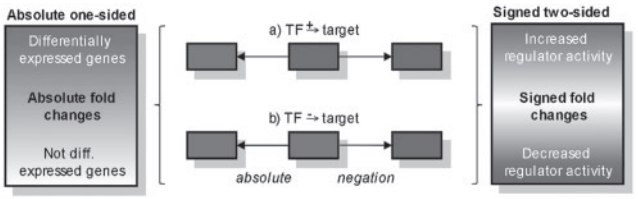

Fig. 2. Preprocessing of the data matrix. Two null hypotheses, the absolute (H_0^{abs}, left) and the signed two-sided (H_0^{sign}, right) can be tested after preprocessing the data matrix of TGs accordingly. For H_0^{abs}, expression profiles are transformed into absolute log fold-changes. For *E.coli*,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

The *absolute-one-sided* (H_o^{abs}) test ignores interaction signs. Enrichment is tested on absolute log fold-changes to evaluate the differential expression of TGs regardless of up- and down-regulation (Fig. 2, left).

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

2.7 Consistency of signs and fold-changes

To evaluate sign annotations, we define consistency as the fraction of the observed fold-changes that conform to the annotated interaction signs, e.g. in case of an activating (or inhibiting) relationship, an up-regulated TF is expected to cause up- (or down-) regulation in a corresponding TG, respectively. A higher consistency value denotes a better correspondence between observation and annotation where perfect or no correspondence is indicated by a consistency of 1 or 0.5, respectively. Values <0.5 indicate that signs and observations mostly disagree. TFs and TGs are considered regulated if they exhibit absolute $\log_2$ fold-changes that exceed a defined threshold (e.g. between 0 and 2). As a second variant, we only consider TFs as regulated if they are deleted or over-expressed in the given experiment.

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 Prill *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 $\log_2$ 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 (Nikiforov, 1994; Siegel, 1956) estimates the maximum possible enrichment by varying the length of the list of genes. In contrast, the HG test (Spiegel, 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, the P-value is computed via:

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

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

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 of 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 (T_{ravg}) and non-targets (nT_{ravg}). The genes ranks were derived by sorting them based on their absolute or signed fold-changes:

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

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):

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

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) 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:

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

where n represents the number of tests applied to calculate the significance of a regulator in a given experiment. Thus, R_{ji} 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 permutation analyses were not performed for GSZ and MGSA.

3 RESULTS

3.1 Detection of TF activity without sign annotations

We first evaluate 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}. Perturbations are only considered effective if the TFs exhibit a fold-change of at least 2 or <0.5. Conversely, substantial fold-changes in non-perturbed (secondary) TFs can be due to a direct or indirect effect from the perturbed (primary) TFs. Such cases are also excluded from the evaluation. In case of negative examples, we vary the fold-change cutoff to explore its influence on the performance of the enrichment tests (Fig. 3). At a higher cutoff, more negative examples are included in the analysis, which leads to a slightly decreased performance but hardly influences the overall ranking of enrichment tests (see Table 3 for AUROCs at a cutoff of 0.5).

In addition, we also combine individual tests into a consensus. The scores in the individual score matrices (Fig. 1) are transformed into ranks and averaged. Although some of the constituent tests hardly perform better than random, the consensus shows consistently good results across all scenarios.

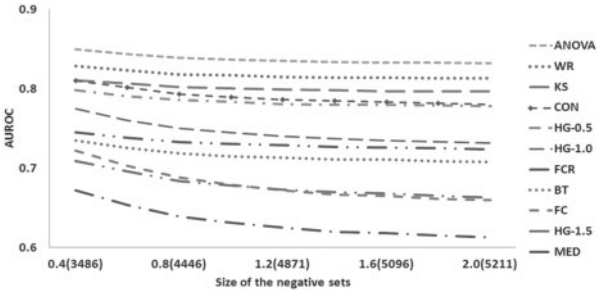

Fig. 3. Dependency of AUROC on the set of negatives (*E.coli*). TFs are only considered as negatives in the AUROC analysis if they exhibit fold-changes of less than a predefined cutoff. The *x*-axis shows the sizes of different negative sets (in brackets) compiled based on different absolute $\log_2$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

3.2 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 TG. H_0^{abs} as applied in the previous section, tests only for differential expression. By using H_0^{sign} instead, we also test whether the fold-changes observed in TF targets are consistent with the interaction sign annotations. Neglecting signs slightly but consistently improves the performance of enrichment tests (Table 3).

To evaluate why the interaction sign annotations do not improve our results, we compare the signs to the observed fold-changes. Activating interactions conform to our expectation, i.e. up- (or down-) regulation of a TF causes up- (or down-) regulation, respectively, of their TGs. In case of inhibiting relationships, we expect opposite fold-changes in TFs and TGs. This occurs only rarely in the data at hand (Fig. 4) and thus explains the reduction in the performance of the signed tests.

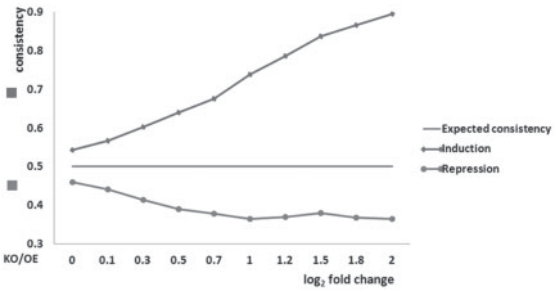

Fig. 4. Consistency of sign annotations and fold-changes (*E.coli*). In case of induction, fold-changes of TFs and their TGs predominantly point into the same direction (red line, consistency >0.5). Consistency increases for TFs and TGs exhibiting higher fold-changes. However, up-regulated TFs rarely cause down-regulation (or vice versa) of TGs in case of repressing relationships (green line, consistency <0.5). This trend is confirmed if only deleted or over-expressed TFs (KO/OE: dots, left side) are considered.

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

	Enrichment tests	*E.coli* TFs		*S.cerevisiae* TFs		Human miRNAs
		(H_o^{abs})	(H_o^{sig})	Y1-(H_o^{abs})	Y2-(H_o^{abs})	(H_o^{abs})
ANOVA	Two-sample ANOVA≡*t*-test	0.86 (± 0.03)	0.66 (± 0.05)	0.71 (± 0.04)	0.71 (± 0.04)	0.84 (±0.03)
GSZ	Gene set Z-score	0.82 (± 0.05)	0.67 (± 0.05)	0.69 (± 0.04)	0.73 (± 0.05)	0.82 (± 0.05)
CON	Consensus of all tests	0.80 (± 0.05)	0.60 (± 0.05)	0.73 (± 0.03)	0.67 (± 0.04)	0.80 (± 0.01)
WR	Wilcoxon's rank sum	0.83 (± 0.05)	0.64 (± 0.05)	0.71 (± 0.03)	0.68 (± 0.04)	0.77 (± 0.03)
HG-0.5	Hypergeometric, cut = 0.5	0.80 (± 0.04)	0.72 (± 0.06)	0.70 (± 0.03)	0.58 (± 0.02)	0.81 (± 0.03)
FCR	Average gene rank	0.74 (± 0.05)	0.53 (± 0.04)	0.71 (± 0.03)	0.68 (± 0.04)	0.75 (± 0.03)
KS	Kolmogorov–Smirnov	0.81 (± 0.06)	0.69 (± 0.04)	0.64 (± 0.04)	0.63 (± 0.04)	0.76 (± 0.03)
BT	Bootstrapping	0.72 (± 0.03)	0.51 (± 0.003)	0.72 (± 0.03)	0.67 (± 0.04)	0.66 (± 0.007)
FC	Average fold-change	0.72 (± 0.04)	0.51 (± 0.03)	0.75 (± 0.03)	0.68 (± 0.04)	0.51 (± 0.004)
MED	Median	0.67 (± 0.05)	0.50 (± 0.03)	0.69 (± 0.03)	0.66 (± 0.03)	0.68 (± 0.03)
HG-1.0	Hypergeometric, cut = 1.0	0.78 (± 0.04)	0.71 (± 0.04)	0.68 (± 0.04)	0.54 (± 0.03)	0.72 (± 0.03)
HG-1.5	Hypergeometric, cut = 1.5	0.73 (± 0.04)	0.67 (± 0.05)	0.72 (± 0.03)	0.56 (± 0.02)	0.50 (± 0.04)
FCRW	Average fold-change rank weight	0.56 (± 0.05)	0.50 (± 0.002)	0.56 (± 0.04)	0.71 (± 0.03)	0.50 (± 0.001)
MGSA-0.5	Model-based gene set analysis	0.63 (± 0.08)	0.61 (± 0.04)	0.53 (± 0.04)	0.53 (± 0.04)	0.48 (± 0.05)

The order reflects the overall ranking of methods (see section 3.6).

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_0^{abs} versus H_0^{sign}). Thereby, we derive the following ordering of methods: ANOVA > GSZ > CON > WR > HG-0.5 > FCR > KS > HG-1.0 > BT > HG-1.5 > FC > 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 Wilcoxons 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

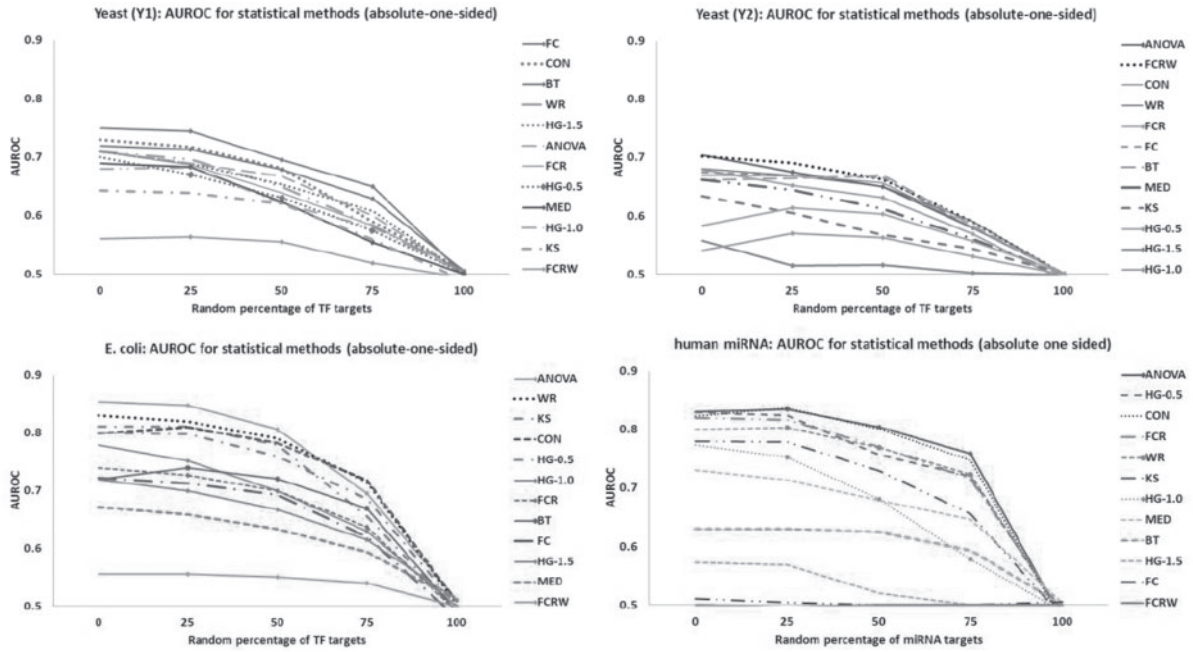

Fig. 5. Progressive randomization of gene regulatory networks. Relationships are randomized in steps of 25% (100% = full randomization). In each case, the average AUROC from 100 randomized networks is shown. The order of curves at 0% corresponds to the order of methods in the legend.

(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: A/07/96865 to H. N.) and DFG (GRK 1563 to H. N.).

Conflict of Interest: none declared

REFERENCES

Ackermann,M. and Strimmer,K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, **10**, 47.

Al-Shahrour,F. *et al*. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.

Baek,D. *et al*. (2006) The impact of microRNAs on protein output. *Nature*, **455**, 64–71.

Barry,W.T. *et al*. (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**, 1943–1949.

Barry,W.T. *et al*. (2008) A statistical framework for testing functional categories in microarray data. *Ann. Appl. Stat.*, **2**, 286–315.

Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.

Bauer,S. *et al*. (2010) GOing Bayesian: model-based gene set analysis of genome-scale data. *Nucleic Acids Res.*, **38**, 3523–3532.

Boorsma,A. *et al*. (2008) Inferring condition-specific modulation of transcription factor activity in yeast through regulon-based analysis of genomewide expression. *PLoS One,* **3**, e3112.

Borda,J. (1781) Memoire sur les elections au scrutin. *Histoire de l'Academie des Sciences*, Paris.

Chen,K. and Rajewsky,N. (2007) The evolution of gene regulation by transcription factors and microRNAs. *Nat. Rev. Genet.*, **8**, 93–103.

Cheng,C. *et al*. (2009) mRNA expression profiles show differential regulatory effects of microRNAs between estrogen receptor-positive and estrogen receptor-negative breast cancer. *Genome Biol.*, **10**, R90.

Chua,G. *et al*. (2006) Identifying transcription factor functions and targets by phenotypic activation. *Proc. Natl Acad. Sci. USA*, **103**, 12045–12050.

Efron,B. and Tibshirani,R.J. (1993) *An Introduction to the Bootstrap*. Chapman & Hall, Boca Raton, FL.

Efron,B. and Tibshirani,R.J. (2006) On testing the significance of sets of genes. *Ann. Appl. Stat*., **1**, 107–129.

Essaghir,A. *et al*. (2010) Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. *Nucleic Acids Res.*, **38**, e120.

Faith,J.J. *et al*. (2008) Many microbe microarrays database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.*, **36**, D866–D870.

Farh,K.K. *et al*. (2005) The widespread impact of mammalian microRNAs on mRNA repression and evolution. *Science*, **10**, 1817–1821.

Friedman,R.C. *et al*. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.

Gama-Castro,S. *et al*. (2011) RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res.*, **39**, D98–D105.

Gatti,D.M. *et al*. (2010) Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics*, **11**, 574.

Georges,S.A. *et al*. (2008) Coordinated regulation of cell cycle transcripts by p53-Inducible microRNAs, miR-192 and miR-215. *Cancer Res.*, **68**, 10105–10112.

Goeman,J.J. and Bühlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

Grimson,A. *et al*. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.

Harris,M.A. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.

He,L. *et al*. (2007) A microRNA component of the p53 tumour suppressor network. *Nature*, **447**, 1130–1134.

Herrgård,M.J. *et al*. (2003) Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res*., **13**, 2423–2434.

Hobert,O. (2008) Gene regulation by transcription factors and microRNAs. *Science*, **319**, 1785–1786.

Hu,H. (2010) An efficient algorithm to identify coordinately activated transcription factors. *Genomics*, **95**, 143–150.

Hu,Z. *et al*. (2007) Genetic reconstruction of a functional transcriptional regulatory network. *Nat. Genet.*, **39**, 683–687.

Huang da,W. *et al*. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

Jiang,Q. *et al*. (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.*, **37**, D98–D104.

Kertesz,M. *et al*. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.

Khatri,P. and Drãghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.

Krek,A. *et al*. (2005) Combinatorial microRNA target predictions. *Nat. Genet.*, **37**, 495–500.

Lehmann,E.L. (1975) *Nonparametric Statistical Methods Based on Ranks*. McGraw-Hill, New York.

Levine,D.M. *et al*. (2006) Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways. *Genome Biol*., **7**, R93.

Liu,Q. *et al*. (2010) TF-centered downstream gene set enrichment analysis: inference of causal regulators by integrating TF-DNA interactions and protein post-translational modifications information. *BMC Bioinformatics*, **11** (Suppl. 11), S5.

MacIsaac,K.D. *et al*. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**, 113.

Mann,H.B. and Whitney,D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.*, **18**, 50–60.

Martin,D. *et al*. (2004) GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol.*, **5**, R101.

Martinez,N.J. and Walhout,A.J. (2009) The interplay between transcription factors and microRNAs in genome-scale regulatory networks. *Bioessays*, **31**, 435–445.

Michoel,T. *et al*. (2009) Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. *BMC Syst. Biol.*, **3**, 49.

Miller,R.G. (1997) *Beyond ANOVA: Basics of Applied Statistics*. Chapman & Hall, Boca Raton.

Mootha,V.K. *et al*. (2003) PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.

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

Naeem,H. *et al*. (2011) MIRTFnet: analysis of miRNA regulated transcription factors. *PLoS One,* **6**, e22519.

Nam,D. and Kim,S.Y. (2008) Gene-set approach for expression pattern analysis. *Brief Bioinform*., **9**, 189–197.

Narendra,V. *et al*. (2011) A comprehensive assessment of methods for de-novo reverse-engineering of genome-scale regulatory networks. *Genomics*, **97**, 7–18.

Nikiforov,A.M. (1994) Algorithm AS 288: exact Smirnov two-sample tests for arbitrary distributions. *Applied Statistics*, **43**, 265–284.

Ott,C.E. *et al*. (2011) MicroRNAs differentially expressed in postnatal aortic development downregulate elastin via 3' UTR and coding-sequence binding sites. *PLoS One*, **6**, e16250.

Papadopoulos,G.L. *et al* (2009) The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res.*, **37**, D155–D158.

Pavlidis,P. *et al*. (2004) Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem. Res.*, **29**, 1213–1222.

Pehkonen,P. *et al*. (2005) Theme discovery from gene lists for identification and viewing of multiple functional groups. *BMC Bioinformatics*, **6**, 162.

Prill,R.J. *et al*. (2010) Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS One*, **5**, e9202.

Ritchie,W. *et al*. (2009) Predicting microRNA targets and functions: traps for the unwary. *Nat. Methods*, **6**, 397–398.

Rivals,I. *et al*. (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407.

Selbach,M. *et al*. (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, **455**, 58–63.

Siegel,S. (1956) *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York.

Sohler,F. and Zimmer,R. (2005) Identifying active transcription factors and kinases from expression data using pathway queries. *Bioinformatics*, **21**, 115–122.

Sood,P. *et al*. (2006) Cell-type-specific signatures of microRNAs on target mRNA expression. *Proc. Natl Acad. Sci. USA*, **103**, 2746–2751.

Spiegel,M.R. (1992) *Theory and Problems of Probability and Statistics*. McGraw-Hill, New York, pp. 113–114.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Törönen,P. *et al*. (2009) Robust extraction of functional signals from gene set analysis using a generalized threshold free scoring function. *BMC Bioinformatics*, **10**, 307.

Tu,K. *et al*. (2009) Combinatorial network of primary and secondary microRNA-driven regulatory mechanisms. *Nucleic Acids Res.*, **37**, 5969–5980.

Xiao,F. *et al*. (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.*, **37**, D105–D110.

Yi,M. *et al*. (2006) Wholepathwayscope: a comprehensive pathway-based analysis tool for high-throughput data. *BMC Bioinformatics*, **7**, 30.

Zeeberg,B.R. *et al*. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28.

Zien,A. *et al*. (2000) Analysis of gene expression data with pathway scores. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 407–417.

Zweig,M.H. and Campbell,G. (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.*, **39**, 561–577.