Discovering Causal Signaling Pathways Through Gene-Expression Patterns

Published online 21 May 2010 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* ***W109–W117***
doi:10.1093/nar/gkq424

Discovering causal signaling pathways through gene-expression patterns

Jignesh R. Parikh[1], Bertram Klinger[2,3], Yu Xia[1,4,5], Jarrod A. Marto[6,7] and Nils Blüthgen[2,3,*]

[1]Bioinformatics Program, Boston University, Boston, MA 02115, USA, [2]Institute of Pathology, Charité Universitätsmedizin Berlin, [3]Institute for Theoretical Biology, Humboldt University Berlin, Germany, [4]Department of Chemistry, Boston University, [5]Department of Biomedical Engineering, Boston University, Boston, MA 02115, USA, [6]Department of Cancer Biology and Blais Proteomics Center, Dana-Farber Cancer Institute and [7]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, MA 02115, USA

Received January 29, 2010; Revised April 28, 2010; Accepted May 6, 2010

ABSTRACT

High-throughput gene-expression studies result in lists of differentially expressed genes. Most current meta-analyses of these gene lists include searching for significant membership of the translated proteins in various signaling pathways. However, such membership enrichment algorithms do not provide insight into which pathways caused the genes to be differentially expressed in the first place. Here, we present an intuitive approach for discovering upstream signaling pathways responsible for regulating these differentially expressed genes. We identify consistently regulated signature genes specific for signal transduction pathways from a panel of single-pathway perturbation experiments. An algorithm that detects overrepresentation of these signature genes in a gene group of interest is used to infer the signaling pathway responsible for regulation. We expose our novel resource and algorithm through a web server called SPEED: Signaling Pathway Enrichment using Experimental Data sets. SPEED can be freely accessed at http://speed.sys-bio.net/.

INTRODUCTION

Signal transduction pathways integrate information about the cellular environment and regulate the activity of a multitude of transcription factors, thereby controlling cellular processes, such as migration, proliferation and differentiation through context-dependent gene expression. While the techniques to measure transcript levels on a genome scale have matured and become affordable, it is not yet possible to get a global picture of the activity of signaling pathways in a high-throughput manner.

Typical strategies to analyze gene-expression profiles include searches for overrepresented biological functions, molecular mechanisms or pathways that the regulated genes are involved in, often utilizing annotation databases like Gene Ontology (1) and pathway databases such as KEGG (2). While these strategies have proven very useful to systematically organize the results, it is often forgotten that these analyses mainly detect pathways that are regulated in response to the perturbation—and not the cause of the observed regulatory pattern. If, for example, the MAPK signaling pathway has been found overrepresented in the regulated genes it only shows that the pathway is regulated in response to a perturbation, but not that the MAPK signaling pathway caused the gene regulation. Although genes regulated by a signaling pathway may involve genes that encode for members of the same pathway (feedback regulators), they may also regulate genes that are members of other pathways (transcriptional cross-talk) (3). Thus, pathway membership of regulated genes is unlikely to unveil the regulating pathway, and other strategies have to be employed in order to probe the origin of a certain gene-expression pattern.

One such computational strategy that is often used is the search of overrepresented binding sites in the promoter region of the regulated genes. This strategy is indeed suitable for finding the transcription factors causing the regulation. However, the method is seriously hampered by the poor specificity of computational prediction of

*To whom correspondence should be addressed. Tel: +49 30 2093 9106; Fax: +49 30 2093 8801; Email: nils.bluethgen@charite.de

transcription factor binding sites, due in large part to the high degeneracy and low information content of the recognized binding sequences (4). In the future, development in experimental strategies such as ChIP-chip or ChIP-seq will likely improve this situation. Still, links between signaling pathways and the transcription factors are only known for a limited number of cases. Thus, even if one has discovered which transcription factor is causing the regulation, in most cases one still cannot pin down the signaling pathway upstream of this transcription factor.

In order to infer signaling pathways that caused the regulation of a group of genes, we developed Signaling Pathway Enrichment using Experimental Data sets (SPEED), a data collection and algorithm that allows for identifying signaling pathways that cause an observed regulatory pattern. The key idea behind SPEED is that the same signaling pathway typically regulates a similar (small) core set of genes in most cell types, while different signaling pathways typically regulate different core sets of genes. We term such a set of consistently regulated genes 'signature genes' for that particular pathway. A list of regulated genes in an experiment can then be compared to a catalog of signature genes for different signaling pathways. In the following section, we present the SPEED web server with application to the analysis of two acute myeloid leukemia (AML) data sets.

MATERIALS AND METHODS

Selection and preprocessing of gene-expression data

Gene-expression data sets were identified manually in the gene-expression omnibus (GEO) database based on their abstracts. We concentrated our search on single-perturbation experiments measured within 4 h after treatment to limit secondary effects. This time span represents the shortest interval, which still allowed collection of a sufficient number of data sets in most pathways. If, however, not enough such data sets were available for a particular pathway, long-term experiments such as stable overexpression or knockdown experiments were included if their effect on signature genes was assumed to be prominent.

Custom R-scripts were written to download and process the selected expression data sets automatically. Data for each micro-array experiment within the data set were subjected to quantile normalization and log-transformation if needed. Finally, a LOESS model for the standard deviation as a function of expression value was trained either on replicate experiments (when available) or on all experiments and was used to compute Z-scores. Z-score $(x) = (x-\mu)/\sigma$, where μ is the sample mean and σ is the standard deviation. The probes were annotated with Entrez Gene identifiers using GEOquery.

Implementation

SPEED runs on an Apache server with the pre-processed micro-array data stored in an SQLite 3.0 database (see Supplementary Figure S2 for database schema). The SPEED algorithm is implemented in Python 2.5 with dependency on the scipy 0.7.0 (http://www.scipy.org/) package. Graphical output is dynamically generated using the Google Chart API (http://code.google.com/apis/chart/).

Clustering

Hierarchical clustering was performed using the statistical package R, clusters were linked using Ward's method. The distance between two pathways was defined as: −log [# common signature genes/min (# signature genes for each set)].

Validation using literature gene lists

Gene lists were extracted directly from the manuscripts or supplementary data and were not further processed (see Supplementary Table S1 for references). Raw data were not available for these lists. Every significant hit in SPEED had a false discovery rate (FDR) < 0.05.

FANTOM4 data processing

The FANTOM4 data set was downloaded as tab delimited text files from the Center for Information Biology gene EXpression (CIBEX) database (http://cibex.nig.ac.jp/cibex2/ExperimentMiame.do?queryExperimentalDesignAccession=CBX47). Each of the 52 transcription factor knockdowns in THP-1 cell line were conducted in three replicates. Probe expression values were divided by appropriate controls and the ratios were log (base 2) transformed. Probes were only considered valid if the transcript was expressed in all three replicates for at least half of the 52 transcription factor knockdowns. Valid probe IDs were converted to Entrez Gene IDs and used as background. Genes with at least 2-fold differential median expression were considered for input to SPEED.

RESULTS

We define signature genes as genes that are consistently regulated by a particular pathway across many experiments. Overrepresentation of such signature genes in a list of differentially expressed genes can then hint at the signaling pathway that caused the regulation. We set up a pipeline to compile publicly available expression data sets from pathway perturbation experiments, store them into a database, extract signature genes, and finally detect significant overlap with a user-supplied gene group via a web server (http://speed.sys-bio.net/, see schematics in Figure 1).

Compilation of a database to define signature genes for signaling pathways

In order to extract signature genes, we selected a set of 11 signaling pathways: TGF-β, H_2O_2, TLR, IL-1, MAPK, PI3K, MAPK+PI3K, Wnt, JAK-STAT, TNF-α and VEGF. For each signaling pathway, we manually searched the GEO database (5) for gene-expression data sets where this signaling pathway is specifically perturbed (see 'Materials and Methods' for selection criteria). The

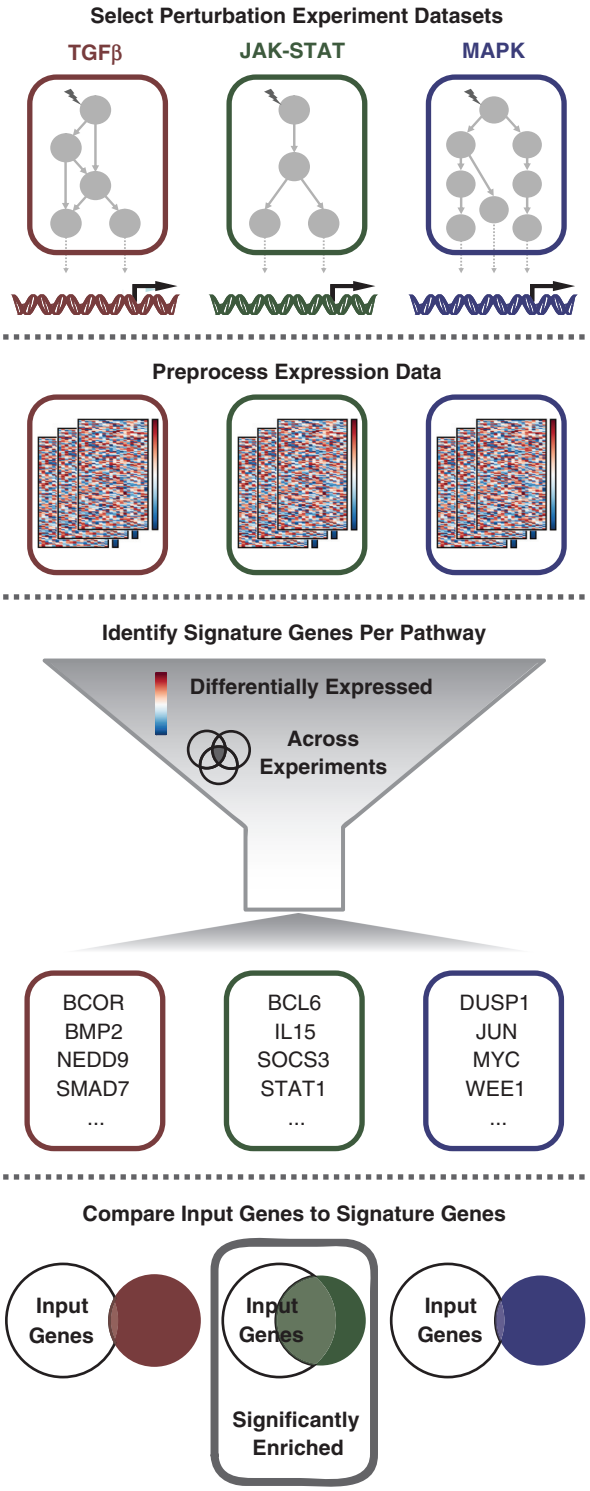

Figure 1. Overview of the SPEED algorithm. The SPEED algorithm is based on the identification of signature genes that are consistently regulated by specific signaling pathways using publicly available micro-array data. Gene-expression data sets from single-pathway perturbation experiments were manually selected from the GEO database. Next, gene-expression values from the selected database were automatically processed using custom R-scripts, and expression changes were stored as *Z*-score rank percentiles in the SPEED database. The SPEED web server extracts signature genes per pathway on the fly based on user-specified parameters describing the level of differential expression (*Z*-score percentile; ex: top 1%) and the level of consistency across experiments (percentage of experimental data sets where a gene is differentially expressed; ex: at least 20%). Users can compare their own gene sets against the extracted signature genes to identify modulated upstream signaling pathways.

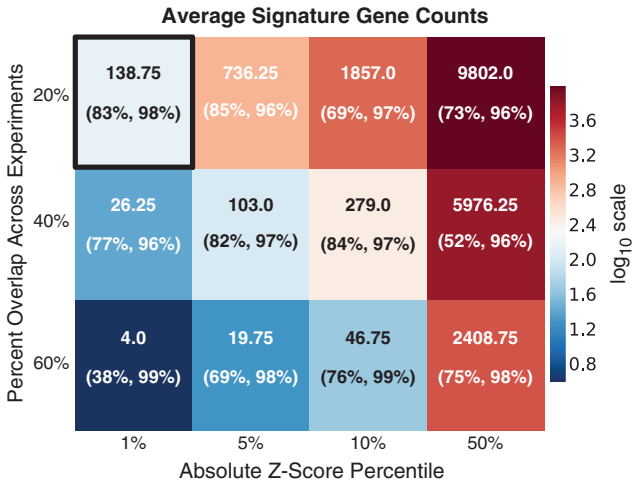

Figure 2. Average number of SPEED signature genes. SPEED signature genes are sensitive to user-specified parameters. Average signature genes per pathway without the uniqueness constraint are listed as a function of *Z*-score percentile and percent overlap across experiments. The heat map corresponds to $\log_{10}$ of the average number of signature genes. The sensitivity and specificity is noted in parenthesis. The values for the default parameter set of *Z*-score percentile $\leq$1% and percent overlap $\geq$20% are boxed. Our choice of default parameter set, as well as recommended parameters (in black text) are determined *ad hoc* based on their biological meaning, resulting number of signature genes and performance metrics. Only the default pathways are considered here and bottom 50% expressed genes are discarded for all calculations.

data sets were automatically downloaded from GEO (6), annotated and normalized. The expression changes per probe were then transformed into *Z*-scores per gene (see 'Materials and Methods' section for details). Subsequently, the gene ID, expression rank and *Z*-score rank were stored in the SPEED database. Currently, 215 sets of micro-array experiments are stored in the database, resulting in almost 6 million expression values for 21 485 human genes.

From this database, signature genes for each pathway can be extracted. We define three criteria for a signature gene: it has to be (i) expressed and (ii) regulated (iii) across several experiments for each pathway. As the stored micro-array data originates from different sources and even platforms, we chose to apply thresholds based on percentiles to select the signature genes. For each of the data sets, genes are defined as expressed if their expression value rank is higher than a pre-determined threshold [criteria (i); default: top 50%). Next, regulated genes are extracted based on the rank of their *Z*-score [criteria (ii); default: top 1%). Finally, genes are selected as signature genes if they are expressed and regulated in more than a pre-determined percentage of data sets for a pathway [criteria (iii); default: >20%). Users can adjust these parameters to create lists of signature genes on the fly.

The mean number of signature genes per pathway using the default parameters is 139. The signature gene lists for any given parameter set can be downloaded as tab delimited text directly from the SPEED web server. Figure 2 lists the mean number of signature genes per pathway for various parameters.

Overlap of signature genes across pathways

Since signaling pathways form highly interconnected networks, the definition of distinct signaling pathways is difficult and often arbitrary. We therefore analyzed the overlap between signature genes between all pairs of pathways and noticed that there is considerable overlap. In order to group similar pathways, we performed hierarchical cluster analysis (Figure 3). Clustering revealed that the major clusters separate pathways involved in immune response from pathways controlling cell cycle and cell growth. Furthermore, VEGF signaling seems very closely related to pathways from both clusters (MAPK+PI3K signaling and TNF-α signaling). Also, signaling triggered by IL-1 and TNF-α shows strong overlap in the signature genes, probably due to the shared downstream IKK-β/NF-κB signaling cascade (7). We therefore added a feature in SPEED to select only those signature genes that are unique to each pathway. We note, however, that under such a uniqueness constraint the signature genes depend strongly on the thresholds chosen, and the predictive power of the algorithm does not improve (Figure 2; Supplementary Figure S1). Additionally, since we expect a strong crosstalk between pathways, non-unique signature genes seem to be a more biologically reasonable choice. We also decided to give the user the possibility to restrict their analysis on a subset of pathways. The default setting restricts the analysis on four major pathways JAK-STAT, MAPK+PI3K, TGF-β and TLR. While users have complete control in their choice of parameters, the selected parameters should be biologically reasonable (Figure 2 describes our recommended parameters).

Detecting overrepresented signature genes

In order to detect overrepresented signature genes, we apply Fisher's exact test with a multiple hypothesis correction. For each pathway, it is tested whether the corresponding signature genes are overrepresented in the user-supplied list. A *P*-value is provided to indicate the significance for the overrepresentation. To adjust for multiple hypotheses, an FDR for the *P*-value (8) is also provided. Users can restrict the analysis to a subset of genes in the database by submitting a list of background genes. For example, if one submits a list of differentially expressed genes from a micro-array study, we would recommend using all significantly expressed genes from that micro-array study as background.

Validation

In order to test how well SPEED can predict modulated signaling pathway activity, we collected gene lists from literature with known signaling pathway perturbation. Note that gene-expression data from these literature sources was not integrated into SPEED. A result was considered to be true positive if the top-ranked pathway prediction matched the actual perturbed pathway and the FDR was below 0.05. Table 1 describes the validation results for the default pathways (see Supplementary Table S1 for details). The overall sensitivity, calculated as the fraction of all literature sources with correct

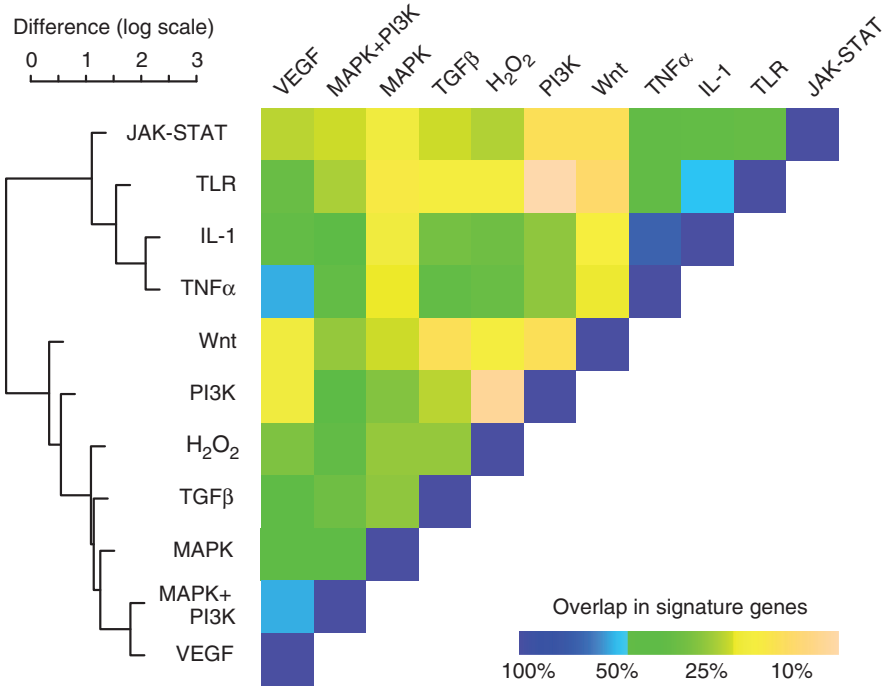

Figure 3. Signaling pathway crosstalk demonstrated by overlapping signature genes. The percent overlap between signature genes between all pairs of pathways is displayed as a heat map with higher overlap suggesting greater crosstalk between the respective signaling pathways. Pathway similarity is calculated as the negative log of the percent overlap and similar pathways are grouped using hierarchical clustering. Two major clusters are realized, separating pathways involved in immune response (JAK-STAT, TLR, IL-1 and TNF-α) from pathways controlling cell cycle.

Table 1. Validation of the SPEED algorithm on gene lists from independent literature sources

Pathway	Test sets	Correct top-ranking events (sensitivity)		Significantly overrepresented events (sensitivity)	
		SPEED (%)	GATHER (%)	SPEED (%)	GATHER (%)
JAK-STAT	10	8 (80)	1 (10)	10 (100)	7 (70)
TGF-β	5	4 (80)	2 (40)	5 (100)	2 (40)
MAPK+PI3K	6	4 (67)	1 (17)	6 (100)	6 (100)
TLR	6	5 (83)	3 (50)	6 (100)	3 (50)
Total	27	21 (78)	7 (29)	27 (100)	18 (65)

SPEED results from 27 literature-derived gene lists for the four default pathways are summarized. As compared to traditional signaling pathway membership analysis using GATHER, SPEED correctly predicts the perturbed signaling pathway as the top-ranking result at a higher rate (78% compared to 29%). SPEED also outperforms GATHER in identifying the correct pathway as one of the significantly overrepresented pathways (FDR $\leq$ 0.05 for SPEED and no threshold for GATHER).

pathway predictions, using default parameters was found to be 78%. Since traditional signaling pathway membership analyses do not aim at predicting causal pathways, SPEED outperforms GATHER (9), which correctly predicts the upstream pathway with a sensitivity of 29%.

Due to limited number of available positive sets in literature, we also conducted leave-one-out cross validation to get a better estimate of sensitivity. Once again, we considered a result to be true positive if the top-ranking predicted pathway was indeed the one that was perturbed. Using default parameter values, the sensitivity was calculated to be 83% of 130 tests. The classification problem here is not binary and therefore the expected sensitivity for random predictions is far <50%. In order to estimate the specificity, the fraction of negative input sets correctly identified as negative, we ran SPEED on 200 negative sets comprised of randomly selected genes from the database. The size of the negative sets was randomly chosen between 50 and 200 genes. The overall specificity and precision (the proportion of true positive results against all positive results) was calculated to be 98 and 96%, respectively. Figure 2 notes the sensitivity and specificity for different parameters.

Comparison with pathway membership databases

Validation using the literature derived gene sets indicated a difference between genes regulated by a pathway and gene products that are members of the pathway. To verify the disparity between regulated genes and pathway members, we searched regulated genes from each SPEED experimental data set for overrepresentation of acting pathway members in KEGG Pathway (2), BioCarta and Panther (10) databases. 39% of all data sets had significant enrichment of the acting pathway members as determined by KEGG Pathway. 9 and 24% of all data sets had significant enrichment of the corresponding pathway in BioCarta and Panther, respectively. These findings suggest some degree of transcriptional feedback where genes regulated by a pathway translate to protein members within the same pathway. However, only 4 and 7% of all data sets had the acting pathway ranked as the top one in KEGG and Panther, respectively. No experimental data set had the acting pathway ranked as the top one in BioCarta (see Supplementary Tables S4 and S5 and Figures 3 and 4, for detailed results). The low top-rank percentage reasserts the unfavorable use of pathway membership databases for identifying pathways regulating a list of genes. Consequently, the SPEED data sets and signature genes are fundamentally different than lists of pathway members found in existing membership databases.

Web server

The scope of the SPEED web server is the identification of causal signaling pathways given a list of input human genes. Thus, we implemented the web server with an unencumbered user interface where the user can input gene lists in widely adopted formats including Entrez Gene ID, gene symbol, Uniprot, GI number, Refseq, Ensembl and IPI. Any other gene identifiers can be easily converted to those formats using available conversion tools such as Clone ID converter (11), DAVID bioinformatics resources (12) or the UniProt ID mapping service (13). Upon submission, the user is directed to the output page where the results are presented as a list of signaling pathways ordered by their significance with number of signature genes present in the input, *P*-values and FDRs for each pathway. Graphical outputs, such as a pie chart representing the distribution of signature genes in the input list and a bar chart describing the relative FDRs, further aid in interpretation of the results (Figure 4). The symbols of the signature genes present in the input are provided with hyperlinks to the respective NCBI Entrez Gene web page. Separate web pages are provided for database access to retrieve signature genes as tab delimited text or to browse through experimental details such as cell type, perturbation strategy and length of perturbation with hyperlinks to the original gene-expression data sources.

In addition to the web interface, we provide download options for the raw data as tab delimited text or as an SQLite database, which is a single cross-platform disk file, together with Python source code to facilitate local programmatic access to SPEED functionality. All SPEED code is open-source and free to use.

In the following sections, we demonstrate the functionality of SPEED through analysis of two AML data sets.

Total genes in list: 800
Total genes in background: 21485

Pathway	Genes in List	Genes in Background	P-Value	FDR
TNFa	52	259	1.09e-23	4.91e-23
IL-1	31	141	8.45e-16	2e-15
TLR	28	181	1.63e-10	2.88e-10
MAPK_PI3K	21	118	2.38e-09	3.15e-09
TGFB	23	142	2.81e-09	3.37e-09
JAK-STAT	20	114	7.36e-09	5.63e-09
VEGF	13	56	1.03e-07	9.19e-08
PI3K_only	13	67	9.48e-07	6.01e-07
MAPK_only	41	559	3.12e-05	1.85e-05
H2OS	9	60	0.000354	0.000269
Wnt	7	83	0.035	0.0239

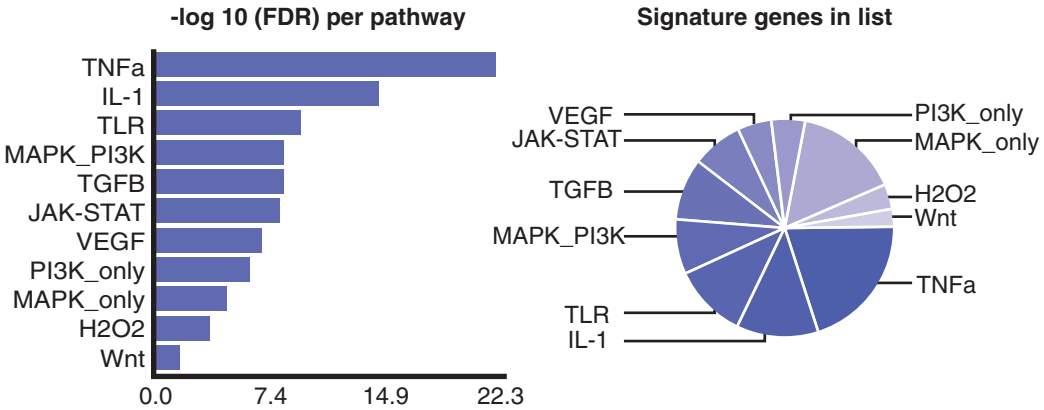

Figure 4. SPEED analysis of downregulated genes in high-risk AML patients with CEBPa mutations. Screenshot of SPEED results shows TNF-α as the top-ranking upstream signaling pathway for target genes downregulated in high-risk AML patients with CEBPa mutations versus CEBPa wild-type. The pathways are sorted by FDR and graphical outputs aid in interpretation of the results.

Identification of signaling pathways downregulating genes in high-risk AML patients with CEBPa mutations

Patients with AML harbor mutations in one or more genes such as FLT3, NPM1, c-KIT and CEBPa (14). Some mutations like internal tandem duplication in FLT3 (FLT3-ITD) are associated with high-risk AML patients and confer poor outcomes (14). Marcucci *et al.* (14) conducted a study to search for prognostic markers in cytogenetically normal AML patients with high-risk molecular features (FLT3-ITD and/or NPM1 wild-type) with or without CEBPa mutations. The study identified 928 genes that were downregulated in CEBPa mutants as compared to wild-type. We searched the downregulated genes for enriched pathway annotations using GATHER (9) and DAVID (12), using standard parameters (Bayes factor $\geq$6 and corrected *P*-value $\leq$0.05 for GATHER and DAVID, respectively). No pathways were identified as containing a significant number of members in the input gene list, which suggests that the downregulated genes have little effect on signaling pathways.

As previously noted, searching for membership enrichment does not provide insight into the signaling pathways that causally regulate a list of genes. We searched the same list of downregulated genes using SPEED (default parameters with all pathways selected) and identified TNF-α as the top-ranking pathway (Figure 4). Interestingly, the top three pathways, TNF-α, IL-1 and TLR, form a single cluster (Figure 3).

Predicting signaling pathways upstream of 24 transcription factors in THP-1 cells

In the recently published FANTOM4 study, the consortium knocked down 52 transcription factors in the AML cell line THP-1 (15) and measured changes in the transcriptome. We reasoned that if the target genes of the transcription factors overlap significantly with signature genes of a signaling pathway, the pathway is likely to be upstream of that transcription factor. We therefore used this data to see whether SPEED can be used to systematically infer relationships between signaling pathways, transcription factors and regulated genes. We extracted 52 lists of differentially expressed genes corresponding to each transcription factor knockdown (see 'Materials and Methods' section for details) and ran SPEED to identify upstream signaling pathways. SPEED identified at least one signaling pathway upstream of 24 transcription factors. Table 2 describes the top-ranking signaling pathway for each transcription factor (see Supplementary Tables S2 and S3 for all predictions). While for some of the factors, the identified pathway is

Table 2. Identification of signaling pathways upstream of transcription factor in THP-1 cells

Top-ranked pathway	Transcription factor
MAPK+PI3K	CEBPA (27), EGR1, (28), ETS2 (29), FLI1 (30), MLLT3, MYBL2 (31), NFE2L1, NFYA, NRAS[a], SP1 (32), ZNF238
JAK-STAT	FOXJ3, FOXP1, GFI1 (33), MXI1, RUNX1 (34), STAT1 (35), YY1
TLR	BMI1, CEBPG, GATA2 (36), IRF8 (37), SPI1 (38)
TGF-β	MYB (39)

SPEED was run programatically on lists of differentially expressed genes following 52 transcription factor knockdowns. For 24 transcription factors, the corresponding differentially expressed gene lists had an overrepresentation of at least one upstream signaling pathway. The top-ranking signaling pathway with a minimum FDR of 0.05 is listed for each of the 24 transcription factors. Literature references for transcription factors known to be associated with identified signaling pathways are noted.
[a]Although NRAS is not a transcription factor, the significant overlap with MAPK+PI3K signature genes serves as validation of the SPEED algorithm because NRAS is upstream of both the MAPK and PI3K signaling pathways (25,26).

known to be upstream (indicated in the table), for others our results are predictions that may provide starting points for experiments.

DISCUSSION

The analysis of gene lists obtained from high-throughput experiments is a classical problem in bioinformatics. Most approaches, however, search for overrepresented functions, processes or pathways in the regulated gene group and are thus only suitable to identify the affected signaling pathways and not the pathways that caused the observed changes in gene expression. With SPEED, we focus on targets of signaling pathways by automatically deriving signature genes, i.e. genes that are typically regulated in a variety of cell types when the activity of a specific pathway is altered. While existing web servers also compare user gene lists with disease related or single-experiment signatures derived from micro-array data (16–21), SPEED, to our knowledge, is the only web server that integrates heterogeneous micro-array data for the purpose of identifying causal signaling pathway relationships. Additionally, SPEED signature gene lists can be readily incorporated into established gene list comparison tools like Gene Set Enrichment Analysis (22).

The number of pathways in SPEED is currently limited by publicly available gene-expression data from pathway perturbation experiments. Currently, SPEED contains 11 signaling pathways. However, the presence of pathway perturbation data in GEO suggests that these 11 pathways are of special interest to researchers and are thus the major signaling pathways responsible for many phenotypes studied using gene-expression technologies. Nevertheless, we expect that the number of SPEED pathways will increase in the future as micro-arrays continue to become more affordable. Furthermore, we can readily incorporate quantitative next-generation sequencing data as the field matures. To facilitate user suggestions, we have included a form on the web server for recommending new pathways or data.

It is important to note that not all pathway perturbation data should be included in SPEED. The reliability and biological meaning of SPEED signatures crucially depends on the nature of the experiments. For example, we aimed at selecting experiments that measured gene expression at early time points in order to reduce any indirect or long-term effects. To further reduce indirect effects, we selected experiments such that single pathways were primarily perturbed. While some perturbations may affect more than a single pathway, the reproducibility over multiple experiments via the overlap parameter can account for noise in the data as long as the pathway is the primarily perturbed one. However, if pathways are strongly coupled or converge on the same transcription factors, such as the MAPK and PI3K/AKT pathways (23), it is necessary to combine them into a single pathway description. The SPEED pathway labeled MAPK+PI3K is an example of combining coupled pathways.

Validation using literature-derived gene lists indicated that SPEED performs well in predicting upstream causal pathway-perturbation events with a sensitivity of 78%. However, it is possible that other cellular processes also regulate the same genes as SPEED signaling pathways and are the actual causal influences. Since the predicted signaling pathways are derived based on correlations with gene signatures, results should be considered as a starting point for further investigation into the actual events leading to the regulated set of genes.

The identification of signaling pathways that cause observed changes in gene expression is important, since aberrant signaling pathways have been implicated in diseases such as diabetes (24) and several cancers (25,26). We exemplify the utility of SPEED by identifying upstream signaling pathways in two AML studies. For high-risk AML patients with and without CEBPa mutations, SPEED identified TNF-α signaling as the top candidate that caused the observed differences between the two groups of patients. Given that high-risk AML patients with CEBPa mutations have better clinical outcomes, SPEED results may be taken as a starting point to narrow down the search for the differences in outcome by focusing on differences in TNF-α signaling.

Using micro-array data from knockdown of transcription factors in an AML cell line, we aimed at identifying links between transcription factors and upstream signaling pathways. We searched for associations between signature genes of signaling pathways and the targets of the transcription factors. For 24 factors, we could establish such associations, of which some are already known, whereas others, like MAPK+PI3K to ZNF238, are novel

predictions. Thus, SPEED facilitates completing the picture of cell signaling events leading to differential expression of genes via a layer of transcription factors.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Rositsa Koleva and Eric D. Smith for helpful discussions.

FUNDING

Germany's Federal Ministry for Education and Research (BMBF) (grant Forsys Partner to N.B.); the European Commission (grant number CancerSys HEALTH-F4-2008-223188 to N.B.); National Science Foundation Integrative Graduate Education and Research Traineeship (IGERT) (grant number DGE-0654108 to J.P.). Funding for open access charge: National Science Foundation Integrative Graduate Education and Research Traineeship (IGERT) (grant number DGE-0654108).

Conflict of interest statement. None declared.

REFERENCES

1. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
2. Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.
3. Legewie,S., Herzel,H., Westerhoff,H.V. and Bluthgen,N. (2008) Recurrent design patterns in the feedback regulation of the mammalian signalling network. *Mol. Syst. Biol.*, **4**, 190.
4. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.
5. Barrett,T. and Edgar,R. (2006) Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*. *Methods Mol. Biol.*, **338**, 175–190.
6. Sean,D. and Meltzer,P.S. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, **23**, 1846–1847.
7. Vallabhapurapu,S. and Karin,M. (2009) Regulation and function of NF-kappaB transcription factors in the immune system. *Annu. Rev. Immunol.*, **27**, 693–733.
8. Bluthgen,N., Kielbasa,S.M. and Herzel,H. (2005) Inferring combinatorial regulation of transcription in silico. *Nucleic Acids Res.*, **33**, 272–279.
9. Chang,J.T. and Nevins,J.R. (2006) GATHER: a systems approach to interpreting genomic signatures. *Bioinformatics*, **22**, 2926–2933.
10. Thomas,P.D., Campbell,M.J., Kejariwal,A., Mi,H., Karlak,B., Daverman,R., Diemer,K., Muruganujan,A. and Narechania,A. (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.*, **13**, 2129–2141.
11. Alibes,A., Yankilevich,P., Canada,A. and Diaz-Uriarte,R. (2007) IDconverter and IDClight: conversion and annotation of gene and protein IDs. *BMC Bioinformatics*, **8**, 9.
12. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
13. Consortium,U. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
14. Marcucci,G., Maharry,K., Radmacher,M.D., Mrozek,K., Vukosavljevic,T., Paschka,P., Whitman,S.P., Langer,C., Baldus,C.D., Liu,C.G. *et al.* (2008) Prognostic significance of, and gene and microRNA expression signatures associated with, CEBPA mutations in cytogenetically normal acute myeloid leukemia with high-risk molecular features: a Cancer and Leukemia Group B Study. *J. Clin. Oncol.*, **26**, 5078–5087.
15. Suzuki,H., Forrest,A.R., van Nimwegen,E., Daub,C.O., Balwierz,P.J., Irvine,K.M., Lassmann,T., Ravasi,T., Hasegawa,Y., de Hoon,M.J. *et al.* (2009) The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat. Genet.*, **41**, 553–562.
16. Culhane,A.C., Schwarzl,T., Sultana,R., Picard,K.C., Picard,S.C., Lu,T.H., Franklin,K.R., French,S.J., Papenhausen,G., Correll,M. *et al.* GeneSigDB–a curated database of gene expression signatures. *Nucleic Acids Res.*, **38**, D716–D725.
17. Wu,J., Qiu,Q., Xie,L., Fullerton,J., Yu,J., Shyr,Y., George,A.L. Jr and Yi,Y. (2009) Web-based interrogation of gene expression signatures using EXALT. *BMC Bioinformatics*, **10**, 420.
18. Zhang,S.D. and Gant,T.W. (2008) A simple and robust method for connecting small-molecule drugs using gene-expression signatures. *BMC Bioinformatics*, **9**, 258.
19. Cahan,P., Ahmad,A.M., Burke,H., Fu,S., Lai,Y., Florea,L., Dharker,N., Kobrinski,T., Kale,P. and McCaffrey,T.A. (2005) List of lists-annotated (LOLA): a database for annotation and comparison of published microarray gene lists. *Gene*, **360**, 78–82.
20. Newman,J.C. and Weiner,A.M. (2005) L2L: a simple tool for discovering the hidden significance in microarray expression data. *Genome Biol.*, **6**, R81.
21. Rhodes,D.R., Kalyana-Sundaram,S., Mahavisno,V., Varambally,R., Yu,J., Briggs,B.B., Barrette,T.R., Anstet,M.J., Kincead-Beal,C., Kulkarni,P. *et al.* (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*, **9**, 166–180.
22. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *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.
23. Tullai,J.W., Schaffer,M.E., Mullenbrock,S., Kasif,S. and Cooper,G.M. (2004) Identification of transcription factor binding sites upstream of human genes regulated by the phosphatidylinositol 3-kinase and MEK/ERK signaling pathways. *J. Biol. Chem.*, **279**, 20167–20177.
24. Jauregui,A., Mintz,D.H., Mundel,P. and Fornoni,A. (2009) Role of altered insulin signaling pathways in the pathogenesis of podocyte malfunction and microalbuminuria. *Curr. Opin. Nephrol. Hypertens*, **18**, 539–545.
25. Gottesman,M.M. (1994) Report of a meeting: molecular basis of cancer therapy. *J. Natl Cancer Inst.*, **86**, 1277–1285.
26. Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.
27. Tomizawa,M. and Saisho,H. (2007) Insulin-like growth factor (IGF)-II regulates CCAAT/enhancer binding protein alpha expression via phosphatidyl-inositol 3 kinase in human hepatoblastoma cell lines. *J. Cell Biochem.*, **102**, 161–170.
28. Sgambato,V., Pages,C., Rogard,M., Besson,M.J. and Caboche,J. (1998) Extracellular signal-regulated kinase (ERK) controls immediate early gene induction on corticostriatal stimulation. *J. Neurosci.*, **18**, 8814–8825.
29. Yang,B.S., Hauser,C.A., Henkel,G., Colman,M.S., Van Beveren,C., Stacey,K.J., Hume,D.A., Maki,R.A. and Ostrowski,M.C. (1996) Ras-mediated phosphorylation of a conserved threonine residue enhances the transactivation activities of c-Ets1 and c-Ets2. *Mol Cell Biol*, **16**, 538–547.
30. Jinnin,M., Ihn,H., Mimura,Y., Asano,Y., Yamane,K. and Tamaki,K. (2005) Matrix metalloproteinase-1 up-regulation by hepatocyte growth factor in human dermal fibroblasts via ERK

signaling pathway involves Ets1 and Fli1. *Nucleic Acids Res.*, **33**, 3540–3549.
31. Hanada,N., Lo,H.W., Day,C.P., Pan,Y., Nakajima,Y. and Hung,M.C. (2006) Co-regulation of B-Myb expression by E2F1 and EGF receptor. *Mol. Carcinog.*, **45**, 10–17.
32. Milanini-Mongiat,J., Pouyssegur,J. and Pages,G. (2002) Identification of two Sp1 phosphorylation sites for p42/p44 mitogen-activated protein kinases: their implication in vascular endothelial growth factor gene transcription. *J. Biol. Chem.*, **277**, 20631–20639.
33. Rodel,B., Tavassoli,K., Karsunky,H., Schmidt,T., Bachmann,M., Schaper,F., Heinrich,P., Shuai,K., Elsasser,H.P. and Moroy,T. (2000) The zinc finger protein Gfi-1 can enhance STAT3 signaling by interacting with the STAT3 inhibitor PIAS3. *EMBO J.*, **19**, 5845–5855.
34. Sheng,Z., Wang,S.Z. and Green,M.R. (2009) Transcription and signalling pathways involved in BCR-ABL-mediated misregulation of 24p3 and 24p3R. *EMBO J.*, **28**, 866–876.
35. Hu,X. and Ivashkiv,L.B. (2009) Cross-regulation of signaling pathways by interferon-gamma: implications for immune responses and autoimmune diseases. *Immunity*, **31**, 539–550.
36. Liu,W., Zhu,Z.Q., Wang,W., Zu,S.Y. and Zhu,G.J. (2007) Crucial roles of GATA-2 and SP1 in adrenomedullin-affected expression of tissue factor pathway inhibitor in human umbilical vein endothelial cells exposed to lipopolysaccharide. *Thromb. Haemost.*, **97**, 839–846.
37. Laricchia-Robbio,L., Tamura,T., Karpova,T., Sprague,B.L., McNally,J.G. and Ozato,K. (2005) Partner-regulated interaction of IFN regulatory factor 8 with chromatin visualized in live macrophages. *Proc. Natl Acad. Sci. USA*, **102**, 14368–14373.
38. Ishii,J., Kitazawa,R., Mori,K., McHugh,K.P., Morii,E., Kondo,T. and Kitazawa,S. (2008) Lipopolysaccharide suppresses RANK gene expression in macrophages by down-regulating PU.1 and MITF. *J. Cell Biochem.*, **105**, 896–904.
39. Classen,S., Zander,T., Eggle,D., Chemnitz,J.M., Brors,B., Buchmann,I., Popov,A., Beyer,M., Eils,R., Debey,S. *et al.* (2007) Human resting CD4+ T cells are constitutively inhibited by TGF beta under steady-state conditions. *J. Immunol.*, **178**, 6931–6940.