Published online 20 April 2020 *Nucleic Acids Research, 2020, Vol. 48, Web Server issue* ***W307–W312***
doi: 10.1093/nar/gkaa236

SPEED2: inferring upstream pathway activity from differential gene expression

Mattias Rydenfelt[1,2,†], Bertram Klinger[1,2,†], Martina Klünemann[1,2] and Nils Blüthgen[1,2,*]

[1]Institute of Pathology, Charite - Universitätsmedizin Berlin, Berlin 10117, Germany and [2]IRI Life Sciences, Humboldt University of Berlin, Berlin 10115, Germany

Received February 17, 2020; Revised March 25, 2020; Editorial Decision March 29, 2020; Accepted April 16, 2020

ABSTRACT

Extracting signalling pathway activities from transcriptome data is important to infer mechanistic origins of transcriptomic dysregulation, for example in disease. A popular method to do so is by enrichment analysis of signature genes in e.g. differentially regulated genes. Previously, we derived signatures for signalling pathways by integrating public perturbation transcriptome data and generated a signature database called SPEED (Signalling Pathway Enrichment using Experimental Datasets), for which we here present a substantial upgrade as SPEED2. This web server hosts consensus signatures for 16 signalling pathways that are derived from a large number of transcriptomic signalling perturbation experiments. When providing a gene list of e.g. differentially expressed genes, the web server allows to infer signalling pathways that likely caused these genes to be deregulated. In addition to signature lists, we derive 'continuous' gene signatures, in a transparent and automated fashion without any fine-tuning, and describe a new algorithm to score these signatures.

INTRODUCTION

When interpreting transcriptome data, it is often important to infer which signalling pathway triggered a particular gene expression program. A common method to infer pathway activity is to score gene sets. However, most gene sets contain pathway members, and not downstream genes and are not suitable to score signalling pathway activity. In addition, collections of target or signature genes are being used to score pathway activity—such as the Hallmark collection of the Molecular Signatures Database (MSigDB) (1)—yet they are typically derived from one experimental context or cell type.

About a decade ago, we introduced the concept of systematically deriving consensus signatures from signalling perturbation experiments by integrating a large body of available experimental data (2). We have previously demonstrated the benefits of this approach (2,3) compared to signatures derived from single experiments or pathway membership in inferring causal links between differential expression pattern and activated pathway. With the resulting gene signature database SPEED (Signalling Pathway Enrichment using Experimental Datasets) now reaching the 10-year mark and still serving a continuously growing user base we decided to develop an upgrade: SPEED2.

Our new implementation presents a number of substantial improvements: (i) expanding the number of pathways from 11 to 16, (ii) almost tripled the amount of data used for training from 215 to 640 data sets, (iii) using 'continuous' gene signatures spanning the full transcriptional range allowing users to simultaneously test their gene lists for both up- and down regulation, (iv) providing statistical scores to further filter the predictions for highest confident genes and (v) embedding those functions in a newly developed modern web interface.

MATERIALS AND METHODS

For SPEED2, we manually collected 640 publically available transcriptome experiments where signalling pathways were perturbed. These data sets were integrated into ranked gene signature lists for each pathway. The resulting ranked signatures were then used to score gene lists and report enrichment results. An overlap between the top signature genes and the query list is calculated and provided as table. These four steps and the implementation of the web server are described in detail in the following.

Data acquisition and differential expression scoring

We collected manually curated publicly accessible microarray experiments stored in the Gene Expression Omnibus database (GEO (5,6)). We specifically selected those experiments where a perturbation was conducted such that it was ensured to drive only a single pathway (e.g. a single specific ligand/knockdown of a key mediator) and focused on experiments where the transcriptional changes are not yet

*To whom correspondence should be addressed. Tel: +49 30 2093 92390; Email: nils.bluethgen@charite.de
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

dominated by secondary responses (i.e. short term stimulation). Eligible pairs of experiments consisting of control and single perturbation experiments were downloaded and normalized after which fold-changes were calculated and transformed into z-scores as described previously (2). A list of all experiments that were used, as well as the raw and unnormalized data can be downloaded from the Help page of the SPEED2 website.

Generating continuous gene signatures

For each perturbation transcriptome data set, we discarded probesets whose expression was below the median expression value, as well as probes with no matching (or obsolete) Entrez Gene ID. If a probe corresponded to multiple Entrez Gene IDs, we picked the lowest ID, and if multiple probesets matched to the same Entrez Gene ID, we kept only the most highly expressed probesets for further analysis. Since perturbation experiments could be either of activatory or of inhibitory nature, we chose, by convention, *z*-values to be positive for up-regulated genes in activating perturbation experiments. To normalise the different magnitudes of z-values in different experiments, we calculated a gene-wise score per experiment by mapping the ranked z-values to a uniform interval [–1,1]. For each pathway, we then calculated the average score per gene across all N experiments for a pathway, and compared it with a mean of N uniformly distributed variables on [–1,1] as null model (Bates distribution). The resulting *P*-values quantify if a given gene is consistently activated/repressed by the pathway's activity and are the core of the SPEED2 analysis pipeline.

SPEED2 analysis

Using the pathway-specific *P*-values calculated for each gene, we calculated a uniformly distributed score between –1 and 1 that ranks genes based on their significance and whether they are on average up- or down-regulated. These genome-wide ranked scores per pathway are used to perform enrichment analysis on the user-provided gene list. To measure enrichment, the SPEED2 web interface supports two types of statistical tests: 'Bates test', detecting shifts in mean rank away from 0 compared to the Bates distribution on [–1,1], and approximate χ^2-test, detecting shifts in variance away from the expected value for a uniform distribution on [–1,1], using a normal approximation.

Candidate genes

To extract representative candidates we also generated signature genes per pathway, which we defined as the top 300 genes in each signature that had a FDR-corrected *P*-value below ($q < 0.05$). After running the SPEED2 analysis, the web server returns a table with the overlap of the query genes and the signature genes that can be starting points for further experimental analyses or validation.

Implementation

SPEED2 runs on an Apache 2 linux web server, with load balancing for the calls to the API. The functionality on the client site of the website is implemented using JavaScript (js), making use of the JQuery, Bootstrap, DataTables js libraries. The backend is implemented in R, and exposes a REST API by using the Plumber R library. All calls to the API are using the http post method. Query gene lists are limited to 500. The R-functions and the signature data base are also available as R-package speed2 available on https://github.com/molsysbio/speed2.

RESULTS

Performing signalling pathway enrichment analysis

Based on previous ideas (2), we have developed a data base and web service to perform signalling pathway enrichment analysis. The basis of this enrichment analysis is a pathway-specific signature data base. The focus of this data base are signalling pathways that are triggered by extra-cellular cues, such as hormones, receptors or external stresses. To compile this data base, we manually compiled gene expression data sets where pathways were perturbed and changes in gene expression was measured shortly afterwards. We manually selected those pathways for which we found a larger number of appropriate perturbation experiments ($\geq$10). This data base consists of ranked gene lists for each of the 16 distinct signalling pathways, each derived from multiple perturbation data sets (between 10 and 113) with transcriptome readout (cf. Figure 1, bottom left). Briefly, we scored consistently up- and down-regulated genes after pathway perturbation. The ranked lists then contain the most significantly up-regulated genes, as determined by *P*-value, on one end of the list, and the most significantly down-regulated genes mapped to the other end (see Materials and Methods). Afterwards the ranks where converted to scores of a uniform distribution between –1 and 1, where the sign denotes down- and up-regulation, respectively.

At the SPEED2 website, users can enter gene lists of interest, e.g. differentially expressed human genes as gene symbols or Entrez Gene IDs, and SPEED2 quantifies if these genes are enriched for strongly pathway-deregulated signature genes (see Figure 1, top right). To determine pathway signature enrichment, SPEED2 offers two options that should be chosen according to the question. If the gene list contains either up- or down-regulated genes, one can choose the Bates test that quantifies shift in mean rank. In contrast, if the user supplied gene list contains both up- and down-regulated genes, the approximate χ^2-test is more appropriate as this scores highly if the supplied genes are accumulating at both ends of the distribution. In most cases the Bates test is more powerful, however, if the user provides a gene list of about an equal number of up- and down-regulated genes, a scenario which can be identified through the 'barcode' visualization in SPEED2, it is recommend to use the χ^2-test. In Supplementary Figures S1– S3 we compare the Bates test, χ^2-test, and GSEA (7), with no obvious advantage of the latter approach (8).

When the analysis is finished, the results are reported for each pathway as a bar graph denoting mean rank of the query list, as well as a 'barcode' plot, showing the distribution of the query genes in the ranked signatures (see Figure 1). Colors show FDR-adjusted *P*-values. In addition to the

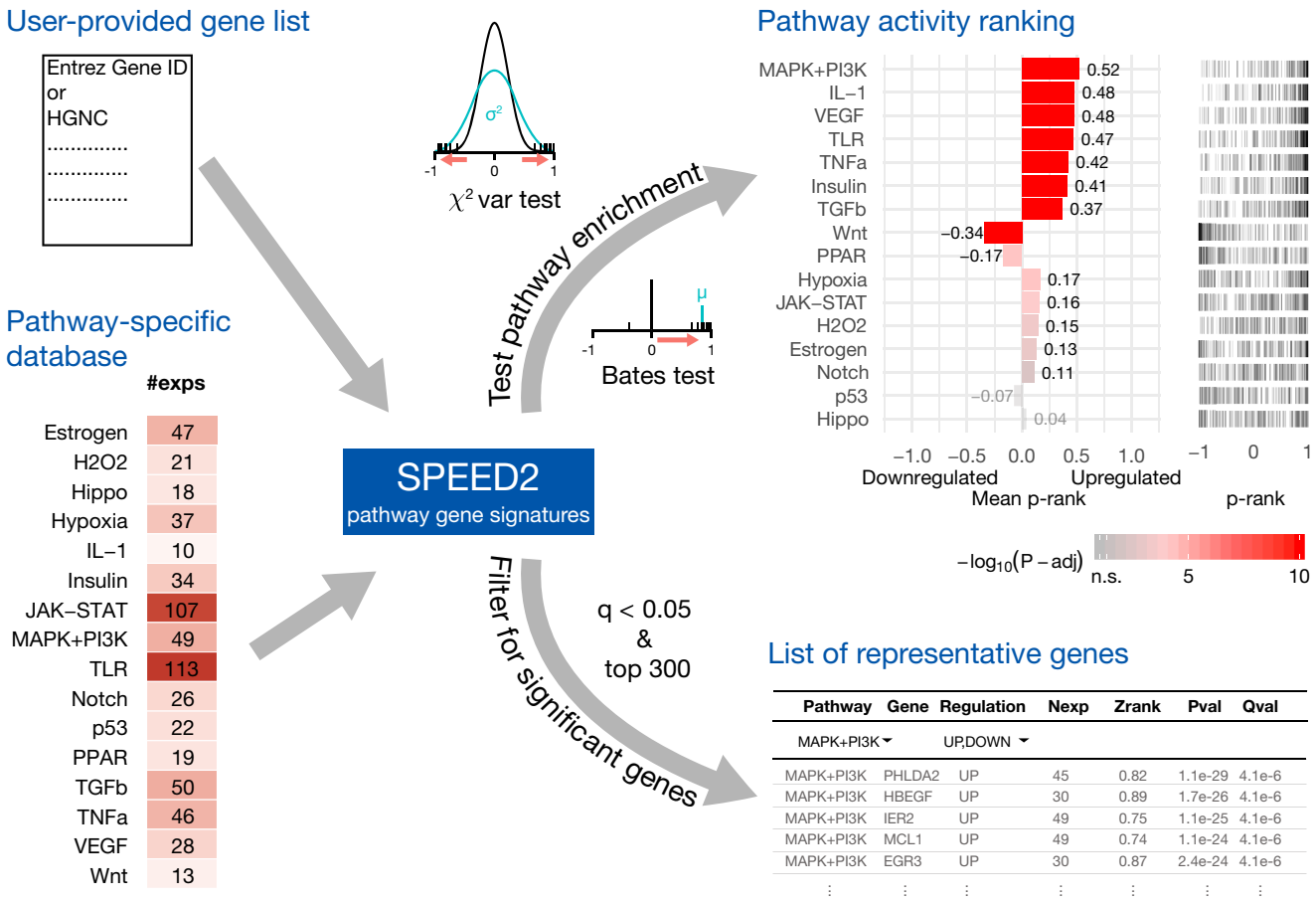

Figure 1. Overview of SPEED2 application We manually selected pathway-specific perturbation experiments and estimated their pathway-relevance. For each experiment *z*-scores are mapped on a scale between –1 and 1 (Zrank) and significance was asserted per gene and pathway by testing the Zranks against a uniform null model (*p*) and corrected for multiple testing (*q*). These pathway-specific significance measures (along with their regulation direction) are now used to evaluate a user-provided gene list for two aspects: (i) testing for pathway enrichment by deviation from the uniform mean (Bates test) or uniform variance (χ^2 var test) on *P*-rank ordered continuous pathway signatures and (ii) filtering the list of genes for pathway-representatives ranked by *q*-value (see main text). The first result gives indications of upstream signalling that might have caused the gene expression change and the second result provides candidate genes to e.g. carry out follow up investigations. We illustrate these SPEED2 outputs on a well defined MAPK target list from Uhlitz *et al*. (4).

visualization of the enrichment, the website offers to download the results as comma separated values (.csv) file, allowing for further analyses. Furthermore, the website reports a table containing all significant signature genes for each pathway that overlap with the query genes. This table can be interactively explored, or also downloaded as .csv file.

Signature characterization

To benchmark SPEED2, we derived 45 independent gene lists for 8 different pathways based on data, not used to generate SPEED2 signatures (Supplementary File 1). For each of these test sets, we determined if the list contained only up- or down-regulated genes, or generally contained target genes irrespective of the direction of regulation by the pathway. We then performed SPEED2 analysis on these lists, using the appropriate statistics (Bates for lists of up- or down-regulated genes, χ^2 for lists containing both up- and down-regulated genes). We marked the number of times the correct pathway was the first, second or third most enriched pathway in SPEED2 and found in 56% of cases the upstream stimuli was ranked first, and in 87% of cases within the top 3 predictions (Figure 2A, left).

Next to independent gene sets we also conducted an analysis for each single experiment of the SPEED2 database for both methods (Supplementary Figures S4 and S5). On average we find similar scoring metrics for the first three ranks as in the independent testset benchmark. We further investigated the dependency on input size and found both scoring methods to be largely robust to input size, with slightly worse performances for small (<50) and large gene sets (>450). Therefore for optimal performance we recommend as input a list of 100–400 genes and have limited the input size to 500. We further noted that on average the Bates test performed slighty better on those benchmarks than the χ^2 test prompting us to suggest the Bates test as the default test on the website.

Due to extensive cross-talk between signalling pathways, stimuli often cross-activate multiple pathways, and thus appearance of multiple top pathways are expected. When visualizing Spearman correlation for mutually significant genes ($P < 0.05$) across pathway-pairs (Figure 2B), we confirmed

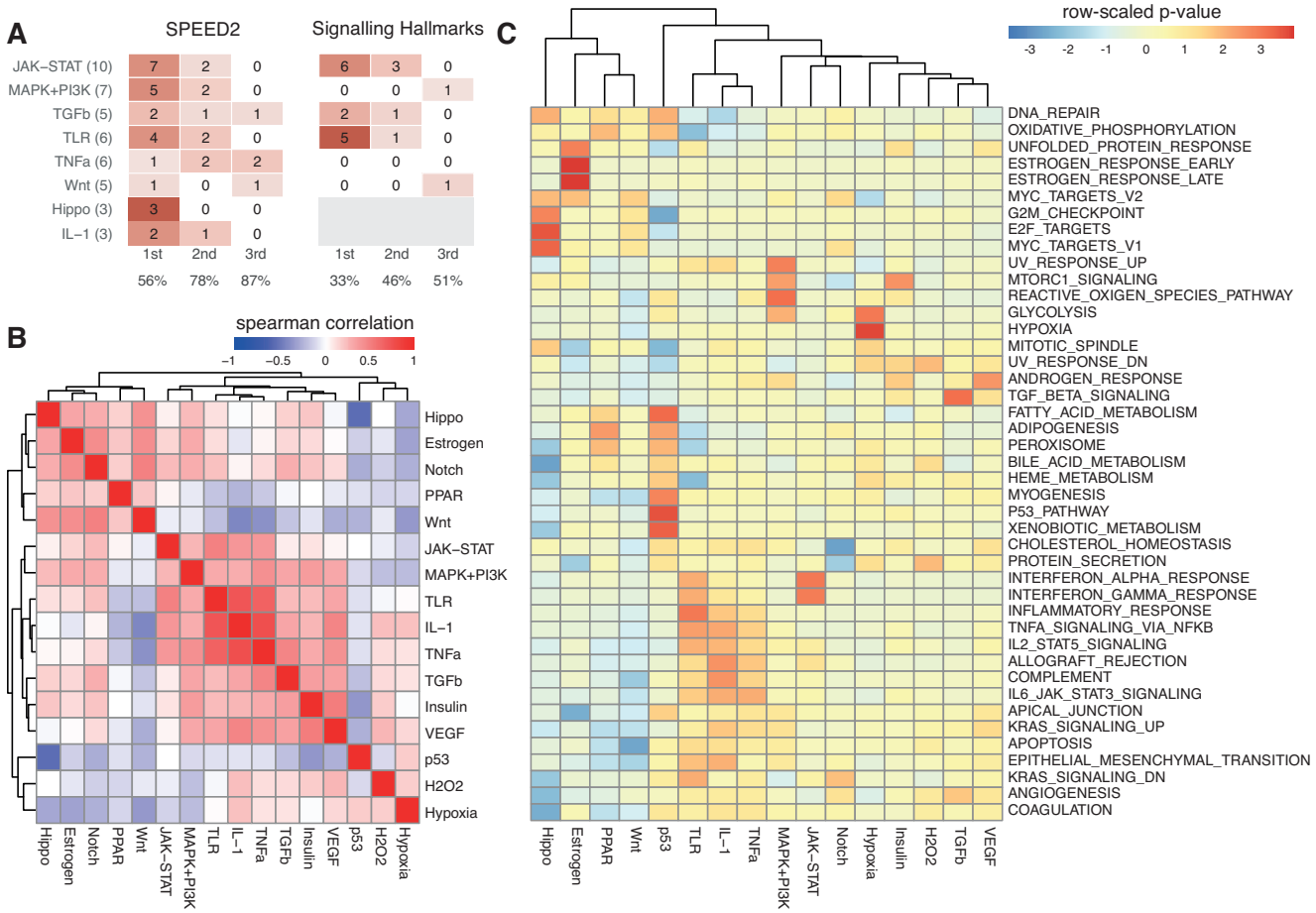

Figure 2. Signature characterization. (**A**) Number of times that regulated pathways of externally curated data sets (total number of benchmarks per pathway in brackets) occurred in the top 3 most enriched pathways in SPEED2 analysis contrasted to the top ranking of the best assigned signalling pathway signatures from the Hallmarks collection using Fisher's exact test (Hippo and IL-1 were not scored as no signalling Hallmark could be assigned). (**B**) Spearman correlation of mutually significant genes ($P < 0.05$) indicates three general signalling groups. (**C**) Scoring of Broad Hallmark signatures by SPEED2 with at least one pathway being more significant than adjusted $P < 0.001$; colors indicate row-scaled adjusted P-value (before scaling sign was set to 1 and –1 for up and down-regulation, respectively), see also Supplementary Figure S1.

that several signatures are highly correlated, with particularly strong correlation between the IL-1, TNFa and TLR pathways.

To further characterize our gene signatures we analysed the MSigDB Hallmark sets with SPEED2 (using the Bates test) finding a substantial agreement for the IL-1, JAK-STAT, TNFa, TLR pathway family, as well as Estrogen, Hypoxia, TGFb and p53 (Figure 2C). For certain other pathways, like Wnt, Notch or MAPK/KRAS, there was little or no agreement between the assigned signatures. This finding is further corroborated, when we performed enrichment analysis on our independent test sets using the MSigDB Hallmark sets (2A, right). We found that enrichment using the signalling-related Hallmark sets generally performed less well compared to SPEED2, particularly for those signatures where SPEED2 and Hallmark signatures diverge, suggesting that SPEED2 signatures are more potent to score signalling pathway deregulation.

To further compare our tool with existing tools and databases we applied our benchmarks on the comprehensive geneset database collection of the Enrichr webtool (9). When scoring the top 4 performing databases representing pathway enrichment: Bioplanet 2019 (10), WikiPathways 2019 Human (11), KEGG 2019 Human (12) and Panther 2016 (13), we note that each tool individually is outperformed by SPEED2 (Supplementary Figure S6). In contrast to SPEED2 those top 4 scoring pathway databases predominantly contain genes encoding for proteins that are important in the signalling relay of the pathway and have little information on downstream transcriptional targets. This indicates that in order to next to pathway membership also encompass causal upstream signalling SPEED2 might be a useful addition to consider for the Enrichr suite.

DISCUSSION

Signaling pathways control the expression of hundreds of target genes by which they influence virtually all major cell fate decisions. Determining causative upstream signalling from gene expression changes is a major strategy to mechanistically understand why cells changed their transcriptome, e.g. in disease. As signaling pathways are mostly regulated post-translationally, direct pathway membership of differentially expressed transcripts is often misleading, render-

ing enrichment analysis with gene ontology (14) and pathway data bases such as Reactome (15) inefficient for this question. A viable strategy however is to determine signature genes that are consistently regulated when a pathway is activated or inhibited. Databases like MSigDB collect signatures of individual or manually curated pathway targets (16). These collections contain important signatures that are helpful in many contexts. However, the quality of the signature varies and depends largely on curation.

About a decade ago, we introduced the concept of an automatic unbiased extraction of signatures by integrating a large number of public transcriptome experiments where pathways were perturbed. This led to the predecessor of SPEED2, the SPEED web tool (https://speed.sys-bio.net/). As the data base of publically available transcriptome data increased dramatically since then, we again set out and collected data from perturbation experiments, and were able to increase the data base by a factor of three. In our earlier version (SPEED), we had MAPK and PI3K as separate pathways, but as they are heavily intertwined and PI3K has primarily post-transcriptional targets, we decided to drop these as separate pathways and retain the MAPK+PI3K pathway. We added seven additional pathways (Estrogen, Hippo, Hypoxia, Insulin, Notch, P53 and PARP) to broaden the scope of potential analysis. Additionally, we decided to use transcriptome-wide ranked signatures that allowed for more sophisticated statistical enrichment analyses instead of co-occurrence. Our benchmarking show that this approach performs well for gene lists that are derived from single perturbation experiments, when the appropriate statistics is used.

Note that when scoring pathway activity from baseline transcriptome measurements, we and others recently established the R-package PROGENy (17), which uses a regression based score trained on perturbation data. This R-package uses gene expression quantification of the whole transcriptome, whereas the SPEED2 analysis conducts an enrichment analysis on a list of genes pre-filtered by the user for e.g. differential expression and uses more distinct pathways. Therefore these two applications complement each other depending on the input data available.

In addition to a web tool, we also developed an R-library that allows to perform SPEED2 analysis programmatically. This library can be downloaded at the SPEED2 website and at https://github.com/molsysbio/speed2. In addition to querying SPEED2 signatures, the R-package also allows to use custom signatures.

With the extension provided by SPEED2 such as the distinction of up and down-regulation, addition of more relevant pathways, increasing data depth and suggesting follow-up candidates we hope to provide an even better service to the scientific community. Additionally, the rich manually curated databases on which SPEED2 is build on will allow to build new tools. For instance, SPEED2 is currently build as an enrichment tool, which allows to visualize the query gene groups in the ranked signatures. A complementarily approach would be to formulate the problem as a classification problem and train a classifier that predicts the most likely deregulated upstream pathway from gene lists.

To conclude, SPEED2 provides a convenient way to score signalling activity for sets of dysregulated genes obtained from transcriptome analysis.

AVAILABILITY AND IMPLEMENTATION

SPEED2 is freely available at https://speed2.sys-bio.net as an online web service. All raw data can be downloaded from the website, and an R-package (speed2) is available at https://github.com/molsysbio/speed2.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

BMBF [ZiSS-Trans, 02NUK047E]. Funding for open access charge: BMBF
Conflict of interest statement. None declared.

REFERENCES

1. Liberzon,A., Birger,C., Thorvaldsdottir,H., Ghandi,M., Mesirov,J. and Tamayo,P. (2015) The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.*, **1**, 417–425.
2. Parikh,J.R., Klinger,B., Xia,Y., Marto,J.A. and Blüthgen,N. (2010) Discovering causal signaling pathways through gene-expression patterns. *Nucleic Acids Res.*, **38**, W109–W117.
3. Cantini,L., Calzone,L., Martignetti,L., Rydenfelt,M., Blüthgen,N., Barillot,E. and Zinovyev,A. (2017) Classification of gene signatures for their information value and functional redundancy. *NPJ Syst. Biol. Appl.*, **4**, 2.
4. Uhlitz,F., Sieber,A., Wyler,E., Fritsche-Guenther,R., Meisig,J., Landthaler,M., Klinger,B. and Blüthgen,N. (2017) An immediate-late gene expression module decodes ERK signal duration. *Mol. Syst. Biol.*, **13**, 928.
5. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
6. Edgar,R., Domrachev,M. and Lash,A. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
7. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R. and Lander,E.S. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
8. Irizarry,R.A., Wang,C., Zhou,Y. and Speed,T.P. (2009) Gene set enrichment analysis made simple. *Stat. Methods Med. Res.*, **18**, 565–575.
9. Kuleshov,M.V., Jones,M.R., Rouillard,A.D., Fernandez,N.F., Duan,Q., Wang,Z., Koplev,S., Jenkins,S.L., Jagodnik,K.M., Lachmann,A. *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, **44**, W90–W97.
10. Huang,R., Grishagin,I., Wang,Y., Zhao,T., Greene,J., Obenauer,J.C., Ngan,D., Nguyen,D.-T., Guha,R., Jadhav,A. *et al.* (2019) The NCATS BioPlanet ? An integrated platform for exploring the universe of cellular signaling pathways for toxicology, systems biology, and chemical genomics. *Front. Pharmacol.*, **10**, 445.
11. Slenter,D.N., Kutmon,M., Hanspers,K., Riutta,A., Windsor,J., Nunes,N., Melius,J., Cirillo,E., Coort,S.L., Digles,D. *et al.* (2017) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.*, **46**, D661–D667.
12. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

13. Mi,H. and Thomas,P. (2009) PANTHER Pathway: An Ontology-Based Pathway Database Coupled with Data Analysis Tools. *Methods Mol. Biol.*, **563**, 123–140.
14. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
15. Jassal,B., Matthews,L., Viteri,G., Gong,C., Lorente,P., Fabregat,A., Sidiropoulos,K., Cook,J., Gillespie,M., Haw,R. *et al.* (2019) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.
16. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdottir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.
17. Schubert,M., Klinger,B., Klünemann,M., Sieber,A., Uhlitz,F., Sauer,S., Garnett,M.J., Blüthgen,N. and Saez-Rodriguez,J. (2018) Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat. Commun.*, **9**, 20.