W78–W83 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* *Published online 2 June 2010*
doi:10.1093/nar/gkq482

R spider: a network-based analysis of gene lists by combining signaling and metabolic pathways from Reactome and KEGG databases

Alexey V. Antonov[1,*], Esther E. Schmidt[2], Sabine Dietmann[1], Maria Krestyaninova[2] and Henning Hermjakob[2]

[1]Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Institute for Bioinformatics and Systems Biology, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany and [2]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received January 29, 2010; Revised April 26, 2010; Accepted May 14, 2010

ABSTRACT

***R* spider is a web-based tool for the analysis of a gene list using the systematic knowledge of core pathways and reactions in human biology accumulated in the Reactome and KEGG databases. *R* spider implements a network-based statistical framework, which provides a global understanding of gene relations in the supplied gene list, and fully exploits the Reactome and KEGG knowledge bases. *R* spider provides a user-friendly dialog-driven web interface for several model organisms and supports most available gene identifiers. *R* spider is freely available at http://mips.helmholtz-muenchen.de/proj/rspider.**

INTRODUCTION

High-throughput technologies enable biological researchers to study hundreds or thousands of genes simultaneously. Genes or proteins are detected that are differentially expressed or co-expressed across varying cellular conditions. However, generating hypotheses about the underlying biological mechanisms based on experimentally derived gene/protein lists remains a non-trivial task for biologists. In 2002, a computerized analysis approach using the Gene Ontology (GO) was proposed to deal with this issue (1,2). Currently, there are over 25 tools performing this type of analysis with some variations (3–13). More recently, computational methods seek to interpret or at least visualize the pathway context of the experimentally derived genes (14–17). In this respect, one should mentioned a landmark procedure proposed recently in (17,18) which goes beyond gene pairs and fully captures the topology of signaling pathways by propagating the perturbations measured at gene levels through the entire pathway. However, the development of rigorous statistical methods for global network inference has been a challenging task.

Recently, we have introduced a network-based computational framework for the interpretation of gene/protein lists derived from high-throughput studies (19,20). Our approach overcomes a major bottleneck of the commonly employed methods for enrichment analysis (21) by providing network models that unite genes from different pathways into a single connected network. A Monte Carlo procedure was employed to estimate the significance of the inferred models, thus providing a rigorous quantitative statistical control (22). A web-based tool, KEGG spider (19), was introduced that exploits the network-based methodology for the exploration of metabolic reactions accumulated in the KEGG database (23). It was demonstrated that KEGG spider provides deeper insight into the genomic basis of metabolism variations in comparison to other tools (19).

Although being a powerful tool, KEGG spider is limited only to metabolism-related genes which cover <10% of the human genome (about 1100 genes). It is clear that many other important cellular processes, such as regulatory and signaling pathways remain uncovered by the inferred network models. On the other hand, the Reactome knowledgebase (24,25) is a dynamically expanding project, which provides high quality expert-authored, peer-reviewed knowledge of human reactions and pathways, covering 3916 human proteins (as of release 30). To provide experimentalists with an efficient web-based tool for the analysis of high-throughput data using Reactome knowledge, we have developed *R* spider, which implements the network-based methodology and exploits the data accumulated in the Reactome knowledgebase to the full extent. *R* spider unites both

*To whom correspondence should be addressed. Tel: +49 89 3187 2788; Fax: +49 (0) 89 3187 3585; Email: alexey.antonov@helmholtz-muenchen.de

Reactome and KEGG knowledge databases covering proteins from signaling and metabolism pathways.

We would like to point out that there are other signaling and metabolic databases available in the public domain like the manually curated BioCarta, NCI or inferred data (26) or (27). *R* spider has the option to switch between Reactome&KEGG, Nature Curated pathways (http://pid.nci.nih.gov/) and BioCarta (www.biocarta.com).

MATERIALS AND METHODS

A global Reactome protein network

Reactome (http://www.reactome.org/) is an expert-authored, peer-reviewed knowledgebase of human reactions and pathways. We used a file in tab-delimited format which specifies protein–protein interaction pairs derived from Reactome data (http://www.reactome.org/download/current/homo_sapiens.interactions.txt.gz). The meaning of 'interaction' is quite broad: two protein sequences occur in the same complex or they occur in the same or neighbouring reaction(s). For the human genome, the global Reactome protein network covers about 3700 proteins (including proteins from non-human species that interact with human proteins) involved in approximately 83 000 unique pairwise interactions (based on release 30).

A global metabolic gene network

The KEGG database is a collection of chemical structure transformation patterns for substrate–product pairs (reactant pairs). A detailed description of the procedure used to construct a global metabolic gene network can be found in ref. (19). The resulting global metabolic gene network links by edges any two genes that are associated with reactions sharing common compounds (from the main reaction pair). For the human genome, the global metabolic gene network covers about 1100 genes involved in approximately 15 000 unique pairwise interactions.

Integral reference network

To unite both networks, the Reactome protein network was transformed into a gene network. As in many cases, several proteins map to the same gene, the resulting gene network has a smaller number of nodes and edges. Once both KEGG and Reactome networks have the same type of node identifiers, they can be united. For the human genome, the resulting integral network covers about 3700 genes involved in approximately 50 000 unique pairwise gene interactions.

Network inference procedure and statistical treatment

Detailed information on the network inference and the Monte Carlo simulation procedure for computing *P*-values can be found in our previously published papers (19,20,28).

Initially, the genes from the input list are mapped to the global reference network. At this point, all nodes from the input list are disconnected. In the first step, all pairs of nodes with distance 1 are connected by edges and connected subnetworks are extracted. The subnetwork with the maximal number of nodes is referred to as an inferred network model D1. In the second step, the disconnected nodes from the input list with distance 2 are connected by edges. The subnetwork with the maximal number of input nodes is inferred and referred to as network model D2. In the next step, the disconnected nodes from the input list with distance 3 are connected by edges and a network model D3 (a subnetwork with the maximal number of input nodes) is inferred. Models D2 and D3 incorporate nodes that are not from the input list but are added to connect input nodes in the network model. We refer to these added nodes as intermediate or missing genes.

Let us assume that we have N genes from the input list to be mapped to the reference network. Next, we refer to the value N as the size of the input list. We infer the network models D1, D2, D3. Let us denote S1, S2, S3 to be the number of input nodes in the inferred network models. We also refer to S1, S2, S3 as the sizes of the respective models D1, D2, D3. Given the number of mapped genes from the input list (N), we consider the sizes of the models (values S1, S2, S3) as statistics. We have to estimate the probability to get models of the same or larger sizes from a randomly generated gene list which has N genes mapped to the reference network.

To generate the background distributions BD1, BD2, BD3 we repeat the following simulation procedure k times, where k specifies the upper significance level. A random gene list Lj of size N (equal to the size of the input list) is generated by sampling genes from global gene network. Index $j = 1 \ldots k$ specifies each of the k random simulations. The network inference procedure described above is applied to the random list Lj and the network models D1, D2, D3 are inferred. Let us denote the size (the number of input genes) of the inferred models D1, D2, D3 for the random list Lj as R1j, R2j, R3j. Thus, after repeating the simulation procedure k times, we get the background distribution R1j ($j = 1 \ldots k$) for models D1, the background distribution R2j ($j = 1 \ldots k$) for models D2, and the background distribution R3j ($j = 1 \ldots k$) for models D3.

To estimate significance of the inferred network model D1 for the input gene list, the value S1 is compared with the distribution R1j. Let n be the number of values from the distribution R1j that are equal or greater than S1. The estimate of P (P-value) of the inferred network model D1 is computed as $P = (n+1)/k$. In the same way the P-values for the model D2 and D3 are estimated.

Statistical treatment plays an important role for the quality control of inferred models. It is clear that given a gene list and a reference network, one can always infer some model, which will cover all genes from the list by relaxing the number of possible intermediate genes. There is a very simple test for any similar tool: the tool must be able to recognize a random gene list and return on average insignificant P-values for the random case. In 20 submissions of different randomly generated gene lists on average only 1 case is expected to be significant at the level of 0.05 (1/20). The estimate of the P-value provided by the Monte Carlo procedure corresponds exactly to the definition of

P-value: the probability to get a model of the same quality for a random gene list.

Enrichment of the reactome and KEGG canonical pathways

To compute enrichment of canonical Reactome and KEGG pathways, we also employed the Monte Carlo procedure. In this case, we randomly draw *k* genes (the number of genes in the input list) 100 times from the set of all genes (or from the background set of genes supplied by the user) and each time we estimate *P*-value based on the hypergeometric distribution for the best (whatever) pathway. Thus, we got a distribution of size 100 of the best *P*-values for a random drawing of k genes which we compare with the *P*-value for the best (whatever) pathway related to our original list. The estimate of the adjusted *P*-value by Monte Carlo procedure is given by the share of random simulations where the best *P*-value was equal or superior (less) than the *P*-value for the best (whatever) pathways related to our original gene list.

RESULTS

R spider (http://mips.helmholtz-muenchen.de/proj/rspider) is a freely available web-based tool that implements a pathway-free statistical framework for the interpretation of gene lists from high-throughput studies. *R* spider is available for several model organisms (*Mus musculus*, *Rattus norvegicus*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Drosophila melanogaster*). In addition, *R* spider has the option to switch to the other available in the public domain signaling pathway databases, Nature Curated pathways (http://pid.nci.nih.gov/) and BioCarta (www.biocarta.com).

R spider has a simple, user-friendly interface. As input it accepts several types of gene or protein identifiers, such as identifiers from 'Entrez Gene' (29), 'UniProt/Swiss-Prot' (30), 'Hugo Gene Symbols', 'UniGene', 'Ensembl' (31), 'RefSeq' (32) and 'Affymetrix' (33). As output, the user obtains network models (D1, D2, D3), where (1, 2, 3) indicates the maximal distance between any two input genes to be considered as 'connected' in the output model. The network model (D1, D2 or D3) represent a connected subnetwork with the maximal number of input genes. *R* spider provides a report on the statistical significance of the inferred network models (D1, D2, D3), as well as a catalog of the enriched Reactome or KEGG pathways. For each model (D1, D2, D3), a link is provided to obtain a graphical visualization. The visualization is performed by the Medusa package (34). We would like to point out that online visualization capabilities are limited. For this reason, we recommend to download the inferred network models as text files (links are provided on the visualization page) and to use freely available packages (Cytoscape, Meduza) for network visualization. Using these programs the users can produce high-quality figures (34,35).

Graphical output

In the graphical output, input genes are represented by rectangles and specified by the input gene Ids. Intermediate genes are represented by triangles and specified by Entrez Gene Symbols. Compounds are represented by circles and specified by compound names (if the length of the name exceeds 10 digits then the compound KEGG id is used). Different colors are used to specify canonical Reactome or KEGG pathways. In general, up to 11 of the most representative pathways (in terms of the number of genes in the model, both input and intermediate genes are counted) are coloured. In most cases, a gene can be associated to multiple pathways. For this reason, *R* spider implements a strict hierarchical procedure for gene coloring. First, pathways are ordered in respect to the number of genes that are present in the model from any given pathway. The most representative pathway will be colored in red. Colored genes (red) are excluded and pathways are reordered considering only the remaining genes. The next most represented pathway will be colored in blue. Colored genes (red and blue) are excluded and pathways are reordered considering only the remaining genes. The procedure will continue until 13 pathways will be colored or there will be no pathway which covers at least two genes. Therefore, colors have a strict hierarchy: red, blue, green and so on. The number before the color indicates the hierarchy order (Figure 1). It is evident that some red genes may also belong to the blue (green and so on) pathway, but not vice versa.

Table: interaction context

For each gene in the reported model, *R* spider provides the full interaction context. This information is summarized in the table 'Interacting Pairs'. In the case of Reactome, there are four types of interactions: 'direct_complex', 'indirect_complex', 'reaction' or 'neighbouring_reaction'. In the case of the KEGG database, interactions represent either a compound (connected genes are assigned to different reactions utilizing the same compound) or, rarely, by a reaction ID (both connected genes catalyze the same metabolic reaction). The edge can be supported by several different interactions, all of which will be reported, and corresponding links to the source data are provided.

Example

We present at our website (http://mips.helmholtz-muenchen.de/proj/rspider/example.html) several hundred examples of analyses by *R* spider of gene lists, which were automatically extracted by text mining from proteomics studies in various biological contexts (36). Here, we present one example in detail to demonstrate the potential benefit of our tool.

Currently, many clinical studies are designed to reveal possible pathogenic mechanisms and novel therapeutic targets for complex diseases with specific phenotypes. The Sézary syndrome, for example, is associated with the aggressive cutaneous T-cell lymphoma/leukemia. In a study by Vermeer *et al.* (37), a high-resolution array-based comparative genomic hybridization was performed on malignant T cells from 20 patients to reveal highly

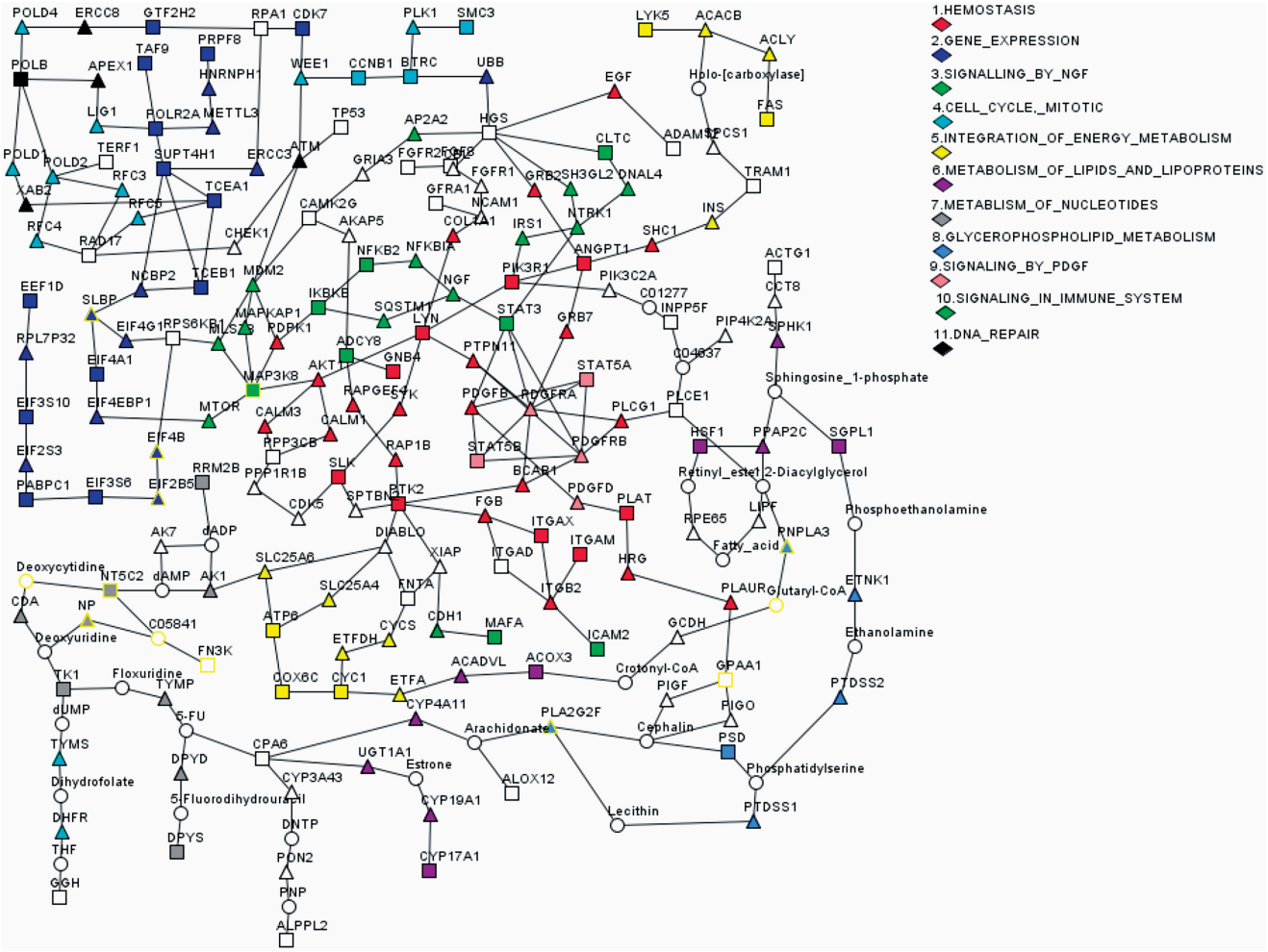

Figure 1. Network model *D3* returned by *R* spider on submission of 360 candidate genes residing in regions with copy number alteration typical of the Sézary syndrome (37). Boxes represent input genes, triangles represent intermediate genes (genes that are added to connect two input genes, for model D3 up to two intermediate genes are allowed between any two input genes), circles represent compounds which are common substrates or products for both connected genes. Diamonds are used to specify the colour of canonical Reactome or KEGG pathways.

recurrent genetic alterations typical for the Sézary syndrome. Minimal common regions with copy number alteration occurring in at least 35% of patients were reported, which comprised in total about 360 candidate genes (see Table 1 in ref. 37).

Only 22 of these genes are mapped to KEGG metabolic pathways. Thus, for comparison, an analysis by KEGG spider reports that the inferred network model is not significant ($P = \sim 0.1$). On the contrary, consideration of the integral reference network that unites both Reactome and KEGG data provides more interesting insights into the possible molecular mechanisms behind genes with copy number alteration in the Sézary syndrome. In this case, 92 out of the 360 genes are mapped to the integral network. Network model D3, which allows up to two missing genes between any two input genes, connects 74 out of the 92 mapped candidate genes into a single non-interrupted network. The model is statistically significant ($P < 0.01$). *R* spider randomly sampled 92 genes from the set of 3700 human genes that constitute the integral reference network for 1000 times; and in 993 cases, the size of the resulting network model D3 was less than 74 genes. Thus, the significance of the model is about 0.01.

R spider provides graphical models. The network model D3 for the considered example, which covers 74 genes ($P < 0.01$), is presented in Figure 1. Proteins from the input list are indicated by rectangles, intermediate proteins by triangles, and chemical compounds are indicated by circles. The colours are used to specify Reactome and KEGG canonical pathways.

In comparison to other available pathway analyses tools, *R* spider provides a global vision of gene functional relations. For example, submission to Onto-express (17) results in reporting of several (~10) enriched pathways with possibility to visualize them separately. This is certainly valuable information. However, the best model (enriched pathway 'Pathways in cancer') covers 19 genes. The relation between pathways as well as the role and relation between genes that are not covered by enriched pathways is not disclosed. Thus, in comparison to Onto-express *R* spider demonstrates that genes residing in regions which frequently have a copy number alteration

in Sézary syndrome are dependent although they belongs to a wide spectrum of signaling and metabolic pathways. In this case the user gets a newly created pathway which covers 74 genes and actually runs through several canonical Reactome and KEGG pathways.

CONCLUSIONS

Various modern genomics technologies result in gene lists. A better understanding of the biological mechanisms, which unite the identified genes, can give clues to a better understanding of the phenomena under study. *R* spider provides a possibility to actively exploit the knowledge of biological processes of various natures accumulated in the Reactome knowledgebase and metabolism related processes in the KEGG database to decipher the mechanisms behind experimentally derived gene lists. A pathway-free statistical framework combined with the most advanced publicly available databases for pathways and reactions makes *R* spider a very attractive tool for interpretation of genomics data.

ACKNOWLEDGEMENTS

We thank Philip Wong for helpful discussions.

FUNDING

Funding for open access charge: European Bioinformatics Institute, Wellcome Trust Genome Campus; The development of Reactome is supported by a grant from the US National Institutes of Health (P41 HG003751) and EU grant LSHG-CT-2005-518254 ''ENFIN''.

Conflict of interest statement. None declared.

REFERENCES

1. Draghici,S., Khatri,P., Martins,R.P., Ostermeier,G.C. and Krawetz,S.A. (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.
2. Khatri,P., Draghici,S., Ostermeier,G.C. and Krawetz,S.A. (2002) Profiling gene expression using onto-express. *Genomics*, **79**, 266–270.
3. 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.
4. Reimand,J., Kull,M., Peterson,H., Hansen,J. and Vilo,J. (2007) g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.*, **35**, W193–W200.
5. Masseroli,M., Martucci,D. and Pinciroli,F. (2004) GFINDer: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining. *Nucleic Acids Res.*, **32**, W293–W300.
6. Martin,D., Brun,C., Remy,E., Mouren,P., Thieffry,D. and Jacq,B. (2004) GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol.*, **5**, R101.
7. Khatri,P., Voichita,C., Kattan,K., Ansari,N., Khatri,A., Georgescu,C., Tarca,A.L. and Draghici,S. (2007) Onto-Tools: new additions and improvements in 2006. *Nucleic Acids Res.*, **35**, W206–W211.
8. Dietmann,S., Georgii,E., Antonov,A., Tsuda,K. and Mewes,H.W. (2009) The DICS repository: module-assisted analysis of disease-related gene lists. *Bioinformatics*, **25**, 830–831.
9. Berriz,G.F., King,O.D., Bryant,B., Sander,C. and Roth,F.P. (2003) Characterizing gene sets with FuncAssociate. *Bioinformatics*, **19**, 2502–2504.
10. Antonov,A.V., Schmidt,T., Wang,Y. and Mewes,H.W. (2008) ProfCom: a web tool for profiling the complex functionality of gene groups identified from high-throughput data. *Nucleic Acids Res.*, **36**, W347–W351.
11. Antonov,A.V., Dietmann,S., Wong,P., Lutter,D. and Mewes,H.W. (2009) GeneSet2miRNA: finding the signature of cooperative miRNA activities in the gene lists. *Nucleic Acids Res.*, **37**, W323–W328.
12. Khatri,P., Bhavsar,P., Bawa,G. and Draghici,S. (2004) Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. *Nucleic Acids Res.*, **32**, W449–W456.
13. Alexa,A., Rahnenfuhrer,J. and Lengauer,T. (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure 2. *Bioinformatics*, **22**, 1600–1607.
14. Adler,P., Reimand,J., Janes,J., Kolde,R., Peterson,H. and Vilo,J. (2008) KEGGanim: pathway animations for high-throughput data. *Bioinformatics*, **24**, 588–590.
15. Reimand,J., Tooming,L., Peterson,H., Adler,P. and Vilo,J. (2008) GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Res.*, **36**, W347–W351.
16. Berger,S.I., Posner,J.M. and Ma'ayan,A. (2007) Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases. *BMC Bioinformatics*, **8**, 372.
17. Draghici,S., Khatri,P., Tarca,A.L., Amin,K., Done,A., Voichita,C., Georgescu,C. and Romero,R. (2007) A systems biology approach for pathway level analysis. *Genome Res.*, **17**, 1537–1545.
18. Tarca,A.L., Draghici,S., Khatri,P., Hassan,S.S., Mittal,P., Kim,J.S., Kim,C.J., Kusanovic,J.P. and Romero,R. (2009) A novel signaling pathway impact analysis 1. *Bioinformatics*, **25**, 75–82.
19. Antonov,A.V., Dietmann,S. and Mewes,H.W. (2008) KEGG spider: interpretation of genomics data in the context of the global gene metabolic network. *Genome Biol.*, **9**, R179.
20. Antonov,A.V., Dietmann,S., Rodchenkov,I. and Mewes,H.W. (2009) PPI spider: a tool for the interpretation of proteomics data in the context of protein-protein interaction networks. *Proteomics*, **9**, 2740–2749.
21. Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.
22. Westfall,P.N. and Young,S.S. (1993) *Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment*. John Wiley & Sons, Inc, New York.
23. Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **27**, 29–34.
24. Matthews,L., Gopinath,G., Gillespie,M., Caudy,M., Croft,D., de,B.B., Garapati,P., Hemish,J., Hermjakob,H., Jassal,B. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, **37**, D619–D622.
25. Vastrik,I., D'Eustachio,P., Schmidt,E., Gopinath,G., Croft,D., de,B.B., Gillespie,M., Jassal,B., Lewis,S., Matthews,L. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, **8**, R39.
26. Kitano,H. and Oda,K. (2006) Robustness trade-offs and host-microbial symbiosis in the immune system4. *Mol. Syst. Biol.*, **2**, 2006.
27. Ma'ayan,A., Jenkins,S.L., Neves,S., Hasseldine,A., Grace,E., Dubin-Thaler,B., Eungdamrong,N.J., Weng,G., Ram,P.T., Rice,J.J. *et al.* (2005) Formation of regulatory patterns during signal propagation in a Mammalian cellular network16. *Science*, **309**, 1078–1083.
28. Antonov,A.V. and Mewes,H.W. (2006) BIOREL: the benchmark resource to estimate the relevance of the gene networks. *FEBS Lett.*, **580**, 844–848.

29. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al*. (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **34**, D173–D180.
30. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M. and Bairoch,A. (2007) UniProtKB/Swiss-Prot. *Methods Mol. Biol*., **406**, 89–112.
31. Hubbard,T.J., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al*. (2007) Ensembl 2007. *Nucleic Acids Res*., **35**, D610–D617.
32. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **35**, D61–D65.
33. Liu,G., Loraine,A.E., Shigeta,R., Cline,M., Cheng,J., Valmeekam,V., Sun,S., Kulp,D. and Siani-Rose,M.A. (2003) NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res*., **31**, 82–86.
34. Hooper,S.D. and Bork,P. (2005) Medusa: a simple tool for interaction graph analysis. *Bioinformatics*, **21**, 4432–4433.
35. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*., **13**, 2498–2504.
36. Antonov,A.V., Dietmann,S., Wong,P., Igor,R. and Mewes,H.W. (2009) PLIPS, an automatically collected database of protein lists reported by proteomics studies. *J. Proteome. Res*., **8**, 1193–1197.
37. Vermeer,M.H., van Doorn,R., Dijkman,R., Mao,X., Whittaker,S., van Voorst Varder,P.C., Gerritsen,M.J., Geerts,M.L., Gellrich,S., Soderberg,O. *et al*. (2008) Novel and highly recurrent chromosomal alterations in Sezary syndrome. *Cancer Res*., **68**, 2689–2698.