VisHiC—hierarchical functional enrichment analysis of microarray data

Darya Krushevskaya[1,2], Hedi Peterson[3,4], Jüri Reimand[1], Meelis Kull[1,4] and Jaak Vilo[1,2,4,*]

[1]Institute of Computer Science, University of Tartu, Liivi 2, 50409 Tartu, [2]Estonian Biocentre, [3]Institute of Molecular and Cell Biology, University of Tartu, Riia 23, 51010 Tartu, and [4]Quretec, Ülikooli 6a, 51003 Tartu, Estonia

ABSTRACT

Measuring gene expression levels with microarrays is one of the key technologies of modern genomics. Clustering of microarray data is an important application, as genes with similar expression profiles may be regulated by common pathways and involved in related functions. Gene Ontology (GO) analysis and visualization allows researchers to study the biological context of discovered clusters and characterize genes with previously unknown functions. We present VisHiC (Visualization of Hierarchical Clustering), a web server for clustering and compact visualization of gene expression data combined with automated function enrichment analysis. The main output of the analysis is a dendrogram and visual heatmap of the expression matrix that highlights biologically relevant clusters based on enriched GO terms, pathways and regulatory motifs. Clusters with most significant enrichments are contracted in the final visualization, while less relevant parts are hidden altogether. Such a dense representation of microarray data gives a quick global overview of thousands of transcripts in many conditions and provides a good starting point for further analysis. VisHiC is freely available at http://biit.cs.ut.ee/vishic.

INTRODUCTION

Microarrays have become the standard way of producing genome-scale measurements of gene expression levels (1). Since the first experimental studies (2), microarrays have been used for answering a large variety of questions, such as characterizing gene expression patterns in tumour cell lines and healthy tissues (3,4), identifying key mechanisms of stem cell differentiation (5), and reconstructing global transcriptional networks in model organisms (6).

Databases like ArrayExpress and GEO (7,8) have become goldmines of transcriptomic information with thousands of publicly available microarray datasets.

Interpretation and visualization is a crucial step of microarray analysis, as measurements are abundant and the level of experimental noise is high (9). A common reasoning behind microarray analysis is 'guilt by association', as genes with similar expression profiles may have common regulatory circuits and functions (10). Unsupervised clustering presented as a heatmap and dendrogram is a common approach for detecting coexpressed groups of genes (11,12). Gene Ontology (GO) annotations are often used for the biological interpretation of detected clusters (13).

Clustering has several well-identified drawbacks that affect interpretation and reproducibility (14). Popular clustering methods rely on input parameters, for example, hierarchical clustering (11) applies a fixed dendrogram cut-off value, and *K*-means (12) require predefining the number (and hence, the structure) of expected groups. Enrichment tools that relate gene groups to GO categories need to be accessed separately, which complicates the analysis of hundreds of clusters. Analysing results of hierarchical clustering is complicated, since each node of the dendrogram represents a potential cluster. Moreover, given the hundreds of potentially relevant datasets in public databases, the manual work would be unreasonable. Data visualization is also technically challenging, since heatmaps with thousands of transcripts hardly fit on computer screens. These problems are still commonly tackled with *ad hoc* means, e.g. removing genes that are 'not interesting' due to constant expression levels.

VisHiC (Visualization of Hierarchical Clustering) is a web server for analysis of gene expression data, that provides agile all-in-one service for hierarchical clustering, functional enrichment analysis and visualization. The tool provides a global overview of a given expression matrix and highlights its most significant functional aspects using GO analysis. VisHiC builds a compact clustering using functional enrichments rather than fixed user-defined

thresholds, by pruning clusters where no enrichments are found.

GO enrichment analysis is a common measure of gene cluster interpretation and a wide range of related tools has been created in recent years (15–17). Several microarray analysis pipelines are available, notably Expression Profiler (18), GeneXPress (19) and AMEN (20) incorporate clustering methods with downstream analysis of annotations, sequence information and protein–protein interactions. The ambiguity of clustering methods has created a need for algorithms that assess multiple clusters (21). Some previously published tools also use functional information for clustering (22–25). More recently, Ovaska *et al.* (26) combine clustering of genes based on semantic similarity of GO with heatmap visualization. However, the above comprise downloadable software that require additional data and expensive local computations. Our web server, on the other hand, provides the latest information from public databases and uses speed-optimized algorithms of HappieClust (27) and g:Profiler (16) to provide fast clustering and functional profiling even for larger datasets. In conclusion, we believe that our server provides an enhanced and useful service to the community.

THE VisHiC SERVER

VisHiC (http://biit.cs.ut.ee/vishic, Figure 1) is a web server for integrated cluster analysis, interpretation and visualization of microarray data that:

(1) performs a fast approximate hierarchical clustering of a user-provided gene expression dataset;
(2) computes functional enrichments of all discovered clusters using GO, pathways and regulatory motifs;
(3) creates a compact heatmap dendrogram of the expression dataset, revealing most important functional enrichments and hiding poorly annotated expression profiles.

The input of VisHiC is a gene expression matrix in plain tab-delimited or Gene Expression Omnibus SOFT format. Alternatively, one may use an expression matrix from our selection of example datasets. VisHiC supports a wide variety of gene, protein and probeset identifiers for human as well as most eukaryotic model organisms. The output of VisHiC is a compact gene expression matrix represented as a heatmap dendrogram, similar to the format used in many gene expression analysis applications. The analysis consists of three consecutive steps as described below.

Novel approximate algorithm allows rapid hierarchical clustering of gene expression data

The first stage of VisHiC analysis involves clustering of the input gene expression matrix.

Agglomerative hierarchical clustering (AHC) organizes the data into a dendrogram, i.e. a tree where every node represents a gene cluster (28). Nodes in the bottom of the hierarchy (i.e. leaf nodes) represent single-gene clusters, all nodes except leaves are made up of two smaller clusters, and the root node contains all genes in the dataset. The AHC algorithm starts from single-gene clusters, iteratively merges most similar neighbours and results in a hierarchical structure of $N-1$ non-trivial clusters given a dataset of N genes.

Computational speed is an important consideration of AHC, as the standard algorithm requires all pairwise distances between expression profiles. This renders to around 200 million distances in case of an average mammalian genome. The VisHiC server incorporates HappieClust, our novel approximate version of the AHC algorithm (27). Instead of computing all pairwise distances, HappieClust takes advantage of pivot-based similarity heuristics to calculate all distances between similarly expressed genes as well as a random subset of more distant pairs. Since only a subset of all pairwise distances is calculated, HappieClust approximates the full AHC based on the pairwise distances that have been calculated during the process. Computational experiments with public microarray data show that HappieClust produces a biologically comparable analysis an order of magnitude faster than standard AHC.

Pearson correlation is the default measure in VisHiC for determining similarity between expression profiles. Alternatively, one may apply the negative correlation measure that detects inverse correlation patterns such as those shared by a repressor and its targets. Absolute correlation is a combination of the two, as it detects both direct and inverse similarity.

Functional enrichment analysis reveals optimal gene clusters of biological relevance

The second stage of VisHiC analysis involves functional enrichment analysis of all detected clusters to infer the optimal clustering.

A common strategy for partitioning a hierarchical clustering involves a dendrogram cut-off. However, it is difficult to provide a biologically plausible cut-off value, as gene expression profiles are not uniformly distributed and a fixed cut-off for different datasets does not guarantee stability.

In this work, we take a different approach and infer clusters using statistical analysis of functional annotations [refer to (29) for a relevant review]. We use our g:Profiler software (16) to profile all discovered clusters for GO terms (13), pathways of Reactome and KEGG (30,31), regulatory motifs of Transfac (32) and microRNA target sites of miRBase (33).

VisHiC applies the cumulative hypergeometric test to detect the significance of a functional annotation α, given that there are k genes in a cluster of n genes with an annotation α, and there are K annotated genes among the total of N genes in the genome:

$$p_\alpha = \sum_{k}^{\min(n,K)} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

To evaluate the total enrichment in a given cluster, VisHiC computes a size-weighted annotation score q
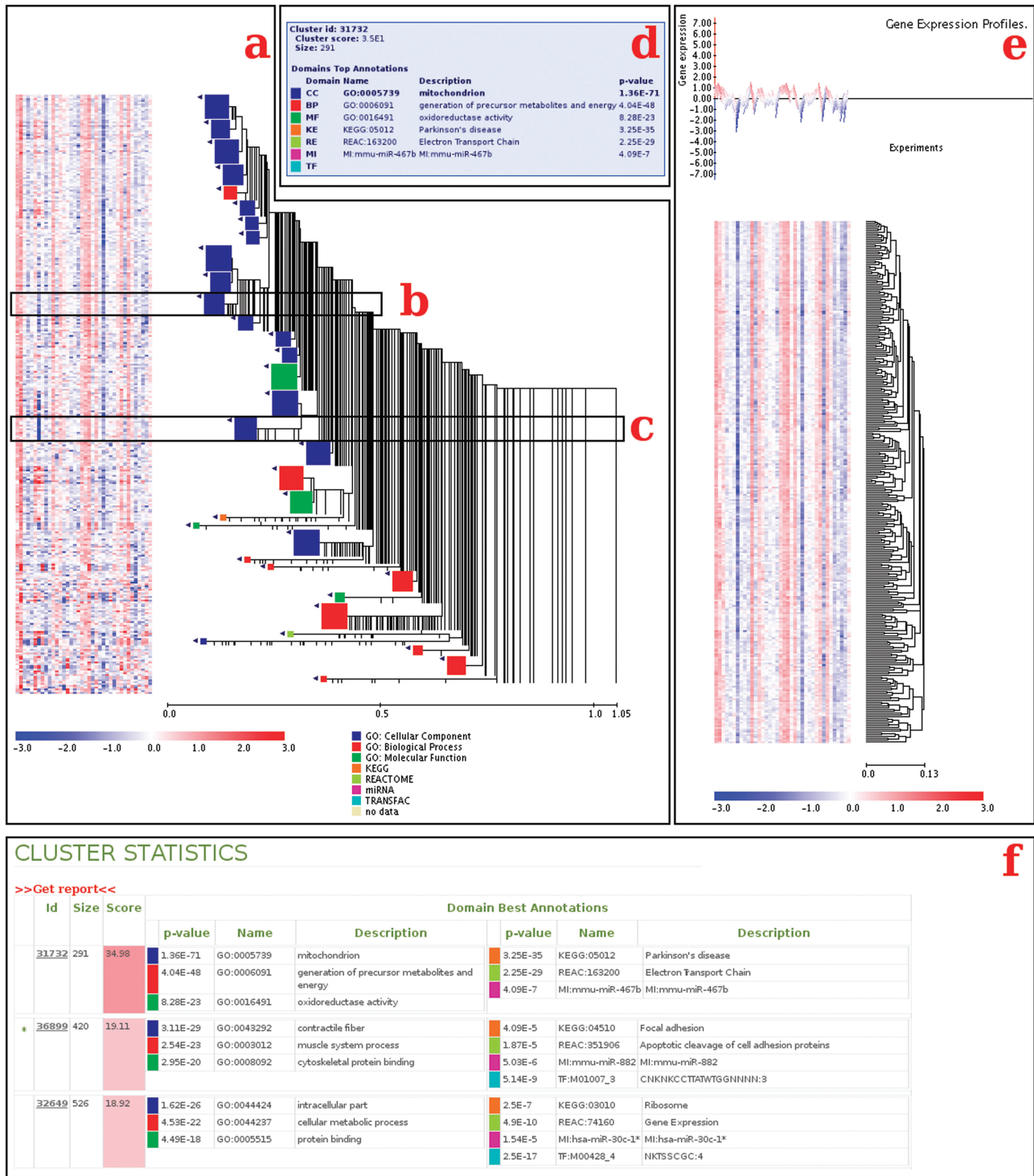
Figure 1. A biological case study with VisHiC. (**a**) Gene expression matrix and annotated dendrogram with significant clusters; (**b**) mitochondrion cluster (ID:31732), (**c**) muscle cluster (ID:36899), (**d**) annotation box of the mitochondrion cluster, appears when moving the mouse over the dendrogram, (**e**) detailed view of the mitochondrion cluster with heatmap, dendrogram and lineplot (**f**) table with functional enrichments, including clusters 31 732 and 36 899. The data presented in the figure comprises microarray measurements of the heart tissue of cardiovascular patients with left ventricular assist device. VisHiC reveals clusters with expected relevant annotations, e.g. mitochondrion, muscle tissue and ribosome (see Results section).

that summarizes enrichments of GO as well as pathways and regulatory motifs:

$$q = \frac{1}{n}\sum_{\alpha} -\log_e (p_\alpha).$$

Alternatively, one may opt for a strategy that assigns the best log P-value to each cluster, giving more preference to clusters with specific annotations:

$$m = \max_{\alpha} -\log_e (p_\alpha).$$

In order to reduce the amount of false positives resulting from numerous enrichment tests, VisHiC computes a special multiple testing correction that accounts for the hierarchical structure of GO (34). Standard corrections such as Bonferroni and Benjamini–Hochberg False Discovery Rate are also applicable.

Enrichment-driven pruning of clustering dendrogram creates a compact view of expression data

The final stage of VisHiC analysis creates a compact and biologically motivated clustering of the expression dataset to reveal its functional essence.

Hierarchical clustering places gene groups in a parent–child structure, where clusters up in the hierarchy naturally contain smaller clusters as subsets. Similarly, the GO comprises a structured vocabulary where smaller groups of specific annotations are contained in large general groups. Hence, one expects to see specific enrichments in child clusters and corresponding general annotations in parent clusters. As the clustering dendrogram contains a spectrum of hierarchically contained clusters from single genes to the whole genome, choosing an optimal cluster involves maximizing certain criteria within a branch.

We have devised the following two-stage greedy algorithm that determines the cluster structure based on functional annotations.

- First, we look for *dense clusters*, i.e. clusters with a high annotation score q, or alternatively, the term with the strongest P-value m. We scan all groups of genes that have functional enrichments, greedily starting from the one that provides the strongest annotation score. A cluster is not considered if any of its child or parent clusters is already a dense cluster. Dense clusters are shown in the final output.
- Second, we detect *sparse clusters*, i.e. groups of genes that have poor or no functional enrichments. We start the analysis from the root of the dendrogram and pass it recursively, compressing all clusters except the ones that contain dense clusters as child nodes. Sparse clusters are cut-off from the dendrogram and corresponding expression profiles are hidden in the heatmap.

Our annotation-driven clustering algorithm is fully automated and does not depend on user-defined cut-offs. Cluster boundaries are determined only from significant enrichments of functional terms. VisHiC excludes small (<5 genes) and large (>1000 genes) clusters from

enrichment analysis for optimal running time. The user may choose a different range of cluster sizes, or disable all compression to view the full expression matrix with all related enrichments. All functional terms that remain significant after multiple testing correction are used for computing the optimal clustering. However, one may apply a more stringent P-value threshold to reduce the number of contributing enrichments and compress the matrix to a greater extent.

The resulting expression matrix is presented as a heat-map of gene activation and repression patterns, complete with a dendrogram that highlights functional groups of coexpressed genes. Colour-coded rectangles in the dendrogram denote dense clusters and related functional categories (GO, KEGG, Reactome, regulatory motifs, microRNA target sites). Cluster-specific functional annotations are additionally presented in a table and also appear when hovering over the dendrogram. The main window displays the compact heatmap with all highlighted clusters, while one may also 'zoom in' to view any cluster separately. In compact view, vertical branch stumps of the dendrogram mark places where sparse clusters are compressed. The user may search for genes of interest, or conduct further analysis via hyperlinks to external resources, e.g. browse-related functional categories via the GO web site or g:Profiler.

Results: expression profiles of heart tissue of cardiovascular patients contain clusters related to muscle, mitochondria and extracellular matrix

We present a case study to demonstrate the use of VisHiC in biological analyses (Figure 1). The example comprises a microarray dataset of myocardial remodelling, including 38 samples from 3 clinical groups of patients with ischemic, non-ischemic and myocardial infarction, taken before and after left ventricular assist device implantation [available in GEO as part of the series GSE974 (35)]. We clustered the dataset, detected optimal clusters with best enrichments and visualized the resulting expression matrix (Figure 1a). We used a custom stringent P-value threshold ($P < 10^{-7}$) and 'best annotation' cluster selection strategy with Pearson correlation measure to compress the matrix into a reasonable publication-sized format.

The best scoring clusters are related to mitochondrion (Figure 1b), muscle tissue (Figure 1c) and extracellular matrix, all of which are expected to be present in heart tissue expression profiles. Mitochondria produce adenosine triphosphate (ATP) and are the primary cellular energy generators. A recent publication underlines the importance of mitochondria in the heart and relates its mutations to heart disorders (36).

The cluster with muscle tissue enrichments (ID:36899, see Figure 1e for expression profiles and Figure 1f for functional annotations) contains 420 probesets for 251 genes and has several strong enrichments (contractile fibre: $P < 10^{-28}$, muscle system process: $P < 10^{-22}$, cytoskeletal protein binding: $P < 10^{-19}$). In addition, our analysis reveals an enrichment for the binding site of serum response factor (SRF) (Transfac M01007,

$P < 10^{-9}$). SRF is a known heart transcription factor which increased expression in congestive heart failure (37).

The case study shows that VisHiC successfully extracts relevant functional aspects of a dataset, and compresses it into an easily perceivable compact format that fits well on screen and paper.

DISCUSSION AND CONCLUSION

VisHiC (http://biit.cs.ut.ee/vishic/) is a public web server for clustering and interpreting gene expression data. The tool is designed to extract the most significant biological features of a microarray dataset in a single run. The main output is a compact global view of the expression matrix with only the most significant clusters shown and less pronounced patterns hidden away, as its interactive format leaves open ends for more detailed analyses. VisHiC provides stability to otherwise ambiguous clustering and performs the labour-intensive task of evaluating hundreds of redundant clusters in a rapid automated manner. The approximate hierarchical clustering and rapid functional analysis guarantee meaningful results even if the datasets are large.

Functional assessment of microarray datasets is an immediate application of VisHiC analysis, as annotations of highlighted clusters should relate to proposed hypotheses. Our approach is likely to be useful for large expression data warehouses, so that first broad overviews could be offered to users who are routinely browsing hundreds of datasets. One may use VisHiC to compare different datasets in the context of experimental conditions, global expression patterns and functional aspects. Integrating expression clusters with other types of experimental data like protein–DNA and protein–protein interactions may provide researchers with additional clues about gene regulation.

ACKNOWLEDGEMENTS

FUNDING

Conflict of interest statement. None declared

REFERENCES

1. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.
2. Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
3. Ross,D.T., Scherf,U., Eisen,M.B., Perou,C.M., Rees,C., Spellman,P., Iyer,V., Jeffrey,S.S., Van de Rijn,M., Waltham,M. *et al.* (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, **24**, 227–235.
4. Ge,X., Yamamoto,S., Tsutsumi,S., Midorikawa,Y., Ihara,S., Wang,S.M. and Aburatani,H. (2005) Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics*, **86**, 127–141.
5. Loh,Y.H., Wu,Q., Chew,J.L., Vega,V.B., Zhang,W., Chen,X., Bourque,G., George,J., Leong,B., Liu,J. *et al.* (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.*, **38**, 431–440.
6. Hu,Z., Killion,P.J. and Iyer,V.R. (2007) Genetic reconstruction of a functional transcriptional regulatory network. *Nat. Genet.*, **39**, 683–687.
7. Parkinson,H., Kapushesky,M., Kolesnikov,N., Rustici,G., Shojatalab,M., Abeygunawardena,N., Berube,H., Dylag,M., Emam,I., Farne,A. *et al.* (2009) ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, **37**, D868–D872.
8. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M. and Edgar,R. (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.
9. Allison,D.B., Cui,X., Page,G.P. and Sabripour,M. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.*, **7**, 55–65.
10. Troyanskaya,O.G. (2005) Putting microarrays in a context: integrated analysis of diverse biological data. *Brief Bioinformatics*, **6**, 34–43.
11. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
12. Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.
13. 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.
14. Garge,N.R., Page,G.P., Sprague,A.P., Gorman,B.S. and Allison,D.B. (2005) Reproducible clusters from microarray research: whither? *BMC Bioinformatics*, **6(Suppl. 2)**, S10.
15. 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.
16. 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**, 193–200.
17. 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.
18. Kapushesky,M., Kemmeren,P., Culhane,A.C., Durinck,S., Ihmels,J., Krner,C., Kull,M., Torrente,A., Sarkans,U., Vilo,J. and Brazma,A. (2004) Expression Profiler: next generation–an online platform for analysis of microarray data. *Nucleic Acids Res.*, **32**, W465–W470.
19. Segal,E., Yelensky,R., Kaushal,A., Pham,T., Regev,A., Koller,D. and Friedman,N. (2004) GeneXPress: a visualization and statistical analysis tool for gene expression and sequence data. *Proceedings of the 11th International Conference on Intelligent Systems for Molecular Biology (ISMB)*. Glasgow, UK.
20. Chalmel,F. and Primig,M. (2008) The Annotation, Mapping, Expression and Network (AMEN) suite of tools for molecular systems biology. *BMC Bioinformatics*, **9**, 86.
21. Torrente,A., Kapushesky,M. and Brazma,A. (2005) A new algorithm for comparing and visualizing relationships between

hierarchical and flat gene expression data clusterings. *Bioinformatics*, **21**, 3993–3999.

22. Adryan,B. and Schuh,R. (2004) Gene-Ontology-based clustering of gene expression data. *Bioinformatics*, **20**, 2851–2852.

23. Okada,Y., Sahara,T., Mitsubayashi,H., Ohgiya,S. and Nagashima,T. (2005) Knowledgeassisted recognition of cluster boundaries in gene expression data. *Artif. Intell. Med.*, **35**, 171–183.

24. Seo,J., Gordish-Dressman,H. and Hoffman,E.P. (2006) An interactive power analysis tool for microarray hypothesis testing and generation. *Bioinformatics*, **22**, 808–814.

25. Dotan-Cohen,D., Melkman,A.A. and Kasif,S. (2007) Hierarchical tree snipping: clustering guided by prior knowledge. *Bioinformatics*, **23**, 3335–3342.

26. Ovaska,K., Laakso,M. and Hautaniemi,S. (2008) Fast Gene Ontology based clustering for microarray experiments. *BioData Min.*, **1**, 11.

27. Kull,M. and Vilo,J. (2008) Fast approximate hierarchical clustering using similarity heuristics. *BioData Min.*, **1**, 9.

28. Jain,A.K., Murty,M.N. and Flynn,P.J. (1999) Data clustering: a review. *ACM Comput. Surv.*, **31**, 264–323.

29. Huang,daW., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

30. Vastrik,I., D'Eustachio,P., Schmidt,E., Joshi-Tope,G., Gopinath,G., Croft,D., de Bono,B., Gillespie,M., Jassal,B., Lewis,S. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, **8**, R39.

31. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.

32. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.

33. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.

34. Reimand,J. (2006) *Gene ontology mining tool GOSt*, MSc thesis, University of Tartu.

35. Hall,J.L., Grindle,S., Han,X., Fermin,D., Park,S., Chen,Y., Bache,R.J., Mariash,A., Guan,Z., Ormaza,S. *et al.* (2004) Genomic profiling of the human heart before and after mechanical support with a ventricular assist device reveals alterations in vascular signaling networks. *Physiol. Genomics*, **17**, 283–291.

36. Fan,W., Waymire,K.G., Narula,N., Li,P., Rocher,C., Coskun,P.E., Vannan,M.A., Narula,J., Macgregor,G.R. and Wallace,D.C. (2008) A mouse model of mitochondrial disease reveals germline selection against severe mtDNA mutations. *Science*, **319**, 958–962.

37. Azhar,G., Zhang,X., Wang,S., Zhong,Y., Quick,C.M. and Wei,J.Y. (2007) Maintaining serum response factor activity in the older heart equal to that of the young adult is associated with better cardiac response to isoproterenol stress. *Basic Res. Cardiol.*, **102**, 233–244.