Published online 22 May 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W305–W311***
doi:10.1093/nar/gkp427

ToppGene Suite for gene list enrichment analysis and candidate gene prioritization

Jing Chen[1], Eric E. Bardes[2], Bruce J. Aronow[2,3] and Anil G. Jegga[2,3,*]

[1]Department of Environmental Health, University of Cincinnati, [2]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center and [3]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA

Received January 28, 2009; Revised April 24, 2009; Accepted May 11, 2009

ABSTRACT

ToppGene Suite (http://toppgene.cchmc.org; this web site is free and open to all users and does not require a login to access) is a one-stop portal for (i) gene list functional enrichment, (ii) candidate gene prioritization using either functional annotations or network analysis and (iii) identification and prioritization of novel disease candidate genes in the interactome. Functional annotation-based disease candidate gene prioritization uses a fuzzy-based similarity measure to compute the similarity between any two genes based on semantic annotations. The similarity scores from individual features are combined into an overall score using statistical meta-analysis. A *P*-value of each annotation of a test gene is derived by random sampling of the whole genome. The protein–protein interaction network (PPIN)-based disease candidate gene prioritization uses social and Web networks analysis algorithms (extended versions of the PageRank and HITS algorithms, and the K-Step Markov method). We demonstrate the utility of ToppGene Suite using 20 recently reported GWAS-based gene–disease associations (including novel disease genes) representing five diseases. ToppGene ranked 19 of 20 (95%) candidate genes within the top 20%, while ToppNet ranked 12 of 16 (75%) candidate genes among the top 20%.

INTRODUCTION

High-throughput genome-wide studies like linkage analysis and gene expression profiling, although useful for classification and characterization, do not provide sufficient information to identify specific disease causal genes. Both of these approaches typically result in hundreds of potential candidate genes, often failing to help researchers in reducing the target genes to a manageable number for further validation. To overcome these limitations, several gene prioritization methods have been developed (1–10). While all of these tools are based on the assumption that similar phenotypes are caused by genes with similar or related functions (2,11–13), they differ by the strategy they adopt in calculating similarity and by the data sources they use (14). Except for ENDEAVOUR (5,14) and ToppGene (10), most of the existing approaches mainly focus on the combination of few data sources. Interestingly, none of these approaches utilize mouse phenotype data explicitly in their prioritization approaches even though the mouse is the key model organism for the analysis of mammalian developmental, physiological and disease processes (15). Additionally, previous reports (16,17) have shown that a direct comparison of human and mouse phenotypes allowed rapid recognition of disease causal genes. In an earlier study (10), we have demonstrated that employing mouse phenotype data in fact improves candidate gene prioritization. Through various examples, we also demonstrated (10) that ToppGene performs better than SUSPECTS (9), PROSPECTR (3) and ENDEAVOUR (5), three commonly used methods in candidate gene prioritization.

Most of the current computational disease candidate gene prioritization methods (1–10) rely on functional annotations, gene-expression data or sequence-based features. The coverage of the gene functional annotations, however, is a limiting factor. Currently, only a fraction of the genome is annotated with pathways and phenotypes (10). While two-thirds of all the genes are annotated by at least one functional annotation, the remaining one-third is yet to be annotated. Recent biotechnological advances such as the high-throughput yeast two-hybrid screen have facilitated building proteome-wide protein–protein interaction networks (PPINs) or 'interactome' maps in humans (18,19). The shift in focus to systems biology in the post-genomic era has generated further interest in PPINs and biological pathways. While protein–protein interactions (PPI) have been used widely to identify novel disease candidate genes (20–24), several recent studies

*To whom correspondence should be addressed. Tel: +1 513 636 0261; Fax: +1 513 636 2056; Email: anil.jegga@cchmc.org

Table 1. Summary of ToppGene suite applications

Application	Description	Input	Output
ToppFun	Detects functional enrichment of input gene list based on Transcriptome (gene expression), Proteome (protein domains and interactions), Regulome (TFBS and miRNA), Ontologies (GO, Pathway), Phenotype (human disease and mouse phenotype), Pharmacome (Drug–Gene associations) and Bibliome (literature co-citation).	Supported identifiers include NCBI Entrez gene IDs, approved human gene symbols, NCBI Reference Sequence accession numbers; single gene list.	HTML output; Tab-delimited downloadable text file; graphical charts
ToppGene	Prioritize or rank candidate genes based on functional similarity to training gene list.	Same as above but with two gene lists (training and test)	HTML output
ToppNet	Prioritize or rank candidate genes based on topological features in protein–protein interaction network.	Same as above	HTML output; Cytoscape-compatible input file; graphical networks
ToppGeNet	Identify and prioritize the neighboring genes of the 'seeds' in protein–protein interaction network based on functional similarity to the 'seed' list (ToppGene) or topological features in protein–protein interaction network (ToppNet).	Single gene list	Same as above

(22,23,25–27) report also using them for candidate gene prioritization.

Since biological networks have been found to be comparable to communication and social networks (28) through commonalities such as scale-freeness and small-world properties, we reasoned that the algorithms used for social and Web networks should be equally applicable to biological networks and developed ToppNet (27). One of the earliest efforts (24) uses a classifier based on several topological features, including degree (number of links to the protein), 1N index (proportion of links to disease-related proteins), 2N index (average 1N index in the neighbors), average distance to disease genes and positive topology coefficient (average neighborhood overlapping with disease genes). A more recent application, Genes2Networks (29), identifies important genes based on a list of 'seed' genes. It generates a Z-score for each 'intermediate' gene from a binomial proportions test to represent its specificity or significance to the 'seed' genes. The former method, independent of known disease-related genes, is used for disease candidate gene identification, especially in cases where there is little or no prior knowledge about the disease. The latter application, on the other hand, uses a 'seed' list as training to score the neighboring genes. It avoids bias toward highly connected 'hub' genes, but the candidate gene is searched in a local network region unlike ToppNet, and the user has to provide the size of the neighborhood region in the network.

Here, we describe a unique, one-stop online assembly of computational software tools (summarized in Table 1 and Figure 1) that enables biomedical researchers to (i) perform gene list enrichment analysis (ToppFun), (ii) perform candidate gene prioritization based on functional annotations (ToppGene), (iii) perform candidate gene prioritization based on protein interactions network analysis (ToppNet) and (iv) identify and rank candidate genes in the interactome based on both functional annotations and PPIN analysis (ToppGeNet). Instructions and 'help' for each of these modules can be accessed from the homepage. The database is updated periodically, and the current status of the data (versions and coverage) can also be accessed from the homepage ('Database details'). Additionally, several examples with stepwise instructions are provided to demonstrate the utility of these applications (see 'Supplementary' section from ToppGene homepage).

TOPPFUN: GENE LIST FUNCTIONAL ENRICHMENT

ToppFun can be used for gene list functional enrichment analysis. It uses as many as 14 annotation categories including GO terms, pathways, protein–protein interactions, protein functional domains, transcription factor-binding sites, microRNAs, gene tissue expressions and literatures. Flexible options are provided to either download results as a tab-delimited file or display as a chart. Hypergeometric distribution with Bonferroni correction is used as the standard method for determining statistical significance.

TOPPGENE: FUNCTIONAL ANNOTATIONS-BASED CANDIDATE GENE PRIORITIZATION

ToppGene works by generating a representative profile of the training genes using as many as 14 features and identifies over-representative terms from the training genes. This forms the first step and is done by using ToppFun (see previous section). The test set genes are compared to this representative profile of the training set or the overrepresented terms from the training genes for all categorical annotations and the average vector for the expression values (Figure 1). For a test gene, a similarity score to the training profile for each of the 14 features

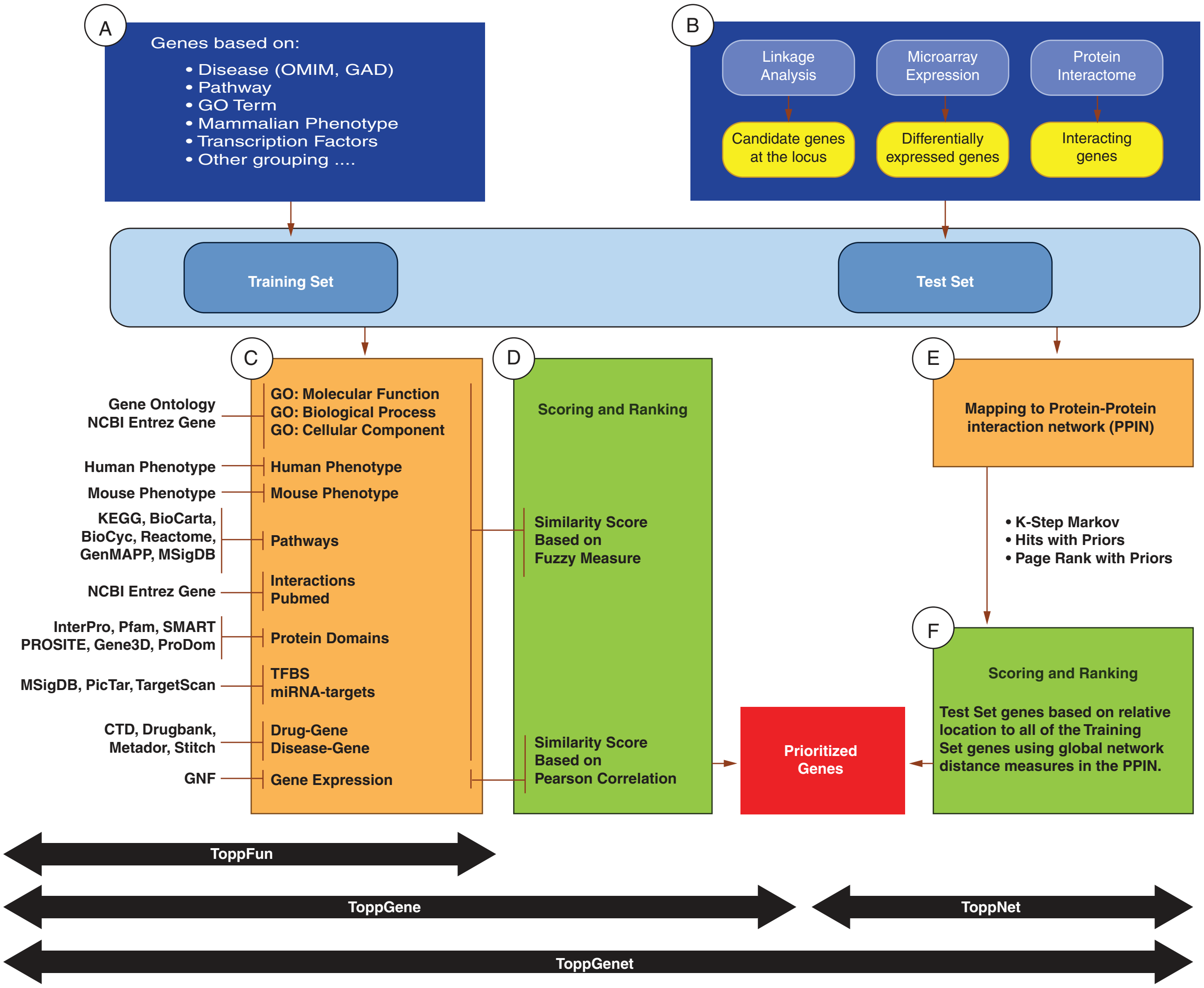

Figure 1. Schematic representation of workflow and methodology in ToppGene Suite applications. (**A**) Genes in the training set are selected based on their attributes or current gene annotations (genes associated with a disease, phenotype, pathway or a GO term). (**B**) The test gene source can be candidate genes from linkage analysis studies or genes differentially expressed in a particular disease or phenotype or genes from the interactome. (**C**) ToppFunEnriched terms of the gene annotations and sequence features, namely, GO: Molecular Function, GO: Biological Process, Mouse Phenotype, Pathways, Protein Interactions, Protein Domains, transcription factor-binding sites, miRNA-target genes, disease-gene associations, drug-gene interactions, and Gene Expression, compiled from various data sources and also used to build the training set gene profile. (C and **D**) ToppGene—a similarity score is generated for each annotation of each test gene by comparing to the enriched terms in the training set of genes. The final prioritized gene list is then computed based on the aggregated values of the 14 similarity scores. (**E** and **F**) ToppNet—Training and test set genes are mapped to a protein–protein interaction network. Scoring and ranking of test set genes are based on the relative location to all of the training set genes using global network-distance measures in the PPIN.

is derived and summarized by the 14 similarity scores. In the case of a missing value (for instance, lack of one or more annotations for a test gene), the score is set to −1. Otherwise, it is a real value in [0, 1]. Different methods are used for similarity measures of categorical (e.g. GO annotations) and numeric (i.e. gene expression) annotations. While a fuzzy-based similarity measure is applied for categorical terms [see Popescu *et al.* (30) for additional details], for numeric annotation, i.e. the microarray expression values, the similarity score is calculated as the Pearson correlation of the two expression vectors of the two genes. The 14 similarity scores are combined into an overall score using statistical meta-analysis. A *P*-value of each annotation of a test gene G is derived by random sampling of the whole genome. The *P*-value of similarity score S_i is defined as:

$$p(S_i) = \frac{\left(\begin{array}{c}\text{Count of genes having higher than } G \text{ in} \\ \text{the random sample}\end{array}\right)}{\left(\begin{array}{c}\text{Count of genes in the random sample} \\ \text{containing annotation}\end{array}\right)}$$

Fisher's inverse chi-square method, which states that $-2\sum_{i=1}^{n}\log p_i \rightarrow \chi^2(2n)$ (assuming p_i values come from independent tests) is then applied to combine the *P*-values from multiple annotations into an overall *P*-value. The final similarity score of the test gene is then obtained by 1 minus the combined *P*-value. For more details, validation and comparison with other related applications; the readers are referred to our previous study (10).

TOPPNET: NETWORK ANALYSIS-BASED CANDIDATE GENE PRIORITIZATION

ToppNet gene prioritization is based on protein–protein interaction network (PPIN) analyses. Based on the observation that biological networks share many properties with Web and social networks (28), ToppNet uses extended versions of three algorithms from White and Smyth (31)—PageRank with Priors, HITS with Priors and K-step Markov—to prioritize disease candidate genes by estimating their relative importance in the PPIN to the disease-related genes. For more details about the protein interaction datasets used, algorithmic details and validation, see our recently published study (27).

TOPPGENET: PRIORITIZATION OF NEIGHBORING GENES IN PPIN

ToppGeNet differs from ToppGene and ToppNet in that the test set is derived from the protein interactome. In other words, for a training set of known disease genes, the test set is generated by mining the protein interactome and compiling the genes either directly or indirectly interacting (based on user input) with the training set. After any overlapping or common genes between test and training sets are removed, interactome-based test set genes can be prioritized using either a functional annotation-based method (ToppGene) or PPIN-based method (ToppNet). The human protein interaction dataset (file 'interactions.gz'), a compilation of PPIs from BIND (32), BioGRID (33) and HPRD (34), is downloaded from NCBI Entrez Gene FTP site (ftp://ftp.ncbi.nih.gov/gene/).

TOPPGENE SUITE IMPLEMENTATION AND ACCESS

The programs of our enrichment and prioritization methods are implemented in JAVA. An open-source JAVA package, FtpBean by Calvin Tai (http://www.geocities.com/SiliconValley/Code/9129), is used to automatically download data and annotation files from FTP servers. BioJava packages are used to process UniProt records and extract related protein domain information. GOLEM (http://function.princeton.edu/GOLEM/download.html) source code is adapted and modified for dealing with ontology annotations. Colt (http://dsd.lbl.gov/~hoschek/colt) and Jakarta Commons-Math libraries (http://jakarta.apache.org/commons/math) are used for statistical analysis. The fuzzy similarity measure and related functions are implemented locally. The user front end of ToppGene Suite is a web application written in JAVA. The application server is Sun GlassFish Enterprise Server v2.1 running on OpenSUSE 10.3 Linux. Speed is a key consideration in the design choices of the ToppGene Suite front end. When the web server is started, most of the data is loaded from a relational database and kept in memory.

ToppGene Suite uses two different relational databases for persistence of data: (i) Oracle Database 10*g* Enterprise Edition Release 10.2.0.3.0 – 64 bit; and (ii) Apache Derby Server - 10.4.2.0. The two databases are used differently. The 'production data' are stored in a Derby database on the same computer as the web server, which gives Derby the advantage that it does not have to fetch large data sets across a network and therefore eliminates network latency for small queries. The Oracle Database, on the other hand, is used for data collection and refresh. The data schemas in Oracle are highly structured according to the generally accepted database practice of Third Normal Form.

ToppGene Suite uses Hibernate (http://www.hibernate.org/) for updating and retrieving data to and from the databases. The back end of ToppGene Suite is a scripted process that automatically downloads data from publicly available data sources [see (10,27) for more details]. The process, also written in JAVA, is launched using a common JAVA utility called Ant (provided by the Apache Foundation).

The gene information, annotation and the interactions data is updated automatically except for pathways (see the 'Database details' section from the homepage of ToppGene Suite for a list of data resources, coverage and version details and dates of last updates). The 'Database details' is a dynamic web page that reads the in-memory data structures and displays the counts and statistics of the live data. As the data are refreshed, the counts and statistics are automatically updated. Users can enter the training and test sets of genes of interest as queries from the interface, and the application will display enriched themes in the training set genes along with annotated prioritized test genes. Alternately, users can enter training sets and use the extended gene list from the PPIN as a test set to rank the genes in the interactome using either functional annotations or network features.

UTILITY OF THE TOPPGENE SUITE

For a more detailed validation study using ToppGene, the readers are referred to our previous study (10). In the present study, to demonstrate the utility of ToppGene Suite, we focused on recently reported GWAS. The aim was to test whether ToppGene and ToppNet are capable of retrieving or prioritizing the GWAS-discovered novel disease genes in a training-test type of analysis. We used 20 gene–disease associations (including novel disease genes) representing five diseases (Bipolar Disorder, Cardiomyopathy, Celiac Disease, Crohns Disease and Obesity; Table 2). For each of these five disorders, we built a training set containing all the genes already known to play a role in that disorder according to the OMIM and NCBI's Entrez Gene records (limiting the search field to 'Disease/Phenotype' and organism 'Homo sapiens') (See 'Supplementary' section from ToppGene homepage). The test set consisted of the GWAS-reported disease gene plus 99 nearest neighboring genes based on

Table 2. Results of the 20 genetic disease prioritizations using ToppGene and ToppNet

Disease	Reference	Gene	ToppGene rank	ToppNet rank
Bipolar disorder	Le-Niculescu *et al.* (35)	*KLF12*	2	15
Bipolar disorder	Le-Niculescu *et al.* (35)	*RORB*	4	18
Bipolar disorder	Le-Niculescu *et al.* (35)	*RORA*	7	13
Bipolar disorder	Le-Niculescu *et al.* (35)	*ALDH1A1*	10	No interaction data
Bipolar disorder	Le-Niculescu *et al.* (35)	*AK3L1*	11	No interaction data
Cardiomyopathy	Dhandapany *et al.* (36)	*MYBPC3*	1	2
Celiac disease	Hunt *et al.* (37)	*SH2B3*	1	8
Celiac disease	Hunt *et al.* (37)	*CCR3*	2	3
Celiac disease	Hunt *et al.* (37)	*IL18R1*	3	29
Celiac disease	Hunt *et al.* (37)	*RGS1*	9	26
Celiac disease	Hunt *et al.* (37)	*TAGAP*	14	No interaction data
Celiac disease	Hunt et al. (37)	*IL12A*	14	10
Crohns disease	Fisher *et al.* (38)	*MST1*	1	27
Crohns disease	Fisher *et al.* (38)	*NKX2-3*	1	27
Crohns disease	Fisher *et al.* (38)	*IRGM*	2	No interaction data
Crohns disease	Villani *et al.* (39)	*NLRP3*	5	1
Crohns disease	Fisher *et al.* (38)	*IL12B*	7	1
Crohns disease	Barrett *et al.* (40) Franke *et al.* (41)	*STAT3*	11	1
Crohns disease	Franke *et al.* (41)	*PTPN2*	30	6
Obesity	Renstrom *et al.* (42)	*MC4R*	1	1
		Mean	6.8	11.75

The gene-disease associations were from recently reported GWAS and include novel disease gene associations. The training sets were compiled using 'phenotype/disease' annotations in NCBI's Entrez Gene records and OMIM. To build the test set genes, we defined the artificial linkage interval to be the set of genes containing the 99 nearest neighboring genes to the novel disease gene based on their genomic distance on the same chromosome.

their location on the same chromosome. ToppGene and ToppNet prioritization results are presented in Table 2. ToppGene ranked 19 of 20 (95%) candidate genes within the top 20%, while ToppNet ranked 12 of 16 (75%) candidate genes among the top 20%. The mean ranks for ToppGene- and ToppNet-based prioritization were 6.8 and 11.75, respectively (excluding four disease genes that lacked interaction data).

LIMITATIONS

ToppGene or any functional annotation-based prioritization method has some limitations. First, when using a training set of genes, the assumption is that the disease genes we have yet to discover will be consistent with what is already known about a disease and/or its genetic basis, which may not always be the case. Second, the annotations and analyses, as well as the prioritization, can only be as accurate as the underlying online sources from which the annotations are retrieved. Similar to functional annotation-based methods, the performance of network-based prioritization methods (ToppNet) is also dependent on the quality of interaction data, which currently suffers from incompleteness and unreliability with missing interactions and false positives.

CONCLUSIONS

Existing disease candidate gene prioritization methodologies mine biological and functional information about candidate genes, and we believe that our ToppGene Suite can complement these existing approaches by applying novel methods that mine mouse phenotype data and PPIN. Through various examples, we demonstrate that ToppGene Suite is capable of identifying true candidate genes. However, it needs to be emphasized that our aim is not to prove that ToppGene Suite-prioritized genes are true disease genes but rather to aid in selection of a subset of most likely disease gene candidates from larger sets of disease-implicated genes identified by high-throughput genome-wide techniques like linkage analysis and microarray analysis. As the functional annotations of human and mouse genes and the quality of PPIN improves, we envisage a proportional increase in the performance of ToppGene Suite and strongly believe that it will be a valuable adjunct to wet lab experiments in human genetics and disease research. We further hypothesize that integrating the rankings obtained using functional annotations and PPIN-based approaches may improve the prioritization of disease genes.

ACKNOWLEDGEMENTS

We acknowledge the help of Ron Bryson, Technical Writer, Division of Biomedical Informatics, CCHMC, Ohio, USA, in editing the manuscript.

FUNDING

State of Ohio Computational Medicine Center (ODD TECH 04-042); National Institutes of Health/National Institute of Diabetes and Digestive and Kidney Diseases (NIH/NIDDK) 1U01 DK70219 (Murine Atlas of a Genitourinary Smooth Muscle Development); PHS Grant P30 DK078392 (Cincinnati Digestive Health

Center). Funding for open access charge: CCHMC, Cincinnati, OH, USA.

Conflict of interest statement. None declared.

REFERENCES

1. Freudenberg,J. and Propping,P. (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, **18(Suppl. 2)**, S110–S115.
2. Turner,F.S., Clutterbuck,D.R. and Semple,C.A. (2003) POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol*., **4**, R75.
3. Tiffin,N., Kelso,J.F., Powell,A.R., Pan,H., Bajic,V.B. and Hide,W.A. (2005) Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res*., **33**, 1544–1552.
4. Adie,E.A., Adams,R.R., Evans,K.L., Porteous,D.J. and Pickard,B.S. (2005) Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, **6**, 55.
5. Aerts,S., Lambrechts,D., Maity,S., Van Loo,P., Coessens,B., De Smet,F., Tranchevent,L.C., De Moor,B., Marynen,P., Hassan,B. *et al*. (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol*., **24**, 537–544.
6. Thornblad,T.A., Elliott,K.S., Jowett,J. and Visscher,P.M. (2007) Prioritization of positional candidate genes using multiple web-based software tools. *Twin Res. Hum. Genet*., **10**, 861–870.
7. Zhu,M. and Zhao,S. (2007) Candidate gene identification approach: progress and challenges. *Int. J. Biol. Sci*., **3**, 420–427.
8. Tiffin,N., Adie,E., Turner,F., Brunner,H.G., van Driel,M.A., Oti,M., Lopez-Bigas,N., Ouzounis,C., Perez-Iratxeta,C., Andrade-Navarro,M.A. *et al*. (2006) Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Res*., **34**, 3067–3081.
9. Adie,E.A., Adams,R.R., Evans,K.L., Porteous,D.J. and Pickard,B.S. (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, **22**, 773–774.
10. Chen,J., Xu,H., Aronow,B.J. and Jegga,A.G. (2007) Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics*, **8**, 392.
11. Goh,K.I., Cusick,M.E., Valle,D., Childs,B., Vidal,M. and Barabasi,A.L. (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.
12. Jimenez-Sanchez,G., Childs,B. and Valle,D. (2001) Human disease genes. *Nature*, **409**, 853–855.
13. Smith,N.G. and Eyre-Walker,A. (2003) Human disease genes: patterns and predictions. *Gene*, **318**, 169–175.
14. Tranchevent,L.C., Barriot,R., Yu,S., Van Vooren,S., Van Loo,P., Coessens,B., De Moor,B., Aerts,S. and Moreau,Y. (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res*, **36**, W377–W384.
15. Clarke,A.R. (1994) Murine genetic models of human disease. *Curr. Opin. Genet, Dev*., **4**, 453–460.
16. Gorgels,T.G., Hu,X., Scheffer,G.L., van der Wal,A.C., Toonstra,J., de Jong,P.T., van Kuppevelt,T.H., Levelt,C.N., de Wolf,A., Loves,W.J. *et al*. (2005) Disruption of Abcc6 in the mouse: novel insight in the pathogenesis of pseudoxanthoma elasticum. *Hum. Mol. Genet*., **14**, 1763–1773.
17. van Bokhoven,H., Celli,J., Kayserili,H., van Beusekom,E., Balci,S., Brussel,W., Skovby,F., Kerr,B., Percin,E.F., Akarsu,N. *et al*. (2000) Mutation of the gene encoding the ROR2 tyrosine kinase causes autosomal recessive Robinow syndrome. *Nat. Genet*., **25**, 423–426.
18. Rual,J.F., Venkatesan,K., Hao,T., Hirozane-Kishikawa,T., Dricot,A., Li,N., Berriz,G.F., Gibbons,F.D., Dreze,M., Ayivi-Guedehoussou,N. *et al*. (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178.
19. Stelzl,U., Worm,U., Lalowski,M., Haenig,C., Brembeck,F.H., Goehler,H., Stroedicke,M., Zenkner,M., Schoenherr,A., Koeppen,S. *et al*. (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.
20. George,R.A., Liu,J.Y., Feng,L.L., Bryson-Richardson,R.J., Fatkin,D. and Wouters,M.A. (2006) Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res*., **34**, e130.
21. Kann,M.G. (2007) Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief Bioinform*., **8**, 333–346.
22. Kohler,S., Bauer,S., Horn,D. and Robinson,P.N. (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet*., **82**, 949–958.
23. Wu,X., Jiang,R., Zhang,M.Q. and Li,S. (2008) Network-based global inference of human disease genes. *Mol. Syst. Biol*., **4**, 189.
24. Xu,J. and Li,Y. (2006) Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics*, **22**, 2800–2805.
25. Chen,J.Y., Shen,C. and Sivachenko,A.Y. (2006) Mining Alzheimer disease relevant proteins from integrated protein interactome data. *Pac. Symp. Biocomput.,* 367–378.
26. Ortutay,C. and Vihinen,M. (2009) Identification of candidate disease genes by integrating Gene Ontologies and protein-interaction networks: case study of primary immunodeficiencies. *Nucleic Acids Res*., **37**, 622–628.
27. Chen,J., Aronow,B.J. and Jegga,A.G. (2009) Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics*, **10**, 73.
28. Junker,B.H., Koschutzki,D. and Schreiber,F. (2006) Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics*, **7**, 219.
29. 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.
30. Popescu,M., Keller,J.M. and Mitchell,J.A. (2006) Fuzzy measures on the gene ontology for gene product similarity. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **3**, 263–274.
31. White,S. and Smyth,P. (2003) Algorithms for estimating relative importance in networks. In *KDD '03: Proc 9th ACM SIGKDD Int. Conf. Knowledge Discov. Data Mining*, ACM, New York, pp. 266–275.
32. Bader,G.D., Betel,D. and Hogue,C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*., **31**, 248–250.
33. Breitkreutz,B.J., Stark,C., Reguly,T., Boucher,L., Breitkreutz,A., Livstone,M., Oughtred,R., Lackner,D.H., Bahler,J., Wood,V. *et al*. (2008) The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res*., **36**, D637–D640.
34. Peri,S., Navarro,J.D., Kristiansen,T.Z., Amanchy,R., Surendranath,V., Muthusamy,B., Gandhi,T.K., Chandrika,K.N., Deshpande,N., Suresh,S. *et al*. (2004) Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*., **32**, D497–D501.
35. Le-Niculescu,H., Patel,S.D., Bhat,M., Kuczenski,R., Faraone,S.V., Tsuang,M.T., McMahon,F.J., Schork,N.J., Nurnberger,J.I. Jr. and Niculescu,A.B. 3rd. (2009) Convergent functional genomics of genome-wide association data for bipolar disorder: comprehensive identification of candidate genes, pathways and mechanisms. *Am. J. Med. Genet. B Neuropsychiatr. Genet*., **150B**, 155–181.
36. Dhandapany,P.S., Sadayappan,S., Xue,Y., Powell,G.T., Rani,D.S., Nallari,P., Rai,T.S., Khullar,M., Soares,P., Bahl,A. *et al*. (2009) A common MYBPC3 (cardiac myosin binding protein C) variant associated with cardiomyopathies in South Asia. *Nat. Genet*., **41**, 187–191.
37. Hunt,K.A., Zhernakova,A., Turner,G., Heap,G.A., Franke,L., Bruinenberg,M., Romanos,J., Dinesen,L.C., Ryan,A.W., Panesar,D. *et al*. (2008) Newly identified genetic risk variants for celiac disease related to the immune response. *Nat. Genet*., **40**, 395–402.
38. Fisher,S.A., Tremelling,M., Anderson,C.A., Gwilliam,R., Bumpstead,S., Prescott,N.J., Nimmo,E.R., Massey,D., Berzuini,C., Johnson,C. *et al*. (2008) Genetic determinants of ulcerative colitis include the ECM1 locus and five loci implicated in Crohn's disease. *Nat. Genet*., **40**, 710–712.
39. Villani,A.C., Lemire,M., Fortin,G., Louis,E., Silverberg,M.S., Collette,C., Baba,N., Libioulle,C., Belaiche,J., Bitton,A. *et al.*

(2009) Common variants in the NLRP3 region contribute to Crohn's disease susceptibility. *Nat. Genet.*, **41**, 71–76.

40. Barrett,J.C., Hansoul,S., Nicolae,D.L., Cho,J.H., Duerr,R.H., Rioux,J.D., Brant,S.R., Silverberg,M.S., Taylor,K.D., Barmada,M.M. *et al.* (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.*, **40**, 955–962.
41. Franke,A., Balschun,T., Karlsen,T.H., Hedderich,J., May,S., Lu,T., Schuldt,D., Nikolaus,S., Rosenstiel,P., Krawczak,M. *et al.* (2008) Replication of signals from recent studies of Crohn's disease identifies previously unknown disease loci for ulcerative colitis. *Nat. Genet.*, **40**, 713–715.
42. Renstrom,F., Payne,F., Nordstrom,A., Brito,E.C., Rolandsson,O., Hallmans,G., Barroso,I., Nordstrom,P. and Franks,P.W. (2009) Replication and extension of genome-wide association study results for obesity in 4923 adults from Northern Sweden. *Hum. Mol. Genet.*, **18**, 1489–1496.