

# T1DBase: integration and presentation of complex data for type 1 diabetes research

Erin M. Hulbert\*, Luc J. Smink<sup>1,\*</sup>, Ellen C. Adlem<sup>1</sup>, James E. Allen<sup>1</sup>, David B. Burdick, Oliver S. Burren<sup>1</sup>, Christopher C. Cavnor, Geoffrey E. Dolman<sup>1</sup>, Daisy Flamez<sup>2</sup>, Karen F. Friery, Barry C. Healy<sup>1</sup>, Sarah A. Killcoyne, Burak Kutlu, Helen Schuilenburg<sup>1</sup>, Neil M. Walker<sup>1</sup>, Josyf Mychaleckyj<sup>3</sup>, Decio L. Eizirik<sup>2</sup>, Linda S. Wicker<sup>1</sup>, John A. Todd<sup>1</sup> and Nathan Goodman

Institute for Systems Biology, Seattle, WA, USA, <sup>1</sup>Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, University of Cambridge, Cambridge, UK, <sup>2</sup>Laboratory of Experimental Medicine, Free University of Brussels (ULB), Brussels, Belgium and <sup>3</sup>Center for Human Genomics, Wake Forest University School of Medicine, Winston-Salem, NC, USA

Received September 12, 2006; Revised October 11, 2006; Accepted October 12, 2006

## ABSTRACT

T1DBase (<http://T1DBase.org>) [Smink *et al.* (2005) *Nucleic Acids Res.*, 33, D544–D549; Burren *et al.* (2004) *Hum. Genomics*, 1, 98–109] is a public website and database that supports the type 1 diabetes (T1D) research community. T1DBase provides a consolidated T1D-oriented view of the complex data world that now confronts medical researchers and enables scientists to navigate from information they know to information that is new to them. Overview pages for genes and markers summarize information for these elements. The Gene Dossier summarizes information for a list of genes. GBrowse [Stein *et al.* (2002) *Genome Res.*, 10, 1599–1610] displays genes and other features in their genomic context, and Cytoscape [Shannon *et al.* (2003) *Genome Res.*, 13, 2498–2504] shows genes in the context of interacting proteins and genes. The Beta Cell Gene Atlas shows gene expression in  $\beta$  cells, islets, and related cell types and lines, and the Tissue Expression Viewer shows expression across other tissues. The Microarray Viewer shows expression from more than 20 array experiments. The Beta Cell Gene Expression Bank contains manually curated gene and pathway annotations for genes expressed in  $\beta$  cells. T1DMart is a query tool for markers and genotypes. PosterPages are 'home pages' about specific topics or datasets. The key challenge, now and in the future, is to

provide powerful informatics capabilities to T1D scientists in a form they can use to enhance their research.

## INTRODUCTION

T1DBase (<http://T1DBase.org>) (1,2) is a public website and database that supports researchers working on the molecular genetics and biology of type 1 diabetes (T1D) susceptibility and pathogenesis. The system collates and organizes data relevant to T1D research from public and private sources, and integrates this information in a research-accessible manner, presenting the results in a form that is useful and enabling for T1D researchers.

Medical science has become data-rich such that researchers working on T1D and other diseases may be overwhelmed by the volume and diversity of available data. It is unreasonable to expect scientists to understand the nuances of all types of biological data now available or to even keep track of the burgeoning number of available data sources and judge their strengths and weaknesses.

T1DBase provides a simplified and integrated T1D-oriented view of this complex data world. We collect and verify data from a large number of data sources, extract relevant information from the collected datasets, integrate this information in a researcher-accessible manner, and present the results in a form that T1D scientists can use without having to know all these details. In many cases, we also supply the full raw data for users who want to do their own analyses.

Datasets in T1DBase include assembled genome sequences for human, mouse and rat; empirically and computationally

\*To whom correspondence should be addressed. Tel: +1 206 331 0077; Fax: +1 206 363 0431; Email: [EHulbert@systemsbiology.org](mailto:EHulbert@systemsbiology.org)

\*Correspondence may also be addressed to Luc J. Smink. Tel: +1 206 732 1200; Fax: +1 206 732 1299; Email: [Luc.Smink@cimr.cam.ac.uk](mailto:Luc.Smink@cimr.cam.ac.uk)

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© 2006 The Author(s).

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.0/uk/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

derived gene and transcript models from high throughput and biological focused efforts; annotations of gene function; gene orthologies; publications linked to genes; T1D candidate genes; genetically identified T1D susceptibility regions; genetic linkage and association studies pertaining to T1D; SNPs and genotypes; gene expression data from arrays, massively parallel signature sequencing (MPSS), and other technologies; pathways; protein–protein and protein–DNA interactions; and information on NOD mouse congenic strains.

These datasets are integrated so that scientists can easily navigate from information they know—say, a T1D candidate gene to information that is new to them—for example, array data showing how that gene’s expression changes during T1D progression. Tools are provided to let the scientist go further, e.g. to find other genes whose expression profile is similar to the original candidate. We are also functionally integrating T1DBase with EPCoNDB (3), the flagship database of the Beta Cell Biology Consortium. Since research in T1D and other diseases is often carried out in mouse and rat models in addition to human subjects and materials, we integrate data for the human, mouse and rat orthologs of every gene so that scientists can move seamlessly among data from different experiments and sources.

All software is open access and can be downloaded from SourceForge under the GNU General Public License or Perl Artistic License. Most datasets are open access. The exceptions are a few datasets whose redistribution is prohibited by the original data source, and private data uploaded by users or provided by collaborators, for which visibility and redistribution are controlled by the user; an example of the latter would be genotypes from unpublished studies.

The core software underlying T1DBase is a generic framework for disease oriented websites, called GDxBase, which can be used for other diseases. The look and feel of the website are controlled through style sheets and templates, which are easily configurable. GDxBase is currently used in projects for Huntington’s Disease (HDBase—ISB) (4), prion disease (PDDb—ISB), type 2 diabetes (T2DBase—Mark McCarthy’s group in Oxford), bloodomics [BxBase—Willem Ouwehand at the Wellcome Trust Sanger Institute (WTSI) and University of Cambridge] and diseases of energy metabolism (GEMDBase—Ines Barroso at WTSI). Of these, only HDBase is publicly visible at present.

## DATA PRESENTATION

A central theme of the system is to provide simple, concise ways to present complex, consolidated data to researchers who are not specific experts in the particular type of data involved, while at the same time providing additional detail to users with greater expertise.

The *Gene Overview* page summarizes all data in T1DBase for the human, mouse and rat orthologs of any given gene. The results are presented in sections, most of which can be expanded and contracted for a general or more detailed view. It provides an overview of gene function, synonyms, orthologs, whether the gene is a T1D candidate gene or located within a genetically identified T1D susceptibility region, functional annotation from the Beta Cell Gene Expression Bank, gene and protein models, descriptions of

pathways containing the gene, gene ontology (GO) terms (5), other diseases for which the gene is a candidate, gene expression from the Beta Cell Gene Atlas and expression in a range of other tissues, T1D-related and other publications related to the gene, and links to external databases. The gene models section graphically displays transcript data from the University of California at Santa Cruz (UCSC) Genome Bioinformatics Site (6), Ensembl (7), Vega (8), and EPCoNDB (3), alongside manually curated transcripts from the Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory (JDRF/WT DIL). A summary of the transcripts from all sources is calculated and visualized as a histogram, the consensus model. There are links to GBrowse (9) for visualizing the gene in its genomic context and Cytoscape (10) for visualizing the gene in the context of interacting proteins and genes.

Sequence-based features from different sources (e.g. transcript models) are mapped to a common genome assembly so they can be accurately integrated. We endeavor to move quickly to new assemblies as they become available. For some data, we also provide mappings to old assemblies.

Gene orthologs are determined by combining data from HomoloGene (11), Mouse Genome Database (12), Rat Genome Database (13), Ensembl (7), Inparanoid (14), OrthoMCL (15) and KEGG (16). We consider each data source in order, and add to the database all consistent orthology information from that source. No attempt is made to weight information based on the number of sources that report the information because many of the sources are not independent.

The *Marker Overview* page integrates SNP data from different data sources. The main sources currently are dbSNP (11) and HapMap (17). As for the Gene Overview page, the results are presented in sections. There are five different sections: SNP summary, population, genotype information, marker neighborhood and literature. The summary section contains mapping information from different data sources and links to the relevant GBrowse position, chromosome and genome build. The population information details the number of populations for which there is genotype data and provides an expandable table with detailed information about each population, the source of the genotype data, and allele frequencies. The genotypes from HapMap and dbSNP are viewable as a table in the genotype section which contains details about the individual, collection, allele, and the actual genotyped alleles. The marker neighborhood uses GBrowse to show the genomic environment by displaying the marker and other features around it. The literature section is presently limited to publications by JDRF/WT DIL authors about JDRF/WT DIL markers.

All human SNP sequence information is extracted from the latest dbSNP build and mapped to the most recent three human genome assemblies. We also extract information on subjects and the actual genotypes from dbSNP and HapMap. The system also has SNP, subject and genotyping data from a recently published high resolution SNP map of the human MHC and adjacent region in a Swedish population (18) and will distribute results from a larger and broader population being studied by the Type 1 Diabetes Genetics Consortium (T1DGC), the latter in collaboration with the T1DGC Coordinating Center at Wake Forest University School of Medicine. Please note that access to the T1DGC dataset is

governed by consortium policy and international privacy regulations and some of these results will not be visible to the general public.

Genotype datasets are stored in Genetics Laboratory Interactive Data and Experimental Repository (GLIDER) databases, a data exchange and database structure for genotype data. Each GLIDER holds the genotype data for individual experiments and/or populations, but the associated marker and subject data for all GLIDERS is stored in one centralized database. This lets authorized users view integrated information for each marker from different data sources, and also determine if a particular subject has been genotyped by different groups. The results from these databases are made available to the user through the Marker Overview page and can be queried using TIDMart described below.

The *Gene Dossier* provides a concise way of seeing data from many datasets and tools for a list of genes. The Gene Dossier presents a tabular view: each row is a gene, and each column summarizes data from one dataset or tool for that gene. There is also a summary column that reports an overall score for T1D-relevance. In essence the Dossier is, a very condensed form of the Gene Overview page that gives scientists a rapid overview of a list of genes. Users can customize the columns that are displayed and the weights used to generate the summary score. The table can be sorted by any column.

*GBrowse* is an open-source genome browser developed as part of the Generic Model Organism Database (GMOD) project (9). It allows zooming and scrolling of sequence features in a region of the genome. TIDBase has preloaded UCSC and Ensembl tracks, JDRF/WT DIL transcript annotations, EPConDB transcripts, SNPs from dbSNP, HapMap and JDRF/WT DIL, and data from a major T1D linkage study (the so-called *quad scan*) by the T1DGC (19). We have also created consensus gene models by combining transcript sources. User features can be uploaded alongside known genes for comparison, which allows for sharing of data. A link for more information about each feature can be found by clicking on the image for that feature. A new plugin has been added to GBrowse which allows users to overlay their own  $D'$  and  $R^2$  data. Linkage disequilibrium input is in the form of a Haploview LD file (20). The plugin itself is a modified version of code developed by Lalitha Krishnan and Lincoln D. Stein, as part of the International HapMap Consortium (21). A further plugin allows for the graphing of *ad hoc* datasets and is particularly relevant to disease risk estimates such as odds ratio.

We have also added the different haplotypes that have been sequenced and annotated for the MHC region (22). The sequences are available for the reference cell lines PGF, COX and QBL. We have calculated some of the deletion/insertion polymorphisms (DIPs) and SNPs for each of the haplotypes compared to the others based on data extracted from the WTSI website. These datasets can be viewed as separate data tracks in GBrowse. In addition, the genes in the MHC have haplotypic information which can be viewed through the Gene Overview page in tabular format.

The *Beta Cell Gene Atlas* shows basal expression of genes in human and rodent  $\beta$  cells, islets, and related cell types and cell lines. Data from approximately twenty microarray experiments and several MPSS runs were combined to create

lists of genes that are enriched in or specific to these cell types. Genes are described as either not expressed or expressed at a high, moderate or low level.

The *Tissue Expression Viewer* shows gene expression for a list of genes across a range of tissues. This viewer combines data from SymAtlas (23), SAGE, and expressed sequence tags using our implementation of a method proposed in (24) and TissueInfo's tissue hierarchy (25).

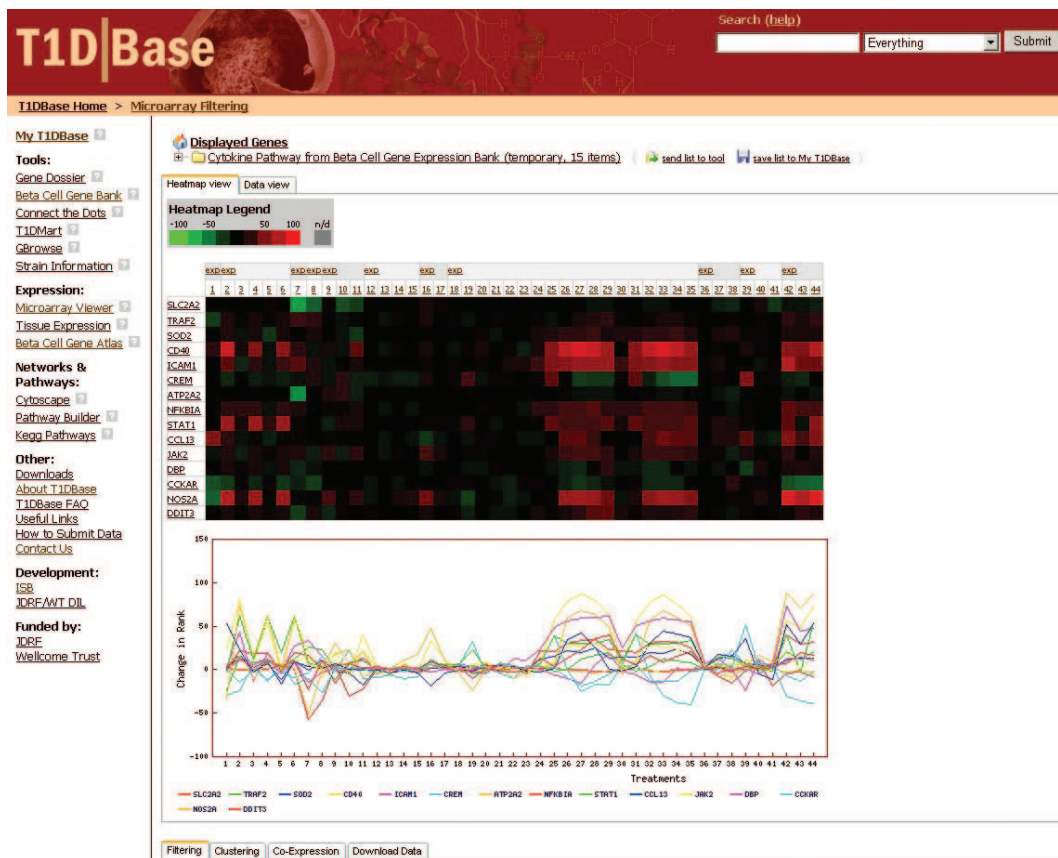
The *Microarray Viewer* shows differential gene expression across experimental conditions from over twenty array experiments. Results are viewed as a heatmap which can be clustered by gene, experimental condition or both (see Figure 1). To provide a valid comparison of gene expression across disparate experiments, we do rank-order normalization of the data and present differential expression as the change in rank. Differential expression can be displayed relative to the experimental control or to basal expression reported in the Gene Atlas. Raw datasets and our subsequent analysis are available for download.

The *Beta Cell Gene Expression Bank* contains manually curated gene and pathway annotations by Decio L. Eizirik, Daisy Flamez, and colleagues for genes expressed in  $\beta$  cells. Gene annotations include information on the gene's function, its localization, disease association (with special focus on T1D and other autoimmune diseases), other interacting proteins, and the phenotype after gene disruption in knockout/transgenic models. Key original references and reviews are provided.

*TIDMart* is a BioMart-style query tool (26) for marker and genotyping data. The user is able to query marker, genotypic, or genomic data as the starting data sources. The data can then be filtered in different ways, including by position, relationship to other genomic features, or subject-based information such as sex and whether the subject is a founder. It is possible to query multiple data sources at the same time including dbSNP, HapMap, and published JDRF/WT DIL markers. Once the results have been filtered, the user can select the columns of information to be output from the matching data. The results can be saved in a list for use in other TIDBase tools or output as plain text, either to a file or to the screen.

*Cytoscape* is an open source program for visualizing and analyzing biological networks (10). Cytoscape can be launched from the Gene Overview page to display a gene in the context of proteins and genes that interact with it or run on a list of genes. In either case, the user has the option of choosing the data sources for the interactions shown in the network. Current data sources include human protein reference database (HPRD) (27), biomolecular interaction network database (BIND) (28), Reactome (29), Molecular INTERaction database (MINT) (30) and predicted interactions from (31).

*PosterPages* are 'home pages' about specific topics or datasets. These pages provide a convenient way to add curated scientific content, data provided by collaborators, or new topics. PosterPages can host supplementary material for publications, serve as a data distribution point for particular experiments, or be used to share data among a group of collaborators. Each page is linked to and from other pages on the site to integrate the data or topic with other content. At present, PosterPages are manually



**Figure 1.** Microarray Viewer showing differential expression of a list of genes across several experiments. The figure shows the default heatmap and profile chart displays of the data. Actual data values can be seen by clicking on the Data view tab. The data can be filtered, clustered, or downloaded by clicking on tabs near the bottom of the figure.

created by T1DBase staff in collaboration with contributing scientists.

Many T1DBase tools are designed to work with lists of genes or markers. Lists are organized in a familiar 'explorer-tree' of folders and sub-folders. Users can create their own lists or use lists that we or other users provide. Lists that we provide include T1D-relevant lists from publications (such as lists of differentially expressed genes from microarray experiments), T1D candidate genes, and genes from genetically identified susceptibility regions (such as genes located within the MHC). The *My T1DBase* section of the site is a central point for creating and manipulating lists, although lists can also be created from the output of various tools.

Site-wide search is available from most pages. Genes, markers, Beta Cell Gene Expression Bank annotations, or T1D candidate regions can be searched using identifiers from public databases, symbols or text. Each search can be run against the entire database or limited by keyword type such as UniGene ID, or dataset such as the Beta Cell Gene Expression Bank. The software uses a combination of keyword recognition and text searching. Search results can be saved in lists and used as input to many other tools on the site.

## DISCUSSION

T1DBase is a unique resource for T1D researchers. By integrating and organizing multiple disparate datasets, T1DBase

makes it possible for scientists to query and explore across the data and find new relationships among the factors that contribute to the complex pathogenesis of T1D. Using data and tools currently in T1DBase, a scientist can start from a gene or pathway of interest, then find all interacting partners, see which of these are expressed in  $\beta$  cells, that show changed expression as T1D progresses, and that are T1D candidate genes. It is a small step to allow gene set enrichment analysis (GSEA) (32) to highlight biological functions shared by the large set of genes that would be identified through this procedure. We are in the process of importing transcription factor binding site information and tools which will enable identification of putative binding sites and transcription factors shared among genes found through this or other procedures which will make it possible to hypothesize further connections among interesting genes. This approach makes sense for other diseases, and we believe that great synergy would result from combining data for multiple diseases.

The opportunities to find novel relationships will grow as more data and tools become available. Of course, this will bring false discovery issues to the fore and will necessitate methods for filtering, classifying, or ranking claimed relationships to reduce the number of false leads. But the key challenge will remain as it is now: providing powerful informatics capabilities to T1D scientists in a form they can use to enhance their research.

## ACKNOWLEDGEMENTS

We gratefully acknowledge the support of our funding bodies: the Juvenile Diabetes Research Foundation, the Wellcome Trust, the Type 1 Diabetes Genetics Consortium, and the National Institute of Diabetes and Digestive and Kidney Diseases. BCH was co-funded by a grant from the Cambridge-MIT Institute. We thank Lee Hood for his thoughtful guidance and the many JDRF/WT DIL lab members whose patient feedback has been invaluable in developing the site. Funding to pay the Open Access publication charges for this article was provided by the Institute for Systems Biology.

*Conflict of interest statement.* None declared.

## REFERENCES

- Smink,L.J., Helton,E.M., Healy,B.C., Cavnor,C.C., Lam,A.C., Flamez,D., Burren,O.S., Wang,Y., Dolman,G.E., Burdick,D.B. *et al.* (2005) T1DBase, a community web-based resource for type 1 diabetes research. *Nucleic Acids Res.*, **33**, D544–D549.
- Burren,O.S., Healy,B.C., Lam,A.C., Schuilenburg,H., Dolman,G.E., Everett,V.H., Laneri,D., Nutland,S., Rance,H.E., Payne,F. *et al.* (2004) Development of an integrated genome informatics, data management and workflow infrastructure: A toolbox for the study of complex disease. *Hum. Genomics*, **1**, 98–109.
- Kaestner,K.H., Lee,C.S., Scearce,L.M., Brestelli,J.E., Arsenlis,A., Le,P.P., Lantz,K.A., Crabtree,J., Pizarro,A., Mazzarelli,J. *et al.* (2003) Transcriptional program of the endocrine pancreas in mice and humans. *Diabetes*, **52**, 1604–1610.
- Goodman,N., McCormick,K., Goldowitz,D., Hockly,E., Johnson,C., Kristal,B., MacDonald,M., Truant,R. and Beuzekom,M.V. (2003) Plans for HDBase—a research community website for Huntington's disease. *Clin. Neurosci. Res.*, **3**, 197–217.
- The Gene Ontology Consortium. (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.*, **34**, D322–D326.
- Hinrichs,A.S., Karolchik,D., Baertsch,R., Barber,G.P., Bejerano,G., Clawson,H., Diekhans,M., Furey,T.S., Harte,R.A., Hsu,F. *et al.* (2006) The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.*, **34**, D590–D598.
- Birney,E., Andrews,D., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cox,T., Cunningham,F., Curwen,V., Cutts,T. *et al.* (2006) Ensembl 2006. *Nucleic Acids Res.*, **34**, D556–D561.
- Ashurst,J.L., Chen,C.K., Gilbert,J.G., Jekosch,K., Keenan,S., Meidl,P., Searle,S.M., Stalker,J., Storey,R., Trevanion,S. *et al.* (2005) The vertebrate genome annotation (Vega) database. *Nucleic Acids Res.*, **33**, D459–D465.
- Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **10**, 1599–1610.
- 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.
- Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvermin,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.
- Blake,J.A., Eppig,J.T., Bult,C.J., Kadin,J.A. and Richardson,J.E. (2006) The mouse genome database (MGD): updates and enhancements. *Nucleic Acids Res.*, **34**, D562–D567.
- de la Cruz,N., Bromberg,S., Pasko,D., Shimoyama,M., Twigger,S., Chen,J., Chen,C.F., Fan,C., Foote,C., Gopinath,G.R. *et al.* (2005) The rat genome database (RGD): developments towards a phenome database. *Nucleic Acids Res.*, **33**, D485–D491.
- O'Brien,K.P., Remm,M. and Sonnhammer,E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**, D476–D480.
- Chen,F., Mackey,A.J., Stoeckert,C.J., Jr and Roos,D.S. (2006) OrthoMCL-DB: Querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.
- Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
- The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
- Roach,J.C., Deutsch,K., Li,S., Siegel,A.F., Bekris,L.M., Einhaus,D.C., Sheridan,C.M., Glusman,G., Hood,L., Lernmark,A. *et al.* (2006) Genetic mapping at 3-kilobase resolution reveals inositol 1,4,5-triphosphate receptor 3 as a risk factor for type 1 diabetes in Sweden. *Am. J. Hum. Genet.*, **79**, 614–627.
- Concannon,P., Erlich,H.A., Julier,C., Morahan,G., Nerup,J., Pociot,F., Todd,J.A. and Rich,S.S. (2005) Type 1 diabetes: evidence for susceptibility loci from four genome-wide linkage scans in 1435 multiplex families. *Diabetes*, **54**, 2995–3001.
- Barrett,J.C., Fry,B., Maller,J. and Daly,M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
- The International HapMap Consortium. (2003) The International HapMap Project. *Nature*, **426**, 789–796.
- Traherne,J.A., Horton,R., Roberts,A.N., Miretti,M.M., Hurler,M.E., Stewart,C.A., Ashurst,J.L., Atrazhev,A.M., Coggill,P., Palmer,S. *et al.* (2006) Genetic analysis of completely sequenced disease-associated MHC haplotypes identifies shuffling of segments in recent human history. *PLoS Genet.*, **2**, E9.
- Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
- Huminiecki,L., Lloyd,A.T. and Wolfe,K.H. (2003) Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. *BMC Genomics*, **4**, 31.
- Skrabanek,L. and Campagne,F. (2001) TissueInfo: high-throughput identification of tissue expression profiles and specificity. *Nucleic Acids Res.*, **29**, E102[E102].
- Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T. and Birney,E. (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.*, **14**, 160–169.
- Mishra,G.R., Suresh,M., Kumaran,K., Kannabiran,N., Suresh,S., Bala,P., Shivakumar,K., Anuradha,N., Reddy,R., Raghavan,T.M. *et al.* (2006) Human protein reference database—2006 update. *Nucleic Acids Res.*, **34**, D411–D414.
- Alfarano,C., Andrade,C.E., Anthony,K., Bahroos,N., Bajec,M., Bantoft,K., Betel,D., Bobechno,B., Boutilier,K., Burgess,E. *et al.* (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res.*, **33**, D418–D424.
- Joshi-Tope,G., Gillespie,M., Vastrik,I., D'Eustachio,P., Schmidt,E., de Bono,B., Jassal,B., Gopinath,G.R., Wu,G.R., Matthews,L. *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, **33**, D428–D432.
- Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecular Interaction database. *FEBS Lett.*, **513**, 135–140.
- Lehner,B. and Fraser,A.G. (2004) A first-draft human protein-interaction map. *Genome Biol.*, **5**, R63.
- 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.