D544–D549 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki095

T1DBase, a community web-based resource for type 1 diabetes research

Luc J. Smink*, Erin M. Helton[1], Barry C. Healy, Christopher C. Cavnor[1], Alex C. Lam, Daisy Flamez[2], Oliver S. Burren, Yang Wang[1], Geoffrey E. Dolman, David B. Burdick[1], Vincent H. Everett, Gustavo Glusman[1], Davide Laneri, Lee Rowen[1], Helen Schuilenburg, Neil M. Walker, Josyf Mychaleckyj[3], Linda S. Wicker, Decio L. Eizirik[2], John A. Todd and Nathan Goodman[1,*]

Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Cambridge Institute for Medical Research, University of Cambridge, Wellcome Trust/MRC Building, Cambridge CB2 2XY, UK, [1]Institute for Systems Biology, Seattle, WA 98103, USA, [2]Laboratory of Experimental Medicine, Free University of Brussels (ULB), Brussels, Belgium and [3]Center for Human Genomics, Wake Forest University School of Medicine, Winston-Salem, NC 27157, USA

Received August 20, 2004; Revised and Accepted October 12, 2004

ABSTRACT

T1DBase (http://T1DBase.org) is a public website and database that supports the type 1 diabetes (T1D) research community. The site is currently focused on the molecular genetics and biology of T1D susceptibility and pathogenesis. It includes the following datasets: annotated genome sequence for human, rat and mouse; information on genetically identified T1D susceptibility regions in human, rat and mouse, and genetic linkage and association studies pertaining to T1D; descriptions of NOD mouse congenic strains; the Beta Cell Gene Expression Bank, which reports expression levels of genes in beta cells under various conditions, and annotations of gene function in beta cells; data on gene expression in a variety of tissues and organs; and biological pathways from KEGG and BioCarta. Tools on the site include the GBrowse genome browser, site-wide context dependent search, Connect-the-Dots for connecting gene and other identifiers from multiple data sources, Cytoscape for visualizing and analyzing biological networks, and the GESTALT workbench for genome annotation. All data are open access and all software is open source.

INTRODUCTION

T1DBase (http://T1DBase.org) is a public website and database that supports the type 1 diabetes (T1D) research community. T1DBase collects information from public sources and collaborating investigators, integrates this information, and presents it in a form that is useful for, and accessible to, T1D researchers. It is analogous to a model organism database but is focused on a specific disease rather than a specific organism. The site contains multiple semi-independent datasets that are curated independently (in some cases by external collaborators), and then unified using integration software developed for this purpose. Figure 1 shows the homepage. All data are open access and all software is open source.

T1DBase is a merger of two separate projects: one at the Institute for Systems Biology (ISB), which was explicitly funded by the Juvenile Diabetes Research Foundation to create a public resource; the other at the Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory (DIL) of the University of Cambridge, whose mission is to develop tools and methods to integrate genomic and genetic data and improve cost-effectiveness in the search for T1D susceptibility genes (1).

T1D is an autoimmune disease in which the insulin-producing pancreatic beta cells are selectively destroyed. T1D is the second most common form of diabetes with a prevalence of 0.4% in Caucasians (OMIM:222100). The disease is caused by a combination of environmental and genetic factors. While most cases are non-familial, disease risk is dramatically higher (15 times) for siblings of an affected individual; many lines of research confirm that the increased risk is at least partially genetic (2). The HLA region on chromosome 6 confers 40–50% of the genetic susceptibility with lesser contributions from the three other known loci: INS (3), CTLA4 (4)

*To whom correspondence should be addressed. Tel: +44 1223 763211; Fax: +44 1223 762102; Email: Luc.Smink@cimr.cam.ac.uk
Correspondence may also be addressed to Nathan Goodman. Tel: +1 206 331 0077; Fax: +1 206 732 1299; Email: natg@shore.net

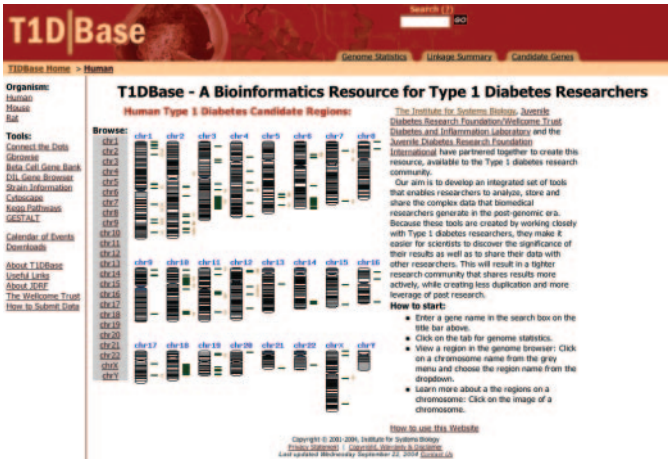

Figure 1. T1DBase homepage. Most pages have a similar format. The T1DBase icon in the top banner returns the user to the homepage. The search box in the upper right corner lets the user search the entire site. Tabs below the search box provide access to related pages. The history links directly below the T1DBase icon allow return to past pages. The navigation bar on the left provides links to the different areas of the site. On the homepage, users can navigate to specific chromosomes or T1D susceptibility regions by clicking on the ideogram. Each page has a footer with links to our privacy policy and other legal fine print.

and LYP/PTPN22 (5). It has been estimated that there could be an additional 50 detectable susceptibility loci which are yet to be identified (W. Y. S. Wang, B. J. Barratt, D. G. Clayton and J. A. Todd, submitted for publication).

The major animal models for T1D are the nonobese diabetic (NOD) mouse, and the BB rat. In addition, research on beta cell function is also carried out on non-diabetic mouse and rat strains.

At present, T1DBase is focused on the molecular genetics and biology of T1D susceptibility and pathogenesis. This represents ~15% of current T1D research and involves more than 1000 investigators, based on a survey of T1D publications.

DATA CONTENT

Annotated genomes

T1DBase provides annotated genome sequence for human, rat and mouse. Currently 32 data tracks are available. The major sources for this information are the public Ensembl (6,7) and UCSC (8) genome databases, augmented by gene annotations produced internally by scientists at DIL and ISB. The DIL, upon dbSNP submission, also publishes its SNPs and primers through T1DBase.

From the large number of data tracks available on Ensembl and UCSC, we have chosen those that are most relevant to T1D investigators. We are in the process of adding tracks and integrating tools not available on Ensembl and UCSC that are of specific interest to our users. Examples appear later.

We are committed to incorporating annotations from scientists in the community. Such annotations are manually reviewed before being added to the database.

T1D prioritized regions

The system has information on prioritized regions in human, rat and mouse.

For human, there are three different kinds of prioritized regions: genetic linkage regions, regions defined through orthology with susceptibility regions in NOD congenic strains and candidate gene regions. There are 20 putative linkage regions, nine orthology regions and numerous candidate gene regions. For linkage and orthology regions, we provide a list of genes in the region. For linkage regions, we also provide a bibliography of publications studying the region, and a summary of LOD scores from the various studies. The candidate genes include ones reported in the literature as having a positive association with T1D, ones reported to be associated with other immune-mediated disorders such as asthma or rheumatoid arthritis that are also candidates for T1D and other genes of interest. The latter includes orthologs of mouse or rat genes associated with T1D, and genes involved in relevant pathways. For genes with literature support, we provide links to the publications.

For mouse, the database contains 27 susceptibility regions defined by the Mouse Genome Database (MGD) (9) with supporting literature. For rat, we have 18 regions and supporting literature from the Rat Genome Database (RGD) (10).

Genetic association studies

We have assembled a dataset of genetic association studies pertaining to T1D in collaboration with the NIH Genetic Association Database (GAD) (11). The dataset covers about 100 genes and 180 publications, including published negative results. Under the collaboration, we carry out literature searches to identify relevant studies and pass the results to GAD for a final quality check and data entry.

NOD strain database

Congenic strains have been created to help identify regions of the NOD genome involved in T1D by introgressing chromosomal regions from resistant strains into the NOD mouse. To visualize the introgressed regions, we have developed a strain database. The strain database has the same data model as the feature database and allows the storage of strain information, such as strain name and its aliases, chromosomes, fine mapping markers and the name of the regions. These data are updated automatically with each update of the NCBI mouse genome build. When an interval is refined, scientists submit the new boundary markers via a webpage and the intervals are recalculated. The intervals are drawn through the Perl GD package. The database and the drawing tools may be of interest to other researchers working with congenic strains. Figure 2 illustrates the way strain information is displayed.

Beta Cell Gene Expression Bank

The Beta Cell Gene Expression Bank is a dataset curated by Decio L. Eizirik and colleagues at the Laboratory for Experimental Medicine at the Free University of Brussels (ULB). There are two main components. One, called the Fast Track, reports the expression level of genes in beta cells under basal conditions and under conditions thought to induce beta cell dysfunction and death in T1D; these data come from a series of microarray experiments conducted in the Eizirik laboratory (12–16). The second component, called the Annotated Track, consists of manual annotation of gene function carried out by beta cell experts; priority is given to genes whose

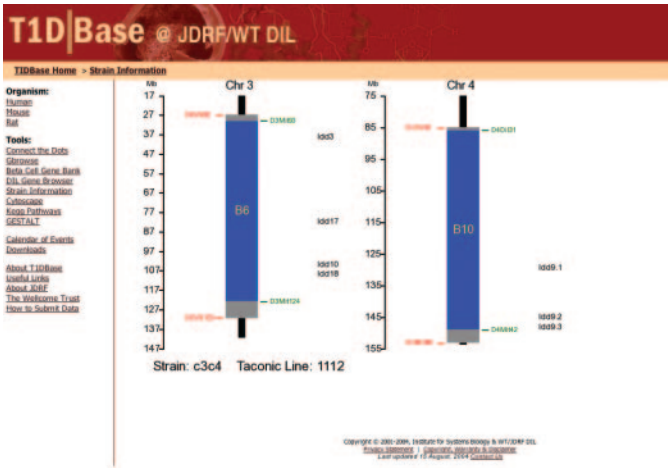

Figure 2. Display of NOD congenic strain A graphical representation of strain c3c4, Taconic line 1112, showing B6 and B10 DNA introgressed into an NOD background as a blue bar. The B6/B10 haplotype boundary markers for the introgressed DNA are shown in green and the NOD boundary markers in red. The regions between known B6 or B10 and known NOD DNA are shaded gray. The Idd designations for the chromosomes shown are displayed in their relative positions. A scale is provided for each chromosome to show genomic location of markers and Idd regions.

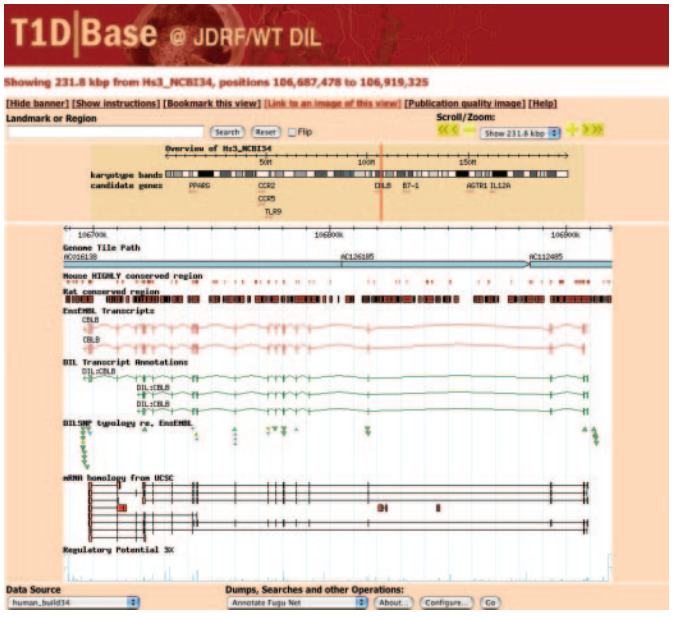

Figure 3. GBrowse display of CBLB gene showing several tracks: the genome tile path, mouse and rat highly conserved regions, Ensembl and DIL transcript annotations, DIL SNP typology, mRNA homology and regulatory potential. Other data track can be switched on using controls on the web page (data not shown).

expression is changed when dysfunction is induced. The annotation includes 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.

The Fast Track currently has data for about 4500 genes from 30 Affymetrix microarray experiments. The Annotated Track contains more than 300 genes at present and is growing at a rate of 40–60 new genes per month.

Gene x tissue expression

This dataset indicates whether a gene is expressed in a limited set of T1D-relevant tissue types, namely, blood marrow, lymph nodes, pancreas, spleen and thymus based on an analysis of UniGene ESTs. This dataset is being replaced by a more comprehensive resource that combines microarray data from the Beta Cell Gene Expression Bank to characterize expression in beta cells and the GNF SymAtlas (17) to characterize expression elsewhere.

Pathways

Genes are linked to pathways in the KEGG (18) and BioCarta (http://www.biocarta.com) databases. The BioCarta pathways are searched using the Cancer Genome Anatomy Project's (CGAP) Pathway Searcher (http://cgap.nci.nih.gov/Pathways/Pathway_Searcher). For KEGG pathways, it is possible to display a table of the genes involved which indicates whether the gene is located within a T1D candidate region.

Links between datasets

When a user accesses a gene from any dataset on the website, a gene page is displayed that provides links to all T1DBase datasets that contain the gene. From this page, the user can also get to GBrowse and most other tools that can manipulate the gene. In addition, the gene page includes links to the following external resources: LocusLink, UniGene, HomoloGene, OMIM, GeneCards and EPConDB. We are in the process of developing similar links within the major tools on the site, so that tools can use information from any dataset to modify how data are visualized or processed.

TOOLS

Generic Genome Browser (GBrowse)

GBrowse (19) is used to visualize genetic and genomic data (Figure 3). Genomic data are extracted from the Ensembl and UCSC genome databases. The Ensembl database is downloaded after each Ensembl release, and the Ensembl API is used to extract the genome features of interest. These are converted into genome feature format (GFF) and loaded into the GBrowse database. From UCSC, certain data types, notably the UCSC mRNA and EST homologies are downloaded, converted into GFF and loaded into the GBrowse database. Currently 32 data tracks are available. Efforts are underway to integrate statistical tools such as selection of tag SNPs (20) and display of D'/R^2 plots for an interval of interest.

An alternative approach to integrating the Ensembl and UCSC data would be to use distributed annotation server (DAS) (21). However the current specification of DAS only allows a limited glyph set, and does not, for instance, allow graphs to be represented.

We make extensive use of the plugin capability provided by GBrowse. A plugin is used to visualize the UCSC dataset of regulatory potential scores (22). This is a very large dataset, which we prefer not to store in our main GBrowse database. Instead, it is imported into a separate database and uses a

plugin to connect GBrowse to the data. Similar plugins are used to visualize Fugu net scores and repeat density plots. We expect to add more plugins as we integrate additional data tracks that do not fit the built-in GBrowse model.

Another plugin facilitates genome annotation. The plugin uses BLAT (23) to align an mRNA sequence to the genome and convert the result into a GFF file. The user can then upload the file and view the annotation in GBrowse. To add the annotation to the permanent database, the user can email the GFF file to T1DBase, the file is then manually verified and loaded into the database.

We also use plugins to allow users to export selected data tracks to a file.

The T1DBase GBrowse provides the T1D research community with a rich genomic data environment by integrating the UCSC and Ensembl genomes and user contributed data.

Search

T1DBase offers a site-wide search capability that works across the multiple datasets present on the site. A technical subtlety is that different kinds of data require different search strategies which the software carries out behind the scenes. Genes are an important special case: the software can search for genes based on a variety of identifiers, including gene names, symbols, LocusLink IDs and UniGene IDs.

The search system is built on the open source Plucene package, a Perl port of the widely used Lucene package (24) (http://www.onjava.com).

Connect-the-Dots

Connect-the-Dots connects identifiers for genes and other entities based on information extracted from multiple data sources. It provides methods for parsing data sources to extract identifiers and connections among identifiers, and loading this information into an internal database. Users can query the database to connect identifiers from any number of sources by following paths composed of the parsed connections. For example, to find literature citations about genes of interest on an Affymetrix chip, a query can connect Affymetrix probeset identifiers to LocusLink identifiers using information from Affymetrix's annotation files and connect the LocusLink identifiers to PubMed identifiers using information in NCBI's LocusLink files. Longer and more complex paths are also possible. Queries are expressed in a special-purpose query language and are translated into SQL by the software.

The system can be used interactively over the Web, or as a batch resource to create specialized translation tables for specific purposes. Many of the translation tables used internally by T1DBase are constructed in this manner.

The current Connect-the-Dots database has information from LocusLink, UniGene (human, mouse and rat), OMIM, IPI, UniProt, HomoloGene, DoTS, several Affymetrix chips, and human and mouse PancChips (pancreas/islet-specific microarrays). The database contains 20 million unique identifiers and 42 million connections extracted from 2 million data source entries.

Cytoscape

Cytoscape (25) is a tool for visualizing and analyzing biological networks, defined broadly to include any collection of interacting bio-molecules. A common use of the software is to display networks of protein–protein and protein–DNA interactions, but it can also be used to display gene networks. A key feature is that Cytoscape can analyze networks in combination with gene expression data, e.g. to discover sub-networks with correlated expression, and annotation data such as Gene Ontology, e.g. to associate sub-networks with biological functions.

Cytoscape can be launched directly from T1DBase, although at present this only works on two demonstration networks. Work is underway to connect Cytoscape to human protein interaction data from HPRD (26), microarray gene expression data from the Beta Cell Gene Expression Bank and other sources and annotations suggesting association with T1D susceptibility.

GESTALT

GESTALT (27) is a workbench for genome annotation that combines automated and manual analysis with an emphasis on rich graphical display of the analysis results. GESTALT can execute a variety of external analysis programs (e.g. for gene recognition) as well as internal analyses (e.g. for compositional complexity analysis). The results are stored in an internal database and can later be retrieved and displayed.

GESTALT analyses have been carried out on most T1D human candidate regions, and the results can be inspected on T1DBase. Several new genes were found through this analysis. For operational reasons, users are not allowed to run their own GESTALT analyses on our website, but can do so on the public GESTALT server at http://db.systemsbiology.net/gestalt/.

IMPLEMENTATION ISSUES

Remapping of features

Local features—meaning annotations that are not in Ensembl or UCSC—are stored in a feature database. The feature database was intended to be a Bio::DB::GFF-shaped database, as used by GBrowse; however, user accountability was required over database inserts, edits and deletes, so various modifications and additions were introduced. The variable GFF field 9 was replaced with a defined set of attributes for each feature type. For each feature, the NCBI build number is linked to the feature's coordinates and these are stored together with the sequence. The database is checked on a daily basis for unmapped features, and the sequences for these features are extracted and mapped onto the genome using BLAT. This storage of sequence also allows for easy remapping after an update of the genome build.

When Ensembl or UCSC issue new releases, we reimport their data and rebuild our GBrowse database from scratch. We then extract local features from the feature database, remap these onto the genome using BLAT and add the remapped features to the GBrowse database.

The remapping process could be made faster through comparison of the new and old genome releases. For genomic regions that are not changed, it can be assumed that all the features contained within the region still have the same coordinates and need not be remapped. However, remapping is currently not a rate-limiting step, and we have not yet attempted this optimization.

Website implementation

The website is implemented in Perl and runs on Linux with the Apache web server. Most of the website uses MySQL as the underlying database engine; the exception is Connect-the-Dots, which uses PostgreSQL due to the complex queries involved. Essentially all web pages are generated by cgi scripts.

We use common Perl modules (CGI, Apache::Session, Template, DBI) for basic web and database functionality, and developed a page template module on top of these to ensure a common look-and-feel. The page template generates the basic look of each page—top banner, side navigation bar and footer material—and handles processing needs such as session tracking, user logins, page titles, error logging and database connections.

ACCESSING THE WEBSITE

Website navigation conforms to standard web paradigms and should be intuitive to web-literate users. A navigation bar on the left-hand side of each page provides links to the different areas of the site, and a search box in the upper right corner allows the user to quickly search for a feature of interest. Each page has tabs that link to closely related pages, such as a link to the Beta Cell Gene Bank from a gene information page. Pages provide links to past pages to make it easier for people to back up or branch their navigation, and we assign meaningful titles to each page so that navigation aids built into most browsers—back and forward buttons and histories—can be used sensibly.

The database is open access and all the data are available for download. The entire database dump can be downloaded, as can all Cytoscape and GESTALT data files. In addition, a few datasets can be downloaded in more succinct form, including definitions of the T1D candidate regions, and summaries of the genes found in these regions. We are working to make more of our content available in convenient formats.

DISCUSSION

Disease researchers require access to a wide variety of data and software tools. T1DBase is an attempt at such an integration, designed with the needs of scientists working on T1D in mind. Providing a comprehensive set of resources in one place accelerates research by reducing the time scientists have to spend searching the web. The integration of data and tools on T1DBase naturally leads the scientist from initial findings to related information.

T1DBase has been designed to be expanded easily; we expect to add more datasets and tools as the project proceeds. An important near-term goal is to add information on protein–protein interactions observed in islets and beta cells, as this is a major area of T1D research. The datasets are reasonably independent of each other and can be curated and managed separately.

While T1DBase is focused squarely on a single disease, the conceptual design should be applicable for many other diseases. We believe that our software is readily adapted for new systems, and welcome the opportunity to work with other disease communities interested in making this happen.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of our funding bodies: the Juvenile Diabetes Research Foundation and the Wellcome Trust. We thank Lee Hood for his thoughtful guidance.

REFERENCES

1. 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. *Human Genomics*, **1**, 98–109.
2. Vyse,T.J. and Todd,J.A. (1996) Genetic analysis of autoimmune disease. *Cell*, **85**, 311–318.
3. Barratt,B.J., Payne,F., Lowe,C.E., Hermann,R., Healy,B.C., Harold,D., Concannon,P., Gharani,N., McCarthy,M.I., Olavesen,M.G. *et al.* (2004) Remapping the insulin gene/IDDM2 locus in type 1 diabetes. *Diabetes*, **53**, 1884–1889.
4. Ueda,H., Howson,J.M., Esposito,L., Heward,J., Snook,H., Chamberlain,G., Rainbow,D.B., Hunter,K.M., Smith,A.N., Di Genova,G. *et al.* (2003) Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature*, **423**, 506–511.
5. Bottini,N., Musumeci,L., Alonso,A., Rahmouni,S., Nika,K., Rostamkhani,M., MacMurray,J., Meloni,G.F., Lucarelli,P., Pellecchia,M. *et al.* (2004) A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. *Nature Genet.*, **36**, 337–338.
6. Birney,E., Andrews,D., Bevan,P., Caccamo,M., Cameron,G., Chen,Y., Clarke,L., Coates,G., Cox,T., Cuff,J. *et al.* (2004) Ensembl 2004. *Nucleic Acids Res.*, **32**, D468–D470.
7. Birney,E., Andrews,T.D., Bevan,P., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cuff,J., Curwen,V., Cutts,T. *et al.* (2004) An overview of Ensembl. *Genome Res.*, **14**, 925–928.
8. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al.* (2003) The UCSC genome browser database. *Nucleic Acids Res.*, **31**, 51–54.
9. Bult,C.J., Blake,J.A., Richardson,J.E., Kadin,J.A., Eppig,J.T. and the Mouse Genome Database Group (2004) The Mouse Genome Database (MGD): integrating biology with the genome. *Nucleic Acids Res.*, **32**, D476–D481.
10. Twigger,S., Lu,J., Shimoyama,M., Chen,D., Pasko,D., Long,H., Ginster,J., Chen,C.F., Nigam,R., Kwitek,A. *et al.* (2002) Rat Genome Database (RGD): mapping disease onto the genome. *Nucleic Acids Res.*, **30**, 125–128.
11. Becker,K.G., Barnes,K.C., Bright,T.J. and Wang,S.A. (2004) The genetic association database. *Nature Genet.*, **36**, 431–432.
12. Cardozo,A.K., Heimberg,H., Heremans,Y., Leeman,R., Kutlu,B., Kruhoffer,M., Orntoft,T. and Eizirik,D.L. (2001) A comprehensive analysis of cytokine-induced and nuclear factor-kappa B-dependent genes in primary rat pancreatic beta-cells. *J. Biol. Chem.*, **276**, 48879–48886.
13. Cardozo,A.K., Kruhoffer,M., Leeman,R., Orntoft,T. and Eizirik,D.L. (2001) Identification of novel cytokine-induced genes in pancreatic beta-cells by high-density oligonucleotide arrays. *Diabetes*, **50**, 909–920.
14. Eizirik,D.L., Kutlu,B., Rasschaert,J., Darville,M. and Cardozo,A.K. (2003) Use of microarray analysis to unveil transcription factor and gene networks contributing to Beta cell dysfunction and apoptosis. *Ann. NY Acad. Sci.*, **1005**, 55–74.
15. Rasschaert,J., Liu,D., Kutlu,B., Cardozo,A.K., Kruhoffer,M., Orntoft,T.F. and Eizirik,D.L. (2003) Global profiling of double stranded RNA- and IFN-gamma-induced genes in rat pancreatic beta cells. *Diabetologia*, **46**, 1641–1657.
16. Kutlu,B., Cardozo,A.K., Darville,M.I., Kruhoffer,M., Magnusson,N., Orntoft,T. and Eizirik,D.L. (2003) Discovery of gene networks regulating cytokine-induced dysfunction and apoptosis in insulin-producing INS-1 cells. *Diabetes*, **52**, 2701–2719.
17. 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.

18. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
19. 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.
20. Johnson,G.C., Esposito,L., Barratt,B.J., Smith,A.N., Heward,J., Di Genova,G., Ueda,H., Cordell,H.J., Eaves,I.A., Dudbridge,F. *et al*. (2001) Haplotype tagging for the identification of common disease genes. *Nature Genet*., **29**, 233–237.
21. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.
22. Kolbe,D., Taylor,J., Elnitski,L., Eswara,P., Li,J., Miller,W., Hardison,R. and Chiaromonte,F. (2004) Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res.*, **14**, 700–707.
23. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
24. Gospodnetic,O. (2003) *Introduction to Text Indexing with Apache Jakarta Lucene*. onjava.com.
25. 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.
26. 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.
27. Glusman,G. and Lancet,D. (2000) GESTALT: a workbench for automatic integration and visualization of large-scale genomic sequence analyses. *Bioinformatics*, **16**, 482–483.