Exploring relationships and mining data with the UCSC Gene Sorter

W.J. Kent, Fan Hsu, Donna Karolchik, et al.

Genome Res. 2005 15: 737-741
Access the most recent version at doi:10.1101/gr.3694705

Resource

Exploring relationships and mining data with the UCSC Gene Sorter

W.J. Kent,[2] Fan Hsu,[2] Donna Karolchik,[2,3] Robert M. Kuhn,[2] Hiram Clawson,[2] Heather Trumbower,[2] David Haussler[1]

[1]*Howard Hughes Medical Institute, University of California, Santa Cruz, California 95064, USA;* [2]*Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064, USA*

In parallel with the human genome sequencing and assembly effort, many tools have been developed to examine the structure and function of the human gene set. The University of California Santa Cruz (UCSC) Gene Sorter has been created as a gene-based counterpart to the chromosome-oriented UCSC Genome Browser to facilitate the study of gene function and evolution. This simple, but powerful tool provides a graphical display of related genes that can be sorted and filtered based on a variety of criteria. Genes may be ordered based on such characteristics as expression profiles, proximity in genome, shared Gene Ontology (GO) terms, and protein similarity. The display can be restricted to a gene set meeting a specific set of constraints by filtering on expression levels, gene name or ID, chromosomal position, and so on. The default set of information for each gene entry—gene name, selected expression data, a BLASTP E-value, genomic position, and a description—can be configured to include many other types of data, including expanded expression data, related accession numbers and IDs, orthologs in other species, GO terms, and much more. The Gene Sorter, a CGI-based Web application written in C with a MySQL database, is tightly integrated with the other applications in the UCSC Genome Browser suite. Available on a selected subset of the genome assemblies found in the Genome Browser, it further enhances the usefulness of the UCSC tool set in interactive genomic exploration and analysis.

[Supplemental material is available online at www.genome.org.]

Since the year 2000, the human genome in the public sector has progressed from a roughly assembled draft that was 85% sequenced to a version in which 99% of the euchromatin has been sequenced and assembled with an exceedingly low error rate. The human gene set has advanced in parallel from a state in which solid mRNA evidence was available on ~25% of the genes and only fragmentary, and often false gene predictions available for the rest, to a point where solid mRNA evidence exists for >75% of human genes. In many cases, the exon and intron structure and the protein translation are known with some confidence, but little is known about the function of the gene. To help address this functional annotation gap, we have developed tools that focus on gene-oriented views of the genome to complement the chromosome-oriented view of the popular UCSC Genome Browser at http://www.genome.ucsc.edu/ (Kent et al. 2002; Karolchik et al. 2003). Of these tools, one of the most significant is the UCSC Gene Sorter (http://www.genome.ucsc.edu/cgi-bin/hgNear).

Genes function and evolve together, and are often best understood by their relationships with each other. The Gene Sorter is a simple but powerful tool that may be used to explore these relationships and collect interesting gene sets. At its heart is a large sorted table with a row for each gene in the organism. Each row contains columns for gene expression, homology, genome position, and other information. The selected gene appears in the top row, and the other rows are sorted such that the genes most closely related in some sense to this gene are nearest the top. The relationships can include shared Gene Ontology (GO) terms (Harris et al. 2004), expression profiles, proximity in genome, and protein similarity measurements, both structural and functional. The Gene Sorter can be configured to display any subset of dozens of columns, which may be filtered to produce restricted gene sets. For example, a user can easily create a gene set containing all kinases on chromosome 21 that are up-regulated in the brain. One can also download protein and various types of DNA sequence associated with a gene set in FASTA format or download the column data in a simple tab-separated format.

The Gene Sorter complements the Genome Browser. By avoiding the clutter of long introns, pseudogenes, transposon-infested gene deserts, and other often uninteresting remnants of the long and winding evolutionary road, the Gene Sorter can often provide much more meaningful information in a single Web page than can the Genome Browser and other chromosome-based viewers. On the other hand, the Gene Sorter is only as good as the underlying gene set on which it is based; this currently limits its application to the human genome and a few well-studied model organisms.

[3]**Corresponding author.**
E-mail donnak@soe.ucsc.edu; fax (775) 703-6375.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3694705.

Results

Description of the Gene Sorter

The Gene Sorter is currently available for the human, mouse, rat, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Saccharomyces cerevisiae* genomes. In the human, mouse, and rat assemblies, the underlying gene set is a synthesis of the SWISS-PROT (Bairoch et al. 2004) (http://us.expasy.org/), GenBank mRNA (Benson et al. 2005) (http://www.ncbi.nih.gov/Genbank/), and RefSeq data (Pruitt et al. 2005) (http://www.ncbi.nlm.nih.gov/

projects/RefSeq/) that are displayed on the UCSC Genome Browser as the Known Genes track. Gene sets for the other organisms were obtained from WormBase (Chen et al. 2005) (http://www.wormbase.org/), FlyBase (Drysdale et al. 2005) (http://flybase.bio.indiana.edu/), the Berkeley *Drosophila* Genome Project (Misra et al. 2002) (http://www.fruitfly.org/), and the *Saccharomyces* Genome Database (Cherry et al. 1998) (http://www.yeastgenome.org/), respectively. The data sets displayed in the columns—which we are constantly expanding—were derived from a wide variety of sources. We update the Gene Sorter database approximately every six months.

By default, the Gene Sorter displays six columns as follows: the row number, the gene name, selected expression data, a BLASTP (Altschul et al. 1997) E-value derived from an alignment between the selected gene and the gene in that row, the genomic position, and a short description of the gene (Fig. 1). Clicking on the label above the column presents a page describing the column contents in detail. Supplemental Table R.1 shows a complete list of the columns available, many of which are hyperlinked to additional information. Clicking on the row number or the gene name selects the gene, moving it to the top of the list. The genomic position field is hyperlinked to the Genome Browser. The description link displays a page containing detailed information about the gene.

The gene details page is shared with the Genome Browser, but has been significantly upgraded as part of the Gene Sorter project. The page is divided into sections that provide descriptive text, sequence data, expression data, protein domains, protein and mRNA structure, links to homologs, and other annotations. The descriptive text is gathered from RefSeq, SWISS-PROT, the model organism databases, GenBank, and the Gene Ontology Consortium. The protein domains are obtained from Pfam (Bateman et al. 2002) and InterPro (Apweiler et al. 2000). The protein structure data is from the Protein Data Bank (PDB) (Berman et al. 2000) and ModBase (Pieper et al. 2004). Secondary mRNA structures are computed with the Vienna tools (Hofacker 2003). The page is fairly large, but is indexed for easy navigation.

The expression data, which are an important part of the Gene Sorter and the shared details page, vary from organism to organism. For each organism, we have tried to locate at least one set of expression data that covers nearly all genes and contains a large number of developmental stages or tissues. For the human and mouse genomes, the primary expression data source is the Gene Expression Atlas 2 from the Genomics Institute of the Novartis Research Foundation (GNF) (Su et al. 2004). It contains two replicates each of 79 different tissues, cell types, and developmental stages in human and 61 in mouse. The Stanford Microarray Database (Ball et al. 2005) provides the *Drosophila* data (Arbeitman et al. 2002), which contain two replicates each of 76 developmental stages. The *C. elegans* data are from the Stuart Kim Lab at Stanford University (Jiang et al. 2001) and contain three replicates of seven developmental stages. The *S. cerevisiae* data are provided by Stanford University (Cho et al. 1998), containing a single replicate of 17 points in the cell cycle. In the current implementation, the microarray data is gene specific rather than transcript specific. We are grateful to all of those who helped create these expression data sets and made them publicly accessible.

By default, the Gene Sorter sorts the displayed genes by their similarity in expression to the selected gene. This similarity is calculated as a weighted sum of differences in log expression ratio values. Genes can also be sorted by protein similarity, location in the genome, name, and shared annotation terms. Supplemental Table R.2 lists the sorting options currently available for the human known genes collection.

The Configuration page controls which columns are displayed and in which order. This page also determines the display configuration of individual columns. By default, the expression columns show only a selection of 10–15 tissues, but can be con-

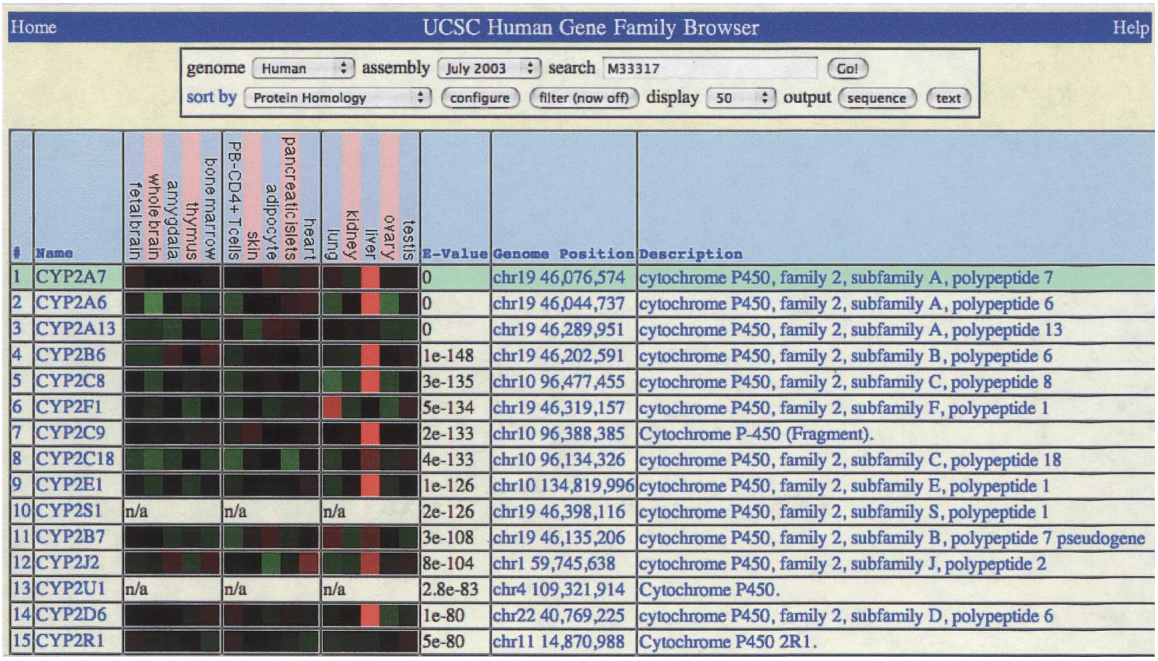

Figure 1. The Gene Sorter main page showing various members of the Cytochrome P450 family 2 sorted by protein homology with the gene *CYP2A7*. This gene family is important for drug metabolism. The selected expression data from the GNF Atlas 2 shows that most, but not all, members are highly expressed in the liver.

figured to display all tissues and even all replicates of all tissues. Absolute expression levels can be shown instead of ratios, and the ratios can be displayed in an alternative yellow/blue coloring scheme that is easier for the color blind to interpret than the traditional red/green colors. The Configuration page also lets the user view splicing and promoter variants of a gene (by default, only the isoform encoding the largest protein is shown). Customized configurations may be named and saved for future use.

Users can define their own columns and upload them via the "Custom Columns" button on the Configuration page. The custom columns format is simple, a short header that defines the column followed by one line of data per gene. The first item on the data line is the gene identifier (a Known Gene ID, RefSeq Accession, Ensembl ID, or any of the other identifiers supported by the Gene Sorter). The remainder of the line contains data to display on that row in the custom column.

Using the Filter page, one can restrict the display to a specific set of genes. This page provides filtering controls for each column, which limit the displayed genes to those that meet the criteria specified for that column. For instance, the expression filter defines minimum and maximum expression levels for each tissue, the Gene Ontology (GO) filter restricts the gene set to those genes associated with certain GO keywords, and the genome position filter narrows the gene set to a section of a particular chromosome. Filters based on identifiers such as gene name, SWISS-PROT ID, and so forth, work in two ways; the user can directly type in a list of identifiers, including wildcards, that will pass the filter, or can upload or paste in longer lists from a file. As with configurations, customized filters may be named and saved for later use.

In the filtered sorted gene list, which is the main display of the Gene Sorter, only the first 50 genes are shown to reduce the load on Web servers and browsers. This number is configurable by a drop-down menu on the main page. This page also provides buttons to display the sequence associated with each gene and to list the genes in a simple tab-separated text format. Regardless of the main display limit, all genes passing the filter are listed in the sequence and tab-separated output. The flexible sequence retrieval mechanism supports protein, mRNA, UTR, genomic (including introns), upstream, and downstream sequence.

Use on biological problems

A common use of the Gene Sorter is to look for other genes that may be involved in the same process as a particular gene. To do this, one types the gene name (for instance, *CYP2A7*) into the search position box and selects a "sort by" option. It is typically a good idea to initially sort the list by name similarity. This displays a list of genes with names that start with the same prefix as the selected gene name. Because biologists frequently assign genes names that correspond to functional families, this can be a quick way to get a handle on what is already known about a gene and its relationships. Displaying the GO column and then sorting on GO similarity is also a very helpful way to learn about existing annotations. However, a gene is likely to relate to uncharacterized genes as well. Sorting by expression similarity and by protein homology can often reveal additional genes that work together with the gene of interest.

Another common usage scenario involves browsing through candidate genes in a region identified by SNP-association studies or other mapping efforts. There are two ways to do this. The user can select a gene in the middle of the region and sort by gene distance. Alternatively, one can configure a filter on the genome position column that restricts the set to the region of interest, and then sort the list by chromosome. The first approach lists genes in the middle of the region first and regions near the outskirts last, but alternates between genes before and after the selected gene. The second approach shows all of the genes in the region, ordered from the start to the end of the region.

Additional filters are available to help users fish for genes of particular interest. For instance, a filter that combines high expression in pancreatic islets cells with GO annotations of membrane proteins can quickly provide a list of candidate genes that might be useful to a researcher exploring the role of autoimmunity in type I diabetes.

The Gene Sorter also can provide interesting data sets for scientists interested in *cis*-regulatory regions. For instance, to collect a data set of putative promoter regions in genes involved in early brain development, one could set up a filter for genes upregulated in the fetal brain, but not the adult brain. Using the sequence button, the user could extract a region ranging from 1000 bases upstream to 50 bases downstream of the annotated transcription start and use this sequence for a DNA motif-finding program such as Weeder (Pavesi et al. 2004).

Discussion

Other good tools exist for browsing and collecting gene sets. The Entrez browser at NCBI (Maglott et al. 2005) (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) is one of the oldest, most popular, and most powerful of such tools. Recently, NCBI has created an Entrez browser on their gene database that—like the Gene Sorter—supports searches on GO terms, chromosome positions, and wildcard searches of various gene-description fields. Searches may be combined via the preview/index function.

The EnsMart data retrieval tool on the Ensembl (Kasprzyk et al. 2004) Web site (http://www.ensembl.org/) is also popular with data-miners. EnsMart searches, which are defined by a series of HTML forms, are more flexible on some fields such as SNPs, but less flexible with the treatment of other fields such as expression. The Genome Alignment and Annotation (GALA) database (Giardine et al. 2003) (http://gala.cse.psu.edu/) is another flexible data-mining tool that is especially good at integrating comparative sequence information. The UCSC Table Browser (Karolchik et al. 2004) (http://www.genome.ucsc.edu/cgi-bin/hgTables) also can be used to integrate comparative sequence information and other information that can be represented as annotation "tracks" on the genome.

Numerous other tools are available for mining microarray data, including the Microarray Explorer (Lemkin et al. 2000) and the Stanford Microarray Database (Gollub et al. 2003). The Pfam, Interpro, and Superfamily (Gough and Chothia 2002) display tools are just a few of the good tools available for classifying and browsing gene families based on protein homology.

While there are advantages to all of these tools, the breadth of the data in the UCSC Gene Sorter, combined with the simple relationship between displayed columns and searched fields, makes it a worthwhile addition to the bioinformatician's data-mining tool collection and provides a useful new view for biologists exploring specific genes.

The Gene Sorter's design facilitates the addition of new columns and sorting types. The introduction of protein/protein in-

teraction data, as well as gene-pathway information, should be especially interesting, because it will add another orthogonal dimension to the existing data. We are also hoping to add chromatin immunoprecipitation and more microarray data, as well as additional links to other useful Web sites.

Conclusions

The UCSC Gene Sorter is a simple, but powerful tool useful both for interactive exploration and for creating gene sets for batch analysis. It integrates data from a wide variety of sources into a single place, where it can be browsed and queried conveniently. The code underlying it is modular, highly configurable, and relatively small. We hope that it will be useful to biologists and bioinformaticians alike.

Methods

The Gene Sorter is a CGI-based Web application written in C that uses a MySQL database. The source code is available from the source link at http://www.soe.ucsc.edu/~kent/. The CGI executable is called hgNear; the source code is in a directory of the same name. The databases may be obtained by following the downloads link at http://www.genome.ucsc.edu/. The Gene Sorter, which is part of our third generation of Web applications at UC Santa Cruz, incorporates many techniques that we learned while extending the Genome Browser.

CGI is the oldest and simplest way to link a Web URL to a program. Generally, when a Web server receives a URL, it translates this into a file that it sends back to the Web browser that requested it. If the URL specifies a CGI instead, the Web server runs the CGI program. The program is passed the values of any fields in a form on the Web page as input, and produces an HTML page as output. Each time the user clicks on a button or follows a link, a CGI is newly invoked.

A challenge faced by all CGI programs is how to remember the interactions a user has had with the Web site so far and to display the particular information the user seeks rather than just a generic Web page. This problem has three solutions. One can have a separate CGI script for each Web page. This is the simplest solution, but has the disadvantage that—over time—the scripts grow to depend on each other in unobvious ways. A second solution is to store all the context information in "hidden" form variables that are embedded in the Web page, but unseen by the user. This solution works well for small- to mid-sized Web applications, but the program state is lost when the user closes the Web browser window. Also, as the information stored in hidden variables becomes large, one is forced to switch from the "get" method—where CGI variables are obvious to other programmers—to the "post" method. This has the side-effect of forcing the user through a confirmation dialog every time the "back" button is used, and makes it more difficult to determine how to create deep links into the Web site. A third solution, which we have adopted, is to store the user's state in a database and save only a single user or session identifier in a cookie or CGI hidden variable. We call this database of user and session variables the "cart." The cart is shared by the Genome Browser, the Table Browser, and the Gene Sorter. By convention, cart variables belonging to the Gene Sorter have the prefix "near."

The cart is the only read/write database in the system. Gene Sorter data are stored in four other databases that are read-only from the Gene Sorter's perspective. The "go" database is loaded directly from the MySQL dumps at http://www.geneontology.org/. Gene associations for supported organisms are loaded into a "goaPart" table in that database. The "swissProt" database is parsed out from SWISS-PROT flat files using the spToDb program developed at UCSC. This program takes about 15 min to produce a highly normalized set of tables that contains everything in the flat files. The microarray expression data are stored in the "hgFixed." database. Each set of expression data is represented by at least two tables, one containing a row for each mRNA sample passed over the array and a second containing a row for each probe in the array. In some cases, additional tables represent the data as ratios as well as absolute values, or condense the data to include just the median value of replicated experiments. The fourth database—which is shared with the Genome Browser—primarily maps various features to locations in the genome. There is a separate version of this database for each assembly of each organism. At the time of this writing, the latest human genome database is hg17, and the most recent mouse version is mm5.

Mapping from the databases to the HTML table on the main Gene Sorter page is relatively simple. For the most part, every cell in every row of the table is filled in with the results of a separate SQL query. A set of configuration files defines the SQL query associated with each column, as well as other column attributes, such as the symbolic name, short and long labels, and the searches that can be applied to the column. The configuration files are stored in a directory hierarchy with three levels, i.e., root, organism, and assembly. Configurations at the organism level override configurations at the root level, and assembly-level configurations override those at the organism level. The configurations are in a simple format that we call ".ra" (for no better reason than it rhymes with the common sequence format ".fa"). The records in a .ra file are separated by blank lines. Each field occupies one line. The first word in the line is the name of the field, and the remainder of the line defines the field contents.

There are currently three types of .ra files. The largest—columnDb.ra files—describe the column configuration using one record per column. The type field in these files is especially important, because it specifies the code the Gene Sorter CGI should execute to display the column. Supplemental Table M.1 lists the available column types. The orderDb.ra file contains a record for each sort method available in the table and also a type field, listed in Supplemental Table M.2. The genome.ra file contains a single record that primarily describes which tables to use for the gene set of a particular organism. The file hgNearData.doc describes the configuration files in detail.

The C source code, which follows object-oriented conventions, contains four important classes of objects, i.e., genes, columns, orderings, and search results. The column object is the most elaborate, containing polymorphic methods (implemented as function pointers) that are set depending on the columnDb.ra type field. These methods are listed in Supplemental Table M.3. The main routine and the simpler column types are implemented in hgNear.c. The more complex column types are each implemented in a separate .c file. Generally, the advFilter method is the most complex to implement, because it can be difficult to make it fast enough to achieve interactive speeds. Currently, the filter is recalculated every time the user submits a page. We may find it necessary to do some caching of the filter results in the future. In all, the size of the Gene Sorter source code is somewhat under 8000 lines, not including the library routines it shares with other http://www.genome.ucsc.edu/ programs.

Acknowledgments

W.J.K. and D.H. were supported by NHGRI grant 1P41HG02371, NCI contract 22XS013A. Additionally, D.H. is supported by the

Howard Hughes Medical Institute. We extend our warmest thanks to the members of the scientific community, too numerous to enumerate, who have contributed to the public data repositories, and without whom this work would not be possible. We also thank the many users of http://www.genome.ucsc.edu/ who have given us valuable feedback at all steps of the design and implementation of our Web-based tools.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., et al. 2000. InterPro—an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16:** 1145–1150.

Arbeitman, M.N., Furlong, E.E., Imam, F., Johnson, E., Null, B.H., Baker, B.S., Krasnow, M.A., Scott, M.P., Davis, R.W., and White, K.P. 2002. Gene expression during the life cycle of *Drosophila melanogaster*. *Science* **297:** 2270–2275.

Bairoch, A., Boeckmann, B., Ferro, S., and Gasteiger, E. 2004. SWISS-PROT: Juggling between evolution and stability. *Brief Bioinform.* **5:** 39–55.

Ball, C.A., Awad, I.A.B., Demeter, J., Gollub, J., Hebert, J.M., Hernandez-Boussard, T., Jin, H., Matese, J.C., Nitzberg, M., Wymore, F., et al. 2005. The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res.* **33:** D580–D582.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. 2002. The Pfam protein families database. *Nucleic Acids Res.* **30:** 276–280.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. 2005. GenBank. *Nucleic Acids Res.* **33:** D34–D38.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28:** 235–242.

Chen, N., Harris, T.W., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., Bradnam, K., Canaran, P., Chan, J., Chen, C., et al. 2005. WormBase: A comprehensive data resource for *Caenorhabditis* biology and genomics. *Nucleic Acids Res.* **33:** D383–D389.

Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., et al. 1998. SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.* **26:** 73–79.

Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., et al. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2:** 65–73.

Drysdale, R.A., Crosby, M.A., and The FlyBase Consortium. 2005. FlyBase: Genes and gene models. *Nucleic Acids Res.* **33:** D390–D395.

Giardine, B., Elnitski, L., Riemer, C., Makalowska, I., Schwartz, S., Miller, W., and Hardison, R.C. 2003. GALA, a database for genomic sequence alignments and annotations. *Genome Res.* **13:** 732–741.

Gollub, J., Ball, C.A. Binkley, G., Demeter, J., Finkelstein, D.B., Hebert, J.M., Hernandez-Boussard, T., Jin, H., Kaloper, M., Matese, J.C., et al. 2003. The Stanford Microarray Database: Data access and quality assessment tools. *Nucleic Acids Res.* **31:** 94–96.

Gough, J. and Chothia, C. 2002. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.* **30:** 268–272.

Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., et al. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32:** D258–D261.

Hofacker, I.L. 2003. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31:** 3429–3431.

Jiang, M., Ryu, J., Kiraly, M., Duke, K., Reinke, V., and Kim, S.K. 2001. Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci.* **98:** 218–223.

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.

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32:** D493–D496.

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.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996–1006.

Lemkin, P.F., Thornwall, G.C., Walton, K.D., and Hennighausen, L. 2000. The microarray explorer tool for data mining of cDNA microarrays: Application for the mammary gland. *Nucleic Acids Res.* **28:** 4452–4459.

Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T. 2005. Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Res.* **33:** D54–D58.

Misra, S., Crosby, M.A., Mungall, C.J., Matthews, B.B., Campbell, K.S., Hradecky, P., Huang, Y., Kaminker, J.S., Millburn, G.H., Prochnik, S.E., et al. 2002. Annotation of the *Drosophila melanogaster* euchromatic genome: A systematic review. *Genome Biol.* **3:** research0083.1–83.22.

Pavesi, G., Mereghetti, P., Mauri, G., and Pesole, G. 2004. Weeder Web: Discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.* **32:** W199–W203.

Pieper, U., Eswar, N., Braberg, H., Madhusudhan, M.S., Davis, F.P., Stuart, A.C., Mirkovic, N., Rossi, A., Marti-Renom, M.A., Fiser, A., et al. 2004. MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.* **32:** D217–D222.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. 2005. NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33:** D501–D504.

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.* **101:** 6062–6067.

Web site references

http://flybase.bio.indiana.edu/; FlyBase Drosophila Genome Database home page.

http://gala.cse.psu.edu/; Genome Alignment and Annotation (GALA) Database home page.

http://genome.ucsc.edu/; UCSC Genome Browser home page.

http://genome.ucsc.edu/cgi-bin/hgNear; UCSC Gene Sorter.

http://www.genome.ucsc.edu/cgi-bin/hgTables; UCSC Table Browser.

http://www.ensembl.org/; Ensembl home page.

http://us.expasy.org/; SWISS-PROT (UniProt) Protein Knowledgebase home page.

http://www.fruitfly.org/; Berkeley Drosophila Genome Project home page.

http://www.geneontology.org/; Gene Ontology Consortium home page.

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi; NCBI Entrez browser.

http://www.ncbi.nih.gov/Genbank/; GenBank Database home page.

http://www.ncbi.nlm.nih.gov/projects/RefSeq/; NCBI Reference Sequences home page.

http://www.soe.ucsc.edu/~kent/; Jim Kent's home page, with links to the Gene Sorter source code.

http://www.wormbase.org/; WormBase home page.

http://www.yeastgenome.org/; Saccharomyces Genome Database home page.

Received January 12, 2005; accepted in revised form February 23, 2005.