GALA, a Database for Genomic Sequence Alignments and Annotations

Belinda Giardine, Laura Elnitski, Cathy Riemer, et al.

Genome Res. 2003 13: 732-741
Access the most recent version at doi:10.1101/gr.603103

GALA, a Database for Genomic Sequence Alignments and Annotations

Belinda Giardine,[1] Laura Elnitski,[1,2] Cathy Riemer,[1] Izabela Makalowska,[4] Scott Schwartz,[1] Webb Miller,[1,3,4] and Ross C. Hardison[2,4,5]

Departments of [1]Computer Science and Engineering, [2]Biochemistry and Molecular Biology, [3]Biology, and [4]Huck Institute for Life Sciences, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

We have developed a relational database to contain whole genome sequence alignments between human and mouse with extensive annotations of the human sequence. Complex queries are supported on recorded features, both directly and on proximity among them. Searches can reveal a wide variety of relationships, such as finding all genes expressed in a designated tissue that have a highly conserved noncoding sequence 5′ to the start site. Other examples are finding single nucleotide polymorphisms that occur in conserved noncoding regions upstream of genes and identifying CpG islands that overlap the 5′ ends of divergently transcribed genes. The database is available online at http://globin.cse.psu.edu/ and http://bio.cse.psu.edu/.

The determination and annotation of complete genomic DNA sequences provide the opportunity for unprecedented advances in our understanding of evolution, genetics, and physiology, but the amount and diversity of data pose daunting challenges as well. Three excellent browsers provide access to the sequence and annotations of the human genome, viz., the human genome browser (HGB) at UCSC (Kent et al. 2002) (http://genome.ucsc.edu/), Map Viewer at the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/), and Ensembl (Hubbard et al. 2002) at the Sanger Centre (http://www.sanger.ac.uk/) and EBI (http://www.ebi.ac.uk/). These sites show known and predicted genes, repetitive elements, genetic markers, and many other types of information as separate tracks in a display using the coordinate system of the human genomic DNA sequence. Thus, users can visually integrate information from all these sources. However, browsers show a single gene or region at a time. They are not designed to support complex queries across multiple types of information simultaneously or multiple genes. For instance, one cannot easily find all genes encoding zinc-finger proteins that are expressed in lung cells, or single nucleotide polymorphisms (SNPs) that occur within a specified distance of such genes. In contrast, recording appropriate annotations in a relational database allows users to query multiple features, thereby supporting searches for nonobvious relationships. The need for greater data integration is leading to new data-mining capacities, such as EnsMart at Ensembl.

Alignments of genomic DNA sequences between species at an appropriate phylogenetic distance are useful for finding candidate functional regions (Hardison and Miller 1993; Gumucio et al. 1996; Hardison et al. 1997; Pennacchio and Rubin 2001). However, this approach is complicated by the variation in divergence rate from locus to locus (Wolfe et al. 1989; Endrizzi et al. 1999; Hardison 2000; DeSilva et al. 2002; Waterston et al. 2002; Hardison et al. 2003). Additional information combined with sequence conservation can refine predictions of functional sequences (Levy et al. 2001). One way to do this is to record both extensive annotations and sequence alignments in a database.

We have developed a database of genomic DNA sequence alignments and annotations, called GALA, to search across tracks of data supplied by current browsers, alignment resources, and databases. Output from GALA queries can be viewed as tracks on the Human Genome Browser, as a table of data with hyperlinks, or as text. When users of the GALA database wish to view alignments, they can choose to display their query output using local alignments with Java (*Laj*), which is a versatile, interactive alignment viewer implemented using Java (Wilson et al. 2001).

[5]Corresponding author.
E-MAIL rch8@psu.edu; FAX (814) 863-7024.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.603103.

RESULTS

The GALA database records information about human genomic DNA (such as genes, repeats, SNPs, CpG islands, recombination frequencies), the structure and function of the products encoded by the genes (including biological process, molecular function, conserved domains, expression patterns), pathological consequences of mutations, and similarity between human and mouse genomic DNA sequences (Table 1). GALA makes extensive use of publicly available datasets, and it incorporates the results of alignments of the human and mouse genomes (Waterston et al. 2002; Schwartz et al. 2003). GALA complements and extends the current resources by allowing simple and complex queries across multiple types of information, effectively integrating evolutionary insights (alignments) with functional and structural annotations. Often, users wish to know about features within a certain distance of other features, rather than specifying precise coordinates. To accomplish this, GALA allows queries for features that are in proximity to or cluster with other features.

Query Page

The fields available for querying appear on a single page organized as scrollable list boxes, drop-down menus, fill-in text boxes, and/or check boxes for each type of information (examples are in Fig. 1A). Having users choose entries from list

13:732–741 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

Table 1. Annotation Statistics and Sources for Fields in the GALA Database

Category	Entries	Source	Example fields from this category	Reference	URL
Genes	35,535	LocusLink at NCBI	Name, type, orientation, exons, coding	Pruitt and Maglott 2001	http://www.ncbi.nlm.nih.gov/LocusLink/
Genes	865	RefSeq at NCBI and HGB	Name, type, orientation, exons, coding	Pruitt and Maglott 2001	http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html http://genome.ucsc.edu/
Gene products and function	17,388	LocusLink at NCBI	Product, biological process, cellular component, molecular function, conserved domain	Pruitt and Maglott 2001	http://www.ncbi.nlm.nih.gov/LocusLink/
Expression data	602,388	UniGene at NCBI	Tissue	Wheeler et al. 2002	http://www.ncbi.nlm.nih.gov/
Genetic disorders	2,802	OMIM	Disorder	Hamosh et al. 2002	http://www.ncbi.nlm.nih.gov/omim/
Alternate gene model: Acembly genes	123,238	Acembly and HGB	Name, type, orientation, exons, coding	J. Thierry-Mieg et al., unpublished	http://www.acedb.org/Cornell/acembly/ http://genome.ucsc.edu/
Alternate gene model: Ensembl genes	27,561	Ensembl and HGB	Name, type, orientation, exons, coding	Hubbard et al. 2002	http://www.ensembl.org/ http://genome.ucsc.edu/
Alternate gene model: Genscan genes	42,737	Genscan and HGB	Name, type, orientation, exons, coding	Burge and Karlin 1997	http://genes.mit.edu/GENSCAN.html http://genome.ucsc.edu/
Alternate gene model: RefSeq genes	16,222	RefSeq and HGB	Name, type, orientation, exons, coding	Pruitt and Maglott 2001	http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html http://genome.ucsc.edu/
Alternate gene model: Twinscan genes	25,744	Twinscan and HGB	Name, type, orientation, exons, coding	Korf et al. 2001	http://genes.cs.wustl.edu/ http://genome.ucsc.edu/
Local alignments	1,585,186	MGSC and HGB	Length, percent identity, gap size, identity step	Waterston et al. 2002; Schwartz et al. 2003	http://bio.cse.psu.edu/ http://genome.ucsc.edu/
Gap free alignments	33,970,427	MGSC and HGB	Length, percent identity	Waterston et al. 2002; Schwartz et al. 2003	http://bio.cse.psu.edu/ http://genome.ucsc.edu/
SNPs	1,956,922	dbSNP at NCBI and HGB	Type, allele, frequency	Sherry et al. 2001	http://www.ncbi.nlm.nih.gov/SNP/ http://genome.ucsc.edu/
Repeats	4,891,898	HGB and Repeat-Masker	Name, class, family	Kent et al. 2002; Smit and Green 1999	http://genome.ucsc.edu/ http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker
CpG islands	26,942	HGB	Name	Kent et al. 2002	http://genome.ucsc.edu/
Transcription factor binding sites	7,655,424	TRANSFAC, Cister, and tffind	Factor name, strand, score	Wingender et al. 2001; Frith et al. 2001	http://www.gene-regulation.com/pub/databases.html http://sullivan.bu.edu/~mfrith/cister.shtml
Recombination rate	8,475	deCODE, Marshfield, Genethon and HGB	Marker, recombination rate, range	Kong et al. 2002; Browman et al. 1998; Hudson et al. 1995	http://www.decodegenetics.com/ http://research.marshfieldclinic.org/genetics/Map_Markers/maps/IndexMapFrames.html http://www.genethon.fr/php/index_us.php http://genome.ucsc.edu/

Note: Users query on fields such as those listed as examples in column 4. The number of entries for each field is subject to change as the source databases update their entries. For all categories except gene products and functions, the number of entries is simply a count of the number of rows in the database table. For gene products and function, the number of entries is the number of gene rows that have data in this category. NCBI, National Center for Biotechnology Information at NIH; OMIM, Online Mendelian Inheritance in Man; HGB, Human Genome Browser at UCSC; and MGSC, Mouse Genome Sequencing Consortium.

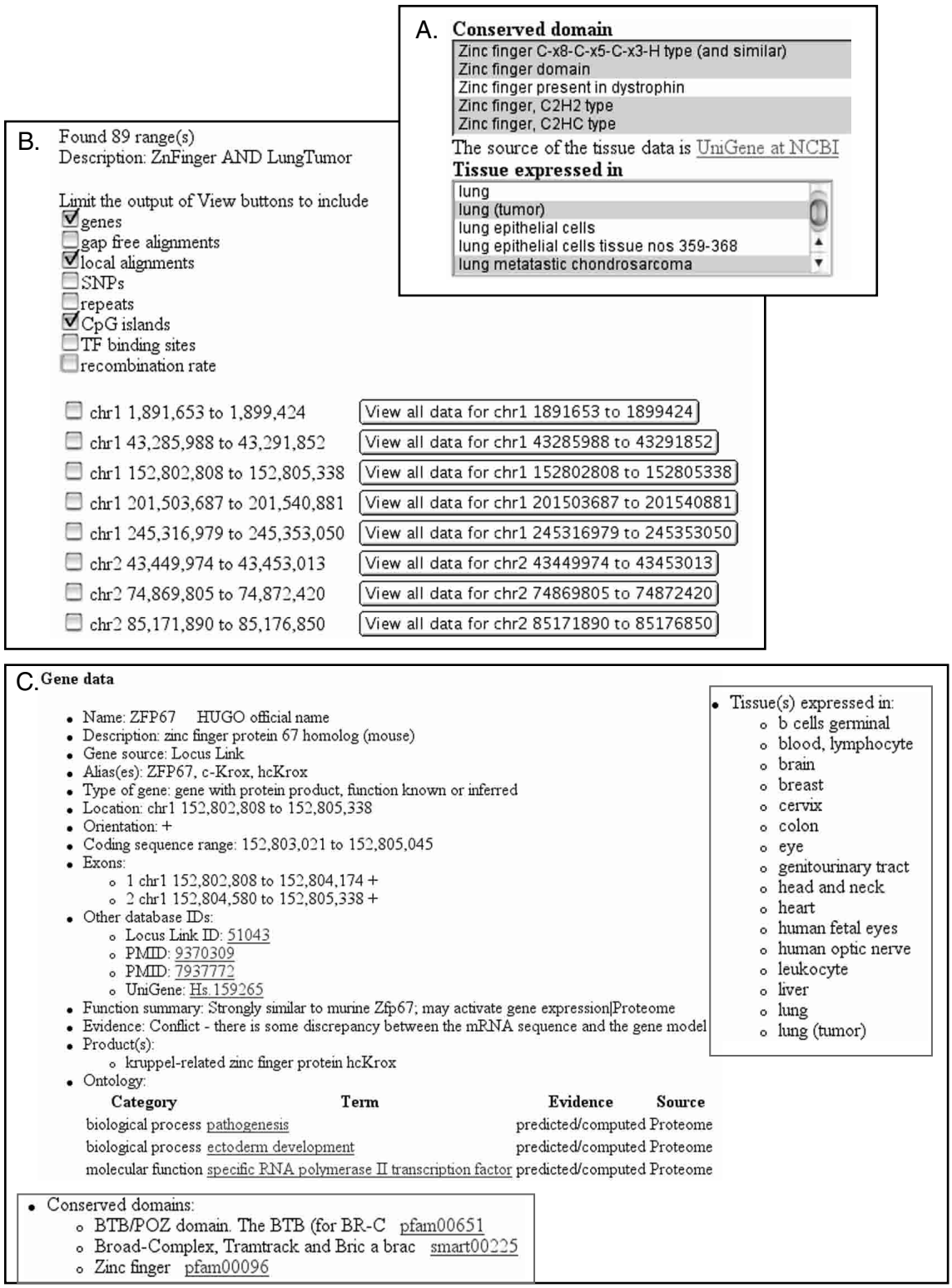

Figure 1 Examples from the GALA query and output pages. (*A*) List boxes, check boxes, and fill-in text boxes are provided for selection of field values on the query page. This panel illustrates the use of list boxes to query for a gene encoding a protein with a zinc finger domain and that is expressed in lung cancers or tumors. (*B*) The GALA "region summary" output screen prompts a user to select the features they would like to see when the corresponding "view" buttons are pressed. (*C*) Selected sections of fields from the data returned for chr 1 152,802,808 to 152,805,338 are shown, including information about the gene and its function, range of tissues in which it is expressed (partial view), and conserved domains in the product.

boxes and drop-down menus for most fields enforces a controlled vocabulary. Many of the fields can be searched for all values by clicking a check box, which is a shortcut to highlighting every item in the field. Fill-in boxes are used to enter coordinates, and they are employed in cases where the list of choices is too long for a scrollable list to be practical (e.g., lists with over 1000 choices). URL hyperlinks adjacent to data fields on the query page point the user to the original source of the data.

Output Formats

GALA provides a variety of choices for the output format of the results. The "region summary" and "gene summary" produce index pages (Fig. 1B) from which one can further explore the selected regions, examining overlap with any or all of the categories available: genes, gap-free alignments, local alignments, SNPs, repeats, CpG islands, or transcription-factor binding sites. Clicking on the "View all data for . . ." button (Fig. 1B) will bring up a page with the information selected (via the check-boxes) for the range returned as the result of a query (Fig. 1C). Plain-text output is provided for off-line examination and importing into other programs.

Three graphical displays are available. The first is a custom user track on the UCSC Human Genome Browser (Kent et al. 2002), which shows an interval returned as a query result in register with any other chosen tracks at the browser (Fig. 2). In this mode, one initially uses GALA to query across multiple fields and genes to find the desired results, and then upon request, the program automatically connects the user to the browser so that its graphical display is used to summarize desired features of the region and its surrounding context. For examining sequence alignments and annotations interactively, query results can be ported automatically to our *Laj* alignment viewer (Wilson et al. 2001), as illustrated in Figure 3. Finally, the distribution of query results across the chromosome can be examined as a bar graph.

Simple Queries

Simple queries are made using one or more fields from the query page. If more than one field is chosen, the output will be an intersection of information from the data tables. An example of a simple query specifies "Find all genes encoding zinc-finger proteins that are expressed in lung tumors or cancers." The user chooses "Zinc finger", "Zinc finger domain" and related terms from the list box for "Conserved domain", and then "lung (tumor)", "lung (carcinoma)", and related terms from "Tissue expressed in", as illustrated in Figure 1A. This finds 89 genes, which are listed on the results page; a partial list of genomic regions containing these is shown in Figure 1B. (The number of records returned is subject to change as the database tables are updated with more information.) One can select a view by genes instead of regions on the query page. The user selects the classes of fields to view on the results page (Fig. 1B). Clicking on the button to "View all data for chr 1 152,802,808 to 152,805,338" returns the requested information, showing that the region contains *ZFP67* (*c-Krox*), which is a Zn finger gene expressed in lung tumors (as expected; Fig. 1C). The results include much additional information, for example, about gene structure, expression in other tissues, proposed functions, and other conserved do-

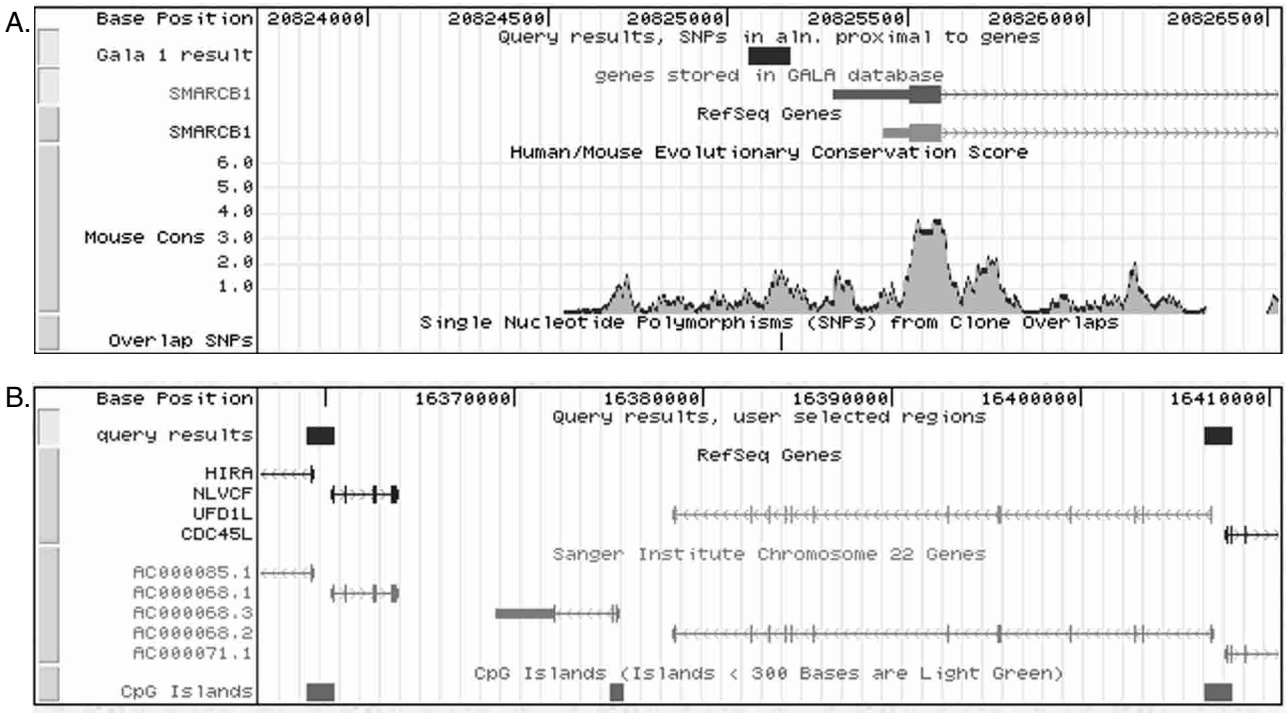

Figure 2 Sample output from GALA visualized in the Human Genome Browser. A clickable button on the output page creates a custom track for the browser that opens in a separate window. (*A*) The region displayed is the 5′ end of the *SMARCB1* gene, which is one of the regions returned by a query for single nucleotide polymorphisms in conserved noncoding regions near the 5′ ends of genes. (*B*) The region displayed includes two of the CpG islands located between divergently transcribed genes, close to positions 38,480,000 and 38,630,000.

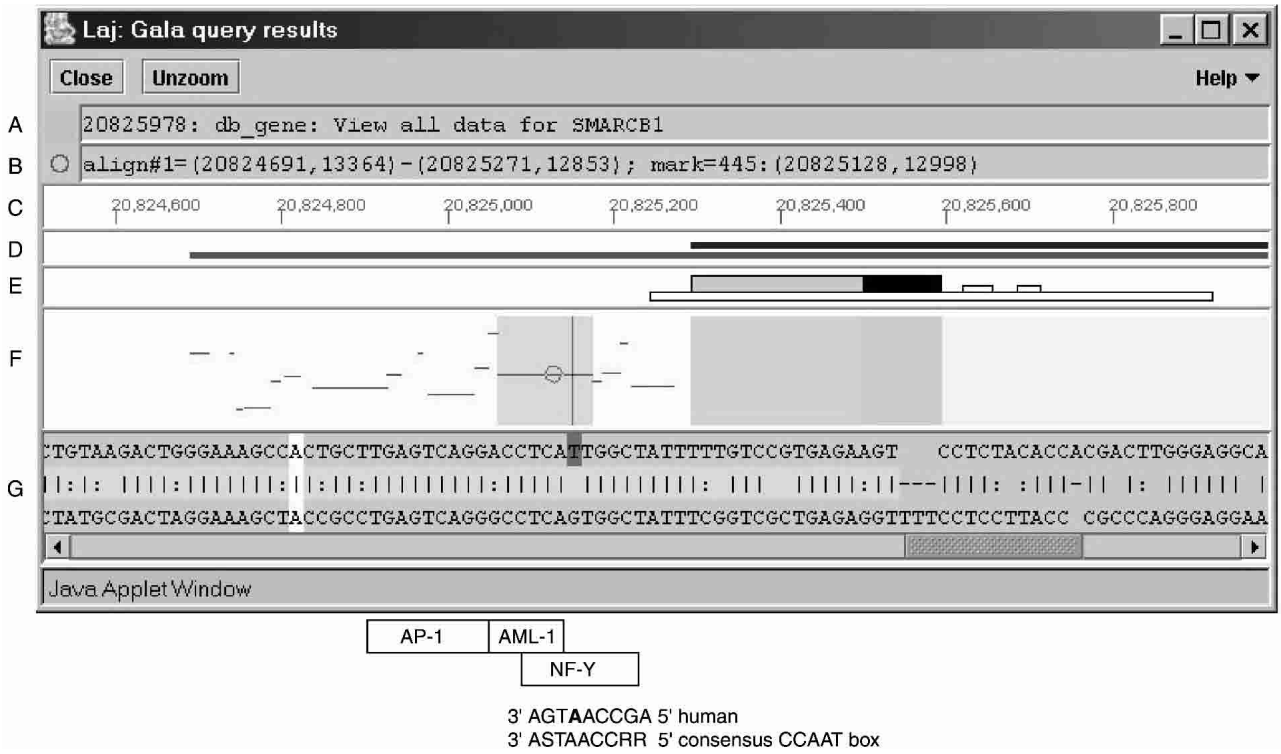

Figure 3 Sample output from GALA visualized in the *Laj* viewer, a Java applet for viewing alignment results. The figure shows a display window with panels for (*A*) the position of the mouse pointer and identification of any objects at that location, (*B*) the position of the moveable circle, (*C*) Human Genome Browser (HGB) coordinates for the region and (*D*) hyperlinks to alignment information (gray) and data for genes (black). (*E*) Icons for genomic features including locations of coding exons (dark-filled taller box), untranslated regions (UTRs) (gray-filled taller box), CpG islands (long, low open box), and simple repeats (short, low open boxes), as well as interspersed repeats when present. (*F*) The percent identity plot of the alignments in the query results showing the positions of the aligning segments in human on the horizontal axis and percent identity of each gap-free segment on the vertical axis. Important features are highlighted with underlays, including coding exons, UTRs, introns, highly conserved noncoding regions (at least 100 bp gap-free and at least 70% identity), and the single nucleotide polymorphisms (SNP). (*G*) The nucleotide-level alignment for the local alignment marked by the circle in *F*. The polymorphic nucleotide is a T (dark gray) in the reference human sequence. Boxes for matches to transcription-factor binding sites in the vicinity of the SNP are drawn below the local alignments with Java (*Laj*) screen shot. For the CCAAT box (binding site for NF-Y), the sequence of the reverse complement of the human sequence is given, as well as the sequence of the consensus binding site. Note that the SNP (in boldface) is part of the consensus-binding site. Alignments that are shown in the percent identity plot (pip) include only the ones that were selected in the query. Thus, if a size threshold was applied, only alignments meeting it appear in the pip. Gray horizontal bars in *D* show the positions of all aligning segments. The nucleotide-level view is obtained by clicking on any alignment in the pip. Names of genes and information about repeats appear in the text box at the top of the page (*A*).

mains in the encoded protein, including hyperlinks to more information (Fig. 1C).

Complex Queries and the History Page

Complex queries use the query page multiple times to formulate a question. The results of individual queries are summarized for the user, and the query itself is placed on the history page. The user selects a method of joining multiple individual queries using options on the history page to obtain the desired results. Queries can be combined by the use of "NOT" statements, intersections, or unions. Furthermore, subtraction can be done on whole regions or limited to overlapping subregions. A particularly versatile feature of GALA is that results from one query that are within a specified proximity of results from another query can be requested from the history page. These features are illustrated in the two following examples.

The first example of a complex query is to search for SNPs in conserved noncoding sequences located proximal to the 5′ ends of genes. Many regulatory regions are conserved above the background level of sequence identity in a region, and one may want to search for those that are close to the 5′ ends of genes. Furthermore, it is possible that some SNPs in conserved noncoding sequences could contribute to changes in gene expression. Thus, this query will return candidate regulatory elements that may be involved in alterations in gene expression.

The complex query begins with three simple queries. First, one asks for all the known genes in the database (currently 33,061 entries). Then, one finds all gap-free alignments between human and mouse of 100 bp in length and 70% or more in identity, which are the criteria used by Loots et al. (2000); this returns 89,847 entries. Because many of these gap-free regions fall within exons, the history page will be used to remove alignments in exons from the query results. The third simple query requests the location of all SNPs in the database (currently 1,936,438 entries).

These three simple queries are then filtered and com-

bined on the history page to produce a dataset that answers the question of interest. First, we restrict alignments to those in proximity to genes; in particular, results from query 2 (highly conserved gap-free alignments) that are within 2000 bp upstream of results from query 1 (all genes) are retained; this is 22,894 regions. Next, genes are subtracted from this set to obtain 12,009 highly conserved, gap-free alignments close to the 5′ ends of genes but not within the genes. Finally, an intersection of the latter set of alignments with results of query 3 (all SNPs) will produce the set of conserved noncoding sequences that are close to the 5′ ends of genes and also have SNPs. The result returns 1053 sequences. One of these falls into the 5′ region of the gene encoding SWI/SNF related, matrix associated actin (*SMARCB1/SNF5*; Online Mendelian Inheritance in Man (OMIM) entry *601607; Figs. 2A, 3). Mutations in the coding exons of this gene are associated with malignant rhabdoid tumors (Versteege et al. 1998).

A second example of a complex query is finding CpG islands between closely spaced, divergently transcribed genes; these are good candidates for promoters. Initially, simple queries are used to find genes (not pseudogenes, which are specified under "type" under "Genes") in a forward orientation (+ strand, which returns 16,842 genes), genes in a reverse orientation (− strand, which returns 16,586 genes), and all CpG islands (which returns 26,942 regions). Then on the history page, the results of the simple queries are combined using the proximity feature to identify CpG islands that are within 2000 bp of the 5′ ends of genes in the two orientations. This gives 6364 and 5868 CpG islands located close to the 5′ ends of genes in forward and reverse orientation, respectively. The intersection of these two sets reveals 1698 ranges containing CpG islands that are within 2000 bp of the 5′ ends of divergently transcribed genes. For example, two of the three CpG islands between positions 16,360,000 and 16,410,000 on chromosome 22 are within the set, and examination in the Human Genome Browser illustrates the divergently transcribed genes flanking them (Fig. 2B). Note that one CpG island in this region does not meet the criteria for proximity to divergently transcribed genes.

Queries to find CpG islands between divergently transcribed genes can be formulated in a variety of ways. An alternative to the one just described is to find all intergenic regions containing CpG islands, and then find the ones that are within 2000 bp of the 5′ ends of genes transcribed in forward direction and also within 2000 bp of the 5′ ends of genes transcribed in reverse direction. This returns slightly different results, because the limitation to intergenic regions removed CpG islands in overlapping genes.

In addition to the methods just discussed for combining results of simple queries on the history page, we also provide a clustering function. Users can select any number of entries in one category to be within a certain distance of entries in another category. One use of the clustering function is to find combinations of transcription factor binding sites. For example, after obtaining NF-E2 binding sites on chromosome 11 from one query (2555 sites) and GATA1 and GATA2 binding sites from a second query (1906 sites), one can find 15 ranges where the NF-E2 binding site is within 50 bp of two binding sites for the GATA proteins. The proximity function also finds ranges where two different features are within a specified distance, but unlike clustering, one cannot specify the number of features (e.g., binding sites for protein A) that must be close to another type of feature (e.g., binding sites for protein B).

Laj Alignment Viewer

Laj (Local Alignments with Java) is a Java applet that allows the user to interactively examine pairwise alignments and associated annotations in a graphical setting (Wilson et al. 2001). The version executed by GALA will display percent identity plots (pips) (Schwartz et al. 2000), text representations of the alignments, a schematic diagram showing the locations of exons and repeats, annotation links to other Web sites containing additional information about particular regions, and colored shading to highlight areas of interest.

Users can easily examine both graphs of alignments and detailed nucleotide-level alignments within the context of annotated genomic features using *Laj*, as illustrated in Figure 3. This shows the human–mouse alignment located 5′ to the human *SMARCB1* gene that also contains a single nucleotide polymorphism in humans, which is one of the regions returned by the first example of a complex query. The pip shows the position and percent identity of each gap-free portion of the alignment, and the icons and other tools show where the alignments are relative to the gene and other features of interest. Clicking on a line representing an alignment will display it in the nucleotide-level view (Fig. 3G). This is done interactively, so users can navigate easily along alignments.

The information viewed by *Laj* can be integrated with other information to draw biological inferences. The SNP is a T (highlighted in dark gray in panel G) in the human reference sequence, and it is a G in the alternative allele. The aligning sequence in mouse is also a G. The SNP is within a match to an NF-Y binding site (a CCAAT box) and is just downstream from good matches to binding sites for AP-1 and AML-1 (Fig. 3). Thus, one could hypothesize that this polymorphism affects expression of the *SMARCB1* gene by altering the binding affinity of NF-Y (or a related protein) and possibly the interactions of DNA-binding proteins in this region. If an investigator has a strong rationale for studying variable expression of the *SMARCB1* gene in human populations, experiments could be designed to test whether the polymorphic site affects expression.

DISCUSSION

GALA is a relational database that allows users to query across different types of information in multiple locations, find features that are proximal to one another, and store and edit previous queries. Users can investigate an enormous variety of issues using GALA; only a few examples were used in this report to illustrate some of the features. Users may want to look for combinations of features expected for regulatory elements, with a complex query such as "Find all highly conserved noncoding sequences within a specified distance of genes expressed in a certain tissue that also have binding sites for a specified set of transcription factors." Others may be interested in evolutionary issues, such as exploring all ancestral repeats that are conserved between human and mouse and are located close to CpG islands. Others may pursue combinations of clinical (e.g., using the field for "disorders" listed in Online Mendelian Inheritance in Man [Hamosh et al. 2002]), biochemical (e.g., using fields for "biological process" and/or "molecular function" of gene product), and genetic (e.g., recombination frequency) features. These rich possibilities should provide new insights that lead to novel experimentation.

GALA complements the existing genome browsers by storing information in a relational database and providing

support for complex queries. It uses the browsers both as the source of some of the data and as one of the ways to display the results of queries graphically, including custom tracks for the Human Genome Browser. The data fields used in GALA are gathered from multiple sources in an automated process that allows quick and efficient updates to the database. Many browsers offer similar categories of data, but their current Web servers are not designed to support complex queries across different data fields or different genes, and they are not currently able to return data on the locations of elements that lie in proximity to one another. These features are incorporated in GALA. Enhanced query capacity and integration are continually being added to the genome browsers, such as EnsMart at Ensemble, Table Browser at the Genome Browser, and Map Viewer at NCBI.

Complex queries are formulated in two stages. The user initially finds the general categories of interest with simple queries and then uses the combination methods on the history page to filter the results. These methods of combining data include intersections and/or unions of data sets, subtracting features not desired in the final output, and extracting features that lie within the same genomic environment with proximity. In addition to searching for proximity of features, users can find clusters of a designated number of features within a specified range. The proximity and clustering features are flexible tools to integrate multiple kinds of information about genomic regions and their conservation.

We show two sample queries that utilize complex querying and return data that are relevant to the study of the human genome. The first query found SNPs in conserved noncoding regions that are proximal to genes. While these results certainly are not conclusive for a role of the SNPs in regulating expression of these genes, they provide candidate sequences that can be examined experimentally. The number of available SNPs has increased dramatically in the past year, and soon it may be common to examine their role in regulatory elements. However, at this time most SNPs are examined as part of the coding regions of genes. The GALA database could become a primary resource for finding SNPs in known and candidate regulatory regions. The second query examines CpG islands that lie between the 5′ ends of divergently transcribed genes. Some of the genes in this category may be in a genomic environment that contains a bidirectional origin of replication, a promoter, and a CpG island, such as has been shown for the *TOP1* gene in humans (Keller et al. 2002).

Improvements planned for GALA include increasing the data recorded (e.g., alignments from multiple species) and expanding the querying capacity. Currently, queries are limited to those that return no more than 7 million entries, but this limitation should be lifted with planned improvements. Regular updates to new genome assemblies and improved datasets are planned.

METHODS

Database Management System

The database is implemented using the Oracle relational database management system version 8i, running on a Sun Blade 100 workstation under the Solaris 8 operating system. The query form Web pages are created dynamically by CGI scripts written in the Perl programming language, which accesses the database via the DBI Perl module.

Genomic DNA Annotation

The human gene annotations are from LocusLink GenBank files found at ftp.ncbi.nih.gov/genomes/H_sapiens/CHR_*. The file ftp.ncbi.nih.gov/genomes/H_sapiens/seq_contig.md is used to calculate the genomic coordinates from the contig coordinates in the GenBank files. The file contains the start points of the contigs in chromosome coordinates, this start point − 1 gives the offset to be added to all the coordinates for genes in that contig. Some of the genes from the RefSeq genes (Pruitt and Maglott 2001) track at UCSC are not included in these files. The missing genes were obtained from UCSC RefSeq (Pruitt and Maglott 2001) track, that is, "refLink" and "refGene" and added to the default set of genes. Each splice variant is entered in the database separately, giving equal status to the variants for querying. This causes the count of rows in the genes table to be larger than the actual number of genes. The database also assigns a unique identifier to each row to use in tying together the annotations for the genes. The LocusLink ID was recorded with all these genes, and this ID was used to match the genes with the annotations.

IDs and accession numbers are used to link the gene entries to information in the resource databases. Data from the resource databases (see below) are automatically entered into the database using programs that parse each file format. We have developed automated procedures to update the data tables, which will be applied to all subsequent updates.

Queries on genes are not limited to the default set of genes; queries on "alternate gene models" are also supported, such as those predicted by Ensembl (Hubbard et al. 2002), Acembly (contributed to the HGB by Jean Thierry-Mieg), GENSCAN (Burge and Karlin 1997), and TWINSCAN (Korf et al. 2001). These alternate gene models do not have the product or function data stored with them, but most of these gene models overlap the database default genes, and annotations for the overlapping default genes can be viewed along with the alternate model genes in the detailed data output.

Information about gene products and their function was imported from the LocusLink database (Pruitt and Maglott 2001) (http://www.ncbi.nlm.nih.gov/LocusLink/). Gene annotations are downloaded into GALA with the following fields: Product, Ontology, Other Annotations, Conserved Domains, and aliases for the gene names. Where possible, gene names are entered into the database by the official nomenclature described at The Human Genome Organization (HUGO) Nomenclature Committee (Povey et al. 2001) (http://www.gene.ucl.ac.uk/nomenclature/). This is assigned as the "principal" name in the database. Also, more aliases and other database IDs for the genes are retrieved with the official names from the file nomeids.txt found at the download site.

Information about the tissue-specific expression profile of each gene with a RefSeq reference is obtained from the UniGene database (Wheeler et al. 2002) (http://www.ncbi.nlm.nih.gov/UniGene/). These data are edited to produce a restricted vocabulary, for example, differences in capitalization, different syntaxes describing tumorigenic tissues, and tissues appended with the word *normal* are changed to a uniform designation.

Information relating human genetic disorders to their associated gene(s) comes from the Online Mendelian Inheritance in Man database, OMIM (Hamosh et al. 2002) (http://www.ncbi.nlm.nih.gov/omim/). GALA uses the OMIM format labeled "Gene Map Key" and stores the data found in "OMIM Gene Map" category. One limitation of the OMIM format is the restriction of no more than three disorders per gene. However, OMIM occasionally lists more than three disorders per gene by separating them with a semicolon. Therefore, upon import into our database, the OMIM data is split at the semicolons and added as separate entries.

Data for SNPs comes from two sources: dbSNP (Sherry et

al. 2001) (http://www.ncbi.nlm.nih.gov/SNP/) and HGB (Kent et al. 2002). dbSNP serves as the main SNP resource when its release date is the same as the human sequence used for annotations. GALA also stores frequency information for the SNP alleles that originates in the dbSNP Sybase table "SubPopAllele". Other fields are filled from the chromosome file in the "ASN1_flat_file" directory. Oracle tracks the HGB designation of TSC (The SNP Consortium) or the National Institutes of Health (NIH) SNP, and these designators appear as choices on the query page.

Information on repeats and CpG islands was imported directly from the HGB (Kent et al. 2002) from annotation tables labeled "chrN_rmsk" and "cpgIsland", respectively. The repeats are found by RepeatMasker (Smit and Green 1999).

Transcription factor binding sites were mapped onto the genome using the programs Cister (Frith et al. 2001) and *tffind* (unpubl.) after the sequence was masked for repeats and coding sequences. The position weight matrices (PWMs) were selected from the TransFac (Wingender et al. 2001) (http://www.gene-regulation.com/index.html) and IMD databases (Dr. Qing Chen, chenq@beagle.colorado.edu) after the datasets were limited to entries containing human, rat, or mouse sequences that have at least 6 nucleotides, not counting Ns, in the core recognition sequence. Some matrices from the TRANSFAC collection do not have a species listing, but are alternative binding sites for common transcription factors (i.e., AP-1); these were included as well.

Recombination data (Hudson et al. 1995; Browman et al. 1998; Kong et al. 2002) were mapped onto the physical coordinates of the June 2002 human genome assembly, as described (Hardison et al. 2003).

Genomic DNA Alignments

Whole genome alignments were computed between the June 2002 assembly of the human genome (Kent and Haussler 2001; Lander et al. 2001) and the February 2002 assembly of the mouse genome (Waterston et al. 2002) as described (Waterston et al. 2002; Schwartz et al. 2003). GALA users can choose "all alignments" to query local alignments of any length, "gap-free alignments" that require a minimum length and percent identity, or "alignments with gaps" that require the user to specify a maximum gap length and identity step across any gaps, in addition to a minimum length and percent identity.

Access

The query page for the database (available from http://globin.cse.psu.edu/ and from http://bio.cse.psu.edu/) is supported by Mac, PC, and Unix Web browsers. Viewing alignments with *Laj* requires Java 1.2 or higher. A Java plug-in is available from Sun Microsystems (http://java.sum.com/products/plugin/) for some systems, but Mac users will need OSX to use *Laj*.

Output Formats

Output from GALA queries can be examined in a variety of formats. One choice is to produce index pages from which one can further explore the selected regions by virtue of overlap with any or all of the categories available. Plain-text output is provided for off-line examination and importing into other programs. Three graphical displays are available: a custom user track on the HGB (Kent et al. 2002), our *Laj* alignment viewer (Wilson et al. 2001), and a bar graph showing the distribution of entries across the chromosome.

Custom tracks for the HGB visualize the results of the query alongside any chosen tracks at the browser. Query results are provided as a clickable button that opens a HGB window slightly larger than the genomic coordinates. Each page is drawn on the fly at HGB and appears in a new window on the browser. Implementation of the custom tracks is described at the URL http://genome.ucsc.edu/goldenPath/help/customTrack.html.

The Java applet *Laj* is an interactive viewer of alignments and associated annotations (Wilson et al. 2001), as illustrated in Figure 3. The hyperlink bars indicate the location of all data for a region; however the alignments shown in the pip correspond only to results of the user's query. Therefore, purple horizontal bars above the percent identity plot may include additional alignments not shown in the pip. Additional features of *Laj*, such as the dot plot, have been disabled from this version of the viewer because of the nature of the alignments. However, the dotplot feature is fully functional when used in conjunction with *lav* output from the PipMaker server (Schwartz et al. 2000; http://bio.cse.psu.edu).

Because GALA provides full annotations for the selected region, it is not practical for *Laj* to display huge sections of the chromosome at once. The time required for GALA to assemble the data files on very large sections of a chromosome and deliver them across the Internet may cause the connection to time out. Also, the information for annotating very large regions could fill the memory on the user's machine. The time and memory limits can be exceeded even if the query returns only two alignments, if they are too far apart. To avoid this problem, GALA checks the size of the region before trying to display it, and if it is longer than 2 Mb, the program prompts the user to select a subregion for display. These pages appear in separate windows, so that one can examine several such ranges at once.

To run the *Laj* applet, the user's Web browser must support at least Java 1.2, which is now available for most computer platforms (a help page with tips for obtaining suitable Java software is provided). We have deliberately avoided using the latest features from newer versions of Java, in order to retain a wider compatibility.

For researchers who wish to use *Laj* to present their own data, a download package containing the program, installation instructions, and additional documentation is available at our Web site (http://bio.cse.psu.edu/). This package can be used to install *Laj* as an applet on a Web server, or to run it in "stand-alone" mode to view local data files. The latter mode does not use a Web browser, though it still requires Java 1.2. It provides more capability to compare and manipulate alignments, but cannot visit linked sites.

History Page

Simple queries can combine one or more fields on the query page to get an intersection of the results. More complex queries are accomplished by using the history page to combine results of simple queries in various ways. The history page has a number of options. For example, queries can be combined by the use of "NOT" statements, intersections, or unions. Furthermore, subtraction can be done on whole regions or limited to overlapping subregions. The database can be used to explore ideas about potential functional regions. The proximity feature is particularly useful for this purpose. Users can choose proximity of their query set within an upstream or downstream region, or both, and restrict an area to lie within, or outside a specified distance from the results of another query. An alternative method of looking at features in combination is through the clustering function. This is useful for finding combinations of transcription factor binding sites. For example, a user could query on all the MyoD binding sites in one query, run a second query for MEF-2 binding sites, and ask for all results that have two or more MEF-2 sites within a certain distance of a MyoD site.

On the history page, users can see, on average, the last 15

queries submitted, which are saved in a browser cookie for up to one week. Additional queries will overload the allowable cookie size and cause the browser to flush the contents of the history page. Therefore, users should click on the "delete selected queries" button to remove unwanted queries before the cookie exceeds the limit. An option to save queries to a separate local file is also provided. Queries that will be repeated after a lapse of 1 wk should be saved this way to prevent their loss when the cookie expires. Both the cookie and the local file save the choices used to formulate the query, rather than its output.

A function to edit previous queries is also available. The old page is opened and a user can simply change the selected fields. Some fields that are in scroll bars should be checked carefully because the default display shows the first entry in the list, even if subsequent entries are selected. The new query will replace the older one on the history page.

The history page is designed to narrow the focus of a query. As more data becomes available for each chromosome, queries will require more time. Thus, a feature on the query page allows a user to select the fields of interest and send a place card to the history page, without actually running the query immediately. This option allows a user to specify and combine preliminary queries, without having to wait for the intermediate steps to run. When the final query is submitted all of the searches are performed.

ACKNOWLEDGMENTS

We thank the individuals and institutions providing the publicly available genomic resources used in constructing this database, and W. James Kent for helpful advice. Our work is supported by PHS grants HG02238 (W.M.), DK27635 (R.H) and HG02325 (L.E.)

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Browman, K.W., Murray, J.C., Sheffield, R.L., White, R.L. and Weber, J.L. 1998. Comprehensive human genetic maps: Individual and sex-specific variation in recombination. *Amer. J. Human Genetics* **63:** 861–869.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

DeSilva, U., Elnitski, L., Idol, J.R., Doyle, J.L., Gan, W., Thomas, J.W., Schwartz, S., Dietrich, N.L., Beckstrom-Sternberg, S.M., McDowell, J.C., et al. 2002. Generation and comparative analysis of approximately 3.3 Mb of mouse genomic sequence orthologous to the region of human chromosome 7q11.23 implicated in Williams syndrome. *Genome Res.* **12:** 3–15.

Endrizzi, M., Huang, S., Scharf, J.M., Kelter, A.R., Wirth, B., Kunkel, L.M., Miller, W. and Dietrich, W.F. 1999. Comparative sequence analysis of the mouse and human Lgn1/SMA interval. *Genomics* **60:** 137–51.

Frith, M.C., Hansen, U. and Weng, Z. 2001. Detection of *cis*-element clusters in higher eukaryotic DNA. *Bioinformatics* **17:** 878–889.

Gumucio, D., Shelton, D., Zhu, W., Millinoff, D., Gray, T., Bock, J., Slightom, J. and Goodman, M. 1996. Evolutionary strategies for the elucidation of *cis* and *trans* factors that regulate the developmental switching programs of the β-like globin genes. *Mol. Phylog. Evol.* **5:** 18–32.

Hamosh, A., Scott, A.F., Amberger, J., Bocchini, C., Valle, D. and McKusick, V.A. 2002. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **30:** 52–55.

Hardison, R. and Miller, W. 1993. Use of long sequence alignments to study the evolution and regulation of mammalian globin gene clusters. *Mol. Biol. Evol.* **10:** 73–102.

Hardison, R., Oeltjen, J., and Miller, W. 1997. Long human–mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. *Genome Res.* **7:** 959–966.

Hardison, R.C. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16:** 369–372.

Hardison, R.C., Roskin, K.M., Yang, S., Diekhans, M., Kent, W.J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., et al. 2003. Co-variation in frequencies of substitution, deletion, transposition and recombination during eutherian evolution. *Genome Res.* **13:** 13–26.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. 2002. The Ensembl genome database project. *Nucleic Acids Res.* **30:** 38–41.

Hudson, T.J., Stein, L.D., Gerety, S.S., Ma, J., Castle, A.B., Silva, J., Slonim, D.K., Baptista, R., Kruglyak, L., Xu, S.H., et al. 1995. An STS-based map of the human genome. *Science* **270:** 1945–1954.

Keller, C., Ladenburger, E.M., Kremer, M., and Knippers, R. 2002. The origin recognition complex marks a replication origin in the human *TOP1* gene promoter. *J. Biol. Chem.* **277:** 31430–31440.

Kent, W.J. and Haussler, D. 2001. Assembly of the working draft of the human genome with GigAssembler. *Genome Res.* **11:** 1541–1548.

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.

Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* **31:** 241–247.

Korf, I., Flicek, P., Duan, D., and Brent, M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17:** S140–S5148.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Levy, S., Hannenhalli, S., and Workman, C. 2001. Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics* **17:** 871–877.

Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., and Frazer, K.A. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288:** 136–140.

Pennacchio, L.A. and Rubin, E.M. 2001. Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.* **2:** 100–109.

Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M. and Wain, H. 2001. The HUGO Gene Nomenclature Committee (HGNC). *Hum. Genet.* **109:** 678–680.

Pruitt, K.D. and Maglott, D.R. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29:** 137–140.

Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. 2000. PipMaker—A Web server for aligning two genomic DNA sequences. *Genome Res.* **10:** 577–586.

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R., Haussler, D., and Miller, W. 2003. Human–mouse alignments with BLASTZ. *Genome Res.* **13:** 103–107.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. 2001. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **29:** 308–311.

Smit, A. and Green, P. 1999. RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html

Versteege, I., Sevenet, N., Lange, J., Rousseau-Merck, M.F., Ambros, P., Handgretinger, R., Aurias, A., and Delattre, O. 1998. Truncating mutations of hSNF5/INI1 in aggressive paediatric cancer. *Nature* **394:** 203–206.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Wheeler, D.L., Church, D.M., Lash, A.E., Leipe, D.D., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Tatusova, T.A., Wagner, L., et al. 2002. Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res.* **30:** 13–16.

Wilson, M.D., Riemer, C., Martindale, D.W., Schnupf, P., Boright, A.P., Cheung, T.L., Hardy, D.M., Schwartz, S., Scherer, S.W., Tsui, L.C., et al. 2001. Comparative analysis of the gene-dense ACHE/TFR2 region on human chromosome 7q22 with the orthologous region on mouse chromosome 5. *Nucleic Acids Res.* **29:** 1352–1365.

Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhauser, R., et al. 2001. The TRANSFAC system on gene expression regulation. *Nucleic Acids Res*. **29:** 281–283.
Wolfe, K.H., Sharp, P.M., and Li, W.H. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* **337:** 283–285.

WEB SITE REFERENCES

http://genome.ucsc.edu/; The human genome browser.
http://www.sanger.ac.uk/; The ensembl genome database.
http://www.ebi.ac.uk/; The ensembl genome database.
http://www.ncbi.nlm.nih.gov/; Site for Map Viewer and GenBank database.
http://www.ncbi.nlm.nih.gov/LocusLink; LocusLink database.
http://www.ncbi.nlm.nih.gov/UniGene/; Unigene database.
http://www.ncbi.nlm.nih.gov/omim/; Online Mendelian Inheritance In Man database.
http://www.ncbi.nlm.nih.gov/SNP/; single nucleotide polymorphism database.
http://www.gene.ucl.ac.uk/nomenclature; Human gene nomenclature.
http://sullivan.bu.edu/~mfrith/cister.shtml; repository and instructions for the Cister program.
http://www.gene-regulation.com/index.html; link to the TRANSFAC database.
http://ftp.genome.washington.edu/RM/RepeatMasker.html; RepeatMasker program.
http://globin.cse.psu.edu/; homepage for GALA link.
http://bio.cse.psu.edu/; homepage for GALA link.
http://genome.ucsc.edu/goldenPath/help/customTrack.html; instructions for custom tracks on the human genome browser.
http://java.sun.com/products/plugin; download Java Plug-in for use with *Laj*.

Received July 9, 2002; accepted in revised form January 24, 2003.