Genome Databases

Advanced article

Winston Hide, *South African National Bioinformatics Institute and University of the Western Cape, Bellville, South Africa*

A genome comprises all of the genetic material in the chromosomes of a particular organism. Genome databases are an organized collection of information that have resulted from the production or mapping of genome (sequence) or genome product (transcript, protein) information.

Article contents

doi: 10.1038/npg.els.0005314

Introduction

A genome is all of the genetic material in the chromosomes of a particular organism. Genome databases are an organized collection of information that have resulted from the production or mapping of genome (sequence) or genome product (transcript, protein) information. The process of making a genome database involves taking information that researchers have generated and organizing it into a database so that biological inferences can be made. Genome databases vary widely, closely reflecting the communities that they serve.

This article focuses on human and model organism databases, but there are several other systems including plant and microbial databases and genome product databases (transcript, protein and structure) that are not covered here. Underlying technologies tend to influence the functionality of a database system and thus have a significant role in delivering understanding of the underlying biology derived from a genome. Some of the technologies that are likely to influence the development of genome databases are described below.

The Value of Genomics is a Function of the Management of the Information it Generates

The rapid accumulation of genome data, including that of the human genome, has been a result of the implementation of high-throughput genome technology. It has meant that a significant new set of information has become available. The utility of that information is related directly to the quality and structure of its organization as a resource. Once organized, the value of the information can be improved markedly by integration and relative crossreferencing between and within biological information systems.

Genome databases and the integration of sequence information

Genome databases contain a variety of biological information. Before the increase in the rate of sequencing that has heralded the human genome, mapping and genetic locus data were the primary information that could be applied at a genome level. For the human genome the Genome Database (GDB), which was originally developed at Johns Hopkins University in Baltimore, Maryland, has been one official central repository for genomic mapping data in *Homo sapiens* since 1990. Although it now contains sequence information, this was not its original focus.

As the field of genomics has moved rapidly from basic genetics and mapping into the sequencing era, genome databases have tended to contain significantly greater proportions of genome sequence. The integration of sequence data with other genomic and biological information, particularly in the higher eukaryotes, has been central to the utility of genome databases. The aim of providing a genome sequence involves the ability to link, for example, a specific phenotype, publication, similar gene or phenotype in another species, disease, genetic locus, experiment or event to a particular molecular sequence. A genome database has the potential to realize this aim and thus provides a powerful link between the biology of an organism and its underlying genome sequence.

Sequence annotation

Sequence annotation is the association of biological information with a specific molecular sequence. The sequencing of any genome results in the production of large amounts of deoxyribonucleic acid (DNA) information. The meaning of this information can be determined only by the process of annotation. The sequence is linked with other genome sequences, usually by sequence comparison, and processed in the light of knowledge that has been determined before the sequencing of the genome in question. A genome database is therefore as good as the quality and implementation of knowledge that is associated with each part of the genome sequence that has been produced.

Model Organisms and Types of Genome Database

Many mapping and sequencing technologies have been developed from studies of nonhuman genomes, such as the bacterium *Escherichia coli*, the yeast *Saccharomyces cerevisiae*, the fruitfly *Drosophila melanogaster*, the roundworm *Caenorhabditis elegans* and the laboratory mouse *Mus musculus*. These experimentally and genetically mutable systems provide models for investigating the complex human genome. Prokaryotic genome projects have tended to be far more diverse in scope and number than have eukaryotic projects. The Genomes OnLine Database yields to date over 170 eukaryotic and 240 prokaryotic ongoing projects, of which at least 78 are published.

Genome databases can be described according to molecules (genome sequence, messenger ribonucleic acid (mRNA), proteins, mutated genes) or on the basis of the organisms that they describe. The complexity and number of such databases reflect the complexity of the systems that they support. (***See*** Genetic Databases; Genetic Databases: Mining.)

Model organisms have been a focus of genomics and therefore of genome databases. As biology is similar in many organisms, working on a model system allows inferences that would not normally be possible for an organism such as a human. Combining the results of genetic and experimental studies with genomic information is necessary to make a useful genome database. The nature of the scientific community that has developed a particular genome project is often reflected in the database assigned to manage that information. A community such as that for the fruitfly *D. melanogaster* or the nematode worm *C. elegans* may have historically worked relatively close together and so have developed a database for each system that has good connectivity between several aspects of genome and nongenome data.

A much larger community, such as that for human, has had to address the problem in a different manner because the community is less close knit, and ‘entry’ points for genome information are much more diverse, resulting in different databases covering different aspects of human genomic information. The principal databases are listed in **Table 1**.

Organism-specific databases

Human

The GDB is among the oldest human genome database systems, and its structure reflects the historic focus of the community on the genetics and mapping of the human genome. The close interaction of the GDB with the human genetics community means that its data have been structured to serve human geneticists.

Data for each gene can include publications, known disease phenotypes, a GeneCard link, a Human Gene Mutation Database and LocusLink, a National Center for Biotechnology Information (NCBI) LocusLink and also rudimentary expression data in terms of the UniGene Cluster code (see **Table 1**). The database is strong in terms of mapping information for each genetic locus, but a sequence-based view of each gene is not available.

Mouse

The Mouse Genome Database (MGD) contains information on mouse genetic markers, molecular segments, phenotypes, comparative mapping data, experimental mapping data and graphical displays for genetic, physical and cytogenetic maps. It was the first online resource for mouse genetic information. MGD is similar in concept to GDB (human) in that it has been focused on support of the mouse research community.

Developed before the advent of large-scale genome sequencing, the MGD also has a strong bias toward nonsequence data. The curators at the Jackson Laboratories in Bar Harbor, Maine, have kept up with developments in genomics to embrace the new data, and also to continue developing pertinent informatics solutions to support the genomics environment. MGD reports contain genes, alleles and phenotypes, molecular probes and segments, mammalian homology and comparative maps, gene expression, strains and polymorphisms, references, accession codes and chromosome committee reports. MGD is a member of the Gene Ontology Consortium (see below).

Roundworm

WormBase is the repository of mapping, sequencing and phenotypic information on the *C. elegans* nematode. The *C. elegans* sequencing project represents one of the earliest efforts to sequence an animal eukaryote. The development of eukaryotic genome databases was spearheaded by this project. AceDB development (see

Table 1 Human Genome Databases

	URL	Notes
Whole-genome databases		
GenomeWeb: Human Genome Resources	http://www.hgmp.mrc.ac.uk/GenomeWeb/human-gen-db-genome.html http://www.hgmp.mrc.ac.uk/GenomeWeb/genome-db.html	A good web page for details of various human and other genome data sources
The Genome Database	http://www.gdb.org/	Database that is not 'sequence-centric'
Online Mendelian Inheritance in Man	http://www3.ncbi.nlm.nih.gov/Omim/	Database that is not 'sequence-centric'
Human Chromosome-Specific WWW servers	http://www.gdb.org/gdb/hgpResources.html#CHROMOSOMES	List of databases for individual human chromosomes
Ensembl	http://www.ensembl.org/	Integrated whole human genome information at the sequence level
Ensembl: Human Genome Central	http://www.ensembl.org/genome/central/	
UCSC Human Genome Browser, commonly known as 'Golden Path Assembly'	http://genome.cse.ucsc.edu/goldenPath/hgTracks.html	Integrated whole human genome information at the sequence level
The Human Genome Guide to Online Information Resources	http://www.ncbi.nlm.nih.gov/genome/guide/human/	Integrated whole human genome information at the sequence level
The Unified Database for Human Genome Mapping	http://bioinformatics.weizmann.ac.il/udb/	Integrated map for each human chromosome includes physical data with links to various databases including GeneCards
Genome Channel	http://compbio.ornl.gov/channel/	Unified query interface for multiple genomes, emphasis on genomes in which US Department of Energy has participated
Single-gene databases		
GeneCards^{TM}	http://bioinformatics.weizmann.ac.il/cards/	Associates human genes, products and diseases
Genatlas	http://www.dsi.univ-paris5.fr/genatlas/	Integrates a growing list of genes, pathways, proteins, diseases and characteristics
RefSeq	http://www.ncbi.nlm.nih.gov:80/LocusLink/refseq.html	An attempt to provide standard reference sequences for analysis
Databases that curate transcript information		
UniGene	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db = unigene	Nonredundant set of gene-oriented clusters with little manual curation. No consensus sequences
TIGR Human Gene Index	http://www.tigr.org/tdb/hgi/index.html	Nonredundant set of transcript-oriented clusters, different approach to Unigene, little manual curation, short consensus sequences
STACKdb^{TM}	http://www.sanbi.ac.za/Dbases.html	Nonredundant set of gene-oriented transcripts containing alternative spliceforms and consensus sequences. Segregated according to tissues or parent mRNA index
BODYMAP	http://bodymap.ims.u-tokyo.ac.jp/	Mouse and human gene expression database, with strong orientation towards transcripts and expression
Human Gene Mutation Database	http://www.hgmd.org	Curated dataset of human gene mutations. Cannot be easily downloaded
LocusLink	http://www.ncbi.nlm.nih.gov/LocusLink/	Provides an integrated query interface to curated sequence and descriptive information about genetic loci

below) has resulted in a stable platform for exploration of methods to provide genome-linked information. Its maturity is reflected in the sophistication of the genome database system, which now supports a completed genome.

WormBase incorporates a large variety of data, including views of the complete sequence, genes, associated single nucleotide polymorphisms, genes of similar expression profile and RNA interference experiments (**Figure 1**). The system has been developed with close integration to the Distributed Annotation Server (DAS) protocol (see below) and as such represents a flagship for future powerful genome database efforts.

Fruitfly

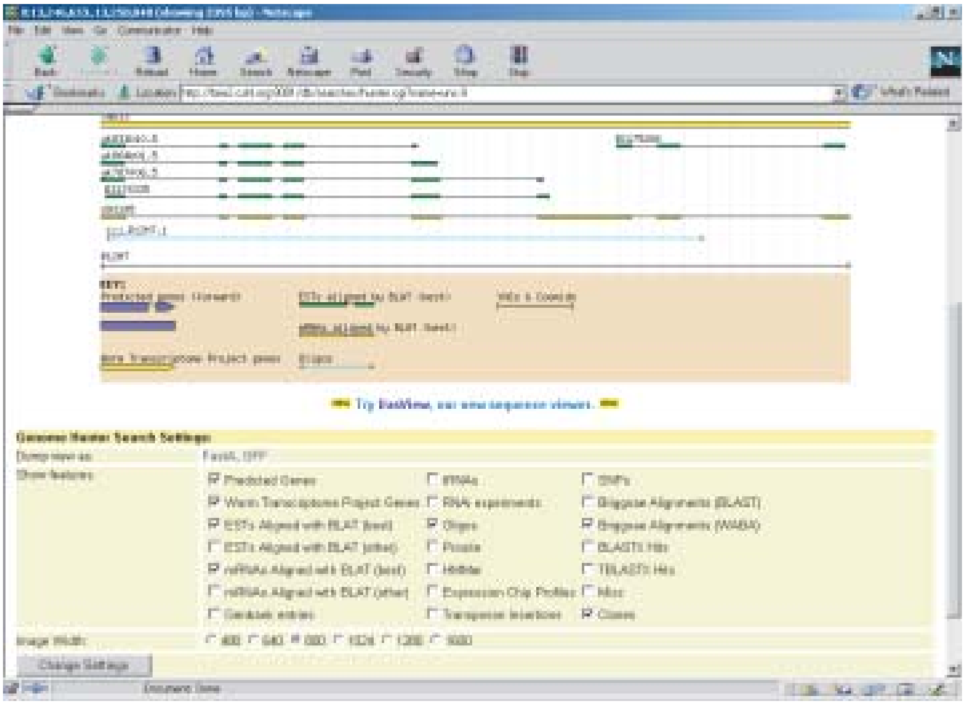
FlyBase is a highly comprehensive database for information on the genetics and molecular biology of *Drosophila*. It includes data from the Drosophila Genome Projects and data curated from the literature (Ashburner and Drysdale, 1994). FlyBase is a very well-integrated genome resource. Among its several features are the ability for users to access cytologic maps, annotated genome, genes, alleles, gene products, genome annotation, protein function, location, process, structure, gene expression, sequences, search genomic sequences and clones, search and order expressed sequence tag (EST) project complementary DNAs, order stocks, browse natural transposons, view anatomy and images, find references and access addresses of people in the community.

The system has been built by the research community of *Drospohila*. Software development for FlyBase has gone hand in hand with experimental development. On completion of the fly genome, annotation was carried out using the concept of a 'jamboree', in which domain experts were shut in a room with software developers to ensure expert curation of the genome. The database reflects the integration of the community that it represents.

Yeast

The Saccharomyces Genome Database (SGD) project collects information and maintains a database of the molecular biology of the yeast *S. cerevisiae*. As one of the oldest eukaryotic genome projects, the SGD contains perhaps some of the most integrated and informative information. It uses the AceDB environment. Once a user has overcome the steep learning curve of the AceDB database structure, it becomes a resource made more powerful by its standardization between organism databases.

The SGD is very powerful in that it holds a completed genome with very well-characterized genes that bear formidable genetic and experimental work. For each gene, the database contains the standard collection of information that is common to other genome database. The SGD also contains the most



Try DasView, our new sequence viewer.

Genome Hunter Search Settings:

Dump view as: FastA, GFF

Show features:

☑ Predicted Genes □ tRNAs □ SNPs
☑ Worm Transcriptome Project Genes □ RNAi experiments □ Briggsae Alignments (BLAST)
☑ ESTs Aligned with BLAT (best) ☑ Oligos ☑ Briggsae Alignments (WABA)
□ ESTs Aligned with BLAT (other) □ Protein □ BLASTX Hits
☑ mRNAs Aligned with BLAT (best) □ Motifs □ TBLASTX Hits
□ mRNAs Aligned with BLAT (other) □ Expression Chip Profiles □ Misc
□ GenBank entries □ Transposon Insertions ☑ Clones

Image Width: ○ 400 ○ 640 ● 800 ○ 1024 ○ 1280 ○ 1600

Change Settings

Figure 1 Genome view of WormBase showing user-selected annotations for gene *unc-9*. Annotation tracks tend to become added with time, allowing the user to refine the evidence and quality of annotation for a particular region of interest.

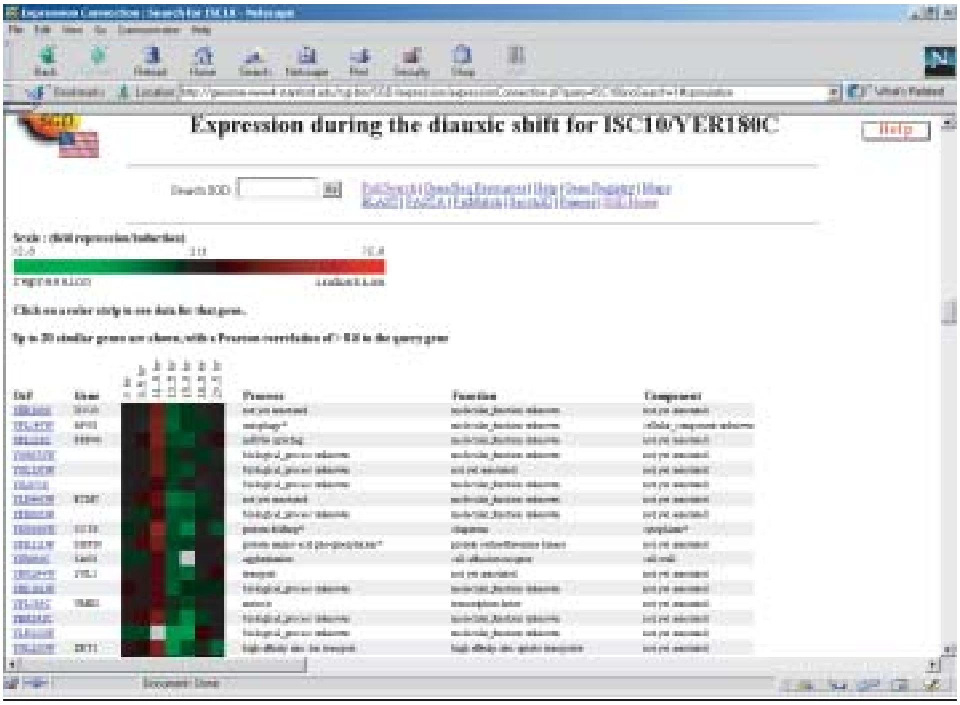


Figure 2 Array gene expression profile for a gene in Sacchromyces Genome Database. Open reading frames (left column) are compared with experiments (top row) and information on relative level of expression (color) and Gene Ontology is presented.

available array expression information that has been consolidated for a genome (**Figure 2**). Genes can be queried electronically by gene name to yield an electronic expression profile that is based on numerous expression array experiments.

Genome Database Systems

Genome databases vary in size, connectivity and complexity as a function of the degree of funding that they command. NCBI and Ensembl, which is funded by the Wellcome Trust, are the largest systems and as such provide the most comprehensive resources. Both are 'sequence-centric' systems. Each 'build' of the human genome assembly is manufactured separately at NCBI and the Ensembl/GoldenPath sites. The systems have the capacity to support several species of genomes and have developed tools and systems for analysis of human, mouse and several other organisms.

NCBI genome resources

> NCBI's website serves an integrated, one-stop, genomic information infrastructure for biomedical researchers. (NCBI Website)

The NCBI system is one of the oldest of the sequence-based resources in that it combines existing information systems onto the new genome assembly information in an integrated manner. However, the genome-centric viewing and integration system itself is very new and still in early development. The NCBI system allows the viewing of genome fragments and also the investigation of human genome sequence from an evidence-based gene-centric viewpoint (**Figure 3**). Links exist between genes in the database and the genome sequence from which they originate.

The entry page shows graphics of all of the chromosomes and allows a search of the data in all of the maps available for that organism. Sequence, cytogenetic, genetic, radiation hybrid and others are included. Terms that can be searched include gene symbol, gene name, marker name, aliases for marker name and text word (e.g. actin) or phrase (e.g. cell adhesion). The linkage between genome views and the richly integrated PubMed/GenBank Entrez database is still very much in development. In the longer term, however, the NCBI genome-based resource is likely to become more powerful.

Ensembl

> Ensembl is a joint project between EMBL–EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on eukaryotic genomes. (Ensembl Website)

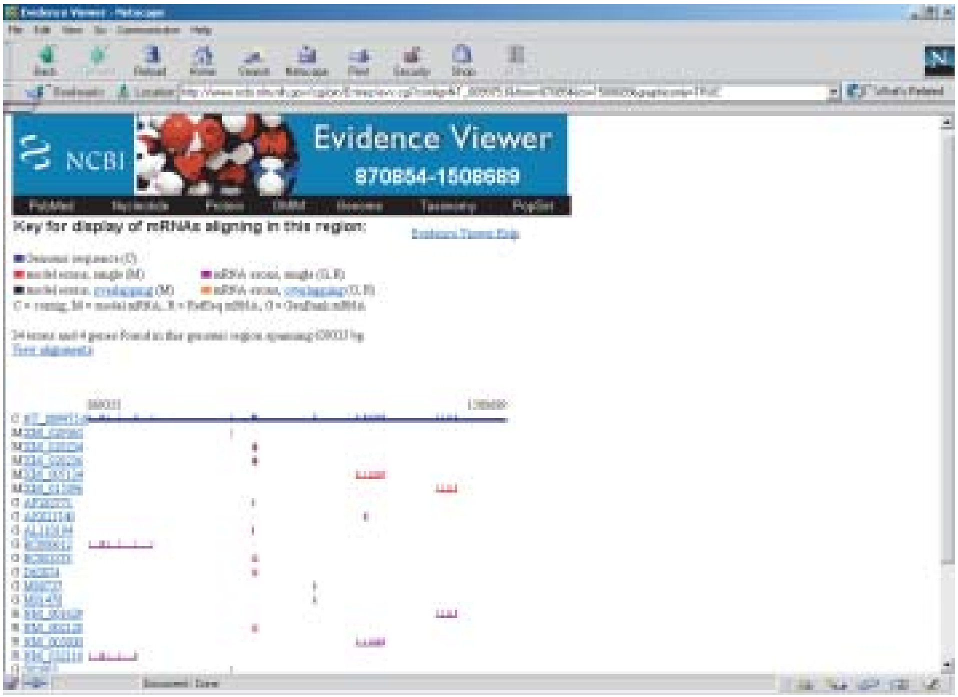


Figure 3 Web-based evidence view at NCBI for genes in a region of the human genome. Use of standardized accessions allows tight integration with other data related to the genome. The map is clickable to allow for links and access to further annotations.

The Ensembl system provides identification of most of the known human genes in the genome sequence, prediction of additional genes, supporting evidence and connections to other resources worldwide using many public genomic databases and tools. The system relies heavily on automated predictions and unsupervised inclusion of supporting evidence, and as such is a guide to the structure and relationships of genes in the genome. From its inception, however, it has been possible to view regions of the genome to include information such as radiation hybrid markers, transcript evidence, gene predictions, exon boundaries, protein products and intron–exon structure of entries (**Figure 4**).

Ensembl design has far less 'legacy' than other systems and, as such, it reflects far easier navigation, accessibility and more powerful integration than the other established databases of human and other systems. It maintains its own accession system but links these accessions to known accessions such as GenBank accession codes and Human Genome Nomenclature Database (HUGO) nomenclature. A gene report links a single page to actual genome views, evidence to support the report, and presumably other information that is pertinent to each entry. Several powerful features are built into the Ensembl system and its development has been rapid. It represents a good starting point for genome viewing and gene structure analysis.

Human Genome Project Working Draft

Golden Path Assembly, located at the University of Santa Cruz, California (UCSC), is the most sequence-centric of the human genome data systems. The system has a powerful feature termed 'tracks' that allow users to submit annotation tracks to the genome at the sequence level (**Figure 5**). The site contains a working draft of the human genome in its most up-to-date state. Because users can add 'tracks', this site has the newest forms of annotation, for example, 'high-resolution haplotypes'. Features include annotation of repeat sequences, transcripts mapped to the genome, Ensembl gene predictions, spliced ESTs, random single nucleotide polymorphisms (SNPs), sequence-tagged site (STS) markers and human mRNAs from GenBank.

Genome Database Technologies

Genome information reflects the complexity of the biological system that it represents. No preexisting technology had a suitable architecture with which to address the construction of databases for biological genome information. Scientists have therefore had to develop systems to organize and to analyze genome information in order to derive biological meaning. The development of genome database systems has

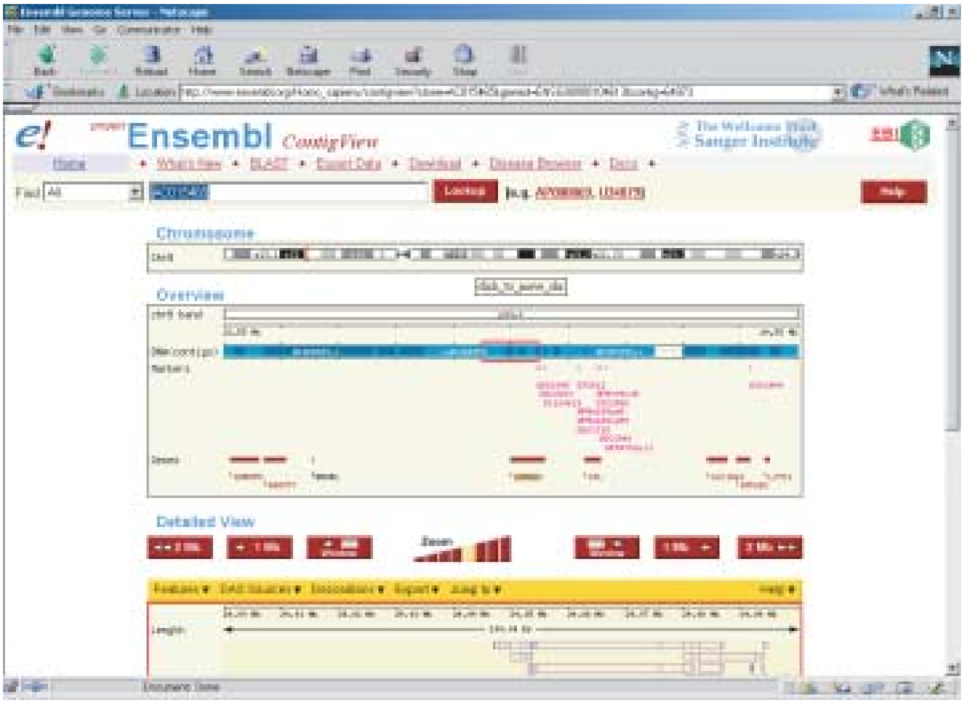


Figure 4 Contig view of human lipoprotein lipase using the Ensembl web viewer. The user can alter the view by zooming, can examine evidence using DAS sources and can link to Ensembl accessions by clicking on the map.

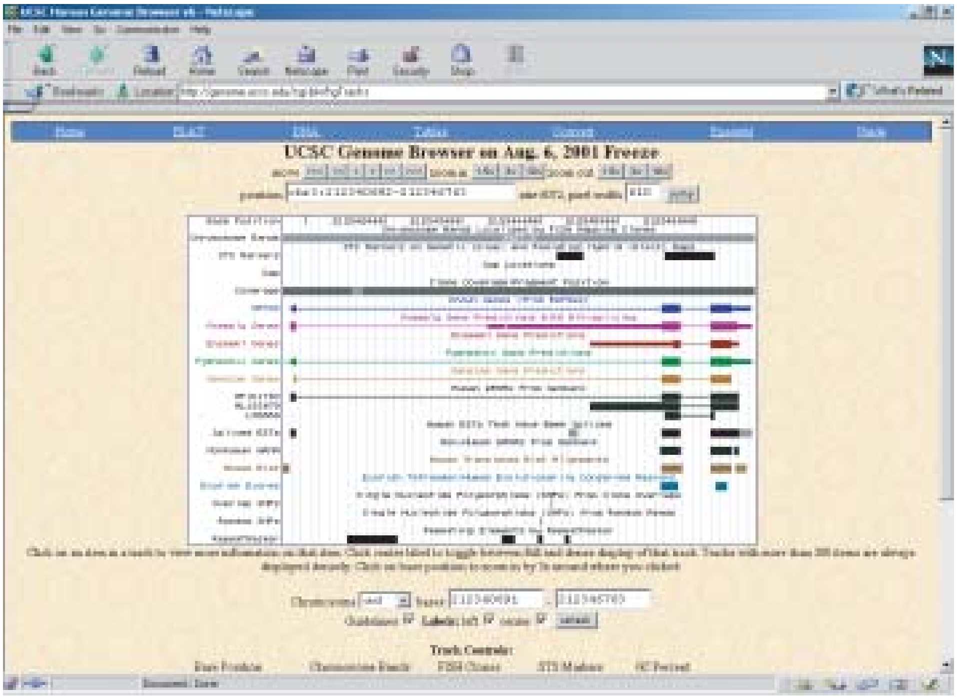


Figure 5 UCSC Genome Browser view showing selected tracks. The UCSC system was the first to allow user-mapped data to
be included with data served by the Genome Browser. Tracks of mapped information (e.g. equivalent fish exons) can be selected using tick boxes below the graphic.

reflected the availability of existing technologies such as relational databases, the World Wide Web and Internet, and scripting languages such as Python and Perl. The setting up of genome databases has also resulted in the development of new systems and languages. The technologies have had to address some consistent problems of database development. Examples of problems include compact, complex and highly interrelated data, and how to let the community interact with information in the database and provide further value to it without degrading the quality of that information.

AceDB

The *Caenorhabditis* community provides an early example of a specific software system that was developed specifically for a particular genome project. AceDB (A *C. elegans* DataBase) is a genome database system that has been developed since 1989 primarily by Jean Thierry-Mieg and Richard Durbin. This system is unique in that it is a software system designed specifically for genome databases and is freely available and distributed. Thus, several genomes are now resident in AceDB systems worldwide. But this system by itself has little value without the data that it contains. AceDB offers the ability to incorporate any form of data and to lay it with context to the genome.

A large user community has grown up around AceDB, which also has a large developer base. It is not a modern system but because of its broad user base it has become a *de facto* standard for a proportion of the genome community. An advantage of the system is that changes to the local copy of information in the database can be sent to a central organizer for redistribution.

DAS

A powerful new approach to the problem of distributing annotation was developed in 2001 (Dowell *et al.*, 2001). DAS is a client–server system, which allows a single client to integrate information from several servers. Implemented as a protocol specification on top of the web page protocol HTTP, the DAS protocol specifies a small number of low-level commands with limited intrinsic semantic content. DAS first allows diverse sources (DAS 'annotation' servers) to overlay annotations on a reference sequence map (e.g. Ensembl, DAS reference server) and then allows these annotations to be seamlessly (but optionally) recruited into the genome view, which is accessed by a DAS client. Ensembl allows DAS sources to be specified for overlay annotation of the reference sequence. Two main annotation servers are supplying DAS-formatted information: the Sanger Institute, which serves gene SNPs and various repeats; and The Institute of Genomic Research (TIGR), which serves tentative human consensus (THC) alignments.

The value of DAS lies in managing the distributed annotation of the genome. Any group with a DAS server can serve up information to the genome community and annotate a genome with respect to their own view of the genome. This provides a solution for the perennial problem of community involvement in genome annotation, as the user can choose to believe the creator of the DAS-served information.

Gene Ontology Consortium

Several lessons have arisen from the development of eukaryotic genome databases such as FlyBase. Perhaps the most important lesson has been that very close communication between the wet bench researcher, the genome scientists and the bioinformatics specialists is essential. Close communication between organism-based data systems also has strong benefits.

Out of close communication has arisen the concept of the 'gene ontology' (Ashburner *et al.*, 2000). Many of the genes that specify core biological functions are shared by eukaryotes that are represented by model organisms. Knowledge of the biological role of such shared proteins in one organism can often be transferred to other organisms. The Gene Ontology Consortium has attempted to provide a dynamic, strictly controlled vocabulary that can be applied to all eukaryotes. This vocabulary addresses biological process, molecular function and cellular component. Model organism databases all take part in the Gene Ontology Consortium and so benefit from shared expertise and vocabularies.

Open Source

Definitions of open source vary widely but represent a model for software and data availability that in loose terms means that all software and data resulting from its use are licensed for open public access and development. Although most of the principal genome databases embrace public distribution and access to source codes of software, each has taken its own track through the open source model.

Software developed at the Ensembl site, and also at several other sites in genome databases, is publicly available. The main difference between the Ensembl system and all others currently in development is that it has been conceived and executed totally in an 'open source' environment using open source tools that are applied commonly across many technologies. It includes the public access to development mailing lists. An open source development methodology, together with the significant funding that supports the project, ensures its long-term viability and broad availability and distribution. NCBI and AceDB use their own open source technologies that have been developed specifically for genome databases.

As the Ensembl team has several different projects undergoing integrated development, associated software is now being written that can integrate back to the genome for human and other genomes. Software examples of available genome annotation viewers developed at the Sanger Institute are Apollo and Artemis (Rutherford *et al.*, 2000). Apollo is a genomic annotation viewer and editor developed for eukaryotic work in a collaboration between the Berkeley Drosophila Genome Project and the Sanger Institute. It provides access to newer software and integration features for the genome projects.

See also

Caenorhabditis elegans Genome Project
Genetic Databases
Human Genome: Draft Sequence

References

Ashburner M, Ball CA, Blake JA, *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25**(1): 25–29.

Ashburner M and Drysdale R (1994) FlyBase – the *Drosophila* genetic database. *Development* **120**: 2077–2079.

Dowell RD, Jokerst RM, Day A, Eddy SR and Stein L (2001) The distributed annotation system. *BioMedCentral Bioinformatics* **2**(1): 7.

Rutherford K, Parkhill J, Crook J, *et al.* (2000) Artemis: sequence visualization and annotation. *Bioinformatics* **16**(10): 944–945.

Further Reading

Date CJ (1990) *An Introduction to Database Systems*, 5th edn. Reading, Boston, MA: Addison-Wesley.

Etzold T, Ulyanov A and Argos P (1996) SRS: information retrieval system for molecular biology data banks. *Methods in Enzymology* **266**: 114–128.

Goffeau A (1997) The yeast genome directory. *Nature* **387**(6632 supplement): 5.

Letovsky S (ed.) (1999) *Bioinformatics: Databases and Systems.* Boston, MA: Kluwer.

Ringwald M, Baldock R, Bard J, *et al.* (1994) A database for mouse development. *Science* **265**(5181): 2033–2034.

Each January issue of *Nucleic Acids Research* is a special database issue.

Web Links

AceDB. Comprehensive information about the AceDB system, its community, tools for use, quick guide and downloads
http://www.acedb.org/

Apollo Gene Annotation Tool. User documentation and downloads for multiple platforms for the Apollo gene annotation tool
http://www.ensembl.org/apollo/apolloguide.html

BioMedCentral.
http://www.biomedcentral.com/1471-2105/2/7

Distributed Annotation System (DAS). Covers latest developments in the Distributed Annotation System community, tools, downloads and developer information
http://www.biodas.org

Ensembl. Central site for the Ensembl genome browsers, helpdesk access, latest announcements and links to data-mining tools for the Ensembl system
http://www.ensembl.org/

Entrez. Entry point to the NCBI retrieval system for searching a growing list of linked databases that include publications, sequences, online texts, diseases and genomes
http://www.ncbi.nlm.nih.gov/Entrez/

Gene Ontology Consortium. Home for the Gene Ontology Consortium. Provides comprehensive information on Gene Ontology projects, downloads, developer tools, publications and links
http://www.geneontology.org

GOLD: Genomes OnLine Database. Resource for comprehensive access to information regarding complete and ongoing genome projects around the world
http://wit.integratedgenomics.com/GOLD/

The Genome Database (GDB). Entry portal to The Genome Database (human)
http://gdbwww.gdb.org/

Human Gene Nomenclature Database (HUGO). Search engine for approved human gene symbols
http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl

Human Gene Mutation Database. Entry point for access to curated information on human gene mutations
http://www.hgmd.org

Mouse Genome Informatics. Entry point for integrated access to data on the genetics, genomics, and biology of the laboratory mouse
http://www.informatics.jax.org/

National Center for Biotechnology Information (NCBI). Homepage for the National Center for Biotechnology Information in the USA. Provides current information on projects, releases and links to the major NCBI projects
http://www.ncbi.nlm.nih.gov/

Open Source. Homepage of the Open Source Initiative. Definitions, licenses and certifications of open source efforts, together with explanations on the nature of open source
http://www.opensource.org

Saccharomyces Genome Database (SGD). Comprehensive entry point for the database of the molecular biology and genetics of the yeast *S. cerevisiae*
http://genome-www.stanford.edu/Saccharomyces/

The Institute for Genomic Research (TIGR). Homepage of the Institute for Genomics Research. Includes links to activities of the institute
http://www.tigr.org/

WormBase. Homepages of the AceDB site for the genome and biology of *C. elegans*. Comprehensive entry point
http://www.wormbase.org