

The UCSC genome browser and associated tools

Robert M. Kuhn, David Haussler and W. James Kent

Submitted: 8th February 2012; Received (in revised form): 9th June 2012

Abstract

The UCSC Genome Browser (<http://genome.ucsc.edu>) is a graphical viewer for genomic data now in its 13th year. Since the early days of the Human Genome Project, it has presented an integrated view of genomic data of many kinds. Now home to assemblies for 58 organisms, the Browser presents visualization of annotations mapped to genomic coordinates. The ability to juxtapose annotations of many types facilitates inquiry-driven data mining. Gene predictions, mRNA alignments, epigenomic data from the ENCODE project, conservation scores from vertebrate whole-genome alignments and variation data may be viewed at any scale from a single base to an entire chromosome. The Browser also includes many other widely used tools, including BLAT, which is useful for alignments from high-throughput sequencing experiments. Private data uploaded as Custom Tracks and Data Hubs in many formats may be displayed alongside the rich compendium of precomputed data in the UCSC database. The Table Browser is a full-featured graphical interface, which allows querying, filtering and intersection of data tables. The Saved Session feature allows users to store and share customized views, enhancing the utility of the system for organizing multiple trains of thought. Binary Alignment/Map (BAM), Variant Call Format and the Personal Genome Single Nucleotide Polymorphisms (SNPs) data formats are useful for visualizing a large sequencing experiment (whole-genome or whole-exome), where the differences between the data set and the reference assembly may be displayed graphically. Support for high-throughput sequencing extends to compact, indexed data formats, such as BAM, bigBed and bigWig, allowing rapid visualization of large datasets from RNA-seq and ChIP-seq experiments via local hosting.

Keywords: UCSC genome browser; bioinformatics; genetics; human genome; genomics; sequencing

INTRODUCTION

The UCSC Genome Browser (<http://genome.ucsc.edu>) [1–5] was created to provide a graphical viewpoint on the very large amount of genomic sequence produced by the Human Genome Project. As the size and coverage of the human genome assembly increased, the focus turned from accumulating and assembling sequence to identifying and mapping functional landmarks, such as known genetic markers, genes, points of regulation and cytological features such as centromeres and telomeres. The UCSC Genome Browser and other visualization tools, such as Ensembl [6] and GBrowse [7] are

widely used by researchers for viewing these genomic annotations.

The client–server model employed by the UCSC Genome Browser has the advantage to the user of offering access to a very large database of information in a uniform interface with no overhead of importing datasets. This model is employed by Ensembl as well, as both projects host a large number of organisms. The databases using the GBrowse software tend to be distributed, with only one or a few organisms at any given site. These model organism databases benefit from the ease with which GBrowse can be set up for a new organism.

Corresponding author: Robert M. Kuhn, Center for Biomolecular Science and Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA, Email: kuhn@soe.ucsc.edu

Robert M. Kuhn, PhD, is Associate Director of the Genome Browser project at the University of California, Santa Cruz, where he has contributed to the Genome Browser since 2003.

David Haussler, PhD, is Investigator at Howard Hughes Medical Institute and Distinguished Professor of Biomolecular Engineering at the University of California, Santa Cruz

W. James Kent, PhD, is Director of the Genome Browser project at the University of California, Santa Cruz and wrote the original Genome Browser code.

Server-side systems, such as the Integrative Genomics Viewer (IGV) [8], use a client-side model that runs a Java application on the user's machine. Not all system administrators will allow users to install Java applications within a firewall. IGV has the advantage of being much faster than web-based browsers, but also does not have the extensive third-party data of the other browsers. IGV has specialized tools for high-throughput sequencing (HTS) not found in the web-based browsers.

The basic paradigm of the UCSC Genome Browser is to show as much high quality, whole-genome annotation data as possible and enable researchers to use their expertise to interpret data themselves. The Genome Browser displays a wide variety of data juxtaposed onto a common coordinate system, providing a high degree of control over what is viewed and how it is configured. Because researchers are frequently confronted with 'fuzzy' or conflicting data, the Genome Browser does not attempt to adjudicate between conflicting views or datasets, such as gene models, for example. Rather than substitute our judgment, we provide the data viewer that presents multiple annotation datasets

side-by-side, allowing expert users to reach their own conclusions (Figure 1).

It is important that users assess multiple lines of evidence before committing to a course of action. For example, are the various gene models (e.g. UCSC, Ensembl, RefSeq, CCDS) in agreement? Is there physical evidence: mRNAs? ESTs? Gencode? ENCODE RNA-seq? The Genome Browser provides quick access to multiple lines of evidence.

The Genome Browser is designed to allow viewing of data at any scale, from single basepair resolution to a whole chromosome, and to fit as much information into the Browser view as the scale permits. Any set of data that can be mapped to genomic coordinates can be presented in the Genome Browser, which is designed to be a gateway to more detailed information. Typically, each item in the Browser view is a live link to a 'details page' containing item-specific information and further links to the original source of the data.

Although the UCSC Genome Browser originated as part of the Human Genome Project, and some 80% of the web traffic on the site is still on the human genome assemblies, a large number of other vertebrate genome assemblies and selected model

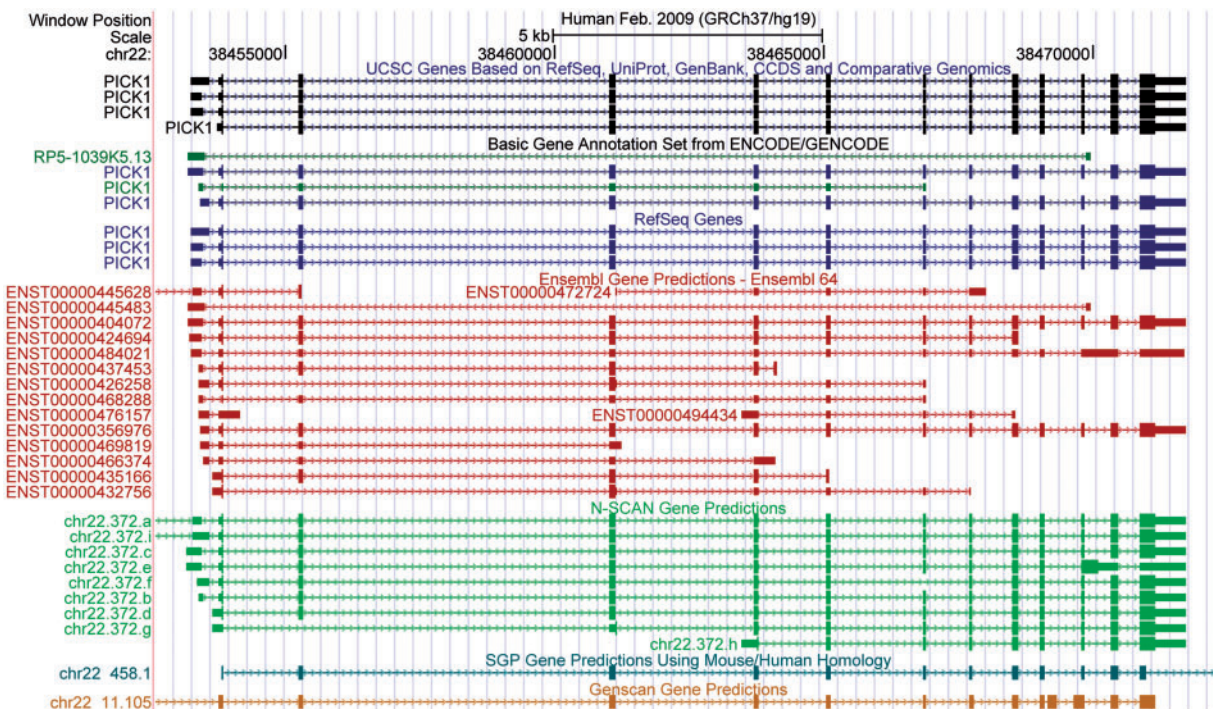


Figure 1: Screenshot of UCSC Genome Browser displaying human PICK1 gene region on chr22 in hg19 assembly. Different gene prediction algorithms predict different annotations in the region. By presenting multiple data sets of similar type, the user is able to more easily evaluate hypotheses. The different tracks often predict different 3'- and 5'-untranslated regions (half-height boxes on ends of annotations), coding regions (full-height boxes), introns (thin line with transcription-direction arrows) or start and end coordinates. The differences may be used to establish a level of confidence in an annotation not obtained from any single method.

invertebrate genomes are also displayed. The same software is used to display the data on every genome assembly, and the interface and navigation of the site are consistent throughout. Much of the data are from external collaborators (including model organism databases) and the Genome Browser links to these other resources. Many users find it convenient to navigate to a gene of interest in the Genome Browser, then use the Browser's direct links to the relevant pages on other sites rather than learn the navigation on multiple websites.

The database underlying the Genome Browser is available for bulk download (see discussion below). All data are freely available to the public, though the original contributors of some data retain a copyright.

While recent publications [2–5] describe the latest advances in the Genome Browser, this article describes an historical context and technical details not previously published. Users of the Genome Browser should gain useful background on the rationale and implementation of the Browser. It should also prove useful for the increasing number of research groups who mirror the UCSC software in local installations to facilitate manipulation and display of sensitive data or of data too voluminous to transmit over the network.

CONTEXT/HISTORICAL BACKGROUND TO UCSC GENOME BROWSER

Jim Kent and David Haussler at the University of California, Santa Cruz played a significant role in the first release of a draft human genome sequence in 2000 [9, 10], which became available from UCSC by bulk download at that time. Several billion bases of DNA in a text file are difficult to interpret, however, and specialized visualization and interpretation tools are required.

Both NCBI [11] and the Ensembl project [12] developed visualization tools at the beginning of the Human Genome Project and others, such as GBrowse [7] appeared soon thereafter. It was a desire to show splicing in *Caenorhabditis elegans* RNAs that formed the basis for what was ultimately to become the UCSC Genome Browser. A computer program called the Intronerator [13] was developed by Jim Kent, which showed the exon–intron structure of *C. elegans* RNAs mapped against genomic coordinates. That basic coordinate-based paradigm persists in the Genome Browser

today. By late 2000, the drawing functions of the Intronerator were incorporated into a web-based computer program, which is today the main graphical viewer of the Genome Browser. This viewer serves as an interface between users and the ever-growing underlying database of genomic data.

Through the years, the project added browsers on other assemblies, with the first mouse and rat browsers being added in 2002 (Figure 2). [Note that throughout this article, the word 'browser' with lower case 'b' is used to refer to either a web client, such as Firefox or Safari or to software and database for viewing a particular genome assembly (such as 'built a browser on the latest dog assembly'). 'Browser' with a capital 'B' refers to the entire Genome Browser website or the Genome Browser graphical viewer.] As the sequencing centers and reference consortia continue to close gaps in genomic sequence, troubleshoot assembly problems and otherwise improve the genome assemblies, these continuing updates in the assemblies have led to shifting coordinates for known sequences and a potential for confusion and error among researchers, particularly when reading literature based on older versions. Because many laboratories invest resources in analyses based on an assembly that may be updated while their work is continuing, UCSC maintains older versions of genome assemblies to ensure continuity for these users. For example, as of this writing, there are four assemblies available on the main public site for both the human reference sequence: hg16 through hg19 (NCBI34 through GRCh37), and for mouse: mm7 through mm10 (NCBI 35 through GRCm38). Still older assemblies are available via the archive link on the main index page, thus providing a history of the genome sequence not available elsewhere. Software tools to convert coordinates from one assembly to another (liftOver) are also provided.

As the sequence assembly of these major organisms has matured, the frequency of release of new assemblies has slowed. However, the release of assemblies for new organisms has increased, and we have tightened the criteria for offering a browser on a new assembly. Typically, we now build a browser only if an assembly has reached 6× genomic coverage and has been deposited in GenBank. Exceptions may be made to represent key nodes in the evolutionary tree or to support inclusion in a multiple alignment. As of April 2012, 127 genome assemblies have been released on 58 organisms (mostly vertebrates), 104

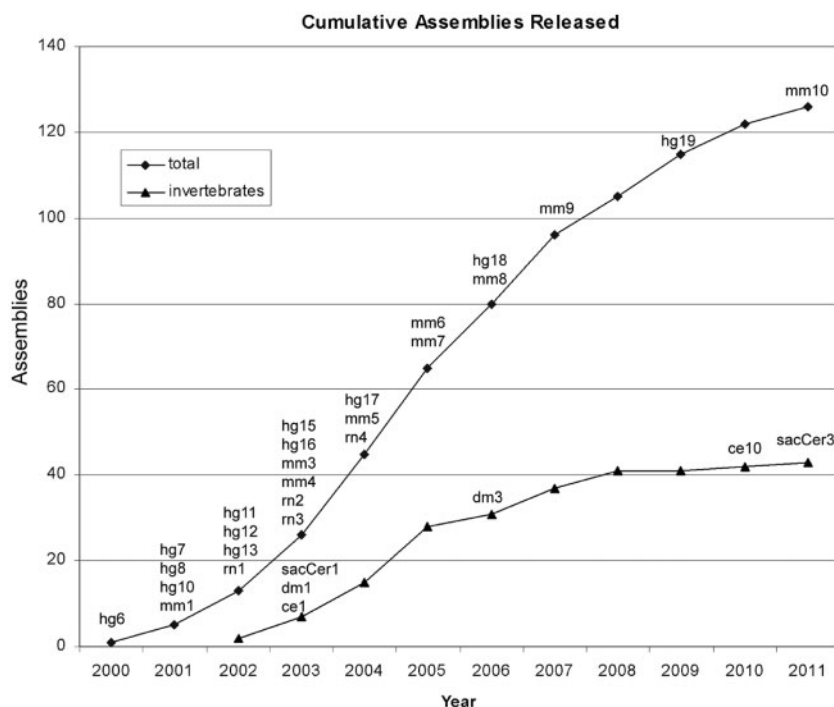


Figure 2: Cumulative number of genome assemblies released in UCSC Genome Browser, showing release dates for key genome assemblies. The steady increase reflects fewer updates of mature assemblies, such as human and mouse, balanced by the increase in new species being completed. Total genome assemblies are marked with all releases of human (hg*), mouse (mm*) and rat (rn*) assemblies. Invertebrate assemblies are marked with first and latest release dates for key model organisms, yeast (sacCer*), *Drosophila melanogaster* (dm*) and *Caenorhabditis elegans* (ce*).

of which are still available on the main public site. Important invertebrate model organism species are also supported (Figure 2).

The Genome Browser software has grown considerably over the years. While the basic Genome Browser interface is a simple coordinate-based map, a variety of supporting tools have grown up around it, with the addition of many new features, often at the request of the user community. The Genome Browser now comprises 35 different programs, many of which are not obvious to the casual user, that communicate between the user and the database. The UCSC Browser offers a fast, multifunctional experience, serving up to 8 million page requests per week to 185 000 unique IP addresses per month.

The footprint of the data underlying the Genome Browser currently amounts to 7.7 Terabytes (Tb) of data in flat files (primarily sequence) and 3.3 Tb of tables in a MySQL database (annotations). In March 2012, interactive or program-driven data usage amounted to 2.2 Tb of data, half of which was served via the Table Browser. Bulk downloads

accounted for an additional 73 Tb of data transmitted during that same period. Usage statistics show that the UCSC Genome Browser and the associated tools are a global resource, with 44% of the traffic from the United States, 36% from Europe (led by the UK, Germany and the Netherlands) and 8% from Asia.

METHODS

Database structure

New assemblies

When a new assembly of genomic sequence is announced, UCSC retrieves the sequence as a fasta file from NCBI along with an AGP file (**A Golden Path**) that describes the sequences and gaps comprising the assembly. The sequence is then typically converted into a compressed format (a .2bit file—each base occupies two bits of memory instead of an entire byte as in a fasta file), which is actively accessed by the Browser when needed.

The construction of a basic genome browser on a new assembly begins with assigning a name. Because assemblies are given a variety of names by the different

sequencing centers and no standard nomenclature exists across all organisms, the UCSC database has standardized on a naming system that is internally consistent, though it has evolved. The first assembly for an organism is given a name in the format *gggSss#* using the first three characters of the genus and species names, with subsequent assemblies incrementing the number. For example, the cow, *Bos taurus*, currently has assemblies *bosTau2* through *bosTau6* on the public site (*bosTau1* having been archived).

Two other types of names continue to be used for organisms whose initial assemblies predate the introduction in 2003 of the six-letter naming scheme. Human assemblies are called *hg###*, because the database and Genome Browser began when UCSC was generating the human assemblies in-house and the **h**uman **g**enome was the only sequence represented. The next several organisms to have browsers were given two-letter names in the format, *gs#*, based on scientific names, until the two-character namespace proved inadequate; mouse (*Mus musculus*) and rat (*Rattus norvegicus*) are *mm#* and *rn#*, respectively.

Data for each organism are stored in a separate database in the MySQL database system, which is built in a modular design. Each genome assembly has a database named with the genome name as described above (*hg#*, *gs#* or *gggSss#*). This database contains all the assembly-specific tables needed to create a display in the browser graphic, including individual tables for annotation datasets and several tables of metadata describing display parameters, configuration options, etc., specific to the assembly (Figure 3).

Several additional databases contain information used by more than one assembly (e.g. data needed to show all the organisms in a pulldown menu). Thus, the *hgcentral* database contains metadata about the assemblies, including genome name, scientific name, location of the .2 bit file, official assembly name and date and other information. Similarly, the *hgFixed* database (not shown in Figure 3) contains global information, such as restriction enzyme recognition sites.

Adding a new assembly

With the large number of organisms now being sequenced and the increasing size of HTS annotation datasets, the UCSC site necessarily becomes more selective in deciding which assemblies and tracks to host. In response to this, the Browser code has been made more portable, to enable others to use the code for other organisms or to mirror the content of the

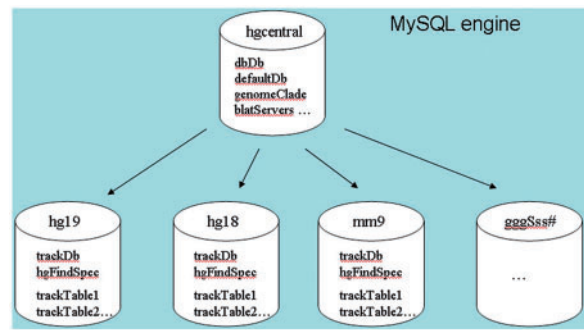


Figure 3: Summary database schema for UCSC Genome Browser. The *hgcentral* database contains tables with metadata about entire genome assemblies. For example, the *dbDb* table has one row for each assembly, specifying assembly name, source, display options and other parameters. Each genome assembly has its own database (lower level), with assembly-specific metadata tables (*trackDb*, *hgFindSpec*) and one or more data tables per displayed Browser track.

UCSC database and Genome Browser in local installations.

The Genome Browser source code is free to noncommercial users and can be downloaded and installed locally. This is useful when privacy considerations, for example, preclude the transmission of data through a network or outside of a local firewall. The Genome Browser can be adapted to organisms not hosted at UCSC without the need for any programming, although the installation does require specific configuration of a Unix system. The process for adding a new genome assembly to an existing installation of the Genome Browser is entirely table- and file-driven.

Briefly, the .2 bit file is stored in a directory on the filesystem. A new assembly is instantiated by appropriate metadata entries in *hgcentral* tables. The addition of a new *gggSss#* database in MySQL and five new tables within it completes the installation of a minimal browser. Among these, two metadata tables, *trackDb* and *hgFindSpec*, specify display options and search parameters, respectively. The *gold* and *gap* tables contain the coordinates from the AGP file of both the contigs or scaffolds and the unsequenced regions from which the assembly was constructed. The *chromInfo* table describes chromosome (or for lower coverage assemblies, scaffold) size and names. Details of the implementation are found in online documentation at <http://genome.ucsc.edu/admin/mirror.html>.

The minimal installation provides basic navigation, the Base Position track with a scale bar at the top of the Browser graphic and the Assembly (from the gold table) and Gap tracks. Two other functionalities are provided automatically by the software in the form of data tracks that, unlike the typical track, are computed on-the-fly and do not rely on a table in the database. The Short Match track allows a user to view the locations of all occurrences of short strings of sequence (2–30 bases) by typing the sequence into the configuration page. The Restriction Enzymes track will display all or a user-selected subset of enzymes with recognition sites in the Genome Browser viewing window.

Standard browser

A standard browser typically includes the minimal browser described above plus data tracks showing lineage-specific repeats (Repeat Masker [14]); GenBank mRNAs and ESTs [15]; physical clones mapped to the genome, such as BACs, YACs, fosmids or STS markers; and gene predictions (e.g. N-SCAN, Ensembl Genes and/or UCSC Genes) [16–18].

All genome assemblies are also released with a BLAT server [19], which allows users to quickly align any sequence of sufficient length (>25 bases) to the reference assembly. These servers load an index of the genome into RAM on the machine and search it with the query sequence, providing a rapid alignment. Dedicated machines at UCSC run BLAT instances for the 104 assemblies on the public site.

Annotation tracks

After a basic browser is built on an assembly, it is usually further populated with annotation data of many types. The database may contain any data that can be mapped to genomic coordinates and therefore displayed in the Genome Browser. Each data set forms a separate track, a self-contained information set that derives from a single experiment or source and represents a particular type of annotation. A data track may be all the mRNAs from GenBank, all DNase hypersensitivity sites in a particular human cell line, all sites with knockout mutations available from a mouse repository or all regions that have homology in another organism. The annotation data represented in the UCSC Genome Browser typically fall into three categories: computed at UCSC, computed elsewhere and displayed in the

Browser or computed and hosted entirely elsewhere (Track Data Hubs).

Tracks generated at UCSC

Most standard browsers are further expanded to include comparative genomic annotations [20] as well as Convert and liftOver capabilities. These functions allow users to navigate between assemblies using sequence homology and synteny. This is useful, especially during times of transition to a new assembly of the same organism, but it also forms a basis for finding regions of orthology between organisms. In this way, for example, users studying a human gene can easily navigate to an orthologous gene in mouse to leverage the results of mouse experiments that cannot be done in humans, such as transfections and knockouts.

All browsers are populated with a set of annotations from GenBank. All mRNAs and ESTs in GenBank are classified either as same species or other species and aligned to the reference assembly in separate tracks. The alignment of 75 million GenBank RNAs and ESTs to the 3 billion bases of the human reference assembly takes ~2 CPU-years of compute time. To keep the tracks current, new mRNA sequences are automatically obtained from GenBank every night and aligned to all 104 genome assemblies. EST sequences are updated weekly.

Other datasets computed at UCSC include the UCSC Genes tracks [18] on human and mouse, BAC-end pairs on the human assembly and Human Proteins mapped to nonhuman genome assemblies using tBLASTn [21]. The Conservation composite track displays the results of the multiz algorithm [22] that aligns the results from up to 46 pairwise Blastz alignments to the reference assembly [20]. It also shows results from the phastCons and phyloP [23, 24] algorithms that compute conservation scores for each basepair in the genome based on those alignments. For each reference assembly, this track typically aligns several close evolutionary relatives to the reference organism as well as human and a small number of other outgroups (Figure 4). Construction of the 47-species multiz track on the hg19 human assembly consumed ~10 CPU-years on the UCSC compute clusters.

For organisms that are poorly annotated due to a paucity of native mRNAs in GenBank, such as horse, pig, chicken, platypus and zebrafish, the Human Proteins track provides a homology-based gene set by aligning proteins in amino acid space

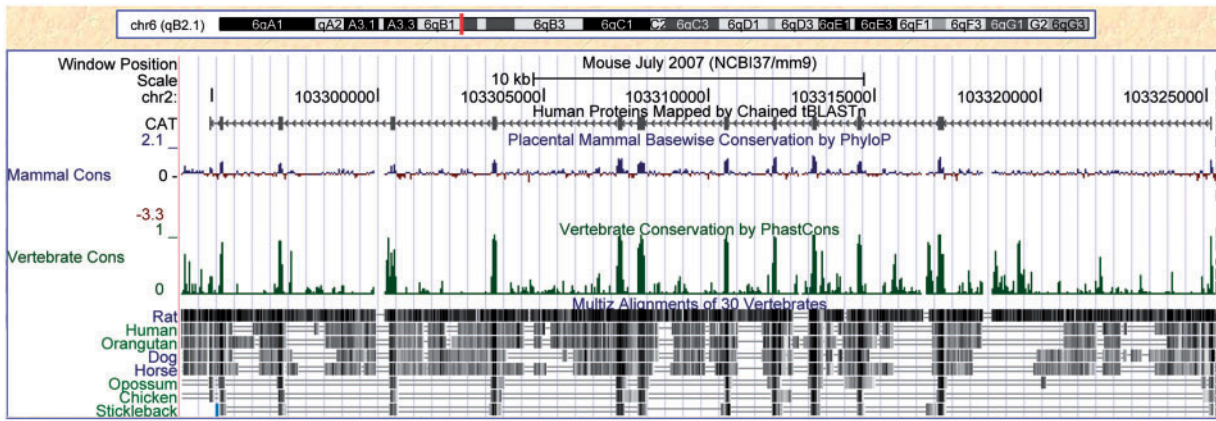


Figure 4: Genome Browser display of CAT gene region on chr2 in mouse mm9 assembly. Shows gene structure (UCSC Genes track), PhyloP, PhastCons scores and multiz alignments. Note the high conservation in exon regions far back into evolutionary time (opossum, chicken, stickleback), while conservation in intron3 regions has been lost.

using tBlastn, which is more useful than DNA alignments for distantly related organisms.

Third-party tracks

A large number of datasets originate outside UCSC and contribute to the Genome Browser's core idea of hosting as many high-quality resources as possible. In most cases, UCSC does not perform significant postprocessing or computation on the data, limiting intervention to converting file formats or other parsing tasks and quality-assurance review. Examples of this type of track include probe sets for commercially available microarrays, human copy-number variation (CNV) data from the Database of Genomic Variants (DGV) [25]; human dimorphism data from DECIPHER [26]; expression data for mouse and human from the GNF Expression Atlas [27]; and segmental duplication data for human, mouse, rat, dog and chicken [28].

The ENCODE project, for which the UCSC Genome Browser is the Data Coordination Center [29], presents a large number of functional annotations: including DNase hypersensitivity sites, indicating open chromatin; histone marks, implicated in gene regulation; and gene expression levels from whole-genome RNA-seq experiments. These data, which are available on the human and mouse assemblies hg19 and mm9, are mapped across multiple cell lines. The resulting tracks represent tissue specificity and developmental mileposts (e.g. embryonic stem cells) for these elements. They can be displayed along with any other tracks on the same assembly, such as GenBank mRNAs or multispecies conservation.

A complete list of tracks available for any assembly can be found by visiting the Gateway page for any genome assembly (<http://genome.ucsc.edu/cgi-bin/hgGateway>) and clicking the button, 'configure tracks and display' or by simply inspecting the track controls beneath the main Browser graphic. The Track Search feature provides keyword lookup.

Examples of data tracks that do undergo further processing or filtering at UCSC include dbSNP [30] and OMIM (Online Mendelian Inheritance in Man) [31]. In these tracks, data from the providers are subdivided into categories to make them more useful to our users. For example, dbSNP data are presented in their entirety in one track, but three other tracks offer subsets: Common Single Nucleotide Polymorphisms (SNPs) (those with minor allele frequency >1%), Flagged SNPs (those identified in dbSNP as 'clinical'—*may* be associated with disease, but use with caution!) and Multiple SNPs (those mapping to more than one genomic location).

Similarly, the OMIM data set has been filtered by UCSC to create three separate tracks, including one track of Allelic Variant SNPs that have phenotypic associations annotated by OMIM. These filtered sets are transmitted to OMIM for redistribution to their licensees. As always, details of how the filtering was done are available by clicking into an item or via the track configuration page.

Users may read about the filtering options available when using tracks by clicking on the small button to the left of the track in the Genome Browser image, or on the label in the track control area below the image. This configuration page gives

users an opportunity to set colors and filters to suit themselves.

For users who do not know exactly which data set contains the information they seek, each data track is accompanied by a description outlining the rationale for the production of the data, implementation details, interpretation guidelines and references to the literature. All of this information is indexed and may be searched by keyword via the Track Search button beneath the Browser graphic. The result is a list of all tracks that have the search term in the documentation and a link to the track description.

Remote content: track data hubs

The ‘track data hub’ feature of the Genome Browser allows users to make their data available on the Browser for the entire UCSC user base. These data tracks are hosted remotely, managed by the contributing lab and may be displayed on any genome assembly. No data are stored at UCSC and only a single row of a data table is ever stored at UCSC. Track hubs are discussed in more detail below. The track data hub mechanism presently serves as the access point to the large amount of data produced by the Epigenomics Roadmap project on the human reference assembly [32], and is available to all users. The data hub functionality is especially useful for those generating Gigabytes of data from HTS experiments, which are too large to upload via the traditional Custom Track method.

TECHNOLOGIES

Display engine

The UCSC Genome Browser data display uses the Common Gateway Interface (CGI) to provide interactive content to Browser users. The Genome Browser communicates to the user via a group of 35 CGI programs written in the C language serving as a front end to the MySQL database and a large number of flat files. The ENCODE project [33, 34] accounts for a significant proportion of the data in the hg18 and hg19 databases.

Different CGIs are used to drive various parts of the Genome Browser, and each sends one or more queries to the MySQL database. For example, hgGateway draws the page on which a user selects the genome assembly of interest, querying the hgcentral database several times in the process of responding to user input. Changes in gateway

pull-down menus are mediated by Javascript when a user chooses species and assembly.

The hgTracks CGI (<http://genome.ucsc.edu/cgi-bin/hgTracks>) is the workhorse that draws the main Browser graphic. Each of the many tracks a user may choose to display is drawn after a query to MySQL that pulls the appropriate track content for the sequence interval. For many tracks, the speed of retrieval of data for a genomic interval is enhanced using a binning scheme as described in Kent *et al.* [1]

The graphic image is computed as a single image, but is then presented in the Genome Browser in slices, one for each track, so that the tracks may be reordered by the user without the need to requery the database. The reordering of track slices in the image is accomplished by Javascript on the client side. Other features that use Javascript in the hgTracks CGI include drag-and-zoom and right-click track configuration, described previously [3].

When hgTracks draws the Genome Browser graphic, it not only draws the annotations in graphical form, but it also generates a client-side image map so that each item in the display becomes a unique clickable link to a details page with specific information about that item (e.g. for an mRNA, the link leads to a page showing the contributing lab, date, tissue and other information; Figure 5A). The user is then one click away from the alignment of the mRNA against the genome (Figure 5A and B). The format of the page is encoded in the CGI, but the detailed information for the specific mRNA is inserted into the page from the database when the item is selected in the graphic. The alignments are precomputed and stored in the database and can be displayed instantly upon demand.

The details page for the items in the UCSC Genes tracks is drawn by its own CGI, hgGene. This CGI populates a large matrix of links leading to further information about the gene (Figure 6). These links lead to content about the gene that is either UCSC-produced, such as orthology predictions (Other Species link), or is externally provided by other databases, such as OMIM [31], GeneCards [35], PubMed, GeneReviews [36], the Human Genome Nomenclature Consortium (HGNC) [37], the Mouse Genome Database at Jackson Lab [38], UniProtKB [39], ExonPrimer, Treefam [40] and more. The hgGene CGI draws the details page for other key model organisms as well as human (mouse,

rat, *Drosophila melanogaster*, *C. elegans*, *Saccharomyces cerevisiae*).

discussion provides an overview of the integrated system functionality.

FEATURES

The newest features of the UCSC Genome Browser, introduced over the last few years, have been reviewed as they appeared [2–4]. The following

Display and navigation

The UCSC Genome Browser has a large number of features designed to make the user's browsing experience fast and intuitive. The display is

Home	Genomes	Genome Browser	Blat	Tables	Gene Sorter	PCR	Session	FAQ	Help
Human Gene TP53 (uc010cnf.1) Description and Page Index									
Description: tumor protein p53 isoform f									
RefSeq Summary (NM_001126117): This gene encodes tumor protein p53, which responds to diverse cellular stresses to regulate target genes that induce cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism. p53 protein is expressed at low level in normal cells and at a high level in a variety of transformed cell lines, where it's believed to contribute to transformation and malignancy. p53 is a DNA-binding protein containing transcription activation, DNA-binding, and oligomerization domains. It is postulated to bind to a p53-binding site and activate expression of downstream genes that inhibit growth and/or invasion, and thus function as a tumor suppressor. Mutants of p53 that frequently occur in a number of different human cancers fail to bind the consensus DNA binding site, and hence cause the loss of tumor suppressor activity. Alterations of this gene occur not only as somatic mutations in human malignancies, but also as germline mutations in some cancer-prone families with Li-Fraumeni syndrome. Multiple p53 variants due to alternative promoters and multiple alternative splicing have been found. These variants encode distinct isoforms, which can regulate p53 transcriptional activity. [provided by RefSeq, Jul 2008].									
Strand: - Genomic Size: 7092 Exon Count: 8 Coding Exon Count: 6									
Page Index	Sequence and Links	UniProtKB Comments	Genetic Associations	CTD	Microarray				
RNA Structure	Protein Structure	Other Species	GO Annotations	mRNA Descriptions	Pathways				
Other Names	GeneReviews	Model Information	Methods						
Data last updated: 2009-10-08									
– Sequence and Links to Tools and Databases									
Genomic Sequence (chr17:7,571,720-7,578,811)			mRNA (may differ from genome)			Protein (214 aa)			
Gene Sorter	Genome Browser	Protein FASTA	Table Schema	CGAP	Ensembl				
Entrez Gene	ExonPrimer	GeneCards	Gepis Tissue	H-INV	HGNC				
Jackson Lab	OMIM	PubMed	Stanford SOURCE	UniProtKB	User annotations				
– Comments and Description Text from UniProtKB									
ID: Q3LRW1 HUMAN									
DESCRIPTION: RecName: Full=Cellular tumor antigen p53;									
FUNCTION: Acts as a tumor suppressor in many tumor types; induces growth arrest or apoptosis depending on the physiological circumstances and cell type. Involved in cell cycle regulation as a trans-activator that acts to negatively regulate cell division by controlling a set of genes required for this process. One of the activated genes is an inhibitor of cyclin-dependent kinases. Apoptosis induction seems to be mediated either by stimulation of BAX and FAS antigen expression, or by repression of Bcl-2 expression (By similarity).									
SUBUNIT: Binds DNA as a homotetramer (By similarity).									
SUBCELLULAR LOCATION: Cytoplasm. Nucleus (By similarity).									
SIMILARITY: Belongs to the p53 family.									

Figure 6: Top of Genome Browser UCSC Genes 'details page' for human TP53 gene (hg19 assembly). This page is the gateway to deep information about the gene. The top section reproduces the RefSeq summary from NCBI, including information at the biochemical, genetic, cellular, physiological and clinical/medical levels, where available. This is followed by an index of this very large page. Links in the boxes lead to individual sections of the page giving details, with links to original data contributors for protein information, microarray gene expression data, pathways and other information. The second section, Sequence and Links to Tools and Databases, features links to content at UCSC (light, greenish background) and external sites (darker, bluish background), including Human Genome Nomenclature Consortium (HGNC), OMIM and others. The third section, Comments and Description Text from UniProtKB, is an example of content of the page from an outside source (each is linked from Page Source section above).

continuously resizable from a single basepair to a full chromosome. Due to this 250-million-fold scale range (for human chr1), various levels of detail are displayed at different scales. Depending on the resolution, the Genome Browser shows as much detail as can fit on the screen. For example, when sufficient screen space is available, the nucleotides, codons or codon numbering are shown (Figure 7A). When zoomed out to larger scale, an mRNA base mismatch may be represented as a colored tick mark and exons shown in alternating colors (Figure 7B). Farther out still, an exon or an entire gene may be represented as a single tick mark (Figure 7C). For genome assemblies having cytological banding coordinates (human, mouse, rat, *D. melanogaster*), a graphic representation of the location is shown above the main Browser graphic (Figure 7C). At all zoom levels, the main Browser graphic shows a scale bar in the center third of the screen (in bases, kilobases or megabases, as appropriate) to assist in orienting the user.

The traditional navigation control buttons above and below the Browser image allow users to zoom in and out and scroll right and left. It is also possible to go directly to a gene or other annotation type (e.g. SNPs) by entering a name or accession number in the position/search box above the Browser graphic. This feature supports accession names for a large number of annotation types, including mRNAs, chromosome bands, SNP rs numbers and more (Figure 7C). Several Genome Browser features make it more intuitive to navigate the genome: Drag-and-zoom, drag-reorder, drag-scroll and right-click navigation collectively give the user a hands-on, mouse-driven way to configure the view (Reviewed in [3]).

Associated tools

In addition to the main graphic image, the UCSC Genome Browser site contains a number of other tools for accessing genomic data (Figure 8). These tools are widely used and account for 42% of the traffic on the UCSC servers. Automated usage of the Table Browser accounts for 13% of the 1.2 million page hits per day. Interactive use of the other tools accounts for the rest: the UCSC Genes details pages, 7%; the Gateway page (for changing/selecting assemblies), 5%; and BLAT, 3%. Genome Browser users focus ~80% of their attention on human assemblies and 12% on mouse, creating more than 1100 new Saved Sessions per month.

Many of the associated tools are particularly well suited for researchers interpreting the results of HTS experiments. Experiments that produce a short-read alignment signal over specific locations (e.g. ChIP-seq, RNA-seq) benefit from the Custom Track or Track Data Hub display features and the bigBed and bigWig [41] data formats. The Table Browser may be used for advanced correlation and intersection analysis of these datasets with each other and with UCSC resident data. For example, the gene expression patterns in two cell lines may be compared for overlap or lack thereof.

Whole-genome or -exome sequencing experiments typically yield data in the form of SNPs or CNVs, which can be viewed in the Genome Browser using Binary Alignment/Map, Variant Call Format (VCF) [42, 43] or Personal Genome SNP formats, via either Custom Tracks or Track Data Hubs.

Custom tracks

The Custom Track feature allows the loading of personal data into the Genome Browser for viewing alongside the rich annotation set of the resident Browser tracks. There are a variety of methods for uploading a track, including copy/pasting into a text box on the page, uploading a file from a local disk drive or input of a URL for a file that is located on a web-accessible server. A special track group (User Tracks), which contains visibility controls just like the Browser-resident tracks, appears at the top of the track controls. These tracks also become fully integrated into the Table Browser for intersection with each other or with resident tracks. Many users find Custom Tracks a convenient way to keep track of experimental data such as SNP results from sequenced samples, or as pointers to genes in a pathway of interest. For example, the variation data that results from sequencing experiments may be uploaded in the Personal Genome SNP format to show the differences between the sample and the reference assembly (Figure 7A). Custom tracks may also be loaded from data in the all-purpose VCF (<http://vcftools.sourceforge.net/specs.html>).

A Custom Track User's Guide is available via a link from the Custom Track input page (<http://genome.ucsc.edu/cgi-bin/hgCustom>), and a large help section that gives access to technical details for uploading and displaying tracks in more than a dozen data formats (<http://genome.ucsc.edu/goldenPath/help/customTrack.html>).

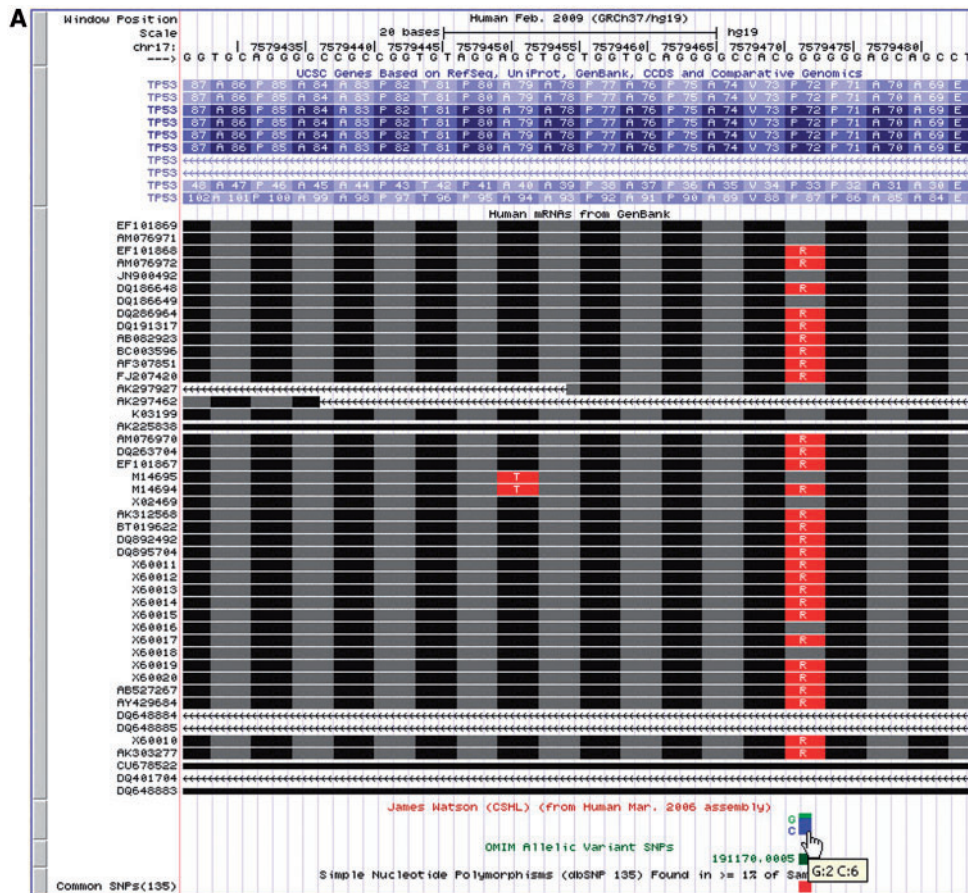


Figure 7: (A) Close-up view (58 bp window) of part of the fourth exon with several tracks displayed. At this resolution, the UCSC Genes track (top track) shows amino acid number and identity for each isoform. The second track, Human mRNAs, which has the ‘show nonsynonymous mRNA codons’ option turned on, shows that most mRNAs differ (arginine) from the reference assembly (proline) at amino acid 72 of the major isoform. Does the reference assembly have a minor allele at this location? At bottom are three tracks with information about SNPs in codon 72 of the major isoforms (two isoforms have different numbering). The James Watson track uses the Personal Genomics SNPs track display format, indicating that 6 C and 2 G nucleotides were found at this location. The coloring shows the relative amounts of the two nucleotides and the mouseover shows the actual read depth. This format may also be used to construct Custom Tracks. The OMIM Allelic Variants SNPs track shows that this SNP has a documented phenotypic association. A link from the details page directly to the relevant record at omim.org leads quickly to the information that this SNP is a polymorphism, not a mutation. The snp135 Polymorphism track at the bottom leads in one click to the information from dbSNP indicating that this SNP is present in the population (of more than 7500 individuals sampled) at nearly 50% frequency. **(B)** Midrange view (930 bp) of (left to right) the fourth, third and part of the second exon with several tracks displayed. The region shown in Figure 7A is in the center (fourth exon). The OMIM AV track has been dragged to top of image. At this resolution, the nucleotides and amino acids in UCSC Genes and mRNA tracks are not labeled: instead, codons appear as dark and light stripes. The mouseover pictured for one isoform of the gene (left side of image in UCSC Gene track) indicates that a click on the double-headed arrow will shift the display to the next, fifth, exon of this isoform (different isoforms have different numbering). The Conservation tracks show that the 3’-end of Exon 4 (conservation graph in Placental Mammal and Vertebrate Conservation tracks slightly left of center in the image) is more highly conserved than the 5’-end, and that conservation, in general, is higher in the exons than in the introns. The SNP track now shows several more polymorphisms, including two non-synonymous amino acid changes (red or intermediate gray in grayscale) and one synonymous amino acid change (green or light gray in grayscale). **(C)** Wide view showing the region (930 kb) around the TP53 gene (highlighted in 80.8 kb region in center). At this scale, it is advantageous to turn off the isoforms, on the configuration page as shown here, by clicking on the minibutton to the left of the track (arrow). The chromosome ideogram above the main Browser graphic shows the location on the chromosome as a red box superimposed on the chromosome bands on the short arm of chr17. The highlighted 80 kb region in the center may be viewed using drag-and-zoom by releasing the mouse button.

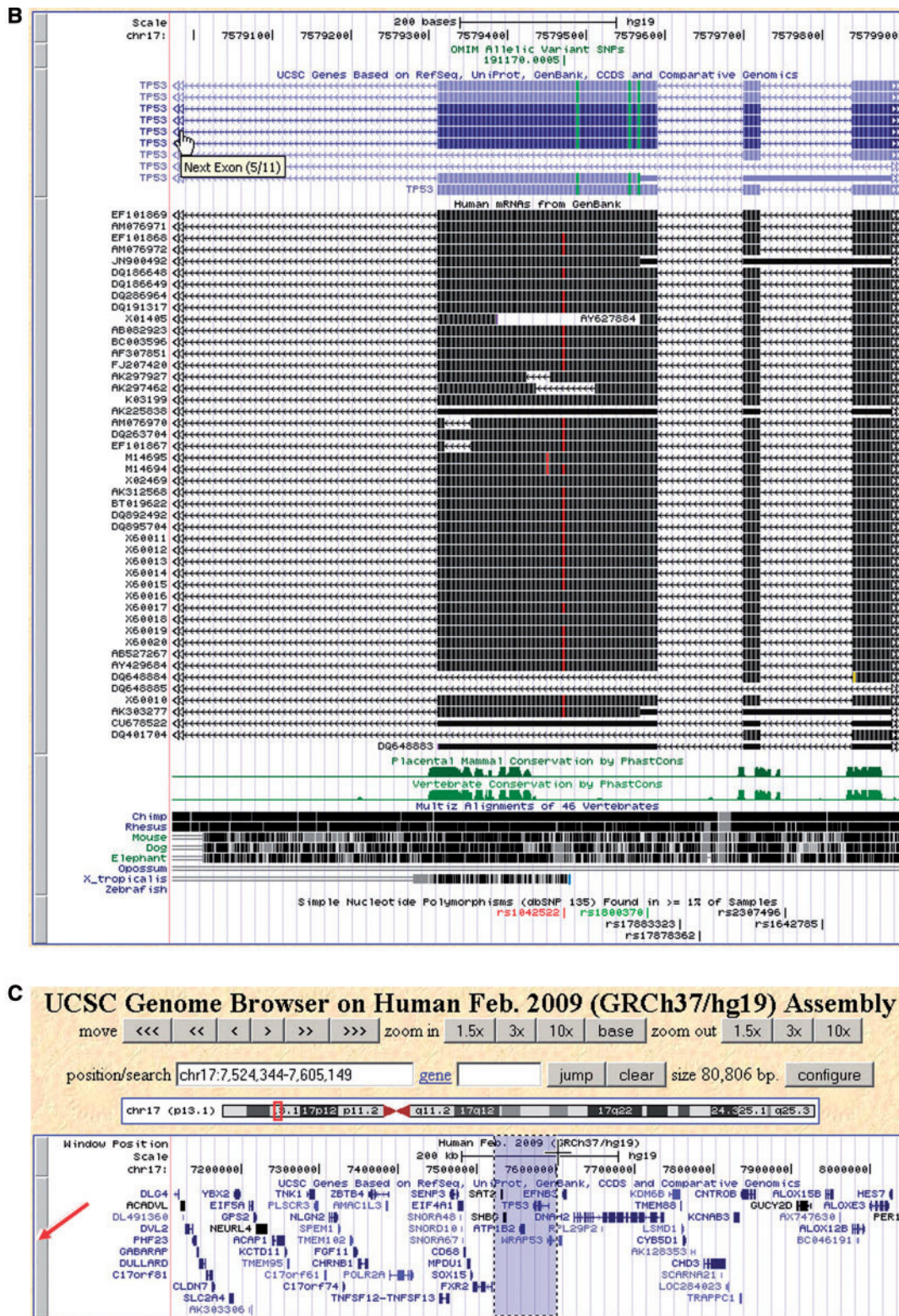


Figure 7: Continued.

Sessions

Configuring the Genome Browser display to show a particular view of the genome can be a time-consuming process. To make a point about a gene

or region, a researcher may need to turn on and configure several different tracks and navigate to the exact location and resolution desired. Sharing this information can require describing a daunting

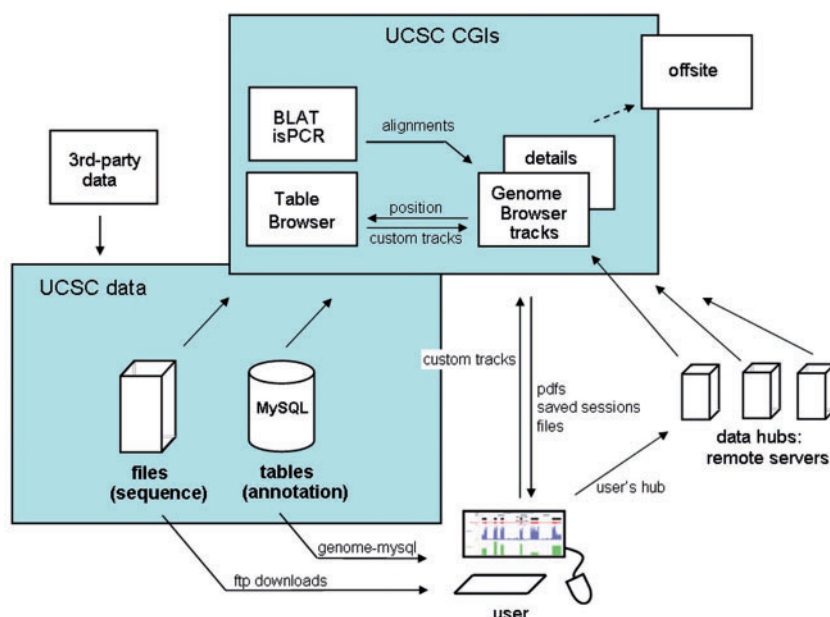


Figure 8: Overview map of UCSC Genome Browser and associated tools. The user interacts with the system via the CGIs (upper shaded area), by clicking in a web browser, such as Firefox, Chrome, Safari or Internet Explorer. The user may upload Custom Tracks. The user receives screen images, saved sessions, data files and pdf image files. The user may also make data available to the Browser via a remote data hub and view the data from other hubs. The CGIs interact with each other as indicated: position information may be sent from the Genome Browser to the Table Browser, and Custom Tracks or tabular data returned. Details about an alignment are available by clicking on an item in the Browser viewer, which in turn may have links to external websites (dotted arrow, upper right). BLAT and isPCR alignments are displayed as tracks in the Browser image. All of the CGIs obtain information from the filesystem or database (lower shaded area), much of which originates with third-party data contributors (upper left). All data are also available to the user directly via ftp download and via direct access to the genome-mysql server (lower left).

list of steps. The Sessions tool (<http://genome.ucsc.edu/cgi-bin/hgSession>) allows a user to save the state of the Genome Browser for later access or for sharing with a colleague. The user can save these settings to a file to be loaded at a later date, or can create a login and save sessions to a personal account. (This is the only place in the Genome Browser where a login is required and is necessary only to allow a user to access, manage and share session information with others.) The user can then easily return to saved sessions, and share them using the associated URL. This is an excellent way to sequester different trains of thought for easy access and future reference. Help is available at <http://genome.ucsc.edu/goldenPath/help/hgSessionHelp.html>.

Track data hubs

Track hubs (<http://genome.ucsc.edu/cgi-bin/hgHubConnect>), like Custom Tracks, allow users to view their own data on the Genome Browser alongside native annotation tracks. However, the

track hubs are posted by the user to web-accessible directories where the Browser software can read the files (Figure 8). This mechanism offers data persistence and enhanced track configurability not provided by Custom Tracks: hub tracks have the full configurability of UCSC-resident tracks such as grouping into composite tracks. Track creators may register their hubs with UCSC to make the content available to any user of the Genome Browser, exposing them to a larger audience or they may be kept private.

Using one of the compressed, indexed data formats, such as bigBed, bigWig [41] or BAM [42], large datasets are stored locally on the server of the hub creators. Only the small portion of the files being requested for display is sent to the Genome Browser, greatly enhancing the speed of the interaction. The really large datasets from most HTS experiments may only be viewed in these compressed formats, as upload times for the full files usually

exceeds http connection limits. A complete description of the process of setting up a data hub may be found online at <http://genome.ucsc.edu/goldenPath/help/hgTrackHubHelp.html>.

Hub creators remain in control of their data, administering them and make them available on their own servers, updating at will. Data hubs are accessed via the 'track hubs' link on the Gateway page or beneath the main Browser graphic. Data from the Roadmap Epigenomics Human Epigenomics Atlas Project [32] hub at Washington University are presented in this fashion.

Table Browser

The Table Browser (<http://genome.ucsc.edu/cgi-bin/hgTables>) is a full-featured user interface on the database tables underlying the Genome Browser. The Table Browser allows filtering of datasets using a variety of criteria, producing information on a specific subset of genes or all the annotations in a particular region of a the genome, for example. Users can also intersect datasets with each other and capture output in a variety of formats, including DNA sequences, flat files or custom tracks in the Genome Browser. A full description of Table Browser capabilities can be found in the Help pages: <http://genome.ucsc.edu/goldenPath/help/hgTablesHelp.html>.

BLAT

BLAT, the Blast-Like Alignment Tool (<http://genome.ucsc.edu/cgi-bin/hgBlat>) allows users to align a sequence of 25 bases or more to the genome [19]. BLAT can be used to look for alignments of short-read output from sequencing machines and can also locate sequences from one organism on the assembly of another. Because some level of mismatch is tolerated, cross-species alignments may be performed provided the species have not diverged too far from each other.

isPCR

The *in silico* PCR tool (<http://genome.ucsc.edu/cgi-bin/hgPcr>) allows the user to enter a pair of potential PCR primers and 'amplify' the genomic DNA. The output is a block of sequence that would be obtained if the reference assembly were the substrate DNA in the experiment. isPCR will detect if the sequence is unique in the genome by giving output for all locations in the genome having the primers in the proper orientation and distance in the genome. For the human and mouse genome assemblies, an optional

target is to use the UCSC Genes track as 'substrate', returning the PCR products that would be obtained using all the spliced mRNA isoforms as template.

Downloads and MySQL access

The full UCSC Genome Browser data set may be downloaded for offline use. Accessed via a link on the home page, or directly via <http://hgdownload.cse.ucsc.edu/downloads.html>, the downloads section is arranged by genome assembly. In this location, the entire contents of any table are available. The genomic databases are also available via a public MySQL server, genome-mysql.cse.ucsc.edu, where users may access the data and run program-driven queries directly on the data.

In addition, several other utilities may be obtained via the downloads page, including the kent source tree (free for noncommercial users), which contains a large number of Unix command-line utilities useful to bioinformaticians. Users may also access files of DNA sequence for the human reference assemblies in which the substitution SNPs from dbSNP are incorporated into the sequence using the IUPAC degeneracy code (<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/snp135Mask/>). The Genome Browser code is free to noncommercial users and can be used to build a local mirror of the Browser.

PDF/PS output

Genome Browser graphic views may be exported as a high-resolution PDF or Postscript file via a button at the top of the Browser window. Both the main Browser graphic and the chromosome ideogram can be prepared as camera-ready graphics for publication in this way.

All links to UCSC Genome Browser functionality mentioned in this publication are collected in Table 1.

Help

The tools of the Genome Browser toolkit are accompanied by extensive online documentation. For example, although loading a custom track can be as simple as typing in a set of coordinates (chrom, chromStart, chromEnd = BED3 format), the options available for Custom Tracks include 14 different supported file formats or track types, described in detail in the Custom Track help pages, and configuration options abound.

Users seeking help have other resources available to them as well. The Genome mailing list archives is a record of all the questions that have been asked and answered on the list since July, 2000. Both the search

Table 1: Links to Browser tools

Genome Browser site: http://genome.ucsc.edu
Selection page for choosing genome assemblies: http://genome.ucsc.edu/cgi-bin/hgGateway
Main Browser graphic display: http://genome.ucsc.edu/cgi-bin/hgTracks
Uploading user data: http://genome.ucsc.edu/cgi-bin/hgCustom
Designating and saving sessions: http://genome.ucsc.edu/cgi-bin/hgSession
Creating and accessing remote data hubs: http://genome.ucsc.edu/cgi-bin/hgHubConnect
Sequence alignment tool, Blat: http://genome.ucsc.edu/cgi-bin/hgBlat
<i>In silico</i> PCR: http://genome.ucsc.edu/cgi-bin/hgPcr
Table Browser: http://genome.ucsc.edu/cgi-bin/hgTables
Building a mirror: http://genome.ucsc.edu/admin/mirror.html
Bulk data downloads: http://hgdownload.cse.ucsc.edu/downloads.html
SNP-masked sequence for human, based on dbSnp release 135: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/snp135Mask/
High-throughput sequence alignment software for VCF format: http://vcftools.sourceforge.net/specs.html

All links to functional tools in the UCSC Genome Browser are collected in this table in the order of appearance in the text.

function and the email address for the mailing list may be accessed from the Home page ('Contact Us' link, <http://genome.ucsc.edu/contacts.html>). Users are encouraged to search the archives before contacting the list, but are also invited to post questions to the list. The Genome mailing list is available for those seeking advice about how to use the Genome Browser, but is not equipped to assist with specific research questions. Online Help links mentioned in this article are collected in Table 2.

FUTURE DIRECTIONS

As the cost of sequencing continues to drop and access to sequencing machines becomes ever more widespread, the number of complete human and other animal genomes available continues to rapidly increase. The task of keeping up with the flood of information is a continuing challenge and the cost of analysis probably already exceeds the cost of sequence acquisition. The UCSC Genome Browser continues to develop tools for visualizing genome-scale data, including expanding the multiz tracks on human and mouse assemblies to include a larger number of organisms. We are also increasing the coverage of the Personal Genomes track on hg19. This track currently maps nine individuals by showing the locations of variation between them and the reference assembly. We are also developing tools to show haplotype linkage information for the typical diploid sample.

The analysis of human variation continues to contribute to our understanding of human populations and their differences. We expect to continue to display results from large sequencing projects, such

Table 2: Links to help documents

Custom tracks: http://genome.ucsc.edu/goldenPath/help/customTrack.html
User saved sessions: http://genome.ucsc.edu/goldenPath/help/hgSessionHelp.html
Data hubs: http://genome.ucsc.edu/goldenPath/help/hgTrackHubHelp.html
Table Browser: http://genome.ucsc.edu/goldenPath/help/hgTablesHelp.html
Searching and contacting mailing lists: http://genome.ucsc.edu/contacts.html

Any links to help documents mentioned in the text are collected here.

as the 1000 Genomes Project [44], structural variation data from DGV [45], dbSNP [30] and dbVar [46], and clinical data such as those reported by Cooper *et al.* [47].

In the future, we expect to implement enhanced tools to assist laboratories using HTS to identify significant SNPs in their samples. Among expected enhancements is a tool to display on the Genome Browser the likely effect of a SNP on a proposed protein product, integrating it with variation annotations already displayed on the Browser. The Ensembl project has released a similar tool, which they call the Variant Effect Predictor [48].

As discussed above, it is possible for researchers to build a local mirror of the Genome Browser and several hundred have done so, including some who use it for genome assemblies not hosted at UCSC. It is not yet as easy to implement as we would like, however, and as a service to those sequencing these organisms (e.g. the Genome 10 K project [49]), we hope to implement a functionality for users to

easily add custom organisms not hosted at UCSC. We hope that will work in a fashion similar to Custom Tracks or Track Hubs and will require no system-level implementation on the user side.

We also hope to deploy a mirror site in the near future in Europe, where traffic from that continent will be redirected. This is expected to overcome transmission latency and make the UCSC website operate faster for those users, as well as for North American users who will benefit from the distributed load. We intend that researchers will have a choice of which server to use and that all functionality will operate exactly as it does on the California site.

Key points

- The UCSC Genome Browser provides a graphical viewer for a large number of sequenced animals with a consistent, coherent interface.
- Many annotation tracks have been mapped to the reference genome sequences, including extensive comparative genomic data.
- Users may display their own data, including HTS data using Custom Tracks and Data Hubs, adjacent to the rich data hosted at UCSC.
- Genome Browser sessions may be saved and shared between users and exported as high-quality graphic files.
- The annotation database is available for download and a large library of software tools for manipulation of DNA sequences is available.

Acknowledgements

The authors wish to gratefully acknowledge data contributors and users worldwide who have contributed data and ideas to the Genome Browser. The staff of the UCSC Genome Browser is also acknowledged, especially Donna Karolchik, Brooke Rhead, Venkat Malladi, Max Häussler, Melissa Cline, Rachel Harte and Pauline Fujita for helpful comments and contributions to the manuscript, and the system administration team, Jorge Garcia, Erich Weiler and Gary Moro. Finally, we thank our Scientific Advisory Board for their consistent support and valuable guidance. The authors receive royalties from the sale of UCSC Genome Browser source-code licenses to commercial entities.

FUNDING

National Human Genome Research Institute (P41HG002371 to D.H., W.J.K. and R.M.K., U41HG004568 to W.J.K., subcontracts P01HG5062, U41HG004269 and U01HG004695 to W.J.K.); subcontracts from the National Institute of Dental and Craniofacial Research (U01DE20057 to R.M.K.); National Institute of Child Health and Human Development (RC2HD064525 to R.M.K.); the National Institute of Environmental Health

Sciences (U01ES017154 to W.J.K.); Howard Hughes Medical Institute (to D.H.). Funding for open access charge: Howard Hughes Medical Institute.

References

1. Kent WJ, Sugnet CW, Furey TS, *et al.* The Human Genome Browser at UCSC. *Genome Res* 2002;**12**:996–1006.
2. Fujita PA, Rhead B, Zweig AS, *et al.* The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* 2011;**39**:D876–82.
3. Dreszer TR, Karolchik D, Zweig AS, *et al.* The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res* 2012;**40**:D918–23.
4. Kuhn RM, Karolchik D, Zweig AS, *et al.* The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res* 2009;**37**:D755–61.
5. Rhead B, Karolchik D, Kuhn RM, *et al.* The UCSC Genome Browser database: update 2010. *Nucleic Acids Res* 2010;**38**:D613–9.
6. Flicek P, Amode MR, Barrell D, *et al.* Ensembl 2012. *Nucleic Acids Res* 2012;**40**:D84–90.
7. Stein LD, Mungall C, Shu S, *et al.* The generic genome browser: a building block for a model organism system database. *Genome Res* 2002;**12**:1599–610.
8. Robinson JT, Thorvaldsdóttir H, Winckler W, *et al.* Integrative genomics viewer. *Nature Biotechnology* 2011;**29**:24–6.
9. Kent WJ, Haussler D. Assembly of the working draft of the human genome with GigAssembler. *Genome Res* 2001;**11**:1541–8.
10. The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* 2001;**409**:934–41.
11. Pruitt KD, Katz KS, Sicotte H, *et al.* Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet* 2000;**16**:44–7.
12. Flicek P, Aken BL, Ballester B, *et al.* Ensembl's 10th year. *Nucleic Acids Res* 2010;**38**:D557–62.
13. Kent WJ, Zahler AM. The intronator: exploring introns and alternative splicing in *Caenorhabditis elegans*. *Nucleic Acids Res* 2000;**28**:91–3.
14. Jurka J, Kapitonov VV, Pavlicek A, *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res* 2005;**110**:462–7.
15. Benson DA, Karsch-Mizrachi I, Lipman DJ, *et al.* GenBank. *Nucleic Acids Res* 2011;**39**:D32–7.
16. Gross S, Brent MR. Using multiple alignments to improve gene prediction. *J Comput Biol* 2006;**13**:379–93.
17. Flicek P, Amode MR, Barrell D, *et al.* Ensembl 2011. *Nucleic Acids Res* 2011;**39**:D800–6.
18. Hsu F, Kent WJ, Clawson H, *et al.* The UCSC Known Genes. *Bioinformatics* 2006;**22**:1036–46.
19. Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res* 2002;**12**:656–64.

20. Kent WJ, Baertsch R, Hinrichs A, *et al.* Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci USA* 2003;**100**:11484–9.
21. Altschul SF, Madden TL, Schaffer AA, *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
22. Blanchette M, Kent WJ, Riemer C, *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* 2004;**14**:708–15.
23. Siepel A, Bejerano G, Pedersen JS, *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005;**15**:1034–50.
24. Siepel A, Pollard KS, Haussler D. New methods for detecting lineage-specific selection. *Proceedings of the 10th International Conference on Research in Computational Molecular Biology (RECOMB 2006)* 2006;**3909**:190–205.
25. Iafate AJ, Feuk L, Rivera MN, *et al.* Detection of large-scale variation in the human genome. *Nature Genetics* 2004;**36**:949–51.
26. Firth HV, Richards SM, Bevan AP, *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet* 2009;**84**:524–33.
27. Su AI, Wiltshire T, Batalov S, *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 2004;**101**:6062–7.
28. Bailey JA, Gu Z, Clark RA, *et al.* Recent segmental duplications in the human genome. *Science* 2002;**297**:1003–7.
29. Raney BJ, Cline MS, Rosenbloom KR, *et al.* ENCODE whole-genome data in the UCSC genome browser (2011 update). *Nucleic Acids Res* 2011;**39**:D871–5.
30. Sherry S, Ward M-H, Kholodov M, *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001;**29**:308–311.
31. Amberger J, Bocchini CA, Scott AF, *et al.* McKusick's Online Mendelian Inheritance in Man (OMIM®). *Nucleic Acids Res* 2009;**37**:D793–6.
32. Bernstein BE, Stamatoyannopoulos JA, Costello JF, *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology* 2010;**28**:1045–8.
33. ENCODE Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004;**306**:636–40.
34. The ENCODE Project Consortium, Birney E, Stamatoyannopoulos J, *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007;**447**:799–816.
35. Safran M, Chalifa-Caspi V, Shmueli O, *et al.* Human gene-centric databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res* 2003;**31**:142–6.
36. Pagon RA, Tarczy-Hornoch P, Baskin PK, *et al.* GeneTests-GeneClinics: genetic testing information for a growing audience. *Human Mutation* 2002;**19**:501–9.
37. Bruford EA, Lush MJ, Wright MW, *et al.* The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res* 2008;**36**:D445–8.
38. Bult CJ, Eppig JT, Kadin JA, *et al.* The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res* 2008;**36**:D724–8.
39. The UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* 2010;**38**:D142–8.
40. Ruan J, Li H, Chen Z, *et al.* TreeFam: 2008 Update. *Nucleic Acids Res* 2008;**36**:D735–40.
41. Kent WJ, Zweig AS, Barber G, *et al.* BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* 2010;**26**:2204–7.
42. Li H, Handsaker B, Wysoker A, *et al.* The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.
43. Danecek P, Auton A, Abecasis G, *et al.* The variant call format and VCFtools. *Bioinformatics* 2011;**27**:2156–8.
44. Kaiser J. A plan to capture human diversity in 1000 genomes. *Science* 2008;**319**:395.
45. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nature Rev Genet* 2006;**7**:85–97.
46. Church DM, Lappalainen I, Sneddon TP, *et al.* Public data archives for genomic structural variation. *Nature Genetics* 2010;**42**:813–4.
47. Cooper GM, Coe BP, Girirajan S, *et al.* A copy number variation morbidity map of developmental delay. *Nature Genetics* 2011;**43**:838–46.
48. McLaren W, Pritchard B, Rios D, *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 2010;**26**:2069–70.
49. Genome 10K Community of Scientists. Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10,000 Vertebrate Species. *J Hered* 2009;**100**:659–74.