Published online 17 December 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D773–D779***
doi:10.1093/nar/gkm966

The UCSC Genome Browser Database: 2008 update

D. Karolchik[1,*], R. M. Kuhn[1], R. Baertsch[1], G. P. Barber[1], H. Clawson[1], M. Diekhans[1], B. Giardine[2], R. A. Harte[1], A. S. Hinrichs[1], F. Hsu[1], K. M. Kober[3], W. Miller[2], J. S. Pedersen[4], A. Pohl[1], B. J. Raney[1], B. Rhead[1], K. R. Rosenbloom[1], K. E. Smith[1], M. Stanke[5], A. Thakkapallayil[1], H. Trumbower[6], T. Wang[1], A. S. Zweig[1], D. Haussler[7] and W. J. Kent[1]

[1]Center for Biomolecular Science and Engineering, University of California Santa Cruz (UCSC), Santa Cruz, CA 95064, USA, [2]Center for Comparative Genomics and Bioinformatics, Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802, USA, [3]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA 95064, USA, [4]The Bioinformatics Centre, Department of Molecular Biology, University of Copenhagen, Denmark, [5]Philip Morris International, Neuchatel, Switzerland, [6]Navigenics, Inc., Redwood City, CA 94061, USA and [7]Howard Hughes Medical Institute, University of California Santa Cruz (UCSC), Santa Cruz, CA 95064, USA

Received September 18, 2007; Revised and Accepted October 17, 2007

ABSTRACT

The University of California, Santa Cruz, Genome Browser Database (GBD) provides integrated sequence and annotation data for a large collection of vertebrate and model organism genomes. Seventeen new assemblies have been added to the database in the past year, for a total coverage of 19 vertebrate and 21 invertebrate species as of September 2007. For each assembly, the GBD contains a collection of annotation data aligned to the genomic sequence. Highlights of this year's additions include a 28-species human-based vertebrate conservation annotation, an enhanced UCSC Genes set, and more human variation, MGC, and ENCODE data. The database is optimized for fast interactive performance with a set of web-based tools that may be used to view, manipulate, filter and download the annotation data. New toolset features include the Genome Graphs tool for displaying genome-wide data sets, session saving and sharing, better custom track management, expanded Genome Browser configuration options and a Genome Browser wiki site. The downloadable GBD data, the companion Genome Browser toolset and links to documentation and related information can be found at: http://genome.ucsc.edu/.

INTRODUCTION

Fundamental to expanding our knowledge of how the human body works in health and in disease is the capability to access and share data produced through experimentation and computational analysis. The University of California, Santa Cruz (UCSC) Genome Browser Database (GBD) (http://genome.ucsc.edu) (1) provides a common repository for genomic annotation data—including comparative genomics, genes and gene predictions; mRNA and EST alignments; and expression, regulation, variation and assembly data—and robust, flexible tools for viewing, comparing, distributing and analyzing the information. Produced and maintained by the Genome Bioinformatics Group at the UCSC Center for Biomolecular Science and Engineering, the GBD focuses primarily on vertebrate and model organism genomes, with an emphasis on comparative genomics analysis.

As of September 2007 the GBD contains data for 11 mammalian species including human, mouse, rat, chimpanzee, rhesus macaque, horse, cow, cat, dog, opossum and platypus; 8 other vertebrates: chicken, lizard (*Anolis carolinensis*), frog (*Xenopus tropicalis*), zebrafish, fugu, tetraodon, medaka and stickleback; and 21 invertebrates including 11 flies, honeybee, *Anopheles* mosquito, five worms, one yeast (*Saccharomyces cerevisiae*) and two deuterostomes—purple sea urchin and sea squirt. For many of the organisms, more than one assembly is provided, and several older archived assemblies may be

*To whom correspondence should be addressed. Tel: +1 831 459 1544; Fax: +1 831 459 1809; Email: donnak@soe.ucsc.edu

found at: http://genome-archive.cse.ucsc.edu/. The GBD stores a collection of annotation data for each assembly, which can be viewed graphically in the UCSC Genome Browser (2) as a series of 'tracks' aligned to the genomic sequence and grouped according to shared characteristics, for example gene predictions, gene expression and variation data. In most instances, each annotation track is represented by a position-oriented table based on genomic sequence coordinates, and may be supplemented by additional non-positional tables that supply related information or link the primary table to other tables in the database. The data are stored in a variety of formats described at: http://genome.ucsc.edu/FAQ/FAQformat.

Minimally, the GBD provides assembly data, comparative genomics annotations, and mRNA, EST and RefSeq (3) gene alignments (when available) from GenBank (4) for each assembly. When available, links are provided to the complementary annotations in two other major genome browsers, Ensembl (5) and NCBI's MapViewer (6). A large set of additional annotations is available for widely studied genomes such as the human and mouse. Assemblies that lack sufficient native RefSeq data alignments and are of sufficient evolutionary distance from the human genome may also include a human proteins annotation that maps human exons using tBLASTn. The organizations and individuals who contributed to the sequencing, assembly, and annotation of featured organisms are acknowledged at: http://genome.ucsc.edu/goldenPath/credits.html; detailed information about the individual annotation tracks may be found in the Genome Browser by clicking the vertical gray or blue bars to the left of the displayed tracks.

UCSC updates the genome assemblies and annotations in the GBD as new releases become available, with priority given to primate and model organism assemblies and annotations that we feel are of widespread interest to GBD users, based on input from our Scientific Advisory Board and feedback received through our mailing lists and user surveys. (The results from a users' survey conducted in May 2007 may be reviewed at: http://genome.ucsc.edu/goldenPath/help/GBsurvey507.html.) RefSeq and mRNA data from GenBank are updated daily; EST data are updated weekly.

In addition to the Genome Browser, several other graphical tools for exploring the data are available from the GBD website, including the Table Browser (7), which provides access for downloading and manipulating the GBD tables as text or tracks; the BLAT sequence-mapping tool (8); the In Silico PCR tool that searches a sequence database with a pair of PCR primers; the Gene Sorter (9) for exploring expression, homology and other gene relationships; the VisiGene *in situ* image browser, the Proteome Browser (10) for viewing related protein information; and the new Genome Graphs tool for uploading and viewing genome-wide data sets. This toolset is accompanied by a comprehensive set of online documentation and FAQs listed at http://genome.ucsc.edu/FAQ/. Online and hands-on training materials are available via the Training link (http://genome.ucsc.edu/training) on the GBD home page.

The GBD data, tools and source may be downloaded from http://hgdownload.cse.ucsc.edu/downloads.html. Instructions for setting up a local server to mirror all or part of the GBD data can be found at http://genome.ucsc.edu/admin/mirror.html.

DATA ACQUISITION AND METHODS

New data

During the year ending September 2007 UCSC added eight new organisms to the GBD: lizard (*A. carolinensis*), horse, platypus, medaka, stickleback and three worms (*Caenorhabditis brenneri*, *C. remanei* and *Pristionchus pacificus*). Nine existing organisms were updated with new assemblies: mouse, cow, cat, fugu, zebrafish, *Drosophila melanogaster*, two worms (*C. elegans*, *C. briggsae*) and sea urchin. As updates are added, older assemblies remain accessible either on the primary website or through the GBD archives.

UCSC Genes—the next generation of Known Genes

In April 2007 UCSC released UCSC Genes (W.J. Kent, manuscript in preparation), an improved version of the existing Known Genes annotation (11), on the March 2006 (Build 36, hg18) human assembly. This annotation, which includes putative non-coding genes as well as protein-coding genes and 99.9% of RefSeq genes, is a moderately conservative prediction set based on data from RefSeq, GenBank and UniProt (12). Each entry requires the support of one GenBank RNA sequence and at least one additional line of evidence, with the exception of RefSeq RNAs, which require no additional evidence. Although some of the transcripts labeled as 'non-coding' in the set may actually code for protein, typically the evidence for the associated protein is weak. Compared to RefSeq, this gene set generally has about 10% more protein-coding genes, approximately five times as many putative non-coding genes, and about twice as many splice variants. As part of the migration to the UCSC Genes annotation, we now use our own UCSC Genes accession numbers as the primary key into the underlying knownGene table, rather than the GenBank mRNA accessions used in previous Known Genes annotations. The base accession numbers remain stable across iterations of the data set, although the suffixes may change to reflect version updates.

A companion annotation to UCSC Genes, the Alt Events track, shows various types of alternative splicing, alternative promoter and other events that result in more than a single transcript from the same gene.

28-Species conservation

UCSC released a new Conservation (13) annotation track on the March 2006 (Build 36, hg18) human genome in June 2007. This track displays multiz (14) multiple alignments of 27 vertebrate species to the human genome, along with measurements of evolutionary conservation across all 28 species and a separate measurement of conservation across the placental mammal subset of species (18 organisms).

Included in the track are 5 new high-quality assemblies—horse, platypus, lizard, stickleback and medaka; 6 new low-coverage mammalian genomes—bushbaby, tree shrew, guinea pig, hedgehog, common shrew and cat; 6 updated assemblies—chimp, cow, chicken, frog, fugu and zebrafish; and 10 assemblies included in the previous version of the Conservation track—rhesus, mouse, rat, rabbit, dog, armadillo, elephant, tenrec, opossum and tetraodon. In addition to the expanded species list, the new Conservation track has been enhanced to include additional filtering of pairwise alignments for each species to reduce paralogous alignments and information about the quality of aligning species sequence included in the multiple alignments downloads. A similar Conservation annotation of at least 30 species is scheduled for release on the July 2007 (Build 37, mm9) mouse assembly in the last quarter of 2007.

Variation and disease

Within the Variation and Repeats annotation group, UCSC has added several new data sets. The simple nucleotide polymorphism (SNP) data from dbSNP (15) Build 126, already available on the human and mouse assemblies, has been added to the chimp, rat, and dog. Updates to SNP Build 128 will be incorporated pending data release from dbSNP. SNP annotations may be filtered by several attributes, including average heterozygosity and weight, location type, class, validation, function and molecule type. The alignments of the SNP's flanking sequences to the genome are displayed on the details page for each SNP; in addition, the hg18 SNP details pages include the chimp and rhesus macaque orthologous alleles. We have also added a HapMap SNPs annotation (16) to the hg17 and hg18 assemblies containing data for 4 million SNPs (dbSNP 125) from four populations, together with the display of orthologous alleles from chimp and rhesus macaque and several options for filtering the data display. The Structural Variation annotation has been expanded to include structural variation data (17), deletions detected by several techniques (18–20) and numerous copy number polymorphism data sets (21–25). The SNP Arrays track displays SNPs available for genotyping with several different microarrays. The Exapted Repeats annotation displays conserved non-exonic elements that have been deposited by characterized mobile elements (26).

Mapping, gene prediction, regulation and expression data

In addition to updating selected existing data sets, we have introduced several new annotations to various assemblies. High-confidence gene annotations from the Consensus Coding DNA Sequence (CCDS) project (http://www.ncbi.nlm.nih.gov/CCDS/) have been added to more human assemblies and to the mouse. The Affymetrix Transcriptome Phase 3 data set (human) (27) displays transcriptome data from tiling Affymetrix GeneChips. The ORegAnno (Open Regulatory Annotation) track (several species) shows literature-curated regulatory regions, transcription factor binding sites and regulatory polymorphisms from the ORegAnno database (28). The ACEScan annotation (human) identifies predicted alternative human-mouse conserved exons from ACEScan (29). The CGAP SAGE track (human, mouse) displays genomic mappings for human LongSAGE tags from the Cancer Genome Anatomy Project (CGAP) (30), using the Serial Analysis of Gene Expression (SAGE) quantitative technique (31). The MGI QTL track (mouse) shows approximate positions of quantitative trait loci based on reported peak LOD scores from the Jackson Laboratory Mouse Genome Informatics group. The zebrafish genome now provides expression data using the Affymetrix Zebrafish GeneChip Genome Array (32).

Mammalian Gene Collection

The track details pages for features in the Mammalian Gene Collection (MGC) (33) Genes annotation track (available for several species) now include extensive information about the MGC clones, including links for ordering the clones.

A new annotation on the hg18 human assembly, ORFeome Clones, shows alignments of human clones from the ORFeome Collaboration (http://www.orfeomecollaboration.org/) (34), a project that aims to be an unrestricted source of fully sequence-validated full-ORF human cDNA clones, with the goal of providing at least one fully sequenced full-ORF clone for each human gene. This annotation is automatically updated daily as new clones become available.

ENCODE

The Genome Browser serves as the data repository for the ENCODE (Encyclopedia of DNA Elements) project (35). The set of human genome annotations available on the UCSC ENCODE portal (http://genome.ucsc.edu/ENCODE/) (36), contributed by members of the ENCODE Consortium, has increased by 40% in the past 12 months, from 130 tracks and 950 tables on 2 assemblies in September 2006 to 199 tracks and 1583 tables on 3 assemblies in September 2007. The ENCODE data sets are now available on the March 2006 (Build 36, hg18) as well as the May 2004 (Build 35, hg17) human assembly.

VisiGene and Gene Sorter data sets

We have updated the VisiGene image sets from the Jackson Lab Mouse Genome Informatics Database and the Allen Brain Atlas, and have added links from the mouse Known Genes and UCSC Genes details pages to *X. laevis* images. The VisiGene probe-processing utilities have been updated to interact with the new UCSC Genes set.

The hg18 Gene Sorter has been expanded to include the Wanker (37), Vidal (38) and Human Protein Reference Database (39) data sets showing the neighborhood of protein interactions surrounding selected genes. The neighborhoods are computed from a genome-wide protein–protein interaction network that connects genes if the proteins they encode have been detected to physically interact in high-throughput experiments. The xxBlastTab tables, which display gene ortholog data for

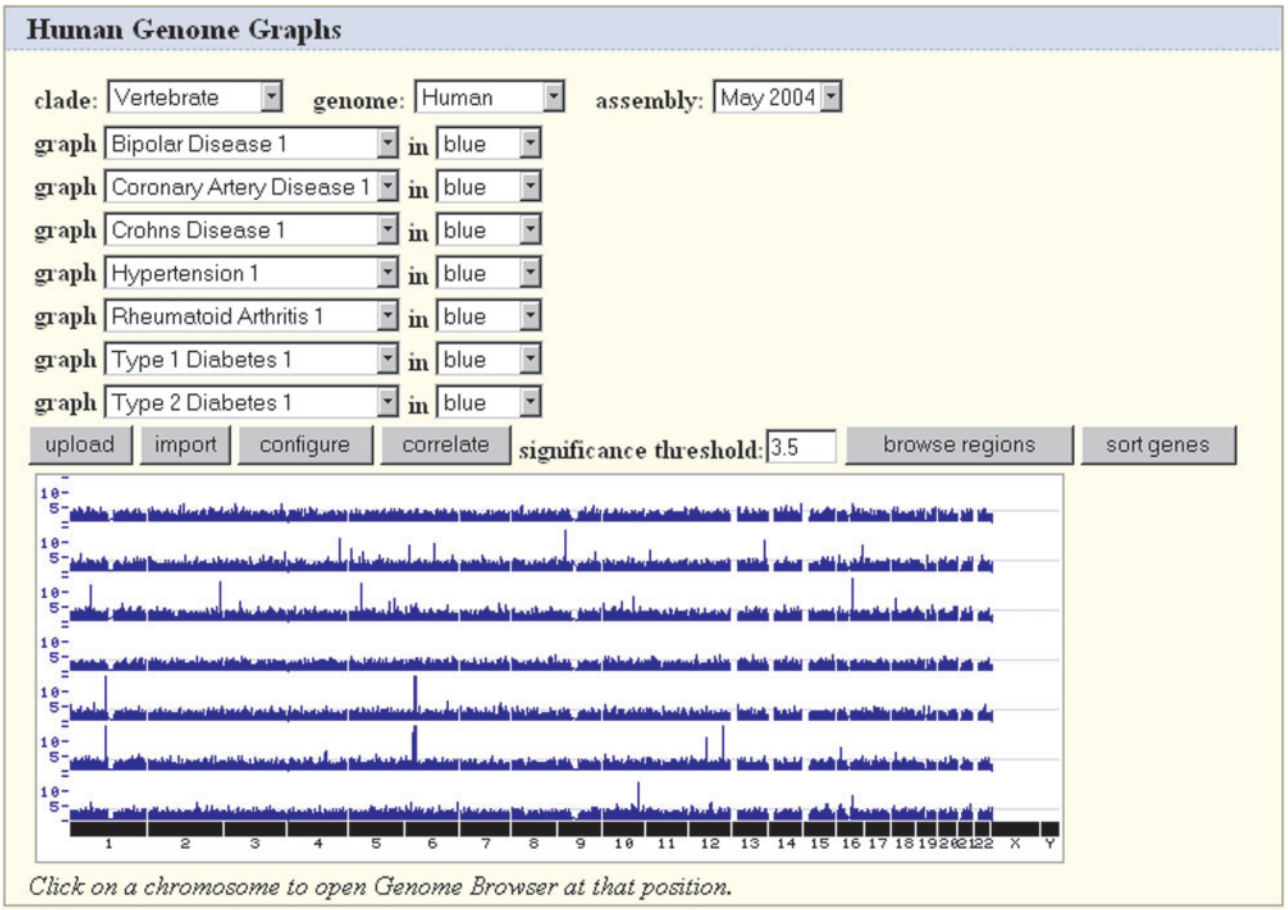

Figure 1. Genome Graphs for the human May 2004 (Build 35, hg17) assembly loaded with data published by the Wellcome Trust Case Control Consortium from a genome-wide association study of seven common diseases (40).

human, mouse and rat in the Gene Sorter and UCSC Genes details pages, have now been filtered for synteny.

NEW FEATURES

The GBD and the Genome Browser toolset are dynamic resources that continually evolve to accommodate new genome assemblies, data types and research requirements. In the past year we have expanded and enhanced many of our Genome Browser tools to improve data browsing and manipulation capabilities for our users and collaborators.

Viewing genome-wide data with Genome Graphs

Genome Graphs (http://genome.ucsc.edu/cgi-bin/hgGenome), a new tool accessible via the 'Genome Graphs' link on the GBD home page, can be used to display genome-wide data sets, for example, the results of genome-wide SNP association studies, linkage studies and homozygosity mapping. Using the Genome Graphs tool, it is possible to upload or import several sets of genome-wide data and display them simultaneously (Figure 1), then accomplish such tasks as restricting the display to only those regions that exceed a set significance threshold, displaying genes via the Gene Sorter that exist in areas where the data meet a given significance threshold, displaying an area of interest in the Genome Browser and calculating the correlation coefficient (R) among the data sets. Both public and personal data sets may be loaded into the tool, and the display can be configured to suit individual needs.

Genomewiki

We have launched a wiki site for sharing information about the UCSC Genome Browser and its data. The wiki—at: http://genomewiki.ucsc.edu—provides an informal forum for our browser users, mirror sites and staff to discuss topics of interest in the genome biology field and exchange usage tips, scripts, programs and notes about mirroring the Genome Browser and working with the Genome Browser source. As with most wiki sites, general users are welcome to edit and add pages after logging in.

Saving and sharing Genome Browser sessions

Users can now save their favorite Genome Browser sessions for reuse and sharing by using a new session management feature, accessible via the 'Session' link (http://genome.ucsc.edu/cgi-bin/hgSession) on the GBD home page and the blue navigation bar at the top of many of the tool web pages. Log-in access to the session features is controlled through the Genome Browser wiki site. Once

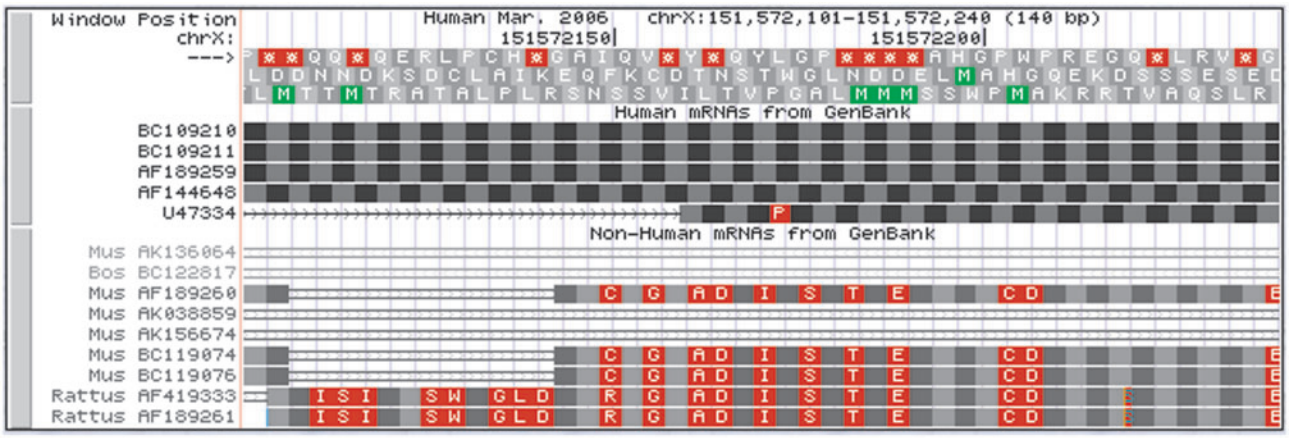

Figure 2. A zoomed-in view of the human and non-human mRNA tracks in the chrX:151 572 101–151 572 240 region on the human March 2006 (Build 36, hg18) genome assembly. In both mRNA tracks, the mRNA coloring options are configured to show nonsynonymous codon differences between the mRNA alignments and the genomic feature at the *top* of the figure. Red indicates codons that differ from the human genomic sequence. The double horizontal lines in the 'Non-Human mRNAs' track highlight areas in which both the mRNA and the genome sequence have an insertion or stretch of non-matching sequence. Note the blue vertical line at the beginning of the bottom Rattus alignment, indicating an insertion at the beginning of the query sequence, also the orange vertical lines with partial peptide insertions in the Rattus alignments.

logged in, the user can save the current Genome Browser session, including the exact position and track combination on display, share the session with another user or keep it private and load one's own saved sessions as well as those shared by others. Saved sessions persist for 1 year after the last access, unless deleted. Custom tracks within sessions persist for at least 48 h after the last time they are viewed.

Managing custom tracks

Custom annotation tracks, a popular Genome Browser feature for several years, allow users to load, display and manipulate personal data in the Genome Browser and Table Browser. The new custom tracks manager (http://genome.ucsc.edu/cgi-bin/hgCustom) makes the use of custom tracks much easier. The management interface can be accessed through the 'add/manage custom tracks' button on the Genome Browser gateway (http://genome.ucsc.edu/cgi-bin/hgGateway) or tracks (http://genome.ucsc.edu/cgi-bin/hgTracks) page. In addition to the data upload options supported in previous versions, users can now load and display multiple custom tracks simultaneously, add to, delete and modify the uploaded custom track set, load and manage tracks from multiple assemblies, and upload description pages for custom tracks. The lifespan of a custom track on the UCSC server has been increased from 8 to 48 h after last access, and we have converted the underlying custom track architecture from a file-based system to a database system to improve the performance.

New display options

We have added several new user-configurable display options to the Genome Browser that expand navigation within gene, mRNA and EST-based tracks and allow increased manipulation of the tracks image and track control groups. These options are controlled on the browser configuration page, which is accessed through the 'configure' button on the Genome Browser gateway or tracks page. When the 'Next/previous item navigation' configuration option is toggled on, gray double-headed arrows display in the Genome Browser tracks image on both sides of the track labels of gene, mRNA and EST tracks (or any standard tracks based on BED, PSL or genePred format). The image window may be shifted to display the next track feature towards the 5' or 3' end of the chromosome by clicking the corresponding left or right arrow. Similarly, the 'Next/previous exon navigation' configuration option displays white double-headed arrows on the both 5' and 3' end of each track item that has exons positioned beyond the edges of the current image. Clicking on one of the arrows shifts the image window to the next exon located towards that end of the feature. Another new configuration option—'Enable track reordering'—allows the user to change the display order of the track groups as well as the order of annotation tracks within the groups, and to move tracks between track groups. This is particularly useful for customizing the browser display to individual research needs or for creating images for publication. Track groups on the Genome Browser tracks page may now be quickly collapsed and expanded by clicking the + and − icons on the left side of the group label.

We have expanded the coloring and display options for mRNA tracks to include several ways to highlight gaps in alignments of query sequences (usually transcripts) to the genome, which frequently indicate a problem with the query sequence or with the genome assembly. Tracks may be colored by genomic, mRNA or nonsynonymous mRNA codons, mRNA bases, or different mRNA bases. Tracks may then be configured, through options on the description page, to display double horizontal lines at locations where both the genome and query sequence have an insertion and to display vertical lines of different colors to distinguish poly(A) tail insertions and insertions at the beginning, end or middle of the query (Figure 2). The new

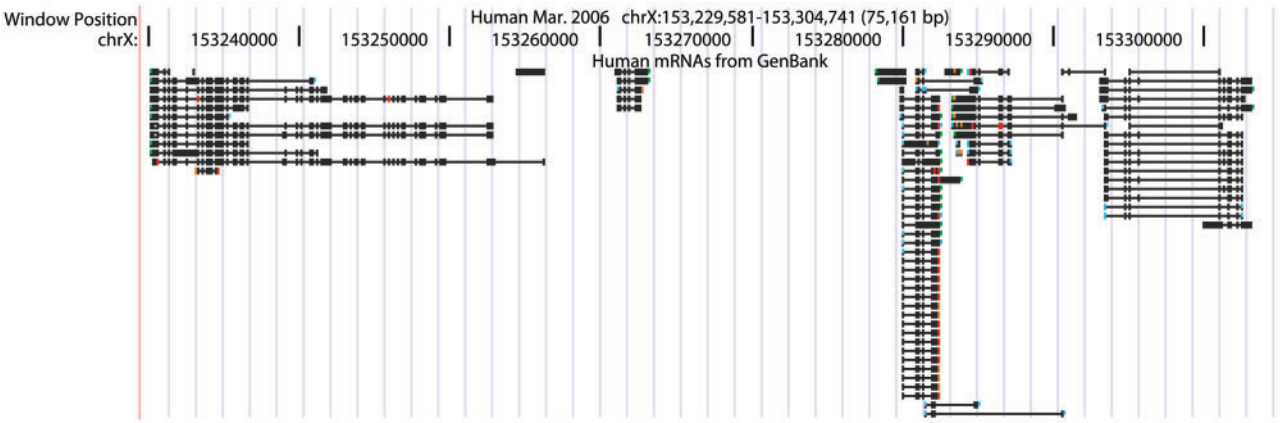

Figure 3. A zoomed-out view of human mRNA alignments in the chrX:153 229 581–153 304 741 region using 'squish' display mode and configured to show nonsynonymous codon differences between the human mRNAs and the genomic sequence. This view is useful for quickly scanning for mRNAs that are free of nonsynonymous regions (i.e. are all-black in color) and have a valid poly(A) tail (green vertical bar).

coloring scheme makes it easier to visually scan a region with hundreds of alignments and pick out regions of interest (Figure 3). The new display options are explained in detail on the mRNA track description pages.

To visually simplify the display of the large number of similar annotation tracks present in the ENCODE track groups, collections of related tracks are now represented by a single 'super-track' control that provides a descriptive overview of the group and lets the user control display characteristics of the entire track set on one page. For example, the Yale ChIP-chip super-track control provides information and access for seven related Yale ChIP-chip annotations.

FUTURE DIRECTIONS

In the upcoming year, UCSC will continue to extend the GBD to include more species and assembly updates—focusing on primates, model organisms and species of critical importance to evolutionary studies—and more annotation data. We may also provide browser access to several of the low-coverage (2×) assemblies currently included in our Conservation tracks. Following the release of UCSC Genes data for the latest mouse assembly in Fall 2007, updates to both the human and the mouse UCSC Genes annotation will be offered ~3–4 times per year. We will continue to expand our collection of human variation, disease-related, expression and genome-wide association data, and plan to explore the incorporation of federated data into the browser. Among the enhancements to our display and data-mining tools, we plan to facilitate the use of custom tracks in the Table Browser, extend display features such as the next/previous item navigation to a larger range of tracks, and add a user-annotated wiki track.

ACKNOWLEDGEMENTS

We would like to thank the many collaborators who have contributed data to our project, our Scientific Advisory Board for their valuable advice and recommendations, and our users for their feedback and support. We would also like to acknowledge the dedicated system administrators who have provided an excellent computing environment: Jorge Garcia, Erich Weiler, Chester Manuel and Victoria Lin. This work was funded by National Human Genome Research Institute (2 P41 HG002371-06 to UCSC Center for Genomic Science, 3 P41 HG002371-06S1 ENCODE supplement to UCSC Center for Genomic Science); National Cancer Institute (Contract No. N01-CO-12400 for Mammalian Gene Collection). TW is a Helen Hay Whitney fellow. Funding to pay the Open Access publication charges for this article was provided by the Howard Hughes Medical Institute.

Conflict of interest statement. D.K., R.M.K., R.B., G.P.B, H.C., M.D., R.A.H., A.S.H., F.H., A.P., B.J.R., B.R., K.R.R., K.E.S., A.T., H.T., A.S.Z., D.H., and W.J.K receive royalties from the sale of UCSC Genome Browser source code licenses to commercial entities.

REFERENCES

1. Kuhn,R.M., Karolchik,D., Zweig,A.S., Trumbower,H., Thomas,D.J., Thakkapallayil,A., Sugnet,C.W., Stanke,M., Smith,K.E. *et al.* (2007) The UCSC Genome Browser database: update 2007. *Nucleic Acids Res.*, **35**, D668–D673.
2. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
3. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
4. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2007) GenBank. *Nucleic Acids Res.*, **35**, D21–D25.
5. Hubbard,T.J.P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
6. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, D5–D12.
7. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC

Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
8. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
9. Kent,W.J., Hsu,F., Karolchik,D., Kuhn,R.M., Clawson,H., Trumbower,H. and Haussler,D. (2005) Exploring relationships and mining data with the UCSC Gene Sorter. *Genome Res.*, **15**, 737–741.
10. Hsu,F., Pringle,T.H., Kuhn,R.M., Karolchik,D., Diekhans,M., Haussler,D. and Kent,W.J. (2005) The UCSC Proteome Browser. *Nucleic Acids Res.*, **33**, D454–D458.
11. Hsu,F., Kent,W., Clawson,H., Kuhn,R., Diekhans,M. and Haussler,D. (2006) The UCSC Known Genes. *Bioinformatics*, **22**, 1036–1046.
12. The UniProt Consortium. (2007) The Universal Protein resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D197.
13. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm and yeast genomes. *Genome Res.*, **15**, 1034–1050.
14. Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.
15. Sherry,S.T, Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
16. The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
17. Feuk,L., Carson,A.R. and Scherer,S.W. (2006) Structural variation in the human genome. *Nat. Rev. Genet.*, **7**, 85–97.
18. Conrad,D.F., Andrews,T.D., Carter,N.P., Hurles,M.E. and Pritchard,J.K. (2006) A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.*, **38**, 75–81.
19. Hinds,D.A., Kloek,A.P., Jen,M., Chen,X. and Frazer,K.A. (2006) Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.*, **38**, 82–85.
20. McCarroll,S.A., Hadnott,T.N., Perry,G.H., Sabeti,P.C., Zody,M.C., Barrett,J.C., Dallaire,S., Gabriel,S.B., Lee,C. *et al.* (2006) Common deletion polymorphisms in the human genome. *Nat. Genet.*, **38**, 86–92.
21. Iafrate,A.J., Feuk,L., Rivera,M.N., Listewnik,M.L., Donahoe,P.K., Qi,Y., Scherer,S.W. and Lee,C. (2004) Detection of large-scale variation in the human genome. *Nat Genet.*, **36**, 949–951.
22. Locke,D.P., Sharp,A.J., McCarroll,S.A., McGrath,S.D., Newman,T.L., Cheng,Z., Schwartz,S., Albertson,D.G., Pinkel,D. *et al.* (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.*, **79**, 275–290.
23. Redon,R., Ishikawa,S., Fitch,K.R., Feuk,L., Perry,G.H., Andrews,T.D., Fiegler,H., Shapero,M.H., Carson,A.R. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.
24. Sebat,J., Lakshmi,B., Troge,J., Alexander,J., Young,J., Lundin,P., Maner,S., Massa,H., Walker,M. *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.
25. Sharp,A.J., Locke,D.P., McGrath,S.D., Cheng,Z., Bailey,J.A., Vallente,R.U., Pertz,L.M., Clark,R.A., Schwartz,S. *et al.* (2005) Segmental duplications and copy number variation in the human genome. *Am. J. Hum. Genet.*, **77**, 78–88.
26. Lowe,C.B., Bejerano,G. and Haussler,D. (2007) Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc. Natl Acad. Sci. USA.*, **104**, 8005–8010.
27. Cheng,J., Kapranov,P., Drenkow,J., Dike,S., Brubaker,S., Patel,S., Long,J., Stern,D., Tammana,H. *et al.* (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, **308**, 1149–1154.
28. Montgomery,S.B., Griffith,O.L., Sleumer,M.C., Bergman,C.M., Bilenky,M., Pleasance,E.D., Prychyna,Y., Zhang,X. and Jones,S.J. (2006) ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics*, **22**, 637–640.
29. Yeo,G.W., Van Nostrand,E., Holste,D., Poggio,T. and Burge,C.B. (2005) Identification and analysis of alternative splicing events conserved in human and mouse. *Proc. Natl Acad. Sci. USA*, **102**, 2850–2855.
30. Riggins,G.J. and Strausberg,R.L. (2001) Genome and genetic resources from the Cancer Genome Anatomy Project. *Hum. Mol. Genet.*, **10**, 663–667.
31. Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
32. Weber,G.J., Choe,S.E., Dooley,K.A., Paffett-Lugassy,N.N., Zhou,Y. and Zon,L.I. (2005) Mutant-specific gene programs in the zebrafish. *Blood*, **106**, 521–530.
33. Gerhard,D.S., Wagner,L., Feingold,E.A., Shenmen,C.M., Grouse,L.H., Schuler,G., Klein,S.L., Old,S., Rasooly,R. *et al.* (2004) The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res.*, **14**, 2121–2127.
34. Collins,J.E., Wright,C.L., Edwards,C.A., Davis,M.P., Grinham,J.A., Cole,C.G., Goward,M.E., Aguado,B., Mallya,M. *et al.* (2004) A genome annotation-driven approach to cloning the human ORFeome. *Genome Biol.*, **5**, R84.
35. The ENCODE Project Consortium *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
36. Thomas,D.J., Rosenbloom,K.R., Clawson,H., Hinrichs,A.S., Trumbower,H., Raney,B.J., Karolchik,D., Barber,G.P., Harte,R.A. *et al.* (2007) The ENCODE project at UC Santa Cruz. *Nucleic Acids Res.*, **35**, D663–D667.
37. Stelzl,U., Worm,U., Lalowski,M., Haenig,C., Brembeck,F.H., Goehler,H., Stroedicke,M., Zenkner,M., Schoenherr,A. *et al.* (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.
38. Rual,J.F., Venkatesan,K., Hao,T., Hirozane-Kishikawa,T., Dricot,A., Li,N., Berriz,G.F., Gibbons,F.D., Dreze,M. *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178.
39. Mishra,G.R., Suresh,M., Kumaran,K., Kannabiran,N., Suresh,S., Bala,P., Shivakumar,K., Anuradha,N., Reddy,R. *et al.* (2006) Human protein reference database—2006 update. *Nucleic Acids Res.*, **34**, D411–D414.
40. The Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.