D590–D598 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj144

The UCSC Genome Browser Database: update 2006

A. S. Hinrichs*, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey[2], R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler[1] and W. J. Kent

Center for Biomolecular Science and Engineering, School of Engineering and [1]Howard Hughes Medical Institute University of California Santa Cruz (UCSC), Santa Cruz, CA 95064, USA and [2]Institute for Genome Sciences and Policy, Duke University, Durham, NC 27708, USA

Received September 15, 2005; Revised and Accepted October 27, 2005

ABSTRACT

The University of California Santa Cruz Genome Browser Database (GBD) contains sequence and annotation data for the genomes of about a dozen vertebrate species and several major model organisms. Genome annotations typically include assembly data, sequence composition, genes and gene predictions, mRNA and expressed sequence tag evidence, comparative genomics, regulation, expression and variation data. The database is optimized to support fast interactive performance with web tools that provide powerful visualization and querying capabilities for mining the data. The Genome Browser displays a wide variety of annotations at all scales from single nucleotide level up to a full chromosome. The Table Browser provides direct access to the database tables and sequence data, enabling complex queries on genome-wide datasets. The Proteome Browser graphically displays protein properties. The Gene Sorter allows filtering and comparison of genes by several metrics including expression data and several gene properties. BLAT and In Silico PCR search for sequences in entire genomes in seconds. These tools are highly integrated and provide many hyperlinks to other databases and websites. The GBD, browsing tools, downloadable data files and links to documentation and other information can be found at http://genome.ucsc.edu/.

INTRODUCTION

As the volume and variety of public sequence and annotation data expand with no end in sight, tools that help the biologist search, view, organize and retrieve public data become ever more useful. The University of California Santa Cruz (UCSC) Genome Browser Database (GBD) (1) and a family of tools for accessing this database, all online at http://genome.ucsc.edu/, provide both interactive and bulk-download access to sequence and annotations for dozens of species, featuring the human genome and several major model organisms. The GBD is also tightly integrated with the MySQL databases underlying the UCSC Proteome Browser (2). The GBD is a collection of related MySQL databases: one database for each genome assembly version supported and several databases containing shared auxiliary information. Within a genome assembly database, data are organized into annotation 'tracks', or collections of annotations anchored to genomic positions and usually accompanied by additional descriptive information. Each genome assembly database contains a wealth of annotation tracks describing the assembly, genomic sequence characteristics, predicted genes, cDNA evidence, repetitive sequences, cross-species homologies, and when available, variation data, expression data, curated gene annotations and more. Although the majority of the annotation tracks are computed at UCSC using public software from UCSC and other institutions, many of the annotations are contributed by collaborators.

UCSC's web-based tools for accessing the GBD are summarized in Table 1. These tools are highly integrated for seamless searching and visualization of annotations; when using the Genome Browser, the Table Browser is only a

*To whom correspondence should be addressed. Tel: +1 831 459 1544; Fax: +1 831 459 1809; Email: angie@soe.ucsc.edu
Present address:
C. W. Sugnet, Affymetrix Inc., Santa Clara, CA 95051, USA

Table 1. Tools available on http://genome.ucsc.edu/ for accessing the GBD

Tool	Description
Genome Browser	Visual display of genome annotations
Table Browser	Flexible, powerful querying of genome and proteome data
Gene Sorter	Gene-focused tool for identifying related genes
Proteome Browser	Visual display of protein properties
BLAT	Rapid search for nucleotide or protein sequence in genome
In Silico PCR	Rapid search for products of primer pairs based on BLAT engine

click away and so on. When possible, links to other databases and websites are provided as well.

Since the time of publication of papers describing the GBD and Table Browser (3) in prior database issues of this journal, many new species and annotations have been added to the GBD, many improvements have been made to the Genome Browser (4) and Table Browser tools and new tools have been created. This paper focuses on additions and enhancements since the Table Browser paper was submitted in September 2003.

NEW DATA AND DATA TYPES

Additional species supported

In the past two years, the GBD has expanded to include many new species: chimp, Rhesus macaque, dog, cow, opossum, chicken, frog (*Xenopus tropicalis*), fugu, zebrafish, Tetraodon, fruitfly (*Drosophila melanogaster* and six other *Drosophila* species), honeybee (*Apis mellifera*), mosquito (*Anopheles gambiae*), sea squirt (*Ciona intestinalis*) and yeast (*Saccharomyces cerevisiae*). For some species, multiple assembly versions have been incorporated into the GBD. We generally provide the two most recent assemblies on the main site (three most recent for the human genome) and move the older assemblies to an archive server (http://genome-archive.cse.ucsc.edu/). The sequencing centers, organizations and individuals who contributed to the assemblies and annotations are acknowledged at http://genome.ucsc.edu/goldenPath/credits.html.

Genome-wide multi-species alignments and conservation scores

Genome-wide multi-species alignment data are now available in the GBD for human, mouse, dog, chicken, zebrafish and fly. The alignments can be downloaded from http://hgdownload.cse.ucsc.edu/downloads.html as Multiple Alignment Format (MAF) files. UCSC generates pairwise cross-species alignments using blastz from the Miller lab at Pennsylvania State University (5) and then distills the blastz alignments to a best-in-genome set using UCSC's axtChain and chainNet tools (6). Multiple pairwise best-in-genome alignments are combined into a multi-species alignment using multiz, also from the Miller lab (7). Conservation scores and conserved regions are computed by UCSC's phastCons (8) using these multiple alignments. A protocol for using the conserved regions and other GBD annotations and tools in a computational screen for non-coding functional elements is described in a recent publication (9).

UCSC Known Genes

The UCSC Known Genes dataset (submitted), currently available in the human, mouse and rat browser databases, is created by a fully automated process that combines protein and mRNA evidence. Proteins from UniProt (10) are paired with associated mRNA sequences from GenBank (11). The mRNA sequences are aligned to the genome and best scoring mRNAs are retained as Known Genes. The Known Genes are heavily cross-referenced to external database identifiers and symbolic names and also serve as a foundation for the UCSC Gene Sorter (12) and Proteome Browser.

Recently the Known Genes process has been updated to achieve higher quality and higher coverage. Candidate genes are filtered much more stringently, discounting alignments that contain frame shift errors or in-frame stop codons. This reduces the number of genes but increases the quality of the gene set. Other areas of improvement are inclusion of RefSeq (13) mRNAs as initial candidates, pairing of proteins and mRNAs, coverage of splice isoforms, weighting of RefSeq and Mammalian Gene Collection (MGC) (14) genes and identification of coding DNA sequence (CDS) regions within mRNA sequences.

The Genome Browser's details page for each known gene contains a wealth of collected information and references, organized into several sections. The quick links section provides many hyperlinks to views of the gene both in UCSC tools and on other websites. When available, descriptions from RefSeq and UniProt are quoted. The sequence section contains links to the mRNA, protein and genomic sequence of the gene. Microarray expression data are displayed graphically when available. Links to Interpro (15) and Pfam (16) provide protein domain and structure information. Known or predicted 3D structures from PDB (17) or ModBase (18) are displayed, with links to those sites for more information. The homolog section contains links to homologs identified by best protein sequence match using blastp. The Gene Ontology (GO) (19) section lists the molecular function, biological process and cellular component terms associated with the gene. The GO terms are retrieved from a local copy of the 'go' database downloaded from http://www.geneontology.org/ (20), to which UCSC adds a table 'goaPart' containing gene associations for supported organisms. The displayed GO terms are linked back to specific queries to the AmiGO website (http://www.godatabase.org/cgi-bin/amigo/go.cgi). The associated mRNA section lists all mRNA sequences whose alignments overlap the gene by at least 12 bases. The pathway section provides links to KEGG (21), NCI/BioCarta (http://www.biocarta.com/), BioCyc (22) and Reactome (23). The Known Genes details pages are also available in the Gene Sorter.

Additions to the Proteome Browser

The Proteome Browser has been updated to support direct access via the Proteome Browser gateway, eliminating the multiple Genome Browser steps required by the previous release. The Proteome Browser now includes proteins of all organisms in the UniProt databases, rather than just the human, mouse and rat protein sets. The databases underlying the UCSC Proteome Browser, Swiss-Prot and proteins, have been renamed as uniProt and proteome respectively as part

of the effort to accommodate UniProt's recent change of its display IDs. Several tables containing information about splice variants and links to Reactome have been added to those databases.

Cross-species protein alignments

Several vertebrate model organisms have a Human Proteins track created as follows. First, human known gene proteins are aligned to the human genome using BLAT (24), in order to identify exon boundaries in the proteins. Then the amino acid sequences of the putative exons are aligned to the model organism's genome using tblastn (25). The exon alignments are chained together where possible. Finally, chained alignments are filtered to retain the single best chain covering at least 60% of the protein query, as well as all chains covering at least 60% of the protein query with at least 60% amino acid identity. Similarly for insect genomes, *D.melanogaster* proteins are aligned to other insects. These chained exonic protein alignments are especially useful in newly sequenced genomes that do not yet have well-developed gene annotations of their own.

ENCODE pages and tracks

UCSC serves as the central repository for genome annotations and genomic data generated by the Encyclopedia of DNA Elements (ENCODE) Consortium (26). The ENCODE pilot phase, in which massive amounts of experimental data were generated on carefully selected regions totaling 1% of the human genome, resulted in the creation of >50 new data tracks consisting of >400 new data tables in the human browser database. The ENCODE tracks include gene annotations, transcription levels, chromatin immunoprecipitation, chromosome features, multiple alignments and conservation of 23 vertebrate species using resequenced regions from the NIH Intramural Sequencing Center (http://www.nisc.nih.gov/), variation data and analyses derived from the original datasets. In each of those categories, multiple experimental methods were used to provide deep coverage and enable comparison of different platforms.

The portal to ENCODE annotations at UCSC is http://genome.ucsc.edu/ENCODE/. It provides general information about the projects, links to NHGRI and other ENCODE sites and also links to pages that simultaneously display both a clickable list of ENCODE regions and a Genome Browser view of the currently selected region.

Other new gene annotations

The MGC is a trans-NIH initiative intending to provide full-length open reading frame clones for human, mouse and rat genes. The GBDs for human, mouse and rat have an MGC Genes track with links to the IMAGE database, through which clones can be ordered, and links to other external databases in addition to the usual sequence and alignment information provided along with mRNA alignments. The zebrafish GBD has a ZGC Genes track with the Zebrafish Gene Collection, a subproject of MGC.

Another significant addition to UCSC's collection of human gene annotations is obtained from the Consensus CDS (CCDS) project (http://www.ncbi.nlm.nih.gov/CCDS/), a collaboration of the European Bioinformatics Institute, the National Center for Biotechnology Information, the Wellcome Trust Sanger Institute and UCSC. CCDS is a high-quality, consistently annotated core set of human protein-coding genes identified by consensus among several sets of annotations: Ensembl (27), Vega (28) and RefSeq. To be included in the CCDS set, coding regions must have identical CDS genomic coordinates in both RefSeq and Ensembl/Vega, must be full-length (beginning with an ATG start codon and ending with a valid stop-codon), must be free of frame shifts, must not overlap with predicted pseudogenes, must have supporting transcripts and protein homology and must use consensus splice sites. Stable, versioned identifiers are assigned to CCDS regions.

Other recent additions of contributed gene annotations or predictions in the human Genome Browser include Vega Genes, Vega Pseudogenes, ECgene (29), Twinscan (30), SGP (31), Geneid (32), Augustus (33), Yale Pseudogenes (34) and Superfamily (35). UCSC's Retroposed Genes track shows processed mRNAs that have been inserted back into the genome since the mouse/human split, including functional genes that have acquired a promoter from a neighboring gene, non-functional pseudogenes and transcribed pseudogenes.

Several genome databases include contributed non-protein-coding RNA gene annotations. For example, the sno/miRNA track in the human Genome Browser contains combined annotations from the miRNA Registry (36) and snoRNA-LBME-DB (37). For mouse, there is a miRNA track with annotations from the miRNA Registry. The EvoFold track in human shows RNA secondary structure predictions made with the EvoFold program (submitted), a comparative method that exploits the evolutionary signal of genomic multiple-sequence alignments for identifying conserved functional RNA structures.

Several gene tracks are imported from major model organism databases: FlyBase (38) genes for *D.melanogaster*, WormBase (39) genes for *C.elegans* and the *Saccharomyces* Genome Database (40) for *S.cerevisiae*.

New expression data

The GBDs for human and mouse include expression data from the Genomics Institute of the Novartis Research Foundation (GNF). The GNF Ratio tracks in human and the GNF U74A, B and C tracks in mouse graphically display the expression values measured in GNF's 2002 study of 91 human and mouse samples (41). The GNF Atlas 2 tracks in human and mouse show GNF's 2004 study of 140 human and mouse samples (42).

The human Genome Browser displays the Affymetrix Transcriptome Project Phase 2 (43) data in the Affy Txn Phase2 track. For the 10 chromosomes 6, 7, 13, 14, 19, 20, 21, 22, X and Y, >74 million 25 bp probes were tiled every 5 bp in non-repeat-masked areas and hybridized to mRNA from 11 different cell lines. The track displays both probe values and 'transfrags' (transcribed fragments) for the 11 cell lines.

The mouse Genome Browser contains data from a study of sex-specific expression in five adult mouse tissues by Rinn *et al*. (44) in the Rinn Sex Exp track.

In the human, mouse, rat and zebrafish Genome Browsers, several tracks do not have associated expression data but provide alignments of consensus/exemplar sequences from several Affymetrix chips to the genomes of human (Affymetrix chips U133, GNF1H, U95), mouse (GNF1M,

U74, MOE430), rat (RG-U34A, RAE230) and zebrafish (Zebrafish). A user can enter a probe ID from a supported chip as a position/search term in the Genome Browser in order to view a probe's genomic context.

The human Genome Browser includes two other new tracks related to expression: First Exon Finder (FirstEF) and Transcription Factor Binding Site (TFBS) conserved. FirstEF shows exon, promoter and CpG window predictions from the FirstEF program (45). The TFBS conserved track contains the location and score of TFBSs conserved in the human/mouse/rat alignment. A binding site is considered to be conserved across the alignment if its score meets a threshold score for its binding matrix in all the three species. matrices are taken from the TRANSFAC database (46).

New variation data

The Simple Nucleotide Polymorphisms (SNPs) track in the human Genome Browser displays simple nucleotide polymorphisms (SNPs as well as small insertions and deletions) from dbSNP (47) and several commercially available genotyping arrays. The track control page for SNPs offers extensive filtering and coloring options to restrict and annotate the display. The user can choose to exclude or color SNPs based on source, molecule type, variant class, validation status, functional class or location type.

The Segmental Dups track, contributed by the Eichler lab at the University of Washington, shows putative genomic duplications in the human genome (48). A new track in human (July 2003 assembly) shows putative copy number polymorphisms collected from four separate studies (49–52). The chicken Genome Browser has SNPs contributed by the Beijing Genomics Institute.

NEW VISUALIZATION FEATURES

Beyond the box: 'wiggle' tracks display continuous-valued data

A new data type allows the storage of one numeric value per base pair position, enabling a graphical display much like a bar chart or a continuous-valued signal across the genome. This data type is called 'wiggle' because of the appearance of the visual display of this data type in the Genome Browser. Numeric values are compressed to spare disk space and time, with a loss of information no greater than that of pixelation in the visual display. Values are stored in binary files that are indexed by database tables; therefore, the values cannot be retrieved from database tables alone. The Table Browser can retrieve values, returning them in a plain text format.

Track control pages for wiggle tracks offer the user a variety of controls over scaling and visual display of the values. The display mode can be set to 'full' for an indication of values by height, or to 'dense' for an indication of values by shade. When in full mode, the graph can be displayed as a bar chart or a sequence of points. A line may be drawn at $y = 0$, an additional line may be drawn at a specified y value and the height and scaling of the values can be adjusted. When zoomed out such that multiple data points must share the same pixel column in the display, the user can choose one of several windowing functions to determine how values are combined for display. The signal may be smoothed within a specified window of 2–16 pixel columns.

Some examples of new tracks that use the wiggle data type are Quality (available when an assembly is released with quality score files), the scores component of the Conservation track in Figure 1 and the GC Percent (GC composition in 5-base windows) track in Figure 2.

Conservation: juxtaposed multi-species alignments and conservation scores

The Conservation track, available for many of the vertebrates as well as *C.elegans*, *D.melanogaster* and *S.cerevisiae*, combines multiple species alignment and per-base-pair conservation scores computed by phastCons and shown in an integrated visual display with special features to highlight differences and gaps at the base pair level. Figure 1 shows a base-level view of the Conservation track. The track controls page for the Conservation track offers all the wiggle display controls for the conservation scores, as well as controls governing the alignment display. Pairwise alignments can be hidden or displayed. Triplets of genomic bases can be highlighted by alternating colors. In the base-level display, bases identical to the reference may be displayed as dots and special marks may be enabled to indicate unaligned bases with a spanning chained alignment.

Chromosome ideogram

For the human, mouse, rat and *D.melanogaster* genomes, a chromosome ideogram displaying cytological staining patterns can be displayed above the main image with a red box indicating the currently viewed region of the chromosome.

New dynamically computed tracks

Two new Genome Browser tracks, Restriction Enzymes and Short Match, are dynamically computed for display only, rather than retrieved from the database; therefore, they are not available from the Table Browser. This dynamic approach is taken because the storage requirements would be prohibitive if the data were precomputed and stored. The Restriction Enzymes track displays target sites for restriction enzymes described in REBASE (53). When the viewing region is zoomed in to the base-level, the restriction sites are displayed with tags showing the cutsites and overhang, with ambiguous bases shown in color (Figure 1). The track control page for Short Match allows the user to input a 2–30 base sequence; the track then displays exact matches of that sequence within the current viewing region. Due to the computational expense and density of these tracks, the Restriction Enzymes track progressively limits the set of enzymes that it aligns when viewing very large regions, up to a maximum of 250 000 bases at which no restriction sites are displayed. The Short Match track limits itself to 1 000 000 matches within the current viewing region.

Enhancements to user custom track support

User custom tracks can be submitted in the new wiggle format for graphical display of numeric data. In addition to http URLs, ftp URLs are now accepted as sources of data. For websites with password protection, the URL format http://username:password@site/ can be used. Individual

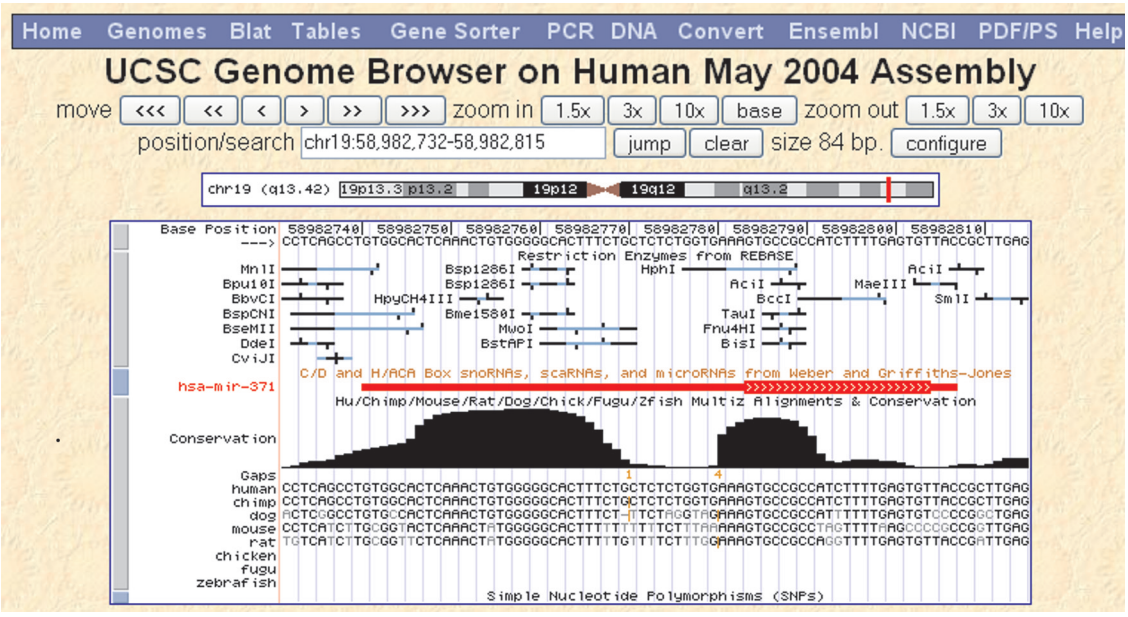

Figure 1. Genome Browser zoomed in to base-level view, showing Base Position, Restriction Enzymes, sno/miRNA, Conservation and SNP tracks at chr19:58, 982, 732-58, 982, 815 in the May 2004 assembly of the human genome. At this location is microRNA hsa-mir-371, portions of which are found to be highly conserved in other mammals by phastCons (8). In the Conservation track, gaps in the human sequence with respect to other species in the alignment are indicated by small orange vertical bars between bases in the other species, with gap lengths indicated by orange numbers in the 'Gaps' row.

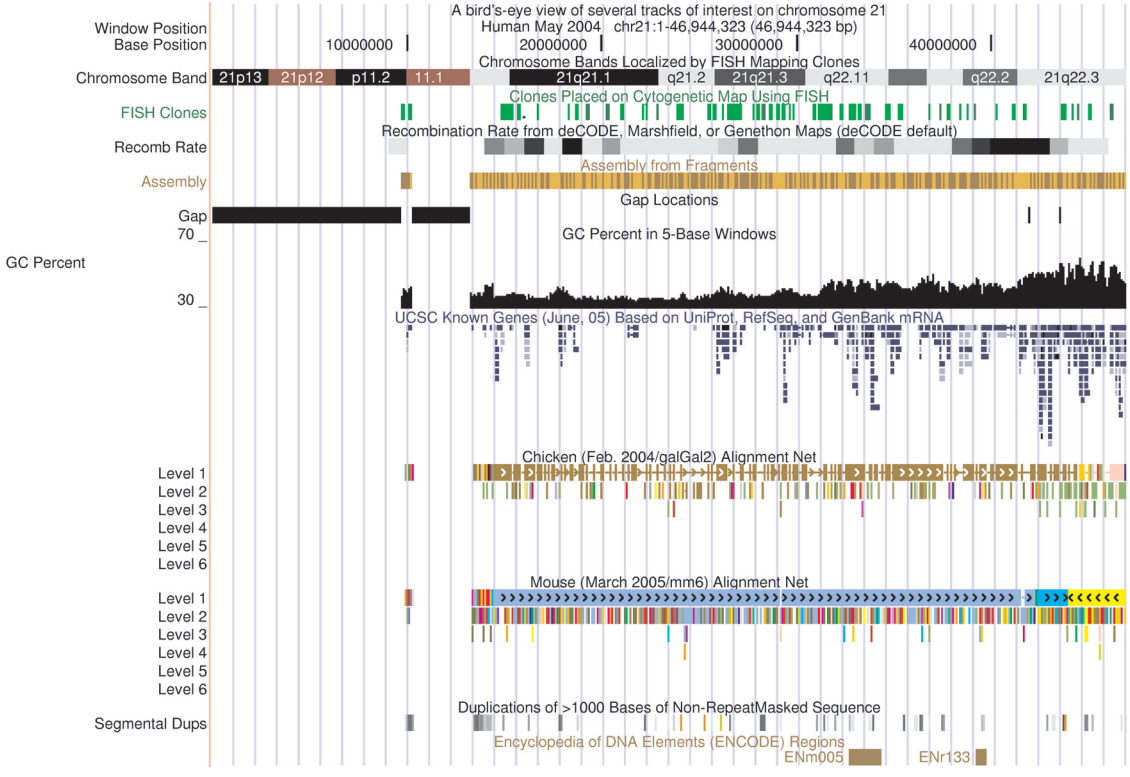

Figure 2. Genome Browser, zoomed out to view all of human chromosome 21 in the May 2004 assembly. A title line and assembly/position line have been added using the Base Position track's new label options. Large gaps show the location of unsequenced heterochromatin. Most tracks are in 'dense' display mode, so that all items are condensed into a single row. The Known Genes track is in 'squish' mode so that individual features are drawn at half-height; this gives an idea of gene density when viewing large regions. The cross-species Net tracks are in 'full' mode so that the hierarchy of alignments to each other species is clear. Human chromosome 21 contains two ENCODE regions (shown in 'pack' mode), in which dozens of additional tracks are available. This image is not a screen snapshot but rather a PostScript image generated by the Genome Browser.

items within custom tracks formatted as Browser Extensible Data can be assigned arbitrary colors by including R, G, B color values in the previously reserved ninth column and including the keyword 'itemRgb=on' in the track description line. Custom track file format details are provided at http://genome.ucsc.edu/goldenPath/help/customTrack.html. The 'Custom Tracks' link on the Genome Browser home page leads to a collection of public custom tracks contributed by various groups. As before, the Table Browser can generate a custom track from the results of a query, either exporting the file or directly importing the track into the user's session for subsequent viewing in the Genome Browser or further querying (including intersection with other tables) in the Table Browser.

UCSC hosts a collection of public custom tracks submitted by Genome Browser users at http://genome.ucsc.edu/goldenPath/customTracks/custTracks.html. Users who would like to share their annotations are encouraged to send custom track files to UCSC.

New display modes

In addition to the previously described display modes 'hide', 'dense' and 'full', many tracks also support two new display modes. 'Pack' mode draws multiple items and their labels on the same horizontal levels when there is room for them to fit side-by-side. 'Squish' mode also packs items side-by-side when possible but saves even more space by omitting labels and drawing items at half-height. Figure 2 contains examples of these two new modes.

New labeling options in gene annotations, mRNA and Base Position tracks

A new option in the track controls page of each protein-coding gene annotation track enables coloring and labeling of amino acids translated from genomic codons, drawn on top of a gene's exons, when zoomed in sufficiently. In mRNA alignment tracks, the alignments can be labeled by genomic codons, mRNA codons, mRNA bases, mRNA codons that differ from genomic codons or mRNA bases that differ from genomic bases. The Base Position track has been enhanced to show amino acids from three frames of translation, when in 'full' display mode and zoomed in sufficiently. Stop codons are highlighted in red and methionine or start codons are highlighted in green. A small arrow to the left of the base values points to the right when viewing base values on the forward strand and to the left when viewing base values on the reverse strand. Clicking on this arrow toggles the strand. The Base Position track controls page now includes options for placing a title, assembly version and/or position at the top of the Genome Browser tracks image (shown in Figure 2).

New PostScript or PDF image generation

The Genome Browser can now generate a publication-quality image in PostScript or PDF formats, exporting a file which can be saved locally. These vector-graphics image formats do not have the grainy appearance of screen snapshots. Figure 2 was generated using this new feature.

Tracks organized by type

Due to the large number of annotation tracks now available (144 for the human July 2003 assembly as of September 2005), track controls are grouped into several categories ('track groups') to ease the task of looking for a track of interest: Mapping and Sequencing, Genes and Gene Predictions, mRNA and expressed sequence tag (EST), Expression and Regulation, Comparative Genomics and Variation and Repeats. For those human assemblies with dozens of ENCODE tracks, additional ENCODE-specific categories partition the tracks into categories of genes, transcript levels, chromatin immunoprecipitation, chromosome structure, comparative genomics and variation.

Track configuration page

A new 'configure' button found on the Genome Browser gateway page and main page (shown in Figure 1) leads to a track configuration page that provides brief descriptions and visibility controls for all tracks, with links to track control pages that provide even more configuration options. To enhance the speed of the Genome Browser display, the section of track controls beneath the image can be turned off using the configuration page.

Composite tracks

Some datasets contain results of multiple parallel experiments; rather than create a separate track for each experiment in such a dataset, we create a single 'composite track' with unified controls. Individual experiments can be selected for display. When a composite track consists of >20 subtracks, extra buttons are added to the controls so that groups of related subtracks can be selected or deselected. For example, the Affy pVal track in the human Genome Browser (ENCODE Chromatin Immunoprecipitation section) consists of 41 subtracks, one for each combination of cell type and time point. Its controls include buttons for selecting all subtracks of each type of cell type and each time point.

NEW TABLE BROWSER INTERFACE AND FUNCTIONALITY

The Table Browser was redesigned and reimplemented in 2004 to offer new functionality and to improve the user interface. The most extensive new enhancement to the Table Browser is the support for queries on related tables in the database. A new button, 'describe table schema', leads to a page that lists tables related to the currently selected table, in addition to describing the columns and content of the currently selected table. When a user edits a filter for a query, the Table Browser displays the button 'Allow Filtering Using Fields in Checked Tables' if the current table has related table(s), allowing the fields of the related table(s) to be incorporated into the filter operation. Similarly, when the output format 'selected fields from primary and related tables' is selected, the Table Browser displays the button 'Allow Selection from Checked Fields' if the current table has related table(s), allowing the user to include columns of related tables in the output of the query.

A new correlation feature quickly computes a linear regression on any two tables that include genomic coordinates.

If a table contains annotations on position ranges instead of numeric scores, then it is first transformed into a vector containing 1 at each base where there is an annotation and 0 where no annotation exists. The position/score vectors of the two tables are then intersected such that scores are retained only where both tables have a score. A linear regression is performed on the two vectors and a results page is returned, giving the Pearson's correlation coefficient (r), several other statistics, scatterplots of the two data vectors and residual versus fitted, and value histograms.

The Table Browser includes special support for the wiggle and multi-species alignment table types. For wiggle tables, query results can be returned as data points or as positions of bases where the track (post filtering and intersection, if specified) contains values. As with ordinary annotation tables, the results can be automatically imported into the user's session as a custom track for viewing in the Genome Browser or further querying in the Table Browser, including intersection with other tracks which enables compound queries. For multi-species alignment tables, query results can be returned either as MAF format, showing the aligned bases, or as position ranges.

When the current table is part of a composite track, a new operation is offered: subtrack merge. This allows a selected subset (or all) of the subtrack tables to be condensed into a single set for querying. For example, the intersection or union of selected subtracks can be queried instead of just one subtrack. For wiggle composite tracks, scores from different subtracks can be combined using a variety of arithmetic operations.

The Table Browser output options have been expanded in the current version. Flat-file output can also be returned as a local file and can be gzip-compressed if desired. In addition, the Table Browser user interface has been streamlined. Instead of leading the user through a sequence of pages offering various choices, the Table Browser now displays most options on its main page, such that parameters of the current query can be seen at a glance and quickly adjusted.

NEW TOOLS

Several new tools offer additional means of querying the database. The Gene Sorter offers a gene-focused interface, allowing retrieval of genes that are similar to a given gene in terms of protein sequence homology, expression data, genomic location, PFAM similarity, GO similarity and other measures. The LiftOver tool, accessible via the Genome Browser's 'Convert' link, translates genomic coordinates from one assembly version into another and also retrieves putative orthologous regions in other species using UCSC's chained and netted alignments. The In Silico PCR utility (isPcr) performs an extremely fast exact-match search of primer pairs.

PROCESS AND INFRASTRUCTURE IMPROVEMENTS

In addition to expanding the data content and enhancing the user interface of the database, much effort has been put into improving the processes by which data are incorporated into the database and tested before release on the public server.

Alignments of RefSeq, MGC, mRNA and EST sequences from GenBank are now carried out by an automated process that regularly downloads updated sequences and aligns them using BLAT to each genome sequence. RefSeq and mRNA sequence alignments are updated every night and EST sequence alignments are updated every weekend. This greatly reduces the delay in displaying new or updated cDNA sequence from GenBank in the UCSC browser, but can also result in the removal or change of a sequence alignment that was previously available.

The Genome Browser source code remains freely available for academic and non-profit use and can be licensed for commercial use. It can be obtained using CVS (http://www.nongnu.org/cvs/) or downloaded as an archive file (http://genome.ucsc.edu/admin/mirror.html#step6, step 6d).

Relationships between tables in the database are now formally stated in the file all.joiner, available in the source code (kent/src/hg/makeDb/schema/all.joiner), enabling automated consistency checks and the Table Browser's new support for relational queries.

About half of the technical staff are dedicated to quality assurance, i.e. testing of software and data integrity. Both automated and manual tests are performed regularly. The site is actively monitored for violations of the usage policy stated on http://genome.ucsc.edu/ in order to keep response time fast for all users.

Mirroring of the site is still supported, although the storage requirements are quite large (~1250 GB as of September 2005) and growing quickly. Two regularly updated mirror sites are maintained at the Medical College of Wisconsin (http://genome.brc.mcw.edu/) and at the Institute for Genome Sciences and Policy at Duke University (http://genome-mirror.duhs.duke.edu/). These mirrors should be used only as an alternative when there are problems accessing the UCSC site.

A public MySQL server is available for occasional direct SQL queries: genome-mysql.cse.ucsc.edu. More information about access to this server, as well as usage guidelines and restrictions, may be found in our FAQ at http://genome.ucsc.edu/FAQ/FAQdownloads#download29. For many queries, especially those involving overlap between genomic coordinates of two tables, the Table Browser is much faster than pure SQL queries.

The Genome Browser documentation set has been extended to include individual user's guides for several of the applications, including the Genome Browser, Table Browser, Gene Sorter and Proteome Browser. The FAQ has been expanded and reorganized to facilitate the location of help topics. We now provide an automatically generated release log, updated daily, that reports releases and updates of genome assemblies, annotation tracks and other data files.

UCSC now offers online and onsite hands-on Genome Browser tutorials and training materials through OpenHelix (http://www.openhelix.com/). A slide presentation and pre-recorded tutorial are available for free download. Quick reference cards describing the Genome Browser and Table Browser are also available.

FUTURE DIRECTIONS

As genomes of ever more species are assembled and incorporated into the GBD, comparative analysis will remain a

focus of our efforts, including measures of conservation and ancestral genome reconstruction. However, because the rate at which genomes are sequenced has now eclipsed UCSC's resources for database production, not every sequenced genome will be incorporated into the database; instead, we plan to focus on the human genome and several major model organisms. The human GBDs will continue to import a wealth of data generated by the ENCODE Project Consortium. ENCODE-motivated improvements to the Genome Browser and Table Browser are also underway. Variation data, especially for human, is another area of anticipated expansion. Improvements are planned in management of user sessions and settings. A new tool for browsing *in situ* images is under development.

CONTACTING US

The genome@soe.ucsc.edu mailing list is a moderated, public forum for general questions about the data and software. Users who wish to learn from others' questions can subscribe to the list at http://www.cse.ucsc.edu/mailman/listinfo/genome. The list archives are searchable via a link on that page. The genome-announce@soe.ucsc.edu list is used solely for announcements such as additions of data, training sessions or server outages; see http://www.cse.ucsc.edu/mailman/listinfo/genome-announce. To report problems accessing the server, or for correspondence not appropriate for the public forum, send email to genome-www@soe.ucsc.edu. Finally, for questions about local installation of the software and/or database, send email to genome-mirror@soe.ucsc.edu.

ACKNOWLEDGEMENTS

The UCSC GBD project is funded by grants from the National Human Genome Research Institute (NHGRI), the Howard Hughes Medical Institute (HHMI) and the National Cancer Institute (NCI). CWS was supported by an HHMI Predoctoral Fellowship. We would also like to thank the many collaborators who have contributed annotation data to our project, as well as our users for their feedback and support. Last but not least, we would like to thank three excellent and dedicated system administrators who have provided a stellar computing environment over the past several years: Jorge Garcia, Patrick Gavin and Paul Tatarsky. Funding to pay the Open Access publication charges for this article was provided by HHMI.

Conflict of interest statement. A.S. Hinrichs, D. Karolchik, R. Baertsch, G.P. Barber, H. Clawson, M. Diekhans, T.S. Furey, R.A. Harte, F. Hsu, A. Pohl, B.J. Raney, K.R. Rosenbloom, C.W. Sugnet, H. Trumbower, D. Haussler and W.J. Kent receive royalties from the sale of UCSC Genome Browser source code licences to commercial entities.

REFERENCES

1. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.
2. 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.
3. 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.
4. 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.
5. Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human-mouse alignments with BLASTZ. *Genome Res.*, **13**, 103–107.
6. Kent,W.J., Baertsch,R., Hinrichs,A., Miller,W. and Haussler,D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA*, **100**, 11484–11489.
7. Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F.A., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H., Green,E.D. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.
8. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
9. Bejerano,G., Siepel,A.C., Kent,W.J. and Haussler,D. (2005) Computational screening of conserved genomic DNA in search of functional noncoding elements. *Nature Methods*, **2**, 535–545.
10. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
11. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2005) GenBank. *Nucleic Acids Res.*, **33**, D34–D38.
12. 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.
13. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.
14. Gerhard,D.S., Wagner,L., Feingold,E.A., Shenmen,C.M., Grouse,L.H., Schuler,G., Klein,S.L., Old,S., Rasooly,R., Good,P. *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.
15. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
16. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
17. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L., Feng,Z. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, **33**, D233–D237.
18. Pieper,U., Eswar,N., Braberg,H., Madhusudhan,M.S., Davis,F.P., Stuart,A.C., Mirkovic,N., Rossi,A., Marti-Renom,M.A., Fiser,A. *et al.* (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.*, **32**, D217–D222.
19. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
20. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D266.
21. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
22. Krummenacker,M., Paley,S., Mueller,L., Yan,T. and Karp,P.D. (2005) Querying and computing with BioCyc databases. *Bioinformatics*, **21**, 3454–3455.
23. Joshi-Tope,G., Gillespie,M., Vastrik,I., D'Eustachio,P., Schmidt,E., de Bono,B., Jassal,B., Gopinath,G.R., Wu,G.R., Matthews,L. *et al.* (2005)

Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*., **33**, D428–D432.
24. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.
25. Altschul,S.F., Gish,W., Miller,W., Meyers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
26. Encode Project Consortium (2004), The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.
27. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al*. (2005) Ensembl 2005. *Nucleic Acids Res*., **33**, D447–D453.
28. Ashurst,J.L., Chen,C.K., Gilbert,J.G., Jekosch,K., Keenan,S., Meidl,P., Searle,S.M., Stalker,J., Storey,R., Trevanion,S. *et al*. (2005) The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res*., **33**, D459–D465.
29. Kim,N., Shin,S. and Lee,S. (2005) ECgene: genome-based EST clustering and gene modeling for alternative splicing. *Genome Res*., **15**, 566–576.
30. Flicek,P., Keibler,E., Hu,P., Korf,I. and Brent,M.R. (2003) Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. *Genome Res*., **13**, 46–54.
31. Wiehe,T., Gebauer-Jung,S., Mitchell-Olds,T. and Guigo,R. (2001) SGP-1: prediction and validation of homologous genes based on sequence alignments. *Genome Res*., **11**, 1574–1583.
32. Parra,G., Blanco,E. and Guigo,R. (2000) GeneID in Drosophila. *Genome Res*., **10**, 511–515.
33. Stanke,M., Steinkamp,R., Waack,S. and Morgenstern,B. (2004) AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res*., **32**, W309–W312.
34. Harrison,P.M., Zheng,D., Zhang,Z., Carriero,N. and Gerstein,M. (2005) Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic Acids Res*., **33**, 2374–2383.
35. Madera,M., Vogel,C., Kummerfeld,S.K., Chothia,C. and Gough,J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res*., **32**, D235–D239.
36. Griffiths-Jones,S. (2004) The microRNA registry. *Nucleic Acids Res*., **32**, D109–D111.
37. Weber,M.J. (2005) New human and mouse microRNA genes found by homology search. *FEBS J*., **272**, 59–73.
38. Drysdale,R.A. and Crosby,M.A. and FlyBase Consortium. (2005) FlyBase: genes and gene models. *Nucleic Acids Res*., **33**, D390–D395.
39. Chen,N., Harris,T.W., Antoshechkin,I., Bastiani,C., Bieri,T., Blasiar,D., Bradnam,K., Canaran,P., Chan,J., Chen,C.-K. *et al*. (2005) WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. *Nucleic Acids Res*., **33**, D383–D389.
40. Christie,K.R., Weng,S., Balakrishnan,R., Costanzo,M.C., Dolinski,K., Dwight,S.S., Engel,S.R., Feierbach,B., Fisk,D.G., Hirschman,J.E. *et al*. (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res*., **32**, D311–D314.
41. Su,A.I., Cooke,M.P., Ching,K.A., Hakak,Y., Walker,J.R., Wiltshire,T., Orth,A.P., Vega,R.G., Sapinoso,L.M., Moqrich,A. *et al*. (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, **99**, 4465–4470.
42. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al*. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
43. Cheng,J., Kapranov,P., Drenkow,J., Dike,S., Brubaker,S., Patel,S., Long,J., Stern,D., Tammana,H., Helt,G. *et al*. (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, **308**, 1149–1154.
44. Rinn,J.L., Rozowsky,J.S., Laurenzi,I.J., Petersen,P.H., Zou,K., Zhong,W., Gerstein,M. and Snyder,M. (2004) Major molecular differences between mammalian sexes are involved in drug metabolism and renal function. *Dev. Cell*, **6**, 791–800.
45. Davuluri,R.V., Grosse,I. and Zhang,M.Q. (2001) Computational identification of promoters and first exons in the human genome. *Nature Genet*., **29**, 412–417.
46. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*., **31**, 374–378.
47. 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.
48. Bailey,J.A., Yavor,A.M., Massa,H.F., Trask,B.J. and Eichler,E.E. (2001) Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res*., **11**, 1005–1017.
49. 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. *Nature Genet*., **36**, 949–951.
50. Sebat,J., Lakshmi,B., Troge,J., Alexander,J., Young,J., Lundin,P., Maner,S., Massa,H., Walker,M., Chi,M. *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.
51. 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., Segraves,R. *et al*. (2005) Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet*., **77**, 78–88.
52. Tuzun,E., Sharp,A.J., Bailey,J.A., Kaul,R., Morrison,V.A., Pertz,L.M., Haugen,E., Hayden,H., Albertson,D., Pinkel,D. *et al*. (2005) Fine-scale structural variation of the human genome. *Nature Genet*., **37**, 727–732.
53. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2005) REBASE—restriction enzymes and DNA methyltransferases. *Nucleic Acids Res*., **33**, D230–232.