The UCSC Genome Browser

The rapid progress of public sequencing and mapping efforts on vertebrate genomes has increased the demand for tools that offer quick and easy access to the data at many levels and facilitate comparative data analysis. The University of California Santa Cruz (UCSC) Genome Bioinformatics Web site at *http://genome.ucsc.edu* provides links to a variety of genome analysis tools, most notably the UCSC Genome Browser (Kent et al., 2002; Hinrichs et al., 2006), a graphical tool for viewing a specified region of a genome and a collection of aligned annotation "tracks." Another tool on the Web site—the UCSC Table Browser—supplies convenient access to the MySql database tables (Karolchik et al., 2003) underlying the Genome Browser annotations. Both browsers support a custom annotation tracks feature that enables users to upload their own data for display and comparison.

The main protocol of this unit (see Basic Protocol) describes how to display and navigate through a specific section of a genome and its associated annotation tracks in the Genome Browser, configure the view to focus on annotations of interest and optimize comparative analysis, link to external information, and download sequence or annotation data. Support Protocol 1 explains the process for creating and displaying a custom annotation track based on the user's own data. Support Protocol 2 provides a basic overview of the UCSC Table Browser, describing the most commonly used functions, how to set up a simple query, and an introduction to some of the advanced features. The Genome Browser annotations and software continually evolve as new data and techniques become available; therefore, it is recommended that the user consult the UCSC Web site (*http://genome.ucsc.edu*) and the current version of the User's Guide (*http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html*) for the latest information on new releases and features.

USING THE UCSC GENOME BROWSER

The Genome Browser software and data may be accessed on the Internet from the UCSC Genome Bioinformatics Group Web site at *http://genome.ucsc.edu*.

Necessary Resources

Hardware

> Unix, Windows, or Macintosh workstation with an Internet connection and a minimum display resolution of 640 × 480 dpi.

Software

> An up-to-date Internet browser, such as Internet Explorer (*http://www.microsoft.com/ie*); Netscape (*http://browser.netscape.com*); Firefox (*http://www.mozilla.org/firefox*); or Safari (*http://www.apple.com/safari*). The Web browser should support JavaScript 1.1 (e.g., Netscape 4.x and higher, Internet Explorer 4.x and higher, and Mozilla 1.0 and higher)

Files

> None

Navigate to the Genome Browser window to a specific genomic position

1. Open the UCSC Genome Bioinformatics Group home page, at *http://genome. ucsc.edu*, in a Web browser.
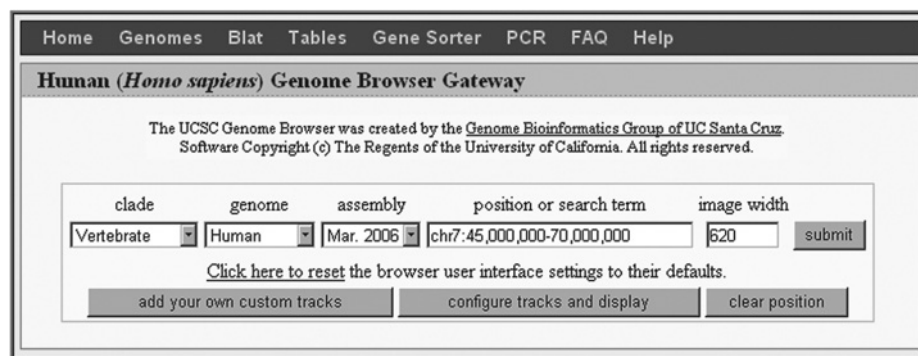
Figure 1.4.1 The Genome Browser Gateway page, set up to span the central portion of chromosome 7 (`chr7:45,000,000-70,000,000`) in the March 2006 human assembly. Custom annotation tracks (Support Protocol 1) can be uploaded by clicking the "add your own custom tracks" button. The initial Genome Browser display may be configured by clicking the "configure tracks and display" button. The lower portion of this page (not shown) displays a description of the selected assembly, relevant links, and examples of queries that may be entered in the "position or search term" box.

The UCSC Genome Bioinformatics home page provides links to the Genome Browser application and a variety of other useful tools: BLAT (Kent et al., 2002), for quickly mapping sequences to a genome assembly; the Table Browser (Karolchik et al., 2004; Hinrichs et al., 2006), for viewing and manipulating the data underlying the Genome Browser; the Gene Sorter (Kent et al., 2005), for exploring relationships (expression, homology, etc.) among groups of genes; VisiGene, for browsing through a large collection of in situ mouse and frog images to examine expression patterns; the Proteome Browser (Hsu et al., 2005), for viewing information about a selected protein; and an in silico PCR tool for rapidly searching a sequence database with a pair of PCR primers. General information about the Genome Browser tool suite can be found in the User's Guide—accessed via the Help link—and the FAQ. From the home page, the user can also download the genomic sequence and annotation data, display contributed custom tracks and older archived data, review a log of released data, and access helpful utilities, training materials, credits for contributors and collaborators, mirror information, and related publications.

2. Click the Genome Browser link in the left-hand sidebar menu to open the Genome Browser Gateway page.

On the Gateway page (Fig. 1.4.1), the user can set the parameters that determine which portion of a genome the Genome Browser will initially display. The bottom portion of the page provides information about the currently selected genome assembly and a list of sample position queries that can be used to open the browser.

Alternatively, the Genome Browser can be accessed by clicking on the BLAT link on the home page and then searching a DNA or protein sequence for regions of homology (step 16).

3. Select the clade, genome, and assembly in which of interest, then type one or more search terms or a set of genomic coordinates into the "position or search term" text box to specify the genome region to display. Click the "submit" button.

The position search supports direct positional queries such as chromosome bands or chromosome coordinate ranges, as well as queries related to genomic features such as gene symbols, mRNA or EST accession numbers, author names, or other descriptive terms likely to occur in GenBank (Benson et al., 2005). The Gateway page shows examples of valid position requests applicable to the selected genome assembly.

If the position query is valid, the Genome Browser will display a page containing an annotation track image specific to the position query, accompanied by navigation controls and display controls (Fig. 1.4.2). Frequently, the position search returns a list of several
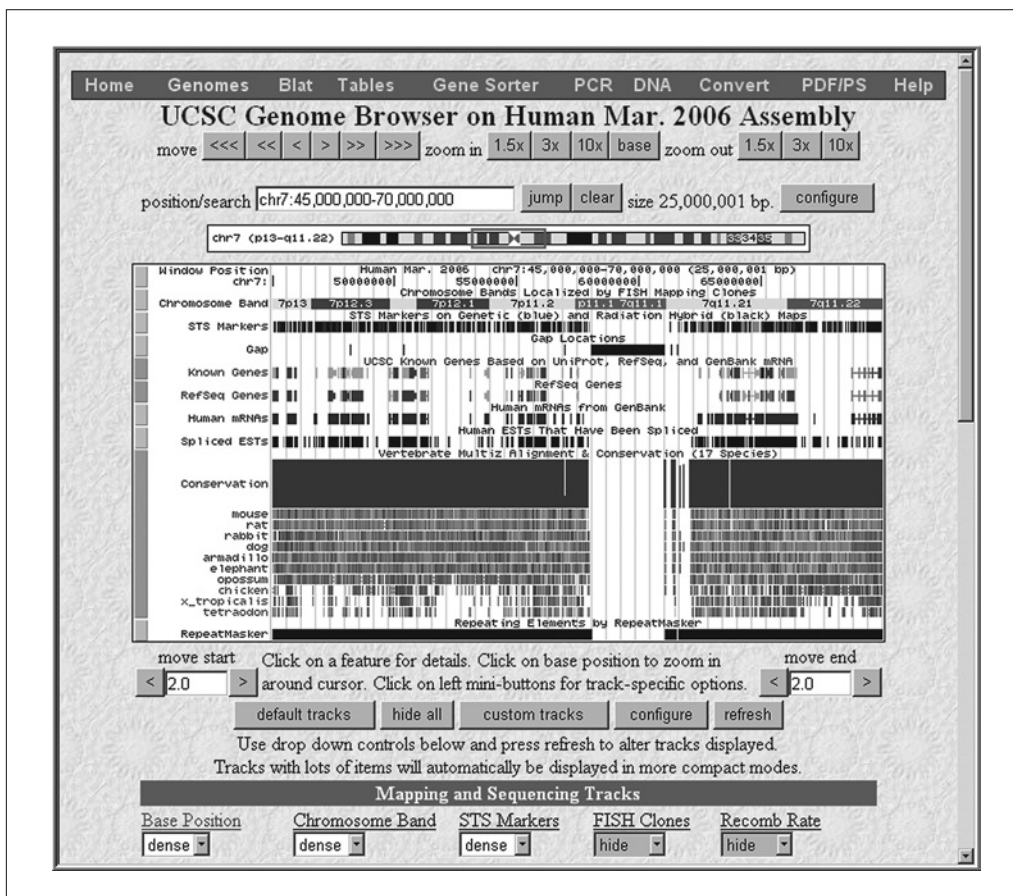
The UCSC Genome Browser

1.4.2

Figure 1.4.2 The Genome Browser annotation track page displaying the `chr7: 45,000,000-70,000,000` region in the March 2006 human assembly (this image differs slightly from the default display). The red rectangle in the ideogram directly above the annotation tracks image indicates the location of the currently displayed region in the overall chromosome. The navigation and configuration buttons are visible above and below the image, and the first line of the Track Controls section can be seen at the bottom of the figure. The name of each annotation track is shown in the left-hand column. The red section of the Chromosome Band track at `p11.1-q11.1` pinpoints the location of the centromere. The Gap track points out gaps in the assembly. The Known Genes and RefSeq Genes tracks give an indication of gene density and gene deserts. At a zoomed-in view, blocks in the Gap track will show gaps between finished clones, while small vertical bars show gaps within draft clones. A horizontal white line through a gap indicates evidence for relative order and orientation of the contigs on either side of the gap.

matches in response to a query rather than immediately displaying the Genome Browser page. When this occurs, click on the item of interest and the Genome Browser will open to that location. Invalid position queries (e.g., withdrawn gene names, abandoned synonyms, misspelled identifiers, and data added after the last Genome Browser database update) will return an error.

A custom annotation track can be uploaded into the Genome Browser by clicking the "add your own custom tracks" button on the Gateway page. For more information on creating and uploading custom annotation tracks, see Support Protocol 1.

To access an older data set that is no longer available from the assembly menu, look in the Genome Browser archives at http://genome-archive.cse.ucsc.edu.

Several aspects of the Genome Browser display can be customized by clicking the "configure tracks and display" button (see step 8).

**Using Biological
Databases**

1.4.3

Browse and configure the annotation tracks display

4. Examine the Genome Browser annotation tracks page (Fig. 1.4.2).

   This image displays a set of annotation tracks aligned beneath a Base Position track (the "ruler") indicating genomic coordinate positions. Tracks are organized into groups reflecting the nature of their data. The first time the Genome Browser is opened, the application's default values are used to configure this display. Any preferences and configurations set during the session will be retained for use in subsequent sessions on the same Web browser. To reset the display to the set of default tracks for the selected assembly, click the "default tracks" button.

   The annotation tracks image is accompanied by controls to configure the display and navigate through the sequence. For selected assemblies, a chromosome band ideogram directly above the image graphically indicates the location of the currently displayed region on the overall chromosome. Custom annotation tracks can be uploaded to the current assembly by clicking the "custom tracks" button below the image (see Support Protocol 1 for more information).

   Figure 1.4.2 shows the annotation track image spanning the central portion of human chromosome 7 defined by the base coordinate range `chr7:45,000,000-70,000,000`. *At this level of detail, the user can get an overview of the coverage and completeness of the region.*

   To generate a high-quality image of this annotation tracks image in PostScript or PDF format, click the PDF/PS link in the top menu bar.

5. Change the display mode of an annotation track by locating the track's controller in the Track Controls section, selecting a display mode from the controller's pull-down menu, and then clicking the "refresh" button.

   Depending on individual display modes, annotation tracks may be hidden from view (hide mode), displayed with all features collapsed into a single line (dense mode), or fully expanded with each feature on a separate line (full mode). Many tracks feature two additional display modes: pack mode, in which each feature is displayed and labeled, but not necessarily on a separate line, and squish mode, which is similar to pack mode, but displays unlabeled features at half-height. To quickly toggle between dense and full modes in the annotation track image, click on the track's label. To hide all the tracks in the display, click the "hide all" button beneath the annotation tracks image.

   By adjusting the display modes of the tracks in the annotation track graphic the user can restrict the display to data of interest, reduce clutter, and improve speed. Dense display mode is useful to get an overview of an annotation or to reduce the space used by a track when the individual feature details of an annotation track are not required. Squished and packed displays show individual feature details of densely populated tracks while conserving space. Use full mode sparingly: in some tracks, the number of features that may potentially align at a selected position can be quite large. When the feature count is excessive in full display mode, the browser displays the first 250 items individually, then groups the remaining items into a single line in dense mode at the bottom of the track.

6. To change the image to a completely new genomic position, type a new set of search terms into the position/search box, then click the "jump" button.

   Figure 1.4.3 shows the finer level of detailed obtained by entering the query `REN100295;REN100309`, *which specifies the region of the HOXA1 gene located between STS markers* `REN100295` *and* `REN100309` *on the March 2006 human genome assembly. Note that the Genome Browser automatically changes the text in the Position box to show the chromosomal position of the resulting display. In most annotation tracks, the aligned regions are represented by vertical bars or blocks. In the Spliced ESTs track shown in this example, the degree of darkness of the block shading corresponds to the number of features aligning to the region. In the mRNA and gene prediction tracks, the thicker regions (usually coding exons) are connected by thin horizontal lines representing gaps (usually spliced-out introns). Thinner blocks on the leading and trailing ends of the aligning regions in gene tracks represent the 5′ and 3′ untranslated regions (UTRs). In full display mode, arrowheads on the connecting lines indicate the direction of transcription.*
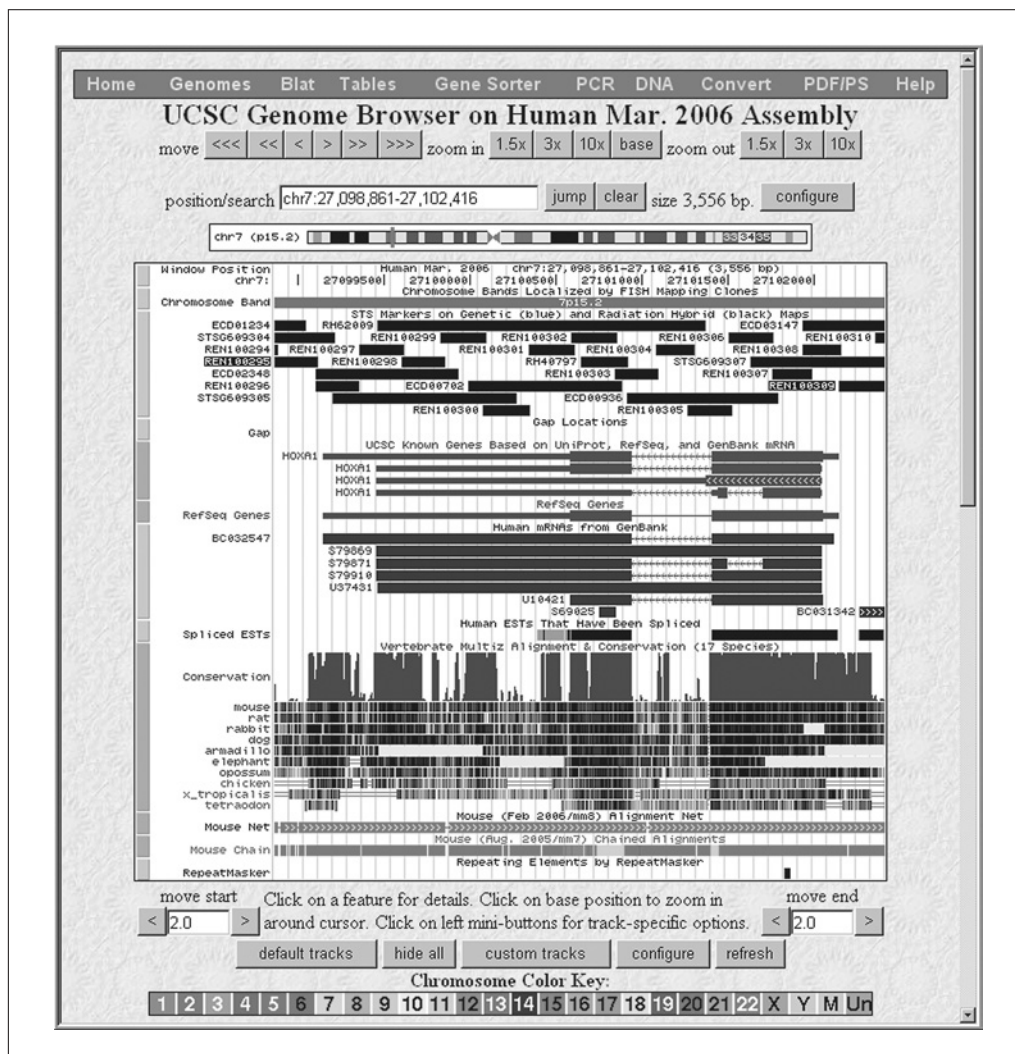
**The UCSC
Genome Browser**

1.4.4

Figure 1.4.3 The Genome Browser annotation track display zoomed in to display the region lying between STS markers `REN100295` and `REN100309` in the *HOXA1* region on human chromosome 7, March 2006 assembly. This display automatically includes 250,000 bases on each side of the query range. The Known Genes and mRNA tracks have been expanded to pack display mode to show individual features. Note the signs of alternative splicing among the mRNAs aligned to the *HOXA1* gene. The Spliced ESTs that indicate active transcription in a region that does not align to an identified gene may provide evidence for an unannotated gene. Further supporting evidence may be found by examining these regions in the cross-species homology tracks for signs of conservation.

Note the comparative genomics annotations displayed in Figure 1.4.3. The Conservation track shows a measure of evolutionary conservation among multiple species, which tends to indicate functional regions of the genome. The lower section of the track shows pairwise alignments of each species to the reference sequence; the top section displays the evolutionary conservation scores assigned by the phastCons program (Siepel et al., 2005). Two pairwise alignment tracks (Mouse Chain and Mouse Net) are also visible in this figure. Chain tracks consist of chained alignment blocks that depict regions in the genome potentially derived from the same segment in a common ancestor, suggesting orthology. A pile-up of overlapping chains in the aligning species represents paralogs, often due to tandem, segmental, or retrotranspositional duplications. Net tracks present an organized view of the chains, displaying the highest-scoring chains spanning the region at the top level, very likely orthologous regions. Lower-scoring chains are used to fill in gaps in the higher-level chains. The block colors in cross-species alignment tracks indicate the chromosomes of origin of the aligning sequence, indexed in the Chromosome Color Key displayed below the annotation tracks image. The color key automatically displays whenever a pairwise alignment track is viewable in the browser.

Using Biological Databases

1.4.5

In Figure 1.4.3 the Known Genes and mRNAs tracks have been expanded to pack display mode to show the individual features.

7. Use the "zoom" and "move" buttons to increase or decrease the breadth of the displayed coordinate range, or to shift one or both ends of the coordinate range to the left or right.

   The navigation buttons are useful for focusing the display on the exact position required. "Zoom" buttons increase or decrease the displayed coordinate range by 1.5-, 3-, or 10-fold. To zoom in by 3-fold on a particular coordinate, click the Base Position track at that location. To quickly zoom in to the base composition of the sequence underlying the current annotation track image, click the zoom-in "base" button. "Move" buttons shift the displayed coordinates in the indicated direction by 10%, 50%, or 95% of the displayed size. To scroll the coordinate position of one side of the track display while holding the position of the opposite end static, click the corresponding "move start" or "move end" arrow button. For example, to preserve the left-hand display coordinate but increase the right-hand coordinate, click the "move end" forward arrow. To increase or decrease the scroll interval, edit the number in the "move start" or "move end" text box.

8. Click the "configure" button above or below the annotation tracks image to access a Web page for changing display characteristics (such as the image width and text size), hiding, showing, or reordering track groups, and displaying the chromosome ideogram, the track controls section, and image labels. Click the "submit" button on this page to apply the changes and return to the annotation tracks page.

   The default display width of the annotation tracks graphic is optimized for smaller monitors with lower resolutions. Most displays are no longer subject to these limitations; in these situations the visible portion of the genome can be increased and the need for screen redraws can be reduced by setting the image width to a larger number.

   Exercise caution when using the "show all" option in the track configuration section: if the group or assembly has a large amount of annotation data, the Web browser session may freeze or terminate before the data sets are loaded.

   If cookies are enabled, modifications made on the configuration page will be retained in future sessions on the same browser until they are reset.

9. Click the vertical button to the left of a displayed track to view additional information about the annotation and (in many cases) to filter the features displayed in the track.

   The description page can also be displayed by clicking the track's label in the "track controls" section.

   Click the button adjacent to the Known Genes track to view an example of a typical description page. This page contains a configuration/filter section (when applicable) followed by a description of the annotation track, information about interpreting and configuring the track display, a discussion of the methods used to collect and compute the data, credits for authors and contributors, associated references, and in this case, restrictions on the use of the data. Additional credits can be found by clicking the Credits link on the home page.

   Most of the tracks in the Genome Browser have filter or configuration options that modify the graphical characteristics or restrict the display to features that match filtering criteria. Filters are useful for focusing attention on relevant features when a track contains large amounts of data. Some of the more complex graphical annotations, such as the continuous value graph ("wiggle") display featured in the Conservation track, offer an extensive set of configuration options. In most cases, detailed configuration information can be found in the "Display Conventions and Configuration" section on the description page.

   Filter and configuration settings are persistent from session to session on the same Web browser. To revert to the original default settings for a track, manually restore the settings on the description page; to return the entire annotation track set to its original settings, click the "Click here to reset" link on the Gateway page.

10. Click on the label of a feature in a track shown in pack or full display mode to view detailed information about the feature and access links to additional information.

The types of information available vary by track. Click on the HOXA1 gene label in the Known Genes track in Figure 1.4.3 to view an extensive collection of information about the gene, including the associated UniProt (Wu et al., 2006) and RefSeq (Pruitt et al., 2005) descriptions, microarray expression data, links to associated information about this gene in several UCSC tools (such as the Gene Sorter, Proteome Browser, and Table Browser) as well as links to related records in external databases, including Online Mendelian Inheritance in Man (OMIM; Hamosh et al., 2005; UNIT 1.2), Entrez Gene (Wheeler et al., 2006), GeneLynx (Lenhard et al., 2001), GeneCards (Safran et al., 2003), AceView, PubMed (Wheeler et al., 2006; UNIT 1.3), the HUGO Gene Nomenclature Committee Database (HGNC; Eyre et al., 2006), the Cancer Genome Anatomy Project (CGAP; Strausberg et al., 2001), PDB (Deshpande et al., 2005), ModBase (Pieper et al., 2006), InterPro (Mulder et al., 2005), Pfam (Finn et al., 2006), the Stanford SOURCE, Jackson Lab, and the Allen Brain Atlas. The page also includes hyperlinks that will display the corresponding protein, mRNA, and genomic sequences for HOXA1. These sequences are a useful source of input into the BLAT tool, which will be discussed in step 16.

The Genome Browser also provides direct links to the Ensembl Browser (Birney et al., 2006; UNIT 1.15) and NCBI's Map Viewer (Wheeler et al., 2006; UNIT 1.5), when available. To view the complementary annotation in one of these browsers, return to the annotation tracks page and click the Ensembl or NCBI link in the top menu bar.

Examine the underlying data and download the sequence and annotation data tables

11. Click the DNA link on the annotation tracks page menu bar to view the DNA sequence underlying the features in the image. This option allows the user to change the formatting and coloring of the text that represents the sequence to highlight features of interest.

The initial window that displays provides options for marking or masking repeats, changing the case of the letters that represent the DNA, showing the reverse complement of the sequence, and displaying additional sequence upstream or downstream of the selected sequence. Click the "extended case/color options" button to display additional font and color configuration options.

The Extended DNA Case/Color Options page is useful for highlighting features within a genomic sequence, pointing out overlaps between two types of features, or masking out unwanted features. In Figure 1.4.4, the configuration has been set to display exons from the Known Genes track in uppercase letters. The Spliced EST track is configured to reflect the level of coverage by setting its color to RGB value (0,64,0). When the Submit button is clicked, the Extended DNA Output window displays exons in uppercase. Nucleotides covered by a single EST appear as dark green, while regions with more EST alignments appear progressively brighter, saturating at four ESTs.

Be careful when requesting complex formatting for a large chromosomal region: when all the HTML tags have been added to the output page, the file size may exceed the limits that the Web browser, clipboard, and other software can display.

12. Click on the Tables link on the annotation tracks page menu bar to examine the database tables underlying the Genome Browser annotation tracks.

The Table Browser tool provides a graphical interface for viewing and manipulating Genome Browser data. Support Protocol 2 gives a brief introduction to using the Table Browser. Additional information can be found in the Table Browser User's Guide accessible from the Help link in the Table Browser top menu bar.

13. Click the Home link on the top menu bar to return to the UCSC Genome Bioinformatics home page, then click the Downloads link on the side bar to display a listing of sequence files and database tables available for downloading.

Using Biological Databases

1.4.7

Figure 1.4.4 An extended DNA Case/Color Options request to display the DNA for the `chr7:27,098,861-27,102,416` region of the March 2006 human assembly. This configuration will show Known Genes in uppercase, all other regions in lowercase, and Spliced ESTs in varying intensities of green, depending on the level of coverage.

The Downloads page contains links to all the Genome Browser assemblies, annotations, and source codes available on the Genome Browser downloads server. To access older assembly versions, it may be necessary to look in the archives (http://genome-archive.cse.ucsc.edu). Data is also downloadable at the Genome Browser FTP site (ftp://hgdownload.cse.ucsc.edu/goldenPath/). FTP or rsync is recommended for large data downloads. All data in the Genome Browser are freely available, except where noted in the README.txt *file specific to a particular downloads directory. The Genome Browser and BLAT source are freely available for academic, noncommercial, and personal use; commercial licensing information can be found via the Licenses link on the home page.*

Convert coordinates in the displayed range to a different assembly using the Convert, LiftOver, or BLAT tools

14. Return to the annotation tracks page. Click the Convert link in the menu bar to convert the coordinates in the displayed range to those of a different assembly.

The coordinate conversion tool is useful for locating the position of a feature of interest in a different genome assembly. Coordinates of features frequently change from one assembly to the next as gaps are closed, strand orientations are corrected, and duplications are reduced. For example, to map the location of a sequence in the May 2004 human assembly to the March 2006 human assembly, open the May 2004 Genome Browser to the desired position, click the Convert link, select the Mar. 2006 option in the New Assembly pull-down menu, then click the "Submit" button. If successful, the Convert tool displays one or more coordinate ranges in the March 2006 assembly to which the May 2004 sequence maps.

15. To convert multiple sets of sequence coordinates between assemblies or to exert control over the parameters used in the conversion, use the LiftOver batch coordinate conversion tool.

   The LiftOver tool can be accessed from the Utilities link on the Genome Bioinformatics home page. Enter the list of coordinate ranges in the large text box, one per line, or upload the list from a file. Detailed information about parameter settings can be found at the bottom of the page, as well as information about a Linux command-line version of the tool.

16. Alternatively, use BLAT to map a sequence to a different assembly:

   a. To determine the sequence underlying the region currently displayed on the annotation tracks image, click the DNA link in the top menu bar on the annotation tracks page, then click the "get DNA" button. For more information on the Get DNA utility, see step 11. To find the sequence of a specific feature within the annotation track image, click on the feature label to display its details page. In most cases, the sequence is available as a link from this page. Note that BLAT limits input to 25,000 bases.

   b. Using the Web browser's copy function, copy the entire sequence onto the clipboard. Return to the previous page and click the BLAT link in the top menu bar.

   c. On the BLAT Web page, paste the sequence into the large text box (Fig. 1.4.5). Select the genome and assembly to which to map the sequence, then click the "submit" button. If successful, BLAT will display a list of search results sorted by score (Fig. 1.4.6).

   d. To view the details of the matching alignments, click the "details" link; to display the sequence in the Genome Browser, click the "browser" link.

   This procedure demonstrates one use of the BLAT search tool. This tool, which can be accessed from the BLAT link on the top menu bar of most Genome Browser pages, is a very fast sequence alignment tool similar to BLAST (UNITS 3.3 & 3.4). For more information on BLAT, refer to the Genome Browser User's Guide.

Figure 1.4.5 A BLAT search set up to align the FASTA sequence in the text box against the March 2006 human genome assembly. This sequence was obtained by copying and pasting the first 600 bases of output from the Get DNA search illustrated in Figure 1.4.4.

Using Biological Databases

1.4.9

Figure 1.4.6 The results returned by the BLAT search shown in Figure 1.4.5. Clicking on the "browser" link for a given line will display the data in the Genome Browser; the "details" link will display a page showing a base-by-base of the alignment to the genome.

CREATING A CUSTOM ANNOTATION TRACK

Custom annotation tracks enable users to upload personal data for temporary use in the Genome Browser and Table Browser. Custom tracks are viewable only on the machine from which they are uploaded, and the data may be accessed only by the users on that machine. Tracks are kept for 48 hr after the last time accessed or until the user switches to a different genome assembly; no permanent archives are created. Optionally, users can make custom annotations viewable by others as well.

Typically, custom annotation tracks are displayed under the corresponding genomic positions on the Base Position track. Each custom track has its own track control and persists even when not displayed in the Genome Browser window (e.g., if the position changes to a range that no longer includes the track).

Since space is limited in the annotation track graphic, many excellent genome-wide tracks must be excluded from the set provided with the browser. A Web page with links to several user-contributed custom tracks can be found by clicking the Custom Tracks link on the home page.

The information in this section provides an overview of custom annotation tracks. For a more detailed discussion of formats and syntax, refer to the Genome Browser custom annotation track documentation Web page at *http://genome.ucsc.edu/goldenPath/help/customTrack.html*.

Necessary Resources

Hardware

> Unix, Windows, or Macintosh workstation with an Internet connection and a minimum display resolution of 640 × 480 dpi.

Software

> Text editor (*APPENDIX 1C*)

Files

> None

1. Format the data set to be analyzed as a tab-separated file using general feature format (GFF) or one of the formats designed specifically for the Human Genome Project or UCSC Genome Browser. These include gene transfer format (GTF), pattern space layout (PSL), browser extensible data (BED), or wiggle format (WIG).

   Each line of data in the file provides display and positional information for a feature line within the displayed annotation track. The browser ignores empty lines and lines starting with a pound sign (#).

   For detailed information on data formats, refer to the Genome Browser's custom annotation track documentation. Data in PSL, GFF, and GTF files must be tab-delimited rather than space-delimited in order to display correctly. More than one data set may be included in an annotation file, but all lines within a single annotation track must be in the same format. Figure 1.4.7 shows examples of data in BED, PSL, and GFF format.

   An easy way to create correctly formatted data for an annotation file is by collecting PSL output from BLAT or downloading data from the Table Browser.

2. Add one or more optional browser lines to the beginning of the formatted data file to specify the configuration of the Genome Browser window in which the custom annotation track will be displayed.

   Browser lines define the genome position to which the browser will initially open, the width of the display, and the configuration of the other annotation tracks that are shown (or hidden) in the initial display. The Genome Browser custom annotation track documentation describes the browser line syntax and options.

   In the sample BED annotation track shown in Figure 1.4.7, the initial display position is set to `chr22:10000000-10007500`, *and all tracks are hidden except the custom annotation track. If the browser position is not explicitly set in the annotation file, the initial display will default to the position setting most recently used by the user, which may not be an appropriate position for viewing the annotation track.*

BED format

```
browser position chr22:10000000-10007500
browser hide all
track name="BED track" description="BED track example" visibility=2 color=0,128,0 useScore=1
chr22 10001000 10005000 itemA 960 + 10001100 10004700 0 2 1567,1488, 0,2512
chr22 10002000 10007000 itemB 200 - 10002200 10006950 0 4 433,100,550,1500
0,500,2000,3500
```

PSL format

```
browser position chr22:13,073,582-13,073,883
track name=PSL track description="PSL example" visibility=2 useScore=1
59 9 0 0 1 823 1 96 +- FS_CONTIG_48080_1 1955 171 1062 chr22
   47748585 13073589 13073753 2 48,20, 171,1042, 34674832,34674976,
59 7 0 0 1 55 1 55 +- FS_CONTIG_26780_1 2825 2456 2577 chr22
   47748585 13073626 13073747 2 21,45, 2456,2532, 34674838,34674914,
59 7 0 0 1 55 1 55 -+ FS_CONTIG_26780_1 2825 2455 2676 chr22
   47748585 13073727 13073848 2 45,21, 249,349, 13073727,13073827,
```

GFF format

```
browser position chr22:10000000-10034000
track name=GFF track description="GFF example" visibility=2
chr22  TeleGene       enhancer 10000000 10001000 500     + . TG1
chr22  TeleGene       promoter 10010000 10010100 900     + . TG1
chr22  TeleGene       promoter 10020000 10025000 800     - . TG2
```

Figure 1.4.7 Sample custom annotation tracks containing BED, PSL, and GFF data formats. To load correctly, the track line data in the PSL and GFF examples must be tab-separated.

3. Add a track line immediately above the formatted data in the file to define the display attributes for the annotation track.

   The track line defines the track's name, description, colors, initial display mode, use score, and associated URL. The Genome Browser custom annotation track documentation contains a complete description of the track line syntax and options. If more than one data set is included in the annotation file, insert a track line at the beginning of each new set of data.

   In Figure 1.4.7 the left-hand label of the BED annotation track is 'BED track'; the center label is 'BED track example'. The track labels will be displayed in green and the features will be fully displayed. Because the useScore *attribute is set to* 1*, the level of shading of each feature will reflect its score value.*

4. Upload the annotation file into the Genome Browser by clicking the "add your own custom tracks" button on the Gateway page (Fig. 1.4.1) or the "custom tracks" button on the annotation tracks page (Fig. 1.4.2).

   If the file is located on a local machine, enter the file name in the "upload" text box in the "URLs or data" section. To open an annotation through a URL or to manually enter the track data, type or paste the information into the large text box in this section. Multiple tracks may be uploaded simultaneously by including all the track data or URLs (on separate lines) in the text box or grouping the tracks into one uploaded file. Figure 1.4.8 shows the custom track that displays when the BED sample track in Figure 1.4.7 is uploaded into the Genome Browser. Optionally, associated track descriptive text may be uploaded or inserted in the "optional track documentation" section.

   To make the annotation file viewable on a different machine or at a different site, put a copy of the file on a Web server and create a custom annotation track URL that allows the file to be uploaded over the Internet. The URL must contain two bits of information specific to the annotation data file: the UCSC genome assembly on which the annotation is based and the URL of the annotation file on the Web site. The Genome Browser FAQ (http://genome.ucsc.edu/FAQ/FAQreleases#release1) lists the UCSC genome assembly codes. The URL can also include the position within the genome to which the Genome Browser should initially open.

   For example, placing the BED track in Figure 1.4.7 in a file named test.bed *on the genome-test.cse.ucsc.edu Web site enables it to be uploaded using the following custom annotation track URL:*

   http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg18&position=chr22&hgt.customText =http://genome-test.cse.ucsc.edu/test.bed.

   This URL sets the assembly database to the hg18 (March 2006) assembly of the human genome, initializes the display position to chromosome 22, and loads the annotation track file http://genome-test.cse.ucsc.edu/test.bed. In this case, the position initialization in the URL is extraneous; it will be overwritten by the position defined in the custom track file.
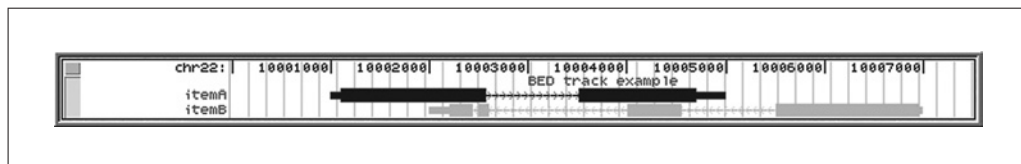

Figure 1.4.8   The annotation track display that results when the BED track example in Figure 1.4.7 is uploaded into the Genome Browser. Note that the lower score value in the ItemB data results in lighter shading of this feature.

USING THE UCSC TABLE BROWSER

The UCSC Table Browser provides a powerful and flexible graphical interface for querying and manipulating the data in the Genome Browser annotation database.

The Table Browser can be used to: (1) retrieve the annotation data or DNA sequence underlying Genome Browser tracks for the entire genome, a specific coordinate range, or a set of accessions; (2) view a list of the tables affiliated with a particular Genome Browser track; (3) view the schema of an annotation table; (4) organize table data into formats that can be used in other applications, spreadsheets or databases; (5) combine data from multiple tables or custom tracks into a single set of output data; (6) filter out certain records in a table based on certain field values; (7) display basic statistics calculated over a selected range of table data; and (8) conduct structured or free-form SQL queries on the annotation data.

The information in this section provides an overview of the Table Browser, which can be accessed on the Internet from the UCSC Genome Bioinformatics home page at *http://genome.ucsc.edu*. For a more detailed discussion of Table Browser options, advanced queries, and several practical examples, refer to the Table Browser User's Guide at *http://genome.ucsc.edu/goldenPath/help/hgTablesHelp.html*.

Necessary Resources

Hardware

> Unix, Windows, or Macintosh workstation with an Internet connection and a minimum display resolution of 640 × 480 dpi.

Software

> An up-to-date Internet browser, such as Internet Explorer (*http://www.microsoft.com/ie*); Netscape (*http://browser.netscape.com*); Firefox (*http://www.mozilla.org/firefox*); or Safari (*http://www.apple.com/safari*). The Web browser should support JavaScript 1.1 (e.g., Netscape 4.x and higher, Internet Explorer 4.x and higher, Mozilla 1.0 and higher)

Files

> None

Set up a simple Table Browser query

1. On the UCSC Genome Bioinformatics home page (Basic Protocol, step 1), click the Table Browser link in the left-hand sidebar menu to display the Table Browser Web page.

   The Table Browser is also accessible from the Tables link in the top menu bar of most Genome Browser pages.

   The top section of the Table Browser Web page (Fig. 1.4.9) contains options for setting up a data query, many of which are optional when conducting simple queries. Each of the options is briefly described at the bottom of the Web page. To view the complete Table Browser User's Guide, click the Help link in the top menu bar.

2. Select the clade, genome, and assembly.

   The clade, genome, and assembly pull-down menus correspond to those found on the Genome Browser Gateway page. The current Genome Browser settings are used when the Table Browser is started from the menu bar on a Genome Browser page.

   For this example, set the clade to "Vertebrate," the genome to "Human," and the assembly to "Mar. 2006."

Figure 1.4.9 The Table Browser tool provides access to the database tables underlying the Genome Browser annotations; in this case, the chromosome 7 data in the knownGene table on the March 2006 human assembly.

3. Select the group, track, and table of interest.

The options in the group and track menus directly correspond to the annotation groups and tracks available in the Genome Browser for the currently selected genome assembly. The track list—which shows all tracks contained in the selected group—automatically updates when a different group is selected.

The table menu lists all the tables in the annotation database that are affiliated with the selected track. Many annotation tracks are based on data from multiple tables joined by common fields. By default, the primary table underlying the track's display in the Genome Browser is listed first.

Click the "describe table schema" button to view the SQL schema for the selected table. The schema page also lists other tables in the annotation database that are joined to the selected table by a particular field, as well as a description of the Genome Browser annotation track associated with the table (when applicable).

The All Tracks and All Tables options in the group menu provide convenient shortcuts if the name of the desired track or table is already known.

For this example, a subset of data in the UCSC Known Genes track will be examined. Select the "Genes and Gene Prediction Tracks" group, the "Known Genes" track, and the "knownGene" (default) table.

4. Specify the query region.

Click the "genome" region setting to view annotation data for the entire genome. To limit the data output to a specific query region, click the "position" region setting and type a query into the adjacent text box. The Table Browser accepts the same types of queries that are valid for the Genome Browser (see Basic Protocol, step 3). Click the "lookup" button to convert a nonpositional query (e.g., an accession or keyword) to a coordinate range.

On genome assemblies with ENCODE annotations (The ENCODE Project Consortium, 2004), an additional "ENCODE" region setting is available that restricts output to data located in ENCODE regions.

For many tables, the query region can be further defined by restricting the output to a set of specific identifiers, such as Known Gene IDs or mRNA accession numbers. Upload the identifiers as a space- or line-separated list by clicking the "paste list" or "upload list" button. For this type of query to return successfully, the identifiers in the list must conform to the format specified for identifiers in the selected table.

For this example, several Known Gene identifiers from chromosome 7 are included in the query. Select the "position" region setting, then type chr7 *in the text box. Click the "paste list" button, then type the following items in the large text box, one per line:* NM_014390, NM_022143, D49487, NM_018077.

5. Select an output format.

   The help text at the bottom of the Table Browser page describes the output formats. Not all options may be available for a given query. The "all fields. . . " format displays the entire set of fields for each record in the output. The "selected fields. . ." format is useful when the user wishes to create output that contains only a subset of fields that will be used as input for further data processing or if the user desires to link in fields from an associated table (see step 7). The "sequence" option returns the sequence underlying the annotation in FASTA format. The GTF, BED, and custom track options are useful for saving the output into a format that can be displayed as a custom track in the Genome Browser. The "data points" format is useful for displaying the conservation scores associated with individual base locations; in contrast, the "MAF" format displays the multiple species alignments underlying the conservation region.

   For this example, choose the "selected fields. . . " output format.

6. Click the "get output" button to submit the query and display the results.

   By default the Table Browser displays the query output in the user's Web browser. To save the data to a file on the local computer, type a file name in the "output file" text box and select the plain or compressed file type option before clicking the "get output" button.

   Many output formats—including the "selected fields. . . " format used in the example— require an additional setup step before the output is displayed. On the setup page associated with our example, check the "name," "chrom", txStart", and "txEnd" boxes, then click the "get output" button. The Table Browser will display the following output:

   ```
   #name        chrom      txStart      txEnd
   NM_014390    chr7       127079437    127519894
   NM_022143    chr7       127454361    127458238
   D49487       chr7       127679296    127682185
   NM_018077    chr7       127737672    127771198
   ```

7. Link in additional data from tables associated with the table being queried.

   The linked tables feature included on the "selected fields. . . " output format setup page provides a convenient way to pull in data from additional tables without having to conduct multiple queries.

   In the previous query, it is easy to display additional data associated with the selected genes by linking in the kgXref table, which provides a convenient cross-reference among gene IDs and information from several different sources such as RefSeq, Swiss-Prot, HGNC, etc. On the selected fields setup page, scroll down to the Linked Tables section, check the box in front of the "hg18/kgXref" table/field, then click the "Allow selection from checked tables" button. At the top of the page, check the "name," "chrom," txStart", and "txEnd" boxes in the hg18.knownGene section, and the "geneSymbol" field in the hg18.kgXref section. Click the "get output" button. The Table Browser will display the HGNC gene symbol associated with each Known Gene record in the output shown in step 6.

Using Biological Databases

1.4.15

8. Click the "summary/statistics" button to display a table of basic statistics about the current query.

The Summary Statistics page profiles data and query characteristics. This information can be useful in determining such information as the percent of bases in a query region that is contained within exons or CpG islands.

Explore advanced query options

9. Create a custom track from a subset of table data using the "custom track" output format option.

The custom track output format allows the user to save query results into a custom annotation file that can be loaded into the Table Browser for further data manipulation or uploaded for display in the Genome Browser.

For this example, repeat steps 1 to 3. Select the "position" region setting and type chr7 *into the position text box. Select the "custom track" output format option, then click "get output". On the custom track setup page, configure the header of the custom track (optional). Select the "Exons plus" option, then click the "get custom track in table browser" button. The track is now loaded into the Table Browser. The data in the track can now be viewed and manipulated by selecting the "Custom Tracks" group option and setting the track list to the name of the user's custom track.*

10. Click the filter "create" button to set up a filter on one or more fields in a data table.

The filter utility allows the user to fine-tune a query to produce a restricted data set that meets a certain set of criteria, such as a minimum threshold or a specific set of IDs or keywords.

For this example, set the clade, genome, and assembly as described in the example in step 2. Select the "Comparative Genomics" group, the "Conservation" track, and the "phastConsNway" table (where N represents the number of species present in the multiple alignment, such as phastCons17way). Set the position to chr7. *On the filter page, set dataValue >0.8, then click the "submit" button. Select the "data points" format, then click "get output". This query will return those bases in the Conservation track that are associated with the peaks where the multiple species conservation score exceeds 0.8 (i.e., regions with a high amount of evolutionary conservation).*

11. Click the intersection "create" button to combine the data from two different tables into a single output file using an intersection or union.

The intersection feature lets the user compare the positions of features in different annotations to identify points of overlap or nonoverlap, establish thresholds for the amount of overlap, and conduct feature-by-feature or base-wise comparisons.

In this example, select the "Custom Track" option in the group menu, then set the track menu to the track created in step 8. Select the "position" region setting and type chr7 *into the position text box. Click the intersection "create" button. On the intersection setup page, select the "Genes and Gene Prediction Tracks" group option, the "Known Genes" track option, and the "knownGene" table option. Retain the defaults for the remaining settings and click "submit." Back on the main Table Browser page, select the "BED" output format, then click "get output." On the output setup page, customize the header (optional), retain the other settings, then click "get BED." The Table Browser will return a list of the conserved bases in the user's custom track that overlap with exons in the Known Genes track.*

Note that the "all fields..." and "selected fields..." output format options are not available when an intersection is active in the current query. Although the intersection utility restricts combinations to two tables, additional tables can be included in an intersection by saving the initial intersection to a custom track, then performing subsequent intersections using the custom track.

GUIDELINES FOR UNDERSTANDING RESULTS

The Genome Browser can be used for genome analysis and interpretation at many different levels. With the annotation track image zoomed out to display several million bases or an entire chromosome, the tool provides a good overview of the coverage and completeness of the region. At a reduced display scale, the Genome Browser is useful for viewing splicing patterns or searching for evidence of previously unidentified genes. By presenting a large collection of annotation tracks in a single view, the browser facilitates interpretations based on a visual correlation of features. However, care must be taken when drawing conclusions. Information presented in the Genome Browser is only as accurate as the underlying data. It is essential to gather supporting evidence when making an analysis, rather than basing judgments on a single track that may contain erroneous or misleading data.

It is important to consider the methods and criteria used to compute an annotation track. Consult the track's description page for a discussion of the sources and methods used to generate the track. In many cases, the page will provide links to additional information about the annotation (such as a seminal publication or related Web site), estimates of accuracy, and caveats for use.

The track details pages are another good source for supporting documentation. Many pages contain links to feature-specific information in external public databases. The OMIM database (*UNIT 1.2*), for example, contains hand-curated experimental literature summaries. LocusLink, GeneLynx, GeneCards, AceView, and PubMed (*UNIT 1.3*) are other good sources for supplementary information.

Many regions−particularly in unfinished areas of a genome−may exhibit discrepancies among the various gene prediction tracks, EST evidence, and cross-species orthology tracks. Tracks generated by gene prediction methods vary considerably in their degrees of sensitivity and specificity. Kent (2002) illustrated some of these differences in a comparison of the correlation of EST, cross-species homology, and ab initio gene prediction tracks with the RefSeq Genes track across the entire genome, along with a similar comparison to the Sanger Centre chromosome 22 gene annotations in the Sanger22 Genes track. It is better to use correlations between EST, cross-species homologies, and ab initio gene predictions to look for evidence of unidentified genes, rather than relying on the information in a single annotation track. There is no gene prediction tool that integrates all the annotation evidence into a single track as yet.

ESTs often exhibit sequencing errors due to the nature of the techniques used. EST databases contain contamination from mRNA and genomic sequence. Because of this, a single unspliced EST should be viewed with considerable skepticism, and alternate splicing predictions should be evaluated by examining the quality of the EST/genomic alignment. Cross-species BLAT alignments that match too perfectly may also be suspect. Those with greater than 97% identity may simply reflect the contamination of one genome by the other.

In several of the annotation tracks generated at UCSC, attempts have been made to filter out data that might provide misleading results. For example, the mRNA and EST alignments on which several of the browser tracks are based are filtered to reduce the presence of pseudogenes, paralogs, and assembly errors. Filtering removes a significant number of alignments in the tracks, particularly very short ones. The Spliced EST track applies additional splicing criteria that greatly reduce the level of contamination from EST databases, although at the expense of eliminating genuine ESTs. Since the maximum intron length allowed by BLAT is 500,000 bases, some ESTs with very long introns are eliminated that otherwise might align.

Using Biological Databases

1.4.17

In summary, good judgment should be used when using any genome-browsing tool. To work effectively in a bioinformatics area subject to errors, it is a good idea to seek supporting data for any unusual findings. Often, the ultimate supporting evidence for a conclusion must be generated in the lab.

COMMENTARY

Background Information

History and development of the UCSC browser

The need for interactive software to search and display a genome at a variety of levels predates the inception of the UCSC Genome Browser. Research on the nematode *C. elegans* in the mid-1990s prompted the creation of A *Caenorhabditis elegans* Database (ACeDB; Eeckman and Durbin, 1995; *UNIT 9.1*) to track strains and genetic crosses. As ACeDB grew in functionality, the software was adopted by the *C. elegans* community, and over the years has been enhanced and extended to support a large number of organisms.

The UCSC Genome Browser was originally developed as an alternative to ACeDB to examine RNA splicing for gene predictions in *C. elegans* (Kent and Zahler, 2000a). This set of Web-based tools—initially called the Intronerator—displayed EST and full-length cDNA tracks from GenBank aligned to the *C. elegans* genomic sequence. The Intronerator was subsequently expanded to include tracks showing homology with *C. briggsae* (Kent and Zahler, 2000b). With the completion of the assembled human genome working draft on the horizon, the software underwent major revisions to accommodate the human genome assembly, which was 30 times larger than that of *C. elegans*. The resulting UCSC Genome Browser retains the speed and performance of its predecessor while displaying the vastly larger data sets of vertebrate genomes. The initial mouse (*Mus musculus*) draft assembly (Waterston et al., 2002) was added to the Genome Browser in 2002, and the browser has subsequently grown to include a large array of genomes: In early 2006, this totaled more than a dozen vertebrates, nearly a dozen species of Drosophila, honeybee, mosquito (*Anopheles gambiae*), two deuterostome species, two nematode species, and a yeast (*Saccharomyces cerevisiae*). Older assemblies are archived as newer versions are released; the UCSC Web site maintains complete assembly archives of the more popular genomes.

In the years since its public debut, the Genome Browser has gained broad popularity among research scientists worldwide for its speed, stability, extensibility, annotation tracks, and the consistency of its user interface. Of the alternative existing tools that provide a somewhat similar functionality to the Genome Browser, the Ensembl Genome Browser (*http://www.ensembl.org/*; *UNIT 1.15*) and the National Center for Biotechnology Information (NCBI) Entrez Map Viewer (*http://www.ncbi.nlm.nih.gov/mapview/*; *UNIT 1.5*) are perhaps the most widely known. The UCSC browser provides links to both of these tools from the menu bar at the top of the annotation tracks page. To support the growing demand for tools that can handle complex analysis of sequence data, the Genome Browser annotation set is continually evolving, including the introduction of several new graphical display types, such as wiggle displays and composite tracks comprised of several related subtracks. Additionally, the Genome Browser application set has grown to include several tools that analyze different aspects of the data: the BLAT alignment tool, Table Browser, Gene Sorter, Proteome Browser, VisiGene, and in silico PCR tool.

How are various annotation tracks determined?

The Genome Browser annotation tracks are grouped by functionality into the following basic categories: mapping and sequencing, phenotype and disease associations, genes and gene predictions, mRNA and EST data, expression and regulation, comparative genomics, variations and repeats, and annotations specific to ENCODE regions. More groupings will undoubtedly be added as scientific technologies improve and analytical efforts branch into new areas. The browser offers a large selection of annotations in each of these categories for the more highly studied genomes, such as the human and mouse; other assemblies may feature only a subset of these annotations. The following discussion highlights some of the tracks featured in the browser; for an indepth description of the tracks, see Kent et al. (2002), Hinrichs et al. (2006), and the individual track description pages.

The browser displays several tracks based on mRNA alignments. The mRNA and EST

sequences are extracted from databases in GenBank (*UNIT 1.3*; *APPENDIX 1B*), and are aligned against the genome using the BLAT search tool (see Basic Protocol). The data undergoes several filtering steps (detailed on the individual track description pages) prior to its presentation in the Genome Browser. As mentioned in the Guidelines for Understanding Results section, these filtering methods reduce the occurrence of misleading and erroneous data in the tracks at the expense of eliminating some genuine data. The mRNA data in the Genome Browser are incrementally updated from GenBank nightly; EST data are updated weekly.

Gene prediction tracks within the UCSC Genome Browser vary in the evidence used for genes they report, their coverage of bases in known coding regions, and their specificity. The Known Genes track (Hsu et al., 2006) is generated at UCSC by a fully automated process that combines protein evidence from UniProt and mRNA evidence from GenBank. The proteins are paired with the associated mRNA sequences, which are then aligned to the genome. The best-scoring mRNAs are retained as the Known Genes set and serve as the basis for the Gene Sorter and Proteome Browser. In 2005 the Known Genes process was updated to increase the quality and coverage through more stringent filtering and the inclusion of more supporting evidence (refer to the Known Genes description page for more details). Other UCSC-generated gene prediction tracks of note include the RefSeq Genes track, based on human RefSeq mRNAs in GenBank that have been aligned against the genome with BLAT and stringently filtered; the Mammalian Gene Collection (MGC) Genes track, showing genes for which high-quality clones are available from the MGC Project, and the Consensus CDS (CCDS) Genes track, which shows a high-quality, consistently annotated core set of human protein-coding genes obtained from the CCDS project (*http://www.ncbi.nlm.nih.gov/CCDS/*) and identified by consensus among the Ensembl, Vega (Ashurst et al., 2005), and RefSeq gene annotation sets.

The Genome Browser provides a wealth of comparative genomics annotations. In addition to the cross-species homology mRNA and EST tracks found in the mRNA and EST group, the Comparative Genomics group contains a wide variety of chain and net pairwise alignment tracks (Kent et al., 2003; Schwartz et al., 2003) that can be used to look for or-

thologous regions between organisms, large-scale rearrangements, duplications and deletions, and processed pseudogenes. The chain tracks can also be used to identify paralogs. The Conservation track, based on a phylogenetic hidden Markov model called "phastCons" (Siepel et al., 2005), displays a measure of evolutionary conservation among a set of species aligned with multiz (Blanchette et al., 2004). This track is highly customizable, allowing the user to adjust the display to the species of interest and vary several of the graph characteristics. The associated Most Conserved track provides an alternative simplified view of the Conservation track that highlights the parts of the genome that are most likely conserved by purifying selection.

The Expression and Regulation group features microarray and expression data from several sources, conserved transcription factor binding sites (TFBS), regulatory potential scores, DNaseI-hypersensitive sites, microRNA regulatory targets, and more. Several high-level map tracks are included in the Mapping and Sequencing section: FISH clones, which shows the locations of FISH-mapped BAC clones from the BAC Resource Consortium (Cheung et al., 2001) along the draft assembly sequence; the Chromosome Bands, which uses the locations of FISH-mapped clones on the cytogenetic map and the assembly to approximate the Giemsa-stained chromosome bands at an 800-band resolution; the Sequence-Tagged Site (STS) Markers track, which displays the positions of markers used in constructing several genetic, radiation hybridization (RH), and yeast artificial chromosome (YAC) maps, as well as markers from the UniSTS database; and the BAC End and Fosmid End Pairs tracks, which show mappings of paired BAC and fosmid end pairs.

The Variation and Repeats section features several simple nucleotide polymorphism (SNP) and HapMap annotations (The International HapMap Consortium, 2003, 2005), regions of putative genomic duplications, and repeats identified by RepeatMasker (Smit, 1999) and Tandem Repeats Finder (Benson, 1999).

Critical Parameters and Troubleshooting

Use caution when interpreting the information displayed in the UCSC Genome Browser, particularly if the chromosomal region under scrutiny is incomplete. The Genome Browser annotation tracks are generated from publicly

Using Biological Databases

1.4.19

available data, and therefore are only as accurate as the data on which they are based. Assembly errors and sequence gaps may occur well into the genome sequencing process due to regions that are intrinsically difficult to sequence, and incorrect data may be propagated into the public databases. The browser cannot fill in sequencing gaps or correctly assign strand information in the absence of good coverage data. Artifactual duplications arise as unavoidable compromises during a build, causing misleading matches in genome coordinates found by alignment. Conclusions about the data should never be made based on the information available in a single track. Instead, gather supporting evidence and correlation from other tracks aligned to the same region to identify problematic areas. Cross-check information in the public databases such as Entrez Gene and OMIM (*UNIT 1.2*).

A common source of confusion among users is the positional differences that result when genome assembly versions are interchanged. New genome versions are added to the UCSC Genome Browser on a regular basis. Unless a feature lies on a completely sequenced and unrevised chromosome, its coordinates are likely to change between one assembly and the next. Often the position of a genomic feature cited in the literature will not coincide with the location displayed in the browser. When faced with such a discrepancy, compare the assembly date of the genome in the reference with that of the genome displayed in the browser. In most cases, the newer assembly will have the most accurate information. When feasible, it is usually best to work with the most current assembly even if it lacks a complete set of annotation tracks. Two procedures are described (see Basic Protocol, steps 14, 15, and 16) that can be used to map the position of genomic sequence in one assembly version to that of a newer version.

UCSC makes a concerted effort to provide uninterrupted Browser and BLAT service to the research community. In the event of the occasional power or equipment failure, there are multiple mirror sites that replicate the UCSC Genome Browser environment. To view a list of actively maintained mirror sites, click the Mirrors link on the UCSC Genome Bioinformatics home page.

Troubleshooting custom annotation track problems

Custom annotation track display problems usually stem from syntax or formatting errors in the annotation track file. A spurious line break in one of the browser, track, or data lines is a frequent source of errors. Another common cause of problems is GFF or GTF data that are separated by spaces rather than tabs. Refer to the troubleshooting section in the Custom Annotation Track section of the User's Guide (*http://genome.ucsc.edu/goldenPath/help/customTrack.html*) for more information.

Suggestions for Further Analysis

The UCSC Genome Bioinformatics home page offers links to several tools that facilitate analysis of the genomic and annotation data underlying the browser's graphical presentation. The Table Browser and BLAT tools were introduced in the main part of the unit (see Basic Protocol). The BLAT tool can be used for a large number of functions, such as finding the genomic coordinates of an mRNA or protein in an assembly, determining the exon structure of a gene, displaying a coding region within a full-length gene, searching for gene family members, or finding homologs of a query from another species. The output of a BLAT or Table Browser search can be saved in a custom track format (PSL or BED, respectively) for direct upload into the browser, or can be downloaded into a spreadsheet or text editor (*APPENDIX 1C*) for further manipulation.

The Gene Sorter

The Gene Sorter is accessible from the top menu bar on most of the browser Web pages. It provides a simple interface for studying the relationship among a group of genes based on protein-level homology, the similarity of gene expression profiles, genomic proximity, or other parameters, which in turn facilitates the study of the evolution of genes and their functions. This tool can be used to gather a collection of genes that share similar properties for statistical analysis or to filter a large group of genes into a small subset of interesting features, based on specific properties.

The VisiGene image browser

The VisiGene image browser is available from a link on the Genome Bioinformatics Group home page. It can be used to browse images from in situ RNA hybridization, reporter genes, and other techniques that show where a gene, enhancer, or promoter is active in an organism. In early 2006 the VisiGene image database contained nearly 100,000 images from several high-throughput gene projects, as well as images from literature curated by the model organism databases.

The Proteome Browser

The Proteome Browser can be accessed from the home page or Genome Browser Known Genes track details pages for selected assemblies. It provides a large variety of information about individual protein characteristics such as polarity, hydrophobicity, amino acid anomalies, domains, exons, and much more, displayed as a series of tracks and histograms. This browser is tightly integrated into the Genome Browser Known Genes annotation and in turn links to several external databases and Web sites containing related information.

The in silico PCR utility

The in silico PCR utility is available from the menu bar on most of the Genome Browser Web pages. It provides a means to quickly search a sequence database with a pair of PCR primers, returning a FASTA output file that contains all sequence in the database that lie between and include the primer pair.

Additional resources for Genome Browser information

In addition to the analytical tools available through the Genome Browser, the track description and details pages provide links to many external resources that present a wealth of related information. For a demonstration of the use of the Genome Browser in comparative genomics analysis, see Bejerano et al. (2005).

Three active mailing lists provide sources for Genome Browser information. The genome@soe.ucsc.edu mailing list provides a moderated discussion forum about the Genome Browser software, databases, genome assemblies, and related tools. The genome-announce@soe.ucsc.edu mailing list provides announcements of data and software releases, system maintenance. Finally, the genome-mirror@soe.ucsc.edu mailing list offers a moderated discussion forum for Genome Browser mirror sites.

Online training materials and tutorials on the Genome Browser are available via the "Training" link on the home page.

In upcoming years, the Genome Browser will provide more support for the visualization and analysis of medical sequence data, including genotype/phenotype data sets, and new types of whole-genome data sets, such as genome-wide ChIP/CHIP data. The authors also plan to expand and enhance the array of comparative genomics analytical tools and data.

Literature Cited

Ashurst, J.L., Chen, C.K., Gilbert, J.G., Jekosch, K., Keenan, S., Meidl, P., Searle, S.M., Stalker, J., Storey, R., Trevanion, S., Wilming, L., and Hubbard, T. 2005. The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res.* 33:D459-D465.

Bejerano, G., Siepel, A.C., Kent, W.J., and Haussler, D. 2005. Computational screening of conserved genomic DNA in search of functional noncoding elements. *Nat. Methods* 2:535-545.

Benson, G. 1999. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* 27:12-17.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. 2005. GenBank. *Nucleic Acids Res.* 33:D34-D38.

Birney, E., Andrews, D., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X.M., Flicek, P., Graf, S., Hammond, M., Herrero, J., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Kokocinski, F., Kulesha, E., London, D., Longden, I., Melsopp, C., Meidl, P., Overduin, B., Parker, A., Proctor, G., Prlic, A., Rae, M., Rios, D., Redmond, S., Schuster, M., Sealy, I., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Stabenau, A., Stalker, J., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C., and Hubbard, T.J. 2006. Ensembl 2006. *Nucleic Acids Res.* 34:D556-D561.

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., Haussler, D., and Miller, W. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* 14:708-715.

Cheung, V.G., Nowak, N., Jang, W., Kirsch, I.R., Zhao, S., Chen, X.N., Furey, T.S., Kim, U.J., Kuo, W.L., and Livier, M. 2001. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* 409:953-958.

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.,

Using Biological Databases

1.4.21

Green, R.K., Flippen-Anderson, J.L., Westbrook, J., Berman, H.M., and Bourne, P.E. 2005. The RCSB Protein Data Bank: A redesigned query system and relational database based on the mmCIF schema *Nucleic Acids Res*. 33:D233-D237.

Eeckman, F.H. and Durbin, R. 1995. ACeDB and Macace. *Methods Cell Biol*. 48:583-605.

The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306:636-640.

Eyre, T.A., Ducluzeau, F., Sneddon, T.P., Povey, S., Bruford, E.A., and Lush, M.J. 2006. The HUGO Gene nomenclature database, 2006 updates. *Nucleic Acids Res*. 34:D319-D321.

Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S.R., Sonnhammer, E.L., and Bateman, A. 2006. Pfam: Clans, Web tools and services. *Nucleic Acids Res*. 34:D247-D251.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 33:D514-D517.

Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., Hillman-Jackson, J., Kuhn, R.M., Pedersen, J.S., Pohl, A., Raney, B.J., Rosenbloom, K.R., Siepel, A., Smith, K.E., Sugnet, C.W., Sultan-Qurraie, A., Thomas, D.J., Trumbower, H., Weber, R.J., Weirauch, M., Zweig, A.S., Haussler, D., and Kent, W.J. 2006. The UCSC Genome Browser database: Update 2006. *Nucleic Acids Res*. 34:D590-D598.

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.

Hsu, F., Kent, W.J., Clawson, H., Kuhn, R.M., Diekhans, M., and Haussler, D. 2006. The UCSC known genes. *Bioinformatics* 22:1036-1046.

The International HapMap Consortium. 2003. The International HapMap Project. *Nature* 426:789-796.

The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* 437:1299-1320.

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., Weber, R.J., Haussler, D., and Kent, W.J. 2003. The UCSC Genome Browser database. *Nucleic Acids Res*. 31:51-54.

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.

Kent, W.J. 2002. BLAT - the BLAST-like alignment tool. *Genome Res*. 12:656-664.

Kent, W.J. and Zahler, A.M. 2000a. The intronerator: Exploring introns and alternative splicing in C. elegans. *Nucleic Acids Res*. 28:91-93.

Kent, W.J. and Zahler, A.M. 2000b. Conservation, regulation, synteny, and introns in a large-scale *C. briggsae-C. elegans* genomic alignment. *Genome Res*. 10:1115-1125.

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.

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. U.S.A.* 100:11484-11489.

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.

Lenhard, B., Hayes, W.S., and Wasserman, W.W. 2001. GeneLynx: A gene-centric portal to the human genome. *Genome Res*. 11:2151-2157.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., Copley, R., Courcelle, E., Das, U., Durbin, R., Fleischmann, W., Gough, J., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McDowall, J., Mitchell, A., Nikolskaya, A.N., Orchard, S., Pagni, M., Ponting, C.P., Quevillon, E., Selengut, J., Sigrist, C.J., Silventoinen, V., Studholme, D.J., Vaughan, R., and Wu, C.H. 2005. InterPro, progress and status in 2005 *Nucleic Acids Res*. 33:D201-D205.

Pieper, U., Eswar, N., Davis, F.P., Braberg, H., Madhusudhan, M.S., Rossi, A., Marti-Renom, M., Karchin, R., Webb, B.M., Eramian, D., Shen, M.Y., Kelly, L., Melo, F., and Sali, A. 2006. MODBASE: A database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res*. 34:D291-D295.

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.

Safran, M., Chalifa-Caspi, V., Shmueli, O., Olender, T., Lapidot, M., Rosen, N., Shmoish, M., Peter, Y., Glusman, G., Feldmesser, E., Adato, A., Peter, I., Khen, M., Atarot, T., Groner, Y., and Lancet, D. 2003. Human gene-centric databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res*. 31:142-146.

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

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., Weinstock,

G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W., and Haussler, D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034-1050.

Smit, A.F. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Gen. Dev.* 9:657-663.

Strausberg, R.L., Greenhut, S.F., Grouse, L.H., Schaefer, C.F., and Buetow, K.H. 2001. In silico analysis of cancer through the Cancer Genome Anatomy Project. *Trends Cell Biol.* 11:S66-S71.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S.E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M.R., Brown, D.G., Brown, S.D., Bult, C., Burton, J., Butler, J., Campbell, R.D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A.T., Church, D.M., Clamp, M., Clee, C., Collins, F.S., Cook, L.L., Copley, R.R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K.D., Deri, J., Dermitzakis, E.T., Dewey, C., Dickens, N.J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D.M., Eddy, S.R., Elnitski, L., Emes, R.D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G.A., Flicek, P., Foley, K., Frankel, W.N., Fulton, L.A., Fulton, R.S., Furey, T.S., Gage, D., Gibbs, R.A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T.A., Green, E.D., Gregory, S., Guigo, R., Guyer, M., Hardison, R.C., Haussler, D., Hayashizaki, Y., Hillier, L.W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D.B., Johnson, L.S., Jones, M., Jones, T.A., Joy, A., Kamal, M., Karlsson, E.K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W.J., Kirby, A., Kolbe, D.L., Korf, I., Kucherlapati, R.S., Kulbokas, E.J., Kulp, D., Landers, T., Leger, J.P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D.R., Mardis, E.R., Matthews, L., Mauceli, E., Mayer, J.H., McCarthy, M., McCombie, W.R., McLaren, S., McLay, K., McPherson, J.D., Meldrim, J., Meredith, B., Mesirov, J.P., Miller, W., Miner, T.L., Mongin, E., Montgomery, K.T., Morgan, M., Mott, R., Mullikin, J.C., Muzny, D.M., Nash, W.E., Nelson, J.O., Nhan, M.N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M.J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K.H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C.S., Poliakov, A., Ponce, T.C., Ponting, C.P., Potter, S., Quail, M., Reymond, A., Roe, B.A., Roskin, K.M., Rubin, E.M., Rust, A.G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M.S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J.B., Slater, G., Smit, A., Smith, D.R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla,

C., Ureta-Vidal, A., Vinson, J.P., Von Niederhausern, A.C., Wade, C.M., Wall, M., Weber, R.J., Weiss, R.B., Wendl, M.C., West, A.P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R.K., Winter, E., Worley, K.C., Wyman, D., Yang, S., Yang, S.P., Zdobnov, E.M., Zody, M.C., and Lander, E.S. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520-562.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L.Y., Helmberg, W., Kapustin, Y., Kenton, D.L., Khovayko, O., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Pruitt, K.D., Schuler, G.D., Schriml, L.M., Sequeira, E., Sherry, S.T., Sirotkin, K., Souvorov, A., Starchenko, G., Suzek, T.O., Tatusov, R., Tatusova, T.A., Wagner, L., and Yaschenko, E. 2006. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 34:D173-D180.

Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Mazumder, R., O'Donovan, C., Redaschi, N., and Suzek, B. 2006. The Universal Protein Resource (UniProt): An expanding universe of protein information. *Nucleic Acids Res.* 34:D187-D191.

Key References

Kent et al., 2002. See above.

A description of the UCSC Genome Browser tool and the underlying conceptual and technical framework.

Hinrichs et al., 2006. See above.

The 2006 update of Kent et al. (2002) that includes software enhancements and additions, new genome assemblies, and new annotations.

Internet Resources

http://genome.ucsc.edu

The UCSC Genome Bioinformatics and Genome Browser home page.

http://hgdownload.cse.ucsc.edu/downloads.html

The UCSC Genome Browser downloads server.

http://genome-mysql.cse.ucsc.edu

The Genome Browser public MySql server.

http://genome.ucsc.edu/goldenPath/help/
    hgTracksHelp.html

The UCSC Genome Browser User's Guide.

http://genome.ucsc.edu/goldenPath/help/
    hgTablesHelp.html

The UCSC Table Browser User's Guide.

http://genome.ucsc.edu/goldenPath/help/
    customTrack.html

Information for constructing and uploading a custom annotation track.

genome@soe.ucsc.edu

Mailing list for questions and discussions about the browser software, database, and genome assemblies.

Using Biological Databases

1.4.23

genome-announce@soe.ucsc.edu

Mailing list for announcements about releases of browser software and data, server maintenance, etc.

genome-mirror@soe.ucsc.edu

Mailing list for questions and discussion about mirroring the UCSC Genome Browser.

Contributed by Donna Karolchik, Angie S. Hinrichs, and W. James Kent
University of California Santa Cruz
Santa Cruz, California