The UCSC Genome Browser: What Every Molecular Biologist Should Know

Mary E. Mangan,[1] Jennifer M. Williams,[1] Robert M. Kuhn,[2] and Warren C. Lathe III[1]

[1]OpenHelix LLC, Bainbridge Island, Washington
[2]Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, California

ABSTRACT

Electronic data resources can enable molecular biologists to quickly get information from around the world that a decade ago would have been buried in papers scattered throughout the library. The ability to access, query, and display these data makes benchwork much more efficient and drives new discoveries. Increasingly, mastery of software resources and corresponding data repositories is required to fully explore the volume of data generated in biomedical and agricultural research, because only small amounts of data are actually found in traditional publications. The UCSC Genome Browser provides a wealth of data and tools that advance understanding of genomic context for many species, enable detailed analysis of data, and provide the ability to interrogate regions of interest across disparate data sets from a wide variety of sources. Researchers can also supplement the standard display with their own data to query and share this with others. Effective use of these resources has become crucial to biological research today, and this unit describes some practical applications of the UCSC Genome Browser. *Curr. Protoc. Mol. Biol.* 107:19.9.1-19.9.36. © 2014 by John Wiley & Sons, Inc.

Keywords: UCSC Genome Browser • primers • custom tracks • variations • SNP • comparative genomics • ENCODE

INTRODUCTION

Sequence databases and software analysis tools are now crucial reagents for molecular biologists. Increasing volumes of sequence data and the diverse annotation types that are necessary to understand them in the genomic context present a challenge to research. The data need to be organized and presented effectively to be optimally useful. Genome browsers accomplish this task in a variety of ways. Some browsers are Web-based, publicly available tools that form a centralized community resource function. Examples of this include the UCSC Genome Browser, ENSEMBL, NCBI's Sequence Viewer, some installations of GBrowse for large research communities, and others (Meyer et al., 2013; Flicek et al., 2013; NCBI Resource Coordinators, 2013; Stein, 2013). Generally managed by well-resourced large project and support teams, they may provide integration of a range of species and data collections. Frequently they also provide customization options such as filtering and/or uploading user data to view in the context of existing data. However, they may lack some flexibility for specific users' needs. Other browsers, or versions of the browsers described above, may be locally installed or stand-alone software tools for users to interact with their species of focus, and with only their own project data, which can be supplemented with appropriate public data sets. These may offer more customization options, and sometimes additional support for alternative file formats, which can be beneficial to users. Also, locally managed security can be an issue for some types of unpublished or human patient–related data. Examples of these types of browsers include Integrative Genome Viewer (IGV), Integrated Genome Browser

Current Protocols in Molecular Biology 19.9.1-19.9.36, July 2014
Published online July 2014 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/0471142727.mb1909s107

(IGB), Gaggle Genome Browser (GGB), and others (Nicol et al., 2009; Bare et al., 2010; Thorvaldsdóttir et al., 2013).

The focus of this unit is the University of California Santa Cruz (UCSC) Genome Browser's organization and tools that provide support for molecular biomedical researchers worldwide (Meyer et al., 2013). Like reagents on a chemical shelf, sequences and associated data need to be obtained, extracted, manipulated, combined, analyzed, and used to further the progress of research.

The UCSC Genome Browser (*http://genome.ucsc.edu*) provides a framework for interpretation of genomic features and elements, with a graphical interface and a user-friendly access mechanism for complex queries as well as batch data retrieval and downloading (Table Browser; Kuhn et al., 2013). Genomic data are visualized on a reference genome sequence framework, with data laid out graphically along the sequence coordinate axis. These datasets are called "annotation tracks" and provide context for the genomic regions (Fig. 19.9.1A). A MySQL database stores the complete data collection, which is used to display the features with a graphical genome viewer interface or can be queried and downloaded with the Table Browser (Fig. 19.9.1B). Additional tools associated with, and integrated into, the UCSC Genome Browser provide access to related specific data types that offer visualization and analysis suitable for special topic areas and investigations (Genome Graphs, in silico PCR, a Variant Annotation Integrator, VisiGene, ENCODE Portal, a Cancer Genome Browser, and more). The two most foundational of these tools are explored in some detail here, which will allow researchers to begin accessing the data efficiently and capably (via the Genome Browser graphical viewer and the Table Browser). Other tools are mentioned briefly to encourage further exploration.

The UCSC Genome Browser provides access to dozens of species' genomic sequences and other data types at the main site in Santa Cruz, California (*http://genome.ucsc.edu*), an official European mirror (*http://genome-euro.ucsc.edu*), and at several regional mirrors intended as disaster backup (*http://genome.ucsc.edu/mirror.html*). The focus here is heavily on mammalian data, but other organisms are available. Other installations of the same software framework have been deployed for other species including archaea (*http://archaea.ucsc.edu/*), malaria parasites (*http://areslab.ucsc.edu/*), HIV (*http://www.gsid.org/*), Cannabis (*http://genome.ccbr.utoronto.ca/*), the Sticklebrowser (*http://sticklebrowser.stanford.edu/*) and more. In each case, the species sets and data types comprising the site may vary, but the basic functionality of the software features is largely the same. The focus of the protocols below is the UCSC main site in Santa Cruz, but mastery of the concepts should enable use of any of the sites.

BASIC PROTOCOL 1

UNDERSTANDING GENOMIC DATA AND FEATURES WITH THE UCSC GENOME BROWSER GATEWAY

To begin to understand the search and display features of the UCSC Genome Browser, access the Gateway. This is the access point to the main browser visualization software. Use nearly any of the major Internet browsers to access the site. A fast connection speed is recommended, but should not be required.

1. Access the UCSC Genome Browser at the URL *http://genome.ucsc.edu*.

   This will display the site homepage. Project description information and news will be in the central page area. Navigation bars (blue) on the left and on the top will provide the entry points for the tools featured in this unit.

2. In the blue navigation areas, click either the top link for Genomes, or the side link for Genome Browser.

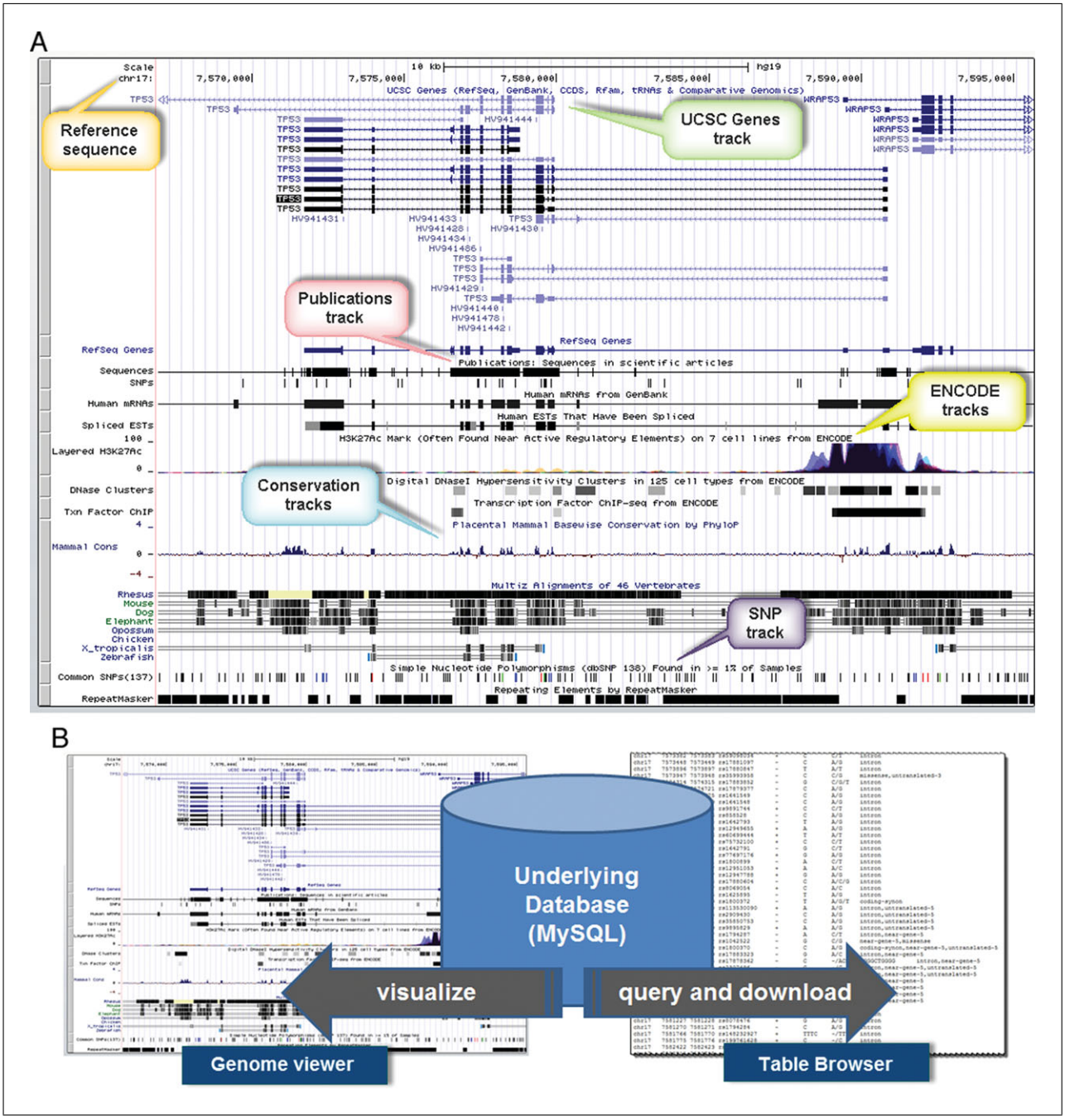


Figure 19.9.1 (**A**) The UCSC Genome Browser organizes the display of genomic information with the official or reference genome sequence as the framework, positioning additional data as "annotation tracks" in the appropriate genomic location to provide context for understanding any genomic region. Default tracks for the human genome assembly February 2009 (GRCh37/hg19) are shown. (**B**) An underlying MySQL database stores genomic sequence data, annotation track details, and auxiliary information which can then be visualized with the graphical Genome viewer interface, or queried and downloaded using the Table Browser interface.

These links provide the same outcome: landing on the Genome Browser Gateway page (Fig. 19.9.2). The Gateway should say Human (Homo sapiens) Genome Browser Gateway. If it does not, then your computer has been here before; click the link "Click here to reset" to return the site to the default settings. Pull-down menu options provide access to other organisms.

3. Examine the Gateway page for the hg19 Human assembly.

   a. A query box area is found at the top ("search term"). It will allow many types of identifiers or coordinates.

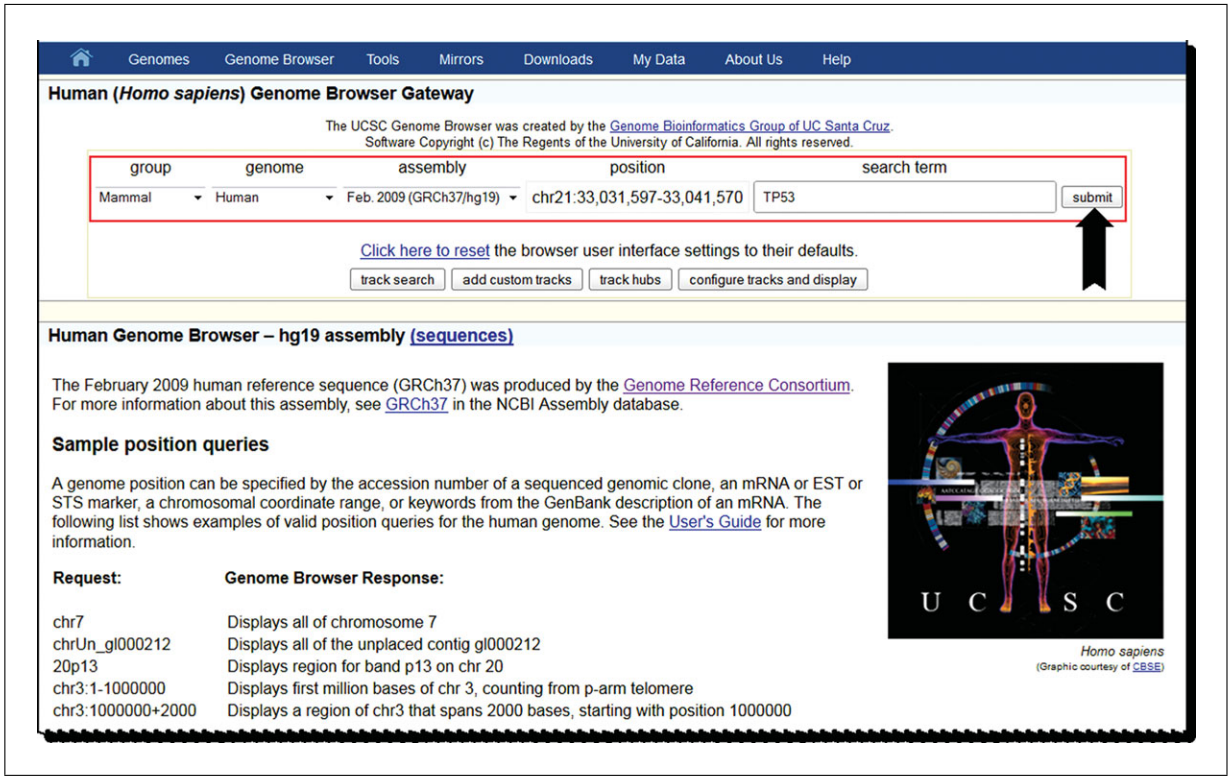


Figure 19.9.2 The Gateway query box provides the entry point for basic text queries of the UCSC Genome Browser and offers access to the Genome viewer. Details and links about the current genome assembly for that species are provided. Helpful formatting notes are offered to guide additional queries.

b. In the lower area, there will be information about the source of the genome assembly, as selected in the pull-down menus above. The current default is the hg19 Human assembly.
c. A section of "Sample position queries" helps to remind users of the appropriate syntax for an inquiry (Request) and the expected outcome (Genome Browser Response).

> *For example, entire chromosomes, nucleotide ranges, or items using names or various identifiers (IDs) provided by many database sources can be employed as query items.*

d. The Assembly Details section provides more detailed information about the specific genomic sequence at hand. Links to statistics about the "build" or version are provided. Notable aspects of the data that may affect interpretation of the results are offered.

4. Focus on the query box area at the top. Select options here to define the query. Examine each option from left to right.

a. *group:* In order to reduce the species list size as more and more species were added, the UCSC Genome Browser team has broken the collection into smaller subsets as a way to bundle the species into useful groups. Access a species of interest by choosing the appropriate item from the group list.
b. *genome:* Select the species to examine.
c. *assembly:* The date in this box represents the date of the official freeze and deposit into GenBank of the official sequence by the sequencing center. The human sequence through the March 2006 assembly was provided by the International Human Genome Consortium, but as of February 2009, the assembly sequence has been obtained from the Genome Reference Consortium (GRC;

Church et al., 2011). Note that different species have different sources of official sequence data. The official assembly sequence is frozen and will not change during the course of one "assembly" date. By default, usually the most current version is shown, but access to earlier versions is available as well. As genome assemblies mature, gaps are filled and mis-assembled regions are corrected, but each assembly should be considered an improvement on the same genome. Sometimes it is useful to reproduce queries done on previous versions, as when trying to replicate information found in older publications. Generally, several assemblies are available for any species from the menu. For species that have had many assembly releases, the oldest assemblies are available in the archives. These are accessed from the homepage left navigation bar link for "Archives."

d. *position:* A default position is given for each assembly, or the last location visited is indicated. Clicking on this chromosomal coordinate causes the position to be copied to the search term box, where it may be edited. The standard representation for chromosome coordinates takes the form of the abbreviation "chr," the chromosome number, a colon, and the nucleotide numbers of the start and end positions for the item, such as: chr21:33,031,597-33,041,570.
e. *search term:* This is where to enter the information to define the region of the genome to examine. Enter gene symbols, names, keywords, authors, genome coordinate nucleotide numbers or ranges, cytological bands, and many types of IDs and accession numbers. Examples are given lower down on the page. Typing in a gene name will lead to suggestions of gene names beginning with the characters typed.

The following are some other features of the query box area that may assist interactions with the browser.

f. *Click here to reset:* This feature will clear any settings in the browser that pertain to the UCSC Genome Browser choices that are made. This may be useful in shared computing environments such as laboratories and libraries when others may have changed features of the display. Occasionally, odd behavior from the software may be encountered and sometimes it will help to reset from this point.
g. *track search:* This button provides access to a search function that is useful for finding the specific data types that are of interest, and that a researcher would like to view on the genome display. The UCSC Genome Browser offers a wide range of information to researchers in the form of annotation tracks, with new tracks and data types being added continuously. This search feature allows researchers to more easily identify specific tracks of data that may be useful to their research.
h. *add custom tracks and track hubs:* These buttons will enable addition of one's own data to the display (see Basic Protocol 3 for more information on custom tracks and Basic Protocol 5 for track hubs).
i. *configure tracks and display:* This button enables access to many features of the subsequent display that can be turned on, turned off, or activated. Tracks can be moved around. Change the size of the text in the displays if desired. This button will also be available on the display page later.

5. Now that the options are familiar, perform a sample query to visualize a genomic location to begin to explore the genomic context. The human gene TP53 is of interest for this example. Because TP53 is an important and medically relevant gene that has been implicated in many cancer types, it is well characterized, and is associated with several data types that can be explored, including publications.

6. On the Gateway page (see Fig. 19.9.2), make these choices:

a. *clade:* Mammal
b. *genome:* Human

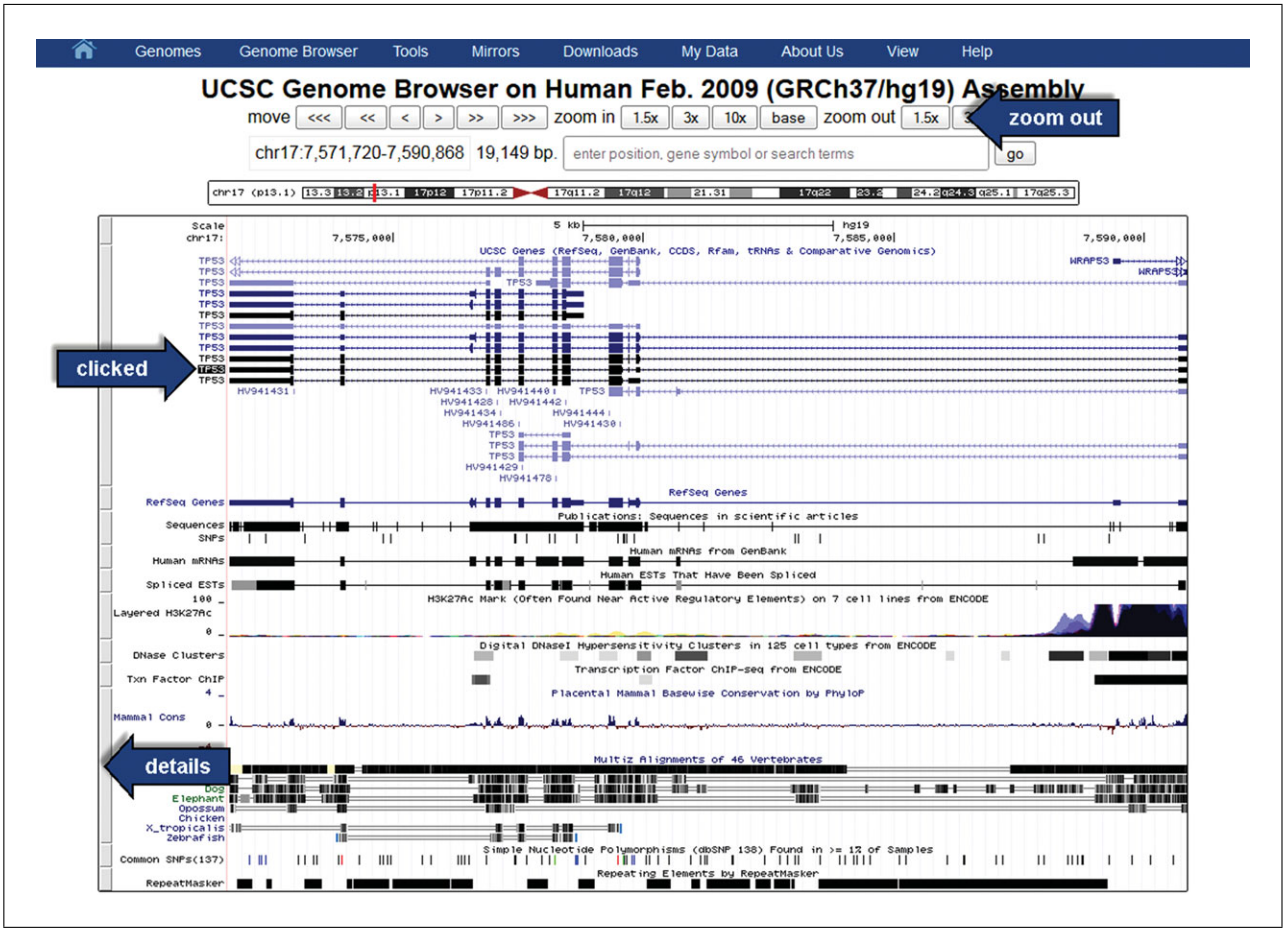


Figure 19.9.3 Overview of a region of the UCSC Genome Browser viewer. The position of the TP53 (uc002gij.3) gene that was clicked from the results is indicated with the arrow (labeled "clicked"). The browser adds a highlight box around the clicked item to help you identify it on the viewer. The arrow (labeled "details") indicates the bar to click for details about the track data. Zooming out (or in) allows users to change the range overview ("zoom out" arrow).

c. *assembly:* Feb. 2009 (GRCh37/hg19)
d. *search term (this search is not case-sensitive):* TP53
e. You may choose the TP53 from the list that appears to go directly to the gene in the viewer, or, if you choose to continue to type without clicking the auto-complete suggestion, you will be taken to a large page of choices for TP53-related results (see step 7 below).
f. Click the "submit" button when the choices are complete.

7. If you chose not to select the suggested gene, a results page will present a list of all the matches for the text string TP53 in the human data collection. The term will be found in related gene names, mRNAs, and more categories, in addition to entries for the TP53 gene. The appropriate type of result for the search needs to be selected. In this case the TP53 gene is the primary interest. Focus on UCSC Genes, a collection created by UCSC from a variety of data sources of genes that have been aligned to the genomic sequence. At this time, there are several entries for TP53. The description text for some items includes the phrase "TP53 target," so those would not correspond to the TP53 sequence itself. Select a sequence that appears to correspond to the full TP53 coding sequence, for example TP53 (uc002gij.3) at chr17:7571720-7590868. Click that link to go to the viewer, which will load that nucleotide range in the view (Fig. 19.9.3), and it will open the UCSC Genes track. The item clicked from the results list will be indicated by highlighting around the name (see the arrow labeled "clicked" shown on the figure).

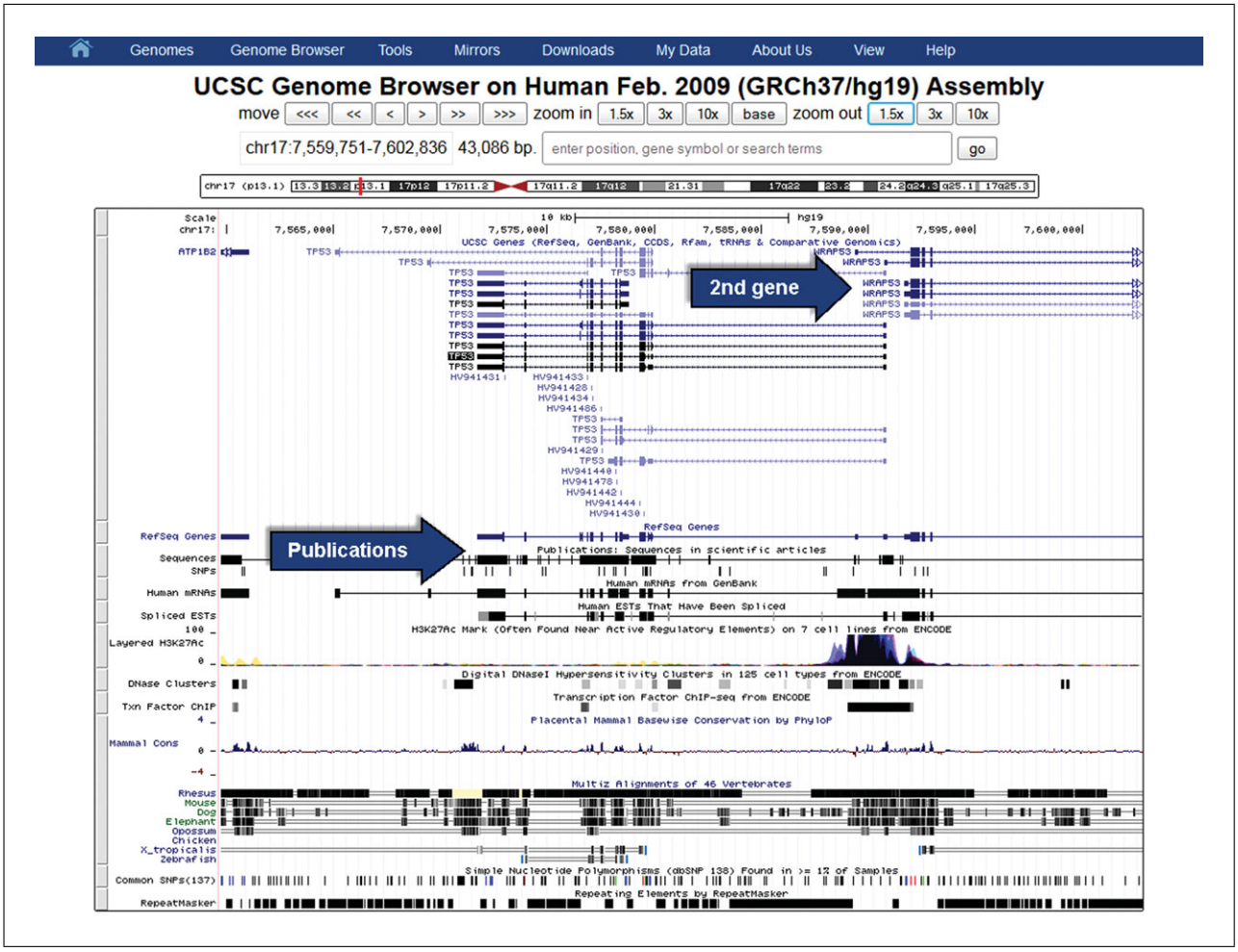


Figure 19.9.4 The neighborhood of the TP53 gene includes a second gene, WRAP53, denoted on the right. The Publications track is highlighted. The Publications track connects this genomic region to instances of that sequence that appear in scientific journal articles.

8. Examine the Genome Viewer in the TP53 region. The Viewer will have reference sequence and tracks in the upper portion of the page, and the lower portion of the page will contain track control menu options.

   The UCSC Genome Browser employs graphical cues to help users understand features in the view. Some short items will be represented with tick marks (SNPs); gene-structure features such as exons and introns are indicated as boxes and lines. On genes such as TP53, the direction of transcription is indicated with arrowheads, exons are indicated as thick boxes, and untranslated transcript features are indicated as half-height boxes. Evolutionary relationships (conservation) are indicated with a two-dimensional display. Any of the display features can be configured by clicking on the gray bars on the left of the image (indicated by the "details" arrow in Fig. 19.9.3), or by locating the annotation track name in the control area below the viewer and clicking the hyperlink. Description pages will provide information about the graphical cues and color codes. Details about individual items in the annotation tracks can be obtained by clicking on those items directly.

9. A great deal of information is provided by the default view (Fig. 19.9.3). To see a view of the larger area in which the TP53 gene is located, click the "zoom out" 1.5x button once. The button is indicted with the "zoom out" arrow in Figure 19.9.3. The new view is shown in Figure 19.9.4. Now it is immediately apparent that the TP53 region of the genome is home to other neighboring genes as well, one of which is indicated with the "2nd gene" arrow in Figure 19.9.4. It can also be seen that this

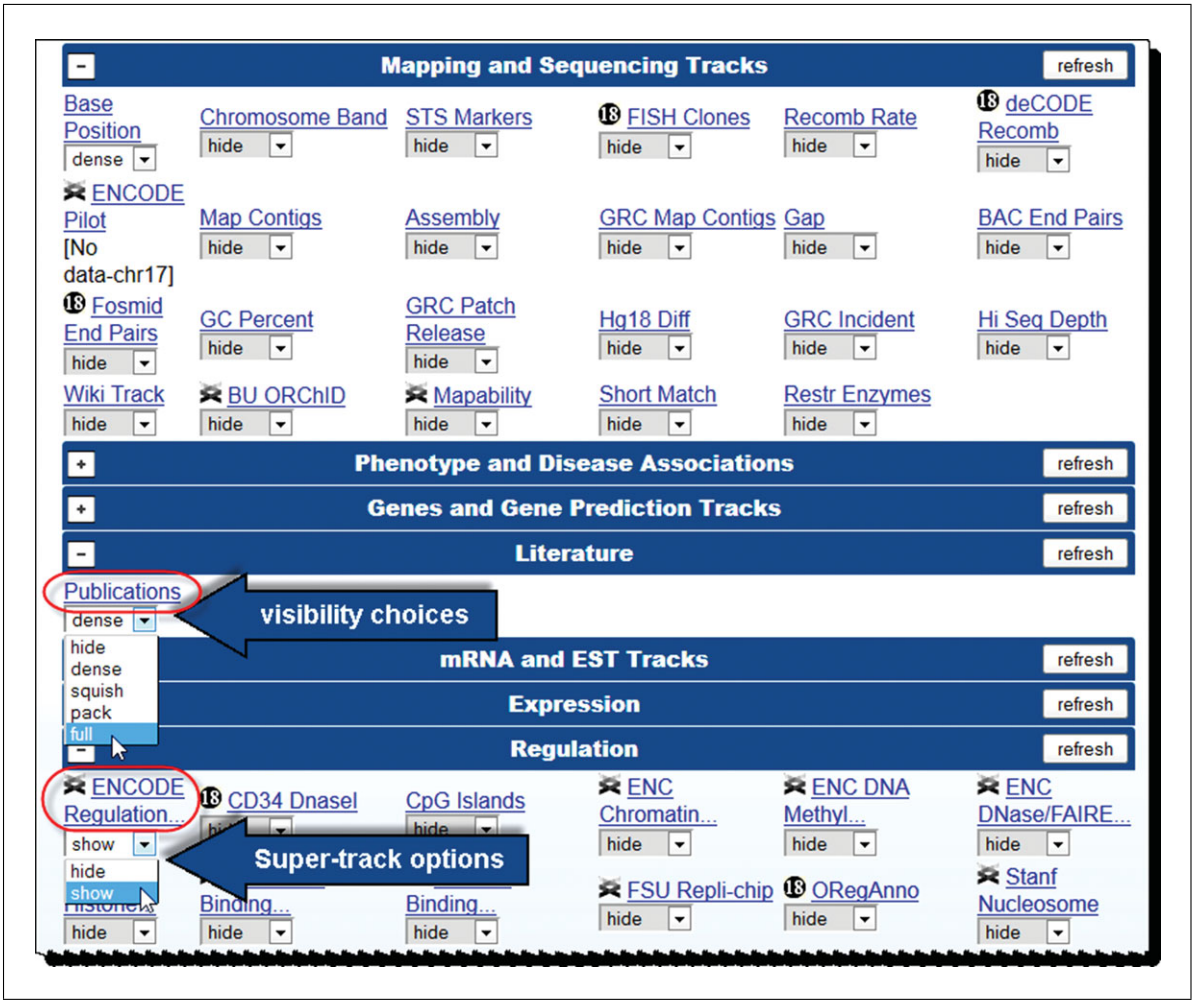


Figure 19.9.5 A portion of the track control menu area is shown. An open menu for a single track offers a variety of display visibility choices. "Hide" removes the track from view; "dense" collapses all the data into a single line; "squish" uses half-sized graphics; "pack" efficiently positions each item (several may share a line if room permits); and "full" puts each data item on a separate line. In the case of a "Super-track" that combines several tracks into a coordinated set, the choices will be "show" and "hide" for the set. However, each component track will still retain the full options. The full options will be available on the hyperlinked description page for the super-track. Choices for the visibility will depend on the data type and one's needs. An oval indicates a clickable hyperlink that provides details about the data contents of the track or the super-track.

area of the human genome is associated with a large body of research publications, as would be expected for such a medically relevant gene. The track for this corpus of literature is indicated with the "Publications" arrow in Figure 19.9.4.

10. The default view shows areas of the genome associated with scientific publications, but it does not show individual titles, or authors. For that, the visibility setting of the Publications track will need to be changed. Examine the track controls section in the lower part of the page (Fig. 19.9.5). The uppermost section is called Mapping and Sequencing Tracks. Below that are other groups of tracks, including the Phenotype and Literature group. In that group, there is a track called Publications. In this case, it seems apparent that this track will have citation information, but if that cannot be determined from the short names, or the user would like to confirm the expectations for the data in that set, clicking the hyperlinked title (circled in figure) provides access to a page that describes the data. The data type, source, any associated publications, graphical and color cues, and more will be provided on the track description page.

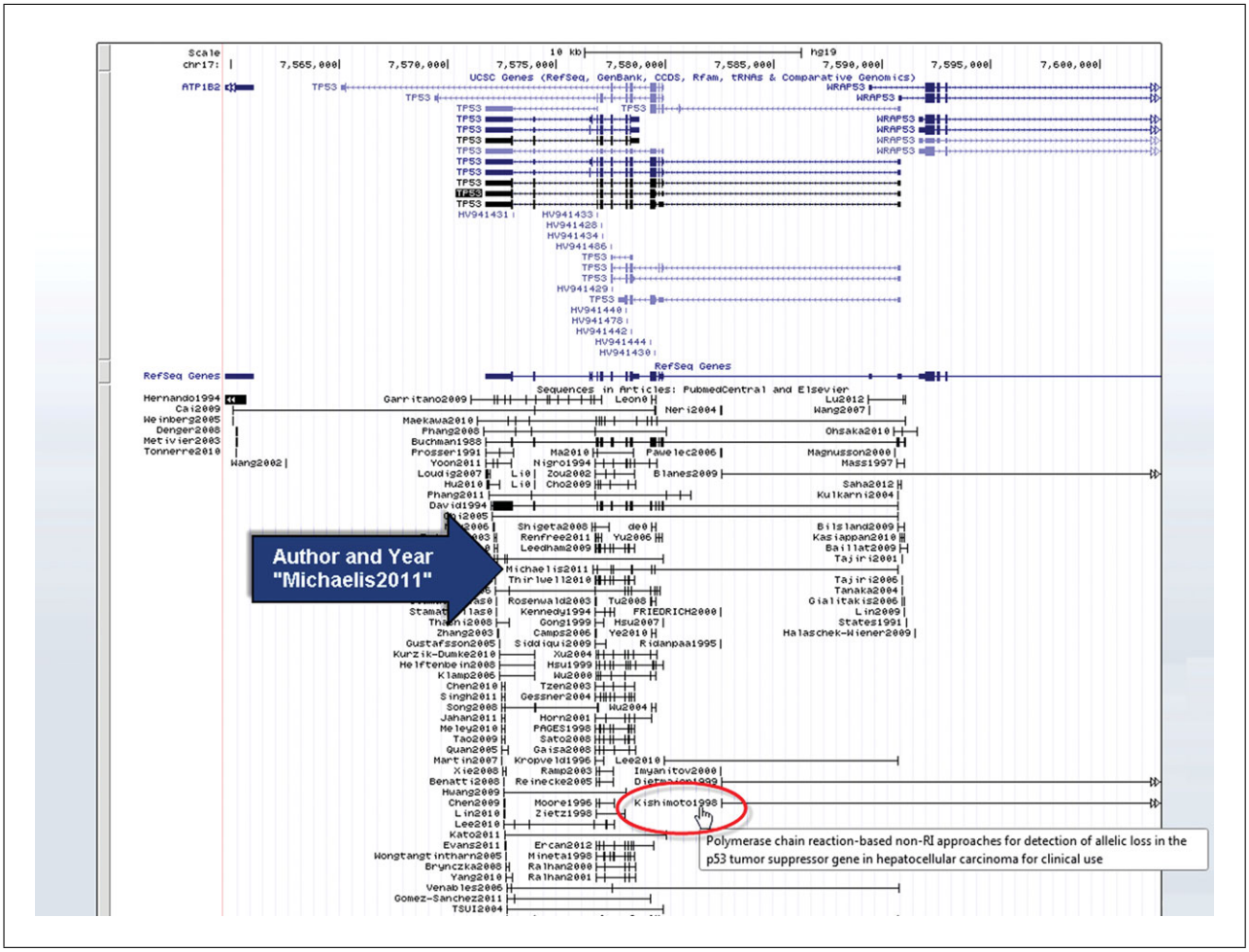


Figure 19.9.6 The data associated with Publications track are shown. The viewer will display a large list of publications aligned to the region that is discussed in the paper. The first author and year of publication are displayed in the view (arrow labeled "Author and Year"). Mousing over an item opens an information popup with the title of the article (circled in the figure).

Pull-down menus for tracks that are turned off are gray ("hide"), while those that are turned on, such as the Publications track here, are white. There are different menu choices for different styles of display. Some are more compressed for smaller graphic views ("dense," "pack," and "squish"), and one is for full display of all the features. For this example, "full" visibility will be chosen, but for different data types, the other choices may be preferred.

11. Select "full" from the "Publications" menu, and then click one of the "refresh" buttons on the page to enable that change and see it in the viewer. The image will now be redrawn in the Viewer (Fig. 19.9.6). The upper section will display a large list of publications aligned to the region that is discussed in the paper. The first author and year of publication are displayed in the view, as indicated with the "author year" arrow in Figure 19.9.6. Mousing over an item opens an information pop-up with the title of the article (circled in the figure). Clicking on a publication provides a page that describes the paper in more detail. As before, links to external sources may provide source information for the citation.

Using the described protocol, publications have been identified that relate directly to a specific region in which the medically relevant TP53 gene resides. Learn more about the data set from the Publications Track Settings page, which will include relevant citations for the source and methods used to obtain this data (Haeussler et al., 2011).

EXPLORE THE ENCODE DATA FOR POTENTIAL REGULATORY ELEMENTS IN A GENOMIC REGION OF INTEREST

The Publications data we saw above could lead to helpful literature associated with the gene or region examined in the viewer. But additional types of genomics data will also be useful to understand a region of interest better. The ENCODE project generated tremendous volumes of genome-wide data with many types of technologies, techniques, and cell lines (Rosenbloom et al., 2013). The Genome Browser was the Data Coordination Center for ENCODE Phase 2, and is the repository for Phase 3. ENCODE data can be explored in the UCSC Genome Browser in large composite tracks that require new visualization strategies and track controls. In this section, we will explore one of these integrated data sets and the associated tracks. For additional details and resources associated with the ENCODE project features at the UCSC Genome Browser, explore the ENCODE Portal (*http://genome.ucsc.edu/ENCODE/*).

1. Back on the view created in Basic Protocol 1 (see Fig. 19.9.6), some features will be adjusted to simplify the view. Return the Publication track to "dense" visibility for now as we continue to examine the region and click a "refresh" button.

2. The view will now contain the default set of tracks and the scale bar. Look at the viewer again and locate the tracks called, in the left label area, "Layered H3K27Ac," "DNase Clusters," and "Txn Factor ChIP" in the lower third of the viewer (Fig. 19.9.7A, red boxed region). These tracks are some of the ENCODE data tracks, as their center labels indicate. They offer details about the presence of a histone modification, DNase hypersensitivity sites, and transcription factor binding sites, respectively. These elements could impact the expression of TP53 or neighboring genes. We can learn more details about the strategies used and data provided in the documentation for the "ENCODE Regulation" track. Scroll down to locate the "ENCODE Regulation" menu in the "Regulation" tracks section of the track controls. (Fig. 19.9.7B).

3. The "ENCODE Regulation" track title link has an adjacent NHGRI icon to indicate that it is a component of the ENCODE project data. The default setting for this menu is "show." Open the menu list to examine the options. You will find that the options are different from the Publications track that we saw before. Here the choices are "show" or "hide" (Fig. 19.9.5). That is because this track is a "super-track," a special type of track developed to offer special controls for collected sets of related information.

   Super-tracks are a special type of track collection. Multiple single tracks (and often, clusters of tracks known as composite tracks) of information can be bundled together into a "super-track" organization with show/hide visibility choices from the main browser page. These tracks may be united by a single technique or technology type, or they could be an integrated collection of data types that could be useful to examine simultaneously. The ENCODE regulation track is the latter type of super-track. Each component track in a super-track is still accessible with its own filters and controls. The underlying tracks are available to examine when the hyperlink for the super-track's name is selected. From the super-track details page, each of the individual tracks visibility may be set, or the accompanying hyperlink for that track will take the user to further options for that track's settings.

4. Examine the underlying component tracks that comprise the ENCODE Regulation super-track.

   a. First, click the hyperlink called "ENCODE Regulation . . . " in the Regulation track group area. (Fig. 19.9.7B).
   b. When the "ENCODE Regulation" Super-track Settings page loads, examine the page layout. The top of the page offers the visibility control pull-down menu for

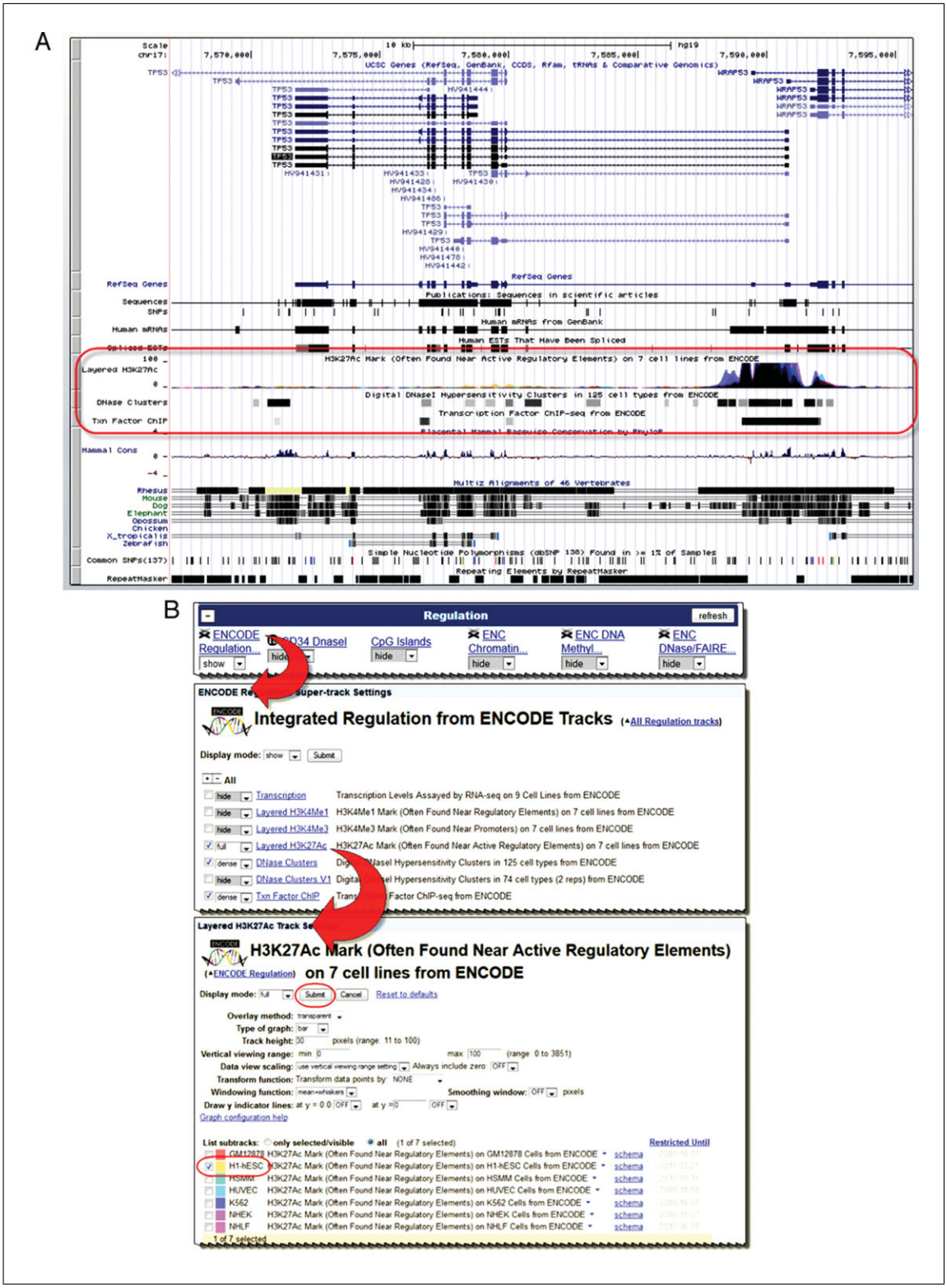


Figure 19.9.7 (**A**) Focus on the ENCODE Regulation super-track area (a histone mark track, a DNase hypersensitivity track, and a transcription factor track, boxed in red). (**B**) The Super-track show/hide menu adjusts the set of tracks as a group (Fig. 19.9.5), but the sub-track controls accessed from the "ENCODE Regulation..." hyperlink offer additional options for setting viewing choices of the component tracks. The ENCODE Regulation sub-track Layered H3K27ac settings page offers various strategies for adjusting the view. (**C**) The default view of the H3K27ac shows all the available cell line histogram signal data (upper). When only one of the cell lines is selected, the browser reflects the choice with yellow graphics only. (**D**) The Txn Factor ChIP track shows signals from various experiments with different transcription factor antibodies (names to the left of a signal) and in different cell lines (letter codes to the right of a signal; red boxed area). For the color version of this figure, go to *http://www.currentprotocols.com/protocol/mb1909*.

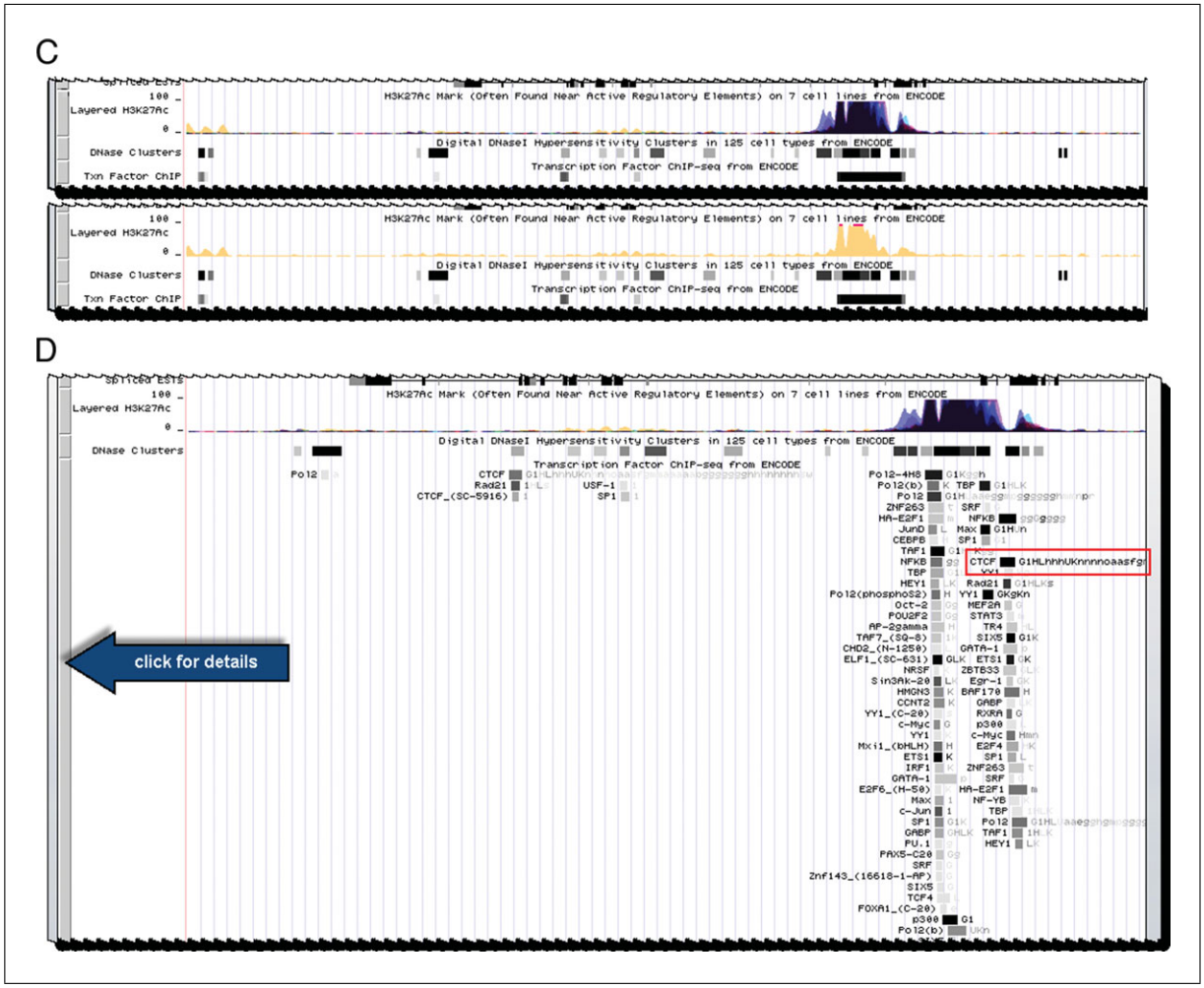


Figure 19.9.7 *Continued*

the super-track display mode. Below that you will find pull-down controls for the individual tracks that comprise this super-track. A click on any of the links next to the pull-down (such as "Layered H3K27Ac") will open a Track Settings configuration page giving access to the individual data tracks.

Some tracks will be visible by default on the browser already (those with white menu displays), and some will be in "hide" mode with gray menus. You can adjust the individual tracks from here using those pull-down options.

5. Use your Back button to return to the Super-track Settings configuration page. Scroll to see the lower portion of the page. The "Description" section of the page provides information about the super-track. This explains the theme of this group of tracks, and provides some details about the display features and sometimes biochemical or bioinformatic methods used to obtain or process the data. You will also find credits and data policy details here for the tracks as a group.

6. From this super-track details page, you can access the individual component tracks. We will examine one that is currently visible, a histone modification mark track. Click on the "Layered H3K27Ac" hyperlink. Notice the many options at the top for setting the data display (Fig. 19.9.7B). The display mode menu now offers the full set of visibility choices (not just show/hide). There are additional settings that can be used to set the type of display and certain thresholds. The cell line display color codes become apparent here, and you can choose to display the specific cell lines of interest.

The description section of the page may provide further guidance about the color codes or other display options, as well as more guidance about the techniques and technologies used to obtain the data. Some may offer additional cell lines or treatments to select for viewing. You may also want to explore publications from the labs that provided the data to learn more about how the data were produced, and more about the settings to consider when you wish to adjust the view to locate relevant features.

7. To examine changes to the settings here, select a single cell line to be shown, and unselect all the others. Choose the yellow H1-hESC cell line with the checkbox (Fig. 19.9.7B). Uncheck all the others. Click the "submit" button at the top to return to the browser and view the track again. Locate the Layered H3K27Ac track in the graphical browser. Note the pattern of the histogram display that shows only yellow now, where before we had multiple overlapping colors (Fig. 19.9.7C). There are similar peaks, but also some differences. Try other cell lines or combinations of cell lines to view. A quick way to return to the specific controls for this track is to use the gray "mini-button" on the left side of the graphic frame. Choose distinct color pairs to help grasp the visualization. Finally, return to check all the cell lines again, click submit, and return to the viewer page.

   Here you can explore the histone mark pattern found among the cell lines, which may correspond to active regulatory elements. Learn more about this histone modification by reading the track details pages and the associated ENCODE publications from the providers of these data. Adjusting the visual field to focus on the items of interest, combined with filtering and zooming, provides a wealth of detail that can assist with planning further benchwork studies including regions to explore in more detail and cell lines to select for study.

8. Now we will adjust another sub-track of the ENCODE Integrated Regulation super-track to examine possible transcription factor binding in that region. Return to the super-track settings page. Note that the "Txn Factor ChIP" track is currently visible on the browser in "dense" mode. Set that menu to "pack" and click the "submit" button to explore these data on the browser.

   You can also do this with the right mouse button, by clicking anywhere in the track.

9. Locate the new expanded data set in the Transcription Factor ChIP-seq from ENCODE area of the browser (Fig. 19.9.7D). Some of the potential transcription machinery protein names may be familiar, such as Pol2 or NFKB. Others may not be so obvious. Also note that there are different shades of the boxes associated with different items. Further, observe the lists of numbers and letters on the right side of each item (red box). To learn more about these features, click the gray mini–button at the left side of this track to go directly to that track configuration page.

   An extensive collection of transcription factor binding signals have been identified, examined in various cell lines, and are apparent in this area. Keep in mind, though, that some of the signal found in this region that could correspond to regulation of the TP53 gene, or to the neighboring WRAP53 gene. Further exploration at higher resolution would be needed to determine that.

10. On the Tnx Factor ChIP Track Settings page, the upper section of the page offers a key to the cell types. The lower area explains more about the work that was performed, and notes that the shading of the display items corresponds to peaks of signal levels associated with the transcription factor occupancy data from the ChIP-Seq experiments processed as described. When you have explored the data enough, set the menu to "hide" and click "submit" to proceed to the next section.

    This quantitative shading shown with the "pack" view can be downloaded as data values if you later use the Table Browser to query for the individual cell lines and signal values. The visual guidance here can help you to assess the potential for binding in a region

by candidate transcription factors. As noted in the track description section, you can also drill down by exploring the ENCODE TF Binding super-track, which offers more subsets of data and additional filters and settings.

SUPPORT PROTOCOL 2

VISUALIZING SNPs IN A GENE OF INTEREST

This protocol can be used to consider whether a gene or regulatory region could be affected by SNPs (Simple Nucleotide Polymorphisms) by visually examining the data. The TP53 gene's reference sequence may be sufficient for a researcher's studies. However, it might also be worthwhile to know if variants in the sequence or regulatory elements have been identified. These could vary by individual, they could be pathogenic or protective medically, or might affect response to treatments, among other things. This protocol will allow a closer look at the gene region and possible polymorphisms. We will focus on the SNP data that are provided by the NCBI's dbSNP short genetic variations collection (National Center for Biotechnology Information Database of Single Nucleotide Polymorphisms, *http://www.ncbi.nlm.nih.gov/SNP*), which can be found in the "Variation" track group area. The dbSNP data consist of single or short nucleotide polymorphisms (and some flanking DNA details), as well as small insertions and deletions (indels), and are mapped to the reference genome at UCSC. In the past, all of the variations were treated as one track of data. As the volume of data has grown, this has been re-organized for some genome assemblies. At this time you will find that the human SNP data are available as 4 tracks, to enable a variety of possible uses. The "All SNPs" track (with a version number, currently snp138) contains the full collection of variations from a release of dbSNP. However, you may instead choose to view the "Common SNPs," "Flagged SNPs," or "Mult. SNPs" subset tracks. Examine the track description page from their hyperlinks to see if you want just common SNPs ($\geq$1% minor allele frequency), those flagged with possible clinical associations, or those that map to multiple locations, respectively. For our purposes, we will examine "All SNPs." Note that older assemblies will have only one SNP track option.

1. Focusing on the TP53 gene region, look for possible variations from the reference genome sequence in the span of the gene from 5′ UTR through 3′ UTR sections. First, note that the SNPs in the default view are the "Common SNPs," as you can see by the left slide label and the descriptive label on the viewer. Back on the main viewer (see Basic Protocol 1), scroll down to the bottom of the page to find the Variations and Repeats group. In that group, find the track for "All SNPs (138)" (or the most recent build number). Click the hyperlink title for "All SNPs (138)" to access the configuration page and the SNP filters.
2. Many options within the SNP data are available to filter and display selected subsets of the data. You can open the menus for the various options by clicking the "+" boxes near the top of the description page. Colors can be set relative to different reference gene sets (such as UCSC Genes or GENCODE or RefSeq, for example). Filters can be engaged based on many types of attributes that come with the SNP data. Coloring options can be chosen based on various features. For now, let us merely change the display mode to "dense" for this track to compare it to the default track of SNPs in the viewer. Set the top menu to "dense" and click "submit" to return to the browser viewer page.
3. Scroll back to the bottom of the viewer to examine the SNP tracks. The "All SNPs" track will display more features than the "Common SNPs" track (Fig. 19.9.8A). The default colors may be more apparent now as well. We will return to the "All SNPs" page to set the colors to help understand the potential roles of these SNPs.

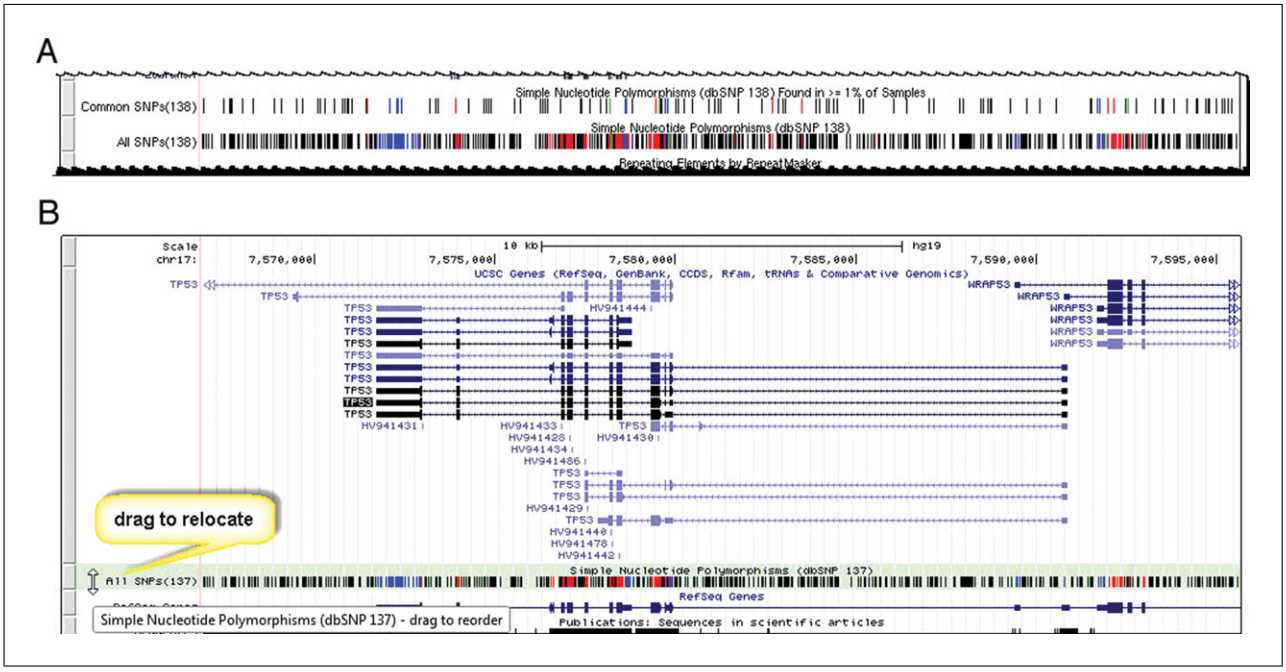


Figure 19.9.8 (**A**) Common SNPs or All SNPs can be shown. The variations from dbSNP can be viewed on the browser using different tracks. A "Common SNPs" track may be the right choice for some investigations. Other times, the "All SNPs" track is suitable. Each of these can be further filtered or set for other visualization needs using the corresponding settings pages. (**B**) Examination of the region of interest and corresponding SNP information may provide useful additional information about variations that affect function. SNPs in the coding regions and in the regulatory regions may offer leads to pursue with further benchwork. Tracks can be dragged to new locations for easier viewing. Here, the SNP track has been dragged to be placed beneath the genes track.

4. Click the gray mini-button at the left side of the "All SNPs" viewer region, or click the "All SNPs" hyperlink to return to that track's configuration page. We will choose to set "Coloring Options," to adjust the views and types of SNPs shown. In this case, open the "Coloring Options" area to access the settings choices. When you arrive, the color settings indicated in the menus that you see are the default settings, the most notable of which is the use of red for SNPs that change an amino acid sequence of a protein. Set the "SNP Feature for Color Specification" to the "Function" drop-down menu. The menus below will determine how individual SNPs will be color coded. Make these choices:

   Unknown: black
   Untranslated: blue
   Locus: black
   Intron: black
   Coding-Synonymous: black
   Splice Site: black
   Splice site: black
   Coding-Non-Synonymous: red

   This will indicate for us visually if an SNP resides in untranslated regions (UTR, blue) or if it would result in a substitution of an amino acid in the protein (non-synonymous, red). Later, you could return to try any color settings you wish, but for this example a simple illustration of differences by UTR and coding functions is all we want to emphasize. Then click "Submit" near the top to return to the viewer with the new color scheme.

It will depend on the zoom level and chromosomal position of the previous view, but the track of SNPs should be present with the specified colors near the bottom of the viewer. Drag this "All SNPs" track closer to the UCSC Genes track by hovering the mouse near the "All SNPs" label area at the left (the track will then have a light green background) and holding your mouse button (Fig. 19.9.8B). Move the track up near the gene area by dragging upwards with your mouse. The relationship between the SNPs and the exons should become more apparent. You may also want to consider the SNPs found within the signals of the ENCODE regulation track data that we explored before. Those tracks can be moved to be adjacent to the SNPs, and zooming in to specific items may be helpful to examine the specific sequences and variations.

Examination of the identified SNPs and sites that overlap a given SNP might offer insights into variations that might be present in different individuals or cell lines, which could direct experimental investigations.

Special note: If you have made changes to the location of tracks and any filter settings, you can save those to return later to the exact same view by using the "Sessions" functions. From the "My Data" menu at the top of a browser page, a "Sessions" option will take you to a "Sessions Management" page. If you create a login for the UCSC Genome Bioinformatics system, you can save your current view, as well as share it with others using a URL that will load the exact details as you have set them for 4 months. Learn more about the Sessions features from the "Help" section.

ALTERNATE PROTOCOL 1

FIND AN EVOLUTIONARILY CONSERVED REGION BETWEEN SEVERAL FISH SPECIES AND HUMAN AND VIEW THE MULTIPLE SEQUENCE ALIGNMENT

Cross-species comparative data may provide clues about important features in genomic sequences. Conserved segments in non-coding regions between distantly related species often indicate to researchers possible important regulatory features. The UCSC Genome Browser has many annotation tracks for studying genomic conservation and continues to release multiple alignment tracks with more species. This query will demonstrate how to view an upstream region of a gene (HOXA7–a homeobox gene with developmental significance and deep vertebrate conservation) and find a conserved segment in selected species and view the alignment.

1. Access the UCSC Genome Browser at the URL *http://genome.ucsc.edu*. In the blue navigation areas, click either the top link for Genomes, or the side link for Genome Browser.

2. The Gateway interface (see Fig. 19.9.2 and Basic Protocol 1) will appear. If prior queries have already been done, fully reset the form to begin this exercise. To do this: "Click here to reset the browser user interface settings to their defaults."

3. Make these choices in the Genome Browser and then click "submit":

   a. *group:* Mammal
   b. *genome:* Human
   c. *assembly:* Feb. 2009 (GRCh37/hg19). Click "submit."

4. First scroll down to below the graphic and click "hide all" to simplify the graphic, and then in the text box at the top of the page ("Position, gene symbol or search term"), type HOXA7, and click on the HOXA7 when it appears, then click the "go" button. Turn the UCSC Genes track on to "pack" if it is not on.

5. To view the conservation track for selected species, in this case fish species, scroll down to the Comparative Genomics tracks section in the annotation tracks controls and locate the link for the Conservation track. Note that a mouse-over on this link reveals the long label: Vertebrate Multiz Alignment & Conservation (100 species

currently, but more species will be included over time). Click the link to access the configuration page to choose species (Fig. 19.9.9) for display in this track.

6. Make the following selections:

   a. *Maximum display mode:* full
   b. *Species selection:* - (minus): turns off "All species"
   c. *Vertebrate:* Select these fish: Tetraodon, fugu, stickleback, medaka, and zebrafish.
   d. *Select subtracks by clade:* uncheck "All species," check Vertebrate
   e. Leave all other parameters as default.
   f. Click Submit.

7. The 100-Way Cons track will be shown in the viewer with the individual fish species in view below the 100-species comparison track (Fig. 19.9.10A). Notice how the conservation is strong in the exons and weak in the introns. Center and zoom on the 5′ region by clicking at the top of the graphic and dragging the mouse to highlight the half-height box of the 5′ region (on the right for this gene, as it aligns to the "bottom" or reverse strand of the reference assembly—if you are more comfortable with the 5′ end on the left, use the Reverse button below the Browser graphic). This should be done for two reasons, first to focus on the object of study, the 5′ untranslated region (UTR), and, second, to view the alignment at the base level. The alignment details for this track can only be viewed when the region is zoomed to less than 30,000 base pairs. Now, nudge the image a little to the left to be sure to have the entire 5′ end. Do this by clicking anywhere in the image, and while holding the mouse button down, drag it to the left. In the display, notice how the conservation drops off as soon as you look outside the transcript. In many genes, the UTR itself will have low conservation. The location in the image can be reproduced exactly by typing in these coordinates: chr7:27,196,057-27,196,391 in the position box above the graphic (Fig. 19.9.10B).

8. Click in the Multiz alignment at the bottom of the track to see the alignment details (Fig. 19.9.10B). Strong conservation in the fish species in the 3′ part of the untranslated region, near the start codon (reverse complement of ATG here: CAT, red oval) can be seen in the alignments. (Fig. 19.9.10C). Obtain DNA sequence for any of these species in this region by clicking the "D" link before the species name in the alignment. Jump to a browser for the orthologous part of the genome assembly for any of the fish by clicking the "B" link.

USE THE UCSC GENOME BROWSER TABLE BROWSER TO QUERY THE UNDERLYING DATABASE AND DOWNLOAD A LIST OF SNPs IN A GENE

BASIC PROTOCOL 2

There may be times when the visual display of a region of interest is sufficient for the work that a researcher needs to accomplish. The visualization allows a simultaneous picture of many types of data mapped to the region. However, there will be other times, or other procedures, which require text-based lists of items that could be analyzed in spreadsheets or in other software tools. The way to access and output this type of data from the UCSC Genome Browser is to use the Table Browser.

The Table Browser relies on the same data seen in the graphical interface (Fig. 19.9.1B), and it pulls the data from the same underlying MySQL database. However, the output can be treated in many more ways, including intersecting with other datasets or filtered using specific criteria of the data. A few of the output options for Table Browser will be examined here.

In this section the human genome will be investigated in the region of a large gene of medical interest, the Duchenne Muscular Dystrophy gene, or DMD. A rapid method to

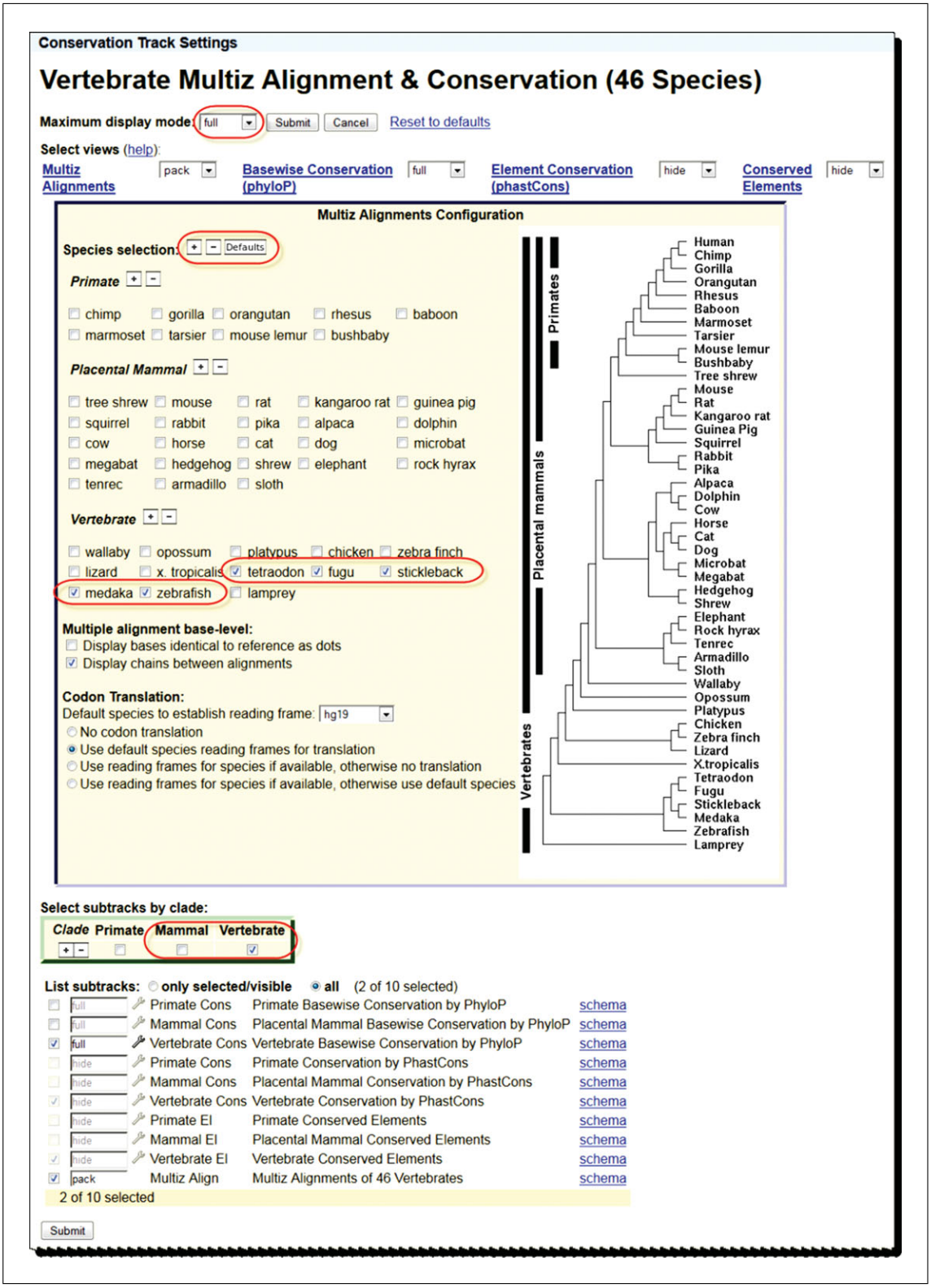


Figure 19.9.9 Track description information is available via hyperlinks on the browser main page or via gray vertical bars on the left side of the browser graphic for that track. They may contain configuration options or filters to customize the data for the display. Here, the 46-Species conservation track offers choices for which species to display. Conservation sets are available for many reference genomes, with the most species mapped to the most recent human and mouse assemblies. New collections are added as new genomes become available for the analyses.

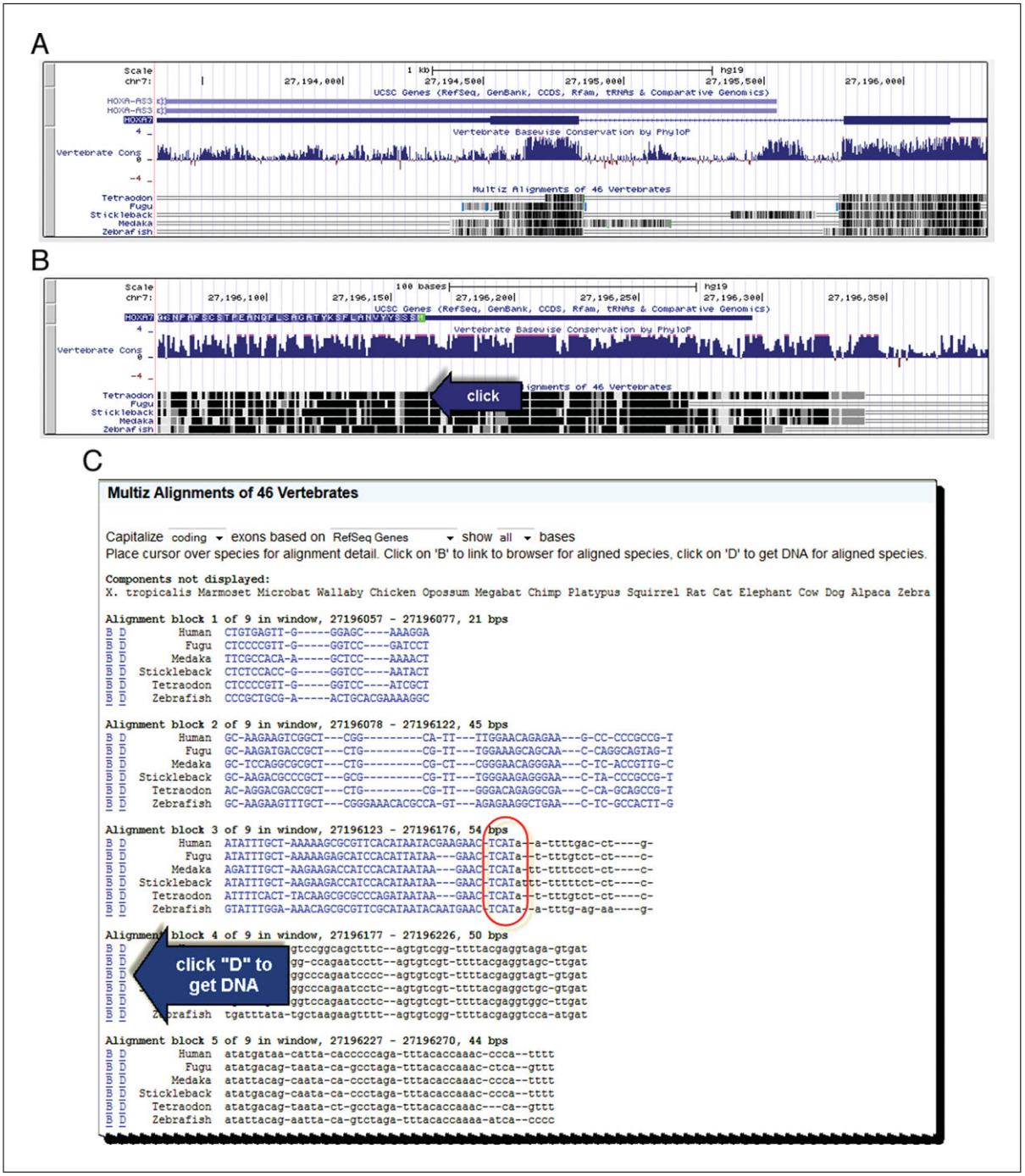


Figure 19.9.10 (**A**) First exon area of the human HOXA7 gene, with comparative data for several fish indicated. (**B**) Zoom in to a sufficient level and then click in the fish species region of the Conservation track to view the alignment. (**C**) Clicking on the conservation area in the display yields the actual alignment data for human and the species that were selected. More details on the display characteristics can be found on the lower portion of the page. Nucleotides of translated codons are in capital letters and colored blue.

extract every SNP in the genomic span of the DMD gene, in just a few steps with the simple Table Browser interface, is described below.

NOTE: At the time of this writing, dbSNP release 138 was available. Later releases may also be used.

1. Access the UCSC Genome Browser at the URL *http://genome.ucsc.edu*. Click either the "Tables" link in the top navigation bar or the "Table Browser" link on the left navigation bar.

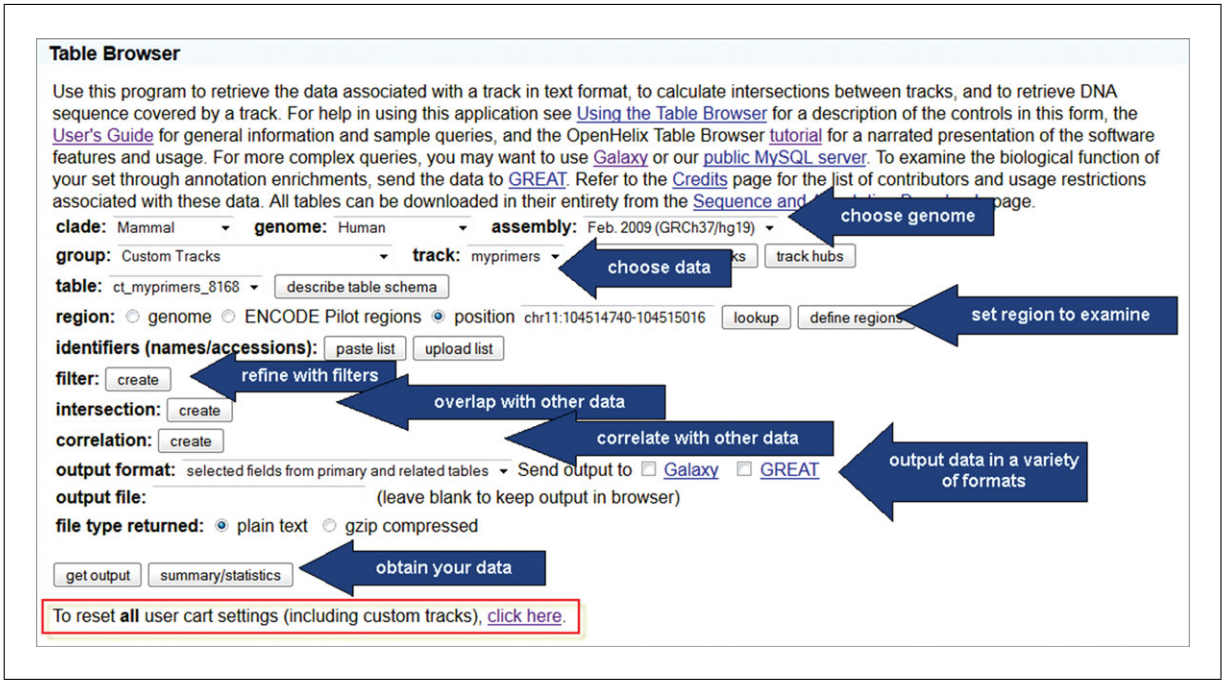


Figure 19.9.11 Overview of the Table Browser interface with steps highlighted (arrows). Many choices for data type, genomic regions, operations, and output of the data are available. To quickly reset any previous choices, filters, or other aspects, click the reset link near the bottom of the form (red box).

2. A form interface (Table Browser) will be presented (Fig. 19.9.11). The choices made on this form will create a query of the database that will retrieve the data. Although it may look daunting at first, some orientation will indicate features of the database that may be recognized from the earlier parts of this unit.

3. The first thing to do is reset the browser. If anyone has been using the software already, the form may not be in its default state. Scroll down the page and click the link that says "To reset all user cart settings (including custom tracks), click here" (Fig. 19.9.11, red boxed area at the bottom). The browser will reset with default settings that include the latest Human genome assembly.

4. The first row of choices should be familiar from Basic Protocol 1. The meaning is the same, and as before one genome and one assembly at a time will be selected to query. In this case select Human and the Feb. 2009 assembly.

5. The next section—"group" and "track"—refers to the annotation tracks seen before on the graphical interface. The pull-down menus allow access to all groups corresponding to the blue-bar groups on the main Genome Browser graphical viewer. Open the "group" menu to examine the tracks. For this example, to pull all the SNPs from a genomic region, choose Variation in the "group" section.

6. When the group choices are made, the tracks options will change to reflect the tracks that are available in that group. In the case of the Human genome a number of tracks are available here. In this case, for a list of SNPs, the top Common SNP choice will be sufficient [Common SNPs (138) at this time].

7. In the next row there are table choices. In the database, data are stored in a collection of different tables. Some tracks have multiple, related tables that contribute to the display in the Browser. Table Browser code can access and assemble data from the different tables in various ways. The code can be used to pull the data for the graphical view into the Genome Browser, or to extract it as text in the Table Browser. For the purposes of this example, SNP data will be pulled from the primary SNP

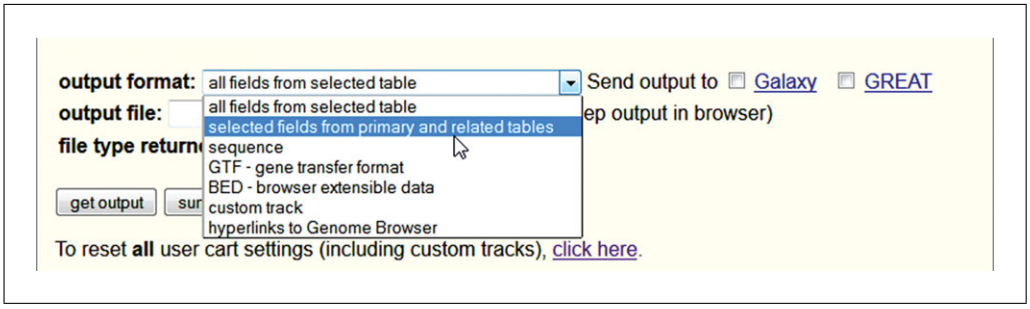


Figure 19.9.12 A variety of output choices are available to obtain results from the Table Browser. Additionally the data can be sent directly to the Galaxy or GREAT tools for further exploration.

table, snp138Common. Click the "describe table schema" button to understand the structure of the table.

8. The next item to determine is the region of the genome to examine. The "region" radio button is set for the whole genome, and could yield a huge table of every SNP in the genome if requested (more than 56 million items in snp138, 14 million in snp138Common). However, the search can be limited to a position, and that is what will be illustrated here, with focus on the DMD gene region. If the nucleotide coordinates are not known, then click the radio button for "position," enter the text `DMD` in the adjacent text box, and click the "lookup" button. To work with a list of regions, enter those using the "define regions" button, or load a list of genes with "paste list," but this will be a simple example of one gene.

9. The "lookup" button will run a query to find coordinates for the gene name(s) entered. The result will be a list of results similar to a position query. One of the DMD genes will be selected for this example. There may be a number of possible splice variants in this case. For this example, choose the one with the following description:

   DMD (uc004ddf.3) at chrX:32534713-33357726—*Homo sapiens* dystrophin (DMD), transcript variant Dp427c, mRNA.

   When that link is clicked, note that it pastes the genome coordinates into the "position" text box. At this point make sure the "position" radio button is chosen.

   Just to summarize what has been done: a species (Human) and the Feb. 2009 assembly was chosen. Data will be pulled from the snp138Common table. Only the SNPs found in the DMD genomic span indicated will be retrieved. At this time, we will skip all the other buttons that can be used to refine a query.

10. Focus on the buttons at the bottom of the form. To quickly get a sense of how many SNPs this might be, the "summary/statistics" button can be clicked to find out how many items this query would produce. At this time, over 4000 SNPs are indicated. This would be extremely difficult to understand quickly in the graphical viewer.

11. Return to the main form from the Summary by using the "Tools . . . Table Browser" menu at the top of the page. It is usually a good idea to avoid using the Back button on the browser.

12. Select an output mode for these data. There are a number of choices for ways to handle it. This can vary based on what the query items were, but for this simple query the menu choices will be similar to those illustrated in Figure 19.9.12.

    To become familiar with the options, users are encouraged to select each one, click the "get output" button, and examine the differences. Large swaths of text data can be

obtained; the actual sequences of the SNPs are available; there are options for data formatted for use in other tools (GTF and BED format); data can be viewed as a custom track; or a huge list of hyperlinks back to the browser viewer can be obtained.

A separate option is to take this whole data set and send it to the GREAT or Galaxy interfaces. GREAT is a separate tool hosted at Stanford University (http://great.stanford.edu/) that predicts functions of cis-regulatory regions. Galaxy is a separate analysis tool hosted at Penn State (http://galaxyproject.org/). Even more complex manipulations can be performed on the data with the analytical tools at Galaxy. Queries can be stored and re-run there.

13. For the purpose of this protocol, just a few items to view will be selected. From the "output format" pull-down menu, choose "selected fields from primary and related tables." This option provides access to secondary, linked tables. Click "get output" for the field choices.

14. Now get a list of the SNPs, with the location, their name or ID, the reference nucleotide and observed altered sequence data, the class of SNP, and the function of the SNP. To obtain this, click the check boxes as shown in Figure 19.9.13A: "chrom," "chromStart," "chromEnd," "name," "refNCBI," "observed," "class," and "func." Additional data from other tables which are found below this section could be accessed, but that is beyond the scope of this unit. Click the "get output" button.

15. A few selected sections of the more than 4000 SNPs in the results are shown in Figure 19.9.13B. The data can be copied and pasted into a text file, or one could go back and use the form to output directly to a file for later use.

ALTERNATE PROTOCOL 2

FOR A LIST OF SNPs, USE THE TABLE BROWSER TO FIND THE CORRESPONDING GENES

The queries so far have begun with the expectation that there is a region of interest, and that a user wants to know more about the features in there. There may be times when there are some items of interest, and the goal would be to see where they map and to obtain additional context about them. For example, there could be a list of SNPs that seem to be important from a genome-wide association study, and it would be worthwhile to see if they occur in certain genes. This example will start with a list of SNPs, and the task will be to obtain more genomic data and annotations around them. The publication illustrated in Figure 19.9.14A, concerning SNPs discovered in linkage with chronic lymphocytic leukemia, will be used as the starting point. This reference, accessible at the URL *http://www.ncbi.nlm.nih.gov/pubmed/18758461* (Di Bernardo et al., 2008).

1. Access the UCSC Genome Browser at the URL *http://genome.ucsc.edu*. Click either the "Tables" link in the top navigation bar or the "Table Browser" link on the left navigation bar.

2. A form interface will appear as described in Basic Protocol 2. If previous queries have already been done, fully reset the form. Click the link near the bottom that says: "To reset all user cart settings (including custom tracks), click here."

3. When the page reloads, start this fresh query. It begins with SNPs, so set the first fields as follows:

   clade: Mammal
   genome: Human
   assembly: Feb. 2009
   group: Variation
   track: Common SNPs (138)
   table: snp138Common

A

Select Fields from hg19.snp138Common

bin	
chrom	Reference sequence chromosome or scaffold
chromStart	Start position in chrom
chromEnd	End position in chrom
name	dbSNP Reference SNP (rs) identifier
score	Not used
strand	Which DNA strand contains the observed alleles
refNCBI	Reference genomic sequence from dbSNP
refUCSC	Reference genomic sequence from UCSC lookup of chrom,chromStart,chromEnd
observed	The sequences of the observed alleles from rs-fasta files
molType	Sample type from exemplar submitted SNPs (ss)
class	Class of variant (single, in-del, named, mixed, etc.)
valid	Validation status of the SNP
avHet	Average heterozygosity from all observations. Note: may be computed on small number of samples.
avHetSE	Standard Error for the average heterozygosity
func	Functional category of the SNP (coding-synon, coding-nonsynon, intron, etc.)
locType	Type of mapping inferred from size on reference; may not agree with class
weight	The quality of the alignment: 1 = unique mapping, 2 = non-unique, 3 = many matches
exceptions	Unusual conditions noted by UCSC that may indicate a problem with the data
submitterCount	Number of distinct submitter handles for submitted SNPs for this ref SNP
submitters	List of submitter handles
alleleFreqCount	Number of observed alleles with frequency data
alleles	Observed alleles for which frequency data are available
alleleNs	Count of chromosomes (2N) on which each allele was observed. Note: this is extrapolated by dbSNP from submitted frequencies and total sample 2N, and is not always an integer.
alleleFreqs	Allele frequencies
bitfields	SNP attributes extracted from dbSNP's SNP_bitfield table

get output | cancel | check all | clear all

B

```
chrX    32658906    32658907    rs12009363    C    A/C    single    intron
chrX    32659110    32659111    rs73621811    G    A/G    single    intron,near-gene-3
chrX    32659122    32659123    rs73621812    A    A/G    single    intron,near-gene-3
chrX    32659168    32659169    rs73467339    A    A/C    single    intron,near-gene-3
chrX    32659423    32659424    rs73467342    C    C/G    single    intron,near-gene-3
chrX    32659511    32659512    rs73621813    G    A/G    single    intron,near-gene-3
chrX    32659552    32659553    rs57815999    A    A/C    single    intron,near-gene-3
chrX    32659591    32659592    rs60180387    C    C/T    single    intron
chrX    32659770    32659771    rs59416281    C    A/C    single    intron,near-gene-5
chrX    32659915    32659916    rs59480298    T    A/T    single    intron,near-gene-5
chrX    32660136    32660137    rs59440959    A    A/C    single    intron,near-gene-5
chrX    32660143    32660143    rs201017064   -    -/T    insertion       intron,near-gene-5
chrX    32660144    32660144    rs200226546   -    -/C    insertion       intron,near-gene-5
chrX    32660273    32660274    rs58234753    C    C/G    single    intron,near-gene-5
chrX    32660304    32660305    rs59854433    C    A/C    single    intron,near-gene-5
chrX    32660482    32660483    rs56842615    A    A/G    single    intron,near-gene-5
chrX    32661757    32661758    rs143461225   T    C/T    single    intron
chrX    32661758    32661759    rs113583679   A    A/G    single    intron
chrX    32662026    32662027    rs73621814    T    C/T    single    intron
chrX    32662103    32662104    rs72470503    A    -/T    deletion        intron
chrX    32662117    32662118    rs72470504    C    C/G    single    intron
chrX    32662121    32662122    rs5928065     C    C/T    single    intron
chrX    32662193    32662194    rs6628728     T    C/T    single    intron
chrX    32662354    32662355    rs34155804    T    A/T    single    missense
chrX    32662588    32662589    rs142646607   T    C/T    single    intron
chrX    32662682    32662683    rs72470505    T    -/G    deletion        intron
chrX    32662744    32662745    rs73467358    C    C/G    single    intron
chrX    32662746    32662747    rs73621815    A    A/G    single    intron
chrX    32715656    32715657    rs6631628     C    C/T    single    intron
chrX    32715674    32715675    rs6631629     A    A/C    single    intron
chrX    32715936    32715937    rs72470512    C    -/G    deletion        intron
chrX    32716109    32716110    rs1800265     C    A/G    single    coding-synon
chrX    32716131    32716132    rs72470514    G    A/C    single    intron
chrX    32716132    32716133    rs72470515    G    C/G    single    intron
chrX    32716158    32716159    rs72470516    A    A/T    single    intron
chrX    32716168    32716169    rs41303183    T    C/T    single    intron
chrX    32716631    32716632    rs57008203    A    A/G    single    intron
chrX    32716635    32716636    rs111765391   C    C/T    single    intron
chrX    32716733    32716734    rs73621820    T    C/T    single    intron
chrX    32716755    32716756    rs12006840    G    G/T    single    intron
chrX    32716762    32716763    rs762    A    G/T    single    intron
chrX    32716854    32716855    rs73621821    C    C/T    single    intron
chrX    32717077    32717077    rs201692931   -    -/TC   insertion       intron
chrX    32717080    32717081    rs72470517    G    -/GA   in-del    intron
chrX    32717133    32717134    rs72470518    T    A/T    single    intron
chrX    32717783    32717784    rs72470519    A    G/T    single    intron
```

Figure 19.9.13 (**A**) The Table Browser interface offers the opportunity to specify data types for the output by selecting fields from the list of available items. (**B**) Output shows samples of the selected items.

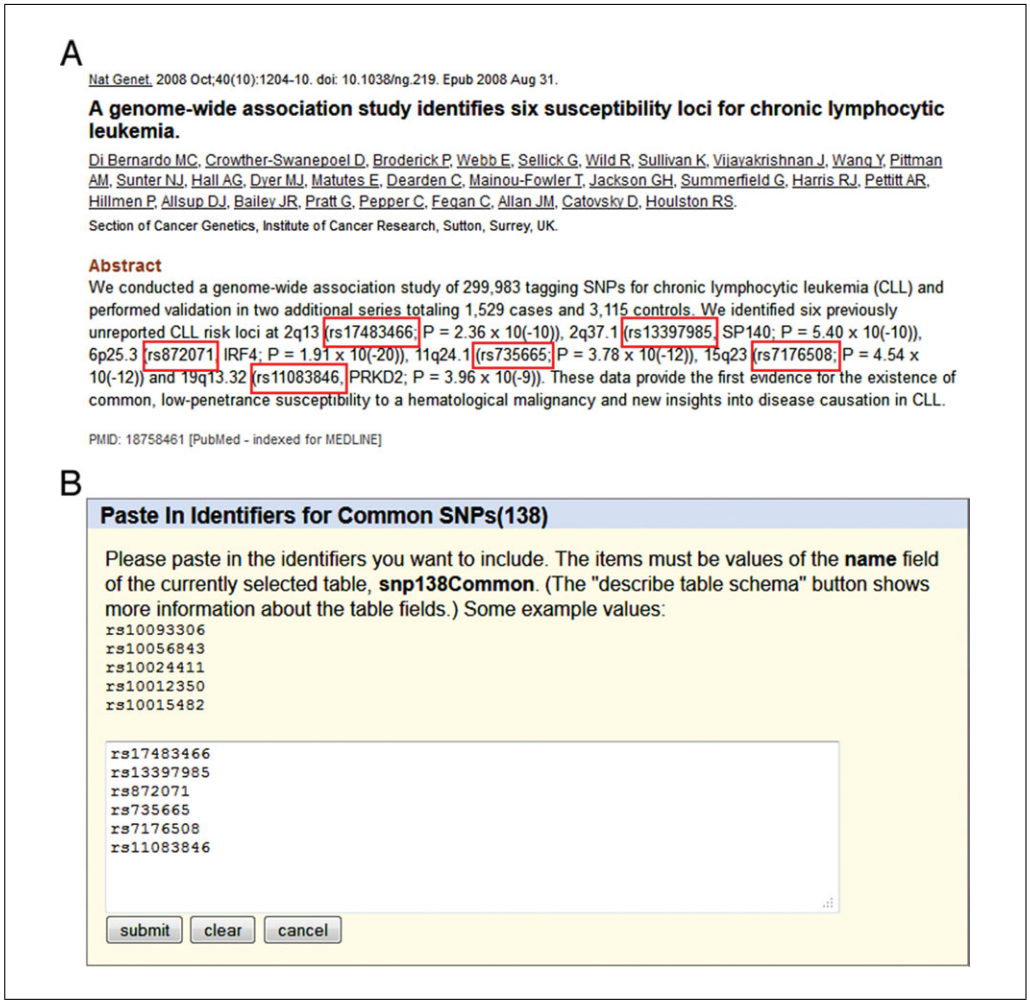


Figure 19.9.14 The Table Browser will accept lists of identifiers as appropriate for specific tables. Here, rsIDs from dbSNP as gleaned from the literature (**A**) can be pasted directly into the Table Browser to limit the search (**B**).

4. In this example, the goal is not just to look for one region, but to look for six places that correspond with the six SNPs in the abstract. This list of SNPs will form the basis of the locations to be found, so create a list of identifiers:

```
rs17483466
rs13397985
rs872071
rs735665
rs7176508
rs11083846
```

   This list could be typed in, or created as a text file on the desktop. The "paste list" option will be shown here (Fig. 19.9.14B). Click the "paste list" button, which can be found in the "identifiers (names/accessions)" row of the table browser interface, and either type or paste in these six items. Click "submit" on the "Paste In Identifiers for SNPs screen" to store them.

5. The Table Browser screen will reappear. To obtain just the genomic locations for each of these SNPs, choose the "output format" as "selected fields from primary and related tables," then click "get output."

6. When asked for fields, select “chrom,” “chromStart,” “chromEnd,” “name.” Click “get output” at the bottom of this table. The results should look like this:

```
#chrom  chromStart  chromEnd   name
chr2    231091222   231091223  rs13397985
chr2    111797457   111797458  rs17483466
chr6    411063      411064     rs872071
chr11   123361396   123361397  rs735665
chr15   70018989    70018990   rs7176508
chr19   47207653    47207654   rs11083846
```

7. With each SNP location now determined, this information can be used to ask which genes contain these SNP locations. Copy the data on the Web page. This will form the basis of the query to find the genes.

8. Return to the Table Browser interface. Reset all user cart settings to clear prior choices (see step 2). The interface should reload with a fresh Table Browser. By default, it will offer Genes and Gene Prediction Tracks, UCSC Genes, and knownGene as the data types to query. In this case, this is what is required, but the goal is to find the genes corresponding to the six SNP locations identified above.

9. In the “region” row, click the “define regions” button. Paste the results of the SNP query in the text box and click “submit.” This should bring the user back to the Table Browser interface. To ask for the output to contain several items stored in different tables from the “output format” pull-down menu, choose “selected fields from primary and related tables.” Click the “get output” button.

10. Select the check boxes for “name,” “chrom,” “txStart,” and “txEnd” from the upper table. This will give the UCSC Identifier name, the chromosome, and the transcription start and end coordinates for the genes. Specifying the transcription boundary may not capture all possible regulatory elements that may be important, but this example will just use the transcription range.

11. Data from linked tables are needed at this time as well. Linked tables are found below the original table (Fig. 19.9.15A). The gene symbol is found in a table called kgXref (known gene cross-reference), which is opened by default below the knownGene table.

12. In the “hg19.kgXref” fields box, click “geneSymbol” and “description.”

13. Click the “get output” button under the top box. The results should show the table of items as selected (Fig. 19.9.15B).

14. Four genes match the SNPs: SP140, ACOXL, IRF4, and PRKD2. Multiple entries for these genes represent different transcripts that would encompass the SNP. Appropriate SNP matches can be determined by comparing the location data. It is also clear that two of the SNPs are not found within the transcription boundaries of a known gene.

There are multiple ways to continue to examine these data and build more intricate queries, but this should offer a taste of how one can obtain data from a list of items. As an interesting exercise, try a similar query using the knownCanonical table (one of the long list of tables that is part of the UCSC Genes Track) as a starting point. The Table Browser interface may also be used from within the Galaxy framework for continued analysis of the data using numerous algorithms available there.

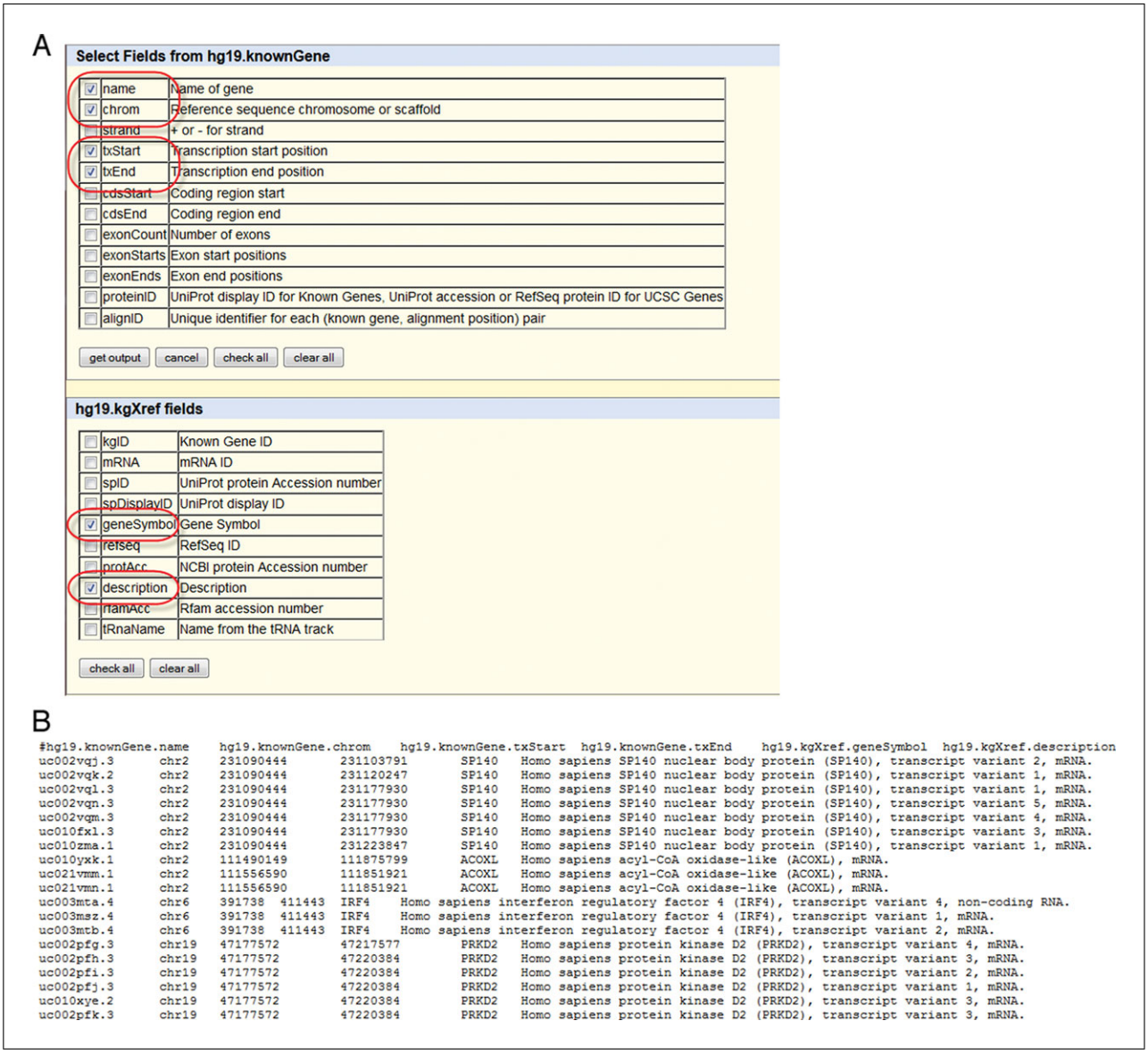


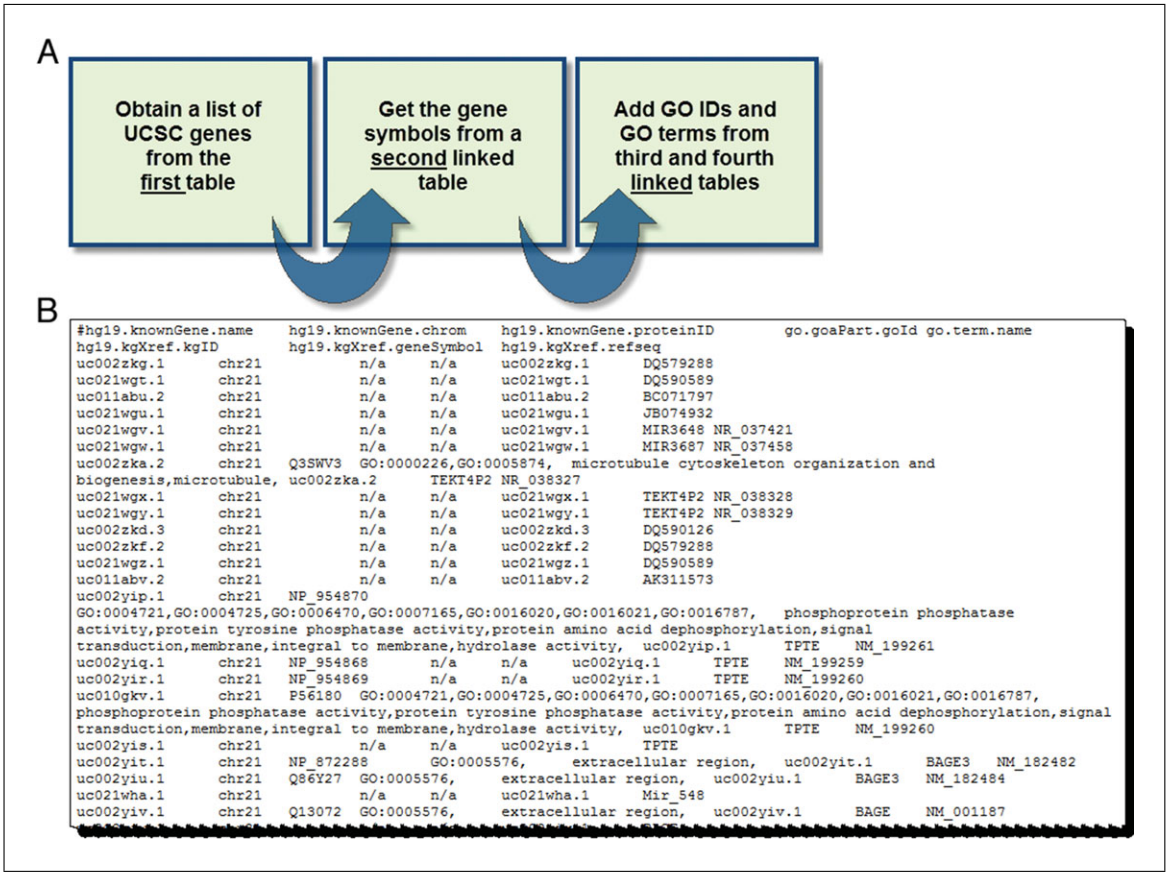
Figure 19.9.15 (**A**) Using the "selected fields from primary and related tables" option, the output of the Table Browser can be configured to join data from several tables. This will include tables linked to the first query tables, which are available to add to the query lower on the selection page. (**B**) Output from main table and linked tables.

ALTERNATE PROTOCOL 3

USE THE TABLE BROWSER TO FIND FUNCTIONAL ANNOTATIONS FOR A LIST OF GENES

There may be times when a researcher has a list of genes and wishes to examine them in more detail. High-throughput studies, database searches, or simply collections from the literature may yield leads of interest for further study. A Table Browser query starting with a gene list can be used to gather information in bulk. This query will begin by collecting all the known genes on human chromosome 21, and then adding Gene Ontology (GO) description terms to the list.

In diagrammatic form, Figure 19.9.16A indicates what will be done. A list of UCSC genes will be obtained from the first table, corresponding gene symbols will be obtained from a second table, and Gene Ontology identifiers and terms will be obtained from additional linked tables.




```
#hg19.knownGene.name    hg19.knownGene.chrom    hg19.knownGene.proteinID        go.goaPart.goId go.term.name
hg19.kgXref.kgID        hg19.kgXref.geneSymbol  hg19.kgXref.refseq
uc002zkg.1      chr21           n/a     n/a     uc002zkg.1      DQ579288
uc021wgt.1      chr21           n/a     n/a     uc021wgt.1      DQ590589
uc011abu.2      chr21           n/a     n/a     uc011abu.2      BC071797
uc021wgu.1      chr21           n/a     n/a     uc021wgu.1      JB074932
uc021wgv.1      chr21           n/a     n/a     uc021wgv.1      MIR3648 NR_037421
uc021wgw.1      chr21           n/a     n/a     uc021wgw.1      MIR3687 NR_037458
uc002zka.2      chr21   Q3SWV3  GO:0000226,GO:0005874,  microtubule cytoskeleton organization and
biogenesis,microtubule, uc002zka.2      TEKT4P2 NR_038327
uc021wgx.1      chr21           n/a     n/a     uc021wgx.1      TEKT4P2 NR_038328
uc021wgy.1      chr21           n/a     n/a     uc021wgy.1      TEKT4P2 NR_038329
uc002zkd.3      chr21           n/a     n/a     uc002zkd.3      DQ590126
uc002zkf.2      chr21           n/a     n/a     uc002zkf.2      DQ579288
uc021wgz.1      chr21           n/a     n/a     uc021wgz.1      DQ590589
uc011abv.2      chr21           n/a     n/a     uc011abv.2      AK311573
uc002yip.1      chr21   NP_954870
GO:0004721,GO:0004725,GO:0006470,GO:0007165,GO:0016020,GO:0016021,GO:0016787,   phosphoprotein phosphatase
activity,protein tyrosine phosphatase activity,protein amino acid dephosphorylation,signal
transduction,membrane,integral to membrane,hydrolase activity,  uc002yip.1      TPTE    NM_199261
uc002yiq.1      chr21   NP_954868       n/a     n/a     uc002yiq.1      TPTE    NM_199259
uc002yir.1      chr21   NP_954869       n/a     n/a     uc002yir.1      TPTE    NM_199260
uc010gkv.1      chr21   P56180  GO:0004721,GO:0004725,GO:0006470,GO:0007165,GO:0016020,GO:0016021,GO:0016787,
phosphoprotein phosphatase activity,protein tyrosine phosphatase activity,protein amino acid dephosphorylation,signal
transduction,membrane,integral to membrane,hydrolase activity,  uc010gkv.1      TPTE    NM_199260
uc002yis.1      chr21           n/a     n/a     uc002yis.1      TPTE
uc002yit.1      chr21   NP_872288       GO:0005576,     extracellular region,   uc002yit.1      BAGE3   NM_182482
uc002yiu.1      chr21   Q86Y27  GO:0005576,     extracellular region,   uc002yiu.1      BAGE3   NM_182484
uc021wha.1      chr21           n/a     n/a     uc021wha.1      Mir_548
uc002yiv.1      chr21   Q13072  GO:0005576,     extracellular region,   uc002yiv.1      BAGE    NM_001187
```

Figure 19.9.16 For certain tables (notably those associated with the UCSC Genes track and its main table, knownGene), several layers of linked tables (presented diagrammatically in panel **A**) may contain information of interest. They can be joined together with successive steps to provide a table of data from the linked tables (**B**). Here, the Gene Ontology (GO) data are pulled from a different table, "go," that is available for linkage from several different assembly-specific databases.

1. Access the UCSC Genome Browser at the URL *http://genome.ucsc.edu*. Click either the "Tables" link in the top navigation bar or the "Table Browser" link on the left navigation bar.
2. The form interface (Table Browser) will appear. If prior queries have already been done, fully reset the form. Click the link near the bottom: "To reset all user cart settings (including custom tracks), click here."
3. Make these choices on the Table Browser:

   clade: Mammal
   genome: Human
   assembly: Feb. 2009
   group: Genes and Gene Prediction Tracks
   track: UCSC Genes
   table: knownGene
4. Choose the "position" radio button and type chr21 as the location. Click "lookup" to quickly add the whole nucleotide range.
5. Leave all other choices as default and choose "selected fields from primary and related tables" in the "output format" menu. Click "get output."
6. On the next page, choose the items for the output. At the top is the primary table selection area. Select "name," "chrom," and "proteinID" for the purposes of this query.

7. Now, add data from linked tables. Using "kgXref" below the knownGene table (Fig. 19.9.15A), which is a cross-reference table for various identifiers, choose the "kgID," "geneSymbol," and "refseq" fields by clicking the check boxes.

8. One of the linked tables is the "go" or Gene Ontology annotation data set. Click the check box next to "go" for the "goaPart" table. Click "Allow Selection from Checked Tables" to view the choices for that table.

9. In the new table choices for "go.goaPart fields," select the field "goId" for GO numerical identifiers.

10. Now add GO terms. In the Linked Tables area, a new "go" database table has appeared because it is linked via go.goaPart. Click the check box next to the "go" table, "term." Click "Allow Selection from Checked Tables" to view the choices for that table.

11. In the "go.term fields" table, check the box for "name."

12. Return to the uppermost box and click "get output" to pull all of the data from this complex query. Data are coming from four tables. When the result is returned, the outcome should be a table of data with columns from "hg19.knownGene," "hg19.hgXref," "go.goaPart," and "go.term" tables (Fig. 19.9.16B).

    It may take a significant amount of time to run this query. Asking for information on all the genes on a whole chromosome (even a small one) can be time-intensive, and joining multiple tables makes this database query very complex.

    Not all genes will have descriptive annotations. Some will have multiple GO annotations. Researchers might choose to focus on transmembrane receptors, or kinases, or extracellular region proteins, depending on the goals of the research. One might also choose to start the query from functional annotations and work the other way to get to genomic regions. The main point is that a list of items can be the starting point and value can be added to the list by extracting additional information from the underlying database.

BASIC PROTOCOL 3

CREATE A SIMPLE CUSTOM TRACK IN THE UCSC GENOME BROWSER TO DISPLAY DATA

Clones, primers, SNPs, or anything else for which genomic location information is available can be overlaid on the UCSC Genome Browser. In this example, imagine that there are three primers for a human gene, Iceberg, that you wish to mark on the Browser graphic for future reference. This protocol walks through the basic steps to generate a track showing those data or any other data that can be mapped via coordinates on the reference assembly. There are essentially four steps: (1) describe how the browser should look when a track is opened; (2) define the track features like name, color, and visibility; format the data in the appropriate columns with position, identifiers, or names, shading level of individual data items, and direction; (3) upload the track; and (4) view it. This does not require any programming skills. It merely requires that the text be put in the correct format.

Full details and additional complexity regarding this topic (including other more complicated styles of track display such as histograms in the "wiggle" format) can be found on the UCSC Genome Browser help documentation at *http://genome.ucsc.edu/goldenPath/help/customTrack.html*.

1. This protocol starts with a text editor or any sort of spreadsheet program. The format used in this example is called BED, for Browser Extensible Data, which is the primary format used by the UCSC Genome Browser team, but over a dozen formats are supported.

2. First, tell the **browser** how to look when it opens a custom track. For this example, the display should focus on the coding area of the Iceberg gene, with the default number of pixels for the window, hide everything except the UCSC gene track, and show the restriction sites because those may be used to check the PCR products from amplifications. In this case, each of the browser characteristics is defined on a separate line:

```
browser position chr11:104,514,740-104,515,016
browser pix 1000
browser hide all
browser pack knownGene
browser pack cutters
```

3. Next establish some things desired in the **track**. Name the track `myprimers`, which is a set of "primers for the Iceberg gene," which should be shown in "pack" visibility display and be blue in color. Here is text that says that in the proper format:

```
track name=myprimers description=''primers for the
  Iceberg gene'' visibility=pack color=0,0,255
  useScore=0
```

This would be on one single line in the text document without a hard return, but it may not appear that way in this unit. Colors are in RGB notation with 255 levels of each color available; 0,0,255 is blue.

4. Now add the **position and specific features of the data items**, which in this case are primers. Draw them as thick boxes in a nucleotide range, give them appropriate names, shade them (a range of 1 to 1000 indicates the depth of shading; in this case, the items will be the darkest possible), and indicate the 5′ to 3′ direction with arrowheads using the + or – to indicate the strand direction of forward or reverse.

```
chr11 104514750 104514771 endPrimer 1000 +
chr11 104514991 104515016 begRevPrimer 1000 -
chr11 104514910 104514930 midRevPrimer 1000 -
```

5. The fields in use here are a BED 6 file format. Additional types of features can be indicated with additional columns, but that is beyond the scope of this example. Documentation on the custom tracks options provides more details.

Now assemble the entire text in one piece:

```
browser position chr11:105009000-105010000
browser pix 1000
browser hide all
browser pack knownGene
browser pack cutters
track name=myprimers description=''primers for the
  Iceberg gene'' visibility=3 color=0,0,255 useScore=0
chr11 105009540 105009561 endprimer 1000 +
chr11 105009781 105009806 beginprimer 1000 -
chr11 105009700 105009720 middleprimer 1000 -
```

This is the text that can be taken to the UCSC Genome Browser and uploaded as a custom track. A text file can simply be uploaded, or it can be copied and pasted into the interface. There are numerous buttons that can be used to upload the custom track data, but use the first one found on the Gateway page for this example (Fig. 19.9.17A).

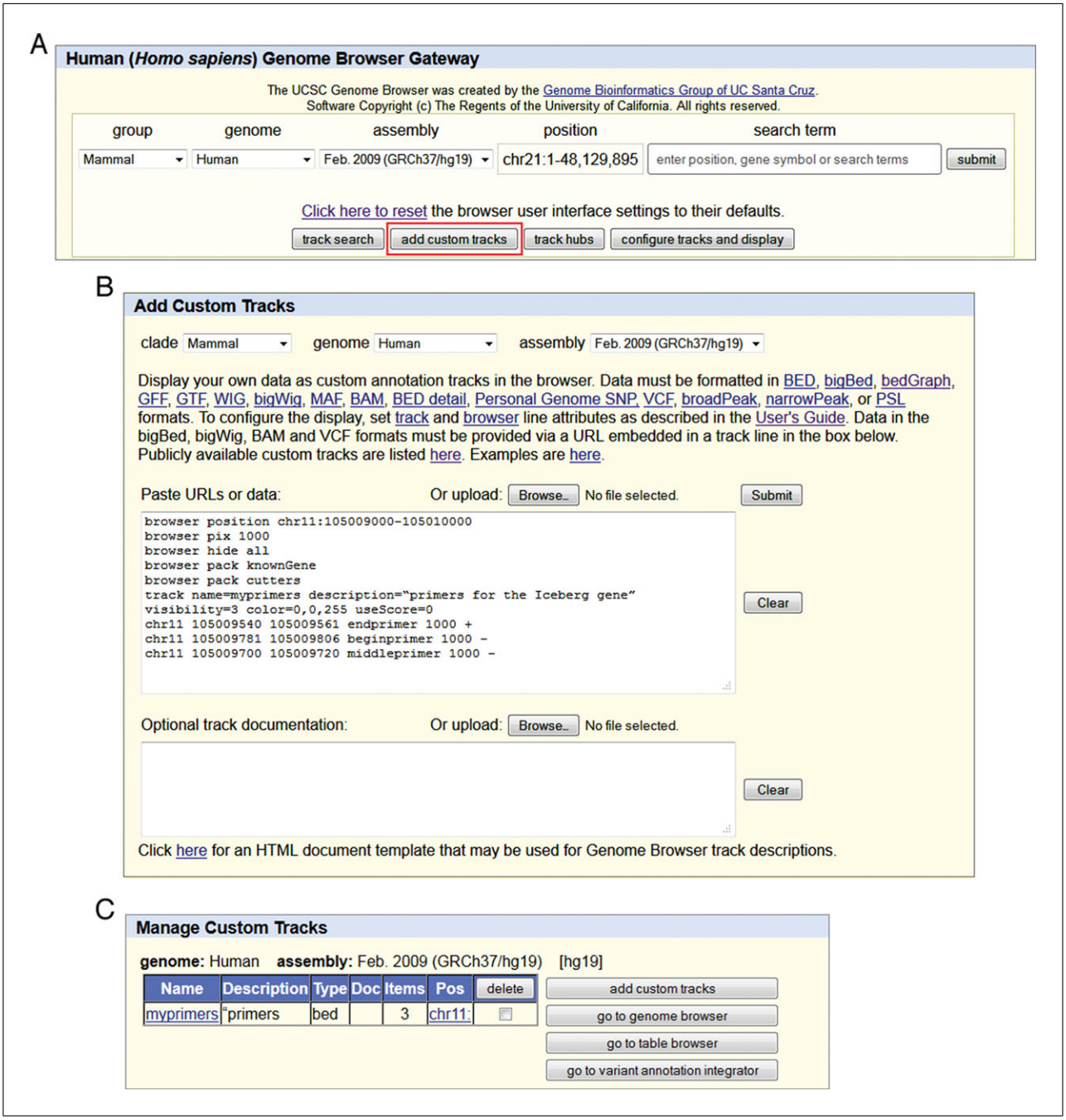


Figure 19.9.17 Users may add data of their own to the Genome Browser via the Custom Tracks feature. (**A**) Access the Custom Tracks feature by selecting the button on the Gateway page. This button is also available on the main browser graphic display page and on the Table Browser interface. (**B**) Data may be uploaded by a variety of methods, including finding a file on a local computer (Browse . . .), pasting in a URL for a Web-accessible file, or pasting in the data directly (i.e., paste URLs or data). Several different file formats are accepted (top). (**C**) All custom tracks that have been uploaded are available under the "manage custom tracks" button on the main browser page (replaces "add custom tracks" when custom tracks have been added).

6. Clicking an "add custom tracks" button provides a new interface for entering the tracks (Fig. 19.9.17B). On the custom track interface, indicate the species and assembly to which the track should apply, though the browser remembers your choices if one arrived here from a browser view of an assembly (hg19 here). Either upload the file or paste the text in the upper text box. Additional information can be added that describes the data in the "Optional track documentation" box. These are the data that would become available to users who click on the items. In this case, that information will not

be added, but it may be something to consider when sharing tracks with colleagues. Paste the text from step 5 above, then click "Submit." A "Manage Custom Tracks" page will provide several ways to use this track now (Fig. 19.9.17C).

7. The track becomes a track in a new blue-bar group, Custom Tracks, in the Genome Browser. (Fig. 19.9.18A, red box). Other tracks can be added to the group, as long as they each have a unique name, but for this example simply continue to explore the options on this one track.

8. At this point, one could examine the data in this track in the genome browser viewer or begin to use it as a query option in the table browser. Start with a look at how it appears in the browser by clicking the button "go to genome browser." The display will be shown in the Genome viewer interface (Fig. 19.9.18A).

   Once a custom track is in place, query and analysis of the data in the context of the other genomic data is possible, as with any other track in the viewer or in the Table Browser. The track gets its own blue-bar group below the Browser graphic along with the usual track controls as seen for any resident track. The track can be shared with others, and everyone can know what primers are available. It is also possible to keep a clone collection for the lab or to show SNPs that a group has discovered. The possibilities are nearly endless.

9. In addition to the graphical view, a custom track is made available in the table browser by default. Use the "Tools" pull-down menu in the upper navigation area to "go to table browser" for those complex Table Browser queries just like any other track (Fig. 19.9.18B).

   The BED format is the most basic and simple format for small data sets. Be aware that there are a couple of data formats that allow more efficient viewing of large datasets. The bigBed and bigWig types are indexed binary formats created from BED and WIG files, respectively. The resulting files enable the Browser to upload only the portions of the files that are needed to display the region in view, rather than the entire file. For large datasets, bigWig and bigBed are considerably faster. For tracks larger than about 100 megabytes, they are required. You can learn more about file formats and converting BED and WIG files to bigBed and bigWig in the user guide (http://genome.ucsc.edu/FAQ/FAQformat.html).

   To view a large number of genome-wide datasets, Track Hubs is another option. Though limited to a specific set of compressed binary indexed formats (bigBed, bigWig, BAM, and VCF), the track hub utility allows data persistence, as well as track configurability and user control of the data that can be quite useful for a large collection of genome-wide datasets. See the user guide for more information (http://genome.ucsc.edu/goldenPath/help/hgTrackHubHelp.html) and the next section for an introduction to Hub functionality.

DISPLAY A TRACK HUB TO VIEW DATA HOSTED BY OUTSIDE PARTIES

BASIC PROTOCOL 4

The Custom Track mechanism is excellent for viewing private data in the Browser and for sharing them with selected colleagues. When data sets get too large, however, they run afoul of the need for the Custom Track to be uploaded in its entirety to UCSC. The track hub mechanism provides the user with a method for storing large files locally in an indexed format from which data are uploaded only as needed for display in limited regions (Raney et al., 2013). Track Hubs also allows for making the data publicly available for all users of the Genome Browser, rather than to only the people who know how to view it. Data are placed on the user's own servers, accessible to the Internet, along with documentation, where the user may update them as needed. An entry in a table at UCSC makes the Hub visible to everyone.

This protocol shows how to view miRNA data from the miRcode Track Hub and explores the mechanism for creating such a hub (Jeggari et al., 2012).

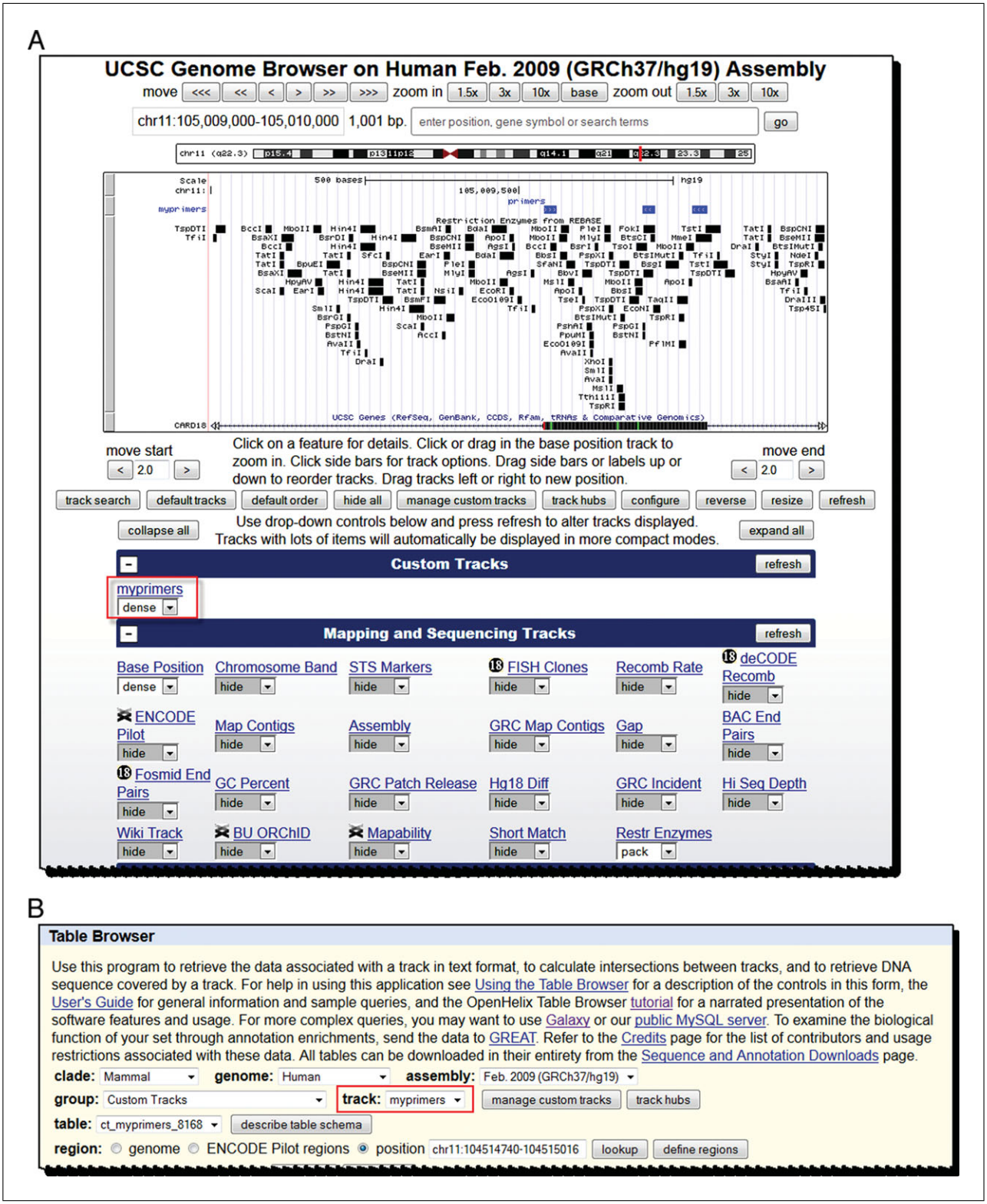


Figure 19.9.18 (**A**) Custom tracks are available for display in the main browser graphic in the same way as resident tracks and may be controlled via pull-down menus in the same way (red box). (**B**) Custom Tracks are also available in the Table Browser (red box) and can be queried in the same fashion as resident tracks. For the color version of this figure, go to *http://www.currentprotocols.com/protocol/mb1909*.

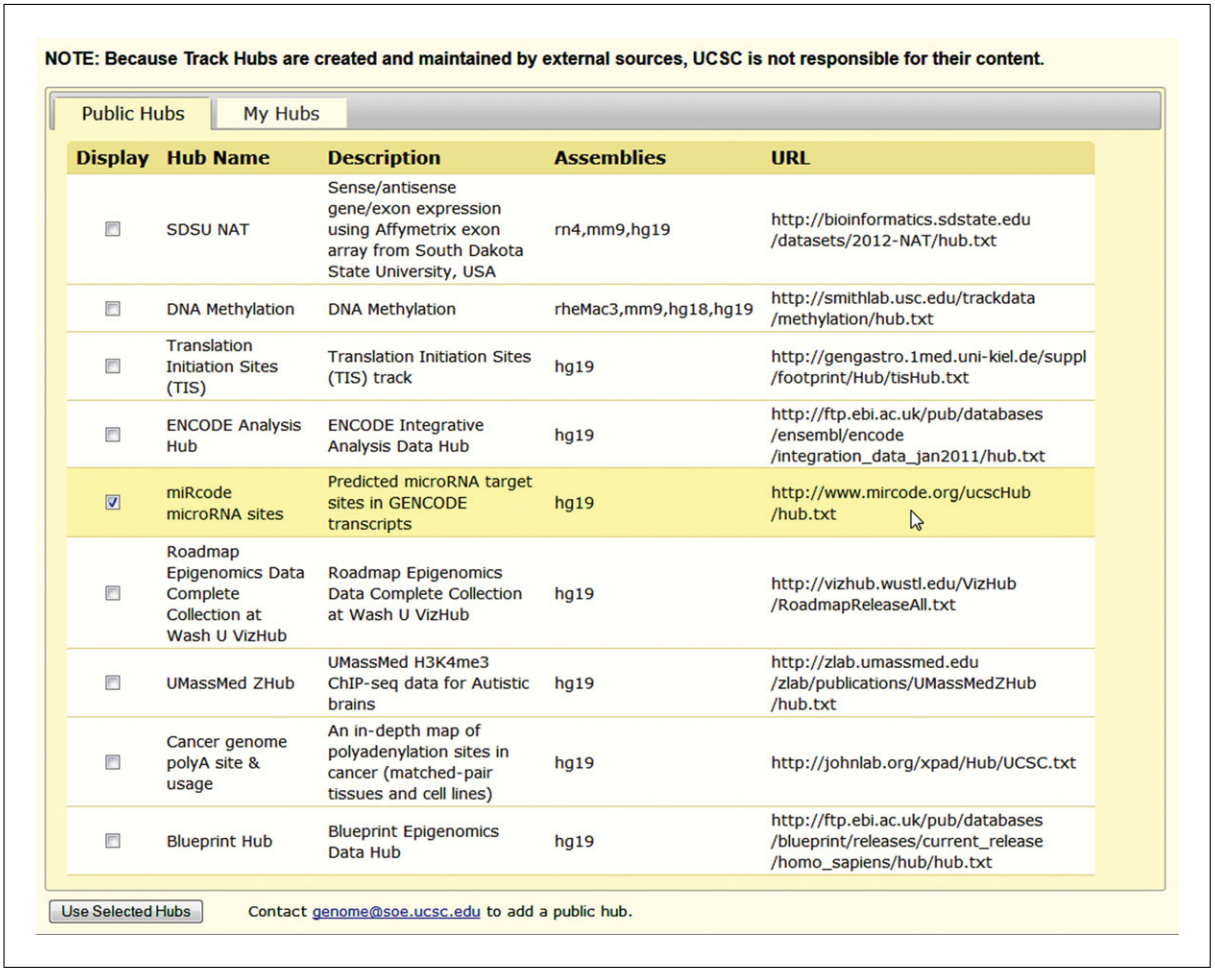


Figure 19.9.19 Track Hub selection page. The cursor arrow indicates the location of the file on the data provider's server that contains instructions for the hub.

1. Navigate to the Track Hubs page:

   a. Starting at *http://genome.ucsc.edu*, click on "Genomes" on the left side of the upper blue bar.
   b. Return to the default settings with the "Click here to reset" link.
   c. Hit "submit" to go to the default position (SOD1 gene).
   d. Click the "hide all" button below the viewer.
   e. Click the "track hubs" button under the viewer.

2. On the Track Hubs page, turn on the miRcode microRNA tracks using the checkbox and click the "use selected hubs" button below (Fig 19.9.19). This Hub is chosen because it is relatively simple and can be used to illustrate the steps for creating your own hub. The other Hubs on the page have varying degrees of complexity, some utilizing composite tracks, where multiple tracks are configured together as a set.

3. Note the new blue-bar track group below the Browser graphic, containing two new tracks from the Hub (Fig. 19.9.20).

   a. Turn on the UCSC Genes track to pack and collapse the "miR Sites High" track to dense, using the pull-downs below the Browser graphic, then "refresh." Note the varying color intensity of the predicted microRNA binding sites on the SOD1 gene.
   b. Navigate to the FGFR2 gene by typing this into the text box above the Browser graphic, then select it from the list when it is displayed. Click "Go." Fig 19.9.20

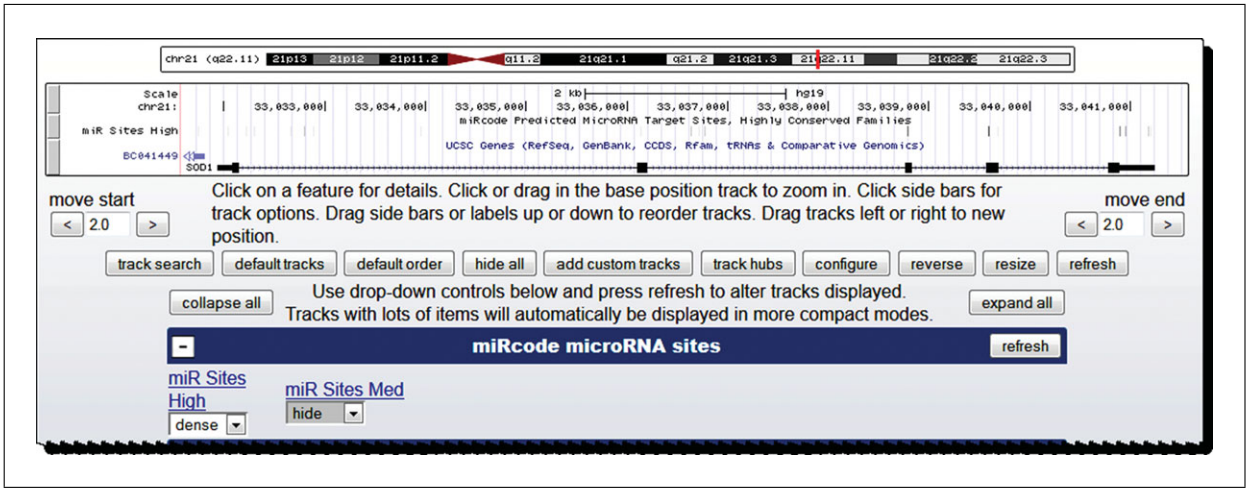


Figure 19.9.20 Track Hub miRcode displayed on hg19 with the UCSC Genes track. Note the new blue-bar track group below the Browser graphic. One of the two miRcode tracks is displayed at "dense" visibility.

shows the resulting view, with the microRNA binding sites displayed. Not the correspondence to the exons in the genes.

4. The authors have provided a description similar to UCSC's resident tracks.

   a. Click on the gray bar to the left of the "miR Sites High" graphic to read their configuration options and explanation of color intensities.
   b. Study the miRcode Track Hub

5. To see how these tracks are configured by the group hosting the data:

   a. Return to the Browser graphic (Use the "Genome Browser" link in the top blue bar on the configuration page and click on "track hubs" below the graphic, or use "MyData . . . Track Hubs" in the pull-down menu).
   b. Copy the URL in the miRcode row, *http://www.mircode.org/ucscHub/hub.txt* (Fig. 19.9.19, cursor arrow) to a new Web browser window. The resulting page is a text file on the data owner's site (Fig. 19.9.21A) describing the new blue bar group: `shortLabel miRcode microRNA sites`; the location of a file defining the genomes on which to display the hub data: `genomeFile genome.txt` and contact information for the Hub's owner.
   c. Copy the text `genomes.txt` and paste it over `hub.txt` in the URL at the top of the browser. Hit the Enter key on your keyboard. The new page (Fig. 19.9.21B) defines which genomes will display the data and the location of the file containing the configuration.
   d. Copy the text `hg19/trackDb.txt` from this page to replace the file-name in the URL as before. The resulting URL, *http://www.mircode.org/ucscHub/hg19/trackDb.txt*, shows two blocks of text defining the two tracks. The key-value pairs define configuration options, e.g., bigDataUrl is the location of the bigBed file with the data, indexed for viewing in the Browser. One track is set to "hide" by default and the other to "pack," which matches the view we obtained when we first turned on the hub (Fig. 19.9.20)
   e. Copy the second item on the first line of text, `mir_sites_highcons` as before, replacing the filename in the URL bar of your Web browser.
   f. Type `.html` after the track name and hit "enter" on your keyboard. The resulting URL should be *http://www.mircode.org/ucscHub/hg19/mir_sites_highcons.html*. Here you will see the page that serves the descriptive text we looked at in step 4a, above.

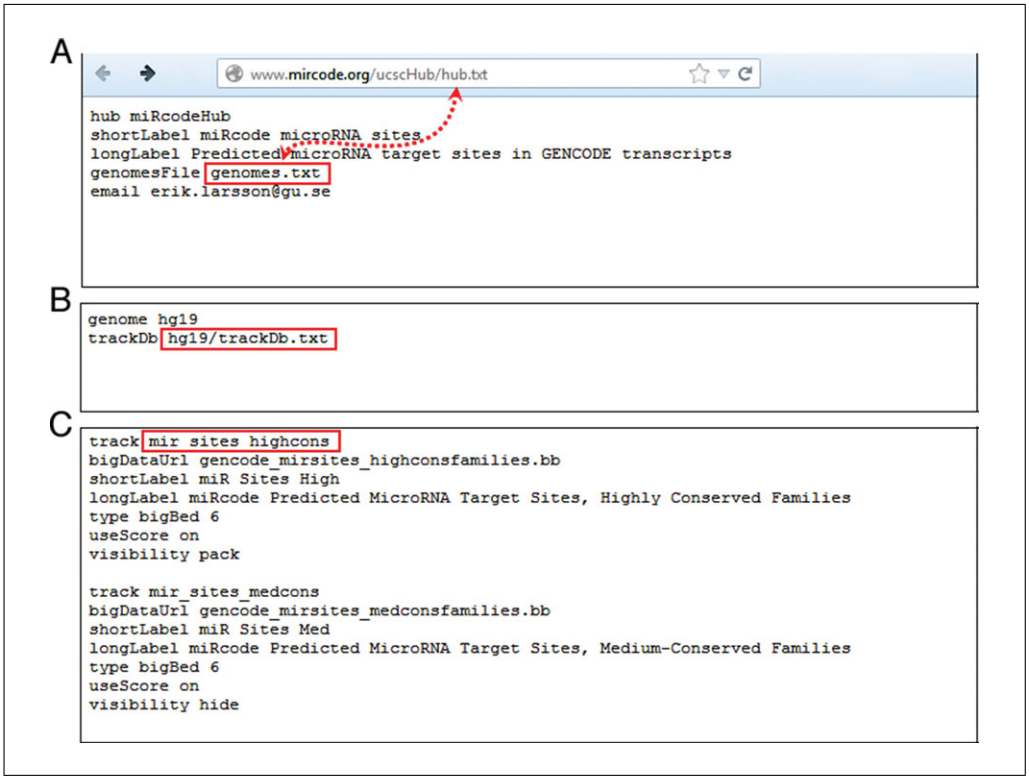


Figure 19.9.21 Text files supporting the miRcode Track Hub. (**A**) The primary page, *http://www.mircode.org/ucscHub/hub.txt*, defines the label for the Hub blue-bar group and shortLabel miRcode microRNA sites, and includes other information required to display the Hub. (**B**) The *http://www.mircode.org/ucscHub/genomes.txt* page can have the names of multiple genome assemblies. Here, only hg19 is referenced. (**C**) The `trackDb.txt` page is in the `hg19` directory and details the configuration of each track. Here, two tracks are defined that determine the tracks in the track controls (shortLabel) of the new blue bar group (Fig. 19.9.20). The bigDataUrl tag identifies the bigBed file that contains the data. It must be in the same directory as the text page, *http://www.mircode.org/ucscHub/hg19*.

Anyone can host a private track hub for their own data by locating files such as those we have seen here on a server of their own and pasting the URL into the box provided under the "My Hubs" tab on the Track Hubs page. A full description of the process is found in the Track Hub User's Guide, http://genome.ucsc.edu/goldenPath/help/hgTrackHubHelp.html. It is also possible to use this mechanism to host a Browser on sequenced genomes not hosted by UCSC. Details for this process are also available in the User's Guide.

COMMENTARY

At this juncture in scientific research, effective use of electronic data resources for query and display of molecular biology data is a requirement. The volume of data available to researchers exceeds the capacity of the traditional literature strategies for understanding many aspects of genomic context. Increasingly the "big data" projects in biology generate data for years before official publications are available, and this data can be accessed and mined much sooner. Electronic data management is definitely an essential skill in this field. This unit, describing many key functions of the UCSC Genome Browser, will enable researchers to make more efficient use of this resource. It also provides the foundations for understanding the data organization and query options that users will need for current and future data sets, and potentially for data generated in their own research. This introduction can only begin to expose the wealth of information available, and to seed ideas for the types of complex visualizations and queries that can be posed on genomic data to enhance and guide laboratory work. Other tools that have not been described here, including in silico PCR for virtual PCR queries, the Variant

Annotation Integrator for analyzing sequence variants, Genome Graphs for visualizing genome-wide association study data, and the Cancer Genome Browser can expand the reach as well. Researchers are encouraged to explore these other interlinked interfaces and portals at the UCSC Genome Browser site. The data can also be used for further exploration employing other genome browsers of different types, including importing data to stand-alone browsers such as IGB, or into tools such as Galaxy (*http://galaxyproject.org*), GenomeSpace (*http://www.genomespace.org*), and other sites and algorithms in the field of bioinformatics. Users can also deepen their understanding of the features and methods at the links for Training on the UCSC Genome Browser homepage to access a variety of additional training and documentation choices. Assistance with all aspects of using this resource is also available in the form of active mailing lists that can be accessed from the "Contact Us" homepage area.

Literature Cited

Bare, J.C., Koide, T., Reiss, D.J., Tenenbaum, D., and Baliga, N.S. 2010. Integration and visualization of systems biology data in context of the genome. *BMC Bioinform.* 11:382.

Church, D.M., Schneider, V.A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.C., Agarwala, R., McLaren, W.M., Ritchie, G.R.S., and Albracht, D. 2011. Modernizing reference genome assemblies. *PLoS Biol.* 9:e1001091

Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. (dbSNP Build ID: 138). Available from: *http://www.ncbi.nlm.nih.gov/SNP/.*

Di Bernardo, M.C., Crowther-Swanepoel, D., Broderick, P., Webb, E., Sellick, G., Wild, R., Sullivan, K., Vijayakrishnan, J., Wang, Y., Pittman, A.M., Sunter, N.J., Hall, A.G., Dyer, M.J., Matutes, E., Dearden, C., Mainou-Fowler, T., Jackson, G.H., Summerfield, G., Harris, R.J., Pettitt, A.R., Hillmen, P., Allsup, D.J., Bailey, J.R., Pratt, G., Pepper, C., Fegan, C., Allan, J.M., Catovsky, D., and Houlston, R.S. 2008. A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. *Nat. Genet.* 40:1204-1210.

Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., García-Girón, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kähäri, A.K., Keenan, S., Komorowska, M., Kulesha, E., Longden, I., Maurel, T., McLaren, W.M., Muffato, M., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H.S., Ritchie, G.R., Ruffier, M., Schuster, M., Sheppard, D., Sobral, D., Taylor, K., Thormann, A., Trevanion, S., White, S., Wilder, S.P., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Harrow, J., Herrero, J., Hubbard, T.J., Johnson, N., Kinsella, R., Parker, A., Spudich, G., Yates, A., Zadissa, A., and Searle, S.M. 2013. Ensembl 2013. *Nucleic Acids Res.* 41:D48-D55.

Haeussler, M., Gerner, M., and Bergman, C.M. 2011. Annotating genes and genomes with DNA sequences extracted from biomedical articles. *Bioinformatics* 27:980-986.

Jeggari, A., Marks, D.S., and Larsson, E. 2012. miRcode: A map of putative microRNA target sites in the long non-coding transcriptome. *Bioinformatics* 28:2062-2063.

Kuhn, R.M., Haussler, D., and Kent, W.J. 2013. The UCSC genome browser and associated tools. *Brief Bioinform.* 14:144-161.

Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B., Raney, B.J., Pohl, A., Malladi, V.S., Li, C.H., Lee, B.T., Learned, K., Kirkup, V., Hsu, F., Heitner, S., Harte, R.A., Haeussler, M., Guruvadoo, L., Goldman, M., Giardine, B.M., Fujita, P.A., Dreszer, T.R., Diekhans, M., Cline, M.S., Clawson, H., Barber, G.P., Haussler, D., and Kent, W.J. 2013. The UCSC Genome Browser database: Extensions and updates 2013. *Nucleic Acids Res.* 41:D64-D69.

NCBI Resource Coordinators. 2013. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 41:D8-D20.

Nicol, J.W., Helt, G.A., Blanchard, S.G. Jr., Raja, A., and Loraine, A.E. 2009. The Integrated Genome Browser: Free software for distribution and exploration of genome-scale datasets. *Bioinformatics* 25:2730-2731.

Raney, B.J., Dreszer, T.R., Barber, G.P., Clawson, H., Fujita, P.A., Wang, T., Nguyen, N., Paten, B., Zweig, A.S., Karolchik, D., and Kent, W.J. 2013. Track Data Hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics* 30:1003-1005.

Rosenbloom, K.R., Sloan, C.A., Malladi, V.S., Dreszer, T.R., Learned, K., Kirkup, V.M., Wong, M.C., Maddren, M., Fang, R., Heitner, S.G., Lee, B.T., Barber, G.P., Harte, R.A., Diekhans, M., Long, J.C., Wilder, S.P., Zweig, A.S., Karolchik, D., Kuhn, R.M., Haussler, D., and Kent, W.J. 2013. ENCODE data in the UCSC Genome Browser: Year 5 update. *Nucleic Acids Res.* 41:D56-D63.

Stein, L.D. 2013. Using GBrowse 2.0 to visualize and share next-generation sequence data. *Brief Bioinform.* 14:162-171.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. 2013. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief Bioinform.* 14:178-192.