Question 3

During a positional cloning project aimed at finding a human disease gene, linkage data have been obtained suggesting that the gene of interest lies between two sequence-tagged site markers. How can all the known and predicted candidate genes in this interval be identified? What BAC clones cover that particular region?

UCSC

One possible starting point for this search is the UCSC Genome Browser home page, at http://genome.ucsc.edu. From this page, select *Human* from the *Organism* pull-down menu in the blue bar at the side of the page, and then click *Browser*. On the Human Genome Browser Gateway page, change the assembly pull-down to *Dec. 2001*. To view a region of the genome between two query terms, enter the terms in the search box, separated by a semicolon. For example, to view the region between STS markers *D10S1676* and *D10S1675*, enter 'D10S1676;D10S1675' in the box marked *position* and press *Submit*. Because both of these markers map to a single position in the genome, the genome browser for the region between those markers is returned (Fig. 3.1).

The *STS Markers* track displays genetically mapped markers in blue and radiation hybrid–mapped markers in black. Click on the *STS Markers* label to expand that track and see each marker listed individually (Fig. 3.2). The markers of interest are called by their alternate names (AFMA232YH9 and AFMA230VA9 in this view) and are at the top and bottom of the interval, respectively (Fig. 3.2, arrows).

The full list of known genes in this display is shown in the *Known Genes* track (Fig. 3.1). These protein-coding genes are taken from the RefSeq mRNA sequences compiled at the NCBI[10] and aligned to the genome assembly using the BLAT program[8]. To export a list of the genes, or other features, in this region, click the *Tables* link in the top blue bar. For more information about a particular gene (such as *MGMT*), click on the gene symbol to get a list of additional links to resources such as Online Mendelian Inheritance in Man (OMIM), PubMed, GeneCards and Mouse Genome Informatics (MGI; Fig. 3.3). Many tracks, including *Acembly Genes, Ensembl Genes* and *Fgenesh++ Genes*, indicate predicted genes (see Question 7).To view the full set of features in any of these categories, click on the title of that track on the left side of the screen in Fig. 3.1. To view brief descriptions of these tracks, as well as others not mentioned, click on the gray box to the left of the track or scroll down to *Track Controls* and click on the title of a feature of interest. Explanations of the gene-prediction programs can be found in Question 7. Reset the browser to its default settings by clicking on the *reset all* button below the tracks.

To see the BAC clones used for sequencing, return to the page illustrated in Fig. 3.1 and click on *Coverage* at the left side of the screen to expand that track. Here BAC clones are listed individually, with finished regions shown in black and draft regions shown in various shades of gray (Fig. 3.4). For details such as size and sequence coverage of a specific clone, click on the clone accession number (such as *AL355529.21*, arrow). From this screen, click on the accession number (as shown in Fig. 3.5) to link to the NCBI Entrez document summary for the clone. The full GenBank entry can be viewed by clicking on *AL355529* on the Entrez document summary page.

According to NCBI naming conventions, this clone is from the RP11 library and has been named 85C15. RP11 is the NCBI designation for RPCI-11, a commonly used human BAC library produced at the Roswell Park Cancer Institute. More information on the naming conventions of genomic sequencing libraries can be found at the NCBI's Clone Registry (Fig. 3.6; http://www.ncbi.nlm.nih.gov/genome/clone/nomenclature.shtml).

Clone ordering information is also available, at http://www.ncbi.nlm.nih.gov/genome/clone/ordering.html.

NCBI

The NCBI MapViewer allows for direct viewing of the region between two markers, as long as both markers are on the master map. If, for example, the master map is a cytogenetic one, one can search chromosome 22 for the region between band numbers 22q12.1 and 22q13.2. If the master map is Gene_Seq, one can view the region between two mapped genes.

Access the Map Viewer home page by starting at the NCBI home page (http://www.ncbi.nlm.nih.gov) and clicking *Human map viewer* in the list on the right-hand side of the page. To view multiple hits on the same chromosome, type in the search terms separated by the word 'OR'. To see the same region between the STS markers D10S1676 and D10S1675, for example, type 'D10S1676 OR D10S1675' in the search box, and hit *Find*. At the top of the resulting page (Fig. 3.7), two red tick marks on the chromosome cartoon indicate that the markers map close to each other on chromosome 10. The search results at the bottom of the page show the alternative names for the two markers (AFMA232YH9 and AFMA230VA9) as well as the maps on which they have been placed. To view both markers at the same time, click on the link for chromosome *10* in the chromosome diagram. Fig. 3.8 shows the region around D10S1676 and D10S1675, with the original queries highlighted in pink. Red lines connect the positions of the marker on the different maps.

The *Maps & Options* link, in the horizontal blue bar near the top of the page, allows the user to customize the maps and region displayed. To view, for example, the known and predicted genes

One can also search for a region between two STS markers using the MapView at Ensembl. Start at the Ensembl Human Genome Browser at http://www.ensembl.org/Homo_sapiens/, click on the idiogram of any chromosome to access the MapView, and enter the marker names in the *Jump to Contigview* section. To use Ensembl to obtain a list of genes (or other annotations) in a defined chromosomal region, click on *Export→Gene List* from any ContigView window (Fig. 1.14, center yellow bar).

in this region, as well as the BAC clones from which the sequence was derived, click on the link to open the *Maps & Options* window (Fig. 3.9). First remove all the maps except *Gene* and *STS* from the *Maps Displayed* box by highlighting them, and selecting *<<REMOVE*. Next, add the *Transcript (RNA), GenomeScan, Component* and *Contig* maps by selecting them from the *Available Maps* box and selecting *ADD>>*. Make the STS map the master by highlighting it, then selecting *Make Master/Move to Bottom*. To limit the view such that only the STSs between D10S1676 and D10S1675 are shown, type the marker names in the *Region Shown* boxes. Hit *Apply* to see the aligned maps. In some cases, it may be useful to select a page size larger than the default of 20 to view more data in the browser window.

Fig. 3.10 shows the maps, as specified in the *Maps & Options* window. The green dots to the right of the STS map show all the maps on which the markers appear. This is a fairly long region of chromosome 10, and not every STS marker is shown. In particular, although there are 611 STSs in this region, only 20 are shown by name in this view. For each known gene, the Genes_Seq map shows all the exons that have been mapped to the genome. Exons for individual known mRNAs are shown on the RNA (Transcript) map. Unless a gene is alternatively spliced, the Genes_Seq and RNA maps will be the same. The GScan (GenomeScan) map shows the NCBI's gene predictions. Any of these genes, known or predicted, are candidates for the disease gene.

The NCBI's assembled contigs, also known as the NT contigs, are found in the Contig map. Blue segments come from finished sequence, orange from draft. These contigs are constructed from the individual GenBank sequence entries shown in the Comp (Component) map. Draft HTG records (phase 1 and 2; see http://www.ncbi.nlm.nih.gov/HTGS/) are displayed in orange and finished HTGs in blue. Most of these GenBank entries are derived from BAC clones. The tiling paths of the BAC clones that were assembled into contigs are clearly visible. One can obtain more details about an entry, including the clone name, by clicking on the accession number to link to Entrez. The clone name is visible directly in the MapViewer if the Comp map is the master. A map can be quickly made the master map by clicking on the blue arrow next to its name.

Because this is a zoomed-out view of the chromosome, individual genes and GenBank entries are difficult to visualize. Zooming in, using the controls in the blue sidebar, will provide a region in more detail. Alternatively, click on the *Data As Table View* in the left sidebar to retrieve all data, including those hidden in this view, as a text-based table (partially shown in Fig. 3.11).
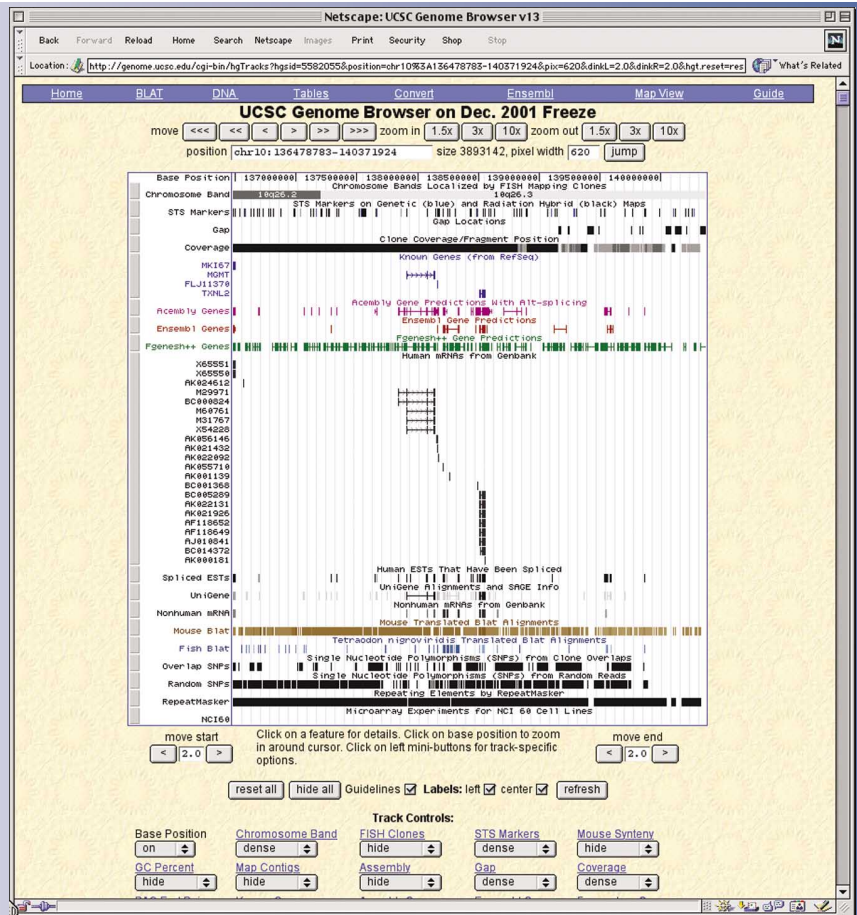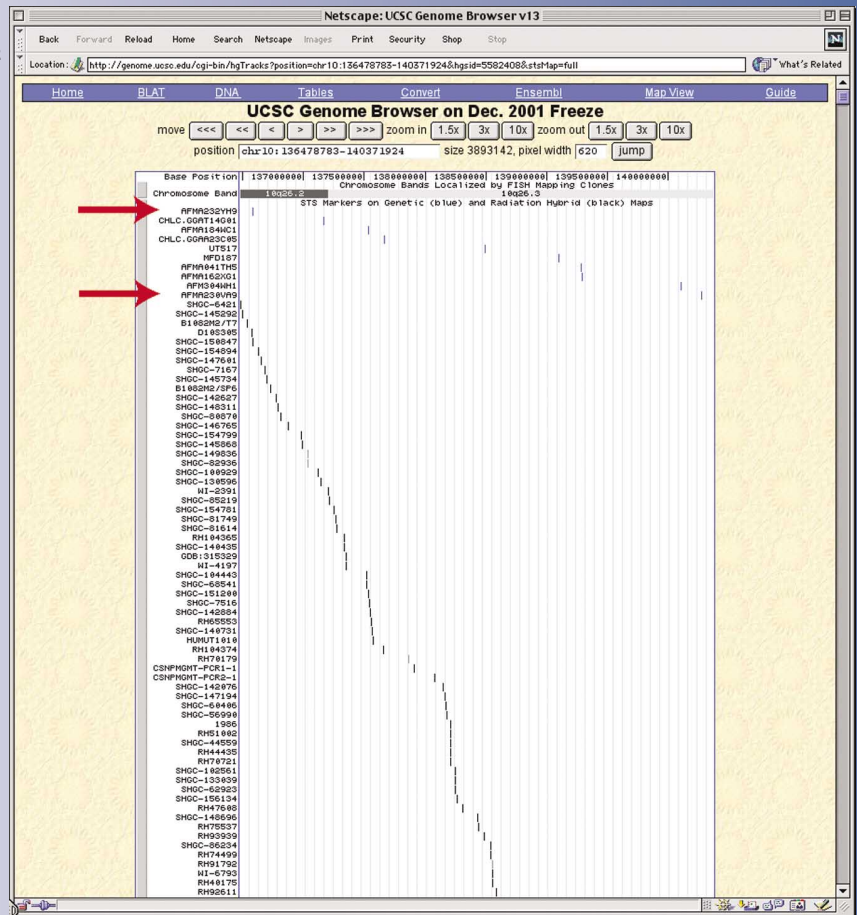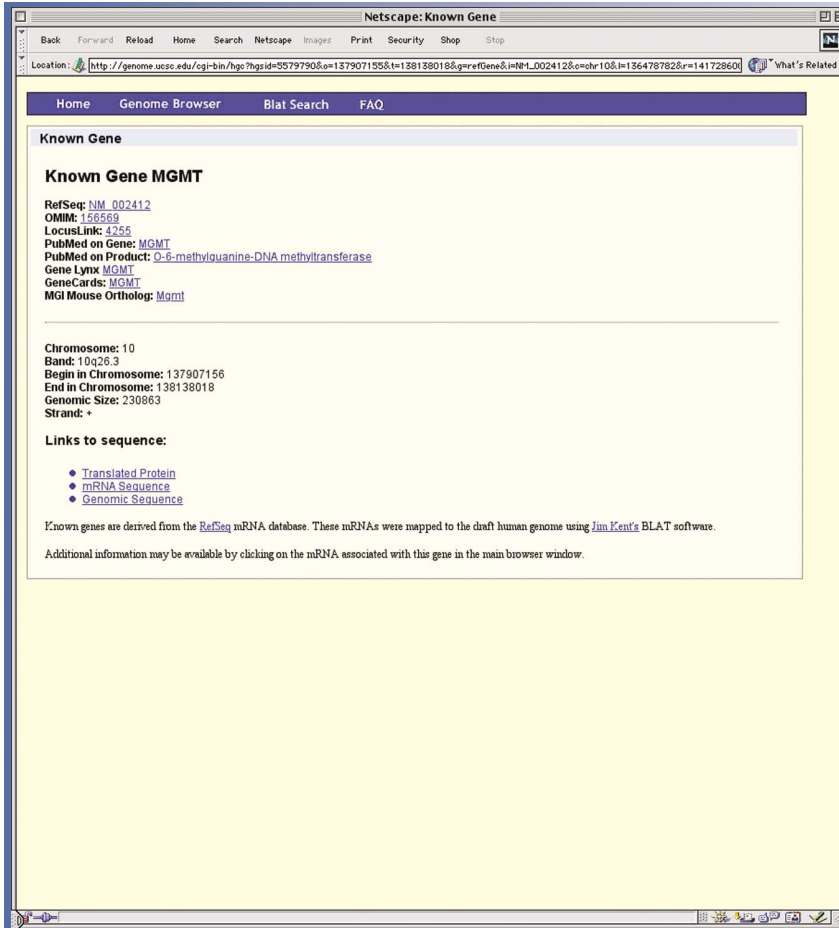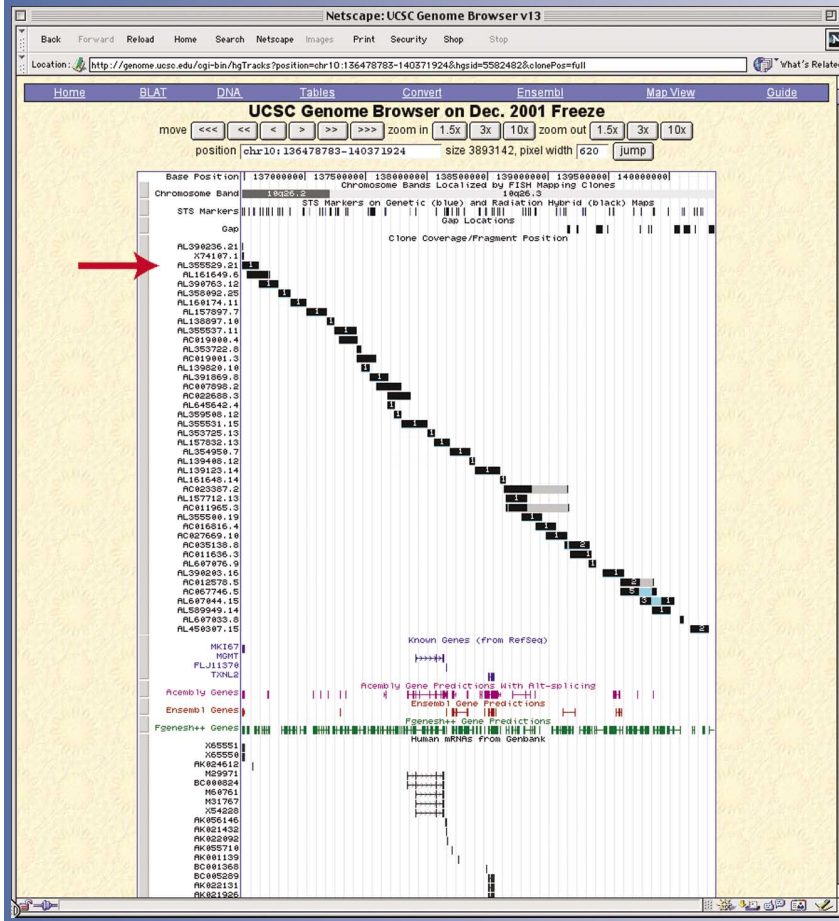
Figure 3.1

Figure 3.2

Figure 3.3

Figure 3.4

Figure 3.5
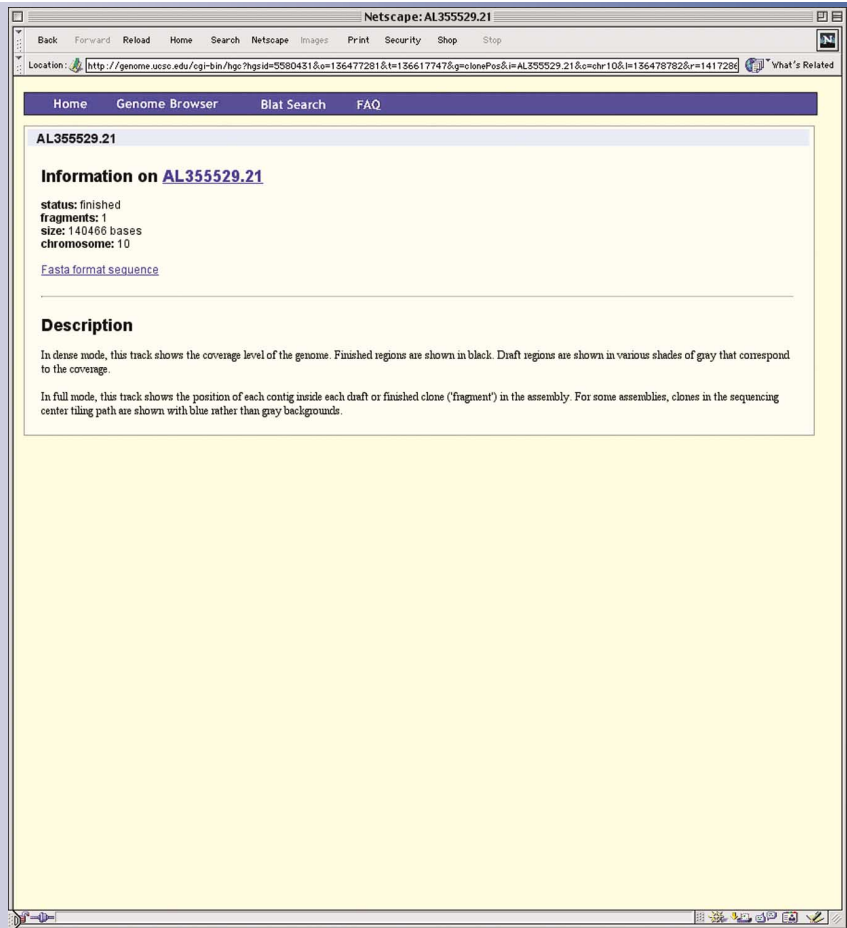
Figure 3.6
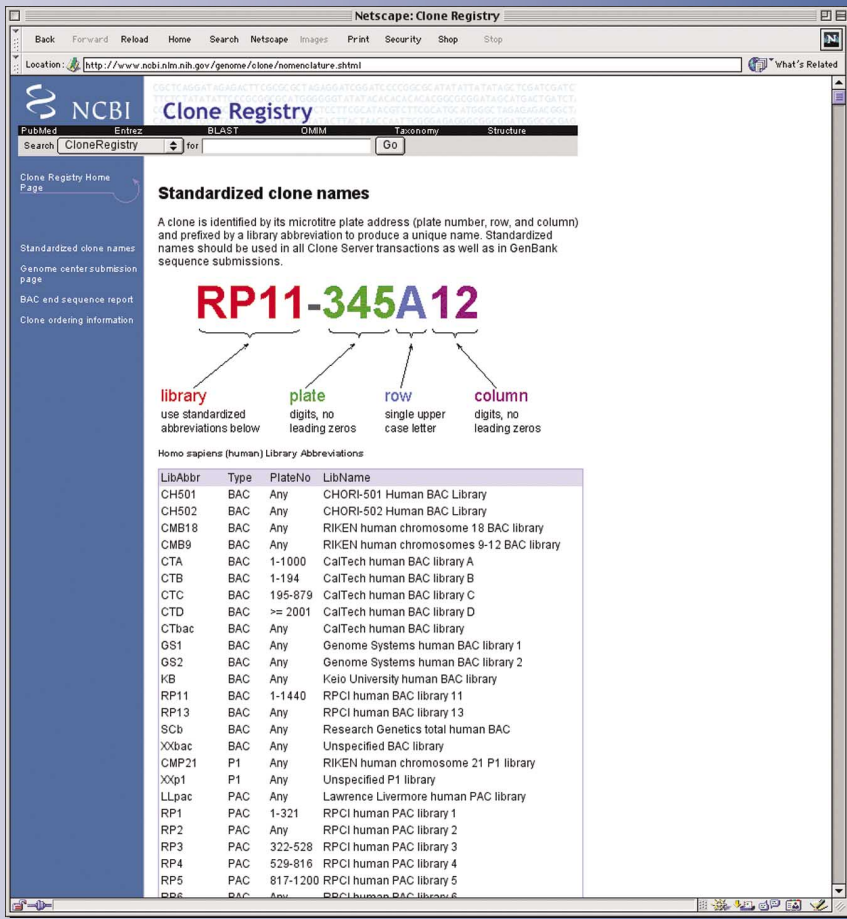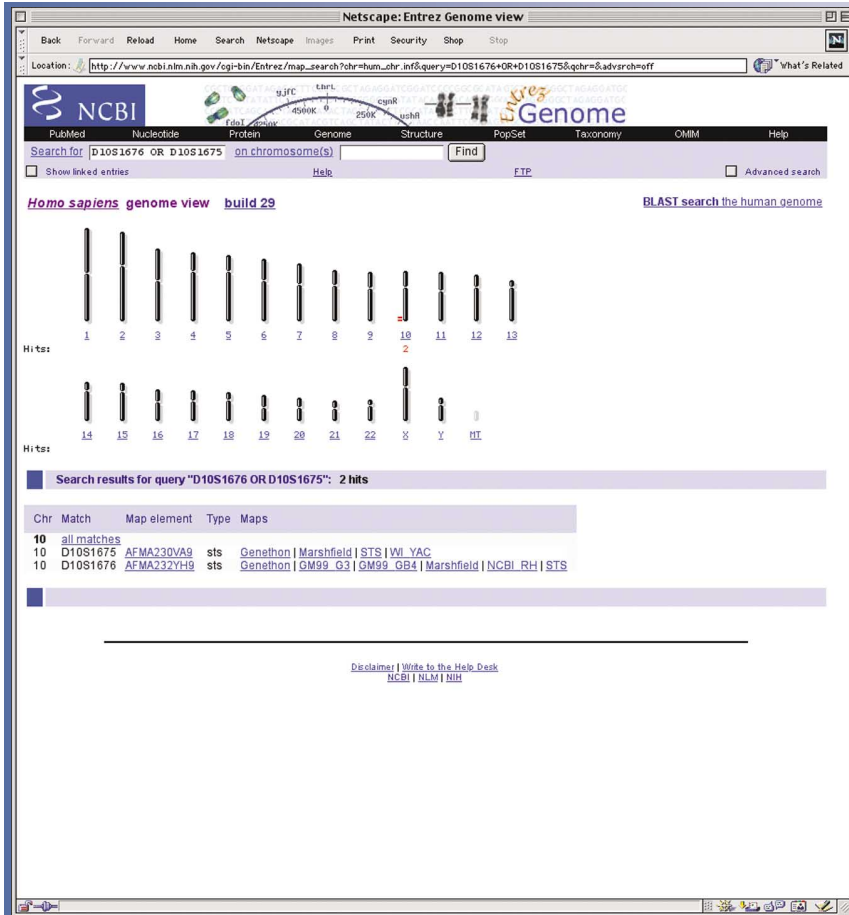
Figure 3.7
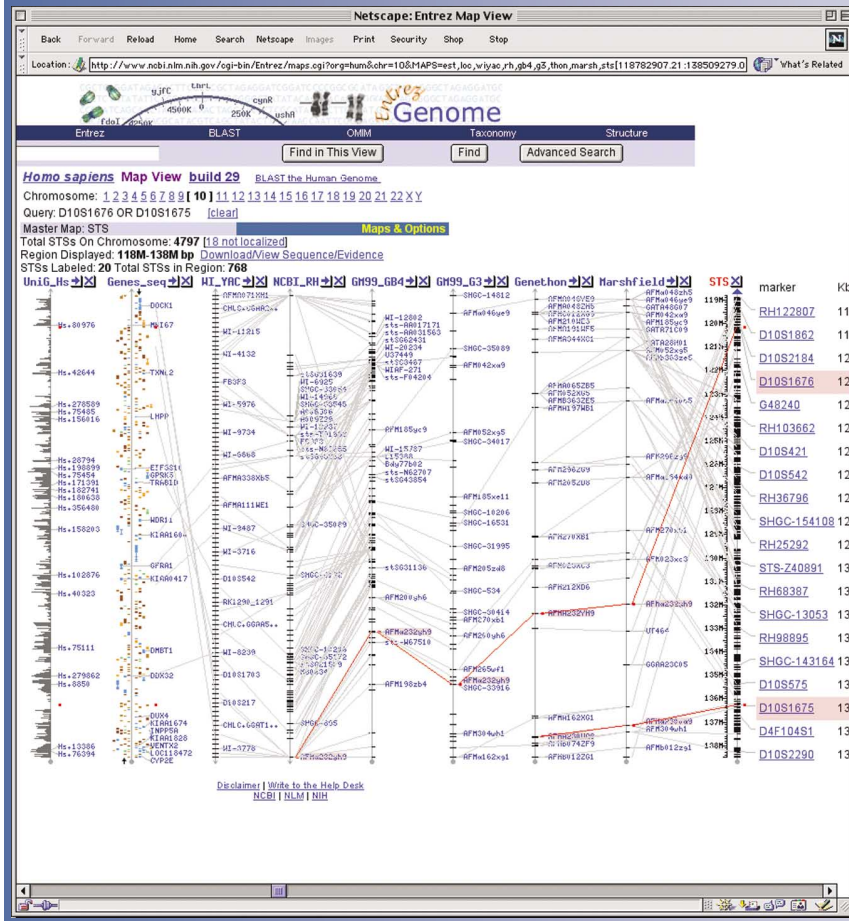
Figure 3.8

Figure 3.9

Figure 3.10

Figure 3.11