Question 11

An investigator has identified and cloned a human gene, but no corresponding mouse ortholog has yet been identified. How can a mouse genomic sequence with similarity to the human gene sequence be retrieved?

For purposes of this example, assume that the user does not already have the human sequence of interest to hand. The first step will be to locate the human gene of interest using the UCSC Genome Browser. Begin by pointing to the UCSC Genome Browser home page, at http://genome.ucsc.edu. Select *Human* from the *Organism* pull-down menu and then click on *Browser*; both are located on the blue navigation bar at the left side of the page. This will take the user to the Human Genome Browser Gateway. Select the *Nov. 2002* version of the UCSC genome assembly, type the gene symbol 'AGPS' into the *position* box, and then click *Submit*. On the resulting page, follow the link for *AGPS* in the *Known Genes* section.

The result of the search on AGPS is shown in Fig. 11.1. In the main figure are a series of 'tracks', which are labeled along the left-hand side. The *Known Gene* track is for *AGPS*, corresponding to the query. Clicking on *AGPS* returns a summary of information on that gene, including the full name of the protein product (alkylglycerone phosphate synthase precursor), a link to the GeneCards database at the Weizmann Institute[20] and links to the translated protein, mRNA and genomic sequences. Focus now on the track labeled *Mouse Cons.* What is shown in this track is the conservation between the human and mouse genomes for 50bp windows in the human genome that have at least 15 bp aligned to mouse. Higher conservation scores are indicated as higher peaks on the *Mouse Cons* track. Note that many of the peaks coincide with exons in the AGPS gene, although some peaks are within introns. More details about this track can be found by clicking on the *Mouse Cons* hyperlink found below the main graphical display.

Click anywhere within the *Mouse Cons* track to view the details of the alignment (Fig. 11.2). This mouse genomic sequence is 279 nt in length and aligns with the human sequence in fifteen blocks. To view the alignment itself, click on the *View details of parts of alignment within browser window* link. On the resulting page (Fig. 11.3), the mouse sequence is shown on top, with the region of alignment in blue. The human genomic sequence is shown next, and a side-by-side alignment of the human and mouse sequences is at the bottom of the web page (not shown).

> The NCBI's UniGene_Mouse map shows alignments of mouse mRNA and EST sequences with the human genome. Add this map using *Maps & Options* (Fig. 3.9). The easiest way to find the mouse ortholog of a human gene is probably to use Ensembl's precomputed *Homology Matches*. These matches, where available, link directly from a human gene to a putative mouse homolog (Fig. 1.9).
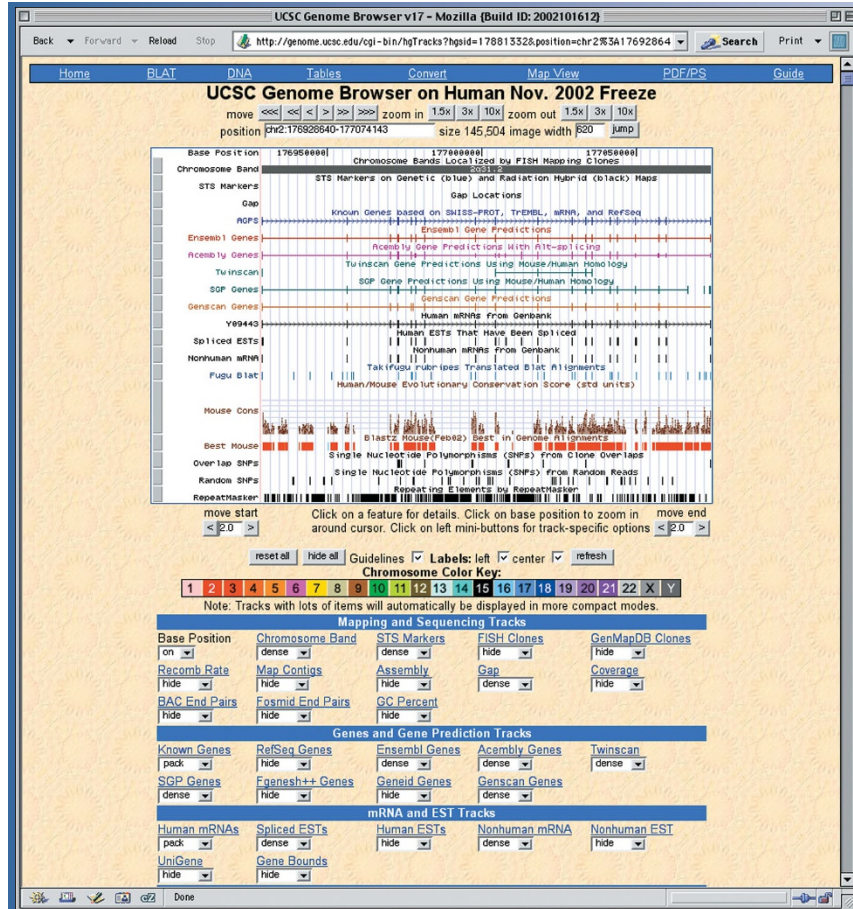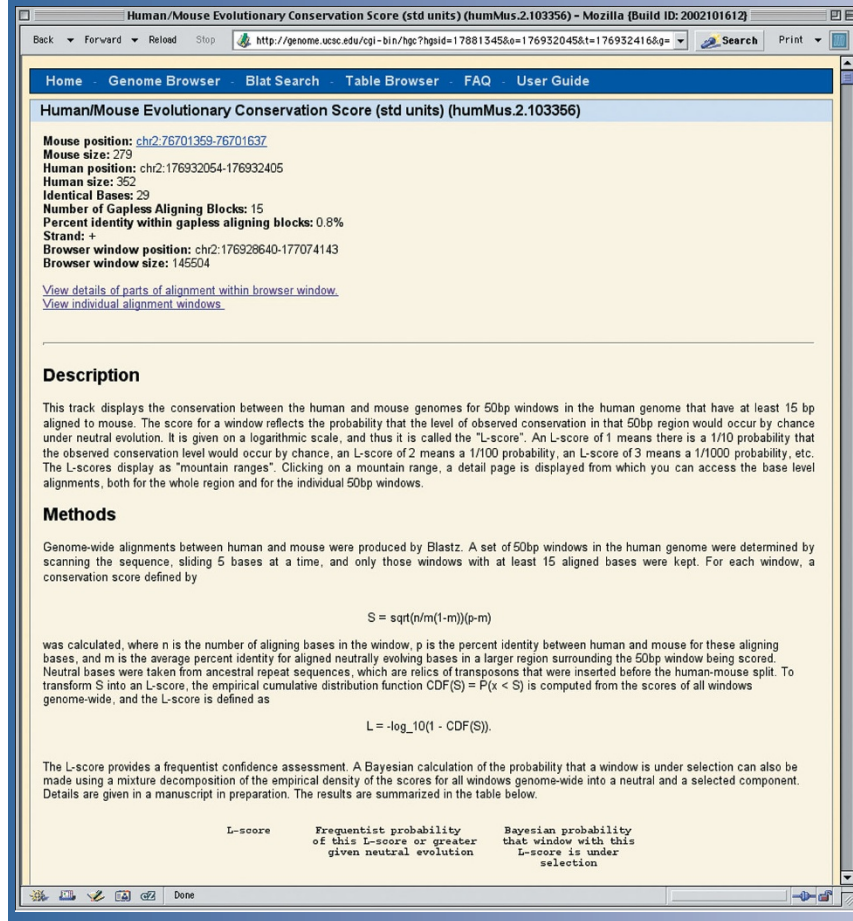
Figure 11.1

UCSC Genome Browser on Human Nov. 2002 Freeze

Figure 11.2

Human/Mouse Evolutionary Conservation Score (std units) (humMus.2.103356)

Mouse position: chr2:76701359-76701637
Mouse size: 279
Human position: chr2:176932054-176932405
Human size: 352
Identical Bases: 29
Number of Gapless Aligning Blocks: 15
Percent identity within gapless aligning blocks: 0.8%
Strand: +
Browser window position: chr2:176928640-177074143
Browser window size: 145504

View details of parts of alignment within browser window.
View individual alignment windows

Description

This track displays the conservation between the human and mouse genomes for 50bp windows in the human genome that have at least 15 bp aligned to mouse. The score for a window reflects the probability that the level of observed conservation in that 50bp region would occur by chance under neutral evolution. It is given on a logarithmic scale, and thus it is called the "L-score". An L-score of 1 means there is a 1/10 probability that the observed conservation level would occur by chance, an L-score of 2 means a 1/100 probability, an L-score of 3 means a 1/1000 probability, etc. The L-scores display as "mountain ranges". Clicking on a mountain range, a detail page is displayed from which you can access the base level alignments, both for the whole region and for the individual 50bp windows.

Methods

Genome-wide alignments between human and mouse were produced by Blastz. A set of 50bp windows in the human genome were determined by scanning the sequence, sliding 5 bases at a time, and only those windows with at least 15 aligned bases were kept. For each window, a conservation score defined by

$$S = \sqrt{n/m(1-m)}(p-m)$$

was calculated, where n is the number of aligning bases in the window, p is the percent identity between human and mouse for these aligning bases, and m is the average percent identity for aligned neutrally evolving bases in a larger region surrounding the 50bp window being scored. Neutral bases were taken from ancestral repeat sequences, which are relics of transposons that were inserted before the human-mouse split. To transform S into an L-score, the empirical cumulative distribution function CDF(S) = P(x < S) is computed from the scores of all windows genome-wide, and the L-score is defined as

$$L = -\log_{10}(1 - CDF(S)).$$

The L-score provides a frequentist confidence assessment. A Bayesian calculation of the probability that a window is under selection can also be made using a mixture decomposition of the empirical density of the scores for all windows genome-wide into a neutral and a selected component. Details are given in a manuscript in preparation. The results are summarized in the table below.

L-score	Frequentist probability of this L-score or greater given neutral evolution	Bayesian probability that window with this L-score is under selection

Figure 11.3