Question 7

How would an investigator easily find compiled information describing the structure of a gene of interest? Is it possible to obtain the sequence of any putative promoter regions?

doi:10.1038/ng1195

One place to initiate this search is at UCSC's Genome Browser, at http://genome.ucsc.edu. For purposes of this example, consider the gene encoding pendrin (*PDS*), a protein associated with developmental abnormalities of the cochlea, sensorineural hearing loss and diffuse thyroid enlargement (goiter).

From the UCSC home page, choose *Human* from the pull-down *Organism* list, and click on *Browser*. The user is now at the Human Genome Browser Gateway. The search in this case is simple: select *Nov. 2002* from the *assembly* pull-down menu, type *pendrin* into the *position* box, and then click *Submit*. The returned results indicate one known gene and two mRNA sequences; click on the accession number of the mRNA sequence AF030880 to continue. The user will now be presented with a graphic overview of the region containing this mRNA. To gain a better perspective of the region, click on the *1.5* button next to *zoom out*. Finally, click the *reset all* button on the middle of the page to reset the tracks to their default settings.

Carrying out these steps will produce an output similar to that shown in Fig. 7.1. For the purpose of this question, however, the default settings are not ideal. Using the *Track Controls* at the bottom of the figure, and following the example in Fig. 7.2, set some tracks to *hide* mode (not shown), others to *dense* (all data condensed onto one line) and some to *full* (a separate line for each feature, up to 300). Before considering the actual data within these tracks, a brief discussion of the content and representation of these tracks is warranted. Many were provided to UCSC by outside individuals. Further information on the gene prediction methods briefly discussed below can be found elsewhere[15].

The general convention for the *RefSeq Genes* and predicted gene tracks (Fig. 7.1) is that each coding exon is shown as a tall, vertical bar or block. 5′ and 3′ untranslated regions are shown as shorter vertical bars or blocks.

Connecting introns are shown as very thin lines. The direction of transcription is indicated by the arrows along that thin line.

RefSeq Genes are taken from mRNA reference sequences within LocusLink[10]. These reference sequences have been aligned against the genome using BLAT.

The *Acembly Gene Predictions With Alt-splicing* track is derived from the alignment of human mRNA and EST sequence data against the genome, using the program Acembly. This program attempts to find the best alignment of each mRNA against the genome and considers alternative splice models. If more than one gene model with statistical significance can be produced, each of these is shown in the display. Additional information on Acembly can be found on the NCBI web site at http://www.ncbi.nih.gov/IEB/Research/Acembly/.

The *Ensembl Gene Predictions* track[7] is provided by Ensembl. The Ensembl genes are predicted by a range of methods, including homology to known mRNAs and proteins, *ab initio* gene prediction using GENSCAN and gene prediction HMMs.

The *Fgenesh++ Gene Predictions* come from a method that predicts internal exons by looking for structural features such as donor and acceptor splice sites, putative coding regions and intronic regions both 5′ and 3′ to a putative exon using a dynamic programming algorithm; the method also takes into account protein similarity data[16].

The *Genscan Gene Predictions* derive from a method called GENSCAN, through which introns, exons, promoter sites and poly(A) signals can be identified. Here, the method does not expect the query sequence to represent one and only one gene, so it can make accurate predictions for either partial genes or multiple genes separated by intergenic DNA[11].

The *Human mRNAs from Genbank* track shows alignments between human mRNAs in GenBank and the genome sequence.

The *Spliced ESTs* and *Human EST* tracks show the alignment of ESTs from GenBank against the genome. Because ESTs usually represent fragments of transcribed genes, there is high likelihood that an EST corresponds to an exonic region.

Finally, the *Repeating Elements by RepeatMasker* track shows, as its name would suggest, repetitive elements such as short and long interspersed nuclear elements (SINEs and LINEs), long terminal repeats (LTRs) and low-complexity regions (http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker). It is customary to remove or 'mask' these elements before applying a gene prediction method to a nucleotide sequence.

Returning to the example shown in Fig. 7.2, notice that most of the tracks return a nearly identical gene prediction; as a rule, exons predicted by multiple methods increase the likelihood that the prediction is actually correct and does not represent a 'false positive'. Most of the methods show a 3′ untranslated region, indicated by the heavy, shorter block at the left of the predictions. The Acembly track shows one possible alternative splice in addition to the full-length product shown in the fourth line of that section, a prediction that agrees with those shown in most of the other tracks. The Genscan track extends off to both the right and the left: GENSCAN can be used to predict multiple genes, and this display implies that the method has been applied in this fashion. Note that UCSC also provides other gene prediction tracks, including Known Genes, Twinscan, SGP Genes, and Geneid Genes. For the most comprehensive analysis of genes in this region, all of these tracks should be analyzed.

Although these graphical overviews are useful, the investigator will more often than not want the actual sequence corresponding to these blocks. For this example, the Fgenesh++ prediction will be used as the basis for obtaining raw sequence data, but the steps will be identical regardless of which track is chosen. Click on the track labeled *Fgenesh++ Gene Predictions* to go to a summary page describing the prediction (Fig. 7.3). The

The NCBI also provides gene predictions, computed using the program GenomeScan[17]. These models are shown on the GenomeScan and Gene_Seq maps.

region has sequence similarity to the pendrin gene (which was already known at the beginning of the example). The size and the beginning- and end-points of the prediction are given, and it is indicated that the prediction lies on the plus strand; this was also indicated in Fig. 7.2 by the right-pointing arrows in the intronic regions. To obtain the sequence, click on *Genomic Sequence*. The user will be taken to a query page entitled *Get Genomic Sequence Near Gene*, from which various combinations of coding and untranslated exons, as well as introns, can be obtained (Fig. 7.4). For each of the options, the sequence is returned in FASTA format, with the nucleotide coordinates being given in the definition line.

Promoter returns just the promoter region, as shown in Fig. 7.5.

5′ UTR Exons returns any exons that comprise 5' untranslated region sequence

CDS Exons return coding sequence exons, or sequence that is translated into protein

3′ UTR Exons returns any exons that comprise 3' untranslated region sequence

Introns returns intron sequence

Downstream returns just downstream sequence

The user can also choose whether results are returned in upper or lower case.

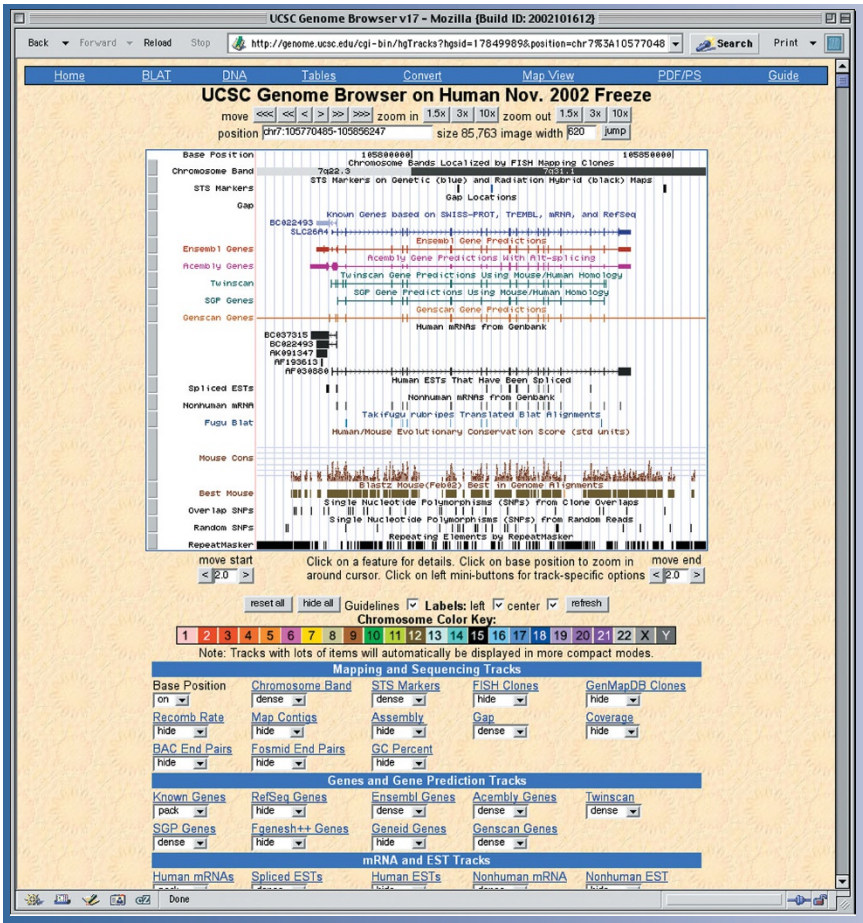


Figure 7.1

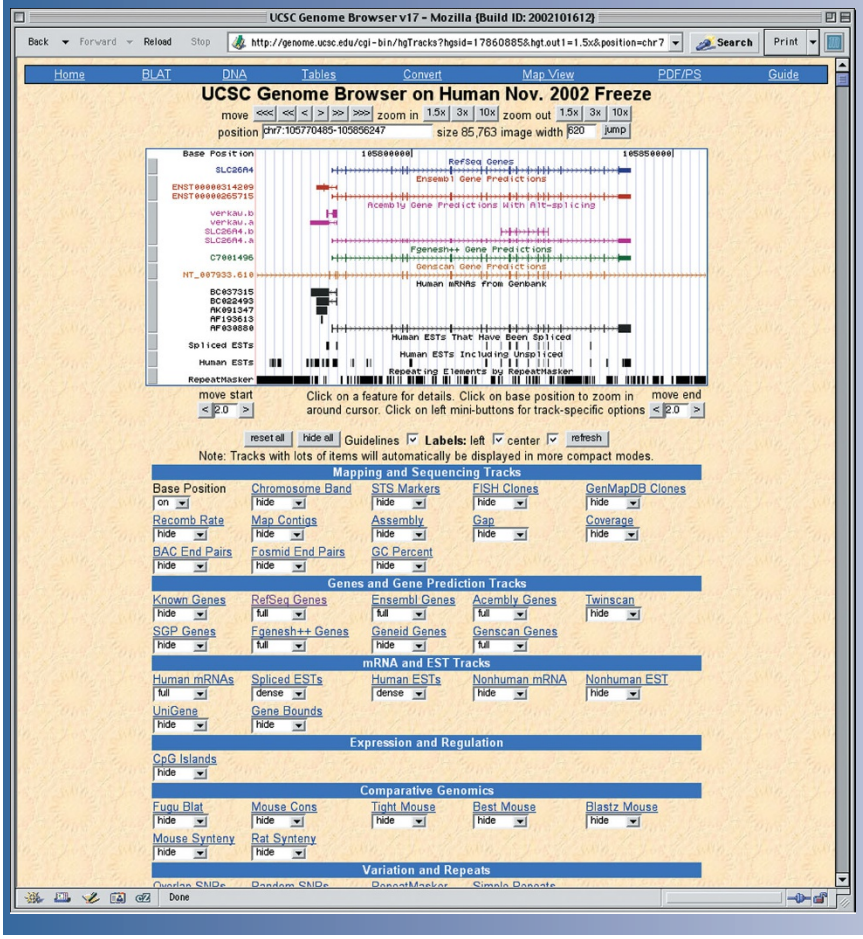


Figure 7.2

Figure 7.3

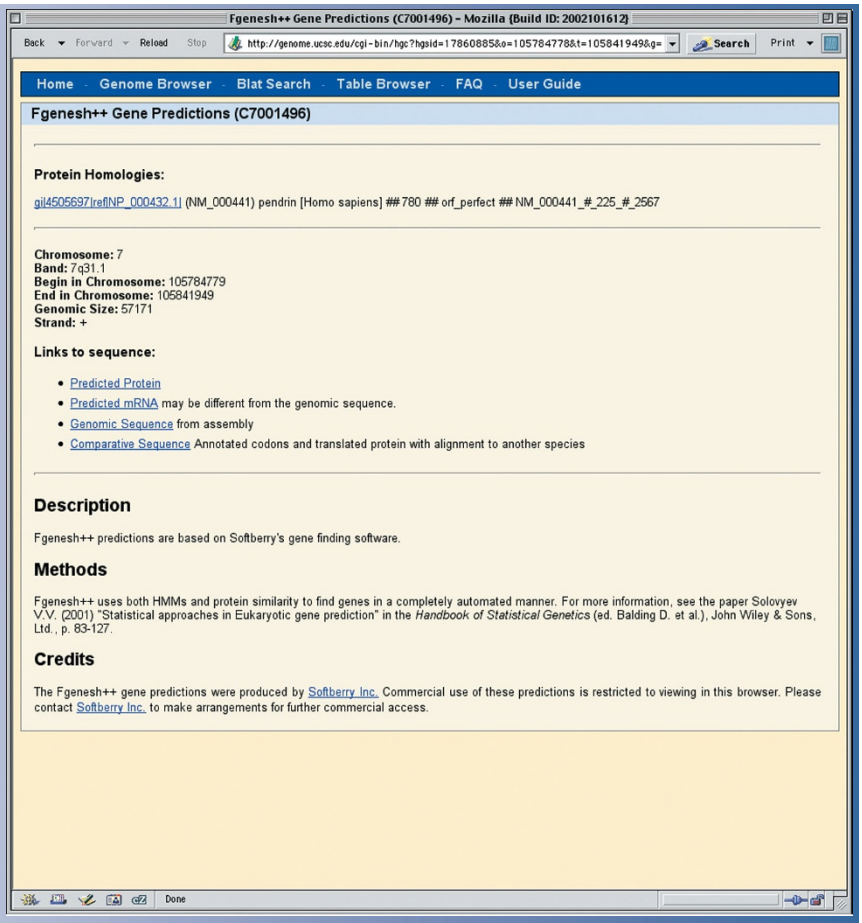
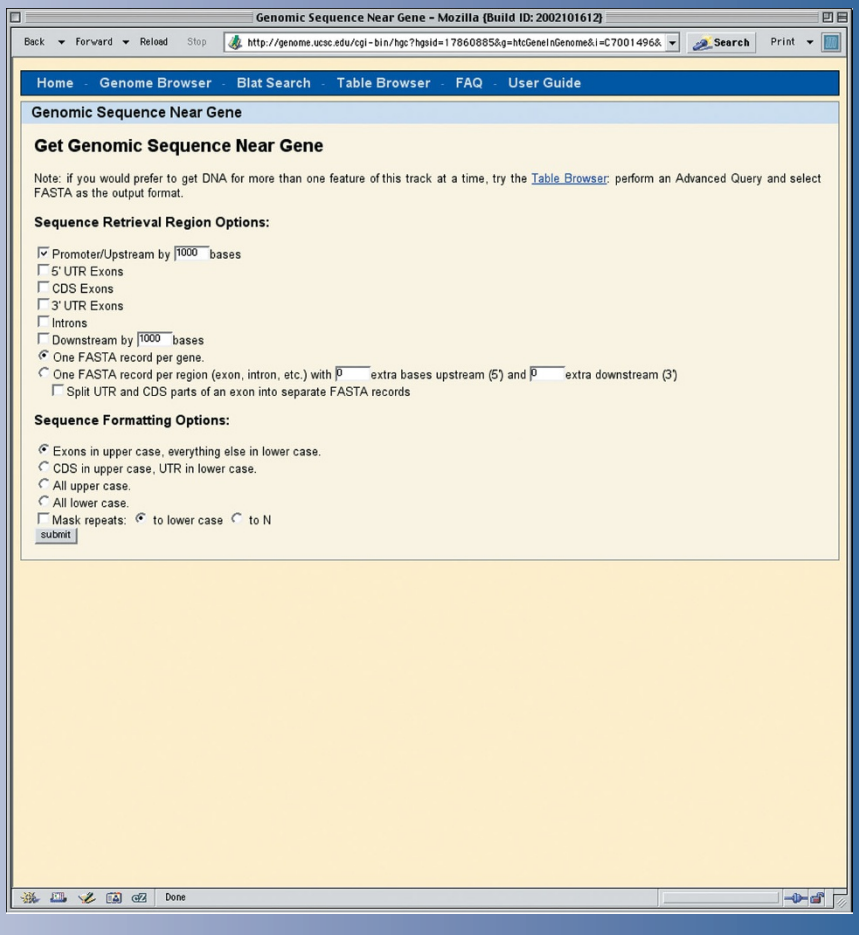


Figure 7.4



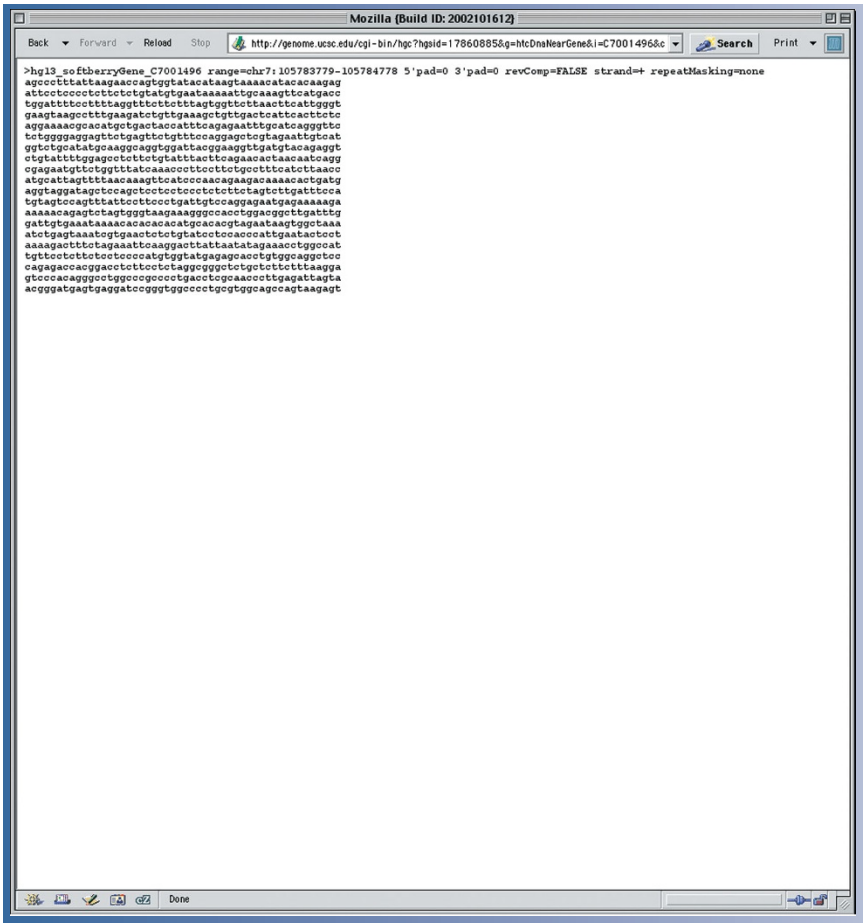


Figure 7.5