Question 5

Given a fragment of mRNA sequence, how would one find where that piece of DNA mapped in the human genome? Once its position has been determined, how would one find alternatively spliced transcripts?

doi:10.1038/ng970

For the purpose of this example, the fragment of mRNA of interest is contained within GenBank accession number BG334944. First, retrieve the nucleotide sequence of this EST using the NCBI's Entrez interface, at <http://www.ncbi.nlm.nih.gov/Entrez/>. Type 'BG334944' into the text box at the top of the page, change the pull-down menu to *Nucleotide* and press *Go*. The resulting page shows one entry, corresponding to accession number BG334944. To retrieve this sequence in FASTA format (a common format for bioinformatics programs), change the pull-down menu on this page to *FASTA* and then press *Text* (Fig. 5.1). A new web page containing only the sequence, in FASTA format, is produced (Fig. 5.2); copy the resulting sequence.

To determine where this sequence maps within the genome, use UCSC's BLAT tool⁸. Begin this search by pointing your web browser to 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 on the side of the page, and then click *Blat*. Paste the FASTA-formatted sequence obtained from Entrez (above) into the large text box on the BLAT search page (Fig. 5.3), change the *Freeze* pull-down menu to *Dec. 2001*, change the *Query* pull-down menu to *DNA* and then press *Submit*. The server will (very quickly) return the search results; in this case, a single match of length 636 is found on the forward strand of chromosome 9 (Fig. 5.4).

To obtain more details on this hit, click the *details* link, to the left of the entry. A long web page is returned, with three major sections: the mRNA sequence (Fig. 5.5, top), the genomic sequence (Fig. 5.5, middle) and an alignment of the mRNA sequence against the genomic sequence (see Fig. 5.9 for an example). In the alignment in Fig. 5.5, matching bases in the cDNA and genomic sequences are colored in darker blue and capitalized. Gaps are indicated in lower-case black type. Light blue upper-case bases mark the boundaries of aligned regions on either side of a gap and are often splice sites.

Returning to the BLAT summary page for this search (Fig. 5.4), click on *browser*. This will produce a graphic representation of where this particular mRNA sequence aligns to the genome (Fig. 5.6). The track labeled *Chromosome Band* indicates that the mRNA maps to 9q34.11. The query sequence itself is represented on the line labeled *Your Sequence from BLAT Search* (arrow, Fig. 5.6). The sequence is shown as being discontinuous: regions of similarity are shown as vertical lines, gaps are shown as thin horizontal lines, and the direction of the alignment is indicated by the arrowheads. The aligned regions of the EST query correspond to the exons of a known gene, shown on the line immediately below (*Known Genes*, here RAB9P40). Typing the EST name, BG334944, directly into a UCSC search box would have generated a similar result to that shown in Fig. 5.6, but part of the purpose of this example is to illustrate the use of BLAT.

Approximately halfway down the graphic is a track labeled *Human ESTs That Have Been Spliced*. This track is at first shown in dense mode, with all the ESTs condensed onto a single line. To see all of the ESTs that align with the genome in this region,

potentially representing differentially spliced transcripts, click on the track's label. This will expand this area of the figure so that each EST occupies a single line (Fig. 5.7). The ESTs are of varying length, but most contain the same exons as the known gene and are (presumably) spliced in the same way. Close inspection indicates that some of the ESTs are missing one or more exons compared with the known gene. Consider the lines marked *BE798864* and *W52533*: the former appears to be missing the fifth exon, whereas the latter is missing the fourth, fifth and sixth exons.

Any of the ESTs can be examined in more detail by clicking on that particular line. Here, click on the line for *BE798864* (arrow, Fig. 5.7) to reach the information page for this EST (Fig. 5.8). The EST is 99.8% identical to the genomic sequence; clicking anywhere on the hyperlinked line in the section marked *EST/Genomic Alignments* returns the actual side-by-side alignment (Fig. 5.9). Differences exist at the ends of the EST, but the sequences are identical in the region surrounding the putative missing exon.

An alternatively spliced mRNA is more likely to be of biological significance when it changes the sequence of the encoded, wildtype protein. To determine whether EST *BE798864* could encode a protein different from that of the known gene (*RAB9P40*), one can simply compare the two sequences directly against each other using the NCBI's BLAST 2 Sequences tool. First, open a new web browser window, because information from the above search will be needed here; this will prevent having to use the browser's *Back* and *Forward* keys excessively and is a good general rule when using multiple web tools. Then access the BLAST home page, at <http://www.ncbi.nlm.nih.gov/BLAST>. Select *BLAST 2 Sequences*, under the header labeled *Pairwise BLAST*. On this page, the user can simply enter accession numbers rather than cutting and pasting sequences into the text boxes. For the EST, simply enter its accession number

Ensembl also displays database hits that overlap with each exon in a transcript. These hits may include proteins as well as ESTs and mRNAs, and may illustrate alternatively spliced products. The hits are shown as green boxes in the TransView (Fig. 13.5), which can be accessed in a number of ways; for example, by clicking on the *View Evidence* box for a transcript on the GeneView (Fig. 1.10). Another good starting point for visualizing alternatively spliced transcripts is the NCBI's Model Maker (follow the *mm* link in Fig. 1.2). The Model Maker displays putative exons from mRNAs, ESTs and gene predictions that align with the genome. Users can select individual exons from these alignments and build a customized gene model. As the Model Maker displays the nucleotide sequence of the model along with its three-frame translation, the effects of adding, modifying or deleting exons can be quickly evaluated.

(BE798864) into the box marked *Enter accession or GI* for Sequence 1. Obtaining the accession number of *RAB9P40* requires going back to the graphic shown in Fig. 5.6 and clicking on the gene's track. Once this has been done, input the gene's accession number (NM_005833) into the box marked *Enter accession or GI* for Sequence 2. Make sure that the *Program* pull-down is set to *blastn* (to compare a nucleotide sequence against another nucleotide sequence, hence the *n* in *blastn*) and click the *Align* button at the bottom of the page to generate the alignment (Fig. 5.10). The sequence corresponding to sequence 1 (the EST) is denoted as the query, whereas the sequence corresponding to sequence 2 (the known gene) is denoted as the subject. The known gene's protein translation is also shown, starting at the end of the third row of the alignment. Examination of the alignment shows that the EST is missing 153 nt (nt 360–512 of the

mRNA), which corresponds to the fifth exon that is missing in BE798864. This gap is in frame, so the EST could encode a homologous yet shorter protein.

Because of the nature of EST sequencing, ESTs often contain sequencing errors at a rate much higher than those of the finished or even draft genomic sequence. It is certainly encouraging that EST BE798864 aligns well with the genomic sequence and that its encoded protein could be in the same frame as that produced from the known gene. In addition, it appears from the UCSC graphic (Fig. 5.7) that other ESTs in this region, such as BE779110, are also missing the fifth exon of *RAB9P40*. All these predictions must, however, be tested computationally by looking at the quality of the EST–genomic alignment as shown above. Final proof of alternative splicing can, of course, only be generated at the laboratory bench.

Figure 5.1

Figure 5.2

Figure 5.3

Figure 5.4

Figure 5.9

Figure 5.10