

---

## CHAPTER 4

---

# Finding, Delineating and Analysing Genes

CHRISTOPHER SOUTHAN

*Oxford GlycoSciences UK Ltd*  
*The Forum, 86 Milton Science Park*  
*Abingdon OX14 4RY, UK*

---

- 4.1 Introduction
  - 4.2 The evidence cascade for gene products
  - 4.3 Shortcomings of the standard gene model
  - 4.4 Locating known genes on the Golden Path
    - 4.4.1 Raw sequence data
    - 4.4.2 Primary accession numbers
    - 4.4.3 Secondary accession numbers
    - 4.4.4 Gene names
    - 4.4.5 Genome coordinates
  - 4.5 Gene portal inspection
  - 4.6 Locating genes which are not present in the Golden Path
  - 4.7 Analysing a novel gene
  - 4.8 Comprehensive database searching
  - 4.9 Conclusions and prospects
- References
- 

### 4.1 INTRODUCTION

This chapter will describe ways to interrogate human genome (HG) data with the results of genetic experiments in order to locate known genes on the current Golden Path (GP) chromosomal assemblies. It will also describe the assessment of evidence for genes that do not yet have experimental support and some analytical choices that may reveal more about them. In addition to some general aspects of gene detection some specific examples will be worked through in some detail. This illustrates technical subtleties that are not easy to capture at the overview level. As an introduction to the HG, GP and gene annotation the following chapter by Semple is recommended. Chapter 2 also provides some useful background on the organization of sequence databases. A caveat needs to be added here that many roads lead to Rome. Some particular ways of hacking through the genome jungle

are implicitly recommended by being used as the examples in this chapter. They will also be restricted to public databases and web tools. These are the personal choices of the author based on an assessment of their availability and utility. Other experts may propose alternative routes to the same information, either using different public resources, locally downloaded datasets, Unix-based tools, commercial software or subscription databases.

Genetic investigations are concerned with discerning the complex relationships between genotype and phenotype. The statement that phenotype is determined by the biochemical consequences of gene expression is equally obvious. However, the reason for making this explicit is to recommend that those performing and interpreting genetic experiments may find it more useful to conceptualize the gene as a cascade of evidence that connects DNA to a protein product rather than abstract ideas about what might constitute a gene locus. The idea of focusing on gene products also makes it easier to design experiments to verify predicted transcripts and proteins. It must also be remembered that many gene products are non-message RNA molecules but they will not be covered in this chapter. Before describing the evidence used to classify gene products it is necessary to define some of the terminology encountered in the literature and database descriptions. These are variously classified as known, unknown, hypothetical, model, predicted, virtual or novel. There are no widely accepted definitions of these terms but their usage in this chapter will be as follows. A known gene product is experimentally supported and would be expected to give close to a 100% identity match to a unique GP location. The term 'unknown' is typically applied to gene products that are supported experimentally but that lack any detectable homology or experimentally determined function. The term 'predicted', also referred to as 'model' or 'hypothetical' by the NCBI, will be reserved for an mRNA or protein ORF predicted from genomic DNA. Virtual mRNAs will refer to constructs assembled from overlapping ESTs that exceed the length of any single component. The term 'novel' has diminishing utility and will simply refer to a protein with no extended identity hits in the major protein databases.

## 4.2 THE EVIDENCE CASCADE FOR GENE PRODUCTS

So what kinds of evidence need to be considered before we assess the likelihood of a stretch of genomic DNA giving rise to a gene product and what kind of numbers can be assigned to these evidence levels? The most solid evidence of a gene is the experimental verification of the protein product by mass spectrometry and/or Edman sequencing. Although these techniques are commonly used to analyse proteins produced by heterologous expression *in-vitro* surprisingly few genes from *in-vivo* or cell line sources have been verified at this level. From the entire SP/TR collection of human proteins only 311 are cross-referenced as having at least a fragment of their primary structure identified directly from a 2D-PAGE experiment (<http://ca.expasy.org/ch2d/>) (Hoogland *et al.*, 2000). Numerous mass spectrometry-based identifications and peptide sequences from human proteins are reported in the literature but little of this data has been formally submitted to the public databases and therefore has not been captured by SwissProt or other secondary databases. However, even this most direct of gene product verifications is rarely sufficient to confirm the entire open reading frame (ORF). For example secreted proteins are characterized by the removal of signal peptides and frequent C-terminal processing. This precludes defining the N and C translation termini by protein chemical means.

The next level down in the evidence cascade is of course an extended mRNA. There are currently 48,681 human mRNAs in GenBank. However transcript coverage is by no means

complete as they collapse down by shared identity to a set of 13,429 human transcripts (excluding splice variants) in the NCBI RefSeq collection (<http://www.ncbi.nlm.nih.gov/LocusLink/RSstatistics.html>) (Pruitt and Maglott, 2001). Although this collection attempts to provide a non-redundant snapshot of gene transcription it must be remembered that they are not all full-length transcripts. If the databases do not contain an extended mRNA the assembly of overlapping and/or clone-end clustered ESTs can be considered as a virtual mRNA (Schuler, 1997). The ESTs have the additional utility that many of them can be ordered as clones. Alternatively, the virtual consensus sequence, backed up by comparisons to the genomic DNA, can be used for PCR cloning. The fact that 94% of known mRNAs are covered by at least one EST makes them strong supporting evidence for a transcript, especially if they include a plausible splice junction and are derived from multiple clones from different tissue cDNA libraries (<http://www.ncbi.nlm.nih.gov/UniGene/>). The TIGR gene indexes are a useful source of pre-assembled virtual sequences that they term tentative human consensus sequences or THCs (Quackenbush *et al.*, 2001). These can also be selected in the UCSC genome display. The use of unspliced ESTs as evidence for a transcribed gene is unreliable as they can arise from genomic contamination. However human EST-to-genome matches for exon detection can be further supported where orthologous ESTs from other vertebrates, such as mouse or rat, match uniquely in the same section of GP. If an assembly of mouse ESTs is consistent with a human gene model then the existence of an orthologous human transcript is strongly implicated.

The protein databases occupy the centre of the evidence cascade for gene products. Those mRNAs that translate to an open reading frame (ORF) are experimentally supported even if they are not full-length and/or there can be ambiguity about the choice of potential initiating methionines. However, the fact that the protein databases have now expanded to include human ORFs derived solely from genomic predictions (described in the next section) means that the evidence supporting them as gene products becomes circular. The highest curation level is provided by SwissProt sequences from the Human Proteomics Initiative set (HPI) ([http://ca.expasy.org/sprot/hpi/hpi\\_stat.html](http://ca.expasy.org/sprot/hpi/hpi_stat.html)). The March 2002 number comprised 7895 unique gene products and 2039 splice variants (O'Donovan *et al.*, 2001). The SwissProt/TrEMBL (SP/TR) total for human proteins in February 2002 was 24,147, including splice variants (<http://www.ebi.ac.uk/protome/HUMAN/interpro/stat.html>). The current Ensembl release, 4.28.1, contains 21,619 proteins classified as 'knowns' by an identity above 95% to a human SP/TR entry (Hubbard *et al.*, 2002). The International Protein Index (IPI) maintains a database of cross references between the data sources SwissProt, TrEMBL, RefSeq and Ensembl. This provides a minimally redundant yet maximally complete set of human proteins with one sequence per transcript (<http://www.ebi.ac.uk/IPI/IPIhelp.html>). The March 2002 release contains 65,082 protein sequences but this includes 28,350 XP RefSeq ORFs predicted by the NCBI which are not supported by mRNAs.

The next level of evidence can be classified as genomic prediction i.e. where a cDNA, a translated ORF and a plausible gene splice pattern can be predicted from a stretch of genomic DNA (Burge and Karlin, 1997). This proceeds more effectively on finished sequence or at least where unfinished sequence contains the exons in the correct order. This is done after filtration of repeats which can be considered as another link in the evidence chain. A very high local repeat density certainly suggests where exons are unlikely but the converse is not true i.e. the absence of repeats does not prove the presence of genes. The shortcomings of *ab initio* gene prediction have been pointed out but the geneticist should at least be aware of possible false positives and false negatives (Guigo *et al.*, 2000). The Ensembl statistics of the ratio of genes

predicted by Genscan over genes with a high evidence-supported threshold is currently 7.5:1 ([http://www.ensembl.org/Homo\\_sapiens/stats/](http://www.ensembl.org/Homo_sapiens/stats/)). Although this clearly represents over-prediction some may be 'genes-in-waiting' which more accumulated evidence may verify, for example by the cloning of an extended mRNA. Looking for a consensus or at least common exons from a number of gene prediction programs with different underlying gene model assumptions can strengthen this type of evidence but this can become a circular argument where the programs are both trained and benchmarked with known genes. For unfinished genomic sequence the presence of gaps and local miss-ordering of contigs within the clone degrades the performance of *ab initio* methods. The most effective way of filtering down genomic predictions without experimental evidence is homology support i.e. the predicted protein shows extended similarity with other proteins. This is described in detail in the Ensembl documentation but in essence all possible protein similarity sections from translated DNA are identified and used to build homology-supported gene predictions using GeneWise (Birney and Durbin, 2000). The advantage of gene detection by homology is that the entirety of protein sequence space can be used. The caveat is that predicted gene products with low similarity to extant proteins would be discarded in this filter, although the entire set of Genscan predictions are preserved for searching in Ensembl and can also be displayed at UCSC.

The next link in the evidence chain is a special case of the similarity principle but in this case utilizing comparisons between the genomes of other vertebrates such as mouse and fish for which extended data are now available (Wiehe *et al.*, 2001). Mouse genome assemblies have recently appeared on the Ensembl and UCSC sites. Although the initial assembly is only ~20% the total depth in the trace archives and HTGS divisions is approaching complete coverage. Cross-species data can be assessed at three levels. The first is a simple DNA similarity on pieces of mouse DNA known to be syntenic from the location of known mouse genes and/or the extended similarity score which, with appropriate masking, locates it uniquely to a human locus. This approach is termed phylogenetic footprinting (Susens and Borgmeyer, 2001). The problem for gene product detection is that this is too sensitive i.e. mouse/human syntenic regions have many conserved similarity 'patches' outside the boundaries of known exons. They are likely to be important for functions not yet understood but are difficult to discriminate from potential coding regions. The second level is mouse BLAT as used on the UCSC site. This goes a step back by doing a translation similarity comparison rather than direct DNA-to-DNA. This makes it more likely to pick up reading frame similarities across exons. The third level is the so-called exofish. By the detection of translation similarities at the amino acid level this is capable of detecting those exons that are conserved between human and fish. This will be more useful when exofish updates to a complete fish genome rather than a partially assembled one.

The last link in the evidence chain, the *in silico* recognition of transcriptional control regions, is circumstantial but is likely to increase in utility (Kel-Margoulis *et al.*, 2002). These could include potential start sites in proximity to CpG islands, promoter elements, transcription factor binding sites, and potential polyadenylation acceptor sites in 3' UTR. When considered in isolation these signals have poor specificity but taken in combination with a consensus gene prediction and conservation of these putative control regions between human and mouse, they can become a useful part of the evidence chain.

In summary there is currently direct experimental evidence for ~15,000 genes and strong evidence to support a lower gene limit of around 25,000. The confirmation rates for the types of evidence listed above has not been calibrated experimentally so we cannot come up with any kind of scoring function to rank gene likelihood. Going to the

extremities of the evidence cascade, for example with the 65,082 ORFs from the IPI or the 62,271 UniGene clusters containing at least two ESTs, would result in a higher upper limit. This uncertainty becomes a key issue for genetic experiments. Let us suppose, for example, that a linkage study has defined a trait within the genomic region bounded by two microsatellite markers. If the lower limit gene number is true then the investigator merely needs to check the annotations from any of the three gene portals to produce a list of gene products between the positioned markers from which to choose candidates for further work. If the upper limit is true this approach has a major limitation because many of the genes between the markers will not be annotated. However, the different levels of gene evidence described above can be visualized in the display tracks of the genome viewers. Consideration of the evidence will enable the geneticist to decide what experiments need to be designed to confirm potential novel gene products. An example of working through this evidence is given in the examples below.

### 4.3 SHORTCOMINGS OF THE STANDARD GENE MODEL

One of the conclusions that could be drawn from the draft human genome sequence was that the standard gene model of a defined gene locus  $\rightarrow$  a single mRNA species  $\rightarrow$  a single protein, is no longer adequate to describe the increasingly complex relationship between the genome and its products. Attempts to fit transcript data into the standard gene model highlight a number of 'grey' areas. The first of these is delineating the extreme 5' and 3' ends of the mRNA transcripts (Pesole *et al.*, 2002; Suzuki *et al.*, 2002). The fact that many mRNAs are labelled as partial is testimony to the difficulty of finding library inserts that are complete at the 5' end. In many cases the mRNAs are considered finished when a plausible ORF has been delineated. However, very few cDNAs are full-length in that they have been 'walked out' to determine the true 5'-most initiation of transcription in the 5' UTR. The same problem applies to the UTR at the 3' end. There may be substantial stretches of 3' UTR extending downstream of the first polyadenylation position at which further cloning attempts have ceased. The problem is compounded by the poor performance of gene prediction programs for 5' and 3' ends. The first step towards resolving uncertainties about transcript extremities, is to survey the coverage of all available cDNA sequences, whether nominally full-length or partial, ESTs and patent sequences. These can often extend the UTR sections. The second grey area concerns pseudogenes. In some cases genomic sequence is so severely degraded that transcription is unlikely. However, from the current pseudogene listing in RefSeq of 1598 loci, at least 30 are recorded as having detectable transcripts (<http://www.ncbi.nlm.nih.gov/LocusLink/statistics.html>). The third grey area is gene product heterogeneity. In some cases there may be alternative upstream initiation methionines or alternatively spliced exons in the 5' UTR. The causes for 3' heterogeneity include variations in the pattern of intron splicing from a pre-mRNA, as well as alternative poladenylation positions inside the 3 UTR. The fourth grey area concerns overlapping genes. As genomic annotation proceeds we can find more examples of this both from gene products reading from opposite strands and same-strand genes in close proximity.

Considering these grey areas as a whole, they can all be seen as deviations from the simple gene model. Many individual examples of such complexities had been documented before the genome draft of May 2001. However, it is only since then that assessments of their overall incidence could be made, most recently for completed chromosomes such as 20 (Deloukas *et al.*, 2001). It is therefore essential for the geneticist to keep an open mind about the extremities and plurality of gene products.

### 4.4 LOCATING KNOWN GENES ON THE GOLDEN PATH

Genes can be located by one of the following: a section of raw sequence data, a primary accession number, a secondary accession number, a similarity search, a gene product name, or a set of Golden Path (GP) coordinates. Each of these has advantages and disadvantages and, although the three gene portals are generally consistent, they may not give the same answers in every case. Bearing in mind that only the first two of these are stable and (almost) free of potential ambiguity it is better to use at least two ways to define and store the results, for example a section of raw sequence and a gene name, or a primary accession number and a set of GP coordinates. The BACE gene will be used as an example of a known gene to locate. The potential complexity of this task is illustrated by the example of the Ensembl gene report for BACE that includes no less than 46 separate terms (Figure 4.1).

#### 4.4.1 Raw Sequence Data

The availability of GP means that most features can now be unambiguously located in the genome with as little as 100 bp. This means that storing a sequence string, preferably with a longer sequence context of 200–1000 bp, is a useful method of locking-on to a genomic location. It is also immune to the vagaries of shifting secondary accession numbers, naming ambiguities or GP sequence finishing that can change the genomic coordinates. Performing nucleotide searches against GP using tools such as BLAT (UCSC) or SAHA (Ensembl) or BLAST (NCBI), means that sequence matches can be quickly located. The disadvantage for raw sequence is that it has to be stored in its entirety, it may contain errors, it needs the operation of a similarity search to be located and similarity matches across repeat containing sections or duplicated regions of the genome need close inspection to sort out. This can be a particular problem for STSs and SNPs

<b>Ensembl gene ID</b>	ENSG00000160610
<b>Genomic Location</b>	<b>View gene in genomic location:</b> <a href="#">22599867 - 120575716 on chr12 (2.6 Mb)</a> on chromosome 11 <b>This gene is located in sequence:</b> <a href="#">AF001672.4.1.136278</a>
<b>Description</b>	BETA-SECRETASE PRECURSOR (EC 3.4.23.5) (BETA-SITE APPOLYMERASING ENZYME)(BETA-SITE AMYLOID PRECURSOR PROTEIN-CLEAVING ENZYME) (ASPARTY-PROTEASE 2) (ASP 2) (A2P2) (MEMBRANE ASSOCIATED ASPARTIC PROTEASE 2) (MEMAPSN2); [Source:SWISSPROT;Acc:P9687]
<b>Prediction Method</b>	This gene was predicted by the Ensembl analysis pipeline from either a GeneWise or GenScan prediction followed by confirmation of the exons by comparisons to protein, cDNA and EST databases
<b>Predicted Transcripts</b>	1: <a href="#">ENST00000292695</a> [ <a href="#">View supporting evidence</a> ] [ <a href="#">View protein information</a> ]
<b>Links</b>	<b>This Ensembl gene corresponds to the following other database identifiers</b> <b>EMBL:</b> <a href="#">AF001672</a> [gen] <a href="#">AF024458</a> [masc] <a href="#">AF130075</a> [gen] <a href="#">AF330192</a> [gen] <a href="#">AF304563</a> [masc] <a href="#">AF304564</a> [gen] <a href="#">AF3888</a> [E gen] <b>GO:</b> <a href="#">GO:001494</a> <a href="#">GO:005522</a> <a href="#">GO:006847</a> <a href="#">GO:006858</a> <a href="#">GO:004398</a> <a href="#">GO:004346</a> GO mapping is inherited from <a href="#">outpost/psprembi</a> <a href="#">Search GeneCards for BACE</a> <b>HUGO:</b> <a href="#">LucusLink:</a> <a href="#">23521</a> [tab] <b>IMR:</b> <a href="#">R10957</a> <b>RefSeq:</b> <a href="#">NM_012104</a> [Transcript] <a href="#">NM_005180</a> [Gene] <a href="#">NM_012104</a> [chr] <b>SWISSPROT:</b> <a href="#">P9687</a> [Transcript] <a href="#">P9687</a> [Gene] <a href="#">P9687</a> [chr] [ <a href="#">Sequence</a> ] <b>SpTREMBL:</b> <a href="#">Q9BY98</a> [Transcript] <a href="#">Q9BY98</a> [Gene] <a href="#">Q9BY98</a> [chr] [ <a href="#">Sequence</a> ] <b>QDUL2:</b> [Transcript] <a href="#">QDUL2</a> [Gene] <a href="#">QDUL2</a> [chr] [ <a href="#">Sequence</a> ] <b>protein_id:</b> <a href="#">AAPI1142</a> [masc] <a href="#">AAPI377E</a> [masc] <a href="#">AAPI1079</a> [gen] <a href="#">AAPI3867</a> [gen] <a href="#">AAIC0247</a> [masc] <a href="#">AAIC0248</a> [gen] <a href="#">AAI0000C</a> [gen]
<b>InterPro</b>	<a href="#">IPR01461</a> Pepsin (A) aspartic peptidase [ <a href="#">View other Ensembl genes with this domain</a> ] <a href="#">IPR001583</a> Eukaryotic and viral aspartic peptidase active site [ <a href="#">View other Ensembl genes with this domain</a> ]
<b>Protein Family</b>	<a href="#">ELSG000000173</a> BETA-SECRETASE PRECURSOR EC 3.4.23.5; BET This cluster contains 2 Ensembl gene members
<b>Export Data</b>	<a href="#">Export gene data in EMBL, GenBank or FASTA</a>

Figure 4.1 The Ensembl gene report page for BACE (release 4.28.1).

if the GP match is in the region of 98 to 95% identity. Within this range it is difficult to discriminate technical sequencing errors from multiple genomic locations or assembly duplication errors. It can also be useful to search the primary genomic data, especially if GP is not complete in that section. For example although BACE is linked by Ensembl to AP001822 as the finished GP sequence, a database search reveals another four matching primary genomic accession numbers from chromosome 11, AP000892 (finished at version 4) with AC020997, AP000685 and AP000761 still unfinished. One less obvious advantage of these five overlapping genomic contigs is that if they proceed to finishing more SNP positions may be revealed. As described below the genome portals capture mRNA entries for most gene products unless they are very recent. However, because of the thin annotation they do not capture sequences from the patent divisions. A BLAST search of gbPAT with any BACE mRNA gives 18 high-identity DNA matches. These are clearly mRNAs that could be usefully compared with all other mRNA sequences for polymorphisms, splice variants or UTR differences. However users should be aware that not only are some of these 18 entries identical versions of the same sequence derived from multiple claims in the patent documents but they may also be identical to a public accession number if the authors and inventors are from the same institution. Another reason for using raw sequence data for gene product checking is because all secondary databases suffer from the snapshot effect where updates lag behind the content of the primary databases. For example the SNP or EST assignments made for BACE in the secondary databases (see below) could be checked by BLAST searches against the updates of dbSNP or dbEST (remember the latest EST data needs to be searched in 'month' as well as dbEST).

#### 4.4.2 Primary Accession Numbers

Because these uniquely define stretches of sequence they are stable except where genomic and occasionally mRNAs, undergo version changes. They can be used in any of the major genome query portals to go directly to a genomic location. The disadvantage is redundancy for mRNAs, short sequence context for some STSs, both redundancy and large multi-gene sequence tracts for genomic mRNA, and very recent accessions may not be indexed in genome builds. If the query fails to connect to a genome feature the sequences can be searched as raw sequence. Taking the BACE example there are eight mRNA accession numbers listed in Figure 4.1 that can be used as a genome portal query. Interrogating UCSC with BACE retrieves nine mRNA entries, LocusLink connects directly to only three but the UniGene cluster Hs.49349 connects to 12. Users need to be aware that although an mRNA accession number can provide a specific route into GP the variable number of links to the genome portals is related to their update frequency.

#### 4.4.3 Secondary Accession Numbers

From Figure 4.1 we can read eight secondary accession numbers that designate protein translations for each of the BACE mRNAs. It also has three RefSeq numbers NM\_012104 for the mRNA, NP\_036236 for the protein and NT\_009151 for the genomic contig. There is one SwissProt accession BACE\_HUMAN (P56817) and one TREMBL splice variant Q9BYB9. The LocusID, 23621, in turn links out to many other accession numbers which point to the BACE genome sequence. These include the Hs.49349 UniGene cluster that includes 336 ESTs with primary accession numbers. Via the LocusLink Variation link

the RefSNP numbers can be located. In this case they consist of 43 intronic SNPs, three within the mRNA, including one (rs539765) which causes an Arg > Cys exchange, and seven SNPs in the 3' UTR. It is possible to use a RefSNP (rs) number to go directly to the SNP location in Ensembl or UCSC. However because of multiple GP matches in Ensembl it is necessary to know the genomic location beforehand.

#### 4.4.4 Gene Names

Including abbreviations Figure 4.1 there are nine synonyms or aliases for this enzyme. This illustrates the problem where gene products are given different names by different authors. The best way to cross-check names, spelling variations and frequency of use, is to search PubMed. Checking title lines only is more specific but does not capture all occurrences. In this case a title search found a new name extension, BACE1, with five citations compared with 22 for BACE. This seems logical since the discovery of the BACE2 paralogue on chromosome 21. However, the Human Gene Nomenclature Committee have not been consistent because they have only listed BACE and BACE2 as official symbols even though they have listed ACE1 as an alias for ACE since the recent discovery of ACE2 (<http://www.gene.ucl.ac.uk/nomenclature/>). The most frequent specific term was 'beta-secretase precursor' at 30 citations. The alternative 'membrane-associated aspartic protease 2' gave eight citations and 'beta-site app cleaving enzyme' was the least frequent with only two. Paradoxically this has been chosen for the LocusLink name. The least specific name was aspartylprotease 2 with two false positives and ASP 2 with 143 title matches, also mostly false positives. The imprecision of name searching was reinforced by checking ASP-2 with three matches and ASP2 with five. Only one was a true positive and two of the citations referred to ASP2 as an odorant-binding protein from the honeybee. The complexity of the aliases for just one gene product makes it clear that any gene name lists, for example as candidate genes to be screened for mutations, must be backed-up by accession numbers and/or raw sequence. It also illustrates the need to cross-check aliases and their spellings when attempting a comprehensive literature search on a particular gene product. The formal sequence-literature links that can be followed in Entrez, LocusLink or SwissProt are not comprehensive because they are dependent on the journal-author-database system that usually only makes these links explicit for a new accession number. Much important literature remains outside this system. Review articles, for example, do not typically include primary accession numbers when describing genes so the specificity of literature searches remains dependent on the name links. Information trawling with gene names can also be done with the standard internet search portal. Putting the term 'beta-site app cleaving enzyme' into the Google search engine gave 249 hits (<http://www.google.com/>). The listing included duplicates but very few false positives.

#### 4.4.5 Genome Coordinates

Since the adoption of a unified GP assembly this method of genomic location has become more reliable but users are advised to check the synchronization of new GP versions between the three portals. Users should refer to the individual portals for the details of using these coordinates but for the BACE example the NCBI showed a region described as 120,533K–120,594K, the Ensembl viewer specified the coordinates as 120549397–120575715 bp (with a zoom setting 120.5 Mb) on chromosome 11 and the UCSC viewer designated the position in the form chr11 : 120545299–120599798.

### 4.5 GENE PORTAL INSPECTION

From the descriptions above it should be possible to locate any known gene or genetic marker such as an STS or a SNP. Descriptions of the genome viewer features for Ensembl, UCSC and NCBI are included in the chapter by Semple. However two examples are included below (Figures 4.2 and 4.3) because they illustrate technical differences and highlight the deviations from the standard gene model. The UCSC display (Figure 4.2) includes 12 mRNA sequences for BACE where Ensembl (Figure 4.1) has included accession number links for only eight. The display in Figure 4.2 also shows there are significant differences in the lengths of the 5' and 3' ends. Clearly AF201468 (5878 bp) and AB032975 (5814 bp) are the longest reads but in fact AB032975 is labelled as a partial CDS because of what may be a sequencing error at the 5' end. The matches to the spliced ESTs together with the rat and mouse mRNAs suggest the 5' UTR may be full-length for these entries i.e. they extend to the start of transcription. This is in contrast to the shorter 5' ends for the majority of mRNAs. A detailed analysis of the 3' ends by EST distribution profiles indicates that the different UTR lengths in this case arise not from incomplete cloning but from three alternative polyadenylation positions (Southan, 2001). Further heterogeneity is illustrated by three splice variants affecting exons 3 and 4. The representative mRNAs are AB050436, AB050437 and AB050438. There is also an alternative protein reading frame from AF161367, a partial mRNA cloned from CD34+ stem cells. Opening up the spliced EST tracks in the viewer shows individual ESTs corresponding to these splice forms. Approximately midway between exons 1 and 2 (from the 5' end) is a spliced EST, AL544727, derived from spleen. This suggests the possibility of another splice form but this would need analysis for canonical splice sites and experimental verification. Similarly an EST from spinal cord AL589586 suggests an alternative exon just on the 5' side of exon 3. Although the rat and mouse mRNAs displayed in Figure 4.2 show the same exon positions as most human sequences there are suggestions of splice variants in non-human ESTs but these tracks were not expandable in the version tested.

The NCBI display for BACE mRNAs and ESTs (Figure 4.3) shows concordance and discrepancies with the UCSC display (Figure 4.2). The exon positions are identical. They include the same RefSeq mRNA and genomic secondary accession numbers. The EST matches are in broad agreement towards the 3' end but two additional potential exon matches are indicated at the 5' end. Although these may be unspliced matches that would need further investigation, one of these coincides with the XM\_084660 reference sequence

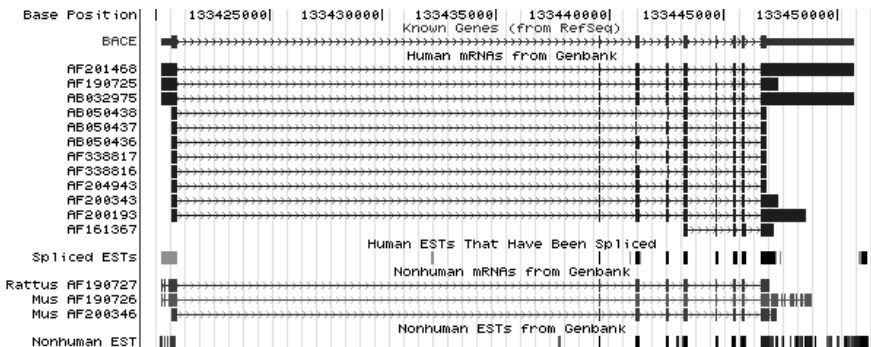
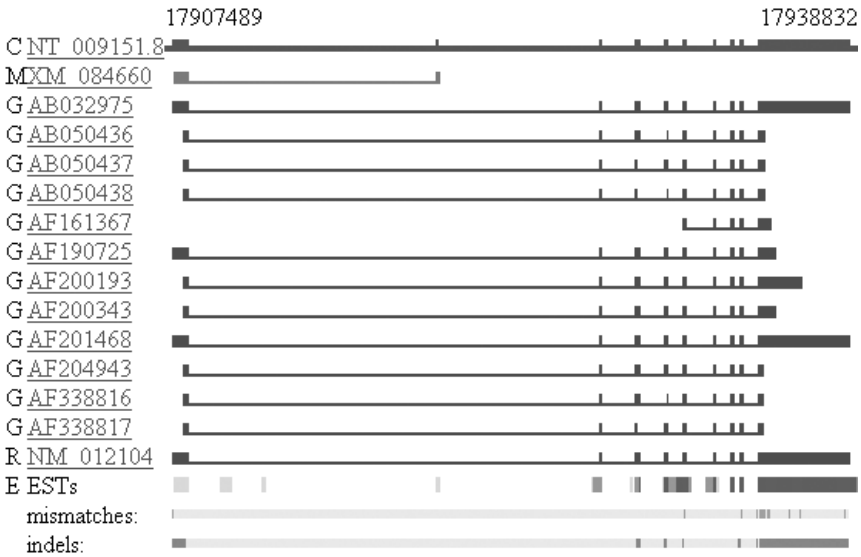


Figure 4.2 The UCSC display for BACE mRNAs and ESTs.



**Figure 4.3** The NCBI display for BACE mRNAs and ESTs.

predicted by NCBI from the contig NT\_009151. There is no mRNA verification for this prediction so it will be of interest to see if additional EST data will appear and, if not, how long this prediction will be maintained as genome annotation. The mismatches and INDEL tracks are a useful feature unique to NCBI. The mismatches within the set of 12 mRNA sequences could represent SNPs or technical sequence errors. The INDELS also show major length discrepancies. In Figure 4.2 these highlight the three splice positions in agreement with UCSC but the INDEL in exon 8 could not be interpreted from the link provided.

#### 4.6 LOCATING GENES WHICH ARE NOT PRESENT IN THE GOLDEN PATH

Estimates suggest the GP is still missing ~2.5% of the genome, there are still small gaps in the unfinished sections and the latest Ensembl release locates only 92% of known proteins (<http://www.ensembl.org/Dev/Lists/announce/msg00070.html>). This means that some genetic markers in close proximity to genes are either not covered by GP or are not fully annotated in unfinished sequence. Two human proteins that have no matches on the current GP version 28 from December 2001 illustrate this problem. The first of these, spP83110 serine protease HTRA3, has an mRNA entry AY040094. The second protein spP83105 serine protease HTRA4 has an mRNA accession but the entire ORF is covered by two long EST reads AL545759 and AL576444. Because it has a full length mRNA HTRA3 has a LocusLink ID of 94031 but no mapping links. Searching HTRA3 by BLASTN against the NCBI nr nucleotide database, containing 1,184,532 sequences, hits only the probable mouse orthologous mRNA, AY037300, at 86% identity within the reading frame. However checking monthly updates at 811,100 sequences reveals a 99% identity to a new genomic entry AC113611 of 190,038 bp from chromosome 4. This sequence was also in the unfinished High Throughput Genomic Sequences (HTGS)

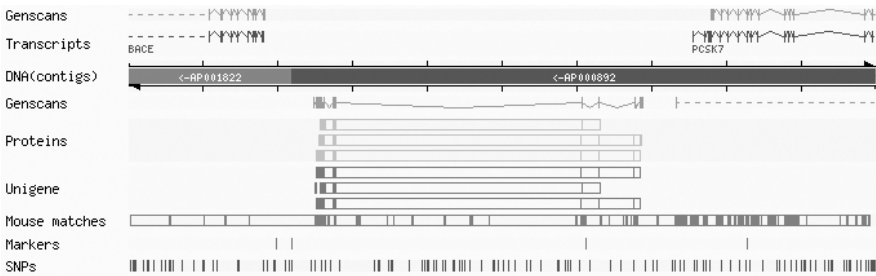
division, with 47,855 sequences, along with the probable rat orthologous genomic section, AC110369, at 87% identity. There were no mouse genome matches from this search. A check on the nucleotide patent databases, with 582,838 sequences, showed a new mRNA match, AX338509 from patent WO0183775. The HTRA3 mRNA has EST matches to UniGene cluster Hs.60440 with four STSs from chromosome 4. Presumably these STSs will be located on GP when the AC113611 genomic sequence is assembled into chromosome 4. Checking the chromosome 4 SNPs at 105,568 sequences by BLAST search, recorded no hits within the 2552 bp mRNA of AY040094 but found over 100 matches within the repeat-masked sections of AC113611. Using the same sequence to BLAST against the 115,608 sequences in the STS division gives eight hits above 95% identity, although only three looked like unique matches. Interestingly the HTRA3 mRNA AY040094 has no STS matches although four chromosome 4 STSs were picked up in the UniGene entry. A possible explanation is that the cluster included clone links to ESTs that extend past the 3' end of the mRNA.

Performing the same database checks for the HTRA4 ESTs, AL545759 and AL576444, produces a different pattern of findings. There were no hits in nr or gbPAT. However, the HTGS search located extended identity hits to no less than four genomic entries. These comprised of three recently sequenced sections of chromosome 8 AC108863, AC105089, AC105088, and a short match to an entry without a chromosomal assignment, AC107926. Checking for HTRA4 in LocusLink could find no IDs because of the absence of a full-length mRNA. It was picked up as the UniGene cluster Hs.322452 with nine ESTs but no mapping information was included even though our search update had located it to chromosome 8. No reading frame SNPs could be detected from the 92,110 chromosome 8 entries. By using the genomic contig, AC108863, (198,743 bp) as a BLAST query only three SNP identity matches were detected, rs1467190, rs2010445 and rs2056170, but three STS markers G60989, G23343, and G04735, were located.

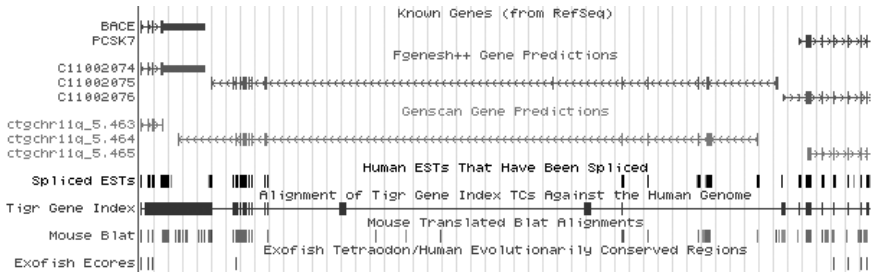
In summary; although these two gene products cannot be located on the latest GP a series of manual database checks have established a mixture of patent mRNAs, unfinished genomic matches, ESTs, STSs and SNPs. It will be interesting to track how soon these features find their way into the GP annotation pipelines. If genetic studies should need this location data in the interim, the searches have established that HTRA3 probably has enough SNPs in the genomic vicinity for association studies, but that there is a very low SNP density in proximity to HTRA4. If the overlapping genomic coverage for HTRA4 could detect all the exons it might be possible to assemble a 'mini golden path' across this particular section. However if it became necessary to re-order and re-assemble the contigs within the unfinished entries this would be a challenging task to perform with web-based tools.

## 4.7 ANALYSING A NOVEL GENE

Sooner or later experimental results will locate a piece of GP where there are no fully annotated known genes. Figures 4.4, 4.5 and 4.6 show selected tracks from the Ensembl, UCSC and NCBI displays between the 3' side of the BACE gene and the 5' end of the next known gene PCSK7. The known genes are marked in brown in Ensembl and blue in UCSC. The latter are mRNA mappings and therefore include the UTR sections. Let us assume a genetic linkage study had found significant associations in this area, either from the two STS markers or the 50 or so SNPs that lie in this interval but are outside the boundaries of the two neighbouring genes. The question immediately arises as to what



**Figure 4.4** The Ensembl display for the unknown gene between BACE (left) and PCSK7 (right).

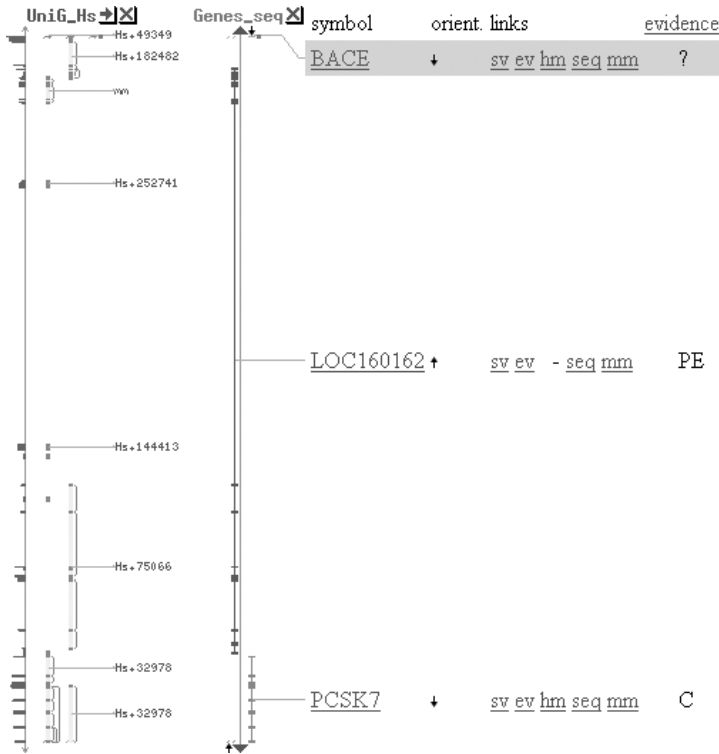


**Figure 4.5** The UCSC display for the unknown gene between BACE (left) and PCSK7 (right).

other gene product(s) might be located between the two knowns. The first step is to check the continuity of this section of GP. This can be done in any of the viewers and in this case there is complete clone overlap across this section.

Inspection of all three displays indicates a possible novel gene product with a variety of supporting evidence. They include gene predictions which include both common and different exon positions. The UCSC Genscan prediction number 464 overlaps with the 3' UTR of BACE making this a less plausible (but still possible) exon. Reading vertically down the Ensembl tracks first we see evidence for three protein homologies (yellow) as judged by the matches in register with the Genscan exon predictions. These are Q96RS9, a novel DZIP3, Q02455 a myosin-like peptide from yeast and P53804 a tetratricopeptide repeat protein. There is the same pattern of exon matches to three UniGene cluster entries (red) Mm.3679 *Mus musculus* for the tetratricopeptide repeat domain protein, Hs.165662 for *Homo sapiens* KIAA0675 unknown protein and Hs.118174 for *Homo sapiens* TTC3 tetratricopeptide repeat domain 3. There is a denser pattern of matches to mouse DNA (pink) that includes many sections outside the Genscan predicted exons.

Moving down the UCSC tracks in Figure 4.5 we see the spliced ESTs (black) in register with Genscan exons. However these identity EST matches are not equivalent to the homology-based UniGene matches in Ensembl. Interestingly the internal exon predicted only by Fgenes has no spliced EST support. Exploring the EST coverage further we see that the (brown) THC tracks include an assembly that matches the predicted exons at the BACE end of the Fgenes++ prediction. The NCBI tracks go into more detail by not only mapping UniGene cluster components directly back to putative genomic exons by



**Figure 4.6** The NCBI display for the unknown gene between BACE (top) and PCSK7 (bottom). The leftmost track shows the EST distribution. The next track to the right marks the UniGene clusters. The central track is the gene prediction for LOC160162 and the gene structure for the N-terminal section of PCSK7 (bottom).

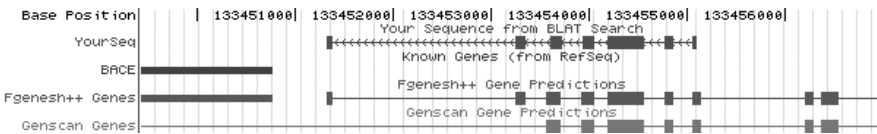
identity matches but also, on the left hand edge, showing an identity block proportional to the number of EST matches. Surprisingly there are five EST clusters which raises the possibility of more than one gene. The mouse BLAT track (brown) is equivalent to the Ensembl (pink) mouse track but the translation mode filters down to fewer features. The exofish track in UCSC (blue) supports just one single exon at the 5' end of the putative novel gene compared with many conserved exons in both gene neighbours. In isolation this would be considered as weak evidence for the gene product. However it could simply mean that this predicted protein is not conserved between fish and human or the puffer fish ORFs are not complete across this section.

Up to this point our analysis of the genomic region between the 3' end of BACE and the 5' end of PKSC7 points strongly to the presence of a gene product on the basis of gene prediction and EST coverage. So where do we go from here? One option is to do some searches with the available mRNA and protein sequence from the Fgenesh++ prediction (numbered C11002075 in Figure 4.5) that can be downloaded from the UCSC site. The result brings us a long way forward in the evidence cascade because we record an 81% protein identity to what is likely to be the recently deposited mouse orthologue mRNA, BC023073. Interestingly this level of similarity should result in this gene passing

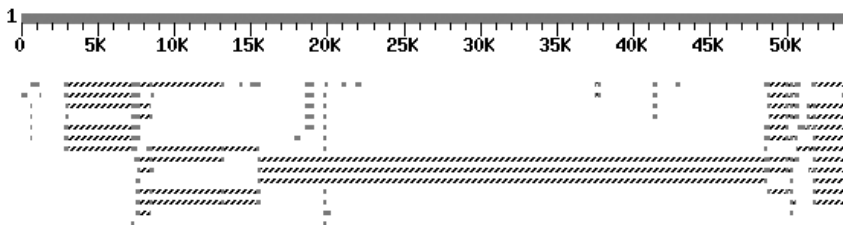
the Genewise threshold for marking a novel gene position (black) in the next release of Ensembl. At these similarity levels we can back-check this mouse sequence against human GP by the very fast BLAT search (<http://genome.ucsc.edu/cgi-bin/hgBlat?command=start>). The result (Figure 4.7) clearly supports both the orientation (3'-to-3' relative to BACE) and seven of the exons from C11002075. However the mouse sequence is clearly missing the 5' end.

The next step involved searching the entire genomic DNA section of 54 kb from which C11002075 was predicted against human ESTs. This was performed using MEGABLAST with a 90% match stringency and masking of the repeat sections in the genomic query section. The result (Figure 4.8) is equivalent in principal to the UniGene clusters in the NCBI viewer but it is easier to pick out the ESTs that bridge several exons. Another reason for doing this analysis is that over 1 million human ESTs have been added to dbEST since the UniGene clusters were built. We can identify three ESTs that cover 35 kb of genomic sequence across three exons and performing the analogous search against mouse ESTs, with an 80% identity cut-off, finds a long EST spanning the four central exons. This gives us more confidence of a single rather than multiple gene products. The next step was to search ESTs against the TIGR THCS to establish if any virtual mRNAs could be found. In fact two of these, THC856832 and THC796698, represented the 5' and 3' ends respectively and to join these assemblies a bridging EST was found, BM055167. By using a web version of the CAP3 assembler (<http://bio.ifom-firc.it/ASSEMBLY/assemble.html>) it was possible to construct an extended virtual mRNA of 2720 bp. This was translated into a protein of 474 amino acids using the translation tool (<http://ca.expasy.org/tools/dna.html>) (Figure 4.9).

So far so good, but what else can we do to verify this putative novel protein *in silico*? The first step is a cross-check for reading frame consistency and species orthologues by performing TBLASTN against all ESTs (Figure 4.10). The results show the complete coverage of the entire ORF by human ESTs but also suggests potential splice variants



**Figure 4.7** The alignment of the mouse protein from BC023073 after a BLAT search against the UCSC GP. The BACE gene is on the left hand side.



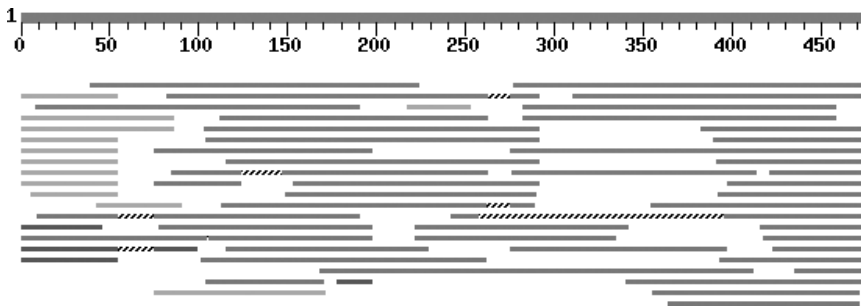
**Figure 4.8** Result of a MEGABLAST search of the genomic sequence between BACE and PCSK7 against human ESTs. The solid lines indicate gaps in the same ESTs. The solid sections are putative exon matches.

```

gcgggctcctgtccctccccactttcctcccggggcgcgggcggggagagcataatggc
agcgtctgaggttgctgggtgttgggccaatgccccagtcctcgggaatctctagttt
atgtgcttccaatcagacgaaggtctccagatggtctaagcaccaaagactctgcaca
gaagcagaagaactcgctctgttgagtgaagttagccaaacaataaccaggagaataa
cagaaatgtccatttggagcactcagagcagaatcctggttcatcagcaggtgacacctc
agcagcgcaccaggtgggttttaggagaaaacttgatagccacagccttgtctttctggc
atggggtctcagctctgatttgaaggatgtggccagcagcaggagaggggggacaca
agcctcggggagcctccatccagtcactcggctctctaaggcaggggtgccatactaa
cagcttgctccaggaattgctctgaagagaaatccccacaacctccatctaaaggaa
ggtaaacggggacacaagcttggatttccgacctgtagtgtctccagcaaatgggggtgaa
ggagtcagagtggtcaggatgatgatcaagatagctcttccctgaagctttctcagaac
attgctgtacagactgacttaagacagctgattcagaggttaaacacagatcaagatatt
gaaaagaatttggataaaaatagacagagagaacctgttgaagagcgttaccaggag
M T E R T L L K E R Y Q E
gtcctggacaaaacagaggcaagtggagaatcagctccaagtgcattaaagcagcttcag
V L D K Q R Q V E N Q L Q V Q L K Q L L Q
caaaaggagagaagaggaaatgaagaatcaccaggagatattaaaggctattcaggatgtg
Q R R E E E M K N H Q E I L K A I Q D V
acaataaagcgggaagaaacaaagaagaagatagagaagaagagaaggatcttctgcag
T I K R E E T K K K I E K E K K E F L T Q
aaggagcaggtctgaaagctgaaattgagaagcttgtgagaagggcagaagagaggtg
K E Q D L K A E I E K L C E K G R R E V
tgggaaatggaactggatagactcaagaatcaggatggcgaataaataggaacattatg
W E M E L D R L K N Q D G E I N R N I M
gaagagactgaacgggcttgaaggcagagatcttatcactagagaccggaaagagtta
E E T E R A W K A E I L S L E S R K E L
ctggtactgaaactagaagaagcagaaaaagagcagaatgtcaccttacttacctcaag
L V L K L E E A E K E A E L H L T Y L K
tcaactcccccaactggagacagttcgttccaaacaggagtgggagacgagatgaaat
S T P P T A L E T V R S K Q E W E T R L N
ggagttcggataatgaaaaagaatggtcgtgaccaatttaatagtcatatcagtttagt
G V R I M K K N V R D Q F N S H I Q L V
aggaacggagccaagctgagcagccttctcaaatccctactccaacttacctccaccc
R N G A K L S S L P Q I P T P T L P P P
ccatcagagacagacttcatgctcaggtgttcaaccagtcctctctggctcctcgg
P S E T D F M L Q V F Q P S P S L A P R
atgcccttctccattggcagggtcacaatgcccatggttatgcccatgcatccccgc
M P F S I G Q V T M P M V M P S A D P R
tctgttcttttccaactcctgaaacctgaccttccagccagccagccttctccacc
S L S F P I L N P A L S Q P S Q P S S P
cttctggctcccagcagaaatagccctggctgggttcccttggcagccccacgg
L P G S H G R N S P G L G S L V S P H G
ccacacatgccccctgcgcctccatcccacctccccagggttggggggtgtaaggct
P H M P P A A S I P P P P G L G V K A
ctcgtgaaactccccggcccaccagtagacaaactggagaagatcctggagaagctg
S A E T P R P Q P V D K L E K I L E K L
ctgaccgggttcccacagtgcaataaggccagatgaccaacattcttcagcagatcaag
L T R F P Q C N K A Q M T N I L Q Q Q I K
acagcagctaccacctggcaggcctgacctggaggaacttatcagttggttgcagca
T A R T T M A G L T M E E L I Q L V A A
cgactggcagaactgagcgggtggcagcaagtactcagccacttggctgcactccgggcc
R L A E H E R V A A S T Q P L G R I R A
ttgttccctgctccactggcccaaatcagtagcccaatgttcttgccttctgcccagtt
L F P A P L A Q I S T P M F L P S A Q V
tcatatctggaaggtcttccatgctccagccacctgtaagctlgtctaattgctgccag
S Y P G R S S H A P A T C K L C L T M C Q
aaactcgtccagcccagtgctcatccaatggcgtgtaccatgtattgcaacaggag
K L V Q P S E L H P M A C T H V L H K E
tgtatcaaatctgggcccagaccaacacaaatgacacttgccttttggcaccctt
C I K F W A Q Q T N T N D T C P F C P T L
aaatgacagactgactggggaggaagaagaagagaactgatgtgaaacggaagcgcgg
K
gttcaagatttctaaaactctatatttatacagtgacatatactcagctgatgacattt
ttattataggtaatgtgtgatagaagctctgtattccaatgttcgtaaatgaaacta
tgtatattatgcagaacagctctgttccccctcatcttgcaattccttgggggatgcag
attgtagggaagatgatgtttagtttggccttgaatattgatatactcctggcccagggtt
ttttcaatacaatataaaaaccacctaggaaacctgctgtgctctcaaggccattctgct
ttggtttggctcagctctagctcatttcttaaggctcatgtatgcagatttaagcct
ggctgctaccctctgctccaaccagatgccttgccttaccgaagcctccagaagcctcag
taattgttaccctctactccaatggataaaaatgagactctgatgtgaggaaaaaaag
taaccctagtagttgaa

```

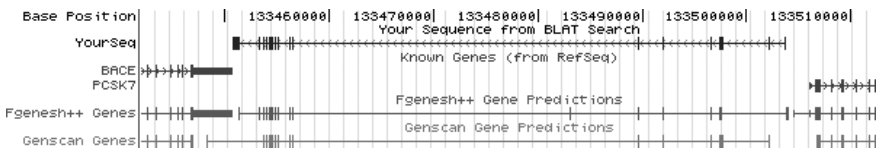
Figure 4.9 Predicted ORF for a novel protein. This was produced by assembling the appropriate assemblies and ESTs into a virtual mRNA. This was then translated to give the putative full-length protein sequence.



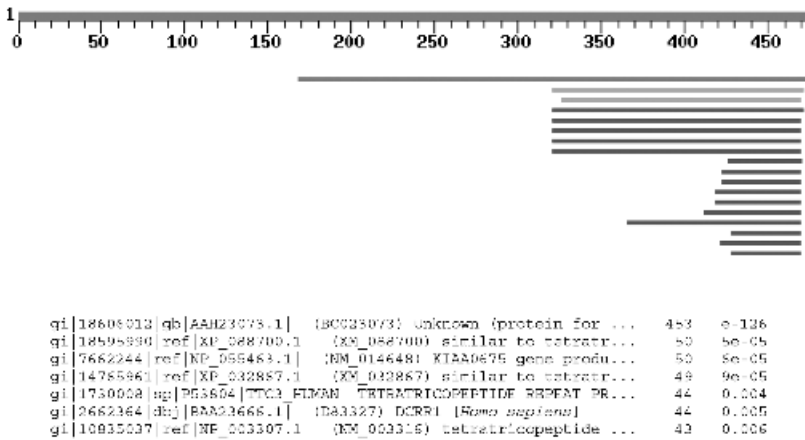
**Figure 4.10** Checking for continuity of reading frame by translation searching (TBLASN) of the unknown ORF against all ESTs. The hatched lines represent deletions in ESTs that could represent splice variants.

in these matches, for example AI351632, represented as hatched lines in Figure 4.10. In addition to a bovine sequence BE75593 we also see a likely orthologous match to AL640079 from a toad. The support for the ORF now seems unassailable. The next step using BLAT again, is to map it back to GP (Figure 4.11). This reveals the matching of 15 exons from putative 5' UTR to 3' UTR. This is consistent with the Fgenes++ prediction at the 5' end but this included two extra exons at the 3' end. The fact that the virtual mRNA butts up very close to both neighbouring genes suggests that this could be a full-length transcript.

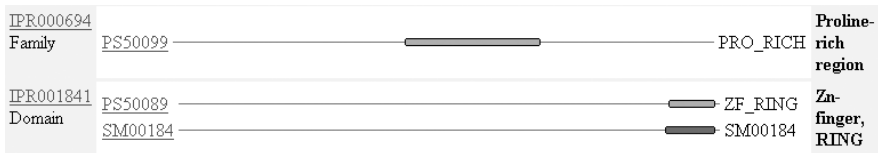
Clearly the analysis of what, for example, might be a candidate disease-associated gene, has to move on from the identification of an ORF to the assignment of function that is both mechanistically plausible and experimentally testable. The subject of assigning functions to new proteins is outside the scope of this chapter. However the two basic steps are a protein database search and motif analysis. The protein search (Figure 4.12) only shows significant similarity scores over the C-terminal section of the protein but the hits include the same proteins assigned as UniGene homologies by Ensembl. A comprehensive domain analysis using InterPro recognizes two domains (Kriventseva *et al.*, 2001; Southan, 2000). One of the domains identified, IPR000694, is a proline-rich domain that may be involved in protein–protein interactions (Figure 4.13). However, the motif recognition specificity is low and therefore this could be a spurious match arising from a general high proline composition. An SRS query shows 1152 of these domains have been recorded in Ensembl (Zdobnov *et al.*, 2002). The second domain, IPR001841, is more specific because it only occurs 187 times in the Ensembl gene set. The RING-finger is a specialized type of Zn-finger of 40 to 60 residues that binds two atoms of zinc, and is probably also involved



**Figure 4.11** Matching the virtual mRNA back against GP using the BLAT search at UCSC. This delineates 15 exons with the gene reading in the opposite orientation to its neighbouring genes, i.e. 3' end to the left, on the same strand.



**Figure 4.12** The sequence similarity scores of the novel ORF against the NCBI non-redundant protein database.



**Figure 4.13** The InterPro domain/protein family analysis result for the novel ORF. The proline-rich domain is defined from a Prosite profile. The zinc finger is defined by both a Prosite profile and a SMART domain.

in mediating protein–protein interactions. They can also bind DNA however, since they contain many Lys, Ser and Thr residues. In fact combining the two domain searches finds intersecting hits (i.e. containing both domains) for only 17 Ensembl proteins. Inspecting the graphical displays shows one of these gene products, ESP0000020915, to be similar in domain orientation and spacing to the novel protein. Unfortunately the trail went cold here because this identifier has been changed in the latest Ensembl release and the SRS link to the protein sequence was dead.

So how did the three major gene portals do? Quite well considering they all included the potential novel gene product as a gene prediction although they disagreed on exon number. They also displayed key supporting evidence in different forms of track annotation. Only a small subset of the display options has been presented here. Was the use of all three portals essential? Strictly speaking we could have accessed sufficient supporting evidence from each one. However to collect all the available data it was necessary to use all three. The other aspect is that each portal has particular facilities that even if not unique at the technical level is easier to use at one portal compared to the other three. Consequently this kind of detailed analysis becomes a *de facto* three-stop-shop. For example the UniGene homology assignments, available from Ensembl, were all correct as judged by the agreement with the protein similarities (red tracks in Figure 4.4). Having said that, one of the direct protein homology assignments (yellow tracks in Figure 4.4),

the myosin-like peptide from yeast, was probably erroneous because of the low complexity of the query protein amino acid composition. In terms of markers, SNPs and genes Ensembl does particularly well for combined export options. The UniGene identity matches on the NCBI display together with the graphical stacks proportional to the number of EST matches are useful but in this case what is likely to be a single transcript was split into four clusters. One of these was illegible on the graphic and two others are dubious because of being unspliced. The UCSC displays were useful to see the two alternative gene models as well as being the only source of the TIGR EST assemblies. Another useful facility on this site is the ability of BLAT to display the hits of any externally constructed model or new database sequence. This can then be compared directly with the other display options (e.g. in Figures 4.6 and 4.11). The NCBI have recently introduced a gene model builder that can reproduce some of the steps above (<http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/ModelMakerHelp.html>).

## 4.8 COMPREHENSIVE DATABASE SEARCHING

The protein matches and the InterPro analysis have already given functional clues about our novel protein. However if this particular gene product was located in close proximity to an SNP with a disease association we would need to find out as much as possible, not only to provide more supporting evidence for the gene product but also testable predictions about function that can be followed up. Performing a comprehensive search is not a trivial exercise since it involves 17 divisions of GenBank and sources of trace data that have not yet been submitted to GenBank. So where do we start? The two large repositories labelled nr protein or nucleotide on the NCBI BLAST server are a useful first choice (<http://www.ncbi.nlm.nih.gov/blast/Blast.cgi>). We have already checked nr protein at 891,607 sequences but we need to compliment this with month, which in this case yields another 61,254 protein sequences but no additional high-scoring hits. The search against nr nucleotide with 1,192,858 sequences records three extended matches. This includes the mouse sequence already described, BC023073, and the primary accession number of the finished genomic section AP000892. The third match, XM\_100696, is a secondary accession number for a reference mRNA sequence predicted by the NCBI Annotation Project from a genomic contig NT\_009151. This is the same prediction labelled LOC160162 in Figure 4.5. There is an accompanying 56-residue predicted ORF that is in the NCBI protein database but has no supporting evidence. Inspection of the genomic location suggests it may be a spurious prediction.

Checking public patented proteins at 88,019 sequences gave no hits. However the patent nucleotide division, gbPAT, at 581,001 sequences, gives three solid hits, AX321627, AX192589 and AX072029. The first of these is a 2114-bp DNA from patent WO0172295. The document indicates this protein was isolated from a lung cancer sample (<http://ep.espacenet.com/>). These hits constitute a partial mRNA level of confirmation for the novel protein but a reciprocal check (i.e. a BLASTN of AX321627 against the nr nucleotide database) indicates this clone may be a chimera from two separate gene products. A search against a commercial patent database, containing 673,453 protein sequences, reveals identity matches for the N-terminal section from patent WO200060077 and a C-terminal identity match from WO200055350, both of which are reported as cancer-associated transcripts (<http://www.derwent.com/geneseq/index.html>). Checking the GSS division by TBLASTN gives four genomic hits; AZ847251 from mouse, AG114530 from chimpanzee, BH306228 from rat and BH406519 from chicken. Using BLAST against the

**TABLE 4.1 Useful Resources for Gene Finding and Analysis**

Site description	URL
Ensembl at EBI/Sanger Centre	<a href="http://www.ensembl.org/">http://www.ensembl.org/</a>
Human Genome Browser at UCSC	<a href="http://genome.ucsc.edu/">http://genome.ucsc.edu/</a>
Map Viewer at NCBI	<a href="http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map_search">http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map_search</a>
Protein Atlas of the genome	<a href="http://www.confirmant.com/">http://www.confirmant.com/</a>
SWISS-2DPAGE database	<a href="http://ca.expasy.org/ch2d/">http://ca.expasy.org/ch2d/</a>
Ensembl 4.28.1 announcement	<a href="http://www.ensembl.org/Dev/Lists/announce/msg00070.html">http://www.ensembl.org/Dev/Lists/announce/msg00070.html</a>
NCBI gene model builder	<a href="http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/ModelMakerHelp.html">http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/ModelMakerHelp.html</a>
UniGene EST clusters	<a href="http://www.ncbi.nlm.nih.gov/UniGene/">http://www.ncbi.nlm.nih.gov/UniGene/</a>
InterPro at EBI	<a href="http://www.ebi.ac.uk/interpro/">http://www.ebi.ac.uk/interpro/</a>
Proteome analysis at EBI	<a href="http://www.ebi.ac.uk/proteome/">http://www.ebi.ac.uk/proteome/</a>
Google general search portal	<a href="http://www.google.com/">http://www.google.com/</a>
RefSeq at NCBI	<a href="http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html">http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html</a>
International Protein Index	<a href="http://www.ebi.ac.uk/IPI/IPIhelp.html">http://www.ebi.ac.uk/IPI/IPIhelp.html</a>
Derwent sequence patent databases	<a href="http://www.derwent.com/geneseq/index.html">http://www.derwent.com/geneseq/index.html</a>
BLAST at NCBI	<a href="http://www.ncbi.nlm.nih.gov/BLAST/">http://www.ncbi.nlm.nih.gov/BLAST/</a>
BLAT at UCSC	<a href="http://genome.ucsc.edu/cgi-bin/hgBlat?command=start">http://genome.ucsc.edu/cgi-bin/hgBlat?command = start)</a>
DAS — distributed annotation	<a href="http://biodas.org/">http://biodas.org/</a>
Exofish at Genoscope	<a href="http://www.genoscope.cns.fr/externe/tetraodon/">http://www.genoscope.cns.fr/externe/tetraodon/</a>
Fgenesh at Sanger Institute	<a href="http://genomic.sanger.ac.uk/gf/Help/fgenesh.html">http://genomic.sanger.ac.uk/gf/Help/fgenesh.html</a>
Expasy translation tool	<a href="http://ca.expasy.org/tools/dna.html">http://ca.expasy.org/tools/dna.html</a>
CAP3 nucleotide assembly tool	<a href="http://bio.ifom-firc.it/ASSEMBLY/assemble.html">http://bio.ifom-firc.it/ASSEMBLY/assemble.html</a>
GeneWise at Sanger Institute	<a href="http://www.sanger.ac.uk/Software/Wise2/">http://www.sanger.ac.uk/Software/Wise2/</a>
Genscan at MIT	<a href="http://genes.mit.edu/GENSCAN.html">http://genes.mit.edu/GENSCAN.html</a>
SSAHA at Sanger Institute	<a href="http://www.sanger.ac.uk/Software/analysis/SSAHA/">http://www.sanger.ac.uk/Software/analysis/SSAHA/</a>

Ensembl mouse peptides detected a C-terminal similarity that is a zinc finger domain match. However both the human and mouse mRNA have unique and solid hits against mouse chromosome 9.40 Mb. This suggests the gene product is derived from this locus although it has not been annotated yet by Ensembl. Interestingly the gene lies between two odour receptors, unlike the human positioning between BACE and PCSK7, showing the position is non-syntenic.

Drawing detailed conclusions from these results is outside the scope of this chapter but the example makes clear how much extra information a comprehensive database search can yield. Was the protein unknown and/or novel? The difficulty of answering this question illustrates the diminishing utility of these terms. The protein has at least one function-related motif that can be recognized at high specificity so it can no longer be classified as an unknown. It remains novel only in the strict sense of not being represented in the current protein databases. It is not novel in the wider sense because both the mRNA and ORF were substantially covered as predicted by sequence data entries in the public and patent databases respectively.

## 4.9 CONCLUSIONS AND PROSPECTS

The geneticist is in the fortunate position of having access to secondary databases and GP genomic viewers of increasing quality, content and utility. This is making the process of finding and analysing gene products easier. However the examples used in this chapter also show that there are many subtle details in genomic annotation and the implications of these will take some time to unravel. This requires comprehensive inspection and may ultimately need experimental verification. The expansion of web-linked interoperativity and interrogation tools means that new options will already be available by the time this is in print. One consequence of these advances could be the perception of a diminished necessity to perform bioinformatic analysis. Although this is true in the sense that secondary databases include an increasing amount of 'pre cooked' bioinformatic data, there is a paradox in that the more sophisticated the public annotation becomes the more important it is to understand the underlying principles. For example, it is important to be able to discriminate between gene products defined by *in-vitro* data or only by *in-silico* prediction.

So what of the future? There are four developments worth highlighting. The first is that the combination of increasing transcript coverage, finished golden path and extensive mouse synteny data will diminish the uncertainty limits of gene numbers. The ability to pick out SNP haplotype blocks in relationship to gene products, already available as tracks on the UCSC display options for chromosome 21 will be a big step forward for association studies (Patil *et al.*, 2001). The proliferation of DAS servers will enable more groups to share their own specialized annotation tracks with the wider community (<http://biodas.org/>). Last but not least defining gene products at the protein level is likely to have a major impact on annotation quality, and efforts are already underway to do this on a genome-wide scale (<http://www.confirmant.com/>).

## REFERENCES

- Birney E, Durbin R. (2000). Using GeneWise in the Drosophila annotation experiment. *Genome Res* **10**: 547–548.
- Burge C, Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**: 78–94.
- Deloukas P, Matthews LH, Ashurst J, Burton J, Gilbert JG, Jones M, *et al.* (2001). The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414**: 865–871.
- Guigo R, Agarwal P, Abril JF, Burset M, Fickett JW. (2000). An assessment of gene prediction accuracy in large DNA sequences. *Genome Res* **10**: 1631–1642.

- Hoogland C, Sanchez JC, Tonella L, Binz PA, Bairoch A, Hochstrasser DF, *et al.* (2000). The 1999 SWISS-2DPAGE database update. *Nucleic Acids Res* **28**: 286–288.
- Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, *et al.* (2002). The Ensembl genome database project. *Nucleic Acids Res* **30**: 38–41.
- Kel-Margoulis OV, Kel AE, Reuter I, Deineko IV, Wingender E. (2002). TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res* **30**: 332–334.
- Kriventseva EV, Biswas M, Apweiler, R. (2001). Clustering and analysis of protein families. *Curr Opin Struct Biol* **11**: 334–339.
- O'Donovan C, Apweiler R, Bairoch A. (2001). The human proteomics initiative (HPI). *Trends Biotechnol* **19**: 178–181.
- Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, *et al.* (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**: 1719–1723.
- Pesole G, Liuni S, Grillo G, Licciulli F, Mignone F, Gissi C, *et al.* (2002). UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Update 2002. *Nucleic Acids Res* **30**: 335–340.
- Pruitt KD, Maglott DR. (2001). RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* **29**: 137–140.
- Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, *et al.* (2001). The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res* **29**: 159–164.
- Schuler GD. (1997). Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J Mol Med* **75**: 694–698.
- Southan C. (2000). Website review: Interpro (the integrated resource of protein domains and functional sites). *Yeast* **17**: 327–334.
- Southan C. (2001). A genomic perspective on human proteases as drug targets. *Drug Discov Today* **6**: 681–688.
- Susens U, Borgmeyer U. (2001). Genomic structure of the gene for mouse germ-cell nuclear factor (GCNF). II. Comparison with the genomic structure of the human GCNF gene. *Genome Biol* **2**: research No. 0017.
- Suzuki Y, Yamashita R, Nakai K, Sugano S. (2002). DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res* **30**: 328–331.
- Wiehe T, Gebauer-Jung S, Mitchell-Olds T, Guigo R. (2001). SGP-1: prediction and validation of homologous genes based on sequence alignments. *Genome Res* **11**: 1574–1583.
- Zdobnov EM, Lopez R, Apweiler R, Etzold T. (2002). The EBI SRS server—recent developments. *Bioinformatics* **18**: 368–373.