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 tool[8]. 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.

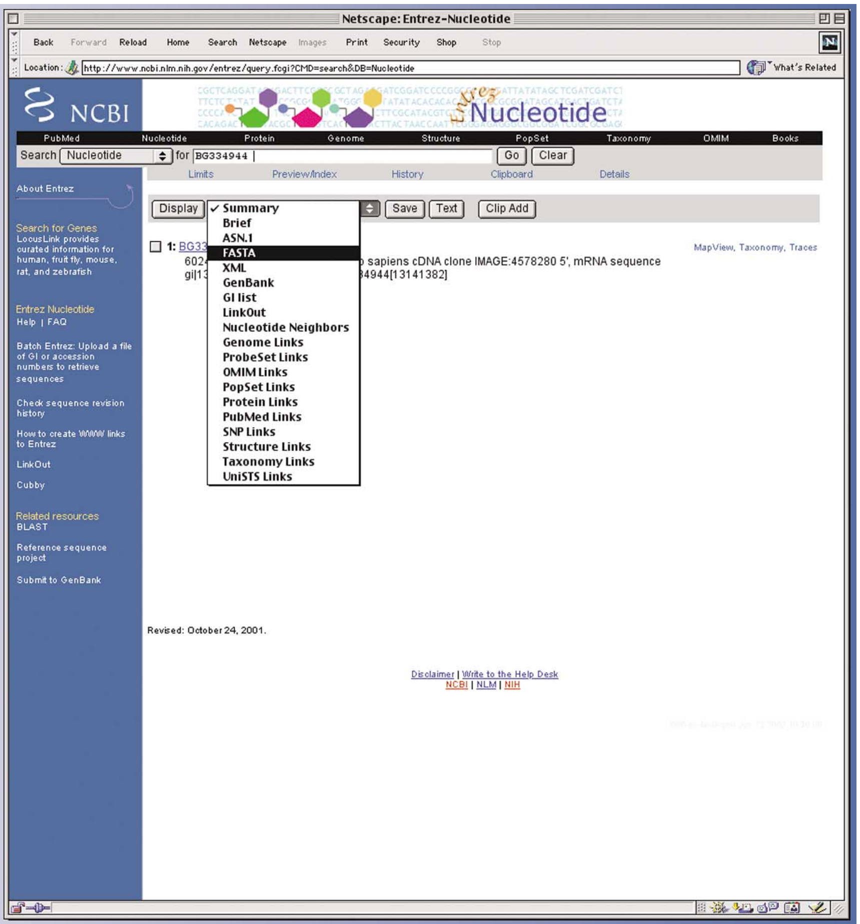
An alternatively spliced mRNA is more likely to be of biological significance when it changes the sequence of the encoded, wildtype protein. To determine whether EST BE798864 could encode a protein different from that of the known gene (*RAB9P40*), one can simply compare the two sequences directly against 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



Netscape: Entrez-Nucleotide
Back Forward Reload Home Search Netscape Images Print Security Shop Stop
What's Related
NCBI
Entrez Nucleotide
PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books
Go Clear
Limits Preview/Index History Clipboard Details
About Entrez
Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish
Entrez Nucleotide
Help | FAQ
Batch Entrez: Upload a file of GI or accession numbers to retrieve sequences
Check sequence revision history
How to create WWW links to Entrez
LinkOut
Cubby
Related resources
BLAST
Reference sequence project
Submit to GenBank
Display
✓ Summary
Brief
ASN.1
FASTA
XML
GenBank
GI list
LinkOut
Nucleotide Neighbors
Genome Links
ProbeSet Links
OMIM Links
PopSet Links
Protein Links
PubMed Links
SNP Links
Structure Links
Taxonomy Links
UniSTS Links
Save Text Clip Add
MapView, Taxonomy, Traces
Disclaimer | Write to the Help Desk
NCBI | NLM | NIH

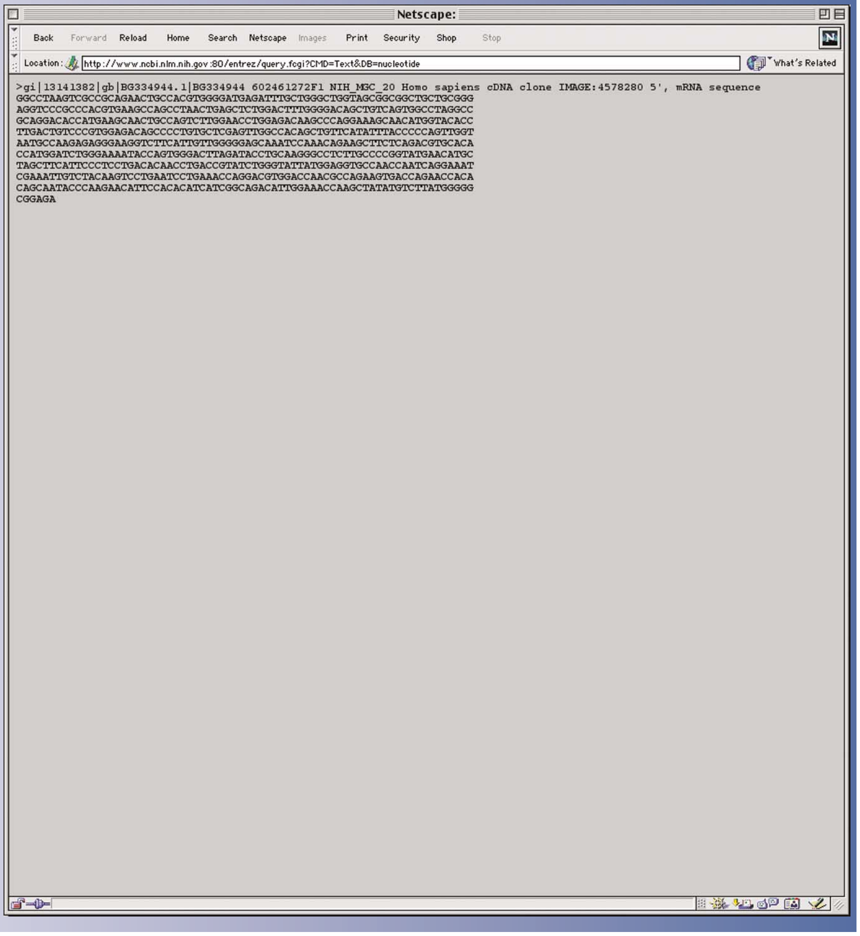
Location: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=Nucleotide

Search Nucleotide for BG334944

□ 1: BG33
602 sapiens cDNA clone IMAGE:4578280 5', mRNA sequence
gi|13 4944[13141382]

Revised: October 24, 2001.

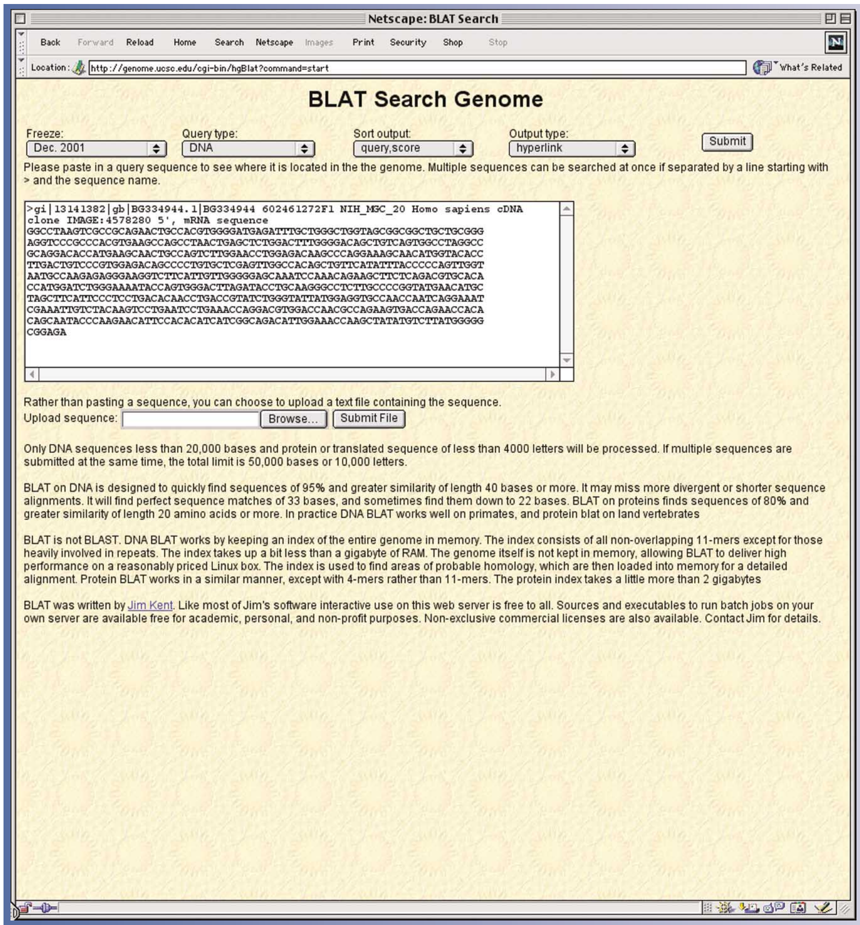
Figure 5.2



Netscape:
Back Forward Reload Home Search Netscape Images Print Security Shop Stop
What's Related

Location: http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?CMD=Text&DB=nucleotide

```
>gi|13141382|gb|BG334944.1|BG334944 602461272F1 NIH_MGC_20 Homo sapiens cDNA clone IMAGE:4578280 5', mRNA sequence
GGCCTAAGTCGCCGCAGAACTGCCACGTGGGGATGAGATTTGCTGGGCTGGTAGCGGCGGCTGCTGCGGG
AGGTCCCGCCCACGTGAAGCCAGCCTAACTGAGCTCTGGACTTTGGGGACAGCTGTCAGTGGCCTAGGCC
GCAGGACACCATGAAGCAACTGCCAGTCTTGGAACCTGGAGACAAGCCCAGGAAAGCAACATGGTACACC
TTGACTGTCCCGTGGAGACAGCCCCTGTGCTCGAGTTGGCCACAGCTGTTCATATTTACCCCCAGTTGGT
AATGCCAAGAGAGGGAAGGTCTTCATTGTTGGGGGAGCAAATCCAAACAGAAGCTTCTCAGACGTGCACA
CCATGGATCTGGGAAAATACCAGTGGGACTTAGATACCTGCAAGGGCCTCTTGCCCCGGTATGAACATGC
TAGCTTCATTCCCTCCTGACACAACCTGACCGTATCTGGGTATTATGGAGGTGCCAACCAATCAGGAAAT
CGAAATTGTCTACAAGTCCTGAATCCTGAAACCAGGACGTGGACCAACGCCAGAAGTGACCAGAACCACA
CAGCAATACCCAAGAACATTCCACACATCATCGGCAGACATTGGAAACCAAGCTATATGTCTTATGGGGG
CGGAGA
```



BLAT Search Genome

Freeze: Dec. 2001 Query type: DNA Sort output: query,score Output type: hyperlink Submit

Please paste in a query sequence to see where it is located in the the genome. Multiple sequences can be searched at once if separated by a line starting with > and the sequence name.

```
>gi|13141382|gb|BG334944.1|BG334944 602461272F1 NIH_MGC_20 Homo sapiens cDNA
clone IMAGE:4578280 5', mRNA sequence
GGCCTAAGTCGCCGCAGAACTGCCACGTGGGGATGAGATTTGCTGGGCTGGTAGCGGCGGCTGCTGCGGG
AGGTCCCGCCCACGTGAAGCCAGCCTAACTGAGCTCTGGACTTTGGGGACAGCTGTCAGTGGCCTAGGCC
GCAGGACACCATGAAGCAACTGCCAGTCTTGGAACCTGGAGACAAGCCCAGGAAAGCAACATGGTACACC
TTGACTGTCCCGTGGAGACAGCCCCTGTGCTCGAGTTGGCCACAGCTGTTCATATTTACCCCCAGTTGGT
AATGCCAAGAGAGGGAAGGTCTTCATTGTTGGGGGAGCAAATCCAAACAGAAGCTTCTCAGACGTGCACA
CCATGGATCTGGGAAAATACCAGTGGGACTTAGATACCTGCAAGGGCCTCTTGCCCCGGTATGAACATGC
TAGCTTCATTCCCTCCTGACACAACCTGACCGTATCTGGGTATTATGGAGGTGCCAACCAATCAGGAAAT
CGAAATTGTCTACAAGTCCTGAATCCTGAAACCAGGACGTGGACCAACGCCAGAAGTGACCAGAACCACA
CAGCAATACCCAAGAACATTCCACACATCATCGGCAGACATTGGAAACCAAGCTATATGTCTTATGGGGG
CGGAGA
```

Rather than pasting a sequence, you can choose to upload a text file containing the sequence.
Upload sequence: Browse... Submit File

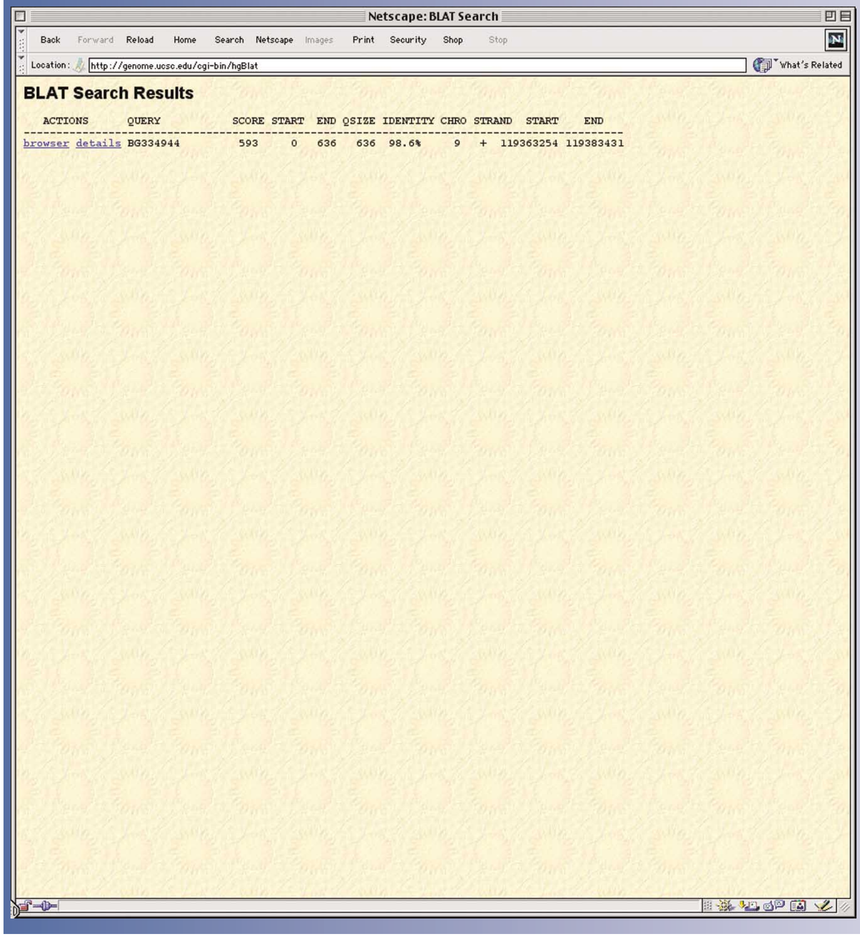
Only DNA sequences less than 20,000 bases and protein or translated sequence of less than 4000 letters will be processed. If multiple sequences are submitted at the same time, the total limit is 50,000 bases or 10,000 letters.

BLAT on DNA is designed to quickly find sequences of 95% and greater similarity of length 40 bases or more. It may miss more divergent or shorter sequence alignments. It will find perfect sequence matches of 33 bases, and sometimes find them down to 22 bases. BLAT on proteins finds sequences of 80% and greater similarity of length 20 amino acids or more. In practice DNA BLAT works well on primates, and protein blat on land vertebrates

BLAT is not BLAST. DNA BLAT works by keeping an index of the entire genome in memory. The index consists of all non-overlapping 11-mers except for those heavily involved in repeats. The index takes up a bit less than a gigabyte of RAM. The genome itself is not kept in memory, allowing BLAT to deliver high performance on a reasonably priced Linux box. The index is used to find areas of probable homology, which are then loaded into memory for a detailed alignment. Protein BLAT works in a similar manner, except with 4-mers rather than 11-mers. The protein index takes a little more than 2 gigabytes

BLAT was written by Jim Kent. Like most of Jim's software interactive use on this web server is free to all. Sources and executables to run batch jobs on your own server are available free for academic, personal, and non-profit purposes. Non-exclusive commercial licenses are also available. Contact Jim for details.

Figure 5.3

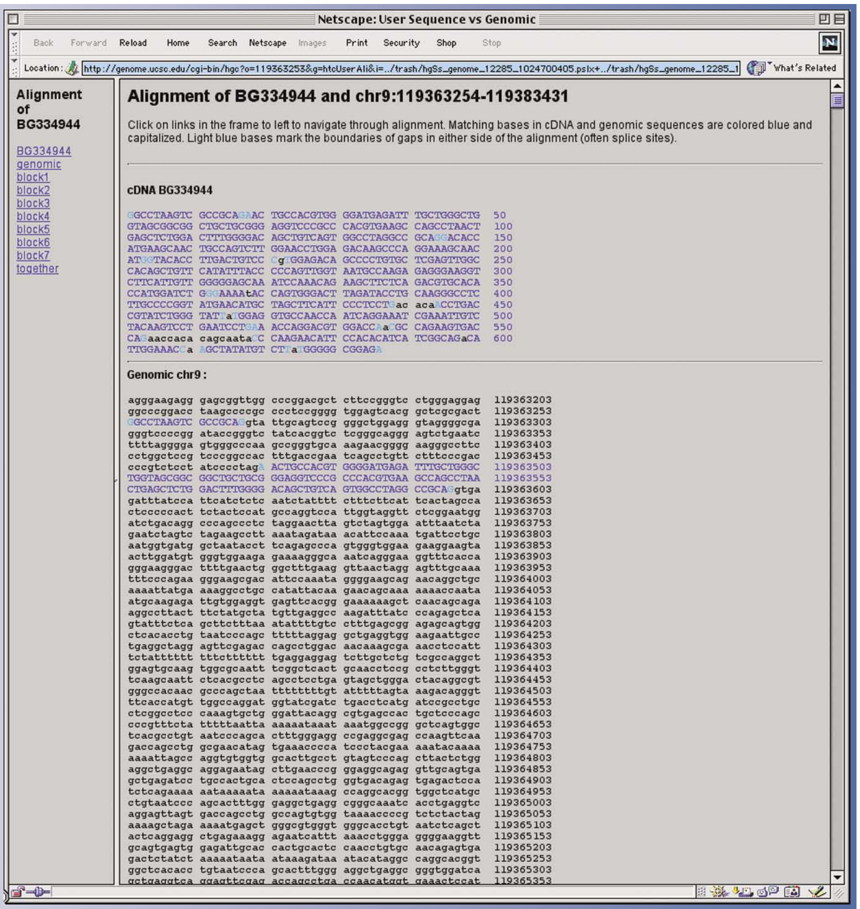


BLAT Search Results

```
   ACTIONS      QUERY           SCORE START  END QSIZE IDENTITY CHRO STRAND  START    END
--------------------------------------------------------------------------------------------
browser details BG334944          593     0  636   636  98.6%      9   +  119363254 119383431
```

Figure 5.4

Figure 5.5



http://genome.ucsc.edu/cgi-bin/hgc?o=119363253&g=htcUserAli&i=../trash/hgSs_genome_12285_1024700405.pslx+../trash/hgSs_genome_12285_1

Alignment of BG334944

BG334944
genomic
block1
block2
block3
block4
block5
block6
block7
together

Alignment of BG334944 and chr9:119363254-119383431

Click on links in the frame to left to navigate through alignment. Matching bases in cDNA and genomic sequences are colored blue and capitalized. Light blue bases mark the boundaries of gaps in either side of the alignment (often splice sites).

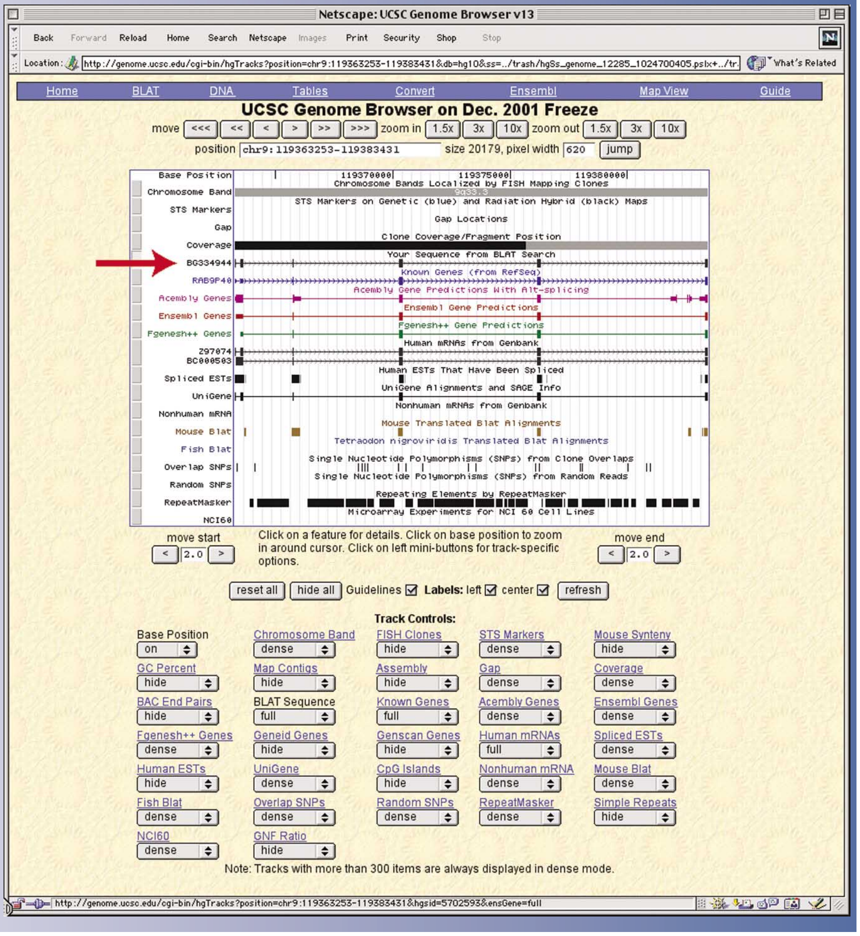
cDNA BG334944

```
GGCCTAAGTC GCCGCAGAAC TGCCACGTGG GGATGAGATT TGCTGGGCTG  50
GTAGCGGCGG CTGCTGCGGG AGGTCCCGCC CACGTGAAGC CAGCCTAACT  100
GAGCTCTGGA CTTTGGGGAC AGCTGTCAGT GGCCTAGGCC GCAGGACACC  150
ATGAAGCAAC TGCCAGTCTT GGAACCTGGA GACAAGCCCA GGAAAGCAAC  200
ATGGTACACC TTGACTGTCC CgTGGAGACA GCCCCTGTGC TCGAGTTGGC  250
CACAGCTGTT CATATTTACC CCCAGTTGGT AATGCCAAGA GAGGGAAGGT  300
CTTCATTGTT GGGGGAGCAA ATCCAAACAG AAGCTTCTCA GACGTGCACA  350
CCATGGATCT GGGAAAAtAC CAGTGGGACT TAGATACCTG CAAGGGCCTC  400
TTGCCCCGGT ATGAACATGC TAGCTTCATT CCCTCCTGac acaACCTGAC  450
CGTATCTGGG TATTaTGGAG GTGCCAACCA ATCAGGAAAT CGAAATTGTC  500
TACAAGTCCT GAATCCTGAA ACCAGGACGT GGACCAaCGC CAGAAGTGAC  550
CAGaaccaca cagcaataCC CAAGAACATT CCACACATCA TCGGCAGaCA  600
TTGGAAACCa AGCTATATGT CTTaTGGGGG CGGAGA
```

Genomic chr9 :

```
agggaagagg gagcggttgg cccggacgct cttccgggtc ctgggaggag  119363203
ggcccggacc taagccccgc ccctccgggg tggagtcacg gctcgcgact  119363253
GGCCTAAGTC GCCGCAGgta ttgcagtccg gggctggagg gtaggggcga  119363303
gggtccccgg ataccgggtc tatcacggtc tcgggcaggg agtctgaatc  119363353
ttttagggga gtgggcccaa gccgggtgca aagaacgggg aagggccttc  119363403
cctggctccg tcccggccac tttgaccgaa tcagcctgtt ctttcccgac  119363453
cccgtctcct atcccctagA ACTGCCACGT GGGGATGAGA TTTGCTGGGC  119363503
TGGTAGCGGC GGCTGCTGCG GGAGGTCCCG CCCACGTGAA GCCAGCCTAA  119363553
CTGAGCTCTG GACTTTGGGG ACAGCTGTCA GTGGCCTAGG CCGCAGgtga  119363603
gatttatcca ttcatctctc aatctatttt ctttcttcat tcactagcca  119363653
ctcccccact tctactccat gccaggtcca ttggtaggtt ctcggaatgg  119363703
atctgacagg cccagccctc taggaactta gtctagtgga atttaatcta  119363753
gaatctagtc tagaagcctt aaatagataa acattccaaa tgattcctgc  119363803
aatggtgatg gctaatacct tcagagccca gtgggtggaa gaaggaagta  119363853
acttggatgt gggtggaaga gaaaagggca aatcagggaa ggtttcacca  119363903
gggaagggac ttttgaactg ggctttgaag gttaactagg agtttgcaaa  119363953
tttcccagaa gggaagcgac attccaaata ggggaagcag aacaggctgc  119364003
aaaattatga aaaggcctgc catattacaa gaacagcaaa aaaaccaata  119364053
atgcaagaga ttgtggaggt gagttcacgg gaaaaaagct caacagcaga  119364103
aggccttact ttctatgcta tgttgaggcc aagatttatc ccagagctca  119364153
gtatttctca gcttctttaa atattttgtc ctttgagcgg agagcagtgg  119364203
ctcacacctg taatcccagc tttttaggag gctgaggtgg aagaattgcc  119364253
tgaggctagg agttcgagac cagcctggac aacaaagcga aacctccatt  119364303
tctatttttt tttctttttt tgaggaggag tcttgctctg tcgccaggct  119364353
ggagtgcaag tggcgcaatt tcggctcact gcaacctccg cctcttgggt  119364403
tcaagcaatt ctcacgcctc agcctcctga gtagctggga ctacaggcgt  119364453
gggccacaac gcccagctaa tttttttgt atttttagta aagacagggt  119364503
ttcaccatgt tggccaggat ggtatcgatc tgacctcatg atccgcctgc  119364553
ctcggcctcc caaagtgctg ggattacagg cgtgagccac tgctcccagc  119364603
cccgtttcta tttttaatta aaaaataaat aaatggccgg gctcagtggc  119364653
tcacgcctgt aatcccagca ctttgggagg ccgaggcgag ccaagttcaa  119364703
gaccagcctg gcgaacatag tgaaacccca tccctacgaa aaatacaaaa  119364753
aaaattagcc aggtgtggtg gcacttgcct gtagtcccag cttactctgg  119364803
aggctgaggc aggagaatag cttgaacccg ggaggcagag gttgcagtga  119364853
gctgagatcc tgccactgca ctccagcctg ggtgacagag tgagactcca  119364903
tctcagaaaa aataaaaata aaaaataaag ccaggcacgg tggctcatgc  119364953
ctgtaatccc agcactttgg gaggctgagg cgggcaaatc acctgaggtc  119365003
aggagttagt gaccagcctg gccagtgtgg taaaaccccg tctctactag  119365053
aaaagctaga aaaatgagct gggcgtgggt gggcacctgt aatctcagct  119365103
actcaggagg ctgagaaagg agaatcattt aaacctggga ggggaaggtt  119365153
gcagtgagtg gagattgcac cactgcactc caacctgtgc aacagagtga  119365203
gactctatct aaaaataata ataaagataa atacataggc caggcacggt  119365253
ggctcacacc tgtaatccca gcactttggg aggctgaggc gggtggatca  119365303
gctgaggtca ggagttcgag accagcctga ccaacatggt gaaactccat  119365353
```

Figure 5.6


Netscape: UCSC Genome Browser v13
Back
Forward
Reload
Home
Search
Netscape
Images
Print
Security
Shop
Stop
Location:
What's Related
Home
BLAT
DNA
Tables
Convert
Ensembl
Map View
Guide
move
<<<
<<
<
>
>>
>>>
zoom in
1.5x
3x
10x
zoom out
1.5x
3x
10x
jump
Base Position
119370000
119375000
119380000
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
9q33.3
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
Gap Locations
Gap
Clone Coverage/Fragment Position
Coverage
Your Sequence from BLAT Search
BG334944
Known Genes (from RefSeq)
RAB9P40
Acembly Gene Predictions With Alt-splicing
Acembly Genes
Ensembl Gene Predictions
Ensembl Genes
Fgenesh++ Gene Predictions
Fgenesh++ Genes
Human mRNAs from Genbank
Z97074
BC000503
Human ESTs That Have Been Spliced
Spliced ESTs
UniGene Alignments and SAGE Info
UniGene
Nonhuman mRNAs from Genbank
Nonhuman mRNA
Mouse Translated Blat Alignments
Mouse Blat
Tetraodon nigroviridis Translated Blat Alignments
Fish Blat
Single Nucleotide Polymorphisms (SNPs) from Clone Overlaps
Overlap SNPs
Single Nucleotide Polymorphisms (SNPs) from Random Reads
Random SNPs
Repeating Elements by RepeatMasker
RepeatMasker
Microarray Experiments for NCI 60 Cell Lines
NCI60
move start
<
>
move end
<
>
reset all
hide all
refresh

http://genome.ucsc.edu/cgi-bin/hgTracks?position=chr9:119363253-119383431&db=hg10&ss=../trash/hgSs_genome_12285_1024700405.pslx+../tr.

UCSC Genome Browser on Dec. 2001 Freeze

position chr9:119363253-119383431 size 20179, pixel width 620

Click on a feature for details. Click on base position to zoom in around cursor. Click on left mini-buttons for track-specific options.

move start 2.0 move end 2.0

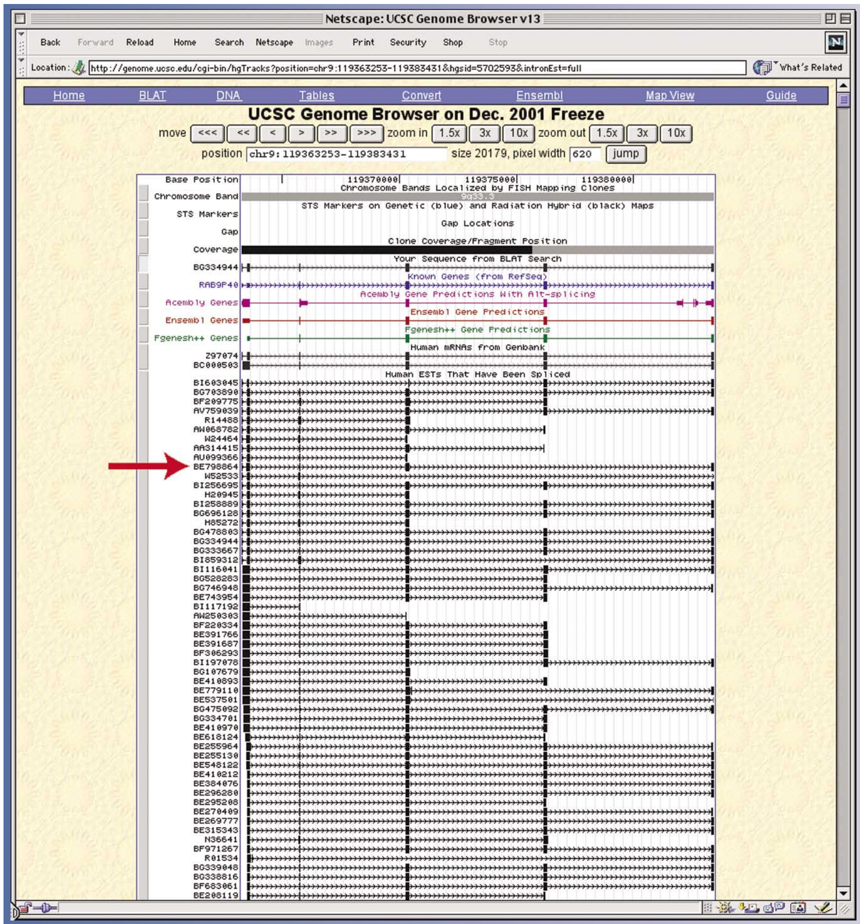
Guidelines ☑ **Labels:** left ☑ center ☑

Track Controls:

Base Position on	Chromosome Band dense	FISH Clones hide	STS Markers dense	Mouse Synteny hide
GC Percent hide	Map Contigs hide	Assembly hide	Gap dense	Coverage dense
BAC End Pairs hide	BLAT Sequence full	Known Genes full	Acembly Genes dense	Ensembl Genes dense
Fgenesh++ Genes dense	Geneid Genes hide	Genscan Genes hide	Human mRNAs full	Spliced ESTs dense
Human ESTs hide	UniGene dense	CpG Islands hide	Nonhuman mRNA dense	Mouse Blat dense
Fish Blat dense	Overlap SNPs dense	Random SNPs dense	RepeatMasker dense	Simple Repeats hide
NCI60 dense	GNF Ratio hide			

Note: Tracks with more than 300 items are always displayed in dense mode.

http://genome.ucsc.edu/cgi-bin/hgTracks?position=chr9:119363253-119383431&hgsid=5702593&ensGene=full

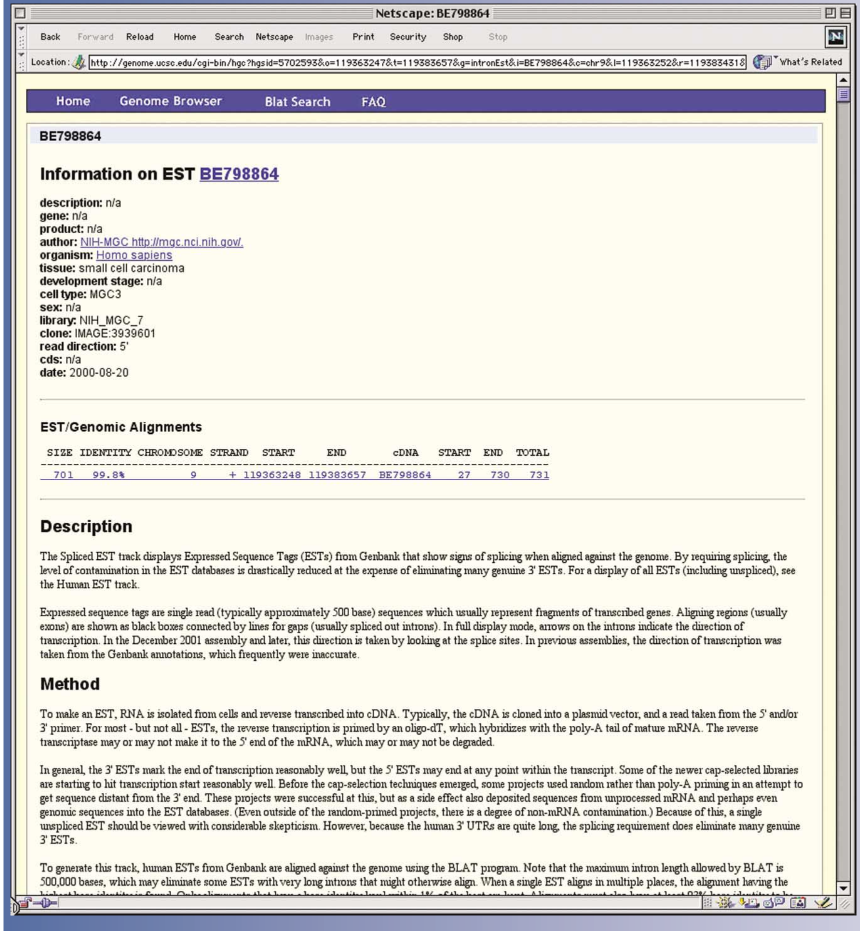


Location: http://genome.ucsc.edu/cgi-bin/hgTracks?position=chr9:119363253-119383431&hgsid=5702593&intronEst=full

UCSC Genome Browser on Dec. 2001 Freeze

position chr9:119363253-119383431 size 20179, pixel width 620

Figure 5.7


Netscape: BE798864
Back Forward Reload Home Search Netscape Images Print Security Shop Stop
What's Related
Home Genome Browser Blat Search FAQ

Location: http://genome.ucsc.edu/cgi-bin/hgc?hgsid=5702593&o=119363247&t=119383657&g=intronEst&i=BE798864&c=chr9&l=119363252&r=119383431&

BE798864

Information on EST BE798864

description: n/a
gene: n/a
product: n/a
author: NIH-MGC http://mgc.nci.nih.gov/.
organism: Homo sapiens
tissue: small cell carcinoma
development stage: n/a
cell type: MGC3
sex: n/a
library: NIH_MGC_7
clone: IMAGE:3939601
read direction: 5'
cds: n/a
date: 2000-08-20

EST/Genomic Alignments

```
 SIZE IDENTITY CHROMOSOME STRAND   START      END        cDNA    START  END  TOTAL
----------------------------------------------------------------------------------
  701   99.8%          9      + 119363248 119383657  BE798864     27   730    731
```

Description

The Spliced EST track displays Expressed Sequence Tags (ESTs) from Genbank that show signs of splicing when aligned against the genome. By requiring splicing, the level of contamination in the EST databases is drastically reduced at the expense of eliminating many genuine 3' ESTs. For a display of all ESTs (including unspliced), see the Human EST track.

Expressed sequence tags are single read (typically approximately 500 base) sequences which usually represent fragments of transcribed genes. Aligning regions (usually exons) are shown as black boxes connected by lines for gaps (usually spliced out introns). In full display mode, arrows on the introns indicate the direction of transcription. In the December 2001 assembly and later, this direction is taken by looking at the splice sites. In previous assemblies, the direction of transcription was taken from the Genbank annotations, which frequently were inaccurate.

Method

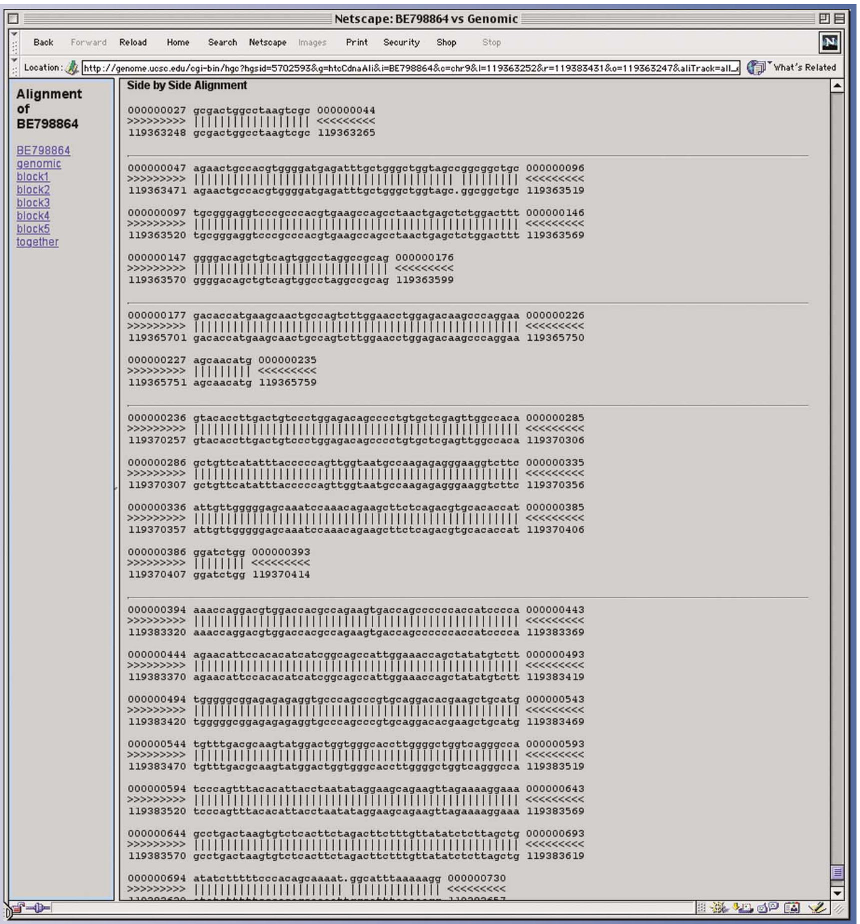
To make an EST, RNA is isolated from cells and reverse transcribed into cDNA. Typically, the cDNA is cloned into a plasmid vector, and a read taken from the 5' and/or 3' primer. For most - but not all - ESTs, the reverse transcription is primed by an oligo-dT, which hybridizes with the poly-A tail of mature mRNA. The reverse transcriptase may or may not make it to the 5' end of the mRNA, which may or may not be degraded.

In general, the 3' ESTs mark the end of transcription reasonably well, but the 5' ESTs may end at any point within the transcript. Some of the newer cap-selected libraries are starting to hit transcription start reasonably well. Before the cap-selection techniques emerged, some projects used random rather than poly-A priming in an attempt to get sequence distant from the 3' end. These projects were successful at this, but as a side effect also deposited sequences from unprocessed mRNA and perhaps even genomic sequences into the EST databases. (Even outside of the random-primed projects, there is a degree of non-mRNA contamination.) Because of this, a single unspliced EST should be viewed with considerable skepticism. However, because the human 3' UTRs are quite long, the splicing requirement does eliminate many genuine 3' ESTs.

To generate this track, human ESTs from Genbank are aligned against the genome using the BLAT program. Note that the maximum intron length allowed by BLAT is 500,000 bases, which may eliminate some ESTs with very long introns that might otherwise align. When a single EST aligns in multiple places, the alignment having the highest base identity is found. Only alignments that have a base identity level within 1% of the best are kept. Alignments must also have at least 93% base identity to be

Figure 5.8

Figure 5.9



http://genome.ucsc.edu/cgi-bin/hgc?hgsid=5702593&g=htcCdnaAli&i=BE798864&c=chr9&l=119363252&r=119383431&o=119363247&aliTrack=all_

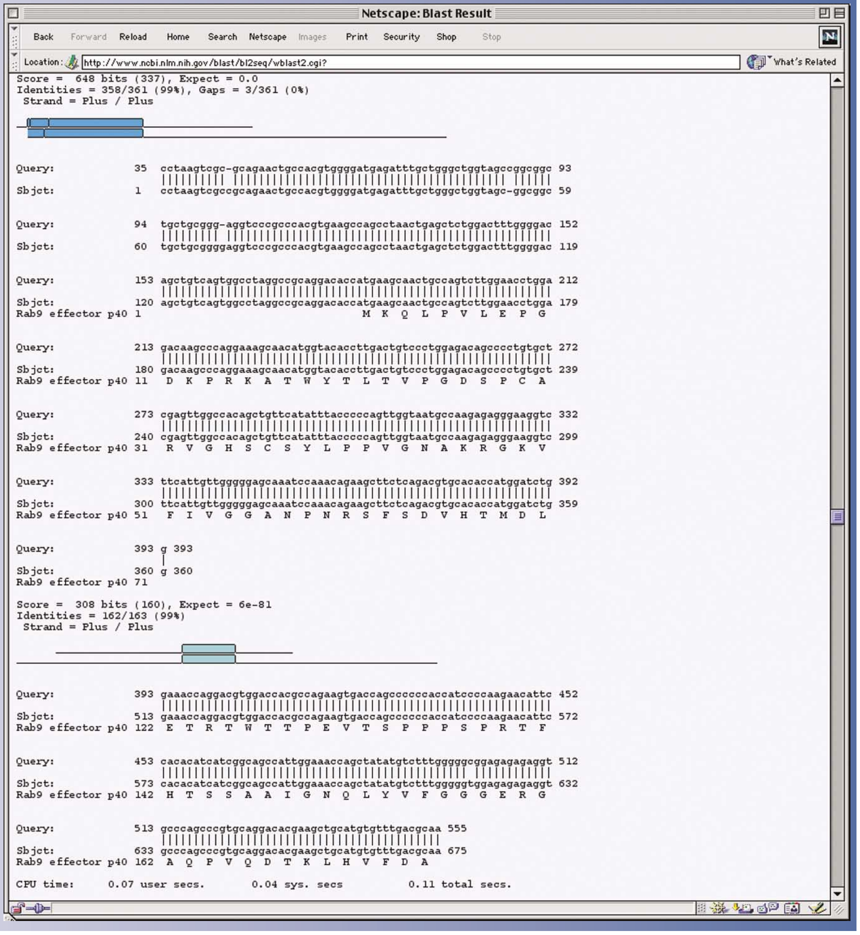
Alignment of BE798864

- BE798864
- genomic
- block1
- block2
- block3
- block4
- block5
- together

Side by Side Alignment

```
000000027 gcgactggcctaagtcgc 000000044
>>>>>>>>> |||||||||||||||||| <<<<<<<<<
119363248 gcgactggcctaagtcgc 119363265


000000047 agaactgccacgtggggatgagatttgctgggctggtagccggcggctgc 000000096
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||| ||||||||| <<<<<<<<<
119363471 agaactgccacgtggggatgagatttgctgggctggtagc.ggcggctgc 119363519

000000097 tgcgggaggtcccgcccacgtgaagccagcctaactgagctctggacttt 000000146
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<<<
119363520 tgcgggaggtcccgcccacgtgaagccagcctaactgagctctggacttt 119363569

000000147 ggggacagctgtcagtggcctaggccgcag 000000176
>>>>>>>>> |||||||||||||||||||||||||||||| <<<<<<<<<
119363570 ggggacagctgtcagtggcctaggccgcag 119363599


000000177 gacaccatgaagcaactgccagtcttggaacctggagacaagcccaggaa 000000226
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<<<
119365701 gacaccatgaagcaactgccagtcttggaacctggagacaagcccaggaa 119365750

000000227 agcaacatg 000000235
>>>>>>>>> ||||||||| <<<<<<<<<
119365751 agcaacatg 119365759


000000236 gtacaccttgactgtccctggagacagcccctgtgctcgagttggccaca 000000285
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<<<
119370257 gtacaccttgactgtccctggagacagcccctgtgctcgagttggccaca 119370306

000000286 gctgttcatatttaccccagttggtaatgccaagagagggaaggtcttc 000000335
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<<<
119370307 gctgttcatatttaccccagttggtaatgccaagagagggaaggtcttc 119370356

000000336 attgttgggggagcaaatccaaacagaagcttctcagacgtgcacaccat 000000385
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<<<
119370357 attgttgggggagcaaatccaaacagaagcttctcagacgtgcacaccat 119370406

000000386 ggatctgg 000000393
>>>>>>>>> |||||||| <<<<<<<<<
119370407 ggatctgg 119370414


000000394 aaaccaggacgtggaccacgccagaagtgaccagcccccaccatcccca 000000443
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<<<
119383320 aaaccaggacgtggaccacgccagaagtgaccagcccccaccatcccca 119383369

000000444 agaacattccacacatcatcggcagccattggaaaccagctatatgtctt 000000493
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<<<
119383370 agaacattccacacatcatcggcagccattggaaaccagctatatgtctt 119383419

000000494 tgggggcggagagagaggtgcccagcccgtgcaggacacgaagctgcatg 000000543
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<<<
119383420 tgggggcggagagagaggtgcccagcccgtgcaggacacgaagctgcatg 119383469

000000544 tgtttgacgcaagtatggactggtgggcaccttggggctggtcagggcca 000000593
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<<<
119383470 tgtttgacgcaagtatggactggtgggcaccttggggctggtcagggcca 119383519

000000594 tcccagtttacacattacctaatataggaagcagaagttagaaaaggaaa 000000643
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<<<
119383520 tcccagtttacacattacctaatataggaagcagaagttagaaaaggaaa 119383569

000000644 gcctgactaagtgtctcacttctagacttctttgttatatctcttagctg 000000693
>>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<<<
119383570 gcctgactaagtgtctcacttctagacttctttgttatatctcttagctg 119383619

000000694 atatcttttcccacagcaaaat.ggcatttaaaaagg 000000730
>>>>>>>>> |||||||||||||||||||||| |||||||||||||| <<<<<<<<<
```

Figure 5.10



http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi?

```
Score =  648 bits (337), Expect = 0.0
Identities = 358/361 (99%), Gaps = 3/361 (0%)
 Strand = Plus / Plus


Query:            35  cctaagtcgc-gcagaactgccacgtggggatgagatttgctgggctggtagccggcggc 93
                      |||||||||| |||||||||||||||||||||||||||||||||||||||||| ||||||
Sbjct:            1   cctaagtcgccgcagaactgccacgtggggatgagatttgctgggctggtagc-ggcggc 59


Query:            94  tgctgcggg-aggtcccgcccacgtgaagccagcctaactgagctctggactttggggac 152
                      ||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:            60  tgctgcggggaggtcccgcccacgtgaagccagcctaactgagctctggactttggggac 119


Query:            153 agctgtcagtggcctaggccgcaggacaccatgaagcaactgccagtcttggaacctgga 212
                      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:            120 agctgtcagtggcctaggccgcaggacaccatgaagcaactgccagtcttggaacctgga 179
Rab9 effector p40 1                                  M  K  Q  L  P  V  L  E  P  G


Query:            213 gacaagcccaggaaagcaacatggtacaccttgactgtccctggagacagcccctgtgct 272
                      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:            180 gacaagcccaggaaagcaacatggtacaccttgactgtccctggagacagcccctgtgct 239
Rab9 effector p40 11   D  K  P  R  K  A  T  W  Y  T  L  T  V  P  G  D  S  P  C  A


Query:            273 cgagttggccacagctgttcatatttaccccagttggtaatgccaagagagggaaggtc 332
                      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:            240 cgagttggccacagctgttcatatttaccccagttggtaatgccaagagagggaaggtc 299
Rab9 effector p40 31   R  V  G  H  S  C  S  Y  L  P  P  V  G  N  A  K  R  G  K  V


Query:            333 ttcattgttgggggagcaaatccaaacagaagcttctcagacgtgcacaccatggatctg 392
                      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:            300 ttcattgttgggggagcaaatccaaacagaagcttctcagacgtgcacaccatggatctg 359
Rab9 effector p40 51   F  I  V  G  G  A  N  P  N  R  S  F  S  D  V  H  T  M  D  L


Query:            393 g 393
                      |
Sbjct:            360 g 360
Rab9 effector p40 71

Score =  308 bits (160), Expect = 6e-81
Identities = 162/163 (99%)
 Strand = Plus / Plus


Query:            393 gaaaccaggacgtggaccacgccagaagtgaccagcccccaccatccccaagaacattc 452
                      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:            513 gaaaccaggacgtggaccacgccagaagtgaccagcccccaccatccccaagaacattc 572
Rab9 effector p40 122  E  T  R  T  W  T  T  P  E  V  T  S  P  P  P  S  P  R  T  F


Query:            453 cacacatcatcggcagccattggaaaccagctatatgtctttgggggcggagagagaggt 512
                      ||||||||||||||||||||||||||||||||||||||||||||||| ||||||||||||
Sbjct:            573 cacacatcatcggcagccattggaaaccagctatatgtctttgggggtggagagagaggt 632
Rab9 effector p40 142  H  T  S  S  A  A  I  G  N  Q  L  Y  V  F  G  G  G  E  R  G


Query:            513 gcccagcccgtgcaggacacgaagctgcatgtgtttgacgcaa 555
                      |||||||||||||||||||||||||||||||||||||||||||
Sbjct:            633 gcccagcccgtgcaggacacgaagctgcatgtgtttgacgcaa 675
Rab9 effector p40 162  A  Q  P  V  Q  D  T  K  L  H  V  F  D  A

CPU time:     0.07 user secs.       0.04 sys. secs         0.11 total secs.
```