# Module 1: Manual Genome Annotation and *de novo* Analysis of Sequence

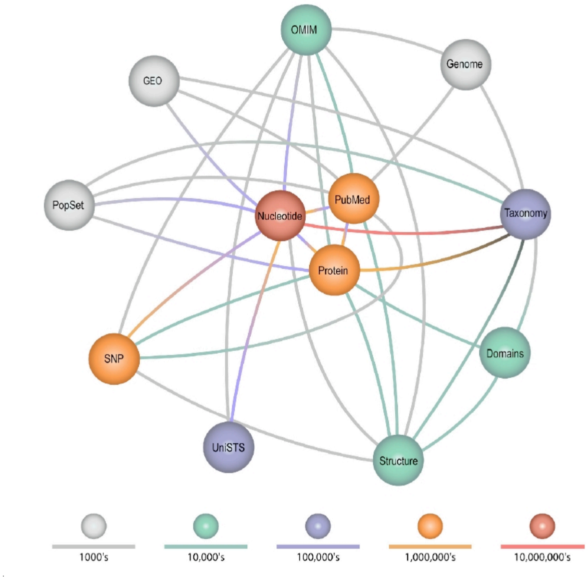
## Aims

- How to access sequence information using Entrez and UniProt.
- General *de novo* analysis of cDNAs, including ORF finder to highlight putative protein products of a cDNA, blastp link from within this to investigate the potential protein products, Spidey to align cDNA to genomic DNA. Clustalw to align similar sequences, view in Jalview and use GeneDoc to produce a graphics file.
- View manually annotated genes in the Vega browser.
- Perform a Blast search using sanger blast server on all finished and unfinished zebrafish clones

## General Introduction

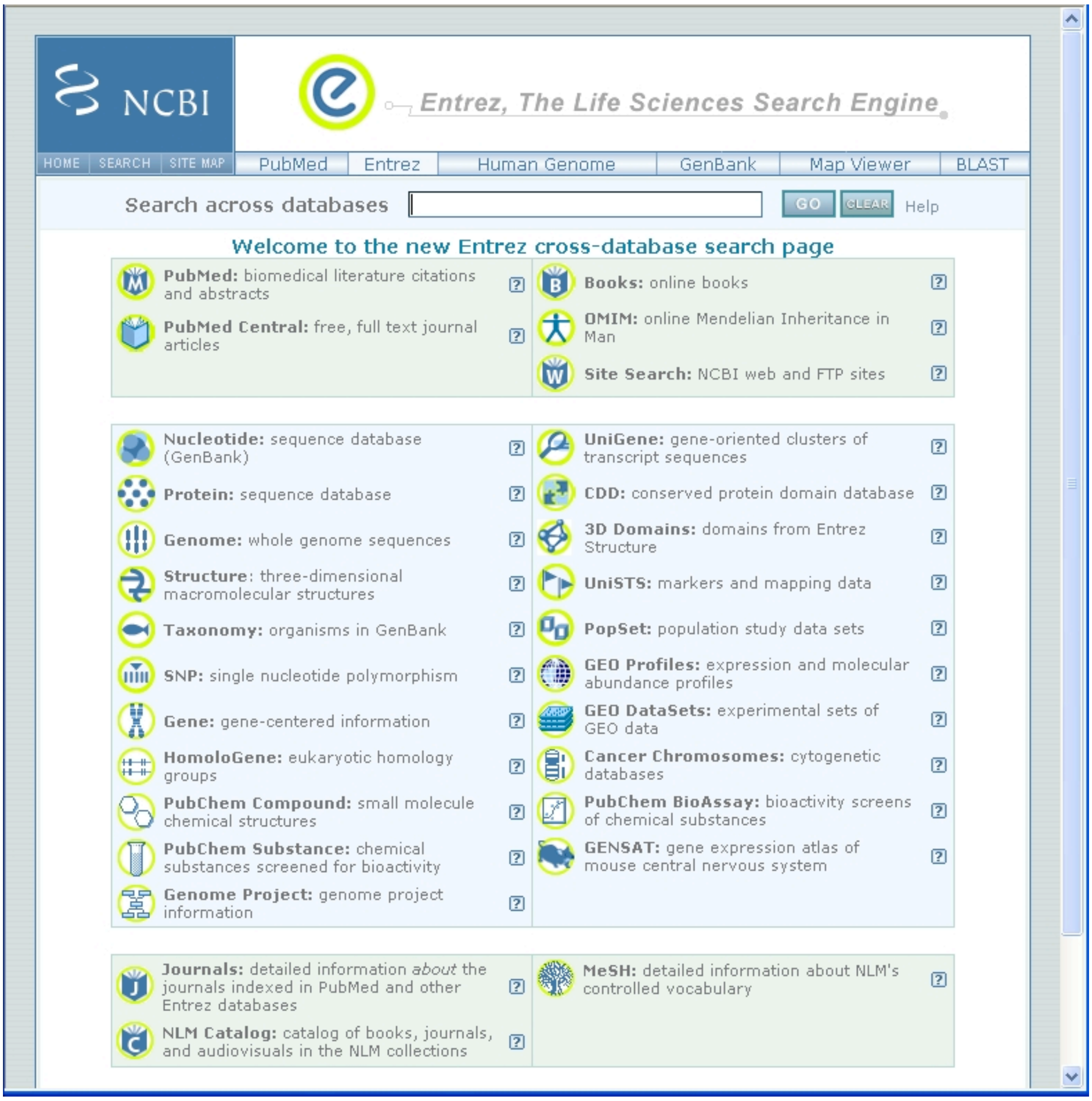
Genomic and protein sequences can be accessed from various databases around the world. I will now introduce the major ways to access this sequence information, namely Entrez at NCBI and UniProt at EBI.

**Entrez** is a search and retrieval system that integrates information from databases at NCBI including the reference sequence (RefSeq) collection. The databases at NCBI include nucleotide sequences, protein sequences, macromolecular structures, whole genomes, and MEDLINE (through PubMed).



Dataflow diagram of the Entrez retrieval system

**Main Entrez entry point:**



Welcome to the new Entrez cross-database search page

**PubMed:** biomedical literature citations and abstracts

**PubMed Central:** free, full text journal articles

**Books:** online books

**OMIM:** online Mendelian Inheritance in Man

**Site Search:** NCBI web and FTP sites

**Nucleotide:** sequence database (GenBank)

**Protein:** sequence database

**Genome:** whole genome sequences

**Structure**: three-dimensional macromolecular structures

**Taxonomy:** organisms in GenBank

**SNP:** single nucleotide polymorphism

**Gene:** gene-centered information

**HomoloGene:** eukaryotic homology groups

**PubChem Compound:** small molecule chemical structures

**PubChem Substance:** chemical substances screened for bioactivity

**Genome Project:** genome project information

**UniGene:** gene-oriented clusters of transcript sequences

**CDD:** conserved protein domain database

**3D Domains:** domains from Entrez Structure

**UniSTS:** markers and mapping data

**PopSet:** population study data sets

**GEO Profiles:** expression and molecular abundance profiles

**GEO DataSets:** experimental sets of GEO data

**Cancer Chromosomes:** cytogenetic databases

**PubChem BioAssay:** bioactivity screens of chemical substances

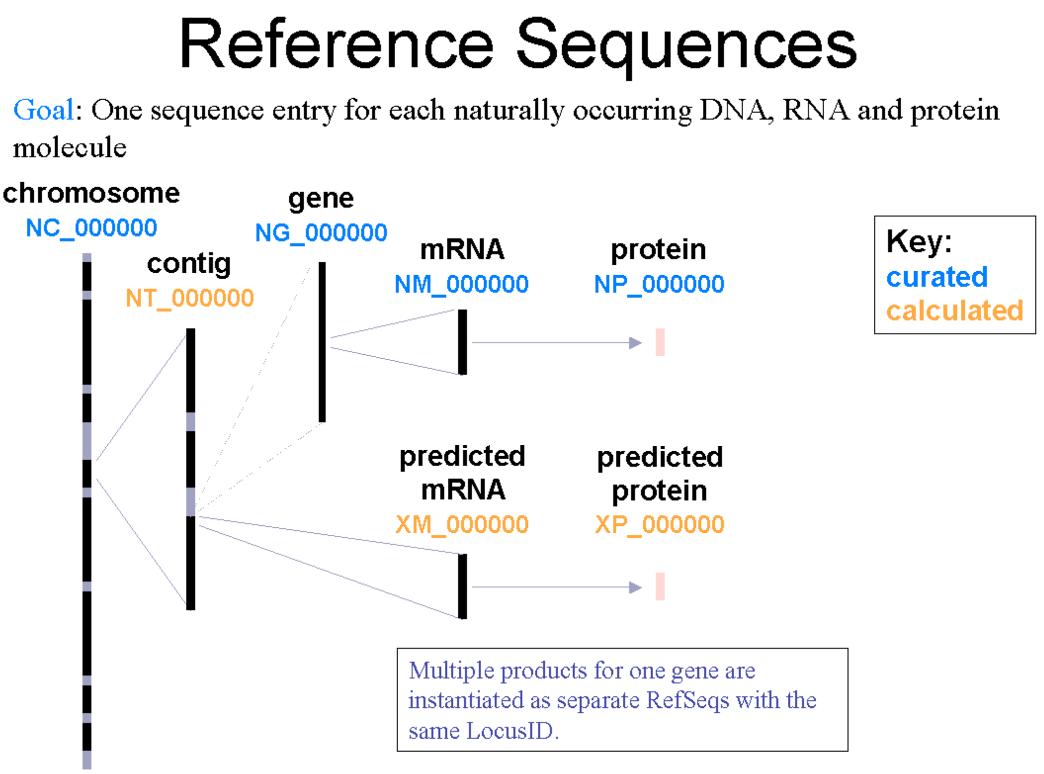
**GENSAT:** gene expression atlas of mouse central nervous system

**Journals:** detailed information *about* the journals indexed in PubMed and other Entrez databases

**NLM Catalog:** catalog of books, journals, and audiovisuals in the NLM collections

**MeSH:** detailed information about NLM's controlled vocabulary

**The Reference Sequence (RefSeq) database** provides a biologically non-redundant collection of DNA, RNA, and protein sequences. Each RefSeq represents a single, naturally occurring molecule from a particular organism. RefSeqs are frequently based on GenBank records but differ in that each RefSeq is a synthesis of information, not a piece of a primary research data in itself. Similar to a review article in literature, a RefSeq is an interpretation by a particular group at a particular time. RefSeqs can be retrieved in several different ways: by searching the Entrez Nucleotide or Protein database, by BLAST searching, by FTP, or through links from other NCBI resources.



# Reference Sequences

Goal: One sequence entry for each naturally occurring DNA, RNA and protein molecule

chromosome
NC_000000

contig
NT_000000

gene
NG_000000

mRNA
NM_000000

protein
NP_000000

predicted mRNA
XM_000000

predicted protein
XP_000000

Key:
curated
calculated

Multiple products for one gene are instantiated as separate RefSeqs with the same LocusID.

**For further information about RefSeq please visit: http://www.ncbi.nlm.nih.gov/RefSeq/key.html#accession**

## Curated Records

| | |
|---|---|
| NC_123455 | Complete genomic sequence (chromosome) |
| NG_123456 | Incomplete genomic sequence |
| NM_123456 | mRNA |
| NP_123456 | Protein derived from NM |
| NR_123456 | Non-coding RNA |

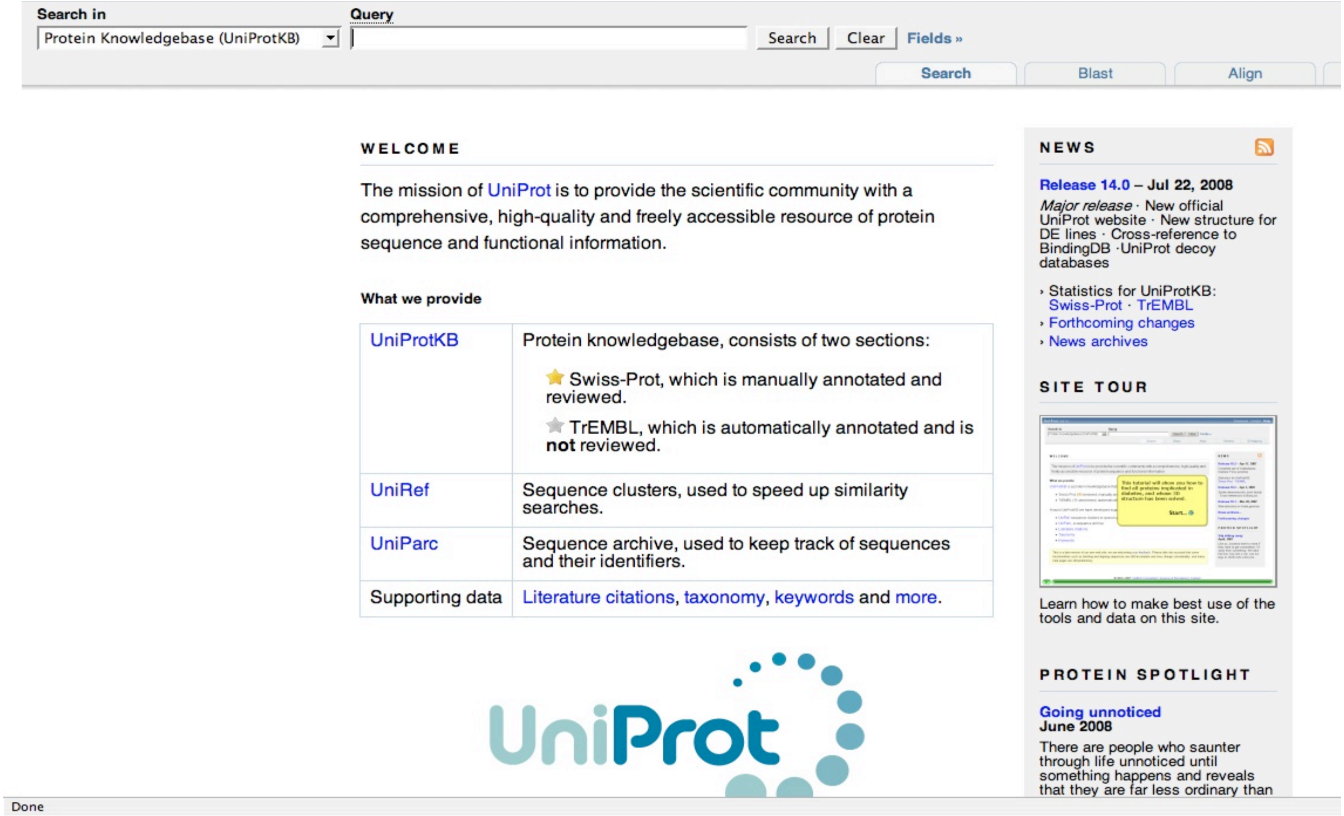
## Model Records

| | |
|---|---|
| NT_123456 | Assembly of BAC data |
| NW_123456 | Assembly of WGS data |
| NZ_ABCD12345678 | Collection of WGS data |
| XM_123456 | mRNA |
| XP_123456 | Protein derived from XM |
| ZP_123456 | Protein derived from NZ |
| XR_123456 | Non-coding RNA |

### The benefits of RefSeq:

- non-redundancy
- explicitly linked nucleotide and protein sequences
- updates to reflect current knowledge of sequence data and biology
- data validation and format consistency
- distinct accession series
- ongoing curation by NCBI staff and collaborators, with reviewed records indicated

**UniProt** (Universal Protein Resource) is the world's most comprehensive catalogue of protein information. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot (a database containing non-redundant, manual high-quality annotation that is cross-referenced to many other databases, with flatfiles containing features and information on individual protein sequences. Each entry has a unique accession number), TrEMBL (an automatic computer-annotated supplement to SwissProt, containing the translations of all coding sequences (CDS) present in the EMBL Nucleotide Sequence Database that are not yet integrated into SwissProt), and The Protein Information Resource (PIR), located at Georgetown University Medical Center (GUMC).



WELCOME

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

What we provide

| | |
|---|---|
| UniProtKB | Protein knowledgebase, consists of two sections:<br>★ Swiss-Prot, which is manually annotated and reviewed.<br>☆ TrEMBL, which is automatically annotated and is **not** reviewed. |
| UniRef | Sequence clusters, used to speed up similarity searches. |
| UniParc | Sequence archive, used to keep track of sequences and their identifiers. |
| Supporting data | Literature citations, taxonomy, keywords and more. |

NEWS

Release 14.0 – Jul 22, 2008

*Major release* · New official UniProt website · New structure for DE lines · Cross-reference to BindingDB ·UniProt decoy databases

› Statistics for UniProtKB: Swiss-Prot · TrEMBL
› Forthcoming changes
› News archives

SITE TOUR

Learn how to make best use of the tools and data on this site.

PROTEIN SPOTLIGHT

Going unnoticed
June 2008

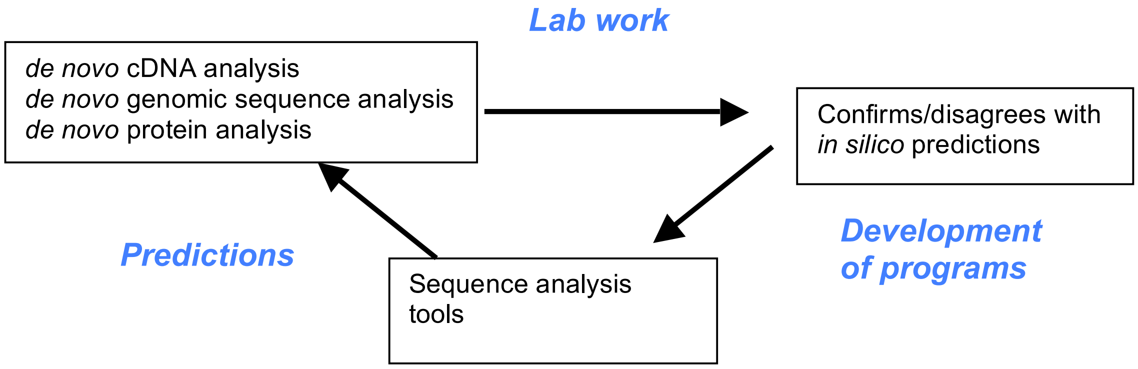
There are people who saunter through life unnoticed until something happens and reveals that they are far less ordinary than

## Module Introduction

The Vertebrate Genome Annotation (Vega) database is a central repository for high quality, frequently updated, manual annotation of vertebrate finished genome sequence. The browser is based on the code from the Ensembl project and the data is produced by the Human and Vertebrate Analysis and Annotation (HAVANA) group at the Wellcome Trust Sanger Institute. Currently the available species are human, mouse, zebrafish, pig and dog. The annotation is undertaken in collaboration and synchronisation with the central zebrafish database ZFIN.

If you have generated a piece of sequence (either cDNA or genomic sequence) you are unlikely to want to wait for the sequence to be annotated automatically by genome viewers. It is possible to annotate the sequence manually, using freely available tools to access various databases. In addition, some of these tools, such as ORF finder, are not used by genome viewers so if you have found a region of interest in Ensembl or NCBI, you may wish to analyse an already annotated sequence using these programs to provide additional information. You may also want to perform *in silico* predictions of expression of a putative protein product, search for similar proteins or align similar sequences.

Such analyses are part of the following flow diagram:


Lab work
*de novo* cDNA analysis
*de novo* genomic sequence analysis
*de novo* protein analysis
Confirms/disagrees with *in silico* predictions
Development of programs
Sequence analysis tools
Predictions

The following is a summary of the DNA and protein analysis programs used in this module, together with a brief outline of their functions:

*ORF Finder (Open Reading Frame Finder)*
A graphical analysis tool that finds open reading frames in a sequence. The putative protein sequences may then be blasted by integral blastp.

*Spidey*
This aligns cDNA to genomic DNA and will output a list of exons together with an alignment and protein translation.

*CLUSTALW*
DNA and protein sequences can be aligned with clustalw and viewed in Jalview.

*GeneDoc*
A desktop package, used to produce graphical images of sequence alignments in many formats.

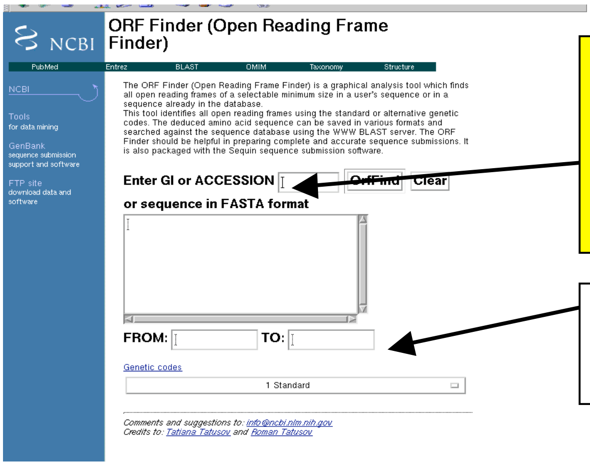
## The Vega database

Vega differs from Ensembl in that it shows annotation from the labour intensive process of manual curation produced by the Human and Vertebrate Analysis and Annotation (HAVANA) group at the Wellcome Trust Sanger Institute (WTSI). Finished genomic sequence is analysed on a clone by clone basis using a combination of similarity searches against DNA and protein databases and a series of *ab initio* gene predictions. Annotation is based on supporting evidence, which is external sequence such as ESTs, cDNAs and protein, and is performed to the standards agreed at the human annotation workshops (HAWK). Vega displays complete chromosomes and specific regions of interest. Grey shading indicates annotation status, with light grey showing partially annotated regions and dark grey showing regions with no annotation. Currently, human chromosomes 1, 3, 6, 7, 8, 9, 10, 13, 14, 16, 17, 18, 19, 20, 22, X and Y have full manual annotation, together with 44 genome-wide ENCODE regions and eight human haplotypes (6-COX, 6-QBL, 6-SSTO, 6-APD, 6-DBB, 6-MANN, 6-MCF, 6-PGF) of the chromosome 6 MHC region. There are also CORF genes on chromosomes 1, 2, 3, 4, 5, 8, 9, 11, 12, 15 and 17. The CORF project (at the WTSI) aims to produce a cDNA clone of each protein coding gene in the human genome, and manual annotation has been used to confirm ORFs and UTRs, to enable the design of PCR primers. Mouse currently has chromosomes 2, 4, 11 and X, together with candidate Insulin Dependent Diabetes (IDD) regions and the DeL36H regions of chromosome 13. Zebrafish has annotation for the majority of genes in ZFIN as well as full chromosome annotation on chrs 1, 2, 4, 5, 8, 9, 10, 13, 18, 19, 20, 22 and 23 currently displayed. Pig shows the MHC (SLA) region on chromosome 7 from Large White Boar and 8Mb of chromosome 17, and dog shows the MHC (DLA) class II region on chromosome 12 from Doberman.

Vega is an important contributor to the conserved CDS (CCDS) project, which is a collaborative effort between the European Bioinformatics Institute (EBI), the National Centre for Biotechnology Information (NCBI), the Wellcome Trust Sanger Institute (WTSI) and the University of California at Santa Cruz (UCSC). The aim of the project is to identify a core set of human protein coding regions that are consistently annotated between the different institutes.

**1.** Does AB006190 have any open reading frames? Use NCBI ORF finder to find ORFs. What potential protein products does this cDNA code for? Use the link from ORF finder to investigate using blastp.

**ORF Finder: www.ncbi.nlm.nih.gov/gorf/gorf.html**



ORF Finder (Open Reading Frame Finder)

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database.
This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software.

**Enter GI or ACCESSION**

**or sequence in FASTA format**

**FROM:** **TO:**

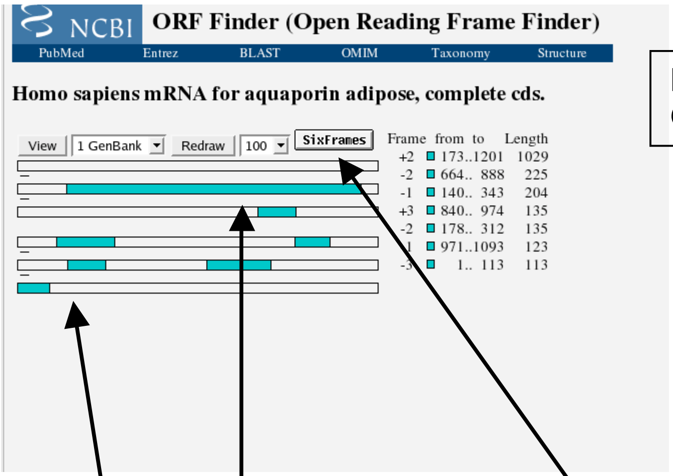
Genetic codes

1 Standard

*Comments and suggestions to: info@ncbi.nlm.nih.gov*
*Credits to: Tatiana Tatusov and Roman Tatusov*

Type in accession number AB006190 into box, or copy and paste sequence in fasta format. **Click on OrfFind**

Select bases and genetic code if required



**ORF Finder (Open Reading Frame Finder)**

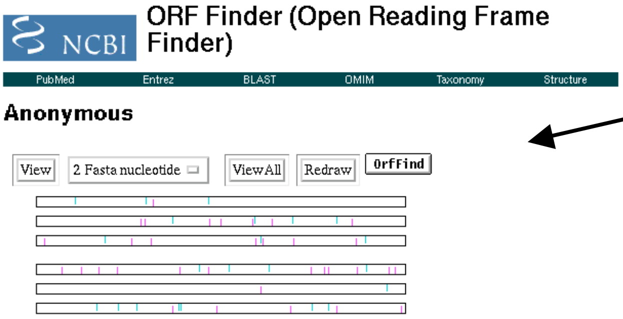
**Homo sapiens mRNA for aquaporin adipose, complete cds.**

| Frame | | from to | Length |
|---|---|---|---|
| +2 | 🟦 | 173..1201 | 1029 |
| -2 | 🟦 | 664.. 888 | 225 |
| -1 | 🟦 | 140.. 343 | 204 |
| +3 | 🟦 | 840.. 974 | 135 |
| -2 | 🟦 | 178.. 312 | 135 |
| 1 | 🟦 | 971..1093 | 123 |
| -3 | 🟦 | 1.. 113 | 113 |

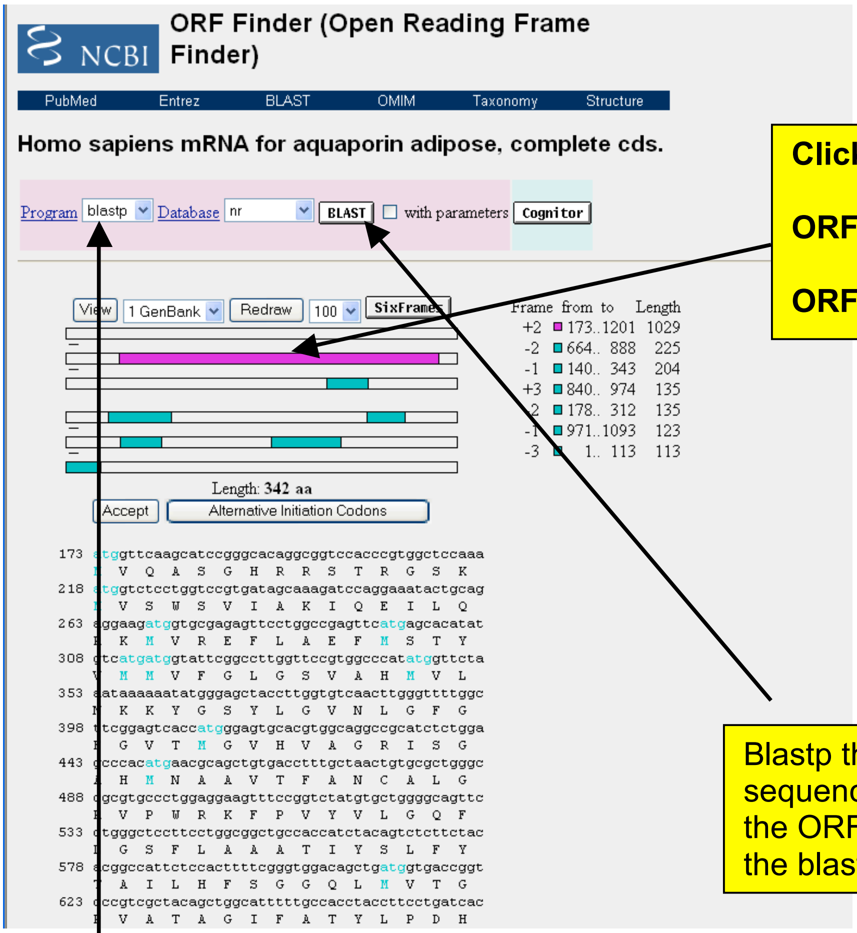
Position and length of ORFs in the sequence

Shaded boxes indicate ORFs. AB006190 has 7 potential ORFs.

Click on 'SixFrames'



Blue lines = start codons
Pink lines=stop codons



Homo sapiens mRNA for aquaporin adipose, complete cds.

| Frame | | from | to | Length |
|---|---|---|---|---|
| +2 | ■ | 173.. | 1201 | 1029 |
| -2 | ■ | 664.. | 888 | 225 |
| -1 | ■ | 140.. | 343 | 204 |
| +3 | ■ | 840.. | 974 | 135 |
| -2 | ■ | 178.. | 312 | 135 |
| -1 | ■ | 971.. | 1093 | 123 |
| -3 | ■ | 1.. | 113 | 113 |

Length: **342 aa**

```
173 atggttcaagcatccgggcacaggcggtccacccgtggctccaaa
    M  V  Q  A  S  G  H  R  R  S  T  R  G  S  K
218 atggtctcctggtccgtgatagcaaagatccaggaaatactgcag
    M  V  S  W  S  V  I  A  K  I  Q  E  I  L  Q
263 aggaagatggtgcgagagttcctggccgagttcatgagcacatat
    R  K  M  V  R  E  F  L  A  E  F  M  S  T  Y
308 gtcatgatggtattcggccttggttccgtggcccatatggttcta
    V  M  M  V  F  G  L  G  S  V  A  H  M  V  L
353 aataaaaaatatgggagctaccttggtgtcaacttgggttttggc
    N  K  K  Y  G  S  Y  L  G  V  N  L  G  F  G
398 ttcggagtcaccatgggagtgcacgtggcaggccgcatctctgga
    F  G  V  T  M  G  V  H  V  A  G  R  I  S  G
443 gcccacatgaacgcagctgtgacctttgctaactgtgcgctgggc
    A  H  M  N  A  A  V  T  F  A  N  C  A  L  G
488 cgcgtgccctggaggaagtttccggtctatgtgctggggcagttc
    R  V  P  W  R  K  F  P  V  Y  V  L  G  Q  F
533 ctgggctccttcctggcggctgccaccatctacagtctcttctac
    L  G  S  F  L  A  A  A  T  I  Y  S  L  F  Y
578 acggccattctccacttttcgggtggacagctgatggtgaccggt
    T  A  I  L  H  F  S  G  G  Q  L  M  V  T  G
623 cccgtcgctacagctggcatttttgccacctaccttcctgatcac
    P  V  A  T  A  G  I  F  A  T  Y  L  P  D  H
```

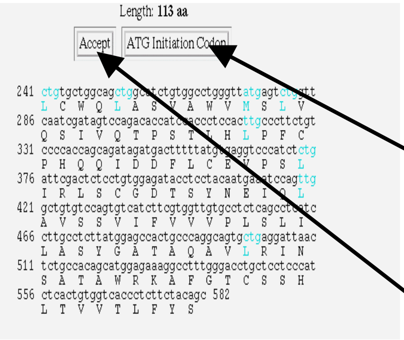
**Click on a shaded**

**ORF to get this view.**

**ORF changes to pink**

Blastp the protein sequence shown below the ORF by clicking on the blast button.

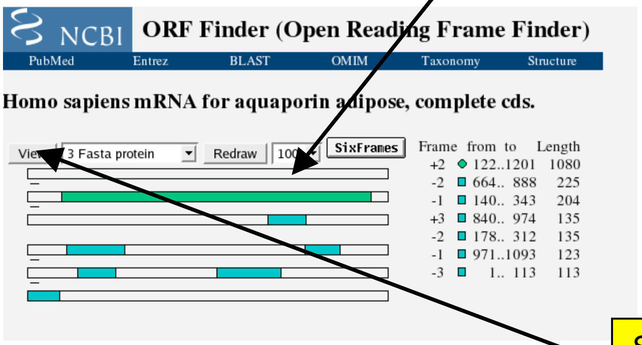
blastp or tblastn and database can be selected



```
241 ctgtgctggcagctggcatctgtggcctgggttatgagtctggtt
    L  C  W  Q  L  A  S  V  A  W  V  M  S  L  V
286 caatcgatagtccagacaccatccaccctccacttgcccttctgt
    Q  S  I  V  Q  T  P  S  T  L  H  L  P  F  C
331 ccccaccagcagatagatgacttttatgtgaggtcccatctctg
    P  H  Q  Q  I  D  D  F  L  C  E  V  P  S  L
376 attcgactctcctgtggagatacctcctacaatgaaatccagttg
    I  R  L  S  C  G  D  T  S  Y  N  E  I  Q  L
421 gctgtgtccagtgtcatcttcgtggttgtgcctctcagcctcatc
    A  V  S  S  V  I  F  V  V  V  P  L  S  L  I
466 cttgcctcttatggagccactgcccaggcagtgctgaggattaac
    L  A  S  Y  G  A  T  A  Q  A  V  L  R  I  N
511 tctgccacagcatggagaaaggcctttgggacctgctcctcccat
    S  A  T  A  W  R  K  A  F  G  T  C  S  S  H
556 ctcactgtggtcaccctcttctacagc 582
    L  T  V  V  T  L  F  Y  S
```

Alternative initiation codons: highlights leucine codons as well as methionine codons.

Click on accept when you are happy with a CDS (the highlighted region will turn green).

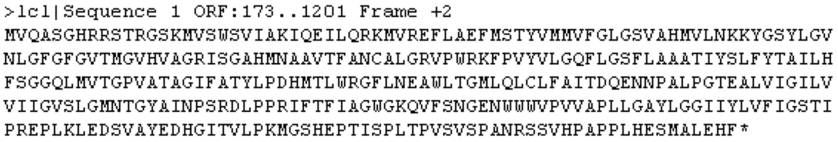


ORF Finder (Open Reading Frame Finder)

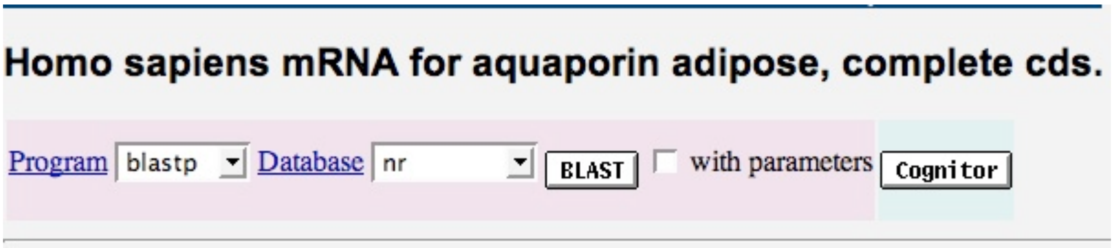
Homo sapiens mRNA for aquaporin adipose, complete cds.

```
Frame  from  to   Length
 +2  ●  122..1201  1080
 -2  ■  664..  888   225
 -1  ■  140..  343   204
 +3  ■  840..  974   135
 -2  ■  178..  312   135
 -1  ■  971..1093    123
 -3  ■    1..  113   113
```

Select the view to be fasta protein and click on view to bring up the sequence in fasta format. This can now be copied and pasted into other applications.



```
>lcl|Sequence 1 ORF:173..1201 Frame +2
MVQASGHRRSTRGSKMVSWSVIAKIQEILQRKMVREFLAEFMSTYVMMVFGLGSVAHMVLNKKYGSYLGV
NLGFGFGVTMGVHVAGRISGAHMNAAVTFANCALGRVPWRKFPVYVLGQFLGSFLAAATIYSLFYTAILH
FSGGQLMVTGPVATAGIFATYLPDHMTLWRGFLNEAWLTGMLQLCLFAITDQENNPALPGTEALVIGILV
VIIGVSLGMNTGYAINPSRDLPPRIFTFIAGWGKQVFSNGENWWWVPVVAPLLGAYLGGIIYLVFIGSTI
PREPLKLEDSVAYEDHGITVLPKMGSHEPTISPLTPVSVSPANRSSVHPAPPLHESMALEHF*
```

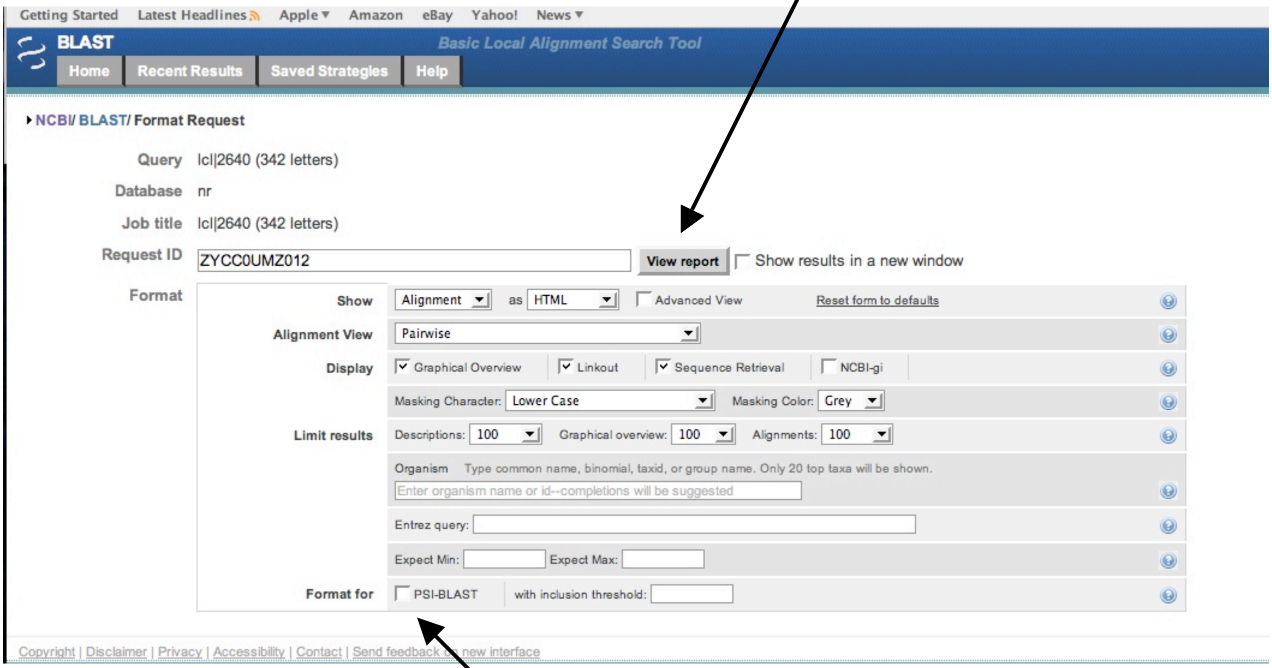


Homo sapiens mRNA for aquaporin adipose, complete cds.

Program blastp Database nr

Click on BLAST button

Click on View report to view BLAST results



Getting Started Latest Headlines Apple Amazon eBay Yahoo! News
BLAST Basic Local Alignment Search Tool
Home Recent Results Saved Strategies Help
NCBI/ BLAST/ Format Request

Query lcl|2640 (342 letters)

Database nr

Job title lcl|2640 (342 letters)

Request ID ZYCC0UMZ012

View report
□ Show results in a new window

Format

Show: Alignment as HTML □ Advanced View

Alignment View: Pairwise

Display: ☑ Graphical Overview ☑ Linkout ☑ Sequence Retrieval □ NCBI-gi

Masking Character: Lower Case Masking Color: Grey

Limit results: Descriptions: 100 Graphical overview: 100 Alignments: 100

Organism Type common name, binomial, taxid, or group name. Only 20 top taxa will be shown.

Enter organism name or id--completions will be suggested

Entrez query:

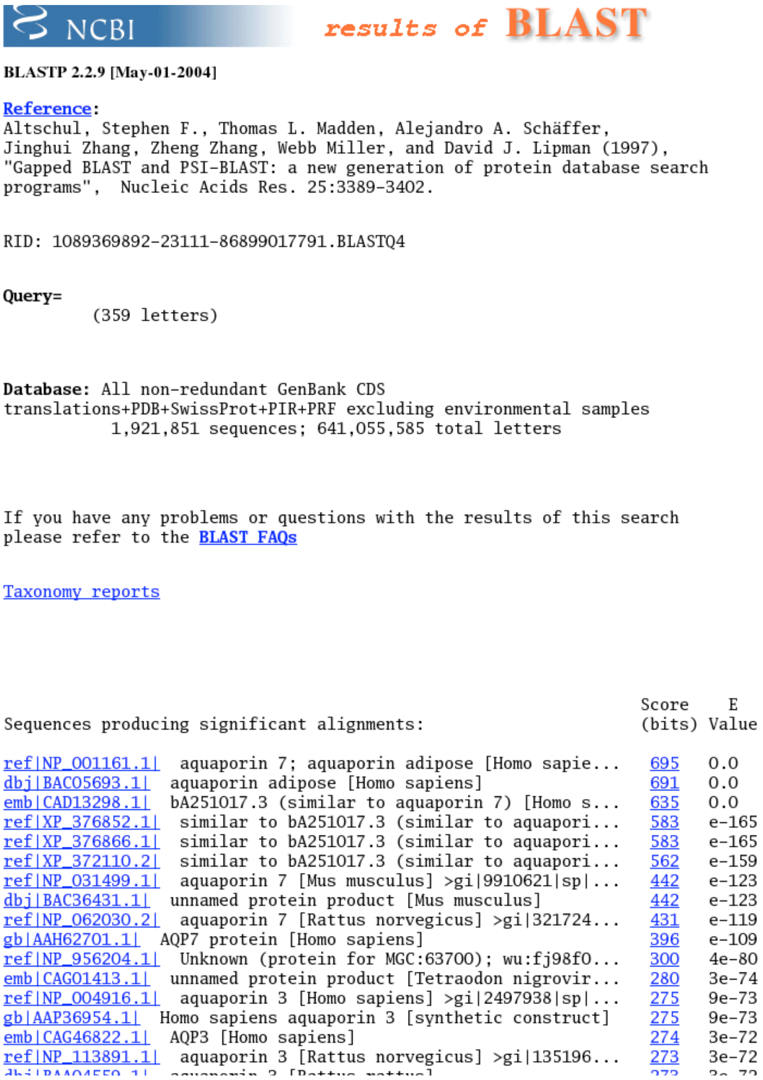
Expect Min: Expect Max:

Format for □ PSI-BLAST with inclusion threshold:

Reset form to defaults
Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback on new interface

Can also format for PSI-Blast if required by checking the PSI-BLAST box

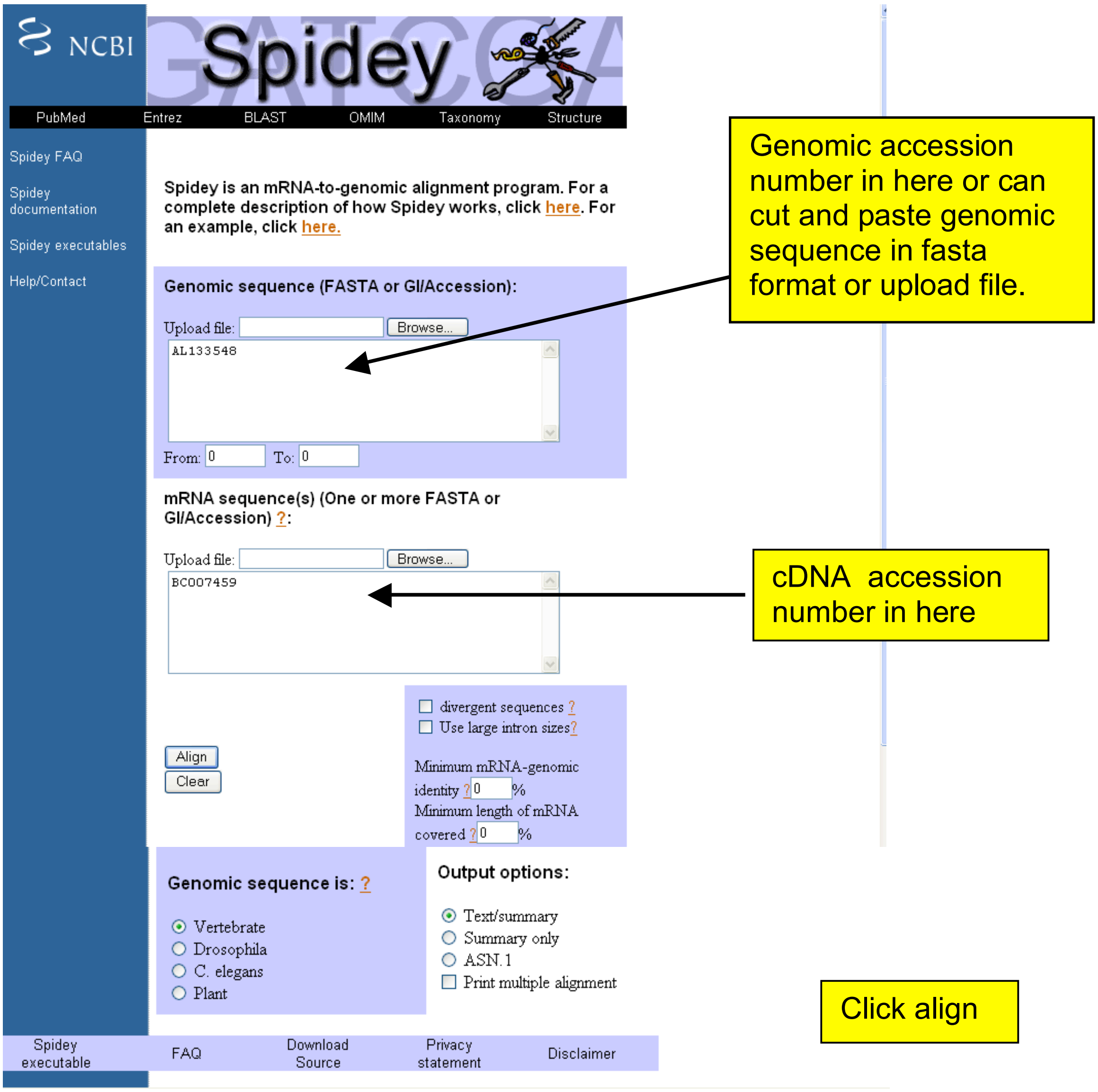
## BLASTP results



```
BLASTP 2.2.9 [May-01-2004]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.


RID: 1089369892-23111-86899017791.BLASTQ4


Query=
         (359 letters)



Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF excluding environmental samples
           1,921,851 sequences; 641,055,585 total letters




If you have any problems or questions with the results of this search
please refer to the BLAST FAQs

Taxonomy reports




                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

ref|NP_001161.1|  aquaporin 7; aquaporin adipose [Homo sapie...   695    0.0
dbj|BAC05693.1|  aquaporin adipose [Homo sapiens]                 691    0.0
emb|CAD13298.1|  bA251017.3 (similar to aquaporin 7) [Homo s...   635    0.0
ref|XP_376852.1|  similar to bA251017.3 (similar to aquapori...   583    e-165
ref|XP_376866.1|  similar to bA251017.3 (similar to aquapori...   583    e-165
ref|XP_372110.2|  similar to bA251017.3 (similar to aquapori...   562    e-159
ref|NP_031499.1|  aquaporin 7 [Mus musculus] >gi|9910621|sp|...   442    e-123
dbj|BAC36431.1|  unnamed protein product [Mus musculus]           442    e-123
ref|NP_062030.2|  aquaporin 7 [Rattus norvegicus] >gi|321724...   431    e-119
gb|AAH62701.1|  AQP7 protein [Homo sapiens]                       396    e-109
ref|NP_956204.1|  Unknown (protein for MGC:63700); wu:fj98f0...   300    4e-80
emb|CAG01413.1|  unnamed protein product [Tetraodon nigrovir...   280    3e-74
ref|NP_004916.1|  aquaporin 3 [Homo sapiens] >gi|2497938|sp|...   275    9e-73
gb|AAP36954.1|  Homo sapiens aquaporin 3 [synthetic construct]    275    9e-73
emb|CAG46822.1|  AQP3 [Homo sapiens]                              274    3e-72
ref|NP_113891.1|  aquaporin 3 [Rattus norvegicus] >gi|135196...   273    3e-72
dbj|BAA04559.1|  aquaporin 3 [Rattus rattus]                      273    3e-72
```

**2.** Align the genomic sequence AL133548 against a piece of cDNA sequence, BC007459, using Spidey.

**SPIDEY:**
**http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/**



Spidey is an mRNA-to-genomic alignment program. For a complete description of how Spidey works, click here. For an example, click here.

Genomic sequence (FASTA or GI/Accession):

Upload file: Browse...

AL133548

From: 0 To: 0

mRNA sequence(s) (One or more FASTA or GI/Accession) ?:

Upload file: Browse...

BC007459

□ divergent sequences ?
□ Use large intron sizes?

Minimum mRNA-genomic identity ? 0 %
Minimum length of mRNA covered ? 0 %

Genomic sequence is: ?

● Vertebrate
○ Drosophila
○ C. elegans
○ Plant

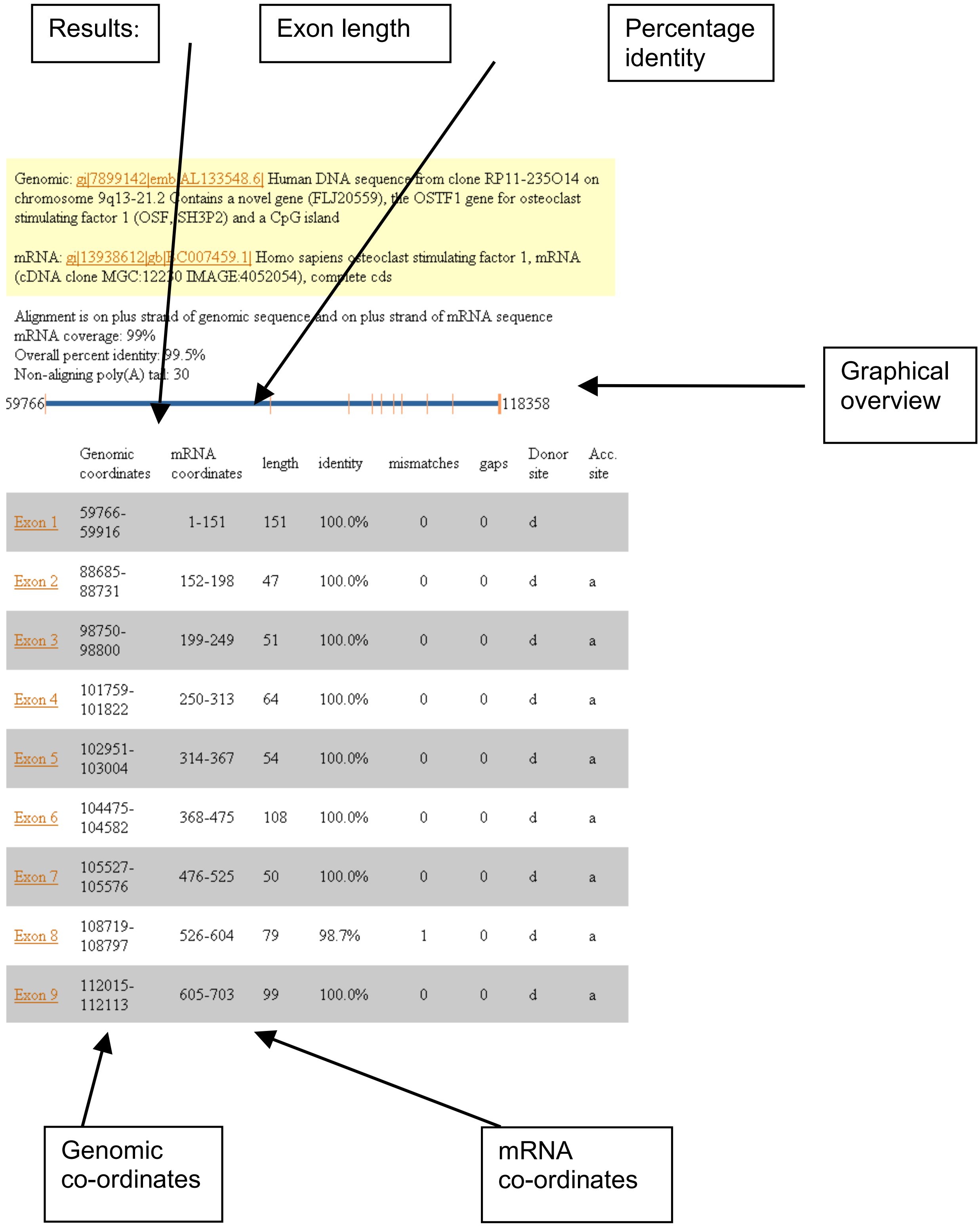
Output options:

● Text/summary
○ Summary only
○ ASN.1
□ Print multiple alignment

Genomic accession number in here or can cut and paste genomic sequence in fasta format or upload file.

cDNA accession number in here

Click align



Results:

Exon length

Percentage identity

Graphical overview

Genomic co-ordinates

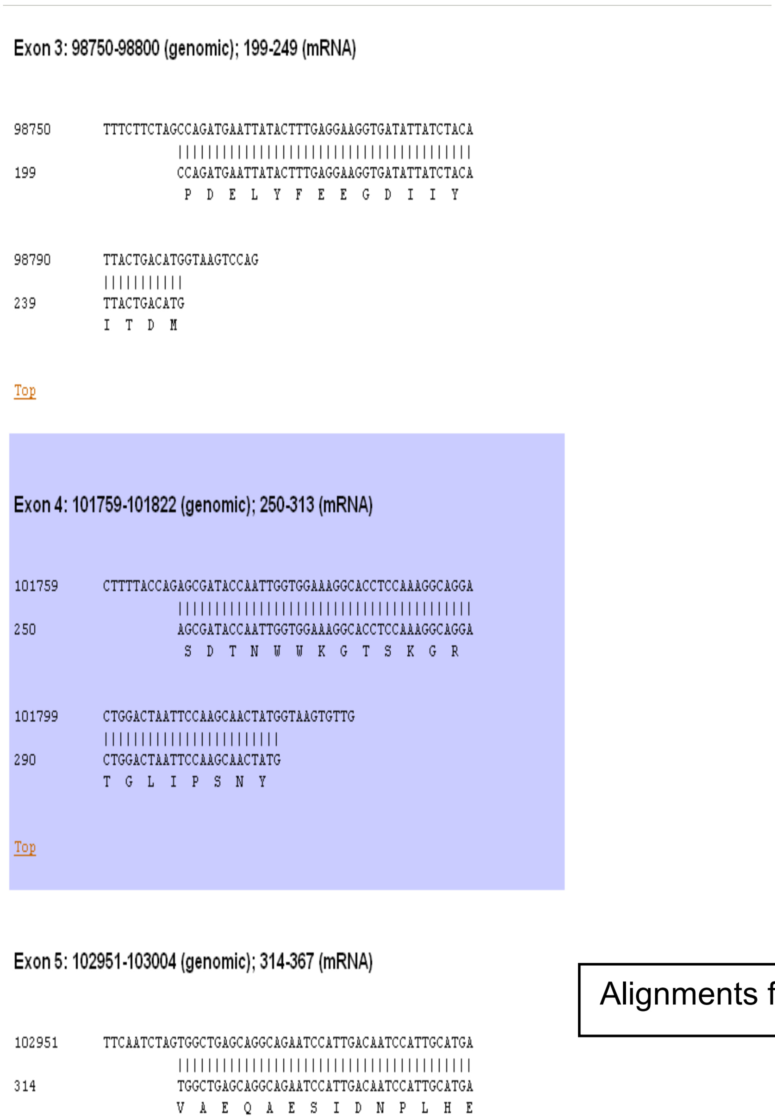
mRNA co-ordinates

Genomic: gi|7899142|emb|AL133548.6| Human DNA sequence from clone RP11-235O14 on chromosome 9q13-21.2 Contains a novel gene (FLJ20559), the OSTF1 gene for osteoclast stimulating factor 1 (OSF, SH3P2) and a CpG island

mRNA: gi|13938612|gb|BC007459.1| Homo sapiens osteoclast stimulating factor 1, mRNA (cDNA clone MGC:12230 IMAGE:4052054), complete cds

Alignment is on plus strand of genomic sequence and on plus strand of mRNA sequence
mRNA coverage: 99%
Overall percent identity: 99.5%
Non-aligning poly(A) tail: 30

| | Genomic coordinates | mRNA coordinates | length | identity | mismatches | gaps | Donor site | Acc. site |
|---|---|---|---|---|---|---|---|---|
| Exon 1 | 59766-59916 | 1-151 | 151 | 100.0% | 0 | 0 | d | |
| Exon 2 | 88685-88731 | 152-198 | 47 | 100.0% | 0 | 0 | d | a |
| Exon 3 | 98750-98800 | 199-249 | 51 | 100.0% | 0 | 0 | d | a |
| Exon 4 | 101759-101822 | 250-313 | 64 | 100.0% | 0 | 0 | d | a |
| Exon 5 | 102951-103004 | 314-367 | 54 | 100.0% | 0 | 0 | d | a |
| Exon 6 | 104475-104582 | 368-475 | 108 | 100.0% | 0 | 0 | d | a |
| Exon 7 | 105527-105576 | 476-525 | 50 | 100.0% | 0 | 0 | d | a |
| Exon 8 | 108719-108797 | 526-604 | 79 | 98.7% | 1 | 0 | d | a |
| Exon 9 | 112015-112113 | 605-703 | 99 | 100.0% | 0 | 0 | d | a |



Exon 3: 98750-98800 (genomic); 199-249 (mRNA)

```
98750       TTTCTTCTAGCCAGATGAATTATACTTTGAGGAAGGTGATATTATCTACA
                      ||||||||||||||||||||||||||||||||||||||||
199                   CCAGATGAATTATACTTTGAGGAAGGTGATATTATCTACA
                       P  D  E  L  Y  F  E  E  G  D  I  I  Y

98790       TTACTGACATGGTAAGTCCAG
            |||||||||||
239         TTACTGACATG
            I  T  D  M
```

Top

Exon 4: 101759-101822 (genomic); 250-313 (mRNA)

```
101759      CTTTTACCAGAGCGATACCAATTGGTGGAAAGGCACCTCCAAAGGCAGGA
                      ||||||||||||||||||||||||||||||||||||||||
250                   AGCGATACCAATTGGTGGAAAGGCACCTCCAAAGGCAGGA
                       S  D  T  N  W  W  K  G  T  S  K  G  R

101799      CTGGACTAATTCCAAGCAACTATGGTAAGTGTTG
            ||||||||||||||||||||||||
290         CTGGACTAATTCCAAGCAACTATG
            T  G  L  I  P  S  N  Y
```

Top

Exon 5: 102951-103004 (genomic); 314-367 (mRNA)

```
102951      TTCAATCTAGTGGCTGAGCAGGCAGAATCCATTGACAATCCATTGCATGA
                      ||||||||||||||||||||||||||||||||||||||||
314                   TGGCTGAGCAGGCAGAATCCATTGACAATCCATTGCATGA
                      V  A  E  Q  A  E  S  I  D  N  P  L  H  E
```

Alignments for each exon:

There are other cDNA to genomic alignment programs available:

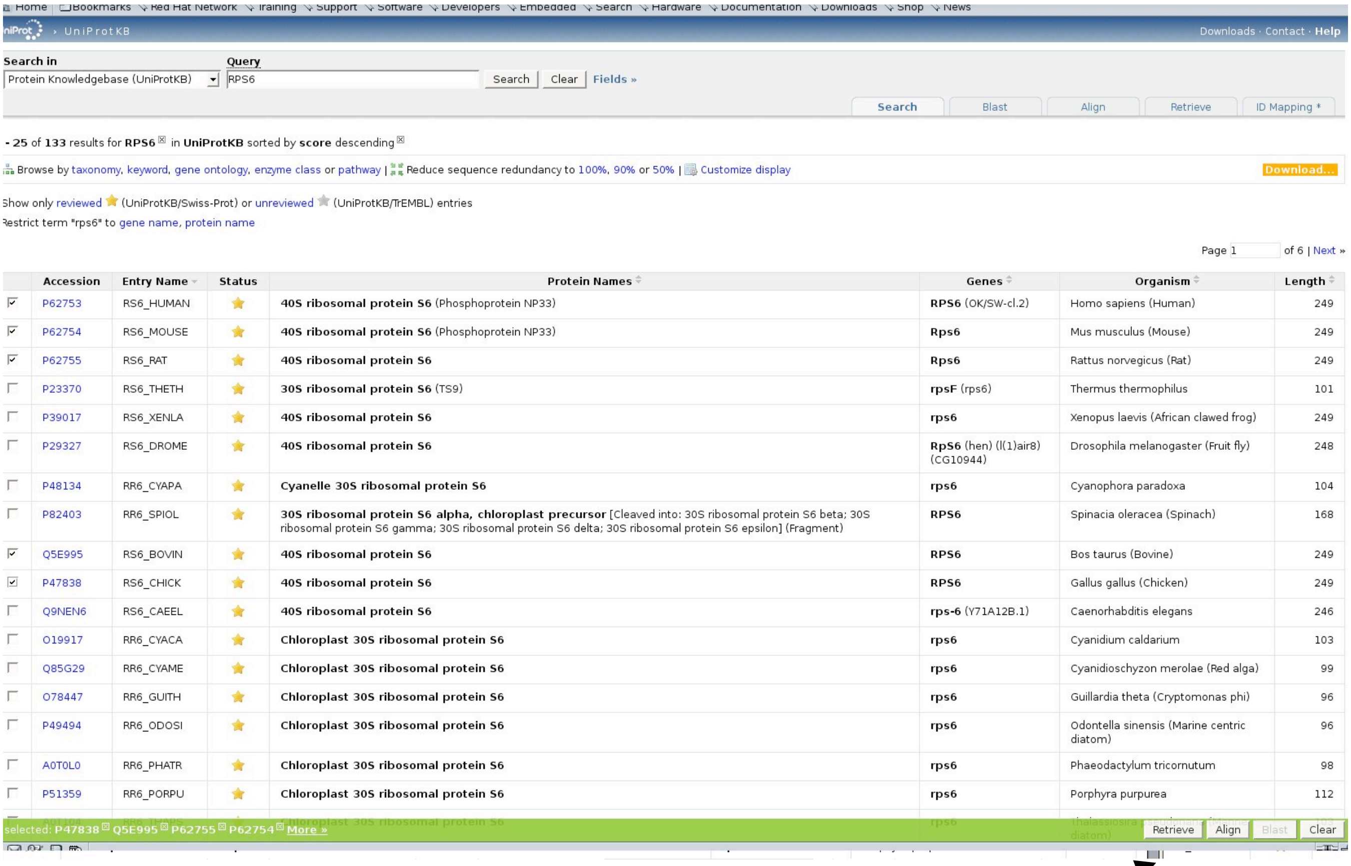
EST2GENOME
http://bioweb.pasteur.fr/seqanal/interfaces/est2genome.html

SIM4
http://pbil.univ-lyon.fr/sim4.html

**3.** Use clustalw to align similar sequences.

To fetch some sequences, used UniProtKB:

Search UniProtKB for RPS6.
Select a human, mouse, rat, cow and chicken sequence.


Home | Bookmarks Red Hat Network Training Support Software Developers Embedded Search Hardware Documentation Downloads Shop News
UniProt › UniProtKB
Downloads · Contact · Help
Search in
Protein Knowledgebase (UniProtKB)
Query
RPS6
Search
Clear
Fields »
Search
Blast
Align
Retrieve
ID Mapping *
Download...

- 25 of 133 results for **RPS6** ☒ in **UniProtKB** sorted by **score** descending ☒

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50% | Customize display

Show only reviewed ★ (UniProtKB/Swiss-Prot) or unreviewed ☆ (UniProtKB/TrEMBL) entries

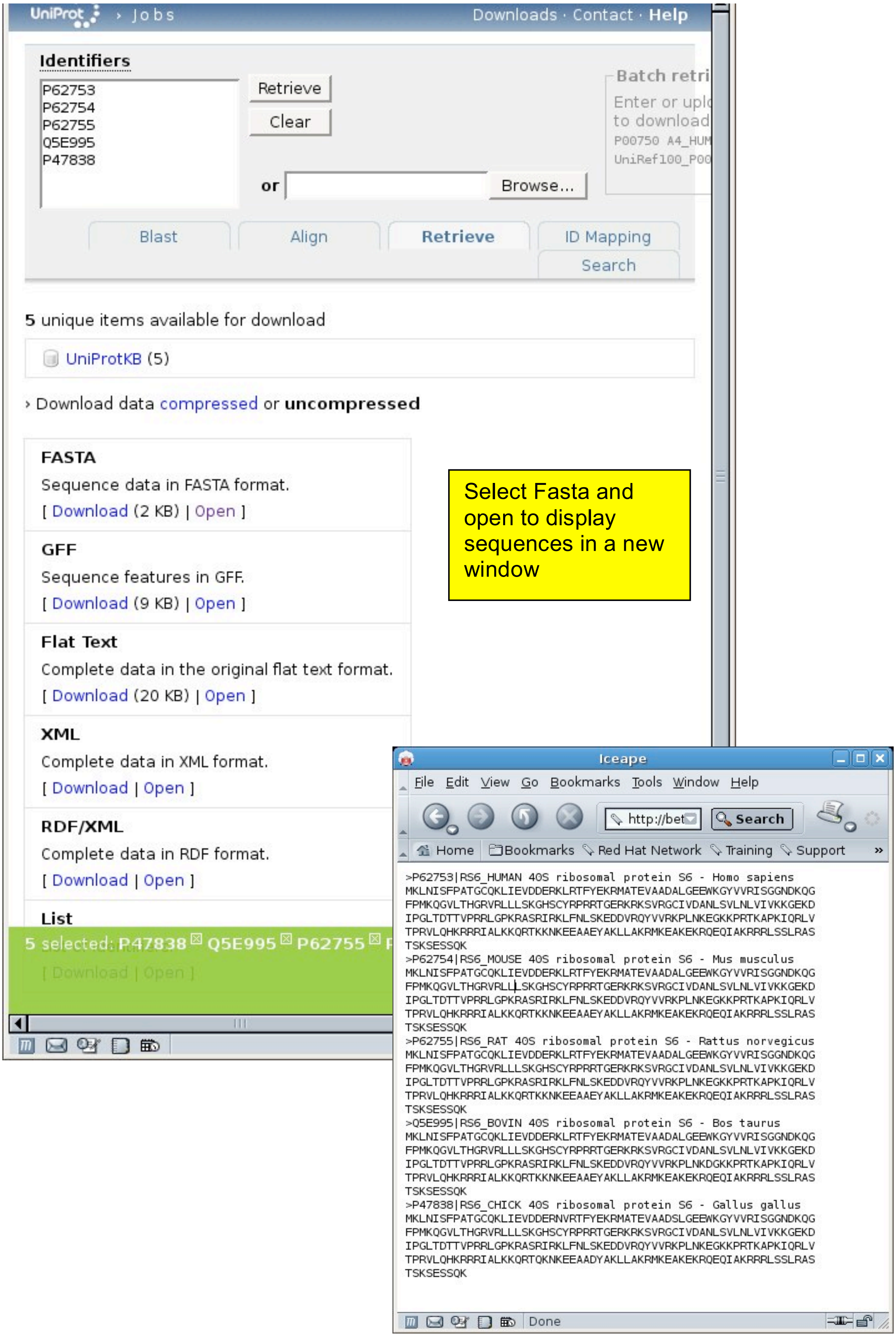
Restrict term "rps6" to gene name, protein name

Page 1 of 6 | Next »

| | Accession | Entry Name | Status | Protein Names | Genes | Organism | Length |
|---|---|---|---|---|---|---|---|
| ☑ | P62753 | RS6_HUMAN | ★ | **40S ribosomal protein S6** (Phosphoprotein NP33) | **RPS6** (OK/SW-cl.2) | Homo sapiens (Human) | 249 |
| ☑ | P62754 | RS6_MOUSE | ★ | **40S ribosomal protein S6** (Phosphoprotein NP33) | **Rps6** | Mus musculus (Mouse) | 249 |
| ☑ | P62755 | RS6_RAT | ★ | **40S ribosomal protein S6** | **Rps6** | Rattus norvegicus (Rat) | 249 |
| □ | P23370 | RS6_THETH | ★ | **30S ribosomal protein S6** (TS9) | **rpsF** (rps6) | Thermus thermophilus | 101 |
| □ | P39017 | RS6_XENLA | ★ | **40S ribosomal protein S6** | **rps6** | Xenopus laevis (African clawed frog) | 249 |
| □ | P29327 | RS6_DROME | ★ | **40S ribosomal protein S6** | **RpS6** (hen) (l(1)air8) (CG10944) | Drosophila melanogaster (Fruit fly) | 248 |
| □ | P48134 | RR6_CYAPA | ★ | **Cyanelle 30S ribosomal protein S6** | **rps6** | Cyanophora paradoxa | 104 |
| □ | P82403 | RR6_SPIOL | ★ | **30S ribosomal protein S6 alpha, chloroplast precursor** [Cleaved into: 30S ribosomal protein S6 beta; 30S ribosomal protein S6 gamma; 30S ribosomal protein S6 delta; 30S ribosomal protein S6 epsilon] (Fragment) | **RPS6** | Spinacia oleracea (Spinach) | 168 |
| ☑ | Q5E995 | RS6_BOVIN | ★ | **40S ribosomal protein S6** | **RPS6** | Bos taurus (Bovine) | 249 |
| ☑ | P47838 | RS6_CHICK | ★ | **40S ribosomal protein S6** | **RPS6** | Gallus gallus (Chicken) | 249 |
| □ | Q9NEN6 | RS6_CAEEL | ★ | **40S ribosomal protein S6** | **rps-6** (Y71A12B.1) | Caenorhabditis elegans | 246 |
| □ | O19917 | RR6_CYACA | ★ | **Chloroplast 30S ribosomal protein S6** | **rps6** | Cyanidium caldarium | 103 |
| □ | Q85G29 | RR6_CYAME | ★ | **Chloroplast 30S ribosomal protein S6** | **rps6** | Cyanidioschyzon merolae (Red alga) | 99 |
| □ | O78447 | RR6_GUITH | ★ | **Chloroplast 30S ribosomal protein S6** | **rps6** | Guillardia theta (Cryptomonas phi) | 96 |
| □ | P49494 | RR6_ODOSI | ★ | **Chloroplast 30S ribosomal protein S6** | **rps6** | Odontella sinensis (Marine centric diatom) | 96 |
| □ | A0T0L0 | RR6_PHATR | ★ | **Chloroplast 30S ribosomal protein S6** | **rps6** | Phaeodactylum tricornutum | 98 |
| □ | P51359 | RR6_PORPU | ★ | **Chloroplast 30S ribosomal protein S6** | **rps6** | Porphyra purpurea | 112 |

selected: **P47838** ☒ **Q5E995** ☒ **P62755** ☒ **P62754** ☒ More »

Click on retrieve



**Identifiers**

P62753
P62754
P62755
Q5E995
P47838

5 unique items available for download

UniProtKB (5)

› Download data compressed or **uncompressed**

**FASTA**
Sequence data in FASTA format.
[ Download (2 KB) | Open ]

**GFF**
Sequence features in GFF.
[ Download (9 KB) | Open ]

**Flat Text**
Complete data in the original flat text format.
[ Download (20 KB) | Open ]

**XML**
Complete data in XML format.
[ Download | Open ]

**RDF/XML**
Complete data in RDF format.
[ Download | Open ]

**List**

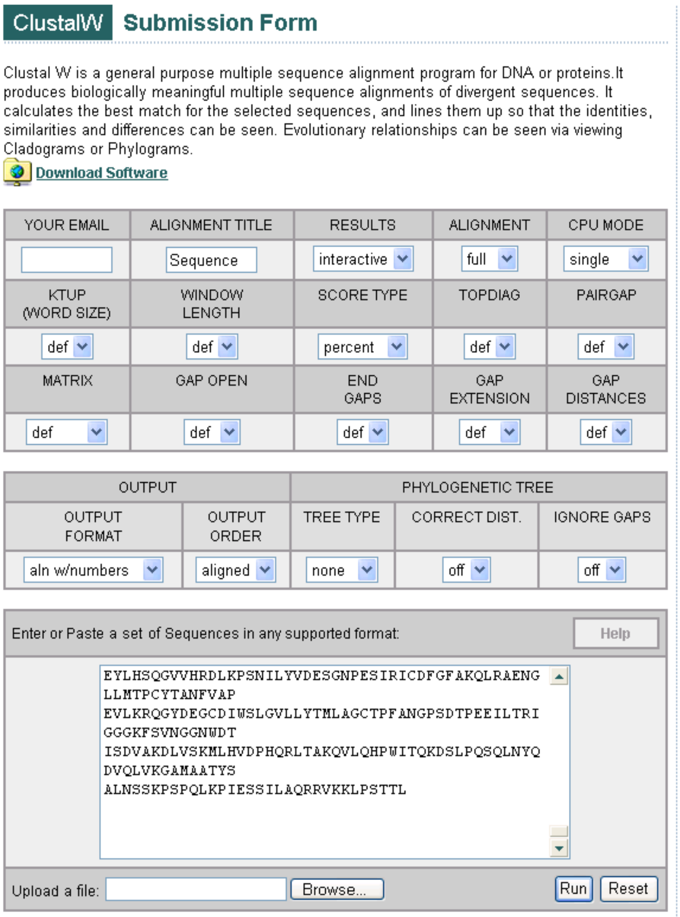
5 selected: P47838 ☒ Q5E995 ☒ P62755 ☒ P

[ Download | Open ]

Select Fasta and open to display sequences in a new window

```
>P62753|RS6_HUMAN 40S ribosomal protein S6 - Homo sapiens
MKLNISFPATGCQKLIEVDDERKLRTFYEKRMATEVAADALGEEWKGYVVRISGGNDKQG
FPMKQGVLTHGRVRLLLSKGHSCYRPRRTGERKRKSVRGCIVDANLSVLNLVIVKKGEKD
IPGLTDTTVPRRLGPKRASRIRKLFNLSKEDDVRQYVVRKPLNKEGKKPRTKAPKIQRLV
TPRVLQHKRRRIALKKQRTKKNKEEAAEYAKLLAKRMKEAKEKRQEQIAKRRRLSSLRAS
TSKSESSQK
>P62754|RS6_MOUSE 40S ribosomal protein S6 - Mus musculus
MKLNISFPATGCQKLIEVDDERKLRTFYEKRMATEVAADALGEEWKGYVVRISGGNDKQG
FPMKQGVLTHGRVRLLLSKGHSCYRPRRTGERKRKSVRGCIVDANLSVLNLVIVKKGEKD
IPGLTDTTVPRRLGPKRASRIRKLFNLSKEDDVRQYVVRKPLNKEGKKPRTKAPKIQRLV
TPRVLQHKRRRIALKKQRTKKNKEEAAEYAKLLAKRMKEAKEKRQEQIAKRRRLSSLRAS
TSKSESSQK
>P62755|RS6_RAT 40S ribosomal protein S6 - Rattus norvegicus
MKLNISFPATGCQKLIEVDDERKLRTFYEKRMATEVAADALGEEWKGYVVRISGGNDKQG
FPMKQGVLTHGRVRLLLSKGHSCYRPRRTGERKRKSVRGCIVDANLSVLNLVIVKKGEKD
IPGLTDTTVPRRLGPKRASRIRKLFNLSKEDDVRQYVVRKPLNKEGKKPRTKAPKIQRLV
TPRVLQHKRRRIALKKQRTKKNKEEAAEYAKLLAKRMKEAKEKRQEQIAKRRRLSSLRAS
TSKSESSQK
>Q5E995|RS6_BOVIN 40S ribosomal protein S6 - Bos taurus
MKLNISFPATGCQKLIEVDDERKLRTFYEKRMATEVAADALGEEWKGYVVRISGGNDKQG
FPMKQGVLTHGRVRLLLSKGHSCYRPRRTGERKRKSVRGCIVDANLSVLNLVIVKKGEKD
IPGLTDTTVPRRLGPKRASRIRKLFNLSKEDDVRQYVVRKPLNKDGKKPRTKAPKIQRLV
TPRVLQHKRRRIALKKQRTKKNKEEAAEYAKLLAKRMKEAKEKRQEQIAKRRRLSSLRAS
TSKSESSQK
>P47838|RS6_CHICK 40S ribosomal protein S6 - Gallus gallus
MKLNISFPATGCQKLIEVDDERNVRTFYEKRMATEVAADSLGEEWKGYVVRISGGNDKQG
FPMKQGVLTHGRVRLLLSKGHSCYRPRRTGERKRKSVRGCIVDANLSVLNLVIVKKGEKD
IPGLTDTTVPRRLGPKRASRIRKLFNLSKEDDVRQYVVRKPLNKEGKKPRTKAPKIQRLV
TPRVLQHKRRRIALKKQRTQKNKEEAADYAKLLAKRMKEAKEKRQEQIAKRRRLSSLRAS
TSKSESSQK
```

http://www.ebi.ac.uk/clustalw/index.htm



## ClustalW Submission Form

Clustal W is a general purpose multiple sequence alignment program for DNA or proteins.It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms.

| | | | | |
|---|---|---|---|---|
| YOUR EMAIL | ALIGNMENT TITLE | RESULTS | ALIGNMENT | CPU MODE |
| | Sequence | interactive | full | single |
| KTUP (WORD SIZE) | WINDOW LENGTH | SCORE TYPE | TOPDIAG | PAIRGAP |
| def | def | percent | def | def |
| MATRIX | GAP OPEN | END GAPS | GAP EXTENSION | GAP DISTANCES |
| def | def | def | def | def |

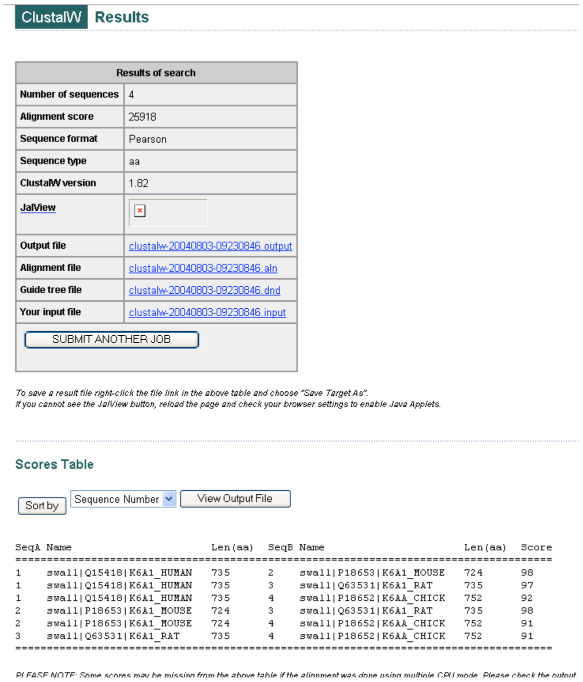
| OUTPUT | | PHYLOGENETIC TREE | | |
|---|---|---|---|---|
| OUTPUT FORMAT | OUTPUT ORDER | TREE TYPE | CORRECT DIST. | IGNORE GAPS |
| aln w/numbers | aligned | none | off | off |

Enter or Paste a set of Sequences in any supported format:

```
EYLHSQGVVHRDLKPSNILYVDESGNPESIRICDFGFAKQLRAENG
LLMTPCYTANFVAP
EVLKRQGYDEGCDIWSLGVLLYTMLAGCTPFANGPSDTPEEILTRI
GGGKFSVNGGNWDT
ISDVAKDLVSKMLHVDPHQRLTAKQVLQHPWITQKDSLPQSQLNYQ
DVQLVKGAMAATYS
ALNSSKPSPQLKPIESSILAQRRVKKLPSTTL
```

Upload a file:

Go into ClustalW. Copy and paste sequences and click on run…….



## ClustalW Results

| Results of search | |
|---|---|
| Number of sequences | 4 |
| Alignment score | 25918 |
| Sequence format | Pearson |
| Sequence type | aa |
| ClustalW version | 1.82 |
| JalView | ☒ |
| Output file | clustalw-20040803-09230846.output |
| Alignment file | clustalw-20040803-09230846.aln |
| Guide tree file | clustalw-20040803-09230846.dnd |
| Your input file | clustalw-20040803-09230846.input |

*To save a result file right-click the file link in the above table and choose "Save Target As".*
*If you cannot see the JalView button, reload the page and check your browser settings to enable Java Applets.*

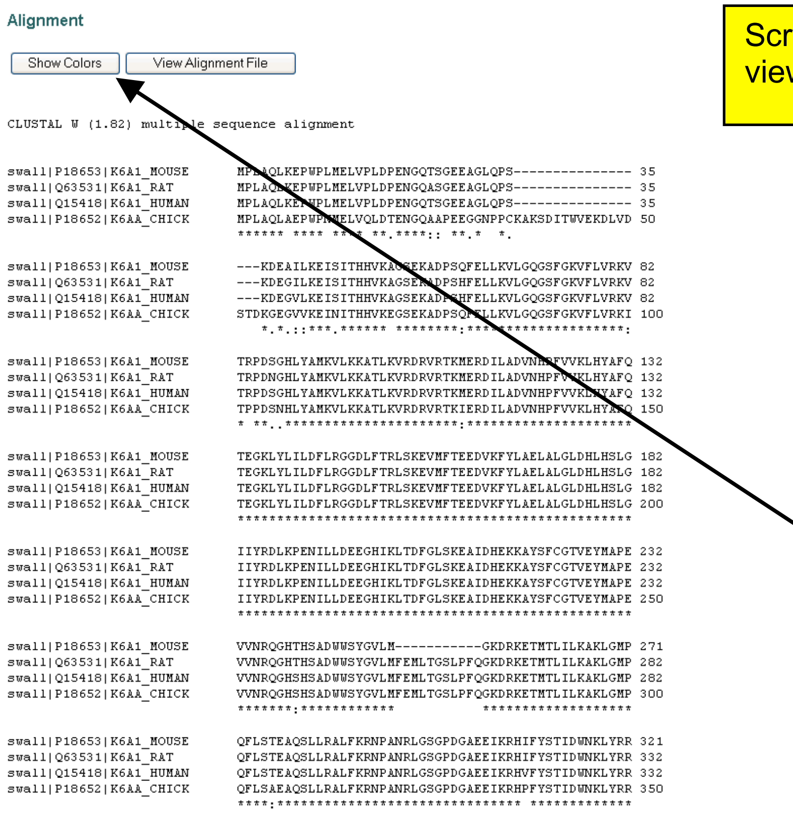
### Scores Table

```
SeqA Name                   Len(aa)  SeqB Name                   Len(aa)  Score
===============================================================================
1    swall|Q15418|K6A1_HUMAN   735      2    swall|P18653|K6A1_MOUSE   724     98
1    swall|Q15418|K6A1_HUMAN   735      3    swall|Q63531|K6A1_RAT     735     97
1    swall|Q15418|K6A1_HUMAN   735      4    swall|P18652|K6AA_CHICK   752     92
2    swall|P18653|K6A1_MOUSE   724      3    swall|Q63531|K6A1_RAT     735     98
2    swall|P18653|K6A1_MOUSE   724      4    swall|P18652|K6AA_CHICK   752     91
3    swall|Q63531|K6A1_RAT     735      4    swall|P18652|K6AA_CHICK   752     91
===============================================================================
```

*PLEASE NOTE: Some scores may be missing from the above table if the alignment was done using multiple CPU mode. Please check the output*

Top of results page

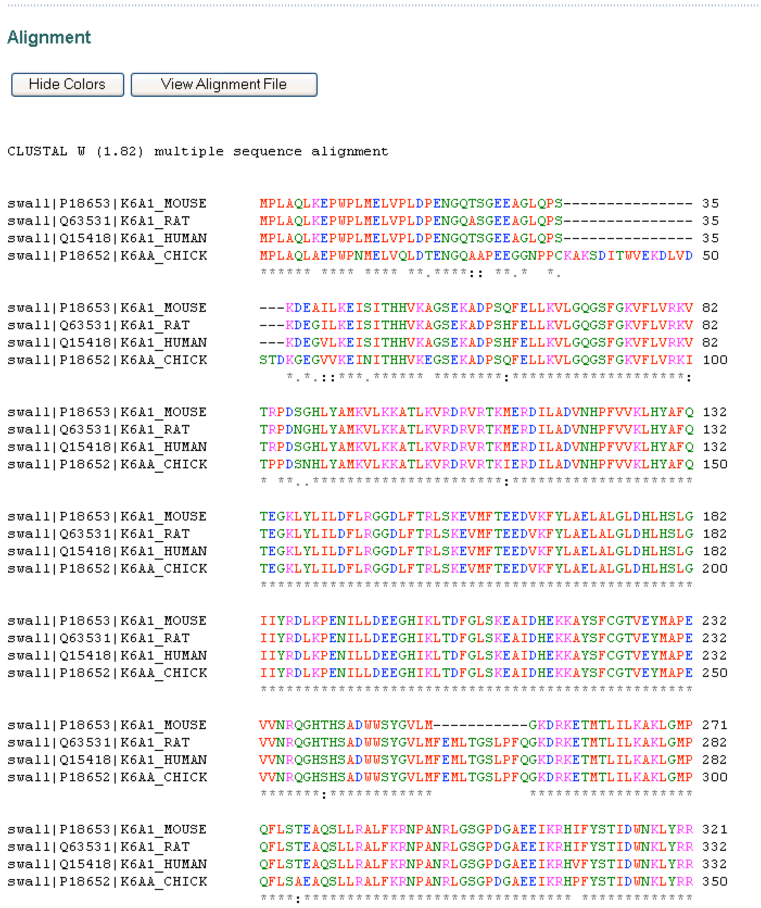
Scroll down to view alignment



Alignment

```
CLUSTAL W (1.82) multiple sequence alignment


swall|P18653|K6A1_MOUSE      MPLAQLKEPWPLMELVPLDPENGQTSGEEAGLQPS--------------- 35
swall|Q63531|K6A1_RAT        MPLAQLKEPWPLMELVPLDPENGQASGEEAGLQPS--------------- 35
swall|Q15418|K6A1_HUMAN      MPLAQLKEPWPLMELVPLDPENGQTSGEEAGLQPS--------------- 35
swall|P18652|K6AA_CHICK      MPLAQLAEPWPNMELVQLDTENGQAAPEEGGNPPCKAKSDITWVEKDLVD 50
                             ****** **** **** **.****:: **.*  *.

swall|P18653|K6A1_MOUSE      ---KDEAILKEISITHHVKAGSEKADPSQFELLKVLGQGSFGKVFLVRKV 82
swall|Q63531|K6A1_RAT        ---KDEGILKEISITHHVKAGSEKADPSHFELLKVLGQGSFGKVFLVRKV 82
swall|Q15418|K6A1_HUMAN      ---KDEGVLKEISITHHVKAGSEKADPSHFELLKVLGQGSFGKVFLVRKV 82
swall|P18652|K6AA_CHICK      STDKGEGVVKEINITHHVKEGSEKADPSQFELLKVLGQGSFGKVFLVRKI 100
                                *.*.::***.****** ********:********************:

swall|P18653|K6A1_MOUSE      TRPDSGHLYAMKVLKKATLKVRDRVRTKMERDILADVNHPFVVKLHYAFQ 132
swall|Q63531|K6A1_RAT        TRPDNGHLYAMKVLKKATLKVRDRVRTKMERDILADVNHPFVVKLHYAFQ 132
swall|Q15418|K6A1_HUMAN      TRPDSGHLYAMKVLKKATLKVRDRVRTKMERDILADVNHPFVVKLHYAFQ 132
swall|P18652|K6AA_CHICK      TPPDSNHLYAMKVLKKATLKVRDRVRTKIERDILADVNHPFVVKLHYAFQ 150
                             * **..**********************:*********************

swall|P18653|K6A1_MOUSE      TEGKLYLILDFLRGGDLFTRLSKEVMFTEEDVKFYLAELALGLDHLHSLG 182
swall|Q63531|K6A1_RAT        TEGKLYLILDFLRGGDLFTRLSKEVMFTEEDVKFYLAELALGLDHLHSLG 182
swall|Q15418|K6A1_HUMAN      TEGKLYLILDFLRGGDLFTRLSKEVMFTEEDVKFYLAELALGLDHLHSLG 182
swall|P18652|K6AA_CHICK      TEGKLYLILDFLRGGDLFTRLSKEVMFTEEDVKFYLAELALGLDHLHSLG 200
                             **************************************************

swall|P18653|K6A1_MOUSE      IIYRDLKPENILLDEEGHIKLTDFGLSKEAIDHEKKAYSFCGTVEYMAPE 232
swall|Q63531|K6A1_RAT        IIYRDLKPENILLDEEGHIKLTDFGLSKEAIDHEKKAYSFCGTVEYMAPE 232
swall|Q15418|K6A1_HUMAN      IIYRDLKPENILLDEEGHIKLTDFGLSKEAIDHEKKAYSFCGTVEYMAPE 232
swall|P18652|K6AA_CHICK      IIYRDLKPENILLDEEGHIKLTDFGLSKEAIDHEKKAYSFCGTVEYMAPE 250
                             **************************************************

swall|P18653|K6A1_MOUSE      VVNRQGHTHSADWWSYGVLM-----------GKDRKETMTLILKAKLGMP 271
swall|Q63531|K6A1_RAT        VVNRQGHTHSADWWSYGVLMFEMLTGSLPFQGKDRKETMTLILKAKLGMP 282
swall|Q15418|K6A1_HUMAN      VVNRQGHSHSADWWSYGVLMFEMLTGSLPFQGKDRKETMTLILKAKLGMP 282
swall|P18652|K6AA_CHICK      VVNRQGHSHSADWWSYGVLMFEMLTGSLPFQGKDRKETMTLILKAKLGMP 300
                             *******:************           *******************

swall|P18653|K6A1_MOUSE      QFLSTEAQSLLRALFKRNPANRLGSGPDGAEEIKRHIFYSTIDWNKLYRR 321
swall|Q63531|K6A1_RAT        QFLSTEAQSLLRALFKRNPANRLGSGPDGAEEIKRHIFYSTIDWNKLYRR 332
swall|Q15418|K6A1_HUMAN      QFLSTEAQSLLRALFKRNPANRLGSGPDGAEEIKRHVFYSTIDWNKLYRR 332
swall|P18652|K6AA_CHICK      QFLSAEAQSLLRALFKRNPANRLGSGPDGAEEIKRHPFYSTIDWNKLYRR 350
                             ****:******************************* *************
```

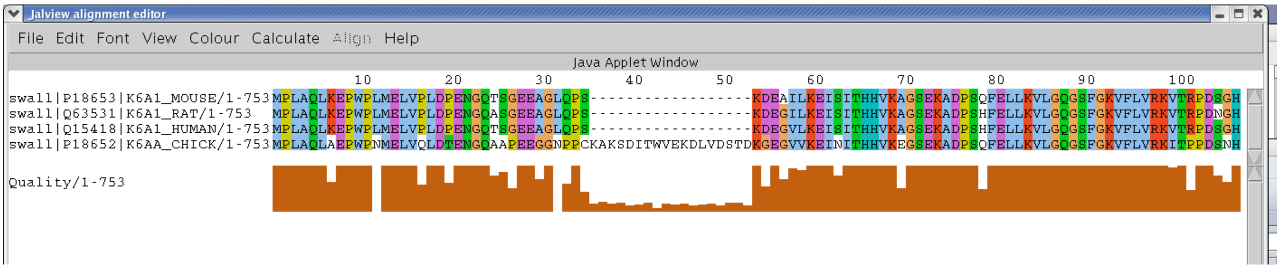
Click on show colours



Alignment

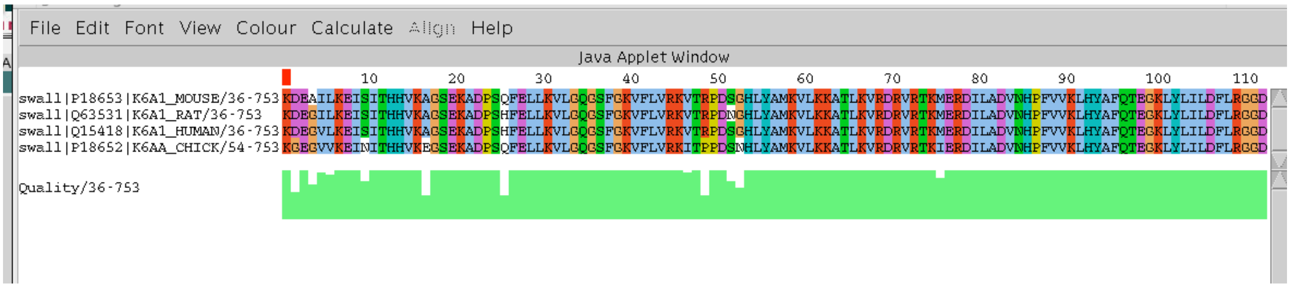
```
CLUSTAL W (1.82) multiple sequence alignment


swall|P18653|K6A1_MOUSE      MPLAQLKEPWPLMELVPLDPENGQTSGEEAGLQPS--------------- 35
swall|Q63531|K6A1_RAT        MPLAQLKEPWPLMELVPLDPENGQASGEEAGLQPS--------------- 35
swall|Q15418|K6A1_HUMAN      MPLAQLKEPWPLMELVPLDPENGQTSGEEAGLQPS--------------- 35
swall|P18652|K6AA_CHICK      MPLAQLAEPWPNMELVQLDTENGQAAPEEGGNPPCKAKSDITWVEKDLVD 50
                             ****** **** **** **.****:: **.*  *.

swall|P18653|K6A1_MOUSE      ---KDEAILKEISITHHVKAGSEKADPSQFELLKVLGQGSFGKVFLVRKV 82
swall|Q63531|K6A1_RAT        ---KDEGILKEISITHHVKAGSEKADPSHFELLKVLGQGSFGKVFLVRKV 82
swall|Q15418|K6A1_HUMAN      ---KDEGVLKEISITHHVKAGSEKADPSHFELLKVLGQGSFGKVFLVRKV 82
swall|P18652|K6AA_CHICK      STDKGEGVVKEINITHHVKEGSEKADPSQFELLKVLGQGSFGKVFLVRKI 100
                                *.*.::***.****** ********:********************:

swall|P18653|K6A1_MOUSE      TRPDSGHLYAMKVLKKATLKVRDRVRTKMERDILADVNHPFVVKLHYAFQ 132
swall|Q63531|K6A1_RAT        TRPDNGHLYAMKVLKKATLKVRDRVRTKMERDILADVNHPFVVKLHYAFQ 132
swall|Q15418|K6A1_HUMAN      TRPDSGHLYAMKVLKKATLKVRDRVRTKMERDILADVNHPFVVKLHYAFQ 132
swall|P18652|K6AA_CHICK      TPPDSNHLYAMKVLKKATLKVRDRVRTKIERDILADVNHPFVVKLHYAFQ 150
                             * **..**********************:*********************

swall|P18653|K6A1_MOUSE      TEGKLYLILDFLRGGDLFTRLSKEVMFTEEDVKFYLAELALGLDHLHSLG 182
swall|Q63531|K6A1_RAT        TEGKLYLILDFLRGGDLFTRLSKEVMFTEEDVKFYLAELALGLDHLHSLG 182
swall|Q15418|K6A1_HUMAN      TEGKLYLILDFLRGGDLFTRLSKEVMFTEEDVKFYLAELALGLDHLHSLG 182
swall|P18652|K6AA_CHICK      TEGKLYLILDFLRGGDLFTRLSKEVMFTEEDVKFYLAELALGLDHLHSLG 200
                             **************************************************

swall|P18653|K6A1_MOUSE      IIYRDLKPENILLDEEGHIKLTDFGLSKEAIDHEKKAYSFCGTVEYMAPE 232
swall|Q63531|K6A1_RAT        IIYRDLKPENILLDEEGHIKLTDFGLSKEAIDHEKKAYSFCGTVEYMAPE 232
swall|Q15418|K6A1_HUMAN      IIYRDLKPENILLDEEGHIKLTDFGLSKEAIDHEKKAYSFCGTVEYMAPE 232
swall|P18652|K6AA_CHICK      IIYRDLKPENILLDEEGHIKLTDFGLSKEAIDHEKKAYSFCGTVEYMAPE 250
                             **************************************************

swall|P18653|K6A1_MOUSE      VVNRQGHTHSADWWSYGVLM-----------GKDRKETMTLILKAKLGMP 271
swall|Q63531|K6A1_RAT        VVNRQGHTHSADWWSYGVLMFEMLTGSLPFQGKDRKETMTLILKAKLGMP 282
swall|Q15418|K6A1_HUMAN      VVNRQGHSHSADWWSYGVLMFEMLTGSLPFQGKDRKETMTLILKAKLGMP 282
swall|P18652|K6AA_CHICK      VVNRQGHSHSADWWSYGVLMFEMLTGSLPFQGKDRKETMTLILKAKLGMP 300
                             *******:************           *******************

swall|P18653|K6A1_MOUSE      QFLSTEAQSLLRALFKRNPANRLGSGPDGAEEIKRHIFYSTIDWNKLYRR 321
swall|Q63531|K6A1_RAT        QFLSTEAQSLLRALFKRNPANRLGSGPDGAEEIKRHIFYSTIDWNKLYRR 332
swall|Q15418|K6A1_HUMAN      QFLSTEAQSLLRALFKRNPANRLGSGPDGAEEIKRHVFYSTIDWNKLYRR 332
swall|P18652|K6AA_CHICK      QFLSAEAQSLLRALFKRNPANRLGSGPDGAEEIKRHPFYSTIDWNKLYRR 350
                             ****:******************************* *************
```

Click on Jalview to view coloured, shaded alignment.



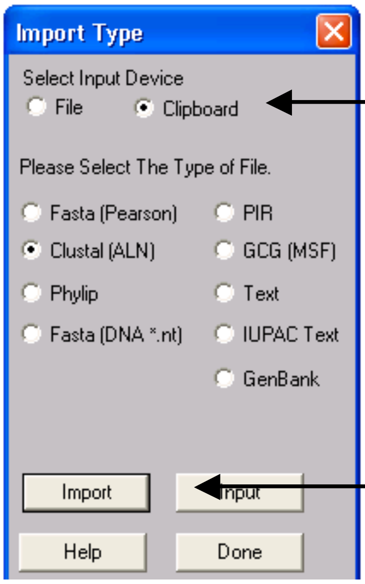
```
                               10        20        30        40        50        60        70        80        90        100
swall|P18653|K6A1_MOUSE/1-753 MPLAQLKEPWPLMELVPLDPENGQTSGEEAGLQPS------------------KDEAILKEISITHHVKAGSEKADPSQFELLKVLGQGSFGKVFLVRKVTRPDSGH
swall|Q63531|K6A1_RAT/1-753   MPLAQLKEPWPLMELVPLDPENGQASGEEAGLQPS------------------KDEGILKEISITHHVKAGSEKADPSHFELLKVLGQGSFGKVFLVRKVTRPDNGH
swall|Q15418|K6A1_HUMAN/1-753 MPLAQLKEPWPLMELVPLDPENGQTSGEEAGLQPS------------------KDEGVLKEISITHHVKAGSEKADPSHFELLKVLGQGSFGKVFLVRKVTRPDSGH
swall|P18652|K6AA_CHICK/1-753 MPLAQLAEPWPNMELVQLDTENGQAAPEEGGNPPCKAKSDITWVEKDLVDSTDKGEGVVKEINITHHVKEGSEKADPSQFELLKVLGQGSFGKVFLVRKITPPDSNH

Quality/1-753
```

To edit the alignment:
Click above the amino acid in the sequence to select the column. From the edit menu select remove sequence to left of selected column………see below



```
                                   10        20        30        40        50        60        70        80        90        100       110
swall|P18653|K6A1_MOUSE/36-753 KDEAILKEISITHHVKAGSEKADPSQFELLKVLGQGSFGKVFLVRKVTRPDSGHLYAMKVLKKATLKVRDRVRTKMERDILADVNHPFVVKLHYAFQTEGKLYLILDFLRGGD
swall|Q63531|K6A1_RAT/36-753   KDEGILKEISITHHVKAGSEKADPSHFELLKVLGQGSFGKVFLVRKVTRPDNGHLYAMKVLKKATLKVRDRVRTKMERDILADVNHPFVVKLHYAFQTEGKLYLILDFLRGGD
swall|Q15418|K6A1_HUMAN/36-753 KDEGVLKEISITHHVKAGSEKADPSHFELLKVLGQGSFGKVFLVRKVTRPDSGHLYAMKVLKKATLKVRDRVRTKMERDILADVNHPFVVKLHYAFQTEGKLYLILDFLRGGD
swall|P18652|K6AA_CHICK/54-753 KGEGVVKEINITHHVKEGSEKADPSQFELLKVLGQGSFGKVFLVRKITPPDSNHLYAMKVLKKATLKVRDRVRTKIERDILADVNHPFVVKLHYAFQTEGKLYLILDFLRGGD

Quality/36-753
```

Create a coloured and shaded figure with GeneDoc (Not available for Mac):
GeneDoc can be downloaded from here: **http://www.psc.edu/biomed/genedoc/**



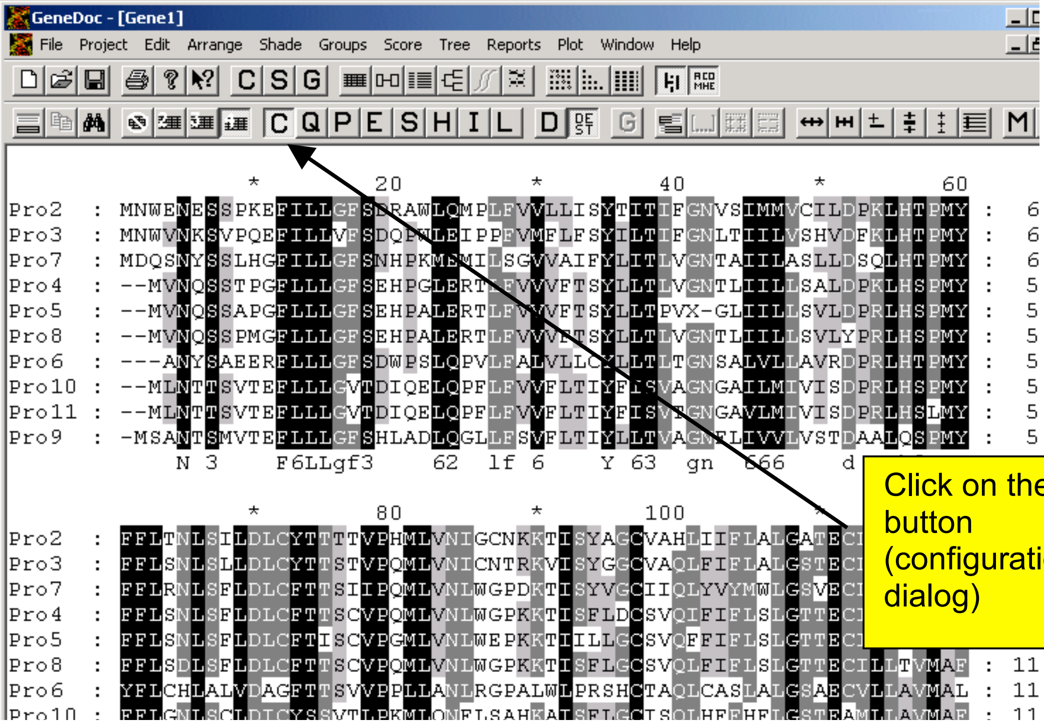
Select Input Device
○ File ● Clipboard

Please Select The Type of File.
○ Fasta (Pearson) ○ PIR
● Clustal (ALN) ○ GCG (MSF)
○ Phylip ○ Text
○ Fasta (DNA *.nt) ○ IUPAC Text
○ GenBank

Copy the aln format alignment from clustalw (**make sure that you include the top line of text**), then open up GeneDoc and select file, import. Select clipboard as the input device and Clustal for the file type.

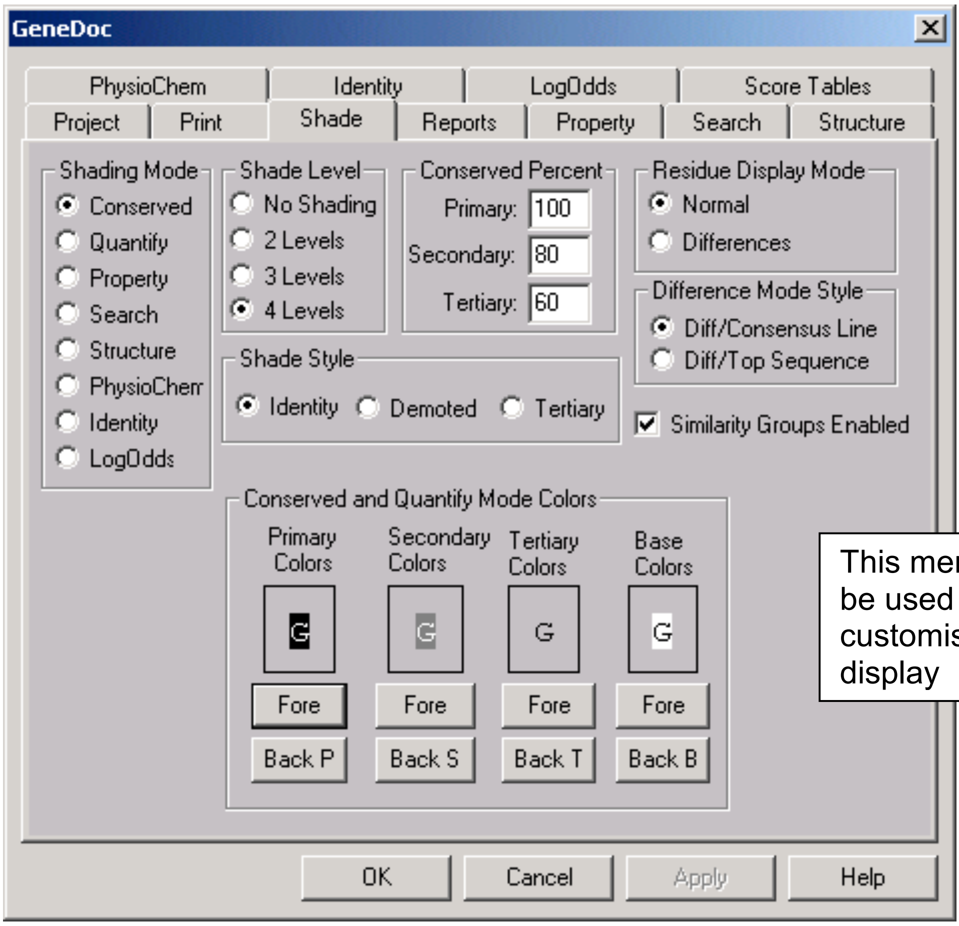
Click on import, then done

**The imported file will be displayed in black and white**



```
                       *        20         *        40         *        60
Pro2   : MNWENESSPKEFILLGFSDRAWLQMPLFVVLLISYTITIFGNVSIMMVCILDPKLHTPMY :   6
Pro3   : MNWVNKSVPQEFILLVFSDQPWLEIPPFVMFLFSYILTIFGNLTIILVSHVDFKLHTPMY :   6
Pro7   : MDQSNYSSLHGFILLGFSNHPKMEMILSGVVAIFYLITLVGNTAIILASLLDSQLHTPMY :   6
Pro4   : --MVNQSSTPGFLLLGFSEHPGLERTLFVVVFTSYLLTLVGNTLIILLSALDPKLHSPMY :   5
Pro5   : --MVNQSSAPGFLLLGFSEHPALERTLFVVVFTSYLLTPVX-GLIILLSVLDPRLHSPMY :   5
Pro8   : --MVNQSSPMGFLLLGFSEHPALERTLFVVVFTSYLLTLVGNTLIILLSVLYPRLHSPMY :   5
Pro6   : ---ANYSAEERFLLLGFSDWPSLQPVLFALVLLCYLLTLTGNSALVLLAVRDPRLHTPMY :   5
Pro10  : --MLNTTSVTEFLLLGVTDIQELQPFLFVVFLTIYFISVAGNGAILMIVISDPRLHSPMY :   5
Pro11  : --MLNTTSVTEFLLLGVTDIQELQPFLFVVFLTIYFISVTGNGAVLMIVISDPRLHSLMY :   5
Pro9   : -MSANTSMVTEFLLLGFSHLADLQGLLFSVFLTIYLLTVAGNFLIVVLVSTDAALQSPMY :   5
                 N 3    F6LLgf3    62  lf 6     Y 63  gn  666    d

                       *        80         *       100         *
Pro2   : FFLTNLSILDLCYTTTTVPHMLVNIGCNKKTISYAGCVAHLIIFLALGATECI
Pro3   : FFLSNLSLLDLCYTTSTVPQMLVNICNTRKVISYGGCVAQLFIFLALGSTECI
Pro7   : FFLRNLSFLDLCFTTSIIPQMLVNLWGPDKTISYVGCIIQLYVYMWLGSVECI
Pro4   : FFLSNLSFLDLCFTTSCVPQMLVNLWGPKKTISFLDCSVQIFIFLSLGTTECI
Pro5   : FFLSNLSFLDLCFTISCVPGMLVNLWEPKKTIILLGCSVQFFIFLSLGTTECI
Pro8   : FFLSDLSFLDLCFTTSCVPQMLVNLWGPKKTISFLGCSVQLFIFLSLGTTECILLTVMAF :  11
Pro6   : YFLCHLALVDAGFTTSVVPPLLANLRGPALWLPRSHCTAQLCASLALGSAECVLLAVMAL :  11
Pro10  : FFLGNLSCLDICYSSVTLPKMLQNFLSAHKALSFLGCTSQLHFFHFLGSTEAMLLAVMAF :  11
```

Click on the C button (configuration dialog)


GeneDoc
PhysioChem | Identity | LogOdds | Score Tables
Project | Print | Shade | Reports | Property | Search | Structure
Fore | Fore | Fore | Fore
Back P | Back S | Back T | Back B
OK | Cancel | Apply | Help

Shading Mode
● Conserved
○ Quantify
○ Property
○ Search
○ Structure
○ PhysioChem
○ Identity
○ LogOdds

Shade Level
○ No Shading
○ 2 Levels
○ 3 Levels
● 4 Levels

Conserved Percent
Primary: 100
Secondary: 80
Tertiary: 60

Residue Display Mode
● Normal
○ Differences

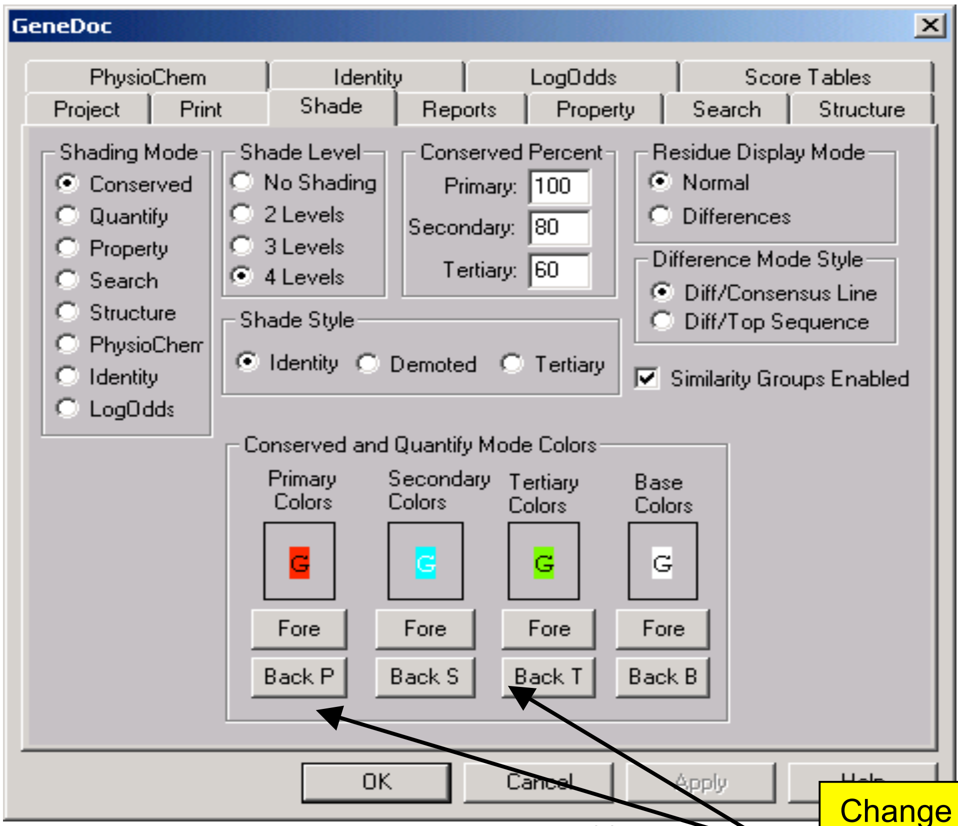
Difference Mode Style
● Diff/Consensus Line
○ Diff/Top Sequence

Shade Style
● Identity ○ Demoted ○ Tertiary

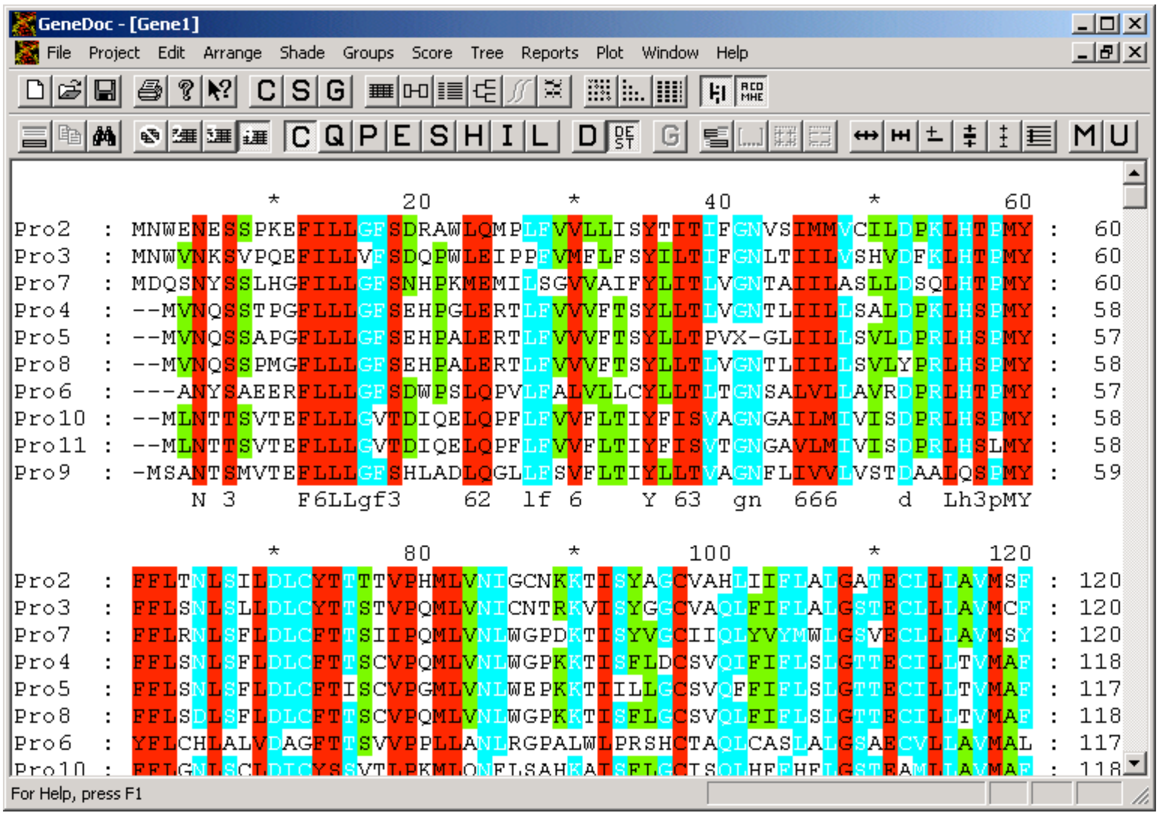
☑ Similarity Groups Enabled

Conserved and Quantify Mode Colors
Primary Colors | Secondary Colors | Tertiary Colors | Base Colors
G | G | G | G

This menu box can be used to customise the display

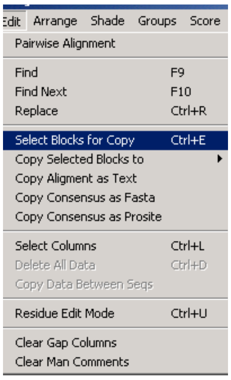

GeneDoc
PhysioChem | Identity | LogOdds | Score Tables
Project | Print | Shade | Reports | Property | Search | Structure
Fore | Fore | Fore | Fore
Back P | Back S | Back T | Back B
OK | Cancel | Apply | Help

Change to colours of your choice


GeneDoc - [Gene1]
File Project Edit Arrange Shade Groups Score Tree Reports Plot Window Help
C S G
C Q P E S H I L D G M U

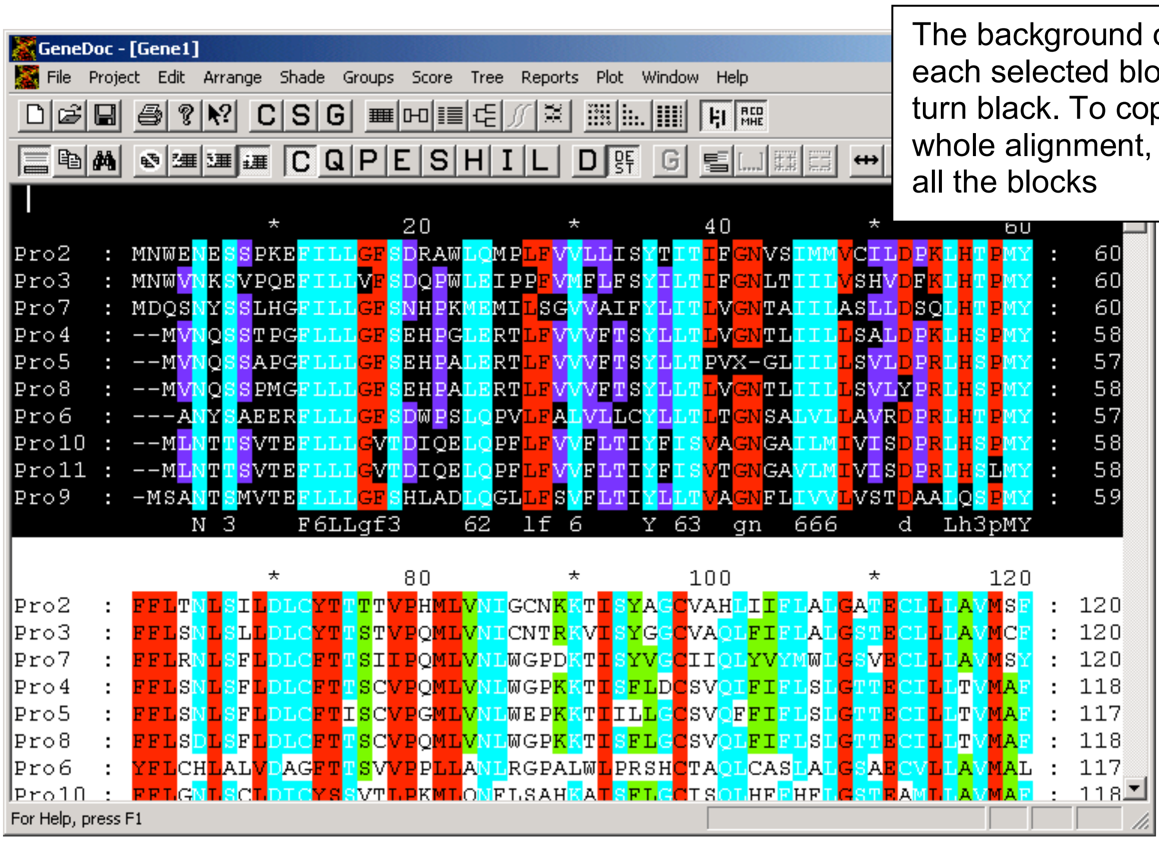
```
                    *        20         *        40         *        60
Pro2   : MNWENESSPKEFILLGFSDRAWLQMPLFVVLLISYTITIFGNVSIMMVCILDPKLHTPMY :  60
Pro3   : MNWVNKSVPQEFILLVFSDQPWLEIPPFVMFLFSYILTIFGNLTIIIVSHVDFKLHTPMY :  60
Pro7   : MDQSNYSSLHGFILLGFSNHPKMEMILSGVVAIFYLITLVGNTAIILASLLDSQLHTPMY :  60
Pro4   : --MVNQSSTPGFLLLGFSEHPGLERTLFVVVFTSYLLTLVGNTLIILLSALDPKLHSPMY :  58
Pro5   : --MVNQSSAPGFLLLGFSEHPALERTLFVVVFTSYLLTPVX-GLIILLSVLDPRLHSPMY :  57
Pro8   : --MVNQSSPMGFLLLGFSEHPALERTLFVVVFTSYLLTLVGNTLIILLSVLYPRLHSPMY :  58
Pro6   : ---ANYSAEERFLLLGFSDWPSLQPVLFALVLLCYLLTLTGNSALVLLAVRDPRLHTPMY :  57
Pro10  : --MLNTTSVTEFLLLGVTDIQELQPFLFVVFLTIYFISVAGNGAILMIVISDPRLHSPMY :  58
Pro11  : --MLNTTSVTEFLLLGVTDIQELQPFLFVVFLTIYFISVTGNGAVLMIVISDPRLHSLMY :  58
Pro9   : -MSANTSMVTEFLLLGFSHLADLQGLLFSVFLTIYLLTVAGNFLIVVLVSTDAALQSPMY :  59
               N 3     F6LLgf3    62  lf 6     Y 63  gn  666     d  Lh3pMY

                    *        80         *       100         *       120
Pro2   : FFLTNLSILDLCYTTTTVPHMLVNIGCNKKTISYAGCVAHLIIFLALGATECLLLAVMSF : 120
Pro3   : FFLSNLSLLDLCYTTSTVPQMLVNICNTRKVISYGGCVAQLFIFLALGSTECLLLAVMCF : 120
Pro7   : FFLRNLSFLDLCFTTSIIPQMLVNLWGPDKTISYVGCIIQLYVYMWLGSVECLLLAVMSY : 120
Pro4   : FFLSNLSFLDLCFTTSCVPQMLVNLWGPKKTISFLDCSVQIFIFLSLGTTECILLTVMAF : 118
Pro5   : FFLSNLSFLDLCFTISCVPGMLVNLWEPKKTIILLGCSVQFFIFLSLGTTECILLTVMAF : 117
Pro8   : FFLSDLSFLDLCFTTSCVPQMLVNLWGPKKTISFLGCSVQLFIFLSLGTTECILLTVMAF : 118
Pro6   : YFLCHLALVDAGFTTSVVPPLLANLRGPALWLPRSHCTAQLCASLALGSAECVLLAVMAL : 117
Pro10  : FFLGNLSCLDLCYSSVTLPKMLQNFLSAHKATSFLGCTSQLHFFHFLGSTEAMLLAVMAF : 118
```

For Help, press F1

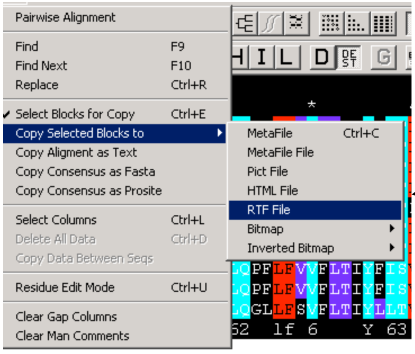


From the edit menu click on Select Blocks for Copy

The background of each selected block will turn black. To copy the whole alignment, select all the blocks


GeneDoc - [Gene1]
File Project Edit Arrange Shade Groups Score Tree Reports Plot Window Help
C S G
C Q P E S H I L D
For Help, press F1

```
                  *        20         *        40         *        60
Pro2   : MNWENESSPKEFILLGFSDRAWLQMPLFVVLLISYTITIFGNVSIMMVCILDPKLHTPMY :  60
Pro3   : MNWVNKSVPQEFILLVFSDQPWLEIPPFVMFLFSYILTIFGNLTIILVSHVDFKLHTPMY :  60
Pro7   : MDQSNYSSLHGFILLGFSNHPKMEMILSGVVAIFYLITLVGNTAIILASLLDSQLHTPMY :  60
Pro4   : --MVNQSSTPGFLLLGFSEHPGLERTLFVVVFTSYLLTLVGNTLIIILSALDPKLHSPMY :  58
Pro5   : --MVNQSSAPGFLLLGFSEHPALERTLFVVVFTSYLLTPVX-GLIIILSVLDPRLHSPMY :  57
Pro8   : --MVNQSSPMGFLLLGFSEHPALERTLFVVVFTSYLLTLVGNTLIIILSVLYPRLHSPMY :  58
Pro6   : ---ANYSAEERFLLLGFSDWPSLQPVLFALVLLCYLLTLTGNSALVLLAVRDPRLHTPMY :  57
Pro10  : --MLNTTSVTEFLLLGVTDIQELQPFLFVVFLTIYFISVAGNGAILMIVISDPRLHSPMY :  58
Pro11  : --MLNTTSVTEFLLLGVTDIQELQPFLFVVFLTIYFISVTGNGAVLMIVISDPRLHSLMY :  58
Pro9   : -MSANTSMVTEFLLLGFSHLADLQGLLFSVFLTIYLLTVAGNFLIVVLVSTDAALQSPMY :  59
             N 3    F6LLgf3    62  lf 6    Y 63  gn  666    d  Lh3pMY

                  *        80         *       100         *       120
Pro2   : FFLTNLSILDLCYTTTTVPHMLVNIGCNKKTISYAGCVAHLIIFLALGATECLLLAVMSF : 120
Pro3   : FFLSNLSLLDLCYTTSTVPQMLVNICNTRKVISYGGCVAQLFIFLALGSTECLLLAVMCF : 120
Pro7   : FFLRNLSFLDLCFTTSIIPQMLVNLWGPDKTISYVGCIIQLYVYMWLGSVECLLLAVMSY : 120
Pro4   : FFLSNLSFLDLCFTTSCVPQMLVNLWGPKKTISFLDCSVQIFIFLSLGTTECILLTVMAF : 118
Pro5   : FFLSNLSFLDLCFTISCVPGMLVNLWEPKKTIILLGCSVQFFIFLSLGTTECILLTVMAF : 117
Pro8   : FFLSDLSFLDLCFTTSCVPQMLVNLWGPKKTISFLGCSVQLFIFLSLGTTECILLTVMAF : 118
Pro6   : YFLCHLALVDAGFTTSVVPPLLANLRGPALWLPRSHCTAQLCASLALGSAECVLLAVMAL : 117
Pro10  : FFLGNLSCLDICYSSVTLPKMLQNFLSAHKATSFLGCTSQLHFFHFLGSTEAMLLAVMAF : 118
```



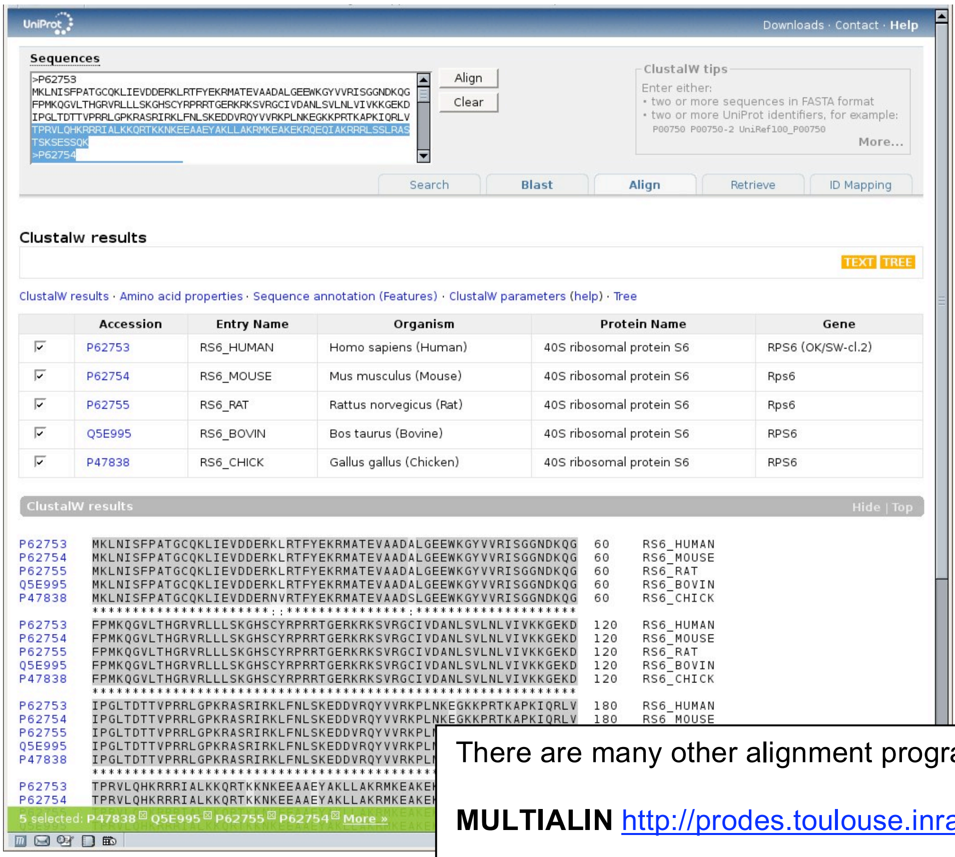
From under the edit menu select Copy Selected Blocks to, RTF file

**Other file types are also available. Use RTF to import into Word, and metafile file to import into PowerPoint**

Import the alignment into Word. RTF format files may be edited in Word

```
Pro2  : MNWENESSPKEFILLGFSDRAWLQMPLFVVLLISYTITIFGNVSIMMV
Pro3  : MNWVNKSVPQEFILLVFSDQPWLEIPPFVMFLFSYILTIFGNLTIILV
Pro7  : MDQSNYSSLHGFILLGFSNHPKMEMILSGVVAIFYLITLVGNTAIILA
Pro4  : --MVNQSSTPGFLLLGFSEHPGLERTLFVVVFTSYLLTLVGNTLIILL
Pro5  : --MVNQSSAPGFLLLGFSEHPALERTLFVVVFTSYLLTPVX-GLIILLSVLDPRLHSPMY :   57
Pro8  : --MVNQSSPMGFLLLGFSEHPALERTLFVVVFTSYLLTLVGNTLIILLSVLYPRLHSPMY :   58
Pro6  : ---ANYSAEERFLLLGFSDWPSLQPVLFALVLLCYLLTLTGNSALVLLAVRDPRLHTPMY :   57
Pro10 : --MLNTTSVTEFLLLGVTDIQELQPFLFVVFLTIYFISVAGNGAILMIVISDPRLHSPMY :   58
Pro11 : --MLNTTSVTEFLLLGVTDIQELQPFLFVVFLTIYFISVTGNGAVLMIVISDPRLHSLMY :   58
Pro9  : -MSANTSMVTEFLLLGFSHLADLQGLLFSVFLTIYLLTVAGNFLIVVLVSTDAALQSPMY :   59

                 *        80         *       100         *       120
Pro2  : FFLTNLSILDLCYTTTTVPHMLVNIGCNKKTISYAGCVAHLIIFLALGATECLLLAVMSF : 120
Pro3  : FFLSNLSLLDLCYTTSTVPQMLVNICNTRKVISYGGCVAQLFIFLALGSTECLLLAVMCF : 120
Pro7  : FFLRNLSFLDLCFTTSIIPQMLVNLWGPDKTISYVGCIIQLYVYMWLGSVECLLLAVMSY : 120
Pro4  : FFLSNLSFLDLCFTTSCVPQMLVNLWGPKKTISFLDCSVQIFIFLSLGTTECILLTVMAF : 118
Pro5  : FFLSNLSFLDLCFTISCVPGMLVNLWEPKKTIILLGCSVQFFIFLSLGTTECILLTVMAF : 117
Pro8  : FFLSDLSFLDLCFTTSCVPQMLVNLWGPKKTISFLGCSVQLFIFLSLGTTECILLTVMAF : 118
Pro6  : YFLCHLALVDAGFTTSVVPPLLANLRGPALWLPRSHCTAQLCASLALGSAECVLLAVMAL : 117
Pro10 : FFLGNLSCLDICYSSVTLPKMLQNFLSAHKAISFLGCISQLHFFHFLGSTEAMLLAVMAF : 118
Pro11 : FFLGNLSYLDICYSTVTLPKMLQNFLSTHKAISFLGCISQLHFFHFLGSTESMLFAVMAF : 118
Pro9  : FFLRTLSALEIGYTSVTVPLLLHHLLTGRRHISRSGCALQMFFFLFFGATECCLLAAMAY : 119
```

You may also align sequences directly from UniProtKB by clicking on the Align button:



```
>P62753
MKLNISFPATGCQKLIEVDDERKLRTFYEKRMATEVAADALGEEWKGYVVRISGGNDKQG
FPMKQGVLTHGRVRLLLSKGHSCYRPRRTGERKRKSVRGCIVDANLSVLNLVIVKKGEKD
IPGLTDTTVPRRLGPKRASRIRKLFNLSKEDDVRQYVVRKPLNKEGKKPRTKAPKIQRLV
TPRVLQHKRRRIALKKQRTKKNKEEAAEYAKLLAKRMKEAKEKRQEQIAKRRRLSSLRAS
TSKSESSQK
>P62754
```

ClustalW tips

Enter either:
- two or more sequences in FASTA format
- two or more UniProt identifiers, for example:
  P00750 P00750-2 UniRef100_P00750

More...

Clustalw results

ClustalW results · Amino acid properties · Sequence annotation (Features) · ClustalW parameters (help) · Tree

| | Accession | Entry Name | Organism | Protein Name | Gene |
|---|---|---|---|---|---|
| ☑ | P62753 | RS6_HUMAN | Homo sapiens (Human) | 40S ribosomal protein S6 | RPS6 (OK/SW-cl.2) |
| ☑ | P62754 | RS6_MOUSE | Mus musculus (Mouse) | 40S ribosomal protein S6 | Rps6 |
| ☑ | P62755 | RS6_RAT | Rattus norvegicus (Rat) | 40S ribosomal protein S6 | Rps6 |
| ☑ | Q5E995 | RS6_BOVIN | Bos taurus (Bovine) | 40S ribosomal protein S6 | RPS6 |
| ☑ | P47838 | RS6_CHICK | Gallus gallus (Chicken) | 40S ribosomal protein S6 | RPS6 |

ClustalW results

```
P62753    MKLNISFPATGCQKLIEVDDERKLRTFYEKRMATEVAADALGEEWKGYVVRISGGNDKQG  60    RS6_HUMAN
P62754    MKLNISFPATGCQKLIEVDDERKLRTFYEKRMATEVAADALGEEWKGYVVRISGGNDKQG  60    RS6_MOUSE
P62755    MKLNISFPATGCQKLIEVDDERKLRTFYEKRMATEVAADALGEEWKGYVVRISGGNDKQG  60    RS6_RAT
Q5E995    MKLNISFPATGCQKLIEVDDERKLRTFYEKRMATEVAADALGEEWKGYVVRISGGNDKQG  60    RS6_BOVIN
P47838    MKLNISFPATGCQKLIEVDDERNVRTFYEKRMATEVAADSLGEEWKGYVVRISGGNDKQG  60    RS6_CHICK
          **********************::***************:********************

P62753    FPMKQGVLTHGRVRLLLSKGHSCYRPRRTGERKRKSVRGCIVDANLSVLNLVIVKKGEKD  120   RS6_HUMAN
P62754    FPMKQGVLTHGRVRLLLSKGHSCYRPRRTGERKRKSVRGCIVDANLSVLNLVIVKKGEKD  120   RS6_MOUSE
P62755    FPMKQGVLTHGRVRLLLSKGHSCYRPRRTGERKRKSVRGCIVDANLSVLNLVIVKKGEKD  120   RS6_RAT
Q5E995    FPMKQGVLTHGRVRLLLSKGHSCYRPRRTGERKRKSVRGCIVDANLSVLNLVIVKKGEKD  120   RS6_BOVIN
P47838    FPMKQGVLTHGRVRLLLSKGHSCYRPRRTGERKRKSVRGCIVDANLSVLNLVIVKKGEKD  120   RS6_CHICK
          ************************************************************

P62753    IPGLTDTTVPRRLGPKRASRIRKLFNLSKEDDVRQYVVRKPLNKEGKKPRTKAPKIQRLV  180   RS6_HUMAN
P62754    IPGLTDTTVPRRLGPKRASRIRKLFNLSKEDDVRQYVVRKPLNKEGKKPRTKAPKIQRLV  180   RS6_MOUSE
P62755    IPGLTDTTVPRRLGPKRASRIRKLFNLSKEDDVRQYVVRKPL
Q5E995    IPGLTDTTVPRRLGPKRASRIRKLFNLSKEDDVRQYVVRKPL
P47838    IPGLTDTTVPRRLGPKRASRIRKLFNLSKEDDVRQYVVRKPL
          ******************************************

P62753    TPRVLQHKRRRIALKKQRTKKNKEEAAEYAKLLAKRMKEAKEK
P62754    TPRVLQHKRRRIALKKQRTKKNKEEAAEYAKLLAKRMKEAKEK
```

5 selected: P47838 ☒ Q5E995 ☒ P62755 ☒ P62754 ☒ More »

There are many other alignment programs available on the web:

**MULTIALIN** http://prodes.toulouse.inra.fr/multalin/multalin.html

**PIMA** http://bioweb.pasteur.fr/seqanal/interfaces/pima-simple.html

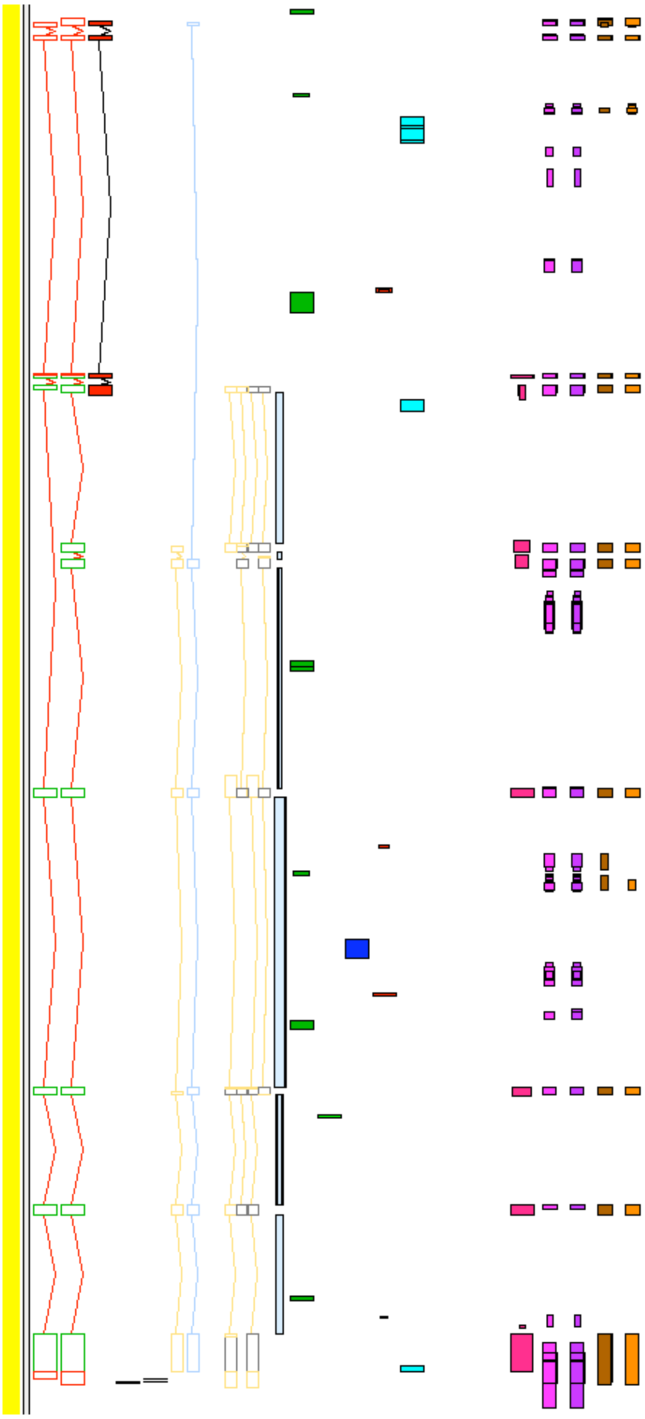
**DIALIGN** http://bioweb.pasteur.fr/seqanal/interfaces/dialign2-simple.html

**DCA** http://bioweb.pasteur.fr/seqanal/interfaces/dca-simple.html

# Manual Genome Annotation

The underlying data for the Vega database is generated by the Havana group.

Below is a screen-shot of the crfb1 locus from the groups' annotation software. Protein coding genes are shown in red and green, whilst non-coding transcripts are shown in red.
Other columns show Blast hits to DNA and protein databases, repeats and gene models generated from the solexa sequenced transcriptome data.

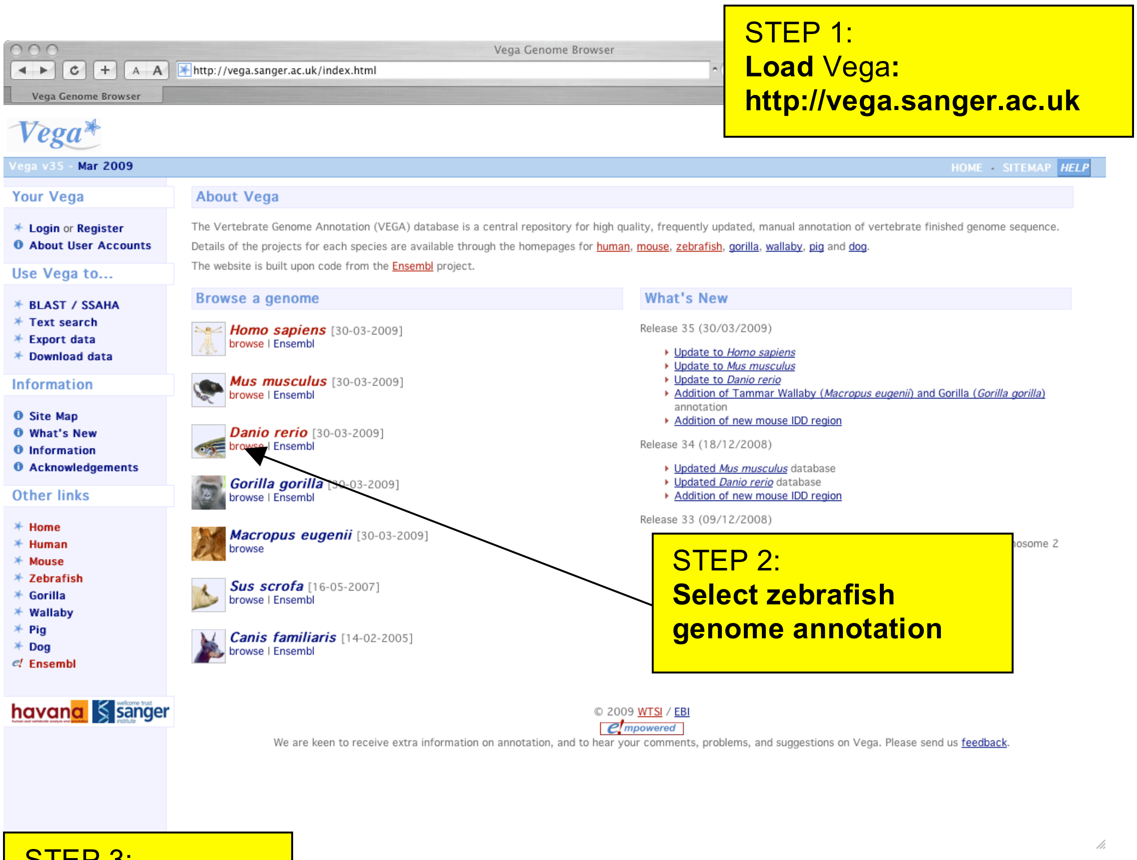


# The Vertebrate Genome Annotation (VEGA) Database

**Worked example:**

**1.** View the crfb1 locus. How many transcripts are there ? What is the supporting evidence ? Export peptide sequence.

STEP 1:
**Load** Vega:
**http://vega.sanger.ac.uk**



About Vega

The Vertebrate Genome Annotation (VEGA) database is a central repository for high quality, frequently updated, manual annotation of vertebrate finished genome sequence. Details of the projects for each species are available through the homepages for human, mouse, zebrafish, gorilla, wallaby, pig and dog.

The website is built upon code from the Ensembl project.

Browse a genome

- *Homo sapiens* [30-03-2009] browse | Ensembl
- *Mus musculus* [30-03-2009] browse | Ensembl
- *Danio rerio* [30-03-2009] browse | Ensembl
- *Gorilla gorilla* [30-03-2009] browse | Ensembl
- *Macropus eugenii* [30-03-2009] browse
- *Sus scrofa* [16-05-2007] browse | Ensembl
- *Canis familiaris* [14-02-2005] browse | Ensembl

What's New

Release 35 (30/03/2009)

- Update to *Homo sapiens*
- Update to *Mus musculus*
- Update to *Danio rerio*
- Addition of Tammar Wallaby (*Macropus eugenii*) and Gorilla (*Gorilla gorilla*) annotation
- Addition of new mouse IDD region

Release 34 (18/12/2008)

- Updated *Mus musculus* database
- Updated *Danio rerio* database
- Addition of new mouse IDD region

Release 33 (09/12/2008)

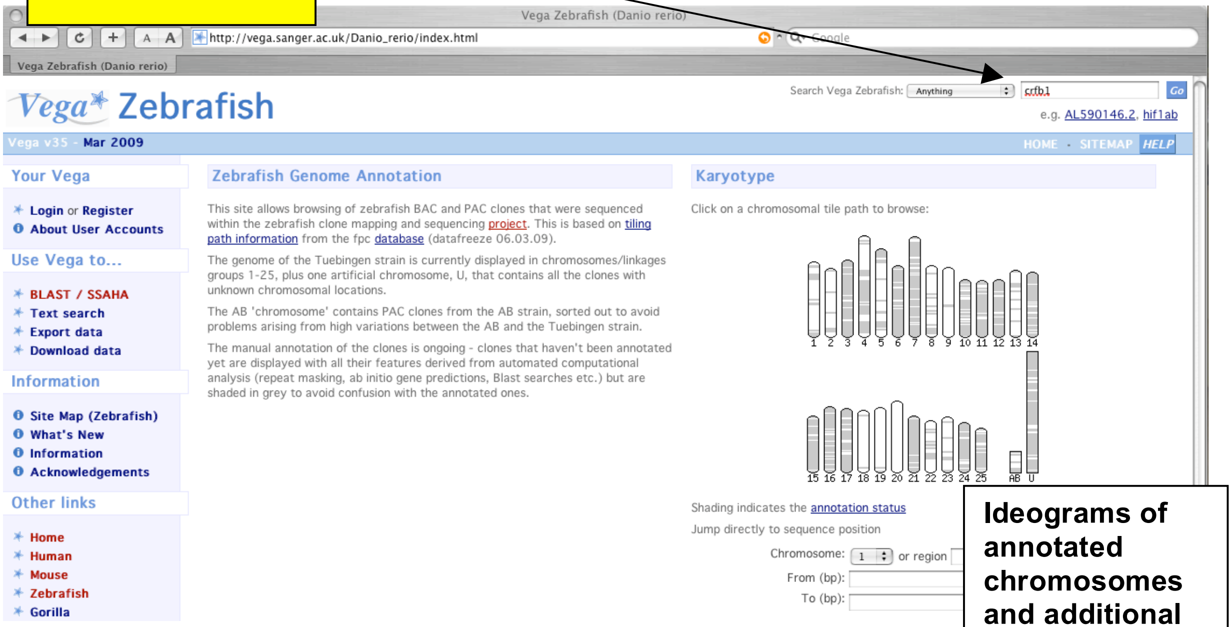
...osome 2

© 2009 WTSI / EBI

We are keen to receive extra information on annotation, and to hear your comments, problems, and suggestions on Vega. Please send us feedback.

STEP 2:
**Select zebrafish genome annotation**

STEP 3:
**Search for gene symbol crfb1**



Zebrafish Genome Annotation

This site allows browsing of zebrafish BAC and PAC clones that were sequenced within the zebrafish clone mapping and sequencing project. This is based on tiling path information from the fpc database (datafreeze 06.03.09).

The genome of the Tuebingen strain is currently displayed in chromosomes/linkages groups 1-25, plus one artificial chromosome, U, that contains all the clones with unknown chromosomal locations.

The AB 'chromosome' contains PAC clones from the AB strain, sorted out to avoid problems arising from high variations between the AB and the Tuebingen strain.

The manual annotation of the clones is ongoing - clones that haven't been annotated yet are displayed with all their features derived from automated computational analysis (repeat masking, ab initio gene predictions, Blast searches etc.) but are shaded in grey to avoid confusion with the annotated ones.

Karyotype

Click on a chromosomal tile path to browse:

Shading indicates the annotation status

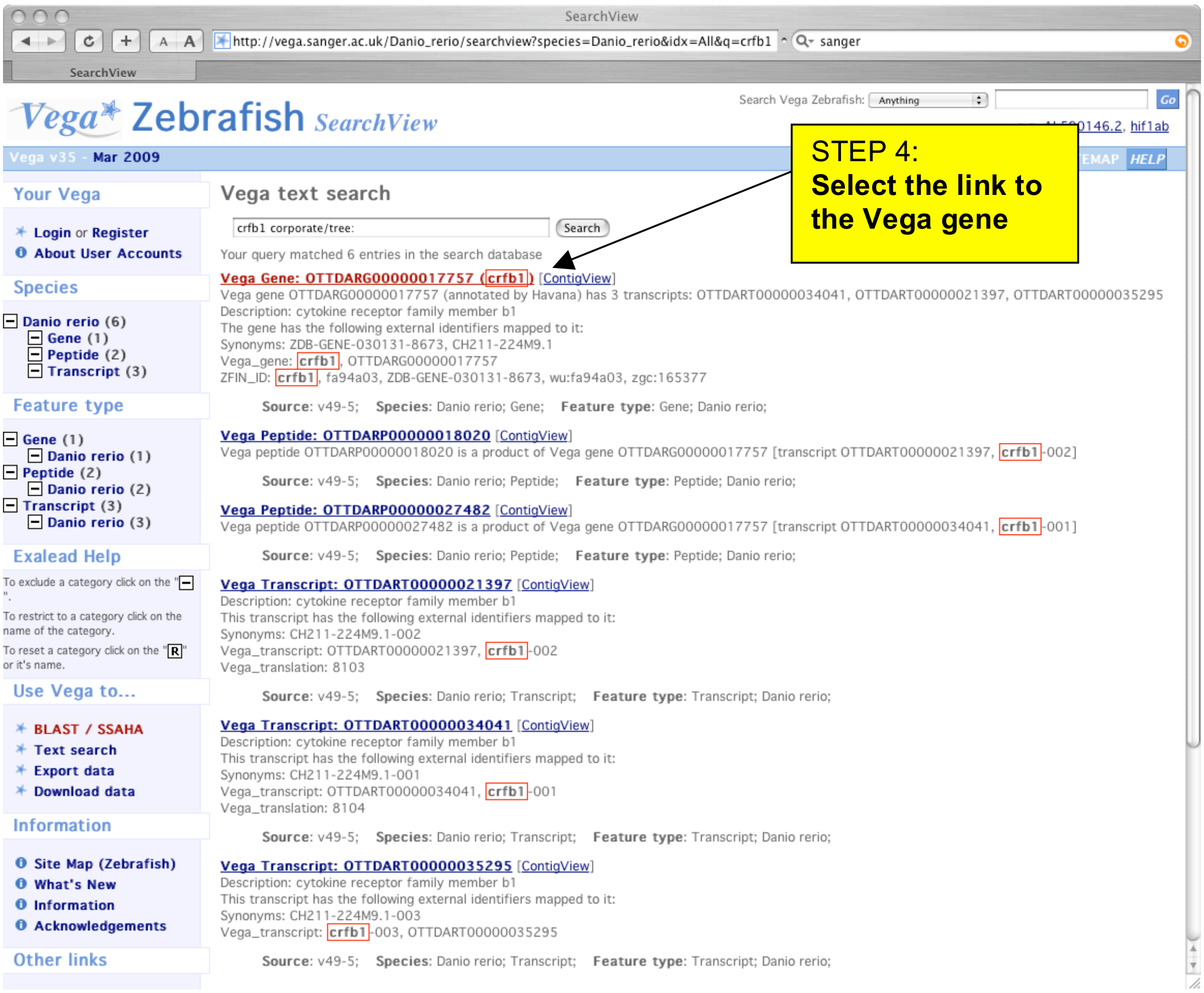
Jump directly to sequence position

Chromosome: 1 or region

From (bp):

To (bp):

**Ideograms of annotated chromosomes and additional information**



To exclude a category click on the "⊟".

To restrict to a category click on the name of the category.

To reset a category click on the "R" or it's name.

Vega text search

crfb1 corporate/tree:

Your query matched 6 entries in the search database

**Vega Gene: OTTDARG00000017757 (crfb1)** [ContigView]
Vega gene OTTDARG00000017757 (annotated by Havana) has 3 transcripts: OTTDART00000034041, OTTDART00000021397, OTTDART00000035295
Description: cytokine receptor family member b1
The gene has the following external identifiers mapped to it:
Synonyms: ZDB-GENE-030131-8673, CH211-224M9.1
Vega_gene: crfb1, OTTDARG00000017757
ZFIN_ID: crfb1, fa94a03, ZDB-GENE-030131-8673, wu:fa94a03, zgc:165377

**Source**: v49-5; **Species**: Danio rerio; Gene; **Feature type**: Gene; Danio rerio;

**Vega Peptide: OTTDARP00000018020** [ContigView]
Vega peptide OTTDARP00000018020 is a product of Vega gene OTTDARG00000017757 [transcript OTTDART00000021397, crfb1-002]

**Source**: v49-5; **Species**: Danio rerio; Peptide; **Feature type**: Peptide; Danio rerio;

**Vega Peptide: OTTDARP00000027482** [ContigView]
Vega peptide OTTDARP00000027482 is a product of Vega gene OTTDARG00000017757 [transcript OTTDART00000034041, crfb1-001]

**Source**: v49-5; **Species**: Danio rerio; Peptide; **Feature type**: Peptide; Danio rerio;

**Vega Transcript: OTTDART00000021397** [ContigView]
Description: cytokine receptor family member b1
This transcript has the following external identifiers mapped to it:
Synonyms: CH211-224M9.1-002
Vega_transcript: OTTDART00000021397, crfb1-002
Vega_translation: 8103

**Source**: v49-5; **Species**: Danio rerio; Transcript; **Feature type**: Transcript; Danio rerio;

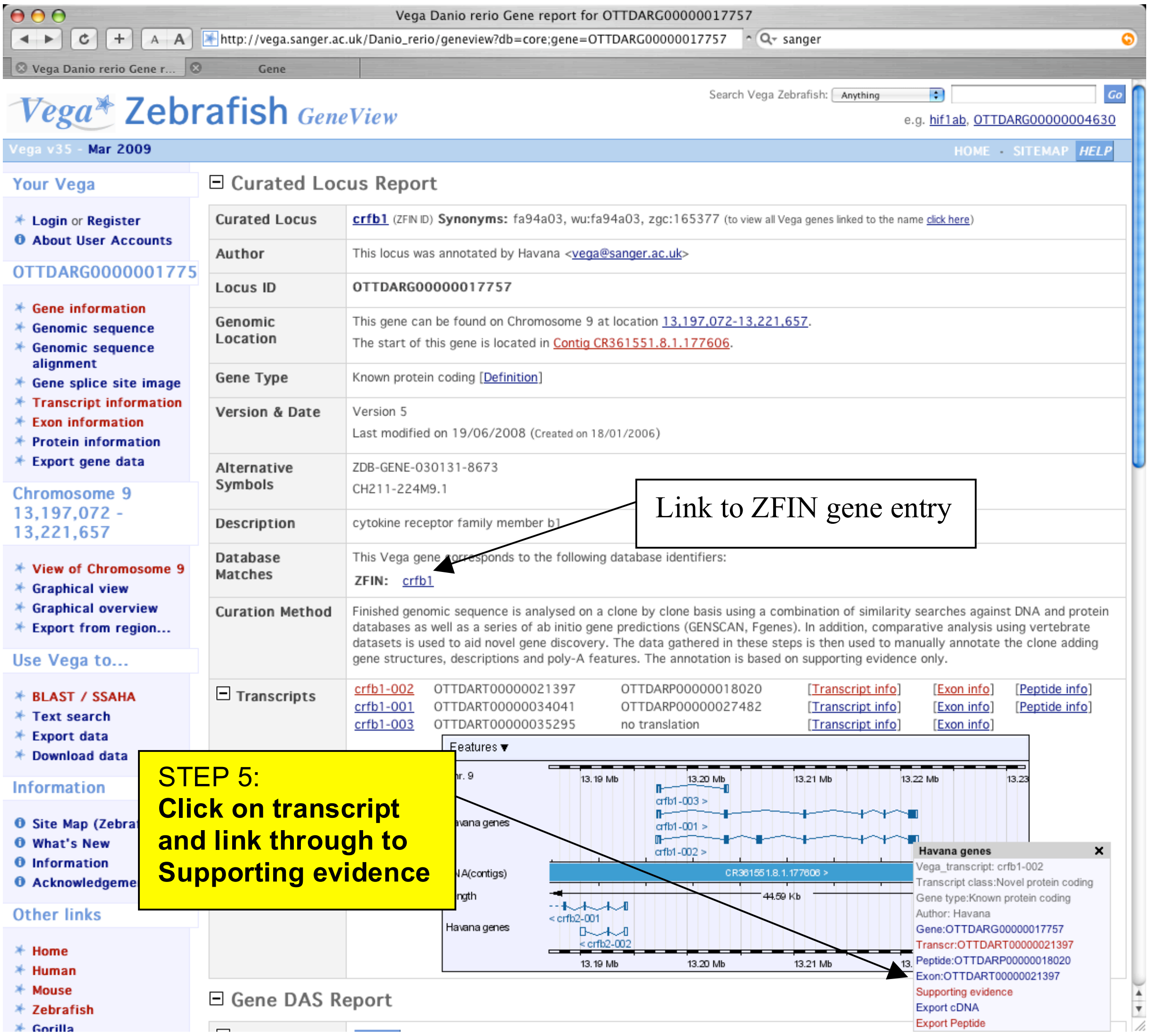
**Vega Transcript: OTTDART00000034041** [ContigView]
Description: cytokine receptor family member b1
This transcript has the following external identifiers mapped to it:
Synonyms: CH211-224M9.1-001
Vega_transcript: OTTDART00000034041, crfb1-001
Vega_translation: 8104

**Source**: v49-5; **Species**: Danio rerio; Transcript; **Feature type**: Transcript; Danio rerio;

**Vega Transcript: OTTDART00000035295** [ContigView]
Description: cytokine receptor family member b1
This transcript has the following external identifiers mapped to it:
Synonyms: CH211-224M9.1-003
Vega_transcript: crfb1-003, OTTDART00000035295

**Source**: v49-5; **Species**: Danio rerio; Transcript; **Feature type**: Transcript; Danio rerio;

STEP 4:
**Select the link to the Vega gene**



Vega Danio rerio Gene report for OTTDARG00000017757

http://vega.sanger.ac.uk/Danio_rerio/geneview?db=core;gene=OTTDARG00000017757

Vega Zebrafish GeneView

Search Vega Zebrafish: Anything Go

e.g. hif1ab, OTTDARG00000004630

Vega v35 - Mar 2009 HOME · SITEMAP HELP

## Curated Locus Report

| | |
|---|---|
| Curated Locus | crfb1 (ZFIN ID) **Synonyms:** fa94a03, wu:fa94a03, zgc:165377 (to view all Vega genes linked to the name click here) |
| Author | This locus was annotated by Havana <vega@sanger.ac.uk> |
| Locus ID | OTTDARG00000017757 |
| Genomic Location | This gene can be found on Chromosome 9 at location 13,197,072-13,221,657.<br>The start of this gene is located in Contig CR361551.8.1.177606. |
| Gene Type | Known protein coding [Definition] |
| Version & Date | Version 5<br>Last modified on 19/06/2008 (Created on 18/01/2006) |
| Alternative Symbols | ZDB-GENE-030131-8673<br>CH211-224M9.1 |
| Description | cytokine receptor family member b1 |
| Database Matches | This Vega gene corresponds to the following database identifiers:<br>ZFIN: crfb1 |
| Curation Method | Finished genomic sequence is analysed on a clone by clone basis using a combination of similarity searches against DNA and protein databases as well as a series of ab initio gene predictions (GENSCAN, Fgenes). In addition, comparative analysis using vertebrate datasets is used to aid novel gene discovery. The data gathered in these steps is then used to manually annotate the clone adding gene structures, descriptions and poly-A features. The annotation is based on supporting evidence only. |

Transcripts

| | | | | | |
|---|---|---|---|---|---|
| crfb1-002 | OTTDART00000021397 | OTTDARP00000018020 | [Transcript info] | [Exon info] | [Peptide info] |
| crfb1-001 | OTTDART00000034041 | OTTDARP00000027482 | [Transcript info] | [Exon info] | [Peptide info] |
| crfb1-003 | OTTDART00000035295 | no translation | [Transcript info] | [Exon info] | |

Havana genes ✕

Vega_transcript: crfb1-002

Transcript class:Novel protein coding

Gene type:Known protein coding

Author: Havana

Gene:OTTDARG00000017757

Transcr:OTTDART00000021397

Peptide:OTTDARP00000018020

Exon:OTTDART00000021397

Supporting evidence

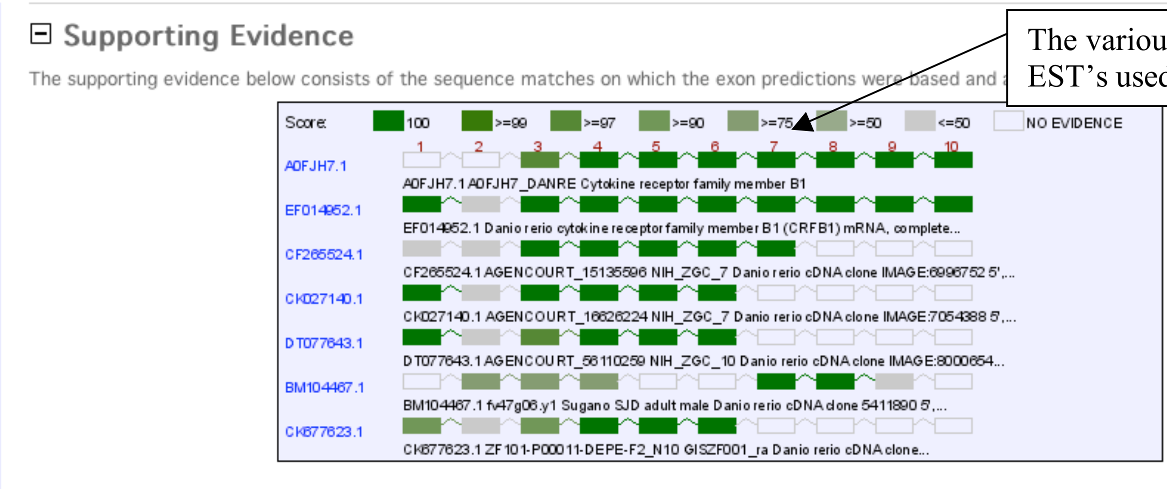
Export cDNA

Export Peptide

Gene DAS Report

Link to ZFIN gene entry

**STEP 5:**
**Click on transcript and link through to Supporting evidence**



## Supporting Evidence

The supporting evidence below consists of the sequence matches on which the exon predictions were based and a

Score: 🟩 100 🟩 >=99 🟩 >=97 🟩 >=90 🟩 >=75 🟩 >=50 ⬜ <=50 □ NO EVIDENCE

A0FJH7.1
A0FJH7.1 A0FJH7_DANRE Cytokine receptor family member B1

EF014952.1
EF014952.1 Danio rerio cytokine receptor family member B1 (CRFB1) mRNA, complete...

CF265524.1
CF265524.1 AGENCOURT_15135596 NIH_ZGC_7 Danio rerio cDNA clone IMAGE:6996752 5',...

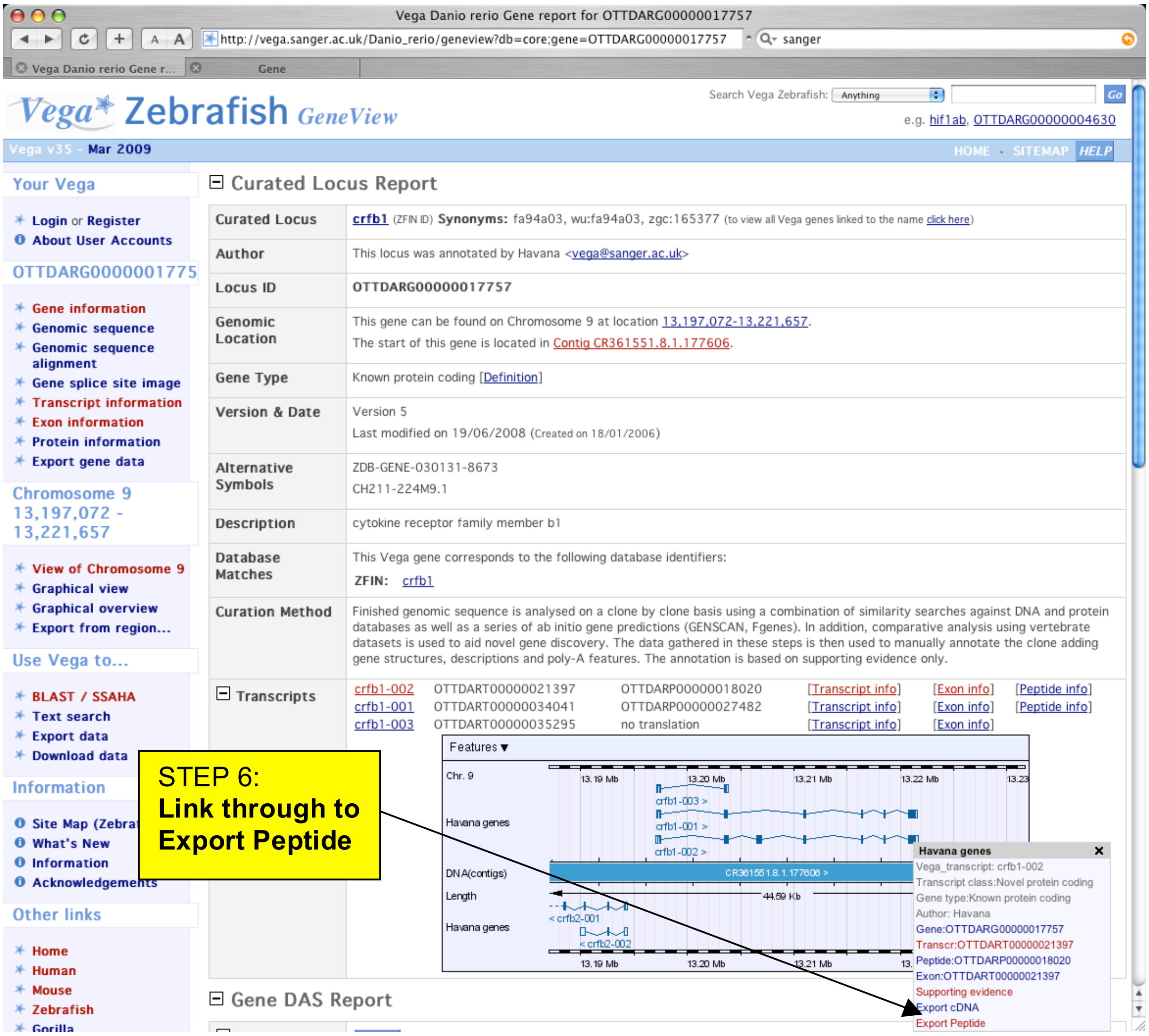
CK027140.1
CK027140.1 AGENCOURT_16626224 NIH_ZGC_7 Danio rerio cDNA clone IMAGE:7054388 5',...

DT077643.1
DT077643.1 AGENCOURT_56110259 NIH_ZGC_10 Danio rerio cDNA clone IMAGE:8000654...

BM104467.1
BM104467.1 fv47g06.y1 Sugano SJD adult male Danio rerio cDNA clone 5411890 5',...

CK677623.1
CK677623.1 ZF101-P00011-DEPE-F2_N10 GISZF001_ra Danio rerio cDNA clone...

The various proteins and EST's used to build variant



Vega Danio rerio Gene report for OTTDARG00000017757

http://vega.sanger.ac.uk/Danio_rerio/geneview?db=core;gene=OTTDARG00000017757

Vega* Zebrafish GeneView

Search Vega Zebrafish: Anything Go

e.g. hif1ab, OTTDARG00000004630

Vega v35 - Mar 2009 HOME · SITEMAP HELP

Your Vega
- Login or Register
- About User Accounts

OTTDARG0000001775
- Gene information
- Genomic sequence
- Genomic sequence alignment
- Gene splice site image
- Transcript information
- Exon information
- Protein information
- Export gene data

Chromosome 9 13,197,072 - 13,221,657
- View of Chromosome 9
- Graphical view
- Graphical overview
- Export from region...

Use Vega to...
- BLAST / SSAHA
- Text search
- Export data
- Download data

Information
- Site Map (Zebra
- What's New
- Information
- Acknowledgements

Other links
- Home
- Human
- Mouse
- Zebrafish
- Gorilla

## ⊟ Curated Locus Report

| | |
|---|---|
| Curated Locus | crfb1 (ZFIN ID) Synonyms: fa94a03, wu:fa94a03, zgc:165377 (to view all Vega genes linked to the name click here) |
| Author | This locus was annotated by Havana <vega@sanger.ac.uk> |
| Locus ID | OTTDARG00000017757 |
| Genomic Location | This gene can be found on Chromosome 9 at location 13,197,072-13,221,657.<br>The start of this gene is located in Contig CR361551.8.1.177606. |
| Gene Type | Known protein coding [Definition] |
| Version & Date | Version 5<br>Last modified on 19/06/2008 (Created on 18/01/2006) |
| Alternative Symbols | ZDB-GENE-030131-8673<br>CH211-224M9.1 |
| Description | cytokine receptor family member b1 |
| Database Matches | This Vega gene corresponds to the following database identifiers:<br>ZFIN: crfb1 |
| Curation Method | Finished genomic sequence is analysed on a clone by clone basis using a combination of similarity searches against DNA and protein databases as well as a series of ab initio gene predictions (GENSCAN, Fgenes). In addition, comparative analysis using vertebrate datasets is used to aid novel gene discovery. The data gathered in these steps is then used to manually annotate the clone adding gene structures, descriptions and poly-A features. The annotation is based on supporting evidence only. |

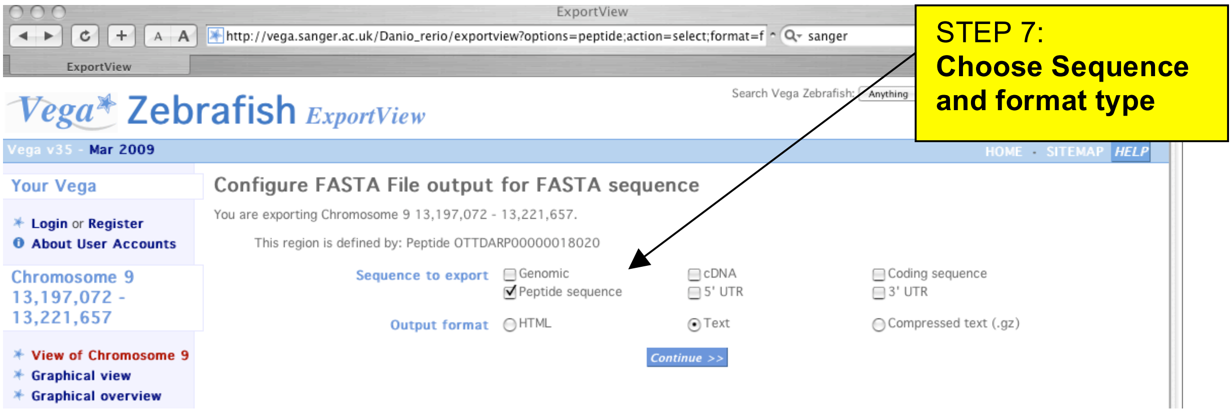
⊟ Transcripts

| | | | | | |
|---|---|---|---|---|---|
| crfb1-002 | OTTDART00000021397 | OTTDARP00000018020 | [Transcript info] | [Exon info] | [Peptide info] |
| crfb1-001 | OTTDART00000034041 | OTTDARP00000027482 | [Transcript info] | [Exon info] | [Peptide info] |
| crfb1-003 | OTTDART00000035295 | no translation | [Transcript info] | [Exon info] | |

Havana genes ✕
Vega_transcript: crfb1-002
Transcript class:Novel protein coding
Gene type:Known protein coding
Author: Havana
Gene:OTTDARG00000017757
Transcr:OTTDART00000021397
Peptide:OTTDARP00000018020
Exon:OTTDART00000021397
Supporting evidence
Export cDNA
Export Peptide

⊟ Gene DAS Report

**STEP 6: Link through to Export Peptide**



ExportView

http://vega.sanger.ac.uk/Danio_rerio/exportview?options=peptide;action=select;format=f

Vega* Zebrafish ExportView

Search Vega Zebrafish: Anything

Vega v35 - Mar 2009 HOME · SITEMAP HELP

Your Vega
- Login or Register
- About User Accounts

Chromosome 9 13,197,072 - 13,221,657
- View of Chromosome 9
- Graphical view
- Graphical overview

## Configure FASTA File output for FASTA sequence

You are exporting Chromosome 9 13,197,072 - 13,221,657.

This region is defined by: Peptide OTTDARP00000018020

Sequence to export: □ Genomic ☑ Peptide sequence □ cDNA □ 5' UTR □ Coding sequence □ 3' UTR

Output format: ○ HTML ● Text ○ Compressed text (.gz)

Continue >>

**STEP 7: Choose Sequence and format type**

## Sanger Blast: Searching for all Finished/Unfinished clones

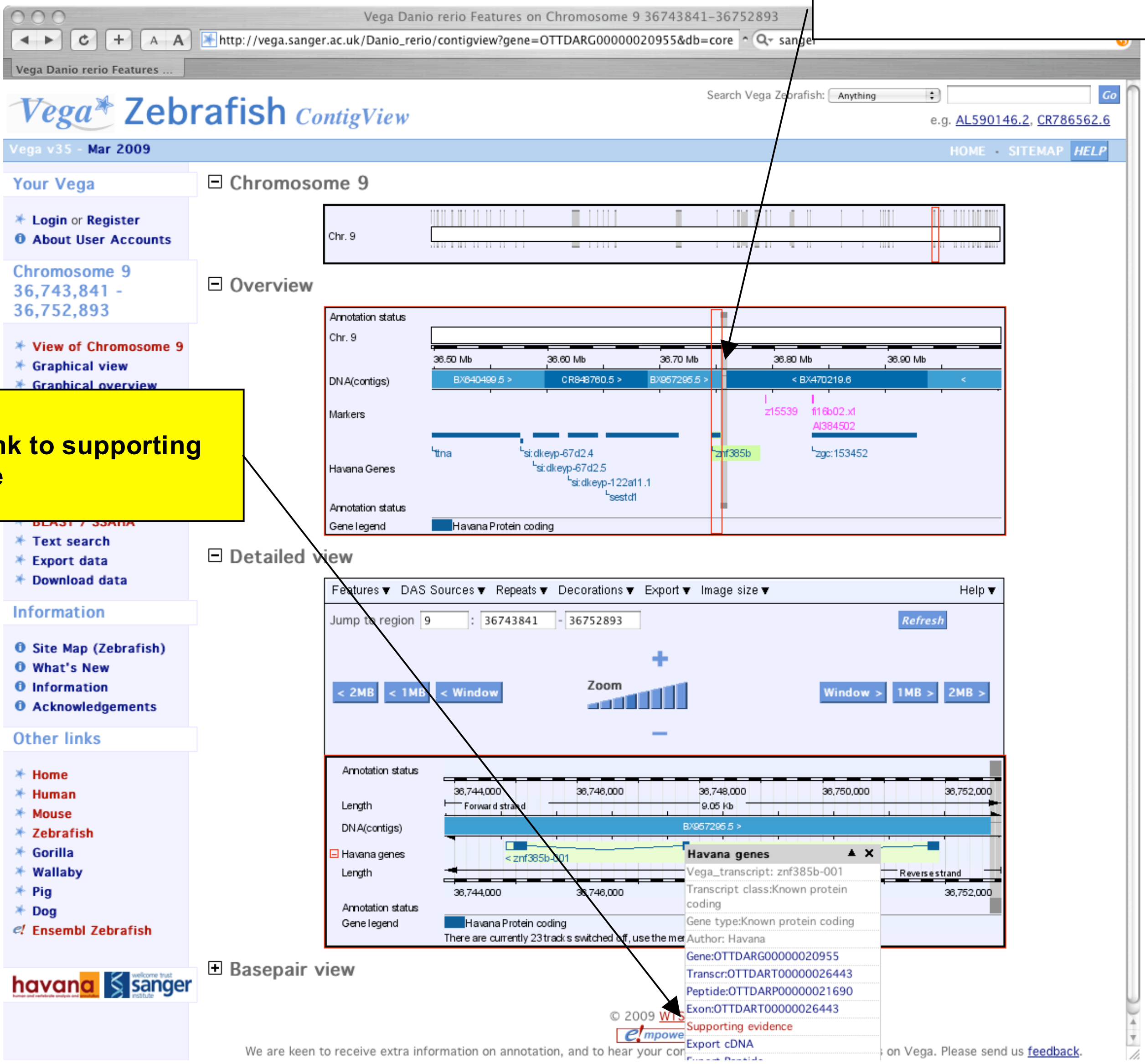
New sequenced clones come through the pipeline on a daily basis. These sequences are submitted to EMBL/GenBank. Although sequenced clones are made public as soon as possible it takes time until they appear in Vega or in a new assembly. The Sanger Institute offers a Blast search page whose target is all the available sequenced clones for zebrafish. This service can be accessed though the Danio rerio project page or directly at:

http://www.sanger.ac.uk/cgi-bin/blast/submitblast/d_rerio

### Worked Example:

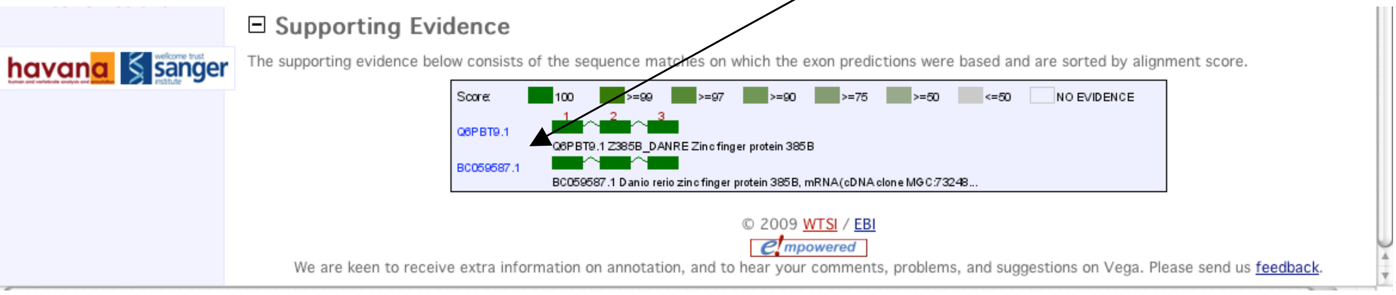
Find which clone the 5' end of the znf385b gene is on ?

5' end of the znf385b gene lies in a gap.



STEP 1:
**Select link to supporting evidence**

STEP 2:
**Select link to external record for BC059587.**



## Supporting Evidence

The supporting evidence below consists of the sequence matches on which the exon predictions were based and are sorted by alignment score.

| Score: | 🟩 100 | 🟩 >=99 | 🟩 >=97 | 🟩 >=90 | 🟩 >=75 | 🟩 >=50 | ▨ <=50 | □ NO EVIDENCE |
|---|---|---|---|---|---|---|---|---|

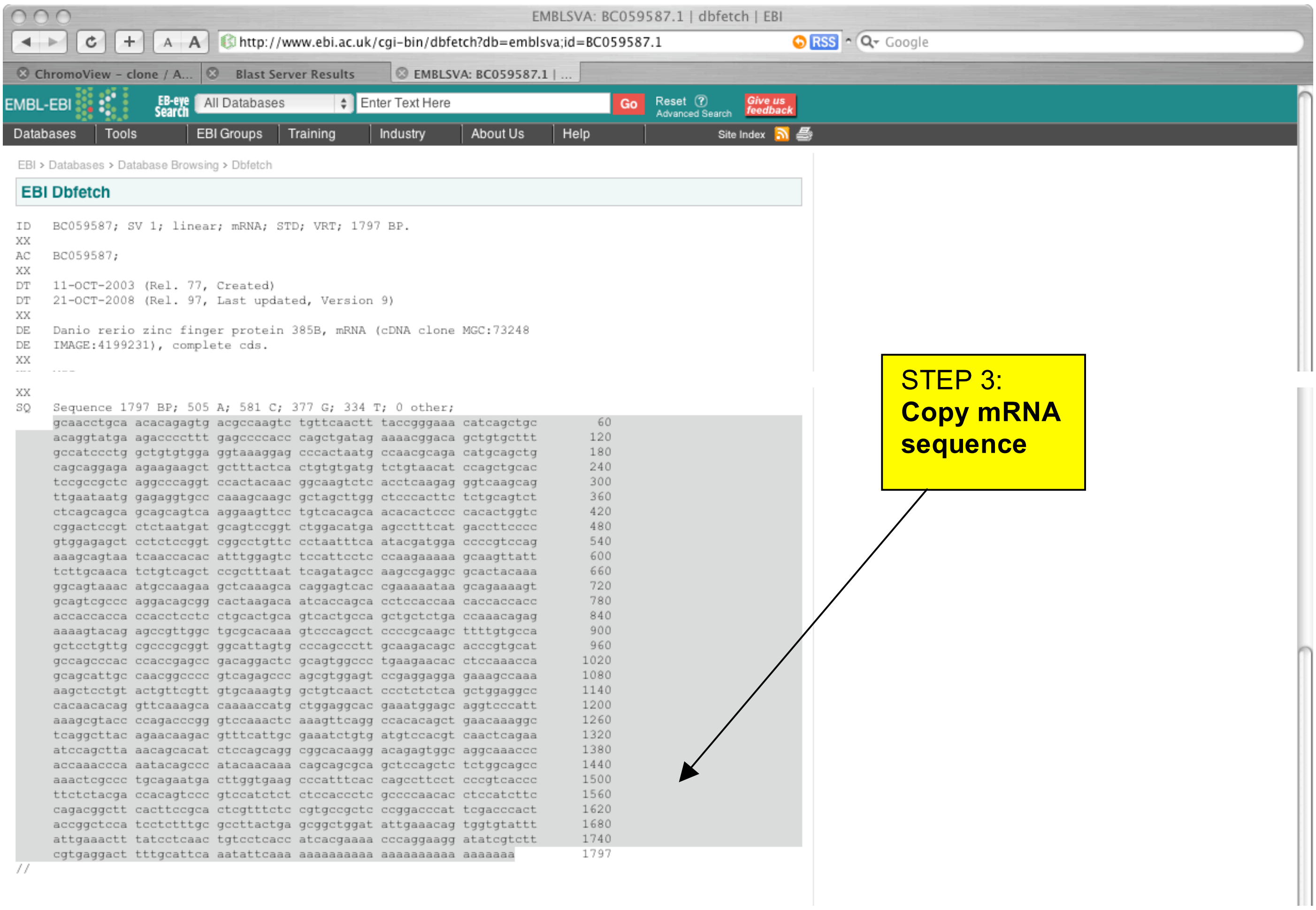
Q6PBT9.1 — Q6PBT9.1 Z385B_DANRE Zinc finger protein 385B

BC059587.1 — BC059587.1 Danio rerio zinc finger protein 385B, mRNA(cDNA clone MGC:73248...

© 2009 WTSI / EBI

e!mpowered

We are keen to receive extra information on annotation, and to hear your comments, problems, and suggestions on Vega. Please send us feedback.

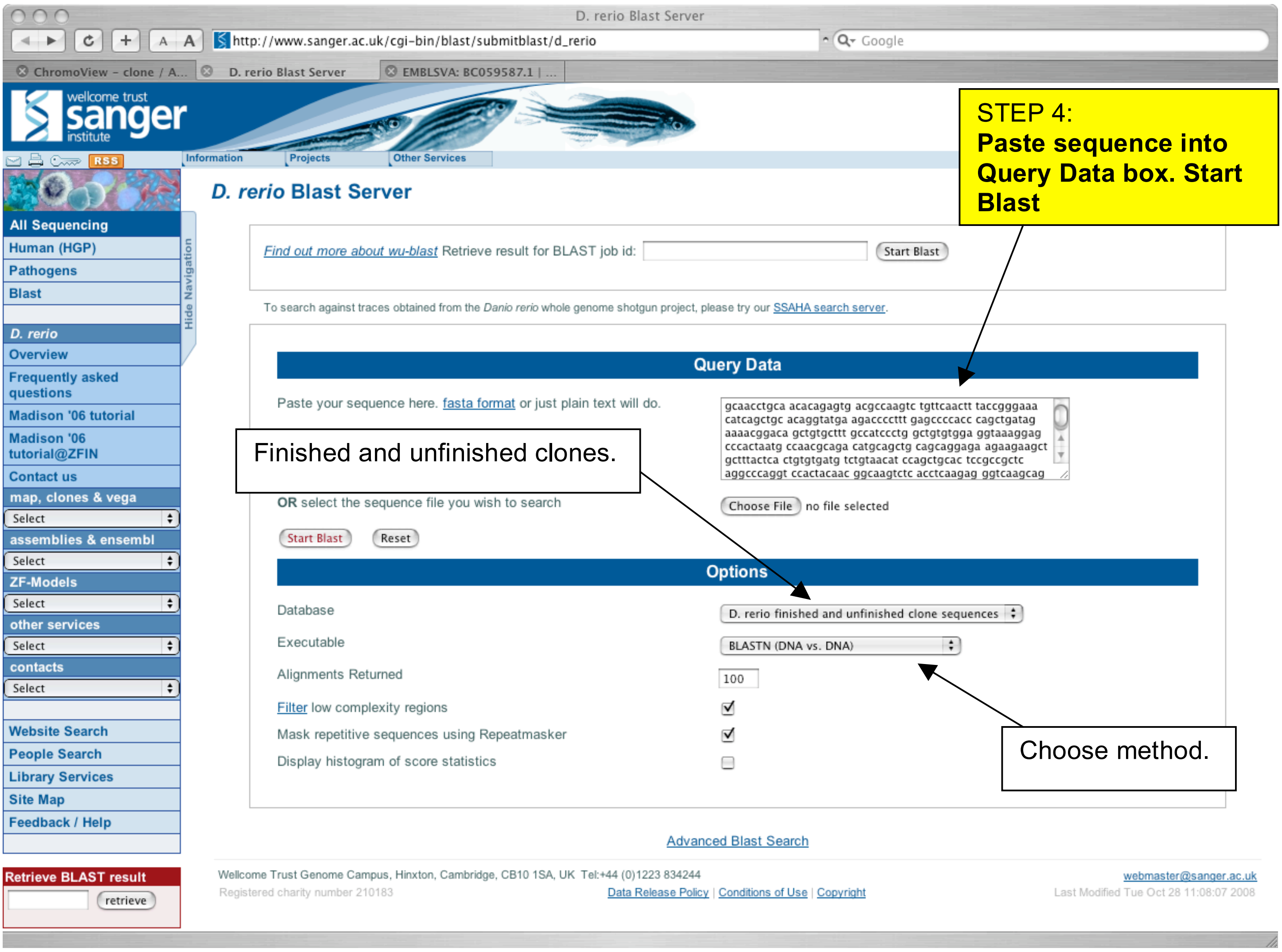


EBI > Databases > Database Browsing > Dbfetch

### EBI Dbfetch

```
ID   BC059587; SV 1; linear; mRNA; STD; VRT; 1797 BP.
XX
AC   BC059587;
XX
DT   11-OCT-2003 (Rel. 77, Created)
DT   21-OCT-2008 (Rel. 97, Last updated, Version 9)
XX
DE   Danio rerio zinc finger protein 385B, mRNA (cDNA clone MGC:73248
DE   IMAGE:4199231), complete cds.
XX

XX
SQ   Sequence 1797 BP; 505 A; 581 C; 377 G; 334 T; 0 other;
     gcaacctgca acacagagtg acgccaagtc tgttcaactt taccgggaaa catcagctgc        60
     acaggtatga agacccettt gagccccacc cagctgatag aaaacggaca gctgtgcttt       120
     gccatccctg gctgtgtgga ggtaaaggag cccactaatg ccaacgcaga catgcagctg       180
     cagcaggaga agaagaagct gctttactca ctgtgtgatg tctgtaacat ccagctgcac       240
     tccgccgctc aggcccaggt ccactacaac ggcaagtctc acctcaagag ggtcaagcag       300
     ttgaataatg gagaggtgcc caaagcaagc gctagcttgg ctcccacttc tctgcagtct       360
     ctcagcagca gcagcagtca aggaagttcc tgtcacagca acacactccc cacactggtc       420
     cggactccgt ctctaatgat gcagtccggt ctggacatga agcctttcat gaccttcccc       480
     gtggagagct cctctccggt cggcctgttc cctaatttca atacgatgga ccccgtccag       540
     aaagcagtaa tcaaccacac atttggagtc tccattcctc ccaagaaaaa gcaagttatt       600
     tcttgcaaca tctgtcagct ccgctttaat tcagatagcc aagccgaggc gcactacaaa       660
     ggcagtaaac atgccaagaa gctcaaagca caggagtcac cgaaaaataa gcagaaaagt       720
     gcagtcgccc aggacagcgg cactaagaca atcaccagca cctccaccaa caccaccacc       780
     accaccacca ccacctcctc ctgcactgca gtcactgcca gctgctctga ccaaacagag       840
     aaaagtacag agccgttggc tgcgcacaaa gtcccagcct ccccgcaagc ttttgtgcca       900
     gctcctgttg cgcccgcggt ggcattagtg cccagccctt gcaagacagc acccgtgcat       960
     gccagcccac ccaccgagcc gacaggactc gcagtggccc tgaagaacac ctccaaacca      1020
     gcagcattgc caacggcccc gtcagagccc agcgtggagt ccgaggagga gaaagccaaa      1080
     aagctcctgt actgttcgtt gtgcaaagtg gctgtcaact ccctctctca gctggaggcc      1140
     cacaacacag gttcaaagca caaaaccatg ctggaggcac gaaatggagc aggtcccatt      1200
     aaagcgtacc ccagacccgg gtccaaactc aaagttcagg ccacacagct gaacaaaggc      1260
     tcaggcttac agaacaagac gtttcattgc gaaatctgtg atgtccacgt caactcagaa      1320
     atccagctta aacagcacat ctccagcagg cggcacaagg acagagtggc aggcaaaccc      1380
     accaaaccca aatacagccc atacaacaaa cagcagcgca gctccagctc tctggcagcc      1440
     aaactcgccc tgcagaatga cttggtgaag cccatttcac cagccttcct cccgtcaccc      1500
     ttctctacga ccacagtccc gtccatctct ctccaccctc gccccaacac ctccatcttc      1560
     cagacggctt cacttccgca ctcgtttctc cgtgccgctc ccggacccat tcgacccact      1620
     accggctcca tcctctttgc gccttactga gcggctggat attgaaacag tggtgtattt      1680
     attgaaactt tatcctcaac tgtcctcacc atcacgaaaa cccaggaagg atatcgtctt      1740
     cgtgaggact tttgcattca aatattcaaa aaaaaaaaaa aaaaaaaaaa aaaaaaa         1797
//
```

STEP 3:
**Copy mRNA sequence**



STEP 4:
**Paste sequence into Query Data box. Start Blast**

Finished and unfinished clones.

Choose method.

## *D. rerio* Blast Server

Find out more about wu-blast Retrieve result for BLAST job id: [ ] Start Blast

To search against traces obtained from the *Danio rerio* whole genome shotgun project, please try our SSAHA search server.

**Query Data**

Paste your sequence here. fasta format or just plain text will do.

```
gcaacctgca acacagagtg acgccaagtc tgttcaactt taccgggaaa
catcagctgc acaggtatga agaccccttt gagccccacc cagctgatag
aaaacggaca gctgtgcttt gccatccctg gctgtgtgga ggtaaaggag
cccactaatg ccaacgcaga catgcagctg cagcaggaga agaagaagct
gctttactca ctgtgtgatg tctgtaacat ccagctgcac tccgccgctc
aggcccaggt ccactacaac ggcaagtctc acctcaagag ggtcaagcag
```

**OR** select the sequence file you wish to search  Choose File no file selected

Start Blast  Reset

**Options**

Database: D. rerio finished and unfinished clone sequences

Executable: BLASTN (DNA vs. DNA)

Alignments Returned: 100

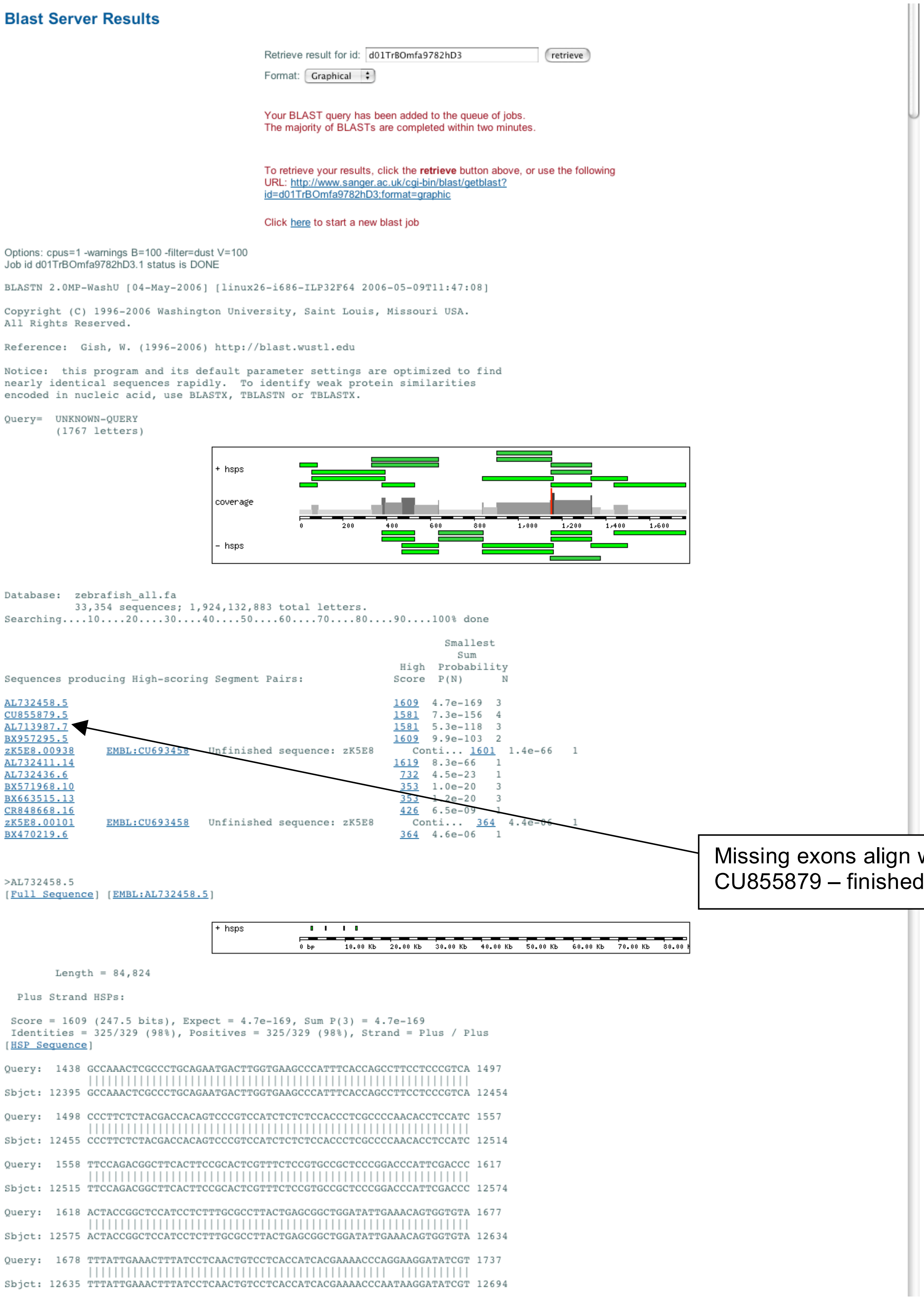
Filter low complexity regions ☑

Mask repetitive sequences using Repeatmasker ☑

Display histogram of score statistics □

Advanced Blast Search

Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK Tel:+44 (0)1223 834244
Registered charity number 210183
Data Release Policy | Conditions of Use | Copyright
webmaster@sanger.ac.uk
Last Modified Tue Oct 28 11:08:07 2008



## Blast Server Results

Retrieve result for id: d01TrBOmfa9782hD3 retrieve

Format: Graphical

Your BLAST query has been added to the queue of jobs.
The majority of BLASTs are completed within two minutes.

To retrieve your results, click the **retrieve** button above, or use the following URL: http://www.sanger.ac.uk/cgi-bin/blast/getblast?id=d01TrBOmfa9782hD3;format=graphic

Click here to start a new blast job

Options: cpus=1 -warnings B=100 -filter=dust V=100
Job id d01TrBOmfa9782hD3.1 status is DONE

```
BLASTN 2.0MP-WashU [04-May-2006] [linux26-i686-ILP32F64 2006-05-09T11:47:08]

Copyright (C) 1996-2006 Washington University, Saint Louis, Missouri USA.
All Rights Reserved.

Reference:  Gish, W. (1996-2006) http://blast.wustl.edu

Notice:  this program and its default parameter settings are optimized to find
nearly identical sequences rapidly.  To identify weak protein similarities
encoded in nucleic acid, use BLASTX, TBLASTN or TBLASTX.

Query=  UNKNOWN-QUERY
        (1767 letters)
```

```
Database:  zebrafish_all.fa
           33,354 sequences; 1,924,132,883 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                Smallest
                                                                  Sum
                                                         High  Probability
Sequences producing High-scoring Segment Pairs:          Score  P(N)      N

AL732458.5                                               1609  4.7e-169  3
CU855879.5                                               1581  7.3e-156  4
AL713987.7                                               1581  5.3e-118  3
BX957295.5                                               1609  9.9e-103  2
zK5E8.00938     EMBL:CU693458   Unfinished sequence: zK5E8     Conti... 1601  1.4e-66   1
AL732411.14                                              1619  8.3e-66   1
AL732436.6                                                732  4.5e-23   1
BX571968.10                                               353  1.0e-20   3
BX663515.13                                               353  1.2e-20   3
CR848668.16                                               426  6.5e-09   1
zK5E8.00101     EMBL:CU693458   Unfinished sequence: zK5E8     Conti...  364  4.4e-06   1
BX470219.6                                                364  4.6e-06   1


>AL732458.5
[Full Sequence] [EMBL:AL732458.5]
```

```
        Length = 84,824

  Plus Strand HSPs:

 Score = 1609 (247.5 bits), Expect = 4.7e-169, Sum P(3) = 4.7e-169
 Identities = 325/329 (98%), Positives = 325/329 (98%), Strand = Plus / Plus
[HSP Sequence]

Query:  1438 GCCAAACTCGCCCTGCAGAATGACTTGGTGAAGCCCATTTCACCAGCCTTCCTCCCGTCA 1497
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 12395 GCCAAACTCGCCCTGCAGAATGACTTGGTGAAGCCCATTTCACCAGCCTTCCTCCCGTCA 12454

Query:  1498 CCCTTCTCTACGACCACAGTCCCGTCCATCTCTCTCCACCCTCGCCCCAACACCTCCATC 1557
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 12455 CCCTTCTCTACGACCACAGTCCCGTCCATCTCTCTCCACCCTCGCCCCAACACCTCCATC 12514

Query:  1558 TTCCAGACGGCTTCACTTCCGCACTCGTTTCTCCGTGCCGCTCCCGGACCCATTCGACCC 1617
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 12515 TTCCAGACGGCTTCACTTCCGCACTCGTTTCTCCGTGCCGCTCCCGGACCCATTCGACCC 12574

Query:  1618 ACTACCGGCTCCATCCTCTTTGCGCCTTACTGAGCGGCTGGATATTGAAACAGTGGTGTA 1677
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 12575 ACTACCGGCTCCATCCTCTTTGCGCCTTACTGAGCGGCTGGATATTGAAACAGTGGTGTA 12634

Query:  1678 TTTATTGAAACTTTATCCTCAACTGTCCTCACCATCACGAAAACCCAGGAAGGATATCGT 1737
             |||||||||||||||||||||||||||||||||||||||||||||||  |||||||||||
Sbjct: 12635 TTTATTGAAACTTTATCCTCAACTGTCCTCACCATCACGAAAACCCAATAAGGATATCGT 12694
```

Missing exons align with CU855879 – finished clone

**Tasks**

1.
How many ORFs does U64564 have and are there any potential protein products and/or domains? (Use OrfFinder) Repeat this for AF177198 and blast the protein product from one of the longest ORFs. What protein hits do you get? Accept this protein and view in fasta format.

2.

Use ClustalW to produce an alignment of human olfactory receptor proteins. Search UniProtKB with OR6 in and select several sequences of your choice. For example: O76002, O76001, Q9GZK4, Q9Y3N9, O76000, P58173. Do the alignment in ClustalW and then copy the aln format alignment into GeneDoc. Open this alignment in GeneDoc and produce an rtf file to import into Word.

3.
Find the rnd1l gene in Vega. Which end falls in the gap ? Can you find the clone that contains the missing exons ?

**Answers:** (These are likely to change due to database updates etc.)

1.

**ORF of U64564**:

There are 6 in the forward frames and 10 in the reverse frames. The longest is in frame one and when blasted there is a hit to myelin oligodendrocyte glycoprotein. The CD search has found an immunoglobulin-like domain.

ORF of AF177198:

There are 39 ORF in total. Blastp of the longest ORF on frame +1 will give hits to the TNL1 (Talin 1). Click on accept and view in fasta protein format. The sequence can be copied and pasted into Word for later use or directly into another program.

**Protein from ORF:**

Various Talin family proteins and B41 and ILWEQ domains are found. Further iterations bring up hits to more distantly related species e.g. Tetraodon, Apis mellifera, Anopheles

Q6NXR9 CDD domains:

DEADc and HELICc domains are detected.

*2.*

*Use ClustalW and GeneDoc to produce an olfactory receptor 6 protein alignment*

Search for OR6 in UniProtKB (or in Entrez).

In Entrez, click on proteins, then select your sequences of choice. Change the display to FASTA and click on 'send to' so the view refreshes.

Copy and paste the sequences into ClustalW. Make sure that the space at the header of each sequence is removed to ensure proper formatting if using Unix. Submit the job.

Copy the ALN format alignment from clustalw, and then open GeneDoc and select file, import. Select clipboard as the input device and aln for filetype. Click on import. Change the colours from within the configuration dialog box and anything else you wish. From the edit menu choose select blocks to copy and click on the blocks. Then select copy selected blocks to, RTF file. Save the file and open in Word.

3.

The 3' end of the rnd1l gene fall in a gap in Vega. Copy the supporting cDNA BC122138 sequence from the EMBL file and blast using Sanger Danio Rerio blast server. The missing exon aligns to finished clone CU862020.