

## Appendix I: Viewing Protein Structures

For some proteins the position of the non-synonymous amino acid substitutions can be visualised in the 3-Dimensional structure of the protein. The structure of SH2D1A has been solved and we will look at it using a piece of standalone software called SwissPDB viewer or Deepview. This software is freely available to download from Expaty (<http://www.expasy.org>).

The first step is to download the file containing the structure of SH2D1A, this can be obtained from the Protein Data Bank (PDB) which is a repository of solved protein structures. The PDB is linked from the Ensembl Gene Report page.

### CHECK GENE ID

Return to the Ensembl Gene Report page for SH2D1A (Ensembl Gene Id = **ENSG00000183918** ) and find the PDB links. Click on ENSP00000360181, then under the external references tab on the left side of the page, click on external identifiers. Scroll down the page to find the PDB links.

SH2 DOMAIN PROTEIN 1A [vi]

**PDB:**



- [1D1Z \[view all locations\]](#)
- [1D4T \[view all locations\]](#)
- [1D4W \[view all locations\]](#)
- [1KA6 \[view all locations\]](#)
- [1KA7 \[view all locations\]](#)
- [1M27 \[view all locations\]](#)

There are six PDB files linked to SH2D1A, these represent different structures. By following these links to PDB you can see that 1D4T is the crystal structure of the XLP Protein SH2D1A in complex with a SLAM peptide.

The screenshot shows the RCSB PDB website interface. At the top right, a yellow box contains the text: "1. Follow the link to the PDB entry for 1D4T". The main content area displays the entry for 1d4t, titled "CRYSTAL STRUCTURE OF THE XLP PROTEIN SAP IN COMPLEX WITH A SLAM PEPTIDE". The entry includes a primary citation, a PubMed abstract, and a molecular description table. The molecular description table lists two molecules: a signaling protein (SH2 DOMAIN) and a signaling lymphocytic activation molecule (SLAM TAIL PEPTIDE).

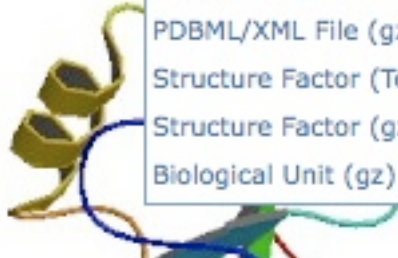
Molecular Description		Hide
Classification:	Signaling Protein	
Structure Weight:	12983.03	
Molecule:	1 CELL SIGNAL TRANSDUCTION MOLECULE SAP	Length: 104
Polymer:	1 Type: polypeptide(L)	
Chains:	A	
Fragment:	SH2 DOMAIN (RESIDUES 1-104)	
Molecule:	SIGNALING LYMPHOCYTIC ACTIVATION MOLECULE	Length: 11
Polymer:	2 Type: polypeptide(L)	
Chains:	B	
Fragment:	SLAM TAIL PEPTIDE (RESIDUES 276 TO 286)	

# 1d4t

 **Display Files** ▾  
 **Download Files** ▾

close

- FASTA Sequence
- PDB File (Text)**
- PDB File (gz)
- mmCIF File
- mmCIF File (gz)
- PDBML/XML File
- PDBML/XML File (gz)
- Structure Factor (Text)
- Structure Factor (gz)
- Biological Unit (gz)

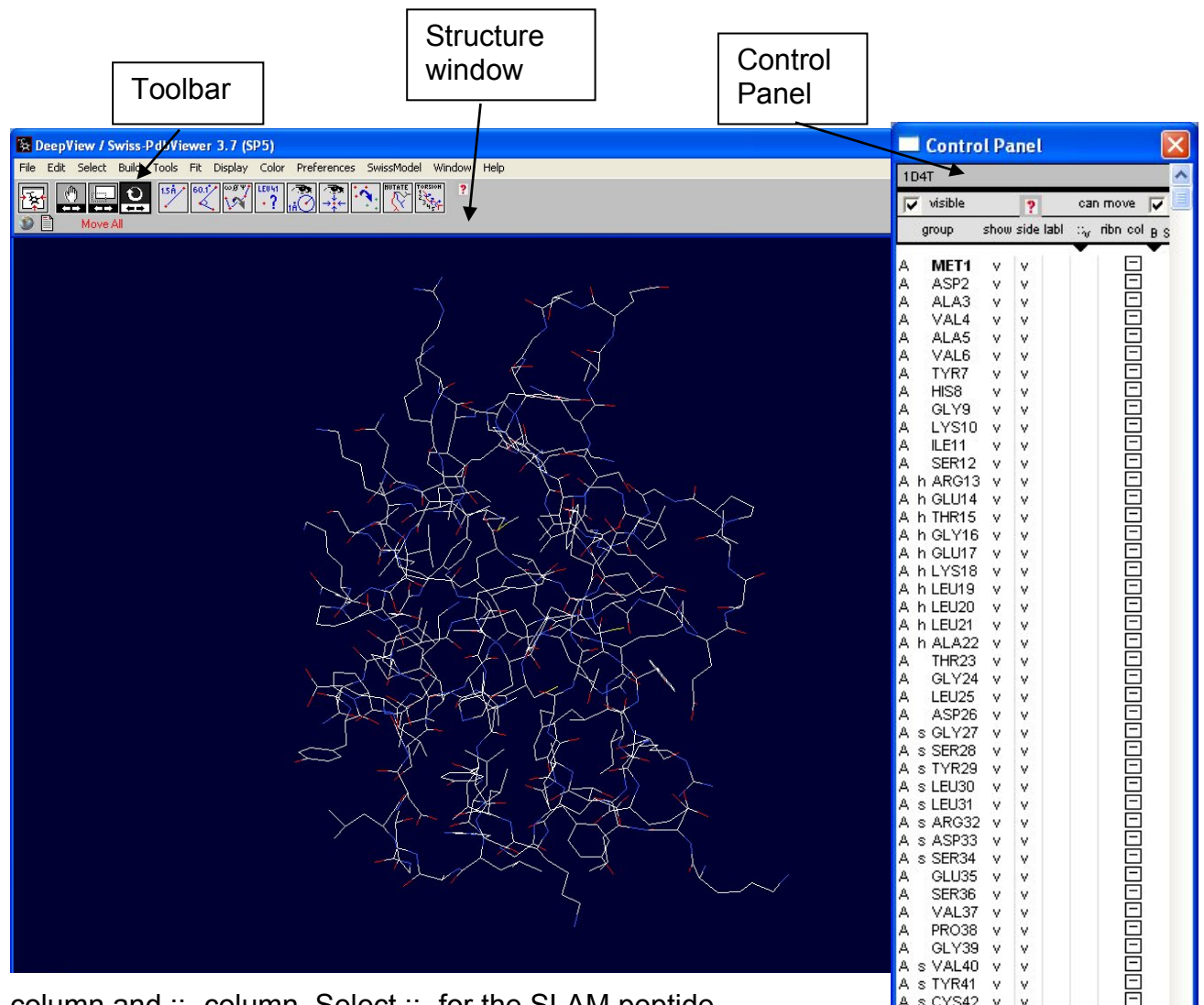


Summary information for the structure

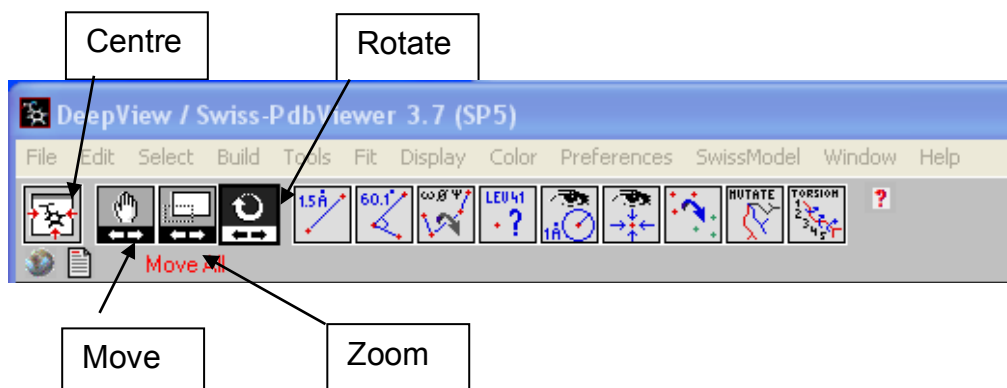
2. Follow the link to Download files and select PDB file (Text)

We now need to open the PDB file in Deepview, a shortcut to which is on the desktop of your PC. To open the file, perform the following steps:

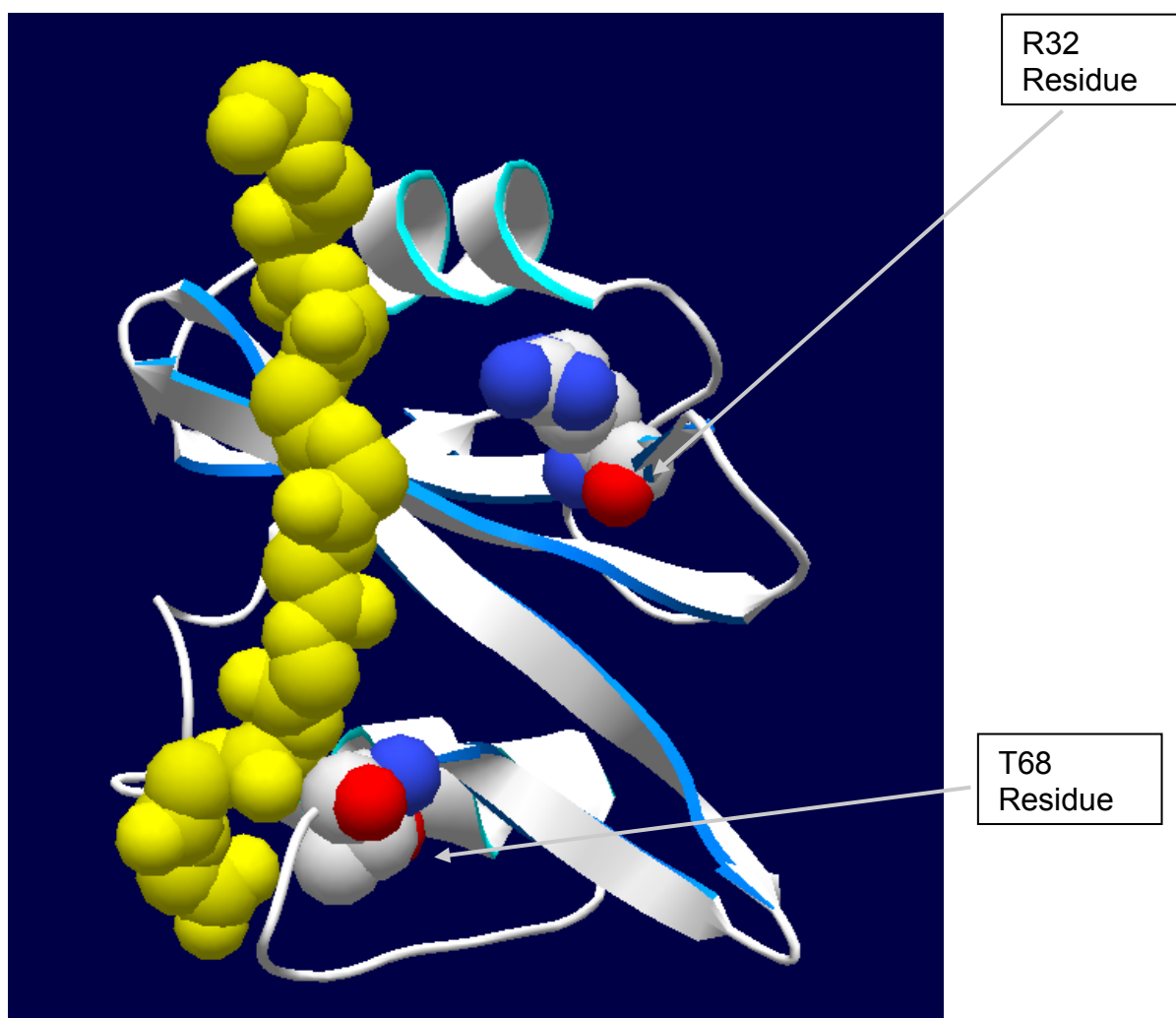
1. Open Deepview
2. From the *File* menu select *Open PDB file* and open the downloaded PDB file (1D4T.pdb)
3. From the *Wind* menu turn on *Control Panel* window



column and ::v column. Select ::v for the SLAM peptide.



4. The crystal structure of the XLP protein SH2D1A in complex with a SLAM peptide. These can be distinguished by assigning different colours to each chain and the polymorphic amino acids – residues 32 and 68 of the SH2D1A chain identified by adding labels:
5. Use the *Control Panel* to colour the SLAM peptide,.
6. Use the *Control Panel* to turn off backbone and side chains except polymorphic residues for SH2D1A and show ribn for all.
7. Use the *Control Panel* to turn on the labels and VDW for the polymorphic residues. This is done by locating the desired residue and clicking on the *labl* column and *::v* column. Select the *::v* for the SLAM peptide.
8. For aesthetics, use *Display* to use open GL rendering and render in solid 3D. Rotate the molecule and zoom in to see position of the residues.



## Appendix II: Comparative Genomics

### I: Genome sequence comparisons and the identification of conserved regions using pre-calculated alignments.

There are now dynamic whole-genome navigation tools available that can be used for visualizing and studying evolutionary relationships between vertebrate and non-vertebrate genomes. These tools have pre-calculated alignments for a variety of sequenced genomes and may obviate the need for completing such alignments yourself.

N.B. Genome assemblies are constantly updated, and you may not be able to access the most recent assemblies when using pre-calculated alignments. If you think that this may affect your results, then please prepare the genome alignments yourself using sequences from relevant species. This can be done at web-sites such as Vista, PipMaker or zPicture, and an example of this follows in this module.

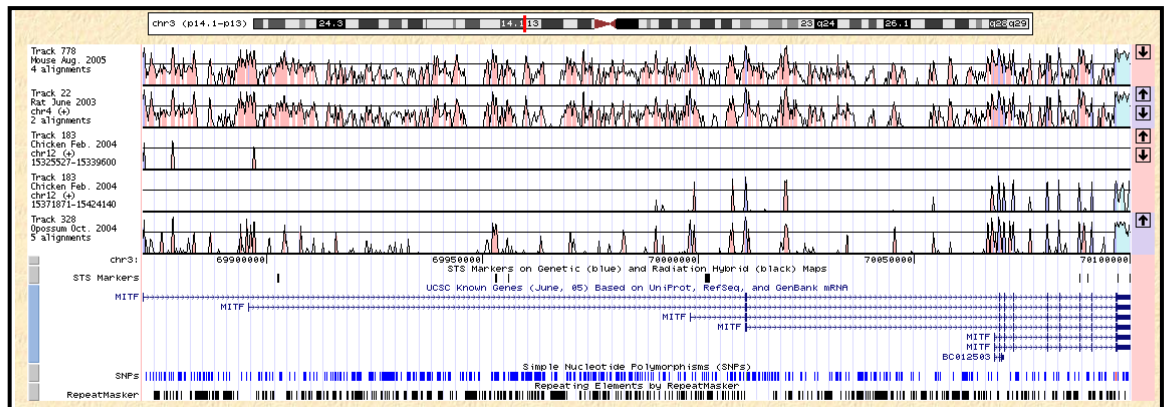
You can identify conserved regions using pre-calculated alignments at the Vista genome browser. Another browser that can be used for such analysis is ECRbrowser.

1. Starting at the Vista homepage select **Vista Browser** under Pre-calculated Whole Genome Alignment option.

**Step 2:**  
Enter the chromosome coordinates for the MITF under Position (chr3:69871323-70100176) and select the option to use VISTA tracks on UCSC browser

1. Press **GO**.

- You will be automatically directed to the UCSC genome browser. Use the controls to displays vista plots for your species of interest and to optimise the UCSC genome browser with your preferred tracks.

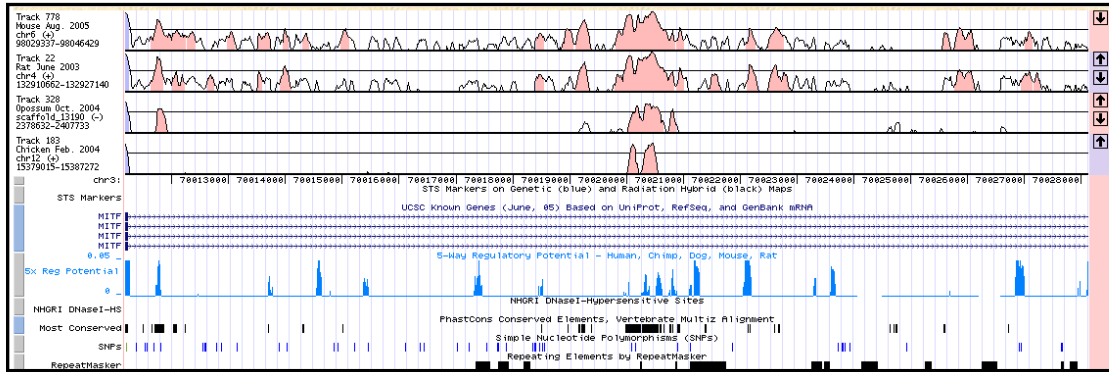


In this example, I have chosen to display the pre-calculated genome alignment results for Human vs: Mouse, Rat, Chicken and Opossum, as well as UCSC known genes, the location of SNPs and Repeats.

The "peaks and valleys" graphs represent percent conservation between aligned sequences at a given coordinate on the base sequence. Multiple alignments that share the base sequence can be displayed simultaneously, one under another. The top and bottom percentage bounds are shown to the right of every row. These bounds can also be adjusted (see how to adjust curve settings). Regions passing a threshold of > 70% identity over 100 bp are coloured. Pink – intragenic conserved regions, blue exonic-conserved regions, light blue – conserved UTR.

The order the species can be altered by using the arrows on the right hand side of the image.

***The main advantage of using the Vista Browser in the UCSC browser is that you can compare the genomic location of a variety of features. For example, most conserved regions identified using PhastCons or 5 x regulatory potential track or even Self-Chain.***



In this example, I have zoomed in on an intragenic conserved element in the *MITF* gene, and also displayed results from the Most Conserved Track, experimentally conserved DNase hypersensitivity sites and the 5 x Regulatory Potential information.

5. Select the track name on the lhs of one of the vista alignments to obtain additional information about the conserved regions. Click on the human-mouse alignment.

You are browsing Human May 2004

aligned with:  
mouse Mouse Aug. 2005  
using the SLAGAN alignment program

human chr3:70,011,178-70,028,129

VISTA tracks on UCSC VISTA Browser

Change Annotation:

Download RefSeq genes Get CNS: **human-mouse**

Location on human	Location on mouse	Alignment
chr3:70,011,178-70,028,129 (+) <a href="#">View sequence</a> <a href="#">Sequence (softmasked)</a> length: 16.95Kb	chr6:98,029,337-98,046,429 (+) <a href="#">Sequence (softmasked)</a> length: 17.09Kb <a href="#">VISTA Browser</a>	Alignment: <a href="#">human-mouse</a> MFA: <a href="#">human-mouse</a> CNS: <a href="#">human-mouse</a> rVISTA: <a href="#">human-mouse</a> PDF: <a href="#">human-mouse</a>

**Step 6:**  
Select Get CNS:human-mouse to access the genome co-ordinates for regions that appear to be conserved between the two species

\*\*\*\*\* Conserved regions - Human May 2004 chr3 (Mouse Aug. 2005 chr6) \*\*\*\*\*

70011178 (98029337) to 70011224 (98029383) = 47bp at 97.8% exon
70011694 (98029787) to 70011882 (98030042) = 256bp at 74.6% noncoding
70011886 (98030046) to 70012234 (98030384) = 363bp at 68.6% noncoding
70012263 (98030412) to 70012397 (98030545) = 133bp at 71.0% noncoding
70012974 (98031100) to 70013086 (98031217) = 119bp at 72.3% noncoding
70013485 (98031649) to 70013583 (98031747) = 102bp at 70.6% noncoding
70013597 (98031748) to 70013721 (98031879) = 133bp at 70.7% noncoding
70013946 (98032081) to 70014039 (98032173) = 94bp at 76.6% noncoding
70014303 (98032422) to 70014403 (98032521) = 102bp at 70.6% noncoding
70014965 (98033056) to 70015106 (98033199) = 141bp at 74.8% noncoding
70018421 (98036313) to 70018546 (98036436) = 126bp at 69.8% noncoding
70018882 (98036770) to 70019059 (98036942) = 178bp at 70.9% noncoding
70019103 (98036964) to 70019385 (98037253) = 282bp at 75.2% noncoding
70019813 (98038164) to 70020758 (98039080) = 952bp at 82.2% noncoding
70020768 (98039109) to 70020857 (98039159) = 90bp at 80.0% noncoding
70020895 (98039203) to 70021012 (98039315) = 119bp at 72.9% noncoding
70022119 (98041038) to 70022268 (98041187) = 150bp at 72.0% noncoding
70022513 (98041415) to 70022767 (98041662) = 253bp at 70.7% noncoding
70025554 (98043937) to 70025694 (98044076) = 140bp at 71.8% noncoding
70025790 (98044151) to 70025961 (98044322) = 173bp at 72.6% noncoding
70025974 (98044335) to 70026119 (98044473) = 145bp at 70.5% noncoding
70027032 (98045161) to 70027149 (98045272) = 118bp at 72.9% noncoding

Total 4254bp at 74.7%

Here you have a list of genome co-ordinates for the conserved regions in human and mouse. This data can be transferred to a Microsoft excel spreadsheet for additional analysis. Such as comparing with the conserved regions identified using an alternative method such as the ECR browser

**zPicture: To manually upload genome sequence to identify conserved regions.**

In this example we will use zPicture <http://zpicture.dcode.org/>. We will upload the human sequence for the MITF directly from the UCSC website; this will automatically create an annotation file for us. We will align this sequence to the orthologous region in the mouse.

The screenshot shows the zPicture web interface. At the top, there is a header with the URL <http://www.dcode.org> and the text "Comparative Genomics Center at Lawrence Livermore National Laboratory". Below this is a navigation bar with "Instructions", "Example: human-rat-fugu and human-rat GATA3 alignment", and "Description". A text box explains that zPicture is a dynamic alignment and visualization tool based on blastz, and that alignments can be submitted to rVista 2.0. A link to "multi-zPicture: multiple sequence alignment tool" is also present.

The main interface is divided into four numbered sections:

- 1 SEQUENCE 1**: "Upload sequence and gene annotation from UCSC Genome Browser". Options include "Paste sequence (in FASTA format)", "FASTA file (.fa)", and "NCBI accession #".
- 2 SEQUENCE 2**: Similar to Sequence 1, but for a second sequence.
- 3 OPTIONAL :: ANNOTATION 1**: "Repeats" section with options for "Repeats are identified by lower-case letters" and "Mask repetitive elements". "Gene annotation (if any)" section with "Paste" and "File" options.
- 4 OPTIONAL :: ANNOTATION 2**: Similar to Annotation 1, but for the second sequence.

At the bottom, there are checkboxes for "Select to run 'fast' BlastZ on microbial-size genomes" and "Select to perform 'chained' (global) blastz alignment", followed by a yellow "SUBMIT" button.

You have the option to align 2, 3 or more sequences

Step 1. Click on the UCSC hyperlink so the co-ordinates for human MITF can

The screenshot shows the UCSC Genome Bioinformatics website. The main header is "UCSC Genome Bioinformatics". Below it is a navigation bar with "Genomes", "Blat", "Tables", "Gene Sorter", "PCR", "Proteome", "FAQ", and "Help". A sidebar on the left contains links for "Genome Browser", "ENCODE", "Blat", "Table Browser", and "Gene Sorter". The main content area is titled "About the UCSC Genome Bioinformatics Site" and contains text about the reference sequence and working draft assemblies for a large collection of genomes.

Step 2. Click on Genomes to access the human genome sequence

The screenshot shows the "Human (Homo sapiens) Genome Browser Gateway". The header includes "Home", "Genomes", "Blat", "Tables", "Gene Sorter", "PCR", "Session", "FAQ", and "Help". Below the header is a text box stating "The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz. Software Copyright (c) The Regents of the University of California. All rights reserved." Below this is a search form with the following fields:

- clade: Vertebrate (dropdown)
- genome: Human (dropdown)
- assembly: Mar. 2006 (dropdown)
- position or search term: MITF (text input)
- image width: 1000 (text input)

There is a "submit" button next to the image width field. Below the search form are three buttons: "add custom tracks", "configure tracks and display", and "clear position". A link "Click here to reset the browser user interface settings to their defaults." is also present.

Step 3. Choose your species (Human) and enter the HGNC name for your gene (MITF), press

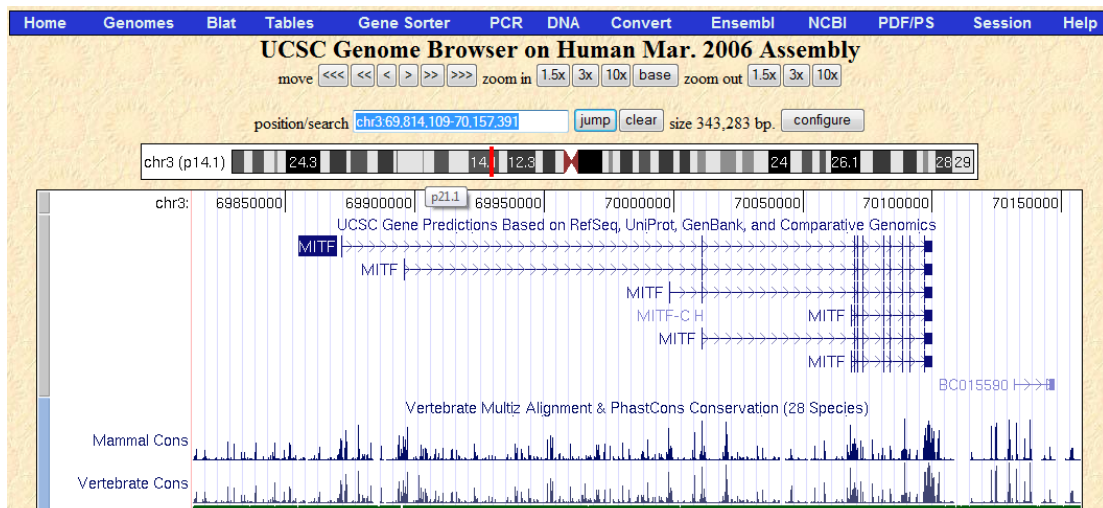


### UCSC Genes

[MITF \(uc003dof.1\) at chr3:70068443-70100177](#) - microphthalmia-associated transcription factor  
[MITF \(uc003dce.1\) at chr3:70068443-70100177](#) - microphthalmia-associated transcription factor  
[MITF \(uc003dod.1\) at chr3:70010946-70100177](#) - microphthalmia-associated transcription factor  
[MITF-C \(uc003doc.1\) at chr3:70008181-70011224](#) - Homo sapiens mRNA for A-type microphthalmia associat  
[MITF \(uc003dob.1\) at chr3:69998132-70100177](#) - microphthalmia-associated transcription factor  
[MITF \(uc003doa.1\) at chr3:69895652-70100177](#) - microphthalmia-associated transcription factor  
[MITF \(uc003dnz.1\) at chr3:69871323-70100177](#) - microphthalmia-associated transcription factor  
[TYR \(uc001pcs.1\) at chr11:88550688-88668575](#) - tyrosinase precursor  
[PIAS3 \(uc001eoc.1\) at chr1:144287345-144297903](#) - protein inhibitor of activated STAT, 3

Step 4. Choose Known Gene MITF with the longest transcript

Step 5. Zoom out 1.5x and copy the chromosome ordinates (Ctrl-C) from the position field (you may want to keep this page open in a separate window).



Return to the zPicture server.

<http://www.dcode.org> - Comparative Genomics Center at Lawrence Livermore National Laboratory

[Instructions](#)    **Example: human-rat-fugu and human-rat GATA3 alignment**    [Description](#)

**zPicture** is a dynamic alignment and visualization tool that is based on **blastz** alignment program utilized by **PipMaker**. zPicture alignments can be automatically submitted to **rVista 2.0** to identify conserved transcription factor binding sites. [Genome Research, 14\(3\), 472-477, \(2004\)](#)

[multi-zPicture: multiple sequence alignment tool](#)

**1**    **SEQUENCE 1**

**Upload** sequence and gene annotation from [UCSC Genome Browser](#)

Paste sequence (in FASTA format)

FASTA file (.fa)  no file selected

NCBI accession #

**2**    **SEQUENCE 2**

**Upload** sequence and gene annotation from [UCSC Genome Browser](#)

Paste sequence (in FASTA format)

FASTA file (.fa)  no file selected

NCBI accession #

**3**    **OPTIONAL :: ANNOTATION 1**

**Repeats:**

Repeats are identified by lower-case letters

Mask repetitive elements

**Gene annotation (if any):**

Paste

File  no file selected

**4**    **OPTIONAL :: ANNOTATION 2**

**Repeats:**

Repeats are identified by lower-case letters

Mask repetitive elements

**Gene annotation (if any):**

Paste

File  no file selected

Select to run "fast" BlastZ on microbial-size genomes  
 Select to perform "chained" (global) blastz alignment

Step 6. Click on Upload to access sequence information from

Alternatively you can paste your sequence of interest into the box, upload from a file (sequences need to be in Fasta format) or input an NCBI accession number.

Request ID: [04271213013871](#)    <http://zpicture.dcode.org/>

**Step 1.** Select a species, genomic interval and genome assembly freeze from the [UCSC Genome Browser](#) (link will open in a new window)

**Step 2.** Describe selected region using the form below

Organism

Assembly

Annotation

Position  (Format: chr7:1000-2000)

**Step 3.** Submit your request for verification

Please read [zPicture instructions](#) or us know ([dcode@ncbi.nlm.nih.gov](mailto:dcode@ncbi.nlm.nih.gov)) if you encounter troubles

Step 7. Select your Organism (human), type of annotation file, and paste the co-ordinates that you copied from the UCSC browser and then submit your request for verification.

TIP: It is essential that the assemblies from which you extracted the genome coordinates for your gene of interest match those in the submission form at zPicture. In this case we are using the Mar. 2006 assembly of the human genome sequence.

Request ID: **08051221815643**

Fetching **Human(hg17)** sequence at **3:69871322-70100176...** ok ([seq1.fa](#))

Fetching **refFlat** gene annotation for this region... ok ([anno1.txt](#))

Forward uploaded data to zPicture **SUBMIT**

A fasta file (seq1.fa) of the genomic sequence has been generated as has an annotation file for the region (anno1.txt)

Step 8. Press submit

The screenshot shows the zPicture web interface with the following elements and callouts:

- 1** **SEQUENCE 1**: A green checkmark and the text "Sequence accepted seq1.fa 228655 bps".
- 2** **SEQUENCE 2**: An "Upload" section with options for "Paste sequence (in FASTA format)", "FASTA file (.fa)", "Choose File" (no file selected), and "NCBI accession #".
- 3** **OPTIONAL :: ANNOTATION 1**: A "Gene annotation (if any):" section with "Paste" and "File" options. The "Paste" option is selected, showing a list of gene annotations.
- 4** **OPTIONAL :: ANNOTATION 2**: A "Gene annotation (if any):" section with "Paste" and "File" options.

Two text boxes provide instructions:

- Top right box:** "Check that the sequence file and annotation file have been uploaded. This procedure can also be repeated for additional species." (Arrows point to the 'SEQUENCE 1' and 'SEQUENCE 2' sections).
- Bottom right box:** "Step 9. Choose to mask human repeats." (Arrows point to the 'Mask repetitive elements' dropdown menus in the 'OPTIONAL :: ANNOTATION 1' and 'OPTIONAL :: ANNOTATION 2' sections).

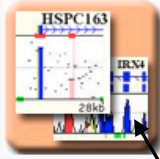
At the bottom of the interface, there are checkboxes for "Select  to run 'fast' BlastZ on microbial-size genomes" and "Select  to perform 'chained' (global) blastz alignment", followed by a yellow **SUBMIT** button.

NOTE: - when the sequences are downloaded from UCSC the sequence is automatically softmasked (repeats are changed to lower case rather than Ns).


Step 10. Repeat the process for the second sequence using the mouse *Mitf* sequence.

Request ID: [08051221815643](#) <http://zpicture.dcode.org/>

**Dynamic visualization:**



**Dot-plot:**



**Update annotation:**

edit [anno1](#) [anno2](#)  
[sequence titles](#)

**rVista 2.0 portal:**

submit alignment to [RVISTA](#)

**Output files:**

list of ECRs in [seq1](#) or [seq2](#)  
 blast-type alignment [seq1\\_seq2.blast](#)  
 blastz alignment [seq1\\_seq2.blastz](#)

**Input files:**

	1	2
sequence	<a href="#">seq1.fa</a>	<a href="#">seq2.fa</a>
seq. masked	<a href="#">seq1.txt</a>	<a href="#">seq2.txt</a>
repeats	<a href="#">seq1.reps</a>	<a href="#">seq2.reps</a>
annotation	<a href="#">anno1.txt</a>	<a href="#">anno2.txt</a>

Contact Ivan Ovcharenko ([ovcharenko1@lnl.gov](mailto:ovcharenko1@lnl.gov)) if you have any questions or suggestions

A variety of results are presented. You can:

- View the alignment of the genome sequences.
- View a dot plot of the sequences.
- Alter the annotation files or sequence titles.
- Analyze the sequences for conserved transcription factor binding sites using rVista.
- View the evolutionarily conserved regions (ECRs) as alignments or lists.

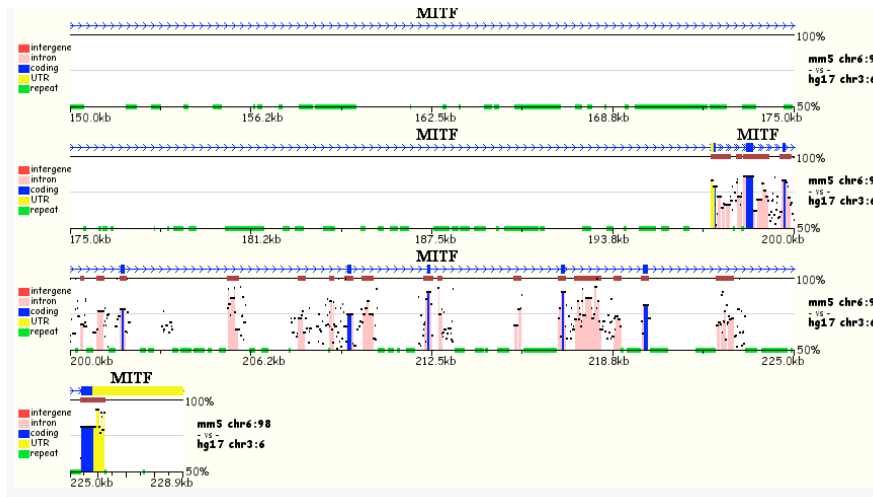
Step 11. Click on the dynamic visualization.

Request ID: [08051221815643](#) <http://zpicture.dcode.org/>

Picture settings	Smooth graph	Base-top switch	Width	ECR length	ECR similarity	Bottom cut-off	Graph height	Remove legend
	<input type="checkbox"/>	<input type="checkbox"/>	25000 bases	at least 100 bases	at least 70 %	50 %	120 pixels	<input type="checkbox"/>

[Refresh](#)

It is possible to alter the visualization. The graph can be smoothed to look like a VISTA plot (smooth graph). The graph can be widened or reduced. The parameters through which an ECR is identified can be altered (ECR length and ECR similarity). The graph height can also be altered.



Step 12. Click on any ECR.

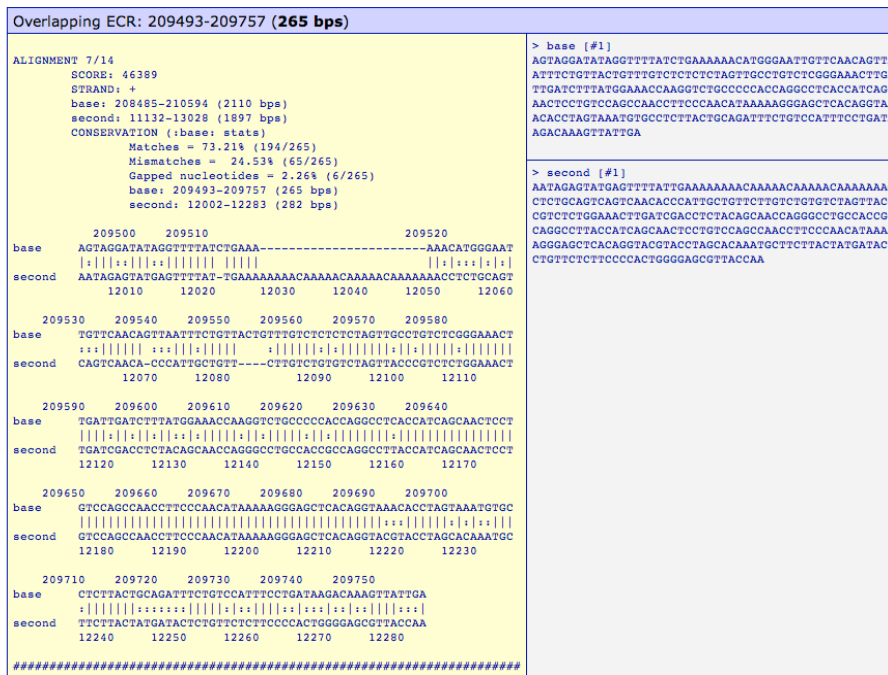
Legend explains colour coding.

Direction of gene transcription is indicated by blue arrows in the gene's structure.

ECRs are displayed in brown.

RepeatMasking is shown in green.

This will give you information about the ECR:



Alignment between the two species for the ECR.

The sequence of the ECR in fasta format (can be used for primer design, BLAST, etc.).

This is the basic form of zPicture. Adapting for more complex analysis is simple. For instance, for more than 2 sequences use multi-zPicture and follow the instructions.

zPicture also allows for regulatory information to be added. Simply return to the results page and click on "submit alignment to rVista" and follow instructions.

**Finally,**

If you are performing large scale comparative analysis you may wish to compare different datasets, for shared or varied features. It is possible to perform such types of analysis *if the dataset are lists of genomic co-ordinates*. For example, you wish to determine which of the “Most Conserved” regions of the human genome are exonic. This type of analysis can be performed using the suite of Comparative Analysis programmes hosted at Galaxy (<http://www.bx.psu.edu/cgi-bin/trac.cgi>)

Unfortunately we don't have time to cover this programme in detail. In brief, this programme allows you to:

- download multiple fragments of the genome sequence simultaneously
- download genomic features from the UCSC genome browser
- perform phylogenetic analysis
- analysis sequence properties and characteristics using the Emboss suite of analysis programmes
- compare different datasets of genomic co-ordinates for overlapping regions, different regions, proximal regions etc.,

Should you wish to know further details about this programme please don't hesitate to ask.

**Selected References**

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I.

VISTA: computational tools for comparative genomics.

Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W273-9.

Nardone J, Lee DU, Ansel KM, Rao A.

Bioinformatics for the 'bench biologist': how to find regulatory regions in genomic DNA. Nat Immunol. 2004 Aug;5(8):768-74.

Ovcharenko I, Loots GG, Hardison RC, Miller W, Stubbs L.

zPicture: dynamic alignment and visualization tool for analyzing conservation profiles. Genome Res. 2004 Mar;14(3):472-7.

Ovcharenko I, Nobrega MA, Loots GG, Stubbs L.

ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes.

Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W280-6.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D.

Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005 Aug;15(8):1034-50. Epub 2005 Jul 15.

## II: Do it yourself!

A variety of different programs can be used to compare genome sequences, with the most commonly used programs being **Zpicture**, **Vista** and **PipMaker**. All of these websites have comprehensive notes that explain the underlying processes that generate the alignments. There are five steps to be completed when conducting comparative genome sequences analysis. They are:

- i) Extracting the base sequence
- ii) Generating an annotation file to highlight the location of exons. These files are automatically generated by zPicture, but must be generated when performing multiple genome sequence alignments in either PipMaker or Vista.
- iii) Extracting additional sequences to compare to the base sequence. This can be done a number of different ways: each of which will be demonstrated to you.
- iv) Masking out repetitive sequences.
- v) Finally, completing the sequence comparison.

**Further details on how to extract, annotate, repeat mask, and align sequences can be found below. Please feel free to complete this task in your own time. Also, any of the instructors will be able to help you with this.**

- i) Extract the base sequence

When analysing genome sequences zPicture there are three possible ways to enter your genomic sequence to be analysed. They are:

- by using genome sequences that you have exported from another source
- by using a GI (accession) number
- by uploading the sequence directly from the UCSC genome browser (as we did in the previous section)

To export the base sequence (**genome sequence 1**) from Ensembl go to the GeneView page for your gene of interest, in this case MITF in human Ensembl.



**Ensembl 52: H.sapiens - Gene summary - Gene: MIF (ENSG00000187098) - Iceape**

Microphthalmia-associated transcription factor [Source: UniProtKB/Swiss-Prot; Acc: P07303]

Location: Chromosome 3: 69,871,276-70,100,177 forward strand

Transcripts: There are 7 transcripts in this gene. [hide transcripts](#)

Transcript	ENST	EST	Protein
MIF-001	ENST00000394355	ENST00000372884	protein_coding
MIF-004	ENST00000314552	ENSP00000324546	protein_coding
MIF-005	ENST00000384351	ENSP00000372880	protein_coding
MIF-006	ENST00000384349	ENSP00000372877	protein_coding
MIF-007	ENST00000352241	ENSP00000395600	protein_coding
MIF-201	ENST00000314528	ENSP00000324443	protein_coding
MIF-202	ENST00000329528	ENSP00000329667	protein_coding

**Gene summary** [help](#)

Name: MIF (HGNC (curated))

Synonyms: bHLH32, MI, W52A [To view all Ensembl genes linked to this name [click here](#)]

CCDS: This gene is a member of the Human CCDS set: [CCDS2813](#), [CCDS43106](#), [CCDS43107](#)

Gene type: Known protein coding

Prediction Method: Gene containing both Ensembl generated transcripts and [Havana](#) manual curation, see [article](#).

Transcripts:

Select your transcript

**Ensembl 52: H.sapiens - Transcript summary - Transcript: MIF-001 (ENST00000394355) - Iceape**

Microphthalmia-associated transcription factor [Source: UniProtKB/Swiss-Prot; Acc: O75030]

Location: Chromosome 3: 69,871,276-70,100,177 forward strand

Gene: This transcript is a product of gene ENSG00000187098 - There are 7 transcripts in this gene. [hide transcripts](#)

Transcript	ENST	EST	Protein
MIF-001	ENST00000394355	ENST00000372884	protein_coding
MIF-004	ENST00000314552	ENSP00000324546	protein_coding
MIF-005	ENST00000384351	ENSP00000372880	protein_coding
MIF-006	ENST00000384349	ENSP00000372877	protein_coding
MIF-007	ENST00000352241	ENSP00000395600	protein_coding
MIF-201	ENST00000314528	ENSP00000324443	protein_coding
MIF-202	ENST00000329528	ENSP00000329667	protein_coding

**Transcript summary** [help](#) [Export data](#)

Exons: 10 Transcript length: 4,814 kbp Translation length: 520 residues

CCDS: This transcript is a member of the Human CCDS set: [CCDS43106](#)

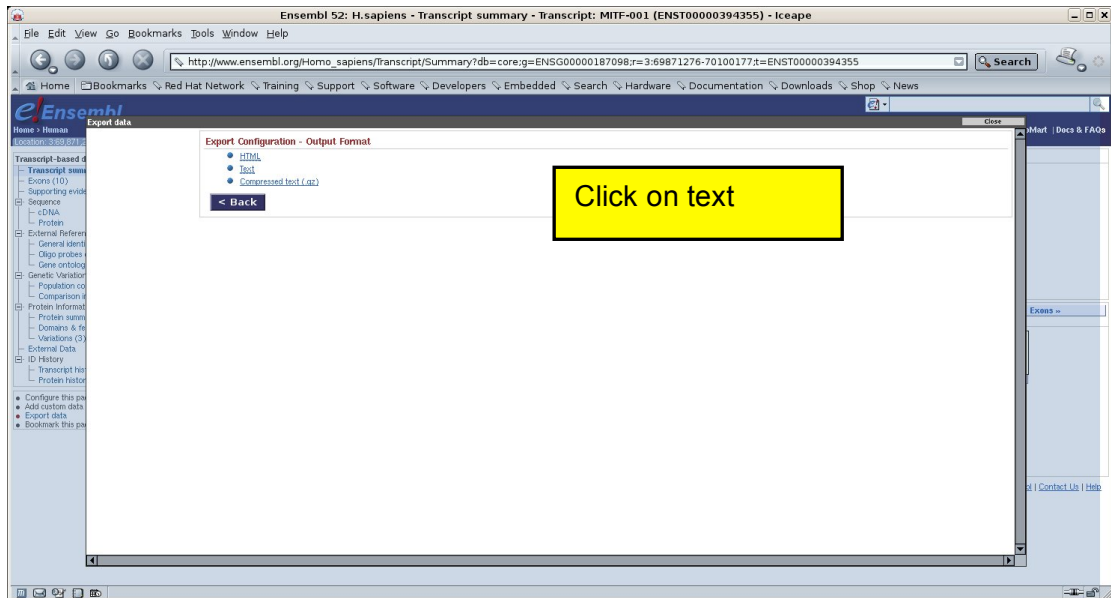
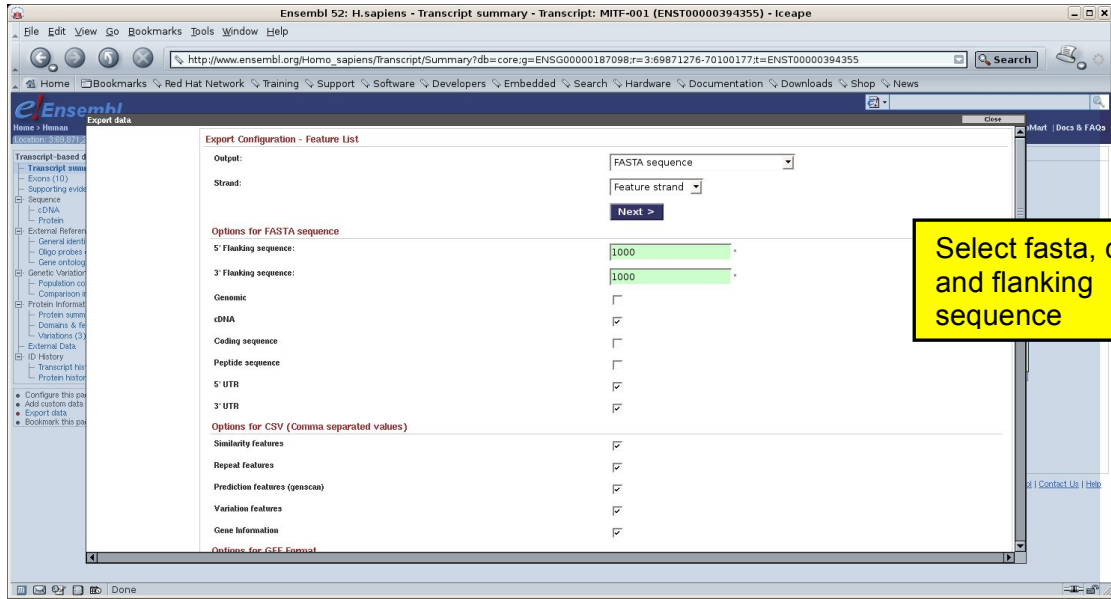
Type: Known protein coding

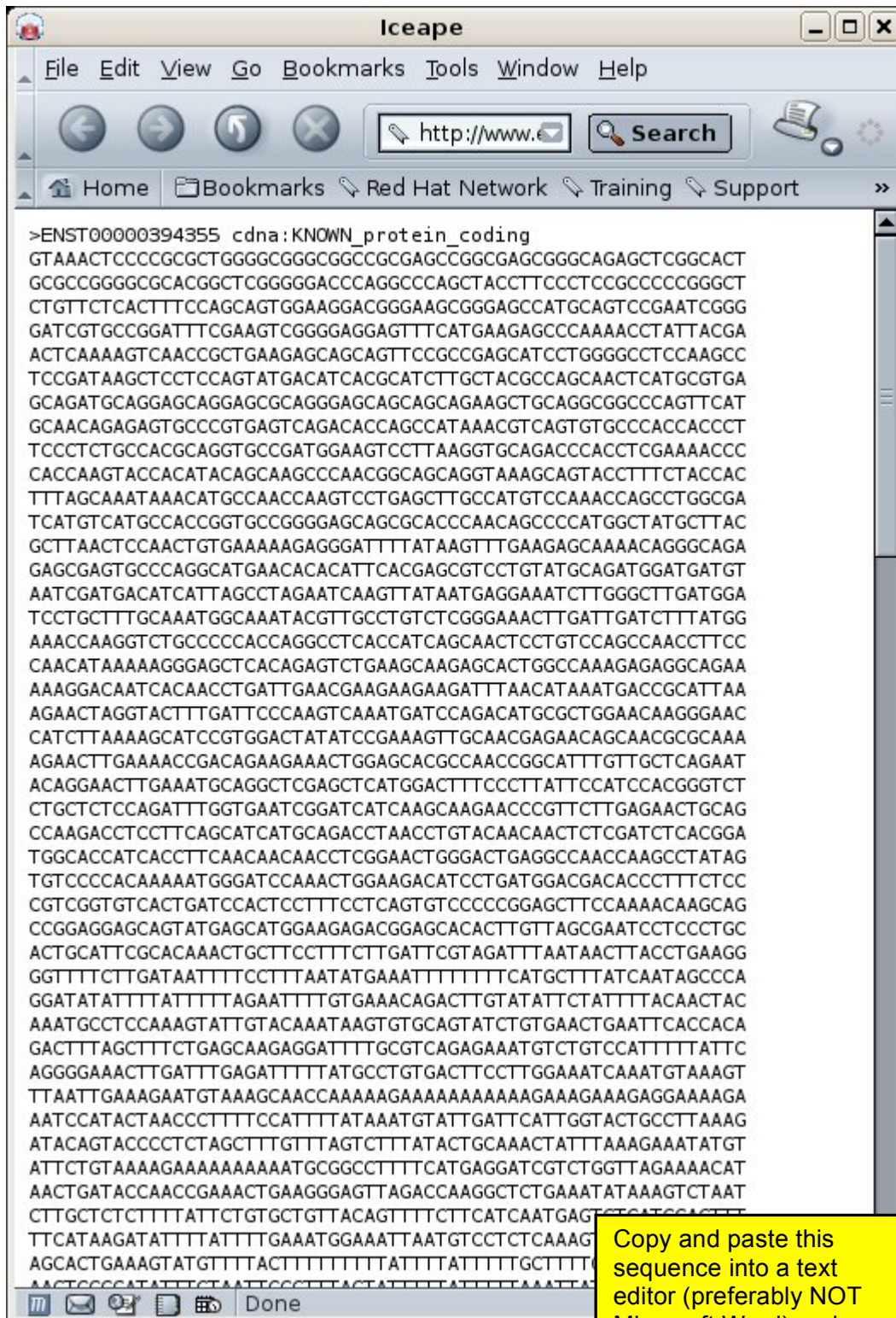
Prediction Method: Manually annotated transcripts (determined on a case-by-case basis) from the [Havana](#) project.

Alternative transcripts: This Vega Havana gene entry corresponds to the following database identifiers: [Havana transcript](#): [MIF](#), [MIF](#) [see all locations]

Ensembl release 52 - Dec 2006 © WTSI / EBI  
Permanent link - View in archive site

Click on export data





Other methods for extracting genomic sequences will be covered in section iii.

ii) Generate an annotation file

(N.B: This is ONLY necessary when using PipMaker and VISTA and will not be completed in today's demonstration)

The annotation file identifies the position of the exons of your gene of interest (**cDNA sequence**) in the base sequence (**gene sequence 1**). The annotation for both PIP and VISTA requires an alignment between the cDNA sequence and the base genomic sequence. There are 3 places do this

- a SPIDEY – see module 3
- b SIM4. Generates the output in the correct format (beware – SIM4 can be a bit temperamental).
- c Ensembl (does it all for you!)

**Using SIM4**

Export the cDNA sequence as above.

Bring up Sim4: <http://pbil.univ-lyon1.fr/sim4.php>



SIM4 addresses the problem of efficiently aligning a transcribed and spliced DNA sequence (mRNA, EST) with a genomic sequence containing that gene, allowing for introns in the genomic sequence (taking into account consensus splice signals) and a relatively small number of sequencing errors.

**cDNA Sequence:**  
 Origin:   
 Sequence name (or bank ID/AC):   
 Strand:  Direct  Reverse  
 If user entered, paste raw sequence below:  
 TGGGAAAAGTTGATGCTCAATAACAGTATAAAACAGCCCTATTTCTTGATAAAAA  
 ATGAC  
 AAATGACTGTCTTGGCGATGCTGGTACTGTAATGTTAATAGTCACCTGCT  
 GTTC  
 GATCCAGCAATAATTTCTGTATGGTCCATAGCACTGTATATTGGATCGATATTA  
 ATGT  
 ATCCCAATGAAATAATCGACTTGTCTTGATAGCCCTATTAAGCATTTGGTTTTTC  
 AC

**Genomic Sequence:**  
 Origin:   
 Sequence name (or bank ID/AC):   
 Strand:  Direct  Reverse  
 If user entered, paste raw sequence below:  
 AAGGA  
 ACATAGAGTTGGATCAGGCTGAATTTATTGATTGGCCCACTAAGTAGGGA  
 TGCCAT  
 TTAATATTGGACCTTGGGAGTTAAAAAATGTTCTAATCATCTATTTGCTTGTTA  
 CCTC  
 AAATATGAATTAAGTCTGTAGTGAAGTGAATGCGCTGATCCCTTGGTTTAA  
 CTAGA  
 TGAAGGGATCCAAAGGCTTAGGGAAGATTGGGA

**SIM4 parameters:**  
 W - word size:   
 X - value for terminating word extensions:   
 N - accuracy of sequences:

Cut and paste cDNA sequence and genomic sequence, scroll down and press submit.

## SIM4 Output


Under this page you can:

1. Take a look at the alignments found by [SIM4](#) ([text format](#)).
2. Visualize the alignments with the [LalnView](#) program (MIME-type: *chemical/x-ain2*).
3. Check out the two nucleotide sequences used in the alignment ([Seq1](#) and [Seq2](#)).

LalnView is a graphical viewer for pairwise sequence alignments. Click [here](#) to take a look at a screenshot. You can download LalnView 2.2 [here](#) (UNIX, Mac and PC versions available).

**Important note:** if you want to visualize the alignments produced by this server with LalnView, you need to have version 2.2 of this program installed on your computer. Earlier versions will not work.

[If you have problems or comments...](#)

 [Back to PBL home page](#)

Choose to view SIM4 alignment

```
seq1 = Seq1, 4618 bp
seq2 = Seq2 (Seq2), 248852 bp

>Seq1 (4618 nucleotides)
>Seq2 (248852 nucleotides)

1-56 (136810-136865) 100% ->
57-306 (149653-149902) 100% ->
307-534 (208341-208568) 100% ->
535-618 (209617-209700) 100% ->
619-714 (211755-211850) 100% ->
715-832 (219570-219687) 100% ->
833-907 (222331-222405) 100% ->
908-983 (226974-227049) 100% ->
984-1131 (229792-229939) 100% ->
1132-4618 (235366-238852) 100%

      0          1          2          3          4          5          6          7          8          9
1 ATCGAGCCGCTTAGAGTTCAGATGTTTCATGCCATGCTCCTTTGAAAGCTT
136810 ATGGAGCCGCTTAGAGTTCAGATGTTTCATGCCATGCTCCTTTGAAAGCTT

      50          51          52          53          54          55          56          57          58          59
51 GTATCT          CAGTCCCGCCGAGCATCCTGGGGCCCTCCAAGCCTC
136860 GTATCTGTA...CAGCAGTCCCGCCGAGCATCCTGGGGCCCTCCAAGCCTC

      100         101         102         103         104         105         106         107         108         109
92 CGATAAGCTCCTCCAGTATGACATCAGCATCTTGCTAGCCAGCAACTC
149688 CGATAAGCTCCTCCAGTATGACATCAGCATCTTGCTAGCCAGCAACTC

      150         151         152         153         154         155         156         157         158         159
142 ATCCCTGACGAGATGCAGGAGCAGGAGCCAGGGACGACGACGCAAGCT
149738 ATCCCTGACGAGATGCAGGAGCAGGAGCCAGGGACGACGACGCAAGCT

      200         201         202         203         204         205         206         207         208         209
192 CCAGGCCGCCAGTTCATGCAACAGAGAGTGCCTTGGAGTCAGACACCAG
149788 CCAGGCCGCCAGTTCATGCAACAGAGAGTGCCTTGGAGTCAGACACCAG

      250         251         252         253         254         255         256         257         258         259
242 CCATAAAGCTCAGTGTGCCACCACCTTCCCTTGCACGCCAGGTGCCG
149838 CCATAAAGCTCAGTGTGCCACCACCTTCCCTTGCACGCCAGGTGCCG

      300         301         302         303         304         305         306         307         308         309
292 ATCGAAGCTCCTTAAG          GTCCAGACCCACCTCGAAAACCCAC
149888 ATCGAAGCTCCTTAAGTA...CAGGTGCAGACCCACCTCGAAAACCCAC
```

Alignment information:  
  
Check the matches are 100% and that the entire cDNA is aligned.

### EXONS

```
> 136810 238852 Seq1
136810 136865
149653 149902
208341 208568
209617 209700
211755 211850
219570 219687
222331 222405
226974 227049
229792 229939
235366 238852
```

Scroll down to bottom and find results in annotation format for PIP and VISTA. Cut and paste into a text editor and save as text file.

## Using Ensembl

Go back to the Export page for MITF (starting from the GeneView page). You will need to use archive for this as the current release (52) does not yet have this functionality.

Ensembl v32 - Jul 2005

Chromosome 3  
69,871,323 - 70,100,174

View of Chromosome 3  
Graphical view  
Graphical overview  
Export information about region  
Export sequence as FASTA  
Export EMBL file  
Export Gene Info in region  
Export SNP info in region  
Export Vega info in region  
View alongside ...  
View Syntenic regions ...  
View region in NCBI browser  
View region in UCSC browser

ENSG00000187098

Gene Information  
Gene splice site image  
Gene variation info.  
Genomic sequence  
Export data  
Transcript information  
Exon information  
Peptide information

Search of Human: Anything Go  
e.g. AC003985.2.1.104492, RH9632, ENSG00000139618

Select region/feature to Export

Choose one or two features from the same chromosome as anchor points and display the region between them. Both features must be mapped to the current Ensembl golden tiling path. If you select "None" for the second feature, the display will be based around the first feature.

Please note that there is an upper limit of 5Mb that we will export.

Region

Chromosome name/fragment

From (type): Gene: ENSG00000187098

To (type): Base pair

Context

Bp downstream: 10000

Bp upstream: 10000

Output format

Output Format: Pipmaker / zPicture format

Continue >>

Fields marked with \* are required

© 2005 WTSI / EBI. Ensembl is available to [download for public use](#) - please see the [code licence](#) for details.

Export as Pip

Show context of 10000 bp either side of feature

Choose either PipMaker/zPicture OR Vista

Export individual files

Ensembl v32 - Jul 2005

Chromosome 3  
69,861,323 - 70,110,174

View of Chromosome 3  
Graphical view  
Graphical overview  
Export information about region  
Export sequence as FASTA  
Export EMBL file  
Export Gene Info in region  
Export SNP info in region  
Export Vega info in region  
View alongside ...  
View Syntenic regions ...  
View region in NCBI browser  
View region in UCSC browser

ENSG00000187098

Search of Human: Anything Go  
e.g. AC003985.2.1.104492, RH9632, ENSG00000139618

Configuring PIP (%age identity plot) output for Pipmaker / zPicture format

PIP format options

Your export has been processed successfully. Please download the exported data by following the links below.

Sequence data: [UHQbTPfCBOTGYAAaOT.fa](#) [FASTA format]  
Annotation data: [UHQbTPfCBOTGYAAaOT.txt](#) [pipmaker format]

OR

Combined file: [UHQbTPfCBOTGYAAaOT.tar.gz](#)

© 2005 WTSI / EBI. Ensembl is available to [download for public use](#) - please see the [code licence](#) for details.

Export individual files (right click and "Save target as..." or similar).

This produces two files, the base genomic sequence and the annotation file, which can be saved as text files and used in any of the comparative analysis packages

### iii) Extracting additional sequences

Repeat this process for each genome sequence required, remembering you can toggle between species using the predicted orthologue section of the GeneView page. However, if the orthologous gene of interest is on the opposite strand, you will need to export your sequence from UCSC (not covered). This is a key difference between exporting in Ensembl and the UCSC.

**iv) Mask out repetitive sequences**

**N.B: Again, this is not necessary when using zPicture and will not be included in today's demonstration.**

For PipMaker, the base genome sequence needs repeatmasking can be done at <http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker> . For VISTA, and zPicture repeats are masked out during the submission process.

## Appendix III: Genotyping and Primer Design

### Genotyping samples

Once you have identified the polymorphisms you are interested in you probably want to look at them in your samples. In recent years there has been an explosion in genotyping techniques and technologies. It is now possible to quickly genotype a single SNP or hundreds of thousands of SNPs in one go. Specialist courses are available to learn about these so we'll present a quick overview here. For most of these applications, the manufacturers provide specialist software for assay design, design assays themselves or offer pre-developed assays.

#### Low Throughput Genotyping

RFLP – PCR of amplicon containing SNP, restriction enzyme digestion and then gel electrophoresis.

<http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/R/RFLPs.html>

Minisequencing – A primer abuts the polymorphism and a sequencing reaction is performed only with labelled dideoxynucleotides (ie only the polymorphic base is sequenced). Can be multiplexed.

[http://www.medsci.uu.se/molmed/PEK/HM\\_Syvanen1999.pdf](http://www.medsci.uu.se/molmed/PEK/HM_Syvanen1999.pdf)

TaqMan – SNP specific probes hybridise to target and fluorescent tags are released by exonuclease activity of Taq polymerase. Assays available off the shelf and also by design. SNPbrowser software is an excellent tool for identifying SNPs of interest with off the shelf assays available.

<http://www.appliedbiosystems.com/>

#### Medium Throughput Genotyping

Mass Spectroscopy – Different SNPs are detected based on the different masses of the polymorphic bases.

[http://www.sequenom.com/applications/high\\_performance\\_genotyping.php](http://www.sequenom.com/applications/high_performance_genotyping.php)

#### High Throughput Genotyping

Parallele – Molecular inversion probes. Oligonucleotide probe central to the process undergoes a unimolecular rearrangement from a molecule that cannot be amplified into a molecule that can be amplified. Up to 20,000 SNPs per reaction.

[http://www.affymetrix.com/technology/mip\\_technology.affx](http://www.affymetrix.com/technology/mip_technology.affx)

Illumina Bead Array – Bead-based microarrays. 317,000 and 550,000 human SNP chip available. Smaller custom arrays with up to 1536 SNPs also available in 8, 16 and 96 sample formats and with 7,600-60,000 SNPs in 12 sample format.

[http://www.illumina.com/technology/life\\_sciences/tech\\_life\\_genotyping.ilmn](http://www.illumina.com/technology/life_sciences/tech_life_genotyping.ilmn)



Affymetrix – Microarray based genotyping by hybridisation to 25mer probes. One chip available with 500,000 human SNPs on. Custom arrays with fewer SNPs also possible.

<http://www.affymetrix.com/products/index.affx>

## Primer Design

Genotyping using techniques such as RFLP and minisequencing require a preliminary amplification of the target SNP and it's surrounding area. The most commonly used tool for primer design is Primer3:

[http://frodo.wi.mit.edu/cgi-bin/primer3/primer3\\_www.cgi](http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi)

The screenshot shows the Primer3 web interface with several key sections:

- Top Navigation:** Links for 'description', 'FAQ', 'contact', and 'code'.
- Input Section:** A text area for pasting a source sequence below (5'→3', string of ACGTNaGn -- other letters treated as N -- numbers and blanks ignored). A dropdown menu is set to 'HUMAN'.
- Options Section:** Three columns of checkboxes for selecting primer types: 'Pick left primer or use left primer below', 'Pick hybridization probe (internal oligo) or use oligo below', and 'Pick right primer or use right primer below (5'→3' on opposite strand)'. The 'Pick right primer' option is checked.
- Advanced Options:** Fields for 'Sequence Id', 'Targets', 'Excluded', and 'Exponent'. A 'Product Size Ranges' field is set to '150-250 100-300 301-400 401-500 501-600 601-700 701-850 851-1000'. A 'Click here to specify the min, opt, and max product sizes only if you absolutely must. Things there is too slow and too computationally intensive for our servers.' link is present.
- General Primer Picking Conditions:** A grid of input fields for parameters like 'Primer Size', 'Primer Tm', 'Product Tm', 'Primer GC%', 'Max Self-Complementarity', 'Max #Prs', 'Inside Target Penalty', 'First Base Index', 'Salt Concentration', and 'Annealing Ckco Concentration'.
- Other Per-Sequence Inputs:** Fields for 'Included Region', 'Start Codon', and 'Exon'.
- Buttons:** 'Pick Primers' and 'Reset Form' buttons are located at the bottom of several sections.

Select the appropriate mispriming library

Enter target sequence

Select options and then hit 'pick primers'

More details of options are available at:

[http://frodo.wi.mit.edu/cgi-bin/primer3/primer3\\_www\\_help.cgi](http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www_help.cgi)

Primer3 Output

```

Using mispriming library kunczy_and_staple.txt
Using 1-based sequence positions

WARNING: Unrecognized base in input sequence

SEQUENCE SIZE: 1428
INCLUDED REGION SIZE: 1428

TARGETS (start, len)*: 349,960

1  Gtaaccgcccactcccggttaagggtctctctgctcagcaccctgtagctggg
61  ccacaggccgcagcagccgcgcagctaacgcttctgtatttttagtagagcaggtttc
121  gccatgttggccatgctggtcttgaccctcgaactcaggtaatccgcccctcggtt
181  cccaaatgctgggttaacaggctgagcactgcgcctgcccagattaaaactttt
241  tttctgtttttttctgagcagagttctgcttttctcatccgggtggagtgcaatggc
301  gccactgggctcactgcacctccaccctccgggtcaagcagctctctctgctcagcc
*****
361  tcccagtagctggaccaccagccaccaccaccaccagctagtttttatttttag
*****
421  tttaggttaagcactcgcctctccgcgaaactggagagctgagcacttggac
*****
481  agtttggcgtgagggtcagctagcactcatccaccaccaccgagggccacta
*****
541  gcccaagtcggtctctgcccaccagtcgcagcagggaactgactatcccccttag
*****
601  cccaaaccagtaactccggagctcgcagcagcctcagcgccttgcggccccc
*****
661  ttgagctcaccacagcctctctccctcttggaccggagcgcctgagcggccgac
*****
721  gcccaattttgtgagcggcgaaggaagctgctgctctccctccctccctccctcc
781  gctgagcctctctccctccctccctccctccctccctccctccctccctccctcc
841  gctgagcctctctccctccctccctccctccctccctccctccctccctccctcc
901  gctgagcctctctccctccctccctccctccctccctccctccctccctccctcc
961  gctgagcctctctccctccctccctccctccctccctccctccctccctccctcc
1021  gctgagcctctctccctccctccctccctccctccctccctccctccctccctcc
1081  gctgagcctctctccctccctccctccctccctccctccctccctccctccctcc
1141  gctgagcctctctccctccctccctccctccctccctccctccctccctccctcc
1201  gctgagcctctctccctccctccctccctccctccctccctccctccctccctcc
1261  gctgagcctctctccctccctccctccctccctccctccctccctccctccctcc
1321  gctgagcctctctccctccctccctccctccctccctccctccctccctccctcc
1381  gctgagcctctctccctccctccctccctccctccctccctccctccctccctcc

KEYS (in order of precedence):
***** target
    
```

No acceptable primers found. Looking at the statistics suggests why this may be.

No Acceptable Primers Were Found

The statistics below should indicate why no acceptable primers were found. Try relaxing various parameters, including the self-complementarity parameters and max and min oligo melting temperatures. For example, for very A-T-rich regions you might have to increase minimum primer size or decrease maximum melting temperature.

Statistics

	con	too	in	in	no	tm	tm	high	high	high	high			
	sid	any	tar	excl	bad	GC	too	too	any	3'	lib	poly	end	
	ered	Na	get	reg	GC	clamp	low	high	compl	compl	sim	X	stab	ok
Left	2691	0	0	0	134	0	273	1923	0	0	0	23	51	185
Right	778	0	0	0	249	0	3	444	0	0	0	19	8	33

Pair Stats:  
 considered 75240, unacceptable product size 75240, ok 0  
 primer3 release 1.0

(primer3\_mwv\_results.cgi v 0.4)

**Primer3**

pick primers from a DNA sequence

Fastq source sequence below (5'→3', string of A/C/G/T/nagtn -- other letters treated as N -- numbers and blanks ignored). FASTA format ok. Please N-out undesirable sequence (vector, ALUs, LINEs, etc.) or use a [Multiple Library \(read file\)](#). HUMAN

Sequence ID: \_\_\_\_\_ A string to identify your output.

Targets: \_\_\_\_\_ E.g. 50.2 means primers to surround the 2 bases at positions 50 and 51. Or mark the source sequence with [ and ] e.g. ...ATCT[CCCC]TCAT... means that primers must flank the central CCCC.

Included Regions: \_\_\_\_\_ E.g. 401,7-6L3 forbids selection of primers in the 7 bases starting at 401 and the 3 bases at 63. Or mark the source sequence with ^ and ^ e.g. ^ATCT^CCCC^TCAT... forbids primers in the central CCCC.

Product Size Range: 1000-1200

Click here to specify the min, max, and max product sizes only if you absolutely must. They often are too slow (and too computationally intensive) for our server.

Number To Return: 5 Max T-Stability: 99

Max Misincorporation: 12.00 Per Max Misincorporation: 24.00

Pick Primers Reset Form

Use the Back button on your browser. Try altering the parameters and trying again. For example, here product size has been increased.

```
Using workspace library hwapr_seq_simple.txt
Using 1-based sequence positions
WARNING: Unrecognized base in input sequence

OL300      TARGET  LEN    TM    GC    MW    3'  T  MW
LEFT PRIMER   273    20    60.38    50.00    5.00    3.00    12.00    ttttgcctcctcagggtggag
RIGHT PRIMER  1413   20    60.62    55.00    4.00    3.00    11.00    agtcttccCGGGGAAAGCTG
SEQUENCE SIZE: 1428
INCLUDED REGION SIZE: 1428

PRODUCT SIZE: 1141, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 0.00
TARGETS (start, len) #: 149,940

1  Gtaacggcaccctccggatccagggtctctctgctcagcaccctggtagtgga

61  cccacgggcacgcacgacgcagcagctawgtgttgtttttttagtagagcagggtttc

121  gcccagttggccatgctggtcttgcaccctgacctcaggtacatccgcgcgctcggtt

181  ccccaaatgctgggtctcaggcgtgagccactgcgcctgocctagattcaaacctctt

241  tttttgtttttttctcagcagcaggtttcgttttttgcctccagggtggagtcctaggt

201  gccccttgggtctccctgcaccctccactccgggttcagggtctctctgctcagcc

361  tcccagtagtgggcctccagcaccgcaccctccacccagctagtttttgtatttttag

421  tagaggtgggttttgccttgttgccagcagcggctctgcacctcctgacctcagtcg

1201  ttgcctcccccagcctctcctccctcttggagcgggagctgctgagcgggctgag

1241  ggcctttttgtgagcggCGAAGGAGGTTGCTGCTCCCTTGGCTCCCGGAAACCTTCC

1321  GACTGGGCTGTCCCGCCGCGGCGAGGC&CTCCGCGGGGGGT&ATTCGGGCTCGG

1381  TTCTGTGCCCGCA&CTTCCCGGtagatcccgccgacctgaa
      ccccccccccccccccccc

KEYS (in order of precedence):
**** target
>>>>> left primer
<<<<<< right primer

ADDITIONAL OL300s

1  LEFT PRIMER      273    20    60.89    50.00    5.00    3.00    11.00    ttttgcctcctcagggtggag
   RIGHT PRIMER   1413   20    60.62    55.00    4.00    3.00    11.00    agtcttccCGGGGAAAGCTG
   PRODUCT SIZE:  1142, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 1.00

2  LEFT PRIMER      273    20    60.89    50.00    5.00    3.00    11.00    ttttgcctcctcagggtggag
   RIGHT PRIMER   1384   20    60.88    50.00    4.00    2.00    10.00    a&AACCC&A&CCCGGAATTA
   PRODUCT SIZE:  1113, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 3.00

3  LEFT PRIMER      261    21    59.87    42.88    4.00    0.00    10.00    c&aggttttgcctttttgctc
   RIGHT PRIMER   1383   20    60.56    55.00    4.00    2.00    10.00    G&ACCC&A&CCCGGAATTA
   PRODUCT SIZE:  1123, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 2.00

4  LEFT PRIMER      273    20    60.38    50.00    5.00    3.00    12.00    ttttgcctcctcagggtggag
   RIGHT PRIMER   1412   19    59.89    57.89    4.00    2.00    11.00    gtttccCGGGGAAAGCTG
   PRODUCT SIZE:  1140, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 0.00
```

Primers successfully designed.

Alternative designs are also listed.

## Exoseq Primers

If you are interested in trying to amplify human exons then it might be possible to skip using Primer3 and use primers that have already been tested. The ExoSeq project at the Wellcome Trust Sanger Institute has already designed and tested primers for many human protein coding genes. Protocols are available at:

<http://www.sanger.ac.uk/humgen/exoseq/protocols.shtml>

You can search for primers for the gene of your choice at:

<http://www.sanger.ac.uk/cgi-bin/humgen/exoseq/search>

## RFLP Design

Some polymorphisms can be detected by RFLP. The best way to check this is using a website to screen your sequence for restriction enzyme sites.

<http://www.firstmarket.com/cutter/cut2.html>

<http://tools.neb.com/NEBcutter2/index.php>

Local sequence file:

GenBank number:

or paste in your DNA sequence: (plain or FASTA format)

```

gcccagtgccagaagagccaaggacaggtacgggtgtcatcacttagacctcaccctgtggagccacacc
ctaggggtggccaatctactcccaggagcagggagggcaggagccaggctggggcataaaaagtcagggc
agagccatctattgctctcatttggcttctgacacaactctctcactagcaacctcaaacagacaccat
ggcgcatctgactcttggagagaagtctgccgttactgccctgtggggcaaggtgaacctggatgaagt
tgggtggtgagggccctggcagggtgggtatcaaggtacaagacaggtttaaggagaccaatagaaactg
ggcatgtggagacagagaagactcttgggttctgataggcaactgactctctctgacctatgggtctatt
ttcccaccttaggctgctgggtggtctaccttggaccagaggtctcttgagtccttggggatctgt
ccactcctaatctattatgacaacctaaagtgaagactcatgcaaaaactactcaatad

```

Standard sequences: # Plasmid vectors  # Viral + phage

The sequence is:  Linear  Circular

Enzymes to use:  NEB enzymes  All commercially available specificities  All specificities  All + defined oligonucleotide sequences  Only defined oligonucleotide sequences

Minimum ORF length to display:  a.a.

Name of sequence:  (optional)

**Earlier projects:**

*Note: Your earlier projects will be deleted 2 days after they were last accessed. You need to have cookies enabled in your browser for this feature to work.*

Disable NEBcutter cookies

s  
ORFs

Cleavage code		Enzyme name code	
▶	blunt end cut	Available from NEB	
◀	5' extension	Has other supplier	
◁	3' extension	Not commercially available	
	cuts 1 strand	*: cleavage affected by CpG meth.	
		#: cleavage affected by other meth.	
		(enz.name): ambiguous site	

NEB cutter returns a list of restriction enzyme sites.

Availability	Display	Zoom	List
All commercial	2 cutters	Zoom in	0 cutters
All	3 cutters	More...	1 cutters
			All sites
			Save all sites
			Flanking enzymes

Local sequence file:

GenBank number:

or paste in your DNA sequence: *(plain or FASTA format)*

```

gccagtgccagaagagccaaggacaggtacggctgtcatcaettagacctcaccctgtggagccacacc
ctaggggtggccaatctactcccaggagcagggcaggagccagggtgggcataaaaagtcagggc
agagccatctattgcttacatttgccttgacacaactgtgttcactagcaacctcaaacagacacccat
ggtgcatctgactccctgTggagaagtctgccgttactgcctgtggggcaaggtgaactggatgaagt
tgggtggtgagggccctcgcagggttgggtatcaaggttacaagacaggtttaaaggagaccaatagaaactg
ggcatgtggagacagagaagactcctgggttcttgataggcactgactctctctgcctattgggtctatt
tccccacccttaggctgctgggtgctacccttgaccacagaggttctttgagtccttggggatctgt
ccactcctgatgctgttatgggcaaccccaaggtgaaggctcatggcaagaaagtgctcggtgc
    
```

The sequence is:  Linear  Circular

Enzymes to use:  NEB enzymes  All commercially available specificities  All specificities  All + defined oligonucleotide sequences  Only defined oligonucleotide sequences

Minimum ORF length to display:  a.a.

Name of sequence:  (optional)

Earlier projects:

*Note: Your earlier projects will be deleted 2 days after they were last accessed. You need to have cookies enabled in your browser for this feature to work.*

Disable NEBcutter cookies

Paste your variant sequence into the box. The polymorphic base has been highlighted in red.

s ORFs

Cleavage code	Enzyme name code
⌞   blunt end cut	Available from NEB
⌞   5' extension	Has other supplier
⌞   3' extension	Not commercially available
cuts 1 strand	*: cleavage affected by CpG meth.
	#: cleavage affected by other meth.
	(enz.name): ambiguous site



NEB cutter returns a list of restriction enzyme sites. The variant base introduces a novel site. This can be used as the basis for an RFLP assay.

Availability:

Display:

Zoom:

List:

## Determining rs# for older SNPs

In older papers interesting SNPs are often identified with reference to their position in a given transcript, eg A49T or 49A>T and not with an rs#. Given the constantly evolving nature of genome annotation it is often not obvious which transcript or start site genic co-ordinates refer to. If you are particularly interested in using them in your study a little detective work is required to find what their rs# is for high throughput assay design and for publication (journals are increasingly requiring rs#).

- 1) Identify the rough location of the SNP. You can do this by finding the primers used to amplify the SNP in the original paper(s). For example you are interested in an IL1B SNP is described as at "position +3953 in exon 5". Firstly identify papers describing this SNP using PubMed (try search terms like "IL1B polymorphism 3953").

The screenshot shows the PubMed search results for the query "IL1B polymorphism 3953". The search results are displayed in a table with two entries:

Item	Author	Title	Journal	PMID
1	Wang CY, Shen YC, Su CH, Lo FY, Lee SH, Tsai HY, Fan SS.	Investigation of the association between interleukin-1beta polymorphism and normal tension glaucoma.	Mol Vis. 2007 May 14;13:719-23.	PMID: 17563722 [PubMed - indexed for MEDLINE]
2	Tsarev VN, Nikolaeva EN.	[The allelic polymorphism of IL-1alpha and IL-beta genes in patients with chronic inflammatory periodontal diseases]	Vestn Ross Akad Med Nauk. 2007;(3):43-7. Russian.	PMID: 17500214 [PubMed - indexed for MEDLINE]

Check the methods for the relevant primer sequences:

### DNA preparation and genotype identification

Blood samples were collected from each subject (5 ml) and genomic DNA was isolated using the Qiagen QiaAmp Blood mini kit (Qiagen, Valencia, CA). IL-1 $\beta$  C(-511)T and C(+3953)T genotyping of genomic DNA were determined with polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP) technique. A 304 bp PCR fragment of the IL-1 $\beta$  (-511) in the promoter region was amplified using the following primers: F5'-TGG CAT TGA TCT GGT TCA TC-3' and R5'-GTT TAG GAA TCT TCC CAC TT-3'. PCR conditions were as follows: a denaturing step of 95 °C for 10 min, then 35 cycles of 95 °C for 45 s, 60 °C for 45 s, 72 °C for 1 min, and a final incubation at 72 °C for 5 min. The products were digested with *Bsu36I* (New England Biolabs, Inc., Beverly, MA) at 37 °C for 3 h and were run on ethidium bromide-stained 2% agarose gel. This gave products that either remained intact (C allele) or were cut into two fragments of 190 and 114 bp (T allele).

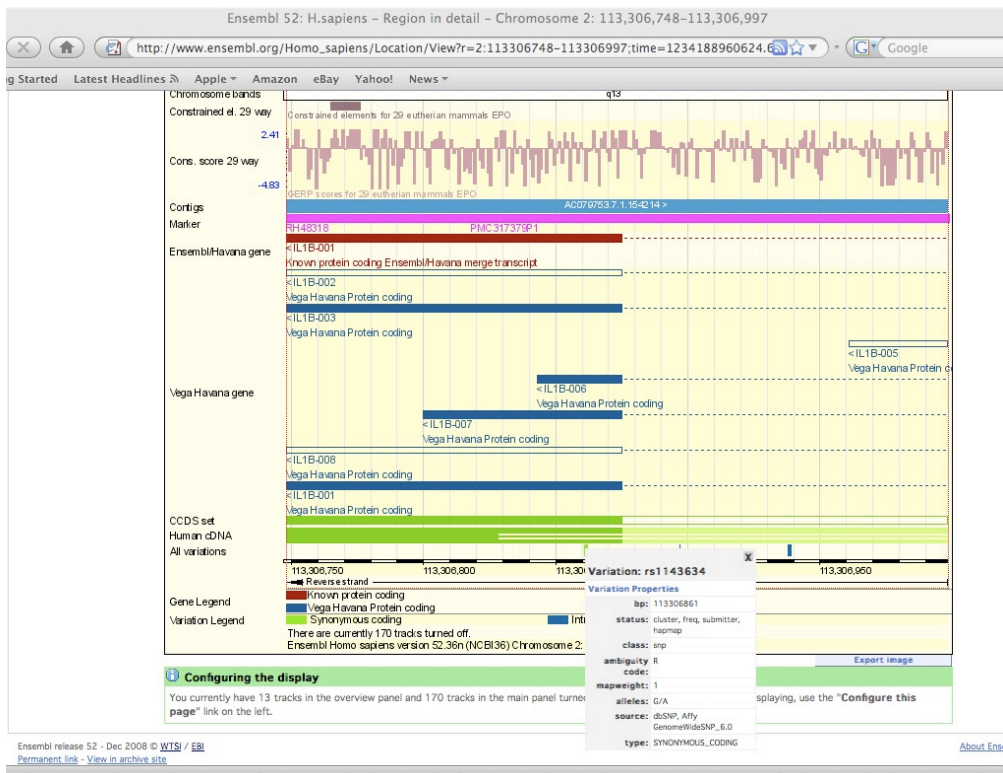
The polymorphic region containing the *TaqI* (New England Biolabs, Inc.) restriction site at position +3953 within exon 5 of the IL-1 $\beta$  gene was amplified using the following primers: F5'-GTT GTC ATC AGA CTT TGA CC-3' and 5'-TTC AGT TCA TAT GGA CCA GA-3'. The PCR conditions were the same as described in the previous. The products were digested with *TaqI* at 65 °C for 3 h. *TaqI* digestions of the 249 bp fragments were cut into two fragment of 135 and 114 bp (allele C) or remained intact (allele T).

- 2) Perform BLAT searches (<http://genome.ucsc.edu/cgi-bin/hgBlat>) to find the genomic location of the primers.

Human BLAT Results											
BLAT Search Results											
ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
<a href="#">browser</a> <a href="#">details</a>	YourSeq	20	1	20	20	100.0%	2	-	113306978	113306997	20

BLAT Search Results											
ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
<a href="#">browser</a> <a href="#">details</a>	YourSeq	20	1	20	20	100.0%	2	+	113306748	113306767	20

- 3) Enter start and finish co-ordinates for fragment (start of one primer, end of the other, whatever spans the biggest area) in Ensembl.
- 4) Simply click on anywhere is chromosome 2 and change the co-ordinates to 133306748 – 11
- 5) Switch on SNPs in the Features list. Click on SNPs to see which one has the right base changes. If there are still multiple options the SNP properties link will give you flanking sequence etc so you can see if for example a restriction enzyme used in the original genotyping assay will work or not.





6) Copy the SNP and flanking sequence from the Ensembl SNPView page, by clicking on variation properties.

8,306,997 Variation: rs1143634

**Variation: rs1143634**

**Variation type** SNP (source [dbSNP](#))

**Synonyms** Affy GenomeWideSNP\_6.0 SNP\_A-8637035

**Alleles** G/A (Ambiguity code: R)  
Ancestral allele: G

**Location** [2:113306861](#) (forward strand)

---

**Variation summary** [help](#) [Ge](#)

---

**Validation status** Proven by **cluster, frequency, submitter** (Feature tested and validated by a non-computational method).  
**HapMap SNP**

**Linkage disequilibrium data** **Links to Linkage disequilibrium data per population:**

<a href="#">CSHL-HAPMAP:HapMap-CEU</a> (Tag SNP)	<a href="#">CSHL-HAPMAP:HapMap-HCB</a> (Tag SNP)	<a href="#">CSHL-HAPMAP:HapMap-JPT</a> (Tag SNP)
<a href="#">CSHL-HAPMAP:HapMap-YRI</a> (Tag SNP)	<a href="#">PERLEGEN:AFD_AFR_PANEL</a>	<a href="#">PERLEGEN:AFD_CHN_PANEL</a> (Tag SNP)
<a href="#">PERLEGEN:AFD_EUR_PANEL</a> (Tag SNP)		

**Flanking Sequence**

```

TGGGTGGACATGGTCTGGGAGTGGTTGAGCCGCTAAATTTCTCAGGGTCACACTCCTGT
TAACAAATGCACGGCCAGTGCATCAATGTGCCATTTCTAGGACCAAGTTGTATAT
TCCTTTTAAATATTTTTTTCACCTGGTGGATCATTTGGCTTAAATTAACCTTCTACTT
TGTTTAAACATGGAGATTAGCAAGCTGCCAGAGCCAGGAGGAAACAGGATGTT
TCCATTTACCTTGTTCCTCATATCCTGTCCCTGGAGGTGGAGAGCTTTCAGTTCAATAG
GACCAGACATCACCAAGCTTTTTGCTGTGAGTCCCGAGCGTGCAGTTCAGTGAATGTA
CAGGTGCATCGTGCACATAAGCCCTGTTATCCCATGTGCRANGAAGATAGTTCTGAAA
TGTGGAGCACATGTTTATAGGTATAAAATCAGAAGGGCAGGCTCCTGAGGCGAGGGGG
CAAAATTTGATTTCTGGAGGACACCTGAGCATATACGGTCAAAGTCTGATGACAAACCC
AGTAGGGATGAAGCTGGGAGTGGGGTGGCTAAGAACACTGGACCTGACACTATTAGCAT
GGGTCCAGCTTCAGGCTATTAATCTGCTCACTGTGGCCGACACAGAGCTACTTAGGTA
AAATGGTATGGTCAACACTAGCCACAGGGAGGTTACGAACCTCTGGTGAATGTA
AGTGAAGGCCCTGAGAAAGAGTGGGGAGTTGCAAATGTCAGTAGCCATCAAGATCTT
CTTTAAGAATAGTTCCACTA
    
```

(Variant highlighted)

7) Paste the SNP sequence into NEBcutter (<http://tools.neb.com/NEBcutter2/index.php>). Change the ambiguity code to the wild type allele and press “submit”. Repeat with the variant allele and look to see if the expected restriction enzyme cuts/doesn’t cut each allele.



# NEBcutter V2.0



This tool will take a DNA sequence and find the large, non-overlapping open reading frames using the E. coli genetic code and the sites for all Type III restriction enzymes that cut the sequence just once. By default, only enzymes available from NEB are used, but other enzymes may be chosen. Just enter your sequence and "submit". Further options will appear with the output. **The maximum size of the input file is 1 Mb and the maximum sequence length is 300 KBases.**

[What's new in V2.0](#)

Local sequence file:

GenBank number:

or paste in your DNA sequence: *(plain or FASTA format)*

```

TGTTTAAAACATGGAGAATTAGCAAGCTGCCAGGAGGCCAGGCAGGGAACCAGGATGTT
TCCATTTACCTTGGTCTCCATATCCTGTCCCTGGAGGTGGAGAGCTTTCAGTTCATATG
GACCAGACATCACCAGCTTTTTTGTCTGTGAGTCCCGGAGCGTCCAGTTCAGTGATCGTA
CAGGGTGCATCGTGCACATAAGCCTCGTTATCCCATGTCTCGAAGAGATAGGTTCTGAAA
TGTGGAGCACATGTTGTTTAGGTATAAAAATCAGAAGGGGAGGCCTCGTGAGGCGAGGCGG
CAAAAATTTGATTTCTTGGAGGACACCTGAGCATATACGGTCAAAGTCTGATGACAAACACC
AGTAGGGATGAAGCTGGGAGTGGGGTGGCTAAGAACAACCTGGACCTGACACTATTAGACAT
GGGTTCCAGCTTTCAGGTCTATTACTGCTCACTGTGGCCGAGCAACAGAGCTACTTAGGTA
AAAATGGTGATGGTCATAACACTAGCCACAGGGAGGTTACGAACCTCTGGTGACAAATGTA
                    
```

Standard sequences:

# Plasmid vectors

# Viral + phage

The sequence is:  Linear  Circular

Enzymes to use:

- NEB enzymes
- All commercially available specificities
- All specificities
- All + defined oligonucleotide sequences
- Only defined oligonucleotide sequences

[define oligos](#)

Minimum ORF length to display:  a.a.

Name of sequence:  *(optional)*

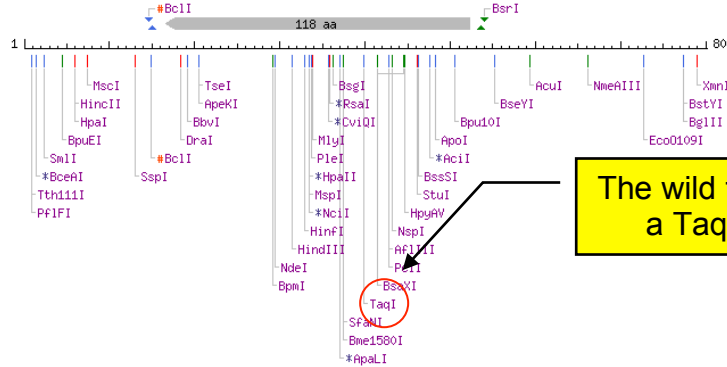


Linear Sequence: *unnamed sequence*

[Help](#) [Comments](#)

Display: - NEB single cutter restriction enzymes  
 - Main non-overlapping, min. 100 aa ORFs  
 GC=46%, AT=54%

Cleavage code	Enzyme name code
⌞   blunt end cut	Available from NEB
⌞   5' extension	Has other supplier
⌞   3' extension	Not commercially available
⌞   cuts 1 strand	*: cleavage affected by CpG meth.
	#: cleavage affected by other meth.
	(enz.name): ambiguous site



The wild type, G, allele creates a Taq I site as expected

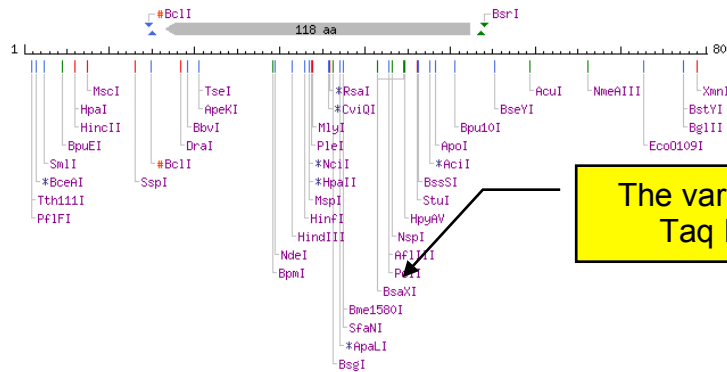


Linear Sequence: *unnamed sequence*

[Help](#) [Comments](#)

Display: - NEB single cutter restriction enzymes  
 - Main non-overlapping, min. 100 aa ORFs  
 GC=46%, AT=54%

Cleavage code	Enzyme name code
⌞   blunt end cut	Available from NEB
⌞   5' extension	Has other supplier
⌞   3' extension	Not commercially available
⌞   cuts 1 strand	*: cleavage affected by CpG meth.
	#: cleavage affected by other meth.
	(enz.name): ambiguous site



The variant, A, allele has no Taq I site as expected

So the IL1B SNP at "position +3953 in exon 5" is now known as rs1143634. NEBcutler can also be used to design your own RFLP tests if you are only interested in genotyping a handful of SNPs. Simply paste both alleles in and see if there is any differential digestion.

