

---

# GENOME ANALYSIS TOOLS

---

A handbook



ASHWINI VASANTH  
JOHNS HOPKINS UNIVERSITY -AAP  
Summer 2021  
AS.410.635.81.SU21 Bioinformatics: Tools for Genome Analysis  
Dr. Sijung Yun, Dr. Sajung Yun

# Contents

Introduction.....	2
Gene Prediction Tools .....	2
ORF Finder .....	2
FEGENESB .....	3
BPROM .....	4
GLIMMER.....	5
GENEMARK.....	7
EasyGene 1.2b .....	9
Neural Network Promoter Prediction .....	10
GENSCAN.....	11
HMMGene .....	12
FGENESH .....	13
AUGUSTUS .....	14
Splign.....	16
BLAT.....	18
BioMart.....	20
R/Bioconductor and BiomaRt.....	21
Galaxy.....	24
IGV (Integrated Genome Viewer) .....	25
BEDtools.....	29
SAMtools .....	30
Next-Generation Sequencing (NGS) analysis.....	32
ChIP-seq data analysis .....	33
RNA-seq data analysis.....	35
Genomic Databases and Genome browsers .....	39
References.....	61

## Introduction

The course “Bioinformatics: Tools for Genome Analysis” introduced me to several tools used by Bioinformaticians worldwide. These included prokaryotic and eukaryotic gene prediction tools – some command line and others with web interface, genome browsers, Galaxy, IGV, R programming language, NGS data analysis etc. This document lists the key tools studied in this course along with a summary of its usage and results and intends to be a ready reckoner for anyone new to the field of Bioinformatics. Also included in the document is a list of specialized databases that host species specific information. The document is linked to the websites associated with the tools and genome browsers.

## Gene Prediction Tools

### ORF Finder

The [ORF Finder](#) looks for Open Reading Frames (ORFs) in the DNA sequence that is entered in the interface. It lists the ORFs, and the protein sequence translated from each ORF. It is a good way to look for protein coding sequences in a newly sequenced DNA.

Open Reading Frame Viewer Help

Sequence Tools ▾ Tracks ▾ ?

ORFs found: 34 Genetic code: 1 Start codon: 'ATG' only

(?) ORFfinder\_7.23.23552411

ORF15 (437 aa) Display ORF as... Mark

Mark subset... Marked: 0 Download marked set as Protein FASTA ▾

Label	Strand	Frame	Start	Stop	Length (nt   aa)
ORF15	+	3	81	1394	1314   437
ORF30	-	2	2359	1454	906   301
ORF26	-	1	1298	498	801   266
ORF20	+	3	3444	>4013	570   189
ORF2	+	1	2098	2640	543   180
ORF17	+	3	1641	2126	486   161
ORF27	-	1	494	78	417   138
ORF4	+	1	2980	3381	402   133
ORF21	-	1	3791	3420	372   123
ORF32	-	2	1057	707	351   116

ORF15 SmartBLAST BLAST

Marked set (0) SmartBLAST best hit titles... BLAST

BLAST Database: UniProtKB/Swiss-Prot (swissprot) ▾

A DNA sequence in FASTA format or an accession ID can be entered in the interface.

After selecting the parameters and submitting the form, a webpage like the one above is displayed that shows the total number of ORFs predicted with the start and stop position of each, and the corresponding protein sequence encoded.

## FEGENESB

**FGENESB** is a bacterial operon and gene finding program. The prediction algorithm is based on Markov chain models. It is a very accurate ab initio prediction program available for bacterial genes.

**USAGE EXAMPLE:** Identify coding sequences (CDS) in the partial sequence of a bacterial strain such as *S. helicoides* strain TABS-2 using ‘bacterial generic’ as the training set. In the example below, nine CDSs are predicted by FGENESB. There are six predicted mRNAs of which two are operons. The mRNAs predicted are enclosed in black rectangles in the screenshot below.

```
Prediction of potential genes in microbial genomes
Prediction of potential genes in microbial genomes
Time: Tue Jan 1 00:00:00 2005
Seq name: Spiroplasma helicoides strain TABS-2, partial sequence
Length of sequence - 5500 bp
Number of predicted genes - 9
Number of transcription units - 6, operons - 2
N Tu/Op Conserved S Start End Score
pairs(N/Pv)
1 1 Op 1 . + CDS 635 - 991 117
2 1 Op 2 . + CDS 998 - 1141 144
3 2 Tu 1 . - CDS 1126 - 1365 73
4 3 Tu 1 . + CDS 1334 - 1978 381
5 4 Tu 1 . + CDS 2242 - 2463 231
6 5 Op 1 . + CDS 2585 - 4003 998
7 5 Op 2 . + CDS 4010 - 4678 423
8 5 Op 3 . + CDS 4703 - 4768 72
9 6 Tu 1 . + CDS 4880 - 5143 169

Predicted protein(s):
>GENE 1 635 - 991 117 118 aa, chain +
MTYSFSFIIERGVQEYDTSKFLISSIASCAFIAHLLFEYFSQIILINQSIIKINTKLKRVT
ARKNFTTENYKVSLDTGEFININSTKINQLADNYFTSIFDISCIIATIISYGFLLTIS
>GENE 2 998 - 1141 144 47 aa, chain +
MLAVMILSLLVLVIPMLMSKIGKKRINVANEENDKFLQTTKDVTNSY
>GENE 3 1126 - 1365 73 79 aa, chain -
MSVNIKPIFIIYPAQYIQQKNIKIKITCPRTTIISSKNLNVDDITFFIFWFLTSNFFDPST
IWLSLFWWFMLQYTQTEL
>GENE 4 1334 - 1978 381 214 aa, chain +
MNIRGLIFTNLISSSVYCFSSSAKALMNIIINHRKVYLSNYQDNKINNNNTVIGEDLTKI
EFKVNDFKYKNSSNLIEKPNLKKGDVKVLYKGSGIGKTTLLKTLFNPSPRSNSQVVV
NEQVEAYDIRSLCSYISQDIVFSKGHLIMLKIANESAEEKQVLSLFELLGQNQLLERLK
PEGLNTKIDDNSSNNFSGGEKQRFSSIIQRGLLENKS
>GENE 5 2242 - 2463 231 73 aa, chain +
MFVDLLASTSEKLTGNRIVFAFEIIALVVSILMITVGMIQNKTSTQGLSALNGGNDELFS
NSKERGMDRTMSI
>GENE 6 2585 - 4003 998 472 aa, chain +
MEENILSLIKQKQKLHNELLKTFKDEELLMSCSKELQDQYKLSKSENVVYFIGEKYKV
GSIKINKEGFGVFKDLNDVQEODYFVFPDSLNKSITTDEVFTVYKESSEERYRANVEDISL
RVKSFLIGEIQPSRDGRFLDFIPSEPGFKNYRIVMINSKDFKLKKDLLVVKVILNVKEKK
LETKIQRIIGDSKNAVRDIISIAYEFNINDPFRNQTLLENADQVAIPINYDEQVKRRLRN
SLVDKNLVITDGSDSKDDDAIYVEKTKDGYKLFAIAIDVSYYVLPFSPLDNTALYRGNS
TYLANKVIPMLPEKLSMGCVSLNPNEDEKDCMVSEMDFDNNGVNMKNKQVYESIMNSKARLT
YKEVNDLFEKVNVSNRDEKEIVDMILLVSKELIHELDKERVSRGSIIDFDVPEEPKIVLDKESNV
DVIVPRDRGVUSERLIENFMVSANESVAQIIFEKNLPYVYRNHGAPKEENLIE
>GENE 7 4010 - 4678 423 222 aa, chain +
LIRALGINVKLTDEKVNPKTIRMALDQISKQIEDQTERDVINVTLKFMEEKAAYSELENI
GHFGLASECYTHFTSPIRYSQDMVWHRYLQYLDKDLRDFKLDDNEKFINKACKINET
EKNVSNAEREVNKVCMAEFMTKHIEKEYEGVVAAVLKPGFLFVQLSNCVEGLIHISELPEF
TEDKTNILVNQNKVFRLGQKVKKVKVQNAADVVKRKRIDFVLV
>GENE 8 4703 - 4768 72 21 aa, chain +
MGEHILLKNNKKAYFNEYILD
>GENE 9 4880 - 5143 169 87 aa, chain +
MNIKKYYEYANYVKQDPTRTRKLLLNKDEIKKILKRVQLENLTIIPLKLYLKGNYAKLEIG
IIGKGGKLIDKRETIKKRDIERRLNKIK
```

## BPROM

**BPROM** is a bacterial sigma 70 promoter prediction program. The interface takes a FASTA sequence as input and generates an output like the one below. The number of promoters predicted along with the locations of the -10 and -35 signals are shown. The algorithm predicts the transcription start site (TSS) of bacterial genes that are regulated by the sigma 70 promoters.

```
>Lactococcus lactis subsp. lactis ptsHI operon, complete sequence
Length of sequence-      2592
Threshold for promoters -  0.20
Number of predicted promoters -    7
Promoter Pos:   225 LDF-  8.79
-10 box at pos.  210 TGGTACAAT Score  78
-35 box at pos.  190 TTGCAA   Score  55
Promoter Pos:  2543 LDF-  5.41
-10 box at pos.  2528 AATTAATAT Score  53
-35 box at pos.  2505 TTGATA   Score  58
Promoter Pos:  1005 LDF-  3.54
-10 box at pos.  990 TGTTAAATT Score  66
-35 box at pos.  973 TTGGCT   Score  33
Promoter Pos:  1860 LDF-  3.46
-10 box at pos.  1845 AGGTATCAT Score  71
-35 box at pos.  1826 TTGCAG   Score  49
Promoter Pos:  1392 LDF-  2.99
-10 box at pos.  1377 TGCTAACATAT Score  67
-35 box at pos.  1352 CTGACG   Score  25
Promoter Pos:  561 LDF-  2.12
-10 box at pos.  546 CAGAATAAT Score  40
-35 box at pos.  527 ATGACT   Score  31
Promoter Pos:  2216 LDF-  0.70
-10 box at pos.  2201 TGGAAGAAT Score  41
-35 box at pos.  2176 ATGAAA   Score  30

Oligonucleotides from known TF binding sites:

For promoter at  225:
  purR: TTTCGTTT at position  200 Score -  6
  purR: ATTTCAAG at position  217 Score -  9
  fnr:  TCAAGAGT at position  220 Score - 13
  nagC: ATATTTTA at position  233 Score -  7
  nagC: ATTTTAGA at position  235 Score -  6
For promoter at  2543:
  rpoS17: AGAGGGAG at position 2483 Score - 10
  fis:   CTCATTAA at position 2499 Score -  9
  argR:  AATTAATA at position 2528 Score - 11
For promoter at  1005:
  crp:  TTAAATTG at position  992 Score - 10
No such sites for promoter at  1860
For promoter at  1392:
  rpoD19: CACCTAAA at position 1391 Score -  6
For promoter at  561:
  argR:  ATAATCAT at position  550 Score -  9
No such sites for promoter at  2216
```

## Sample output:

### BPROM output:

First line - name of your sequence;  
Second and Third lines - LDF threshold and the length of presented sequence  
4th line - The number of predicted promoters  
Next lines - positions of predicted promoters, and their scores with 'weights' of two conserved promoter boxes. Promoter position assign to the first nucleotide of the transcript (Transcription Start Site position).  
After that we present elements of Transcriptional factor binding sites for each predicted promoter (if they found).

### For example:

```
prom Sat Jan 18 21:11:25 EST 2003
Region of E.coli genome between protein_id="AAC76687.1" and protein_id="AAC7668
Length of sequence-        420
Threshold for promoters -  0.20
Number of predicted promoters -      1
Promoter Pos:    145 LDF-  6.02
-10 box at pos.   130 ctttatgat Score  66
-35 box at pos.   109 tttaat   Score  36

oligonucleotides from known TF binding sites:

For promoter at  145:
  fis: TCTTTAAT at position  107 Score -  6
  rpoD17: TTATGATA at position 132 Score -  7
  lexA: ATAAATAAA at position 137 Score - 14
  rpoD17: ATAATAAT at position 141 Score -  8
```

## GLIMMER

**GLIMMER** (Gene Locator and Interpolated Markov ModelER) is a UNIX-based program for finding the genes in microbial DNA (bacteria, archaea, and virus). It makes use of interpolated Markov models (IMMs) to recognize coding regions and to differentiate them from non-coding regions. The software version and the download link can be found [here](#).

**USAGE EXAMPLE:** Given a complete bacterial genome in FASTA format, GLIMMER 3 can be used to find long ORFs in the genome sequence. In bacteria, most long ORFs correspond to actual genes. Hence, the long ORFs can be used to build a training set to analyze the rest of the genome. If working with an unknown DNA sequence, a related genome can be used to build training set. The .predict file produced as the output lists the ORFs with the start and end coordinates and the strand.

Use the command line version of Glimmer to analyze CDSs in a partial sequence from Spiroplasma helicoides strain TABS-2, whose genome was submitted to GenBank on August 23, 2016 (file: sheliprt.fasta). The training set will be the full genome of S. helicoides strain TABS-2 (file: sheli.fasta).

## **UNIX commands:**

```
$ long-orfs -n -t 1.15 sheli.fasta sheli.longorfs  
  
$ extract -t sheli.fasta sheli.longorfs > sheli.train  
  
$ build-icm sheli.icm < sheli.train  
  
$ glimmer3 -o50 -g110 -t30 sheliprt.fasta sheli.icm sheliprt  
  
$ extract -t sheliprt.fasta sheliprt.predict > sheliprt.glimmer
```

## **Output format:**

## **Command usage information:**

**long-orfs [options] <sequence-file> <output-file>** - reads the DNA sequence in the <sequence-file> and outputs the coordinates of the long, non-overlapping ORFs in it.

- -n: do not include header information in the output.
  - -t: only genes with the entropy score less the number specified will be considered.

**extract [options] <sequence-file> <coords>** - read the FASTA-format <sequence-file> and extract from it the subsequences specified by the coordinates. The default value for <coords> is the name of the file containing lines of the form <tag> <start> <stop> [<frame>]..

- -t: omit the last three characters of each output string

**build-icm [options] output\_file < input-file** – read sequences from the standard input and output to the output\_file the interpolated context model (icm) built from them.

**glimmer3 [options] <sequence-file> <icm-file> <tag>** - read DNA sequences in the <sequence-file> and predict genes in them using the <icm-file>. Output goes to the file <tag>.detail and predictions go to <tag>.predict.

- -o <n>: sets the maximum overlap length to <n>. Overlaps that are this short or shorter are ignored.
- -g <n>: set the minimum gene length to <n>.

## GENEMARK

[\*\*GeneMark.hmm prokaryotic\*\*](#) version is a prokaryotic gene prediction tool. The sequence in FASTA format is provided as input to the program. The sets of pre-computed species are available for the user to select as a parameter to the program. The program output provides the list of predicted genes along with the strand, start and end coordinates, and the gene length.

## GeneMark.hmm prokaryotic

Prokaryotic GeneMark.hmm version 1

**Alexander Lukashin and Mark Borodovsky**  
GeneMark.hmm: new solutions for gene finding.  
*Nucleic Acids Research* (1998) 26, pp 1107-1115

Prokaryotic GeneMark.hmm version 2

**John Besemer, Alexandre Lomsadze and Mark Borodovsky**  
GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions.  
*Nucleic Acids Research* (2001) 29, pp 2607-2618

This webpage provides access to gene prediction program GeneMark.hmm prokaryotic (version 3.25) and to the sets of pre-computed species specific algorithm parameters (model parameters). The list of currently supported species is available [here](#). These parameter sets were derived by application of the GeneMarkS that carried out unsupervised training on each genome.

[Browse GeneMark.hmm prokaryotic manual](#)

### Input sequence and Select species

#### Enter sequence (FASTA or multi FASTA format)

```
>Stenotrophomonas maltophilia strain SM45 whole genome shotgun  
sequence  
ACCAAGTCGAAGCGTTCGGCAACTCTTGCCTTGCCCGAGCTCGCCATGCTCGAGCAGCTGCT  
TCTGCTCCTCGGTGTACTTGGCAAGCTCTGCCGCTCGCATACAGCGAGTTCCAGATATTGGCGAATTG  
CTGAAGGCTGTTCGGGCACTCATCTGGCTAACGATGTCCTCGTCTTGAGAGATCTCACCCATGCTC
```

or, upload file:

No file chosen

#### Select species

Stenotrophomonas\_maltophilia\_K279a

#### Action

### Options

Output format for gene prediction	Output options	Optional: results by E-mail
<input checked="" type="radio"/> LST <input type="radio"/> GFF	<input type="checkbox"/> Protein sequence <input type="checkbox"/> Gene nucleotide sequence  <input type="checkbox"/> Coding potential graph (not for multi FASTA) <input type="checkbox"/> PDF <input type="checkbox"/> PostScript	E-mail  Subject GeneMark.hmm prokaryotic <input type="checkbox"/> Compress files

### Advanced options

Switch off gene start related motif(s)

[Contact Us](#) | [Home](#)

## Output file:

```
GeneMark.hmm PROKARYOTIC (Version 3.26)
Date: Sun Jul 25 14:55:43 2021
Sequence file name: seq.fna
Model file name: /home/genemark/parameters/prokaryotic/Stenotrophomonas_maltophilia_K279a/GeneMark_hmm_combined.mod
RBS: true
Model information: Stenotrophomonas_maltophilia_K279a
```

FASTA definition line: Stenotrophomonas maltophilia strain SM45 whole genome shotgun sequence

Predicted genes					
Gene	Strand	LeftEnd	RightEnd	Gene Length	Class
#					
1	-	<3	3407	3405	2
2	-	3400	4050	651	1
3	-	4121	4921	801	1
4	-	4935	5417	483	1
5	-	5410	5745	336	1
6	-	5742	6233	492	2
7	-	6230	6364	135	1

## EasyGene 1.2b

**EasyGene 1.2b** server is a gene prediction tool for prokaryotic DNA. Each prediction comes with a significance score that indicates its likelihood of being a real gene.

## Web-interface:

The screenshot shows the EasyGene 1.2b web interface. At the top, there is a header with the DTU Bioinformatics logo and a message about services being migrated. Below the header, the main title "EasyGene 1.2b Server" is displayed, followed by a subtitle "Gene finding in prokaryotes". The page contains several input fields for sequence submission, organism selection, and R-value cutoff. It also includes sections for instructions, output format, article abstracts, and a submission summary. At the bottom, there are sections for citation, portable version, and getting help.

## Understanding the output:

DTU Bioinformatics  
Department of Bio and Health Informatics

Services are gradually being migrated to <https://services.healthtech.dtu.dk/>.  
Please try out the new site.

Home

**Output format**

**DESCRIPTION**

The output conforms to the [GFF](#) format. For each input sequence the server prints a list of predicted genes, one per line. The columns are:

- sequence**: Input sequence name;
- model**: organism model code (also in plain text in the table head);
- feature**: feature type: CDS or CDSsub (alternative translation start);
- start** and **end**: position in the sequence;
- score**: R-value, indicating how likely the fragment is to be just a non-coding open reading frame rather than a real gene;
- +/-**: predicted start codon;
- odds**: log odds score.

Only the predictions with R-values lower than the selected R-value cutoff (the default is 2) are reported.

The example below shows the EasyGene 1.2 output for the sequence taken from the GenBank entry [AB010576](#), containing *Bacillus subtilis* **ComX**, **ComQ** and **DegQ** genes. All the three genes are predicted as annotated in the database (shown in green), with high confidence, although an alternative translation start is preferred for **comQ** (shown in orange). Two additional genes not annotated in the GenBank entry are also predicted.

**EXAMPLE OUTPUT**

```
#gff-version 2
##source-version easygene-1.2b
##date 2007-08-15
##type DNA
# model: B503 Bacillus subtilis
# sequence: B503 Bacillus subtilis
# feature: CDS
# start: 67 end: 324 score: 0.0271875 + 0 WATG 20.1861
# feature: CDSsub
# start: 55 end: 324 score: 0.031955 + 0 WATG 20.1791
# feature: CDS
# start: 1129 end: 1269 score: 0.0190622 + 0 WATG 15.7102
# feature: CDS
# start: 2327 end: 2491 score: 0.0167943 + 0 WATG 17.2951
# feature: CDS
# start: 300 end: 668 score: 1.43511 - 0 WATG 10.6215
```

**GETTING HELP**

Scientific problems: [Thomas Schou Larsen](#) Technical problems: [Support](#)

This file was last modified Tuesday 3rd January 2017 06:58:02 GMT

## Neural Network Promoter Prediction

It is found in the Berkeley Drosophila Genome Project page. This [program](#) predicts possible transcription promoters for both prokaryotic and eukaryotic DNA sequence. The output shows the start and end positions of the predicted promoters and the score associated with each prediction.

**Berkeley Drosophila Genome Project**  
Searches

**Neural Network Promoter Prediction**

**Home**  
**About BDGP**  
Contact Information, News,  
Citing BDGP

**Projects**  
*D. melanogaster* Release 6  
Genome

**EST Sequencing**

**Drosophila Gene Collection**

**Universal Proteomics  
Resource**

**Gene Disruption Project**

**Expression Patterns**

**modENCODE**

**SNP Map**

**BDGP Resources**

**Download**  
Sequence Data Sets

**Materials**  
Clones, Vectors, Stocks,  
Libraries

**Publications**

**Methods**

**Searches**  
FlyBase All Searches

**Read Abstract Help**

**PLEASE NOTE:** This server runs the 1999 NNPP version 2.2 (March 1999) of the promoter predictor.

Enter a DNA sequence to find possible transcription promoters

Type of organism: Eukaryote @eukaryote  
Include reverse strand? Yes @no  
Minimum promoter score (between 0 and 1): 0.8

Cut and paste your sequence(s) here: Use single-letter nucleotides: (A, C, G, T). You can include multiple sequences if each has a FASTA title line starting with >

Please be patient—promoter prediction takes about 10 seconds per kilobase.

**Training set:** Our training and test sets of human and *Drosophila melanogaster* promoter sequences are available to the community for testing transcription start site predictors. These sites also contain our representative, standardized data sets of human and *Drosophila melanogaster* genes.

**Publications:** Reese MG, 2001. "Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome"; *Comput Chem* **26**(1):51-6.

## Output:

Promoter predictions for 1 prokaryotic sequence with score cutoff 0.80 (transcription start shown in larger font):

Promoter predictions for Spiroplasma :

Start	End	Score	Promoter Sequence
24	69	0.96	ATTTGTTTTTTTATTCTAAATAGTTATTGTATGTC <b>T</b> ATCAACTTC
51	96	0.98	ATTTATGATGCTATCAACCTCTGGTTAAAATTAAAC <b>C</b> ATTGTAATT
92	137	0.96	ATTTGAATTGAATAAAGAAGATAAACCTTAAATGTTT <b>T</b> TCTTAAGAG
126	171	0.95	TGTTTTCTTAAGGTCTATACAGATTAAATTACTTC <b>T</b> GAGTCATG
148	193	0.99	ACAGATTAAATTACTCTGTAGTCATGATTAAATAAA <b>A</b> ATCTTTTA
175	220	0.96	GATTATTAATAAAATCTTACTCTATCTAAACTT <b>G</b> TAATACTTG
188	233	0.92	AATCTTTTACTCTATCTAAACTTTGAATCTTGTAA <b>T</b> AAATGTAA
220	265	0.97	ACTTGTAAATTAAATGTAAGGATAGGCAGATTAAATTGATT <b>T</b> TCTATGTGA
255	300	0.94	GTATTTCTATGATTGTCAGCATAAGCTACAACATTTCAGATCAGC <b>T</b> AGATCAGC
266	311	0.96	GTGATTGTCAGCATAAGCTACAACATTTCAGATCAGCA <b>A</b> CTGTCATT
309	354	0.91	TGTCATTTTACTTTAAGTATTTTAAAAAATTTTAG <b>C</b> TGGTTTGT
332	377	0.83	TTTTAAAAAATTTTAGCTGGTTTGTCTATATTTC <b>T</b> AAATGAT
380	425	1.00	ATTTGTTTGAAATTAGCCATAAAATTAAACATAATCTAGTCATT <b>C</b> ATAATCA
391	436	0.96	AAATTAGCCATAAAATTAAACATAATCTAGTCATT <b>C</b> ATAATCA
404	449	0.99	AAATTAAAAACATAATCTAGTCATT <b>C</b> ATAATCACTTC <b>C</b> TTAAACGC
426	471	0.87	CATTGCAATAATCACTCCCTAAACGCATCTAATTGATA <b>T</b> ATAATTATT
456	501	0.95	CTAATTGATATAATTATTAAATTATATAACTTATT <b>T</b> ATATATTAA
469	514	1.00	AATTATTAAATTATATAACTTATTATATATTAAAA <b>T</b> ATAAAAATA
488	533	0.96	AACTTATTATATTTAAATATAAAATAATCAC <b>A</b> TATTATTT

## GENSCAN

[GENSCAN](#) is one of the first successful eukaryotic gene prediction program. It predicts the locations and the exon-intron structures of genes in sequences up to 1 million base pairs in length. The output of the program includes the genes predicted with the start and stop coordinates, poly(A) tail location, and the strand information. It also shows the sequence of the predicted peptide.

The GENSCAN Web Server at MIT

Identification of complete gene structures in genomic DNA

[?](#)  
[For information about Genscan, click here](#)

Server update, November, 2009: We've been recently upgrading the GENSCAN webserver hardware, which resulted in some problems in the output of GENSCAN. We apologize for the inconvenience. These output errors were resolved.

This server provides access to the program Genscan for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.

This server can accept sequences up to 1 million base pairs (1 Mbp) in length. If you have trouble with the web server or if you have a large number of sequences to process, request a local copy of the program (see instructions at the bottom of this page).

Organism:  Suboptimal exon cutoff (optional):

Sequence name (optional):

Print options:

Upload your DNA sequence file (upper or lower case, spaces/numbers ignored):  No file chosen

Or paste your DNA sequence here (upper or lower case, spaces/numbers ignored):

[Back to the top](#)

GENSCAN was developed by Chris Burge in the research group of Samuel Karlin, Department of Mathematics, Stanford University. The program and the model that underlies it are described in:

## Output:

HMMGene

**HMMGene** is a gene prediction tool for vertebrates and C.elegans. The web-interface provides the user the option to choose the organism, enter the input sequence, and select predict signals that predicts the splice sites and probabilities associated with start/stop codons predicted.

**DTU Bioinformatics**  
Department of Bio and Health Informatics

Services are gradually being migrated to <https://services.healthtech.dtu.dk/>.  
Please try out the new site.

**HMMGene (v. 1.1)**

Prediction of vertebrate and C. elegans genes

---

The server was set up on Feb. 27, 1998. Please mail [krogh@csd.dtu.dk](mailto:krogh@csd.dtu.dk), in case of problems.

Jan 18, 2002  
Error in the processing of annotation in the sequence file corrected

**Instructions**

**Submission of a local file (HTML, 3.0 or higher)**

Organism:

- Human (and other vertebrates)
- C. elegans

**Options:**

- Predict signals
- best prediction

**File with annotation (optional)**

Choose File | No file chosen

Choose File | No file chosen

|

**Submission by pasting sequences:**

Organism:

- Human (and other vertebrates)
- C. elegans

**Options:**

- Predict signals
- best prediction

**Sequences(s) in FASTA format**



**Annotation (optional)**

|

**Restrictions:**  
At most 1000 sequences, at least 50 nucleotides long, at most 10,000,000 nucleotides per submission.

Comments, inquiries and suggestions: send mail to Anders Krogh, [krogh@csd.dtu.dk](mailto:krogh@csd.dtu.dk)

---

**GETTING HELP**

Scientific problems [Anders Krogh](#) Technical problems [Support](#)

This file was last modified Monday 2nd January 2017 13:05:46 GMT

## Sample output with explanation of output format:

**Output from HMMgene**

The output of the program is in **GFF** format, which is a sequence annotation format developed with gene finding in mind. It is very simple and therefore it is easy to develop programs in perl or awk to post-process the output. The following is an example of the form it takes with hmgene.

Note that hmgene only predicts coding regions. That is, the first exon ('firstex' below) is only the coding part of the first coding exon and similarly for the last exon ('lastex' below). Below a 'gene' therefore means the region of the gene from start to stop codon.

```

SEQ1 Hmgene1.1 firstex 692 702 0.347 + 2 bestparse:cds_1
SEQ1 Hmgene1.1 exon_1 2473 2711 0.421 + 1 bestparse:cds_1
SEQ1 Hmgene1.1 exon_2 2499 2609 0.401 + 1 bestparse:cds_1
SEQ1 Hmgene1.1 exon_3 10377 10563 0.861 + 2 bestparse:cds_1
SEQ1 Hmgene1.1 exon_4 11843 11891 0.857 + 2 bestparse:cds_1
SEQ1 Hmgene1.1 exon_5 12383 12396 0.593 + 0 bestparse:cds_1
SEQ1 Hmgene1.1 exon_6 12396 12521 0.593 + 1 bestparse:cds_1
SEQ1 Hmgene1.1 exon_7 13332 13415 0.926 + 1 bestparse:cds_1
SEQ1 Hmgene1.1 exon_8 13351 13603 1.000 + 0 bestparse:cds_1
SEQ1 Hmgene1.1 exon_9 13603 13643 1.000 + 0 bestparse:cds_1
SEQ1 Hmgene1.1 exon_10 14121 14408 0.999 + 0 bestparse:cds_1
SEQ1 Hmgene1.1 exon_11 14483 14579 0.877 + 1 bestparse:cds_1
SEQ1 Hmgene1.1 exon_12 14983 15030 0.835 + 0 bestparse:cds_1
SEQ1 Hmgene1.1 lastex 15643 15704 0.087 + 0 bestparse:cds_1
SEQ1 Hmgene1.1 CDS 692 15704 0.132 + . bestparse:cds_1

```

(the real list is tab separated)

**Columns**

1. Sequence identifier
2. Program name
3. Prediction (see table below for the meaning).
4. Beginning
5. End
6. Score between 0 and 1
7. Strand: +\$ for direct and -\$ for complementary
8. Frame (for exons it is the position of the donor in the frame)
9. Group to which prediction belongs. If several CDS's are found they will be called cds\_1, cds\_2, etc. 'bestparse' is there because alternative predictions will also be available (see below).

The score that comes with all the exons as well as the entire gene ('CDS' above) is a probability, so a value close to one means that the program is fairly certain. (See 'Known Bugs'.) The program also outputs some comment lines which are preceded by '#'.  
**Predictions**

Name	Meaning
firstex	The coding part of the first coding exon starting with the first base of the start codon.
exon_N	The Nth predicted internal coding exon.
lastex	The coding part of the last coding exon ending with the last base of the stop codon.
singlesex	The coding part of an exon in a gene with only one coding exon.
CDS	Coding region composed of the exon predictions prior to this line.
START	Predicted start codon with position of first and last base (only with signal option).
STOP	Predicted stop codon with position of first and last base (only with signal option).
DON	Predicted donor site with position of the base before and after the splice site (only with signal option).
ACG	Predicted acceptor site with position of the base before and after the splice site (only signal option).

## FGENESH

**FGENESH** is an HMM-based gene structure prediction program that can predict multiple genes considering both strands. The DNA sequence in FASTA format is input to the program along with the organism information. The output is quite detailed with the number of predicted genes, exons, strand information. For each predicted gene, the positions of the TSS, CDSs, and the poly(A) sites are provided. The predicted protein sequence for each of the genes can also be seen.

The screenshot shows the FGENESH service page on the Softberry website. The top navigation bar includes 'Home', 'Run Programs Online', and a search bar. The main content area is titled 'Services Test Online' and features the 'FGENESH' service. A brief description states: 'Used in more than 2800 publications' and cites 'Reference: Solovyev V, Kosarev P, Seledsov I, Vorobeyev D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol*. 2006;7, Suppl 1:P 10.1-10.12.' Below this, a note says: 'HMM-based gene structure prediction (multiple genes, both chains). The Fgenesh gene-finder was selected as the most accurate program for plant gene identification. Plant Molecular Biology (2005), 57, 3, 445-460. "Five ab initio programs (FGENESH, GeneMark.hmm, GENSCAN, GlimmerR and Graill) were evaluated for their accuracy in predicting maize genes. FGENESH yielded the most accurate and GeneMark.hmm the second most accurate predictions" (FGENESH identified 11% more correct gene models than GeneMark on a set of 1353 test genes.)' A text input field contains the sequence: 'gt13907843|eTNG\_000007.1| Homo sapiens genomic beta globin region (HBB@) on chromosome 11 GAATTCTAACTCCCCCTCAACCCTAACGTACCGCCATTGGATATAAAGATGTGT'. A note below says: 'Alternatively, load a local file with sequence in Fasta format:'. A file input field is labeled 'Local file name: [Browse]'. A note states: 'Select organism specific gene-finding parameters: Human (Homo sapiens) [Show example result] [Review]'. A note at the bottom left says: 'Most gene finding parameters presented here were trained by Softberry for its own use and distribution, using proprietary and publicly available data. Some of the parameters were created for our academic customers, including Broad Institute/MIT, Washington University, University of Minnesota and The Institute for Genomic Research (TIGR).'. A note at the bottom right says: 'Your use of Softberry programs signifies that you accept Terms of Use'. A note at the very bottom says: 'Last modification date: 24 Oct 2016'.

## Sample output:

Example: Homo sapiens genomic beta globin region (HBB@) on chromosome 11						
<a href="#">Show picture of predicted genes in PDF file</a>						
FGENESH 2.6 Prediction of potential genes in Homo_sapiens genomic DNA						
Time : Sun Jul 25 18:43:27 2021						
Seq name: gil13907843 ref NG_000007.1  Homo sapiens genomic beta globin region (HBB@) on						
Length of sequence: 73308						
Number of predicted genes 10: in +chain 10, in -chain 0.						
Number of predicted exons 21: in +chain 21, in -chain 0.						
Positions of predicted genes and exons: Variant 1 from 1, Score:180.171887						
G	Str	Feature	Start	End	Score	ORF
						Len
1	+	TSS	19456		-7.09	
1	+	1 CDSf	19541	-	19632	16.13
1	+	2 CDSi	19755	-	19977	13.37
1	+	3 CDSL	20833	-	20961	3.34
1	+	PoLA	21055			1.12
2	+	TSS	34446		-7.09	
2	+	1 CDSf	34531	-	34622	13.42
2	+	2 CDSi	34745	-	34967	21.52
2	+	3 CDSL	35854	-	35982	2.92
2	+	PoLA	36043			1.12
3	+	TSS	39382		-7.09	
3	+	1 CDSf	39467	-	39558	13.42
3	+	2 CDSi	39681	-	39903	21.52
3	+	3 CDSL	40770	-	40898	3.66
3	+	PoLA	40959			1.12
4	+	TSS	44415		-8.69	
4	+	1 CDSf	45995	-	46151	16.58
4	+	2 CDSi	46997	-	47100	-1.94
4	+	PoLA	47243			1.12
5	+	TSS	54707		-4.39	
5	+	1 CDSf	54790	-	54881	13.44
5	+	2 CDSi	55010	-	55232	17.01
5	+	3 CDSL	56425	-	56535	2.53
5	+	PoLA	56931			1.12
6	+	TSS	62104		-6.59	
6	+	1 CDSf	62187	-	62278	12.99
6	+	2 CDSi	62409	-	62631	20.06
6	+	3 CDSL	63482	-	63610	9.54
6	+	PoLA	63718			1.12

## AUGUSTUS

**AUGUSTUS** is a gene prediction program for eukaryotic sequences. The program takes as input the sequence in FASTA format along with the organism information. For every predicted gene, the program lists the transcript coordinates, start/stop positions, CDSs, coding sequence, and the predicted protein sequence.



## Splign

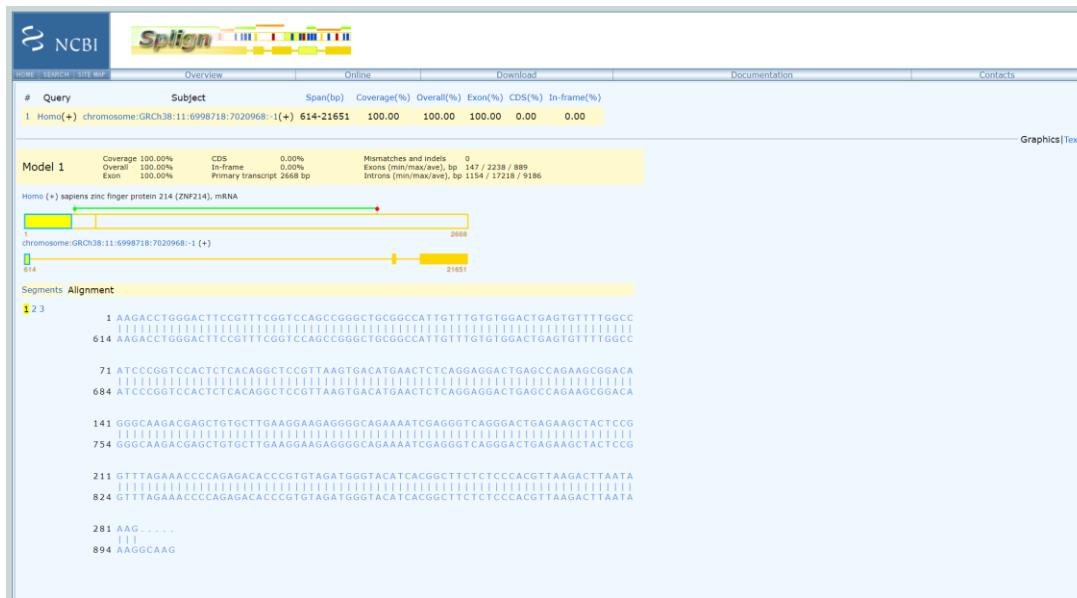
[Splign](#), hosted by NCBI aligns cDNA to genomic DNA and identifies splice-junction locations, frameshifts, alternate gene models where possible. Splign also supports cross-species alignments. The output of the program is the gene model for the input cDNA sequence.

**Example:** Given the cDNA and the genomic sequences, Splign can be used to find the mRNA and CDS coordinates in the genomic DNA.

mRNA locations: 614-896, 18114- 18260, 19414-21651 (3 exons, with exon 1 being the 5'UTR)

CDS locations: 303-430, 431-2124.

Three exons can be seen in the output below, with the green line indicating the coding sequence (CDS starts in Exon2 and continues into exon 3). Exon1 is the 5'UTR.



The figure shows the SpliceN web interface with the following details:

- Header:** NCBI logo, SpliceN logo, and tabs for Overview, Online, Download, Documentation, and Contacts.
- Table:** A table showing search results for "Homo(+) chromosome:GRCh38:11-6998718-7020968-1(+)" with columns: #, Query, Subject, Span(bp), Coverage(%), Overall(%), Exon(%), CDS(%), In-frame(%). The values are: 1, Homo(+), chromosome:GRCh38:11-6998718-7020968-1(+), 614-21651, 100.00, 100.00, 100.00, 0.00, 0.00.
- Model 1:** A yellow box containing coverage statistics: Coverage 100.00%, CDS 0.09%, Overall 100.00%, In-frame 0.09%, and Primary transcript 2668 bp. It also lists mismatch counts: 0 mismatches and 0 indels, and exon/intron coordinates: Exons (min/max/avg), bp 147 / 2238 / 889 and Introns (min/max/avg), bp 1154 / 17218 / 9186.
- Genomic View:** A diagram showing the genomic region from 614 to 21651. A red arrow indicates the direction of transcription. A blue box highlights the primary transcript. A green box highlights the coding sequence (CDS).
- Segments:** A table showing segments aligned to the genomic region. Segments 1, 2, and 3 are listed with their corresponding genomic coordinates and sequences.

Text view of the output:

#	Query	Subject	Span(bp)	Coverage(%)	Overall(%)	Exon(%)	CDS(%)	In-frame(%)
1	Homo(+)	chromosome:GRCh38:11:6998718:7020968:-1(+)	614-21651	100.00	100.00	100.00	0.00	0.00

#	Query	Subject	Idty	Len	Q.Start	Q.Fin	S.Start	S.Fin	Type	Details
+1	Homo	chromosome:GRCh38:11	1	283	1	283	614	896	<exon>GC	M283
+1	Homo	chromosome:GRCh38:11	1	147	284	430	18114	18260	AG<exon>GT	M147
+1	Homo	chromosome:GRCh38:11	1	2238	431	2668	19414	21651	AG<exon>	M2238

When an EST-sequence is available instead of a full-length cDNA, a genomic [BLAST](#) can be performed to align the EST to the genomic sequence in order to determine the location of the gene. Using this method, the full structure of the gene is not available.

## BLAT

[BLAT](#) (BLAST-like alignment tool) at UCSC is similar to NCBI's Splign. It aligns sequences to an entire genome. But unlike Splign, BLAT can align the query sequence to anywhere in the genome and it is not necessary to specify a region of genomic DNA for alignment purposes. It does not however have a graphical interface. The user can select the organism and assembly for the input sequence. The output provides the user with links to open the elements in the UCSC genome browser and provides the score, percent identity, start and end coordinates, chromosome, strand, and span information.

Genomes Genome Browser Tools Mirrors Downloads My Data Projects Help About Us

### Human BLAT Search

#### BLAT Search Genome

Genome:  Search all Human Assembly: Dec. 2013 (GRCh38/hg38) Query type: BLAT's guess Sort output: query.score Output type: hyperlink

All Results (no minimum matches)

Paste in a query sequence to find its location in the genome. Multiple sequences may be searched if separated by lines starting with '>' followed by the sequence name.

**File Upload:** Rather than pasting a sequence, you can choose to upload a text file containing the sequence.  
Upload sequence:  No file chosen

Only DNA sequences of 25,000 or fewer bases and protein or translated sequence of 10000 or fewer letters will be processed. Up to 25 sequences can be submitted at the same time. The total limit for multiple sequence submissions is 50,000 bases or 25,000 letters.  
A valid example is `>cccggAACCAAGGACCTCGGGGTGGCTAACCC` (human SOD1).

The **Search all** checkbox allows you to search all genomes at the same time. Search all is only available for default assemblies and attached hubs with dedicated BLAT servers. The new dynamic BLAT servers are not supported, and they are noted as skipped in the output. See our [BLAT All FAQ for more information](#).

The **All Results** checkbox disables minimum matches filtering so all results are seen. For example, with a human dna search, 20 is minimum matches required, based on the genome size, to filter out lower-quality results. This checkbox can be useful with short queries and with the tiny genomes of microorganisms.

For locating PCR primers, use [In-Silico PCR](#) for best results instead of BLAT.

## Output:

Genomes Genome Browser Tools Mirrors Downloads My Data Projects Help About Us

### Human (hg19) BLAT Results

#### BLAT Search Results

Go back to [chrX:15,578,261-15,621,068](#) on the Genome Browser.

Custom track name:  Custom track description:

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN	
browser	details	YourSeq	1123	1	1127	1127	100.0%	chr7	+	22766819	22771617	4799
browser	details	YourSeq	45	347	626	1127	90.6%	chr1	-	6965981	7238975	272995
browser	details	YourSeq	25	835	859	1127	100.0%	chr9	-	4178871	4178895	25
browser	details	YourSeq	25	719	751	1127	89.3%	chr13	-	71492085	71492116	32
browser	details	YourSeq	22	1108	1127	1127	100.0%	chr4	+	124365926	124365947	22
browser	details	YourSeq	21	521	541	1127	100.0%	chr4	-	13939846	13939866	21
browser	details	YourSeq	21	1042	1062	1127	100.0%	chrX	+	53470821	53470841	21
browser	details	YourSeq	21	1069	1089	1127	100.0%	chr1	+	86450246	86450266	21
browser	details	YourSeq	20	841	860	1127	100.0%	chr1	-	174444164	174444183	20

#### Help

Missing a match?  
[What is chr\\_alt & chr\\_fix?](#)

## BioMart

BioMart tool is found in the top navigation panel of the [Ensembl](#) homepage. It allows users to download data sets based on certain criteria. The current build version (as of August 2021) is 104. “Filters” present on the left menu helps restrict the query by many parameters like region, gene, phenotype, gene ontology, variants etc. Clicking on the “Attributes” in the left menu shows us to select attribute categories. Here we can select the columns that we want in our data set. Once we have finalized on the filters and attributes, we can view the results by clicking on the “Results” button on the top left of the page. Options to download the data sets in TSV, HTML, CSV, XLS formats are available. The data set can also be previewed on the browser. Clicking on “Count” tells us the number of items in our data set.

A worked example of a scenario can be found [here](#). This would familiarize one to the usage of BioMart. The example uses an older build version though.

The screenshot shows the Ensembl BioMart search interface. The top navigation bar includes links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. A search bar at the top right is labeled "Search all species...". The main search area has tabs for New, Count, and Results. On the left, a sidebar shows the dataset selected: Human genes (GRCh38.p13) and the filters applied: With MIM morbid ID(s): Only. Under Attributes, options like Gene stable ID, Gene stable ID version, Transcript stable ID, and Transcript stable ID version are listed. The main search area has a section titled "Please restrict your query using criteria below (If filter values are truncated in any lists, hover over the list item to see the full text)". It includes fields for REGION, GENE (with options to limit by external references or input a list), and AFFY HC G110 probe ID(s).

The previous builds are archived and is available [here](#).

The screenshot shows the Ensembl Archive BioMart search interface. The top navigation bar includes links for BioMart, Downloads, Help & Docs, and Blog. A search bar at the top right is labeled "Search all species...". The main search area has tabs for New, Count, and Results. On the left, a sidebar shows the dataset selected: Dataset 64 / 56305 Genes (Mouse genes (GRCm38.p6)) and the filters applied: Chromosome/scaffold: 11, Band Start: E2, Band End: E2, Transcript count >=: 7, and With RefSeq peptide ID(s): Only. Under Attributes, options like Gene stable ID, Gene stable ID version, Transcript stable ID, Transcript stable ID version, and RefSeq peptide ID are listed. The main search area includes sections for EXTERNAL (with checkboxes for various IDs like CCDS ID, ChEMBL ID, and PDB ID), GO (with checkboxes for GO term accession, name, definition, evidence code, and domain), GOSlim GOA (with checkboxes for GOSlim GOA Accession(s) and Description), and External References (with checkboxes for various databases and IDs like Reactome ID, RefSeq mRNA ID, and RefSeq ncRNA ID).

## R/Bioconductor and BiomaRt

R is an open-source programming tool to mine data from many different data sources.

The [Bioconductor Project](#) is an open-source initiative to use R. Using R, genome scale data analysis can be accomplished more easily than by repeated lookups in the various genome browsers.

BiomaRt is a Bioconductor package that uses R to access data from data marts like Ensembl. An initial setup to use BiomaRt requires installation of R studio, Bioconductor tool biocLite, and use biocLite to install biomaRt.

The R commands are:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("biomaRt")
```

To use the library the following command needs to be run in R studio:

```
library(biomaRt)
```

One of the commonly used commands is getBM. It has four arguments – attributes, filters, values, and mart. Attributes are columns that we want in the output, filters are the limit the result set, values are specific for filters, and mart is the data source.

Example (Screen shot from R studio):

```
#use ensembl database
ensembl = useMart("ensembl")
listDatasets(ensembl)
ensembl = useDataset("hsapiens_gene_ensembl", mart=ensembl)
getBM(attributes=c("ensembl_transcript_id", "hgnc_symbol"), filters="entrezgene_id", values="6928",
      mart=ensembl)
```

A worked example using BiomaRt is below:

Search OMIM.org for "huntington's disease". The first five entries all have this or a similar phrase in the title. Record the five identifiers (six-digit numbers) of those five records. The corresponding biomaRt filter name for these identifiers is "mim\_morbid\_accession". Use biomaRt to retrieve two tables with the following attributes, limiting to the five MIM values you found:

First table: Entrez Gene ID, HGNC symbol, Ensembl Gene ID

Second table: HGNC symbol, Ensembl Gene ID, Ensembl Transcript ID

The first five entries in the OMIM database for the search string “huntington's disease” are shown in the screenshot below. Their record identifiers correspond to 603218, 604802, 143100, 606438, and 607136.

Screen shot of the first five records from the OMIM database search:

**Table 1:**

**R code for Table 1:**

```
library(biomaRt)
listMarts()

#use ensembl database
ensembl = useMart("ensembl")
listDatasets(ensembl)
ensembl = useDataset("hsapiens_gene_ensembl",mart=ensembl)

filters = listFilters(ensembl)
filters[1:100,]
attributes = listAttributes(ensembl)
attributes[1:300,]

#Table 1
omim_ids = c("603218", "604802", "143100", "606438", "607136")
getBM(attributes=c("entrezgene_id", "hgnc_symbol", "ensembl_gene_id"),
      filters=c("mim_morbid_accession"),
      values=list(omim_ids),
      mart = ensembl)
```

Table1 Output:

	entrezgene_id	hgnc_symbol	ensembl_gene_id
1	3064	HTT	ENSG00000197386
2	5621	PRNP	ENSG00000171867
3	57338	JPH3	ENSG00000154118
4	6908	TBP	ENSG00000112592

**Table 2:**  
**R code for table 2:**

```

library(biomaRt)
listMarts()

#use ensembl database
ensembl = useMart("ensembl")
listDatasets(ensembl)
ensembl = useDataset("hsapiens_gene_ensembl",mart=ensembl)

filters = listFilters(ensembl)
filters[1:100,]
attributes = listAttributes(ensembl)
attributes[1:300,]

#Table 2
omim_ids = c("603218", "604802", "143100", "606438", "607136")
getBM(attributes = c("hgnc_symbol", "ensembl_gene_id", "ensembl_transcript_id"),
      filters = c("mim_morbid_accession"),
      values = list(omim_ids),
      mart = ensembl)

```

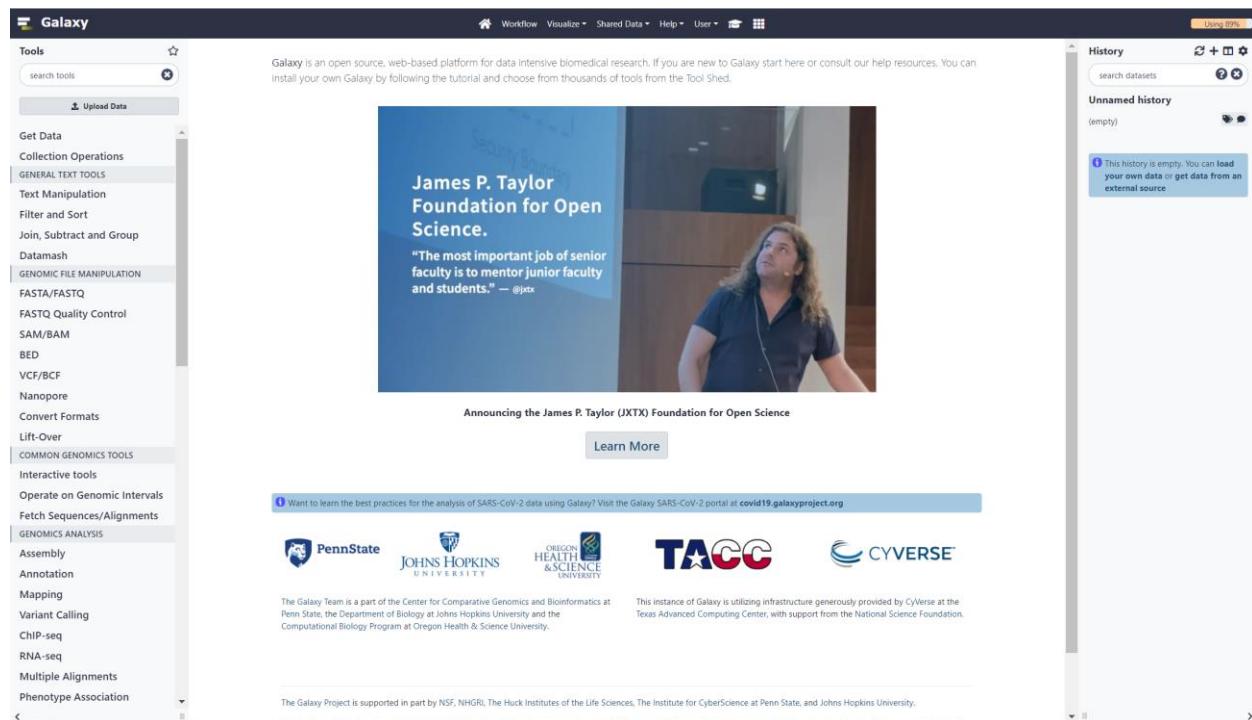
Table 2 output:

	hgnc_symbol	ensembl_gene_id	ensembl_transcript_id
1	HTT	ENSG00000197386	ENST00000680239
2	HTT	ENSG00000197386	ENST00000680956
3	HTT	ENSG00000197386	ENST00000680360
4	HTT	ENSG00000197386	ENST00000681528
5	HTT	ENSG00000197386	ENST00000647962
6	HTT	ENSG00000197386	ENST00000649900
7	HTT	ENSG00000197386	ENST00000680291
8	HTT	ENSG00000197386	ENST00000355072
9	HTT	ENSG00000197386	ENST00000648150
10	HTT	ENSG00000197386	ENST00000506137
11	HTT	ENSG00000197386	ENST00000512909
12	HTT	ENSG00000197386	ENST00000510626
13	HTT	ENSG00000197386	ENST00000649131
14	HTT	ENSG00000197386	ENST00000509618
15	HTT	ENSG00000197386	ENST00000650588
16	HTT	ENSG00000197386	ENST00000650595
17	HTT	ENSG00000197386	ENST00000513639
18	HTT	ENSG00000197386	ENST00000513326
19	HTT	ENSG00000197386	ENST00000509043
20	HTT	ENSG00000197386	ENST00000509751
21	HTT	ENSG00000197386	ENST00000512068
22	HTT	ENSG00000197386	ENST00000513806
23	HTT	ENSG00000197386	ENST00000508321
24	PRNP	ENSG00000171867	ENST00000430350
25	PRNP	ENSG00000171867	ENST00000379440
26	PRNP	ENSG00000171867	ENST00000424424
27	PRNP	ENSG00000171867	ENST00000457586
28	JPH3	ENSG00000154118	ENST00000537256
29	JPH3	ENSG00000154118	ENST00000301008
30	JPH3	ENSG00000154118	ENST00000284262
31	JPH3	ENSG00000154118	ENST00000563609
32	TBP	ENSG00000112592	ENST00000421512
33	TBP	ENSG00000112592	ENST00000230354
34	TBP	ENSG00000112592	ENST00000423353
35	TBP	ENSG00000112592	ENST00000636632
36	TBP	ENSG00000112592	ENST00000446829
37	TBP	ENSG00000112592	ENST00000392092
38	TBP	ENSG00000112592	ENST00000540980
39	TBP	ENSG00000112592	ENST00000616883

# Galaxy

[Galaxy](#) is an open-source framework integrating command-line tools and database information giving the users the choice to perform various complex analysis on genomic data without the need to learn command line syntax. Users can create accounts, generate workflows of their analysis, and share them with peers. It is used by both beginners and seasoned programmers due to its ease of use and integration with other data sources like UCSC.

Galaxy interface looks like the one in the screen shot below:



The tutorials on Galaxy also called screencasts can be found [here](#). A detailed walk through of a scenario is found at [Galaxy 101](#).

**Example:** Use Galaxy or any other tool (hg38) to find all UCSC flagged SNPs in (hg38/db147) on chromosome 1. Be sure to import the data as a BED file. Describe the result (number of lines, what's in columns, etc.).

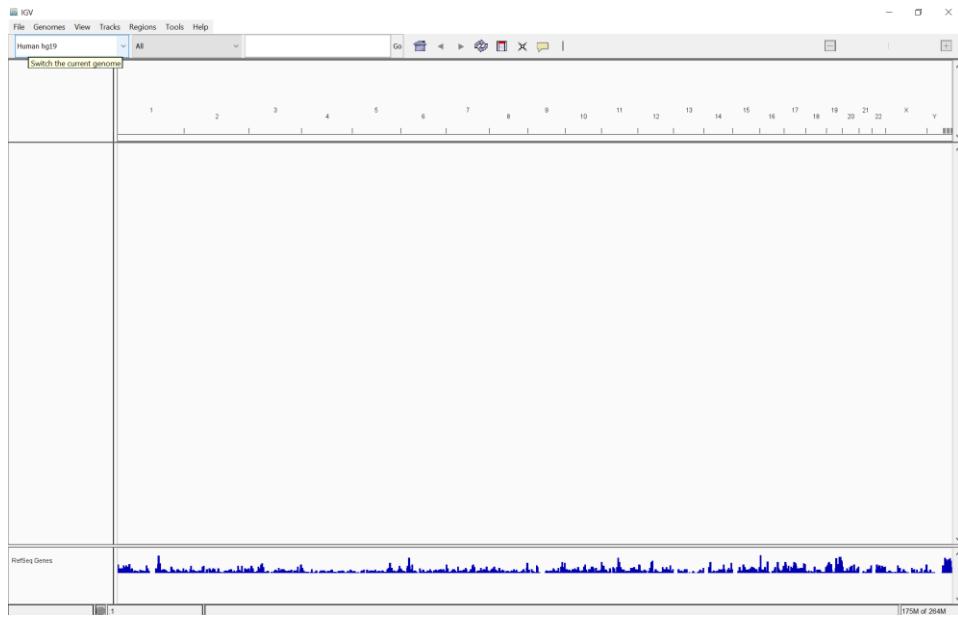
Click on Get Data -> UCSC Table Browser. Put in the necessary details in the Table Browser and import the data into Galaxy. The BED file has 8787 lines. The columns in the BED file are chrom (name of the chromosome on which the feature is present), start (starting position

of the feature in the chromosome), end (ending position), name (name of the BED feature), score (range 0 to 100), and strand (either '+' or '-').

## IGV (Integrated Genome Viewer)

[\*\*IGV\*\*](#) available through the Broad Institute is like the other genome browsers, with the added advantage that it can be locally downloaded and can be run as a standalone application. It allows us to view at high resolution for small regions and low resolution at wider regions. IGV is designed to handle very large genomic-sized data sets. IGV allows for custom data to be uploaded in BED and WIG file formats.

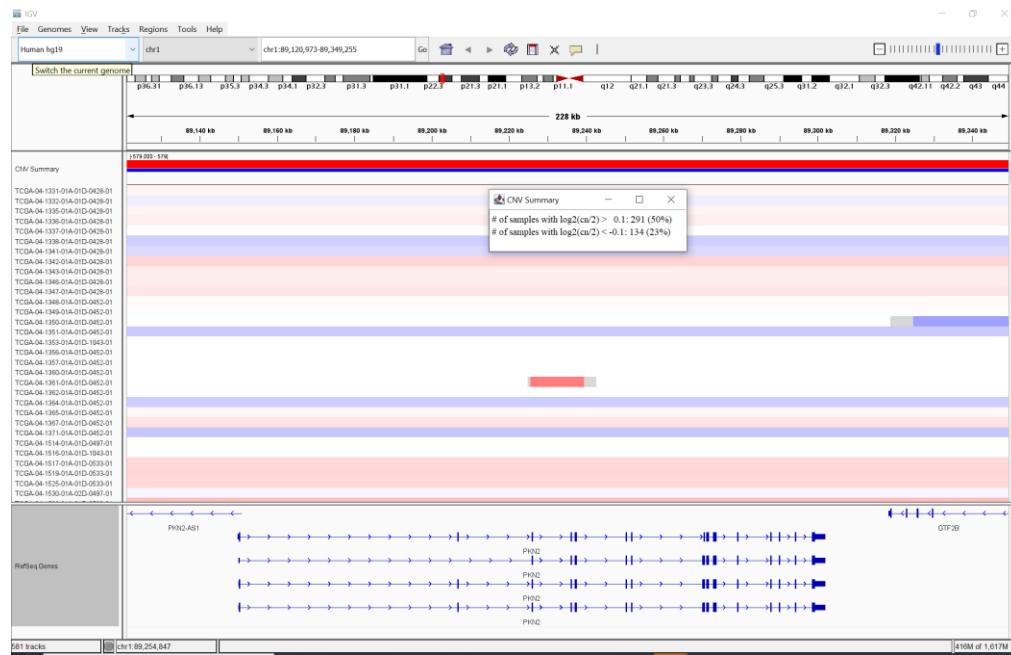
## IGV desktop interface:



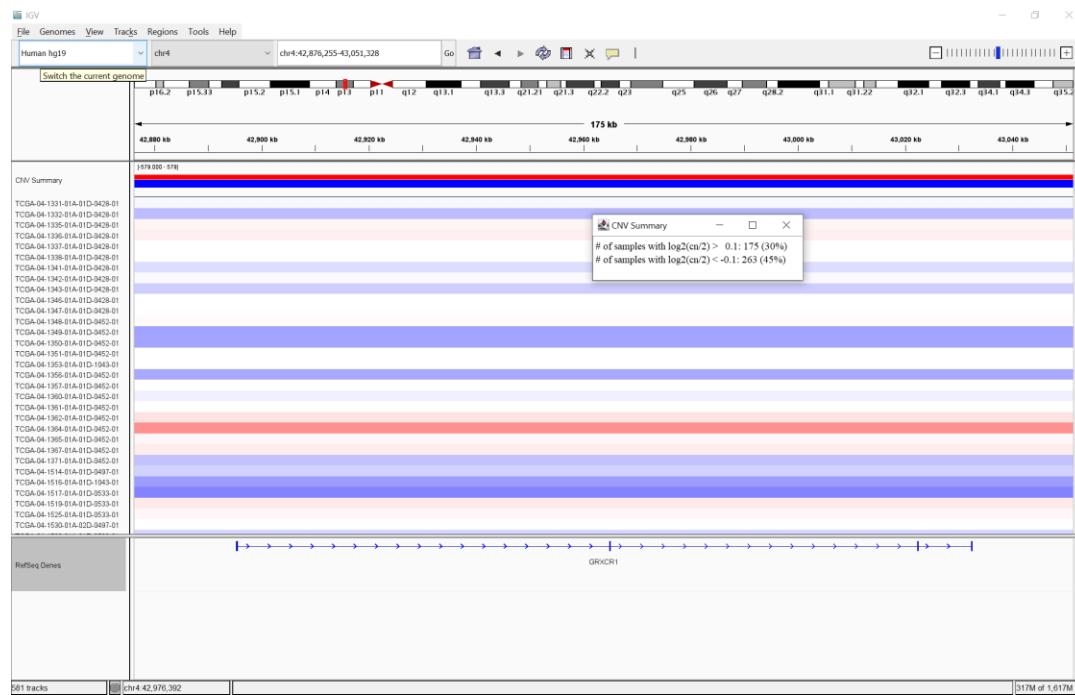
**Example:** Using IGV, load the Firehose (TCGA) data from January 28, 2016 for OV-TP (ovarian cancer) using hg19. The following four genes have been shown to be associated with CNV in some forms of ovarian cancer: PKN2, GRXCR1, PRKN (or PARK2), and PPIAL4A. In a table, qualitatively evaluate the CNV summary (minus germline) in each gene's region (e.g. "even blue and red", "twice as much blue as red", "overwhelmingly red").

Gene	Red %	Blue %	
PKN2	50	23	For the PKN2 gene, red is twice as much as blue, indicating that in 50% of patients the CNV increased, and 23% patients showed decrease in CNV.
GRXCR1	30	45	For GRXCR1 gene, the blue is more than red, which means that there are more patients with decreased CNV (45%) than with an increase in CNV (30%).
PRKN	17	64	For PRKN gene, the CNV shows overwhelmingly blue, the blue is greater than red by more than 3 times, indicating that in majority of the patients (64%) there was a decrease in CNV compared to 17% of patients that showed increase in CNV.
PPIAL4A	60	2%	For the PPIAL4A gene, CNV shows overwhelmingly red with 60% patients showing an increase in CNV, and only 2% showing a decrease in CNV.

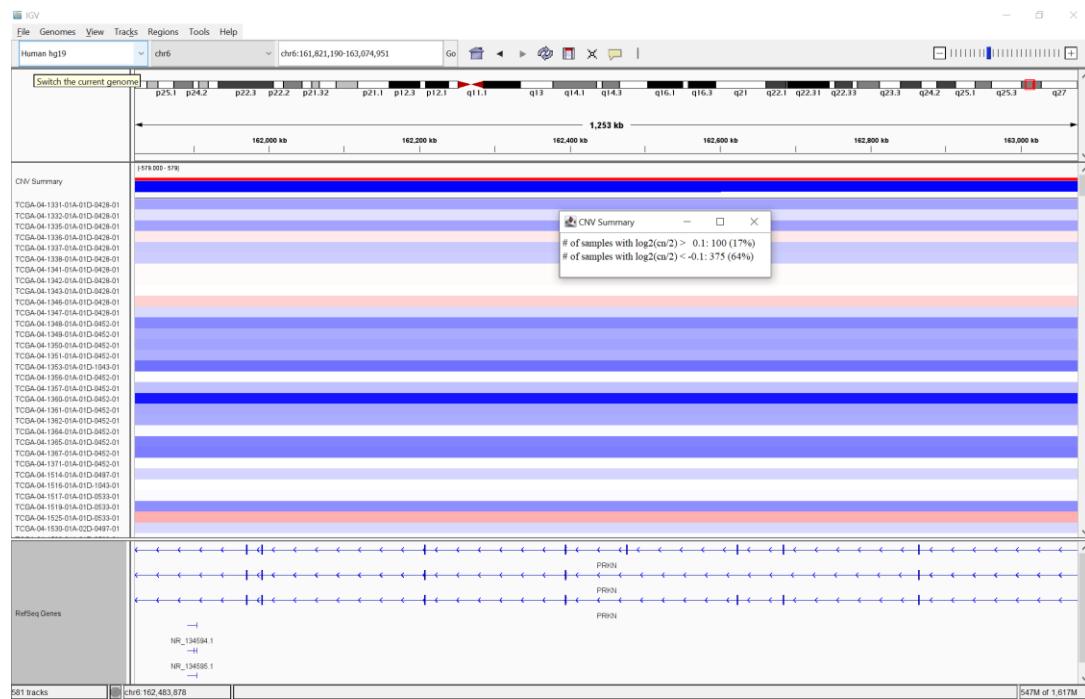
## PKN2 gene CNV summary:



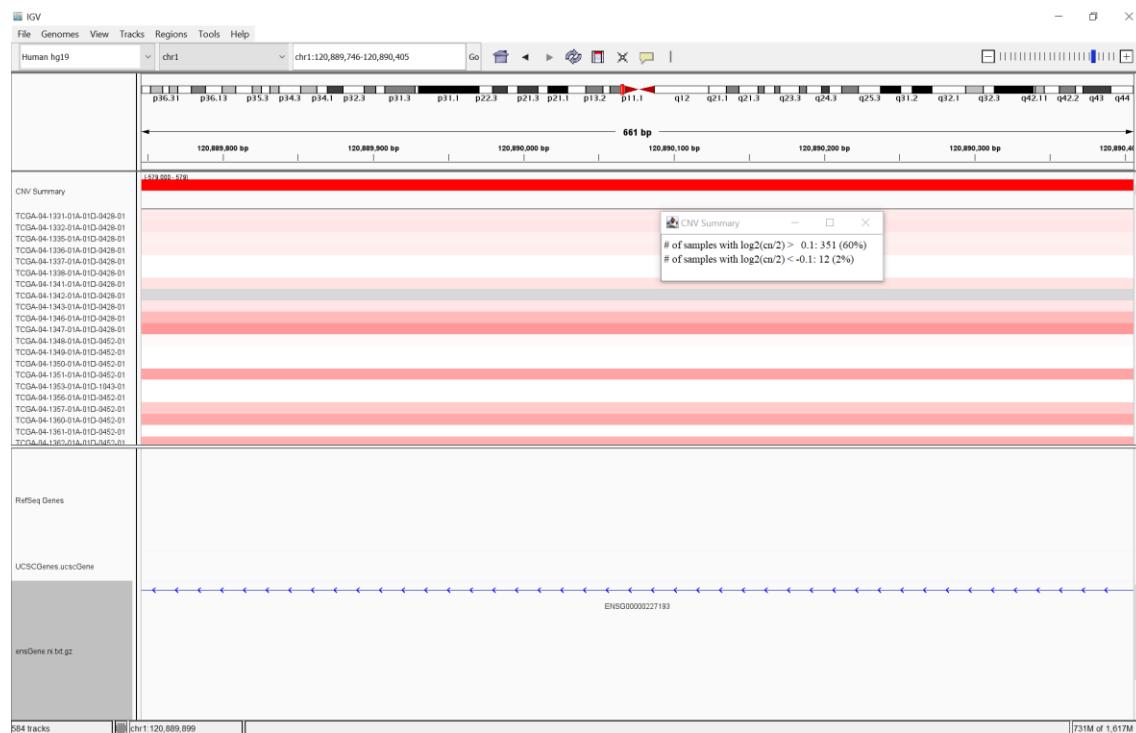
## GRXCR1 gene CNV summary:



## PRKN gene CNV summary:



## PPIAL4A gene CNV summary:



## BEDtools

Bedtools is a popular collection of tools that allow us to perform operations on genomic data. Bedtools are available in Galaxy, but command-line version is also available. For Mac OS and Linux, instructions to install BEDtools can be found [here](#).

BEDtools intersect is a popular interval manipulation operation, with a detailed documentation found [here](#).

**Example:** Find the closest exons to gaps on chromosome 20. To do so, run the following command.

```
bedtools closest -a hs_chr20_gaps.bed -b hs_chr20_refseq.bed >  
hs_chr20_nearest_exon_gap.bed
```

This creates an output file called hs\_chr\_20\_nearest\_exon\_gap.bed

The operations that can be performed with bedtools are shown in the screen shot below:

```
bedtools: flexible tools for genome arithmetic and DNA sequence analysis.  
usage:   bedtools <subcommand> [options]  
  
The bedtools sub-commands include:  
  
[ Genome arithmetic ]  
intersect      Find overlapping intervals in various ways.  
window         Find overlapping intervals within a window around an interval.  
closest        Find the closest, potentially non-overlapping interval.  
coverage       Compute the coverage over defined intervals.  
map            Apply a function to a column for each overlapping interval.  
genomcov       Compute the coverage over an entire genome.  
merge          Combine overlapping/nearby intervals into a single interval.  
cluster        Cluster (but don't merge) overlapping/nearby intervals.  
complement    Extract intervals _not_ represented by an interval file.  
shift          Adjust the position of intervals.  
subtract      Remove intervals based on overlaps b/w two files.  
slop           Adjust the size of intervals.  
flank          Create new intervals from the flanks of existing intervals.  
sort           Order the intervals in a file.  
random         Generate random intervals in a genome.  
shuffle        Randomly redistribute intervals in a genome.  
sample         Sample random records from file using reservoir sampling.  
spacing        Report the gap lengths between intervals in a file.  
annotate      Annotate coverage of features from multiple files.
```

## SAMtools

SAMtools are tools for alignments in the SAM format. It is available to be downloaded [here](#). The same is also available through Galaxy interface. SAMtools can be used to call SNPs in genomic alignments. BAM files can be displayed in IGV.

Some of the SAMtools commands are:

SAM-to-BAM conversion:

Step 1: Create a BAM file using *view* command. Below, -b specifies that the output is a BAM file, -S indicates that the input is a SAM file.

```
samtools view -bS filename.sam >filename.bam
```

Step 2: Sorting the BAM file using the *sort* command. In Galaxy, *SAMtools* does the sorting.

```
samtools sort filename.bam filename.sorted
```

Step 3: Indexing the sorted BAM file. The *index* command will create the BAM index file. Both the BAM and the BAI (BAM index file) must be downloaded. IGV will look for the .bai file with the same name as the BAM file in the same folder.

Below are the various operations that can be performed using command line SAMtools:

```
Usage: samtools <command> [options]

Commands:
-- Indexing
dict          create a sequence dictionary file
faidx         index/extract FASTA
fqidx         index/extract FASTQ
index          index alignment

-- Editing
calmd         recalculate MD/NM tags and '=' bases
fixmate       fix mate information
reheader      replace BAM header
targetcut     cut fosmid regions (for fosmid pool only)
addreplacerg  adds or replaces RG tags
markdup       mark duplicates

-- File operations
collate       shuffle and group alignments by name
cat           concatenate BAMs
merge          merge sorted alignments
mpileup        multi-way pileup
sort           sort alignment file
split          splits a file by read group
quickcheck    quickly check if SAM/BAM/CRAM file appears intact
fastq          converts a BAM to a FASTQ
fasta          converts a BAM to a FASTA

-- Statistics
bedcov        read depth per BED region
depth          compute the depth
flagstat       simple stats
idxstats      BAM index stats
phase          phase heterozygotes
stats          generate stats (former bamcheck)

-- Viewing
flags          explain BAM flags
tview          text alignment viewer
view           SAM<->BAM<->CRAM conversion
depad          convert padded BAM to unpadded BAM
```

## Next-Generation Sequencing (NGS) analysis

Protocol to perform NGS analysis to identify SNPs (variant calling) in NGS data from the 1000 Genomes Project (reference genome hg19). All the steps below can be performed within [Galaxy](#) or using command-line tools.

- Two FASTQ files from the 1000 Genome Project is used. The FASTQ files represent data from paired-end sequencing experiment. The forward reads usually are in the file ending in '\_1', and the reverse reads are in the file ending in '\_2'. Links to the sample data files are given below.
- **Data upload to Galaxy:** Load both files into Galaxy using the **Upload file** tool, choosing **Paste/Fetch data**, and pasting in the given ftp links. The data type should be set to 'fastq' and the genome should be set to 'hg19'.

### Forward reads:

[ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/HG00117/sequence\\_read/SRR044234\\_1.filt.fastq.gz](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/HG00117/sequence_read/SRR044234_1.filt.fastq.gz)

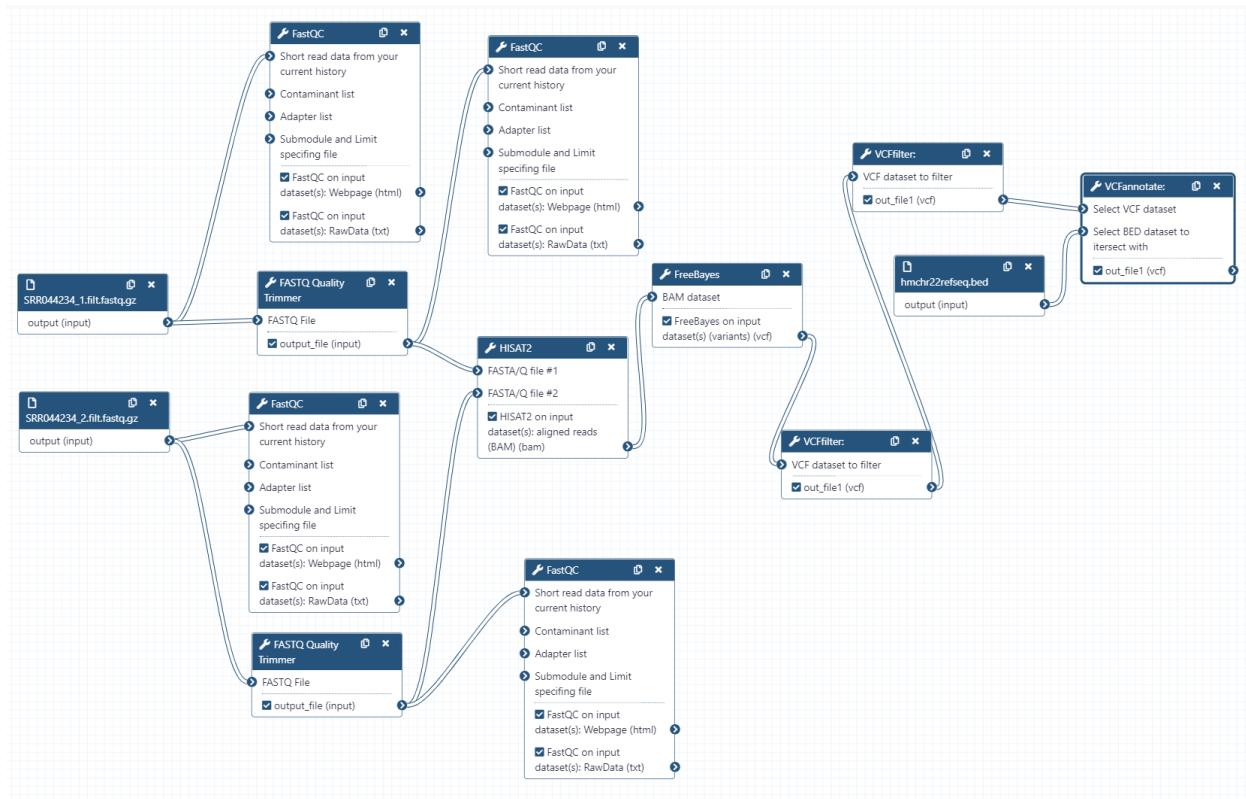
### Reverse reads:

[ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/HG00117/sequence\\_read/SRR044234\\_2.filt.fastq.gz](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/HG00117/sequence_read/SRR044234_2.filt.fastq.gz)

- **Determine quality encoding:** Run **FASTQC** on both files. If the quality encoding is found to be Sanger/Illumina 1.9, update the file type to 'fastqsanger'. If the quality encoding is found to be something other than Sanger/Illumina 1.9, use **FASTQ Groomer** to convert the files to Sanger/Illumina 1.9 encoding. At the end, we will get two FASTQ files in Sanger/Illumina 1.9 encoding.
- **Trim low-quality bases:** Either **FASTQ Quality Trimmer** or **Trimmomatic** tool is used to remove low quality bases from each file. A window size of 4 bases and average quality in the window can be set based on our requirement (20 or 30).
- **FASTQC** is rerun on the trimmed data to ensure that low quality bases were removed.
- **Align reads to reference genome:** There are many alignment tools to align the FASTQ files to the reference genome. **BWA**, **Bowtie2**, or **HISAT** are different alignment tools. Since data is from a paired-end experiment, we must align the reads as paired-end. This step generates the alignments in BAM format.
- **Identify variants:** There are many variant-calling approaches. Some of these tools are FreeBayes, SAMtools (both available in Galaxy), [\*\*GATK Unified Genotyper\*\*](#), [\*\*GATK haplotype Caller\*\*](#), and [\*\*Platypus\*\*](#). The output can be restricted to a specific chromosome to make the result file more manageable. (Ex: Limit the output to chr22:0-51304566)

- **Filter and annotate variants:** Use the **VCFfilter** tool to filter for variants that show heterozygosity (estimated allele frequency = 0.5) and have more than 10 reads covering them (total read depth > 10). The tag IDs for these parameters can be found in the header of the VCF file. To annotate which genes the variants are in, we need to bring in RefSeq genes in BED format from **UCSC Main**. Then, using the **VCFannotate** tool we can intersect the filtered VCF file with the BED annotations.

Workflow diagram from Galaxy:



## ChIP-seq data analysis

The raw ChIP-seq unaligned reads in the FASTQ format are uploaded into Galaxy. The data is then groomed (FASTQC) and trimmed (Trimmomatic) after which it is aligned to a reference genome. Bowtie2 and BWA are a couple of alignment programs that can perform the alignment. The next step is to perform ‘peak calling’ which identifies the peaks in the reads that are aligned to the reference genome. The peaks represent regions in the DNA that are bound to

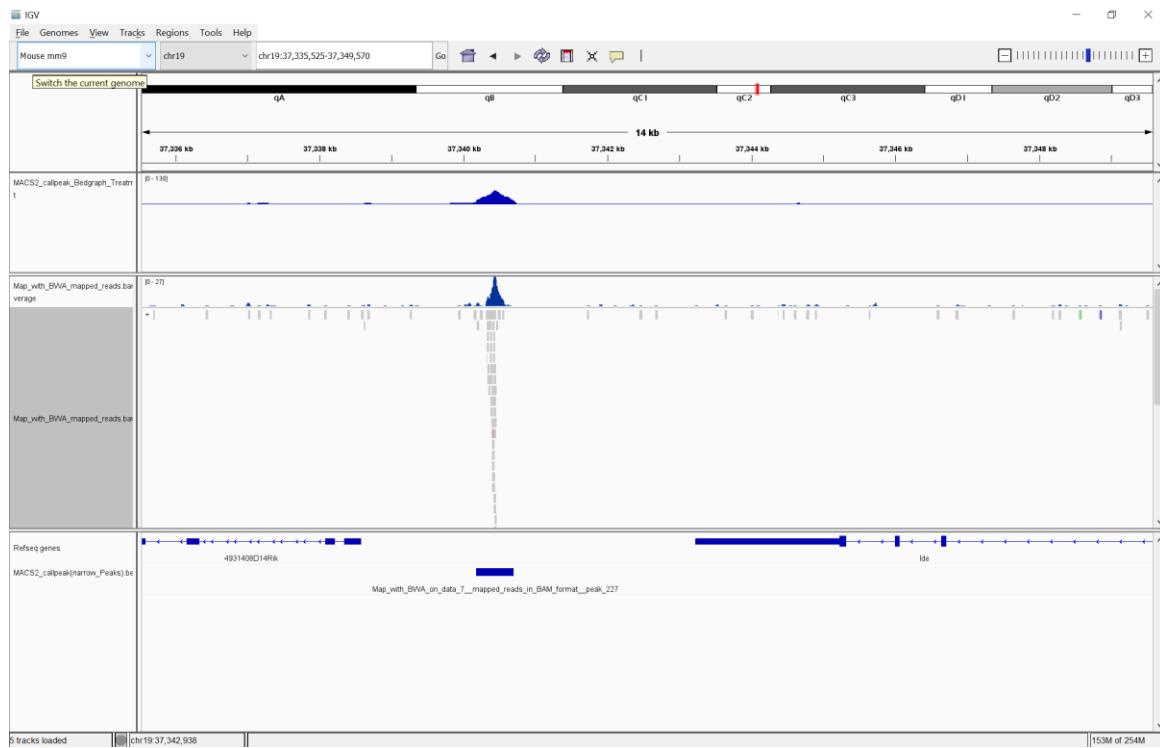
the proteins that was immunoprecipitated. MACS2 and Sicer are two popular peak callers. The input to the program is a BAM file, and the output produced is a BED (Browser extensible data) or WIG (Wiggle) format file. The peak caller used depends on factors such as presence of replicates, control experiments, single/paired end, and type of protein immunoprecipitated. The necessary tools are available in Galaxy or can be executed from command-line. If we need to determine the binding changes under certain conditions or cell types, differential peak calling is to be used. A ChIP-seq tutorial in Galaxy is available [here](#).

Protocol to perform ChIP-seq data analysis (All the necessary tools are available in Galaxy):

- Import the FASTQ file of ChIP-seq reads from mouse (mm9). The full dataset was downsampled to the subset reads from chr19.
- Run **FASTQC** to determine the quality score encoding.
- Run **FASTQ Groomer** to convert the file to Sanger/Illumina 1.9 phred encoding ONLY IF NEEDED.
- Run **Trimmomatic** and set the minimum phred score in a 4 nt sliding window to 25.
- Re-run **FASTQC** to check the quality scores and encoding scheme.
- Run **Map with BWA** with default settings (make sure to select for single-end reads), aligning against the mouse genome version mm9.
- Run **MACS2 callpeak** on the BAM file, setting the Effective genome size to the mouse genome. Choose “Scores in bedGraph files (--bdg)” option under additional outputs. Use default settings for the rest of the parameters and leave the control field blank.
- Load the MACS2 Bedgraph Treatment file and narrow Peaks BED file from previous step and aligned BAM file to IGV or UCSC.
- Focus on chromosome 19 and look for a gene locus that has ChIP peaks nearby.

The MACS2 results can be viewed in IGV browser. It shows the ChIP peaks. There are two genes in this vicinity – the Ide (insulin degrading enzyme) gene and the RIKEN cDNA 4931408D14 gene. We can analyze ChIP-seq data pertaining to histone modifications that has greater coverage over the entire gene region using the MACS2 ‘broad peaks’ option. ‘Narrow peaks’ can be for transcription factor data since the region that is bound would be narrower (Mistry & Khetani, n.d.).

BEDgraph treatment file, MACS2 narrow peaks file, and the aligned BAM file viewed in the IGV browser:



## RNA-seq data analysis

RNA-seq analysis is used to study the transcriptome of an organism that have a reference genome available. RNA-seq reads are aligned to the genome to determine the origination of the RNA molecule. Many tools in the TopHat suite are widely used to map and analyze RNA-seq data. TopHat suite also includes the Bowtie aligner. In case of organisms with no well-annotated genome, de novo assemblers need to be used (Ex: Trinity, available in Galaxy). These tools align the RNA-seq reads to each other based on overlapping segments. This generates a full RNA molecule.

Other tools like Cufflinks (available in Galaxy), assembles the transcripts, and can perform differential expression analysis and/or study the regulation in RNA-seq sample data. The accepted\_hits.bam file from TopHat is provided as input to Cufflinks. The output is a GTF file with gene locations and names. There are two additional outputs – a table with gene expression and another with transcript expression levels.

**Cuffcompare** is used to compare two or more GTF files from Cufflinks. This allows one to compare the gene expression under various conditions like cell types, developmental stages etc.

**Cuffdiff** performs statistical analysis of differential expression between two datasets. It can also be used to perform differential expression of transcripts, exon expression, differential usage of transcription start sites.

Cuffcompare and Cuffdiff are part of the Cufflinks tools available in Galaxy.

**DESeq2** – is a tool that determines differentially expressed features from count tables (obtained using featureCounts).

A Galaxy tutorial performing de novo transcriptome reconstruction with RNA-Seq is available [here](#). The steps involved in the protocol are listed below:

➤ **Data upload.**

file type: fastqsanger, genome:mm10

[https://zenodo.org/record/583140/files/G1E\\_rep1\\_forward\\_read\\_%28SRR549355\\_1%29](https://zenodo.org/record/583140/files/G1E_rep1_forward_read_%28SRR549355_1%29)

[https://zenodo.org/record/583140/files/G1E\\_rep1\\_reverse\\_read\\_%28SRR549355\\_2%29](https://zenodo.org/record/583140/files/G1E_rep1_reverse_read_%28SRR549355_2%29)

[https://zenodo.org/record/583140/files/G1E\\_rep2\\_forward\\_read\\_%28SRR549356\\_1%29](https://zenodo.org/record/583140/files/G1E_rep2_forward_read_%28SRR549356_1%29)

[https://zenodo.org/record/583140/files/G1E\\_rep2\\_reverse\\_read\\_%28SRR549356\\_2%29](https://zenodo.org/record/583140/files/G1E_rep2_reverse_read_%28SRR549356_2%29)

[https://zenodo.org/record/583140/files/Megakaryocyte\\_rep1\\_forward\\_read\\_%28SRR549357\\_1%29](https://zenodo.org/record/583140/files/Megakaryocyte_rep1_forward_read_%28SRR549357_1%29)

[https://zenodo.org/record/583140/files/Megakaryocyte\\_rep1\\_reverse\\_read\\_%28SRR549357\\_2%29](https://zenodo.org/record/583140/files/Megakaryocyte_rep1_reverse_read_%28SRR549357_2%29)

[https://zenodo.org/record/583140/files/Megakaryocyte\\_rep2\\_forward\\_read\\_%28SRR549358\\_1%29](https://zenodo.org/record/583140/files/Megakaryocyte_rep2_forward_read_%28SRR549358_1%29)

[https://zenodo.org/record/583140/files/Megakaryocyte\\_rep2\\_reverse\\_read\\_%28SRR549358\\_2%29](https://zenodo.org/record/583140/files/Megakaryocyte_rep2_reverse_read_%28SRR549358_2%29)

file type: gtf, genome:mm10

[https://zenodo.org/record/583140/files/RefSeq\\_reference\\_GTF\\_%28DSv2%29](https://zenodo.org/record/583140/files/RefSeq_reference_GTF_%28DSv2%29)

➤ Run FastQC on the forward and reverse read files to assess the quality of the reads.

➤ Run Trimmomatic to trim off low quality bases from the ends of the reads.

- “Single-end or paired-end reads?”: Paired-end (two separate input files)

- “Input FASTQ file (R1/first of pair)”: G1E\_rep1 forward read

- “Input FASTQ file (R2/second of pair)”: G1E\_rep1 reverse read

- “Perform initial ILLUMINACLIP step?”: No

➤ Re-run FASTQC on trimmed ends to inspect the differences.

**Mapping** – performed to understand the positions of the reads within the genome.

➤ Run HISAT2 on one forward/reverse read pair with the following settings:

- “Source for the reference genome”: Use a built-in genome

- “Select a reference genome”: Mouse (*Mus Musculus*): mm10
  - “Single-end or paired-end reads?”: Paired-end
    - “FASTA/Q file #1”: Trimmomatic on G1E\_rep1 forward read (R1 paired)
    - “FASTA/Q file #2”: Trimmomatic on G1E\_rep1 reverse read (R2 paired)
    - “Specify strand information”: Forward(FR)
  - “Advanced options”
    - “Spliced alignment options”
      - “Penalty for non-canonical splice sites”: ‘3’
      - “Penalty function for long introns with canonical splice sites”: ‘Constant [f(x) = B]’
      - “Constant term (B)": '0.0'
      - “Penalty function for long introns with non-canonical splice sites”: ‘Constant [f(x) = B]’
      - “Transcriptome assembly reporting”: ‘Report alignments tailored for transcript assemblers including StringTie’
- Run HISAT2 on the remaining forward and reverse read pairs with the same parameters.

#### **De novo transcriptome reconstruction:**

- Run Stringtie on the HISAT2 alignments using the default parameters.
  - Use batch mode to run all four samples from one tool form.
  - “Specify strand information”: Forward (FR)

#### **Transcriptome assembly:**

- Run Stringtie-merge on the Stringtie assembled transcripts along with the RefSeq annotation file.
  - “Transcripts”: all four Stringtie assemblies.
  - “Reference annotation to include in the merging”: RefSeq\_reference\_GTF
- Run GFFCompare on the Stringtie-merge generated transcriptome along with the RefSeq annotation file.
  - “GTF inputs for comparison”: output of Stringtie-merge
  - “Use Reference Annotation”: Yes
    - “Choose the source for the reference annotation”: History
    - “Reference Annotation”: RefSeq\_reference\_GTF
  - “Use Sequence Data”: Yes
    - “Choose the source for the reference list”: Locally cached
    - “Using reference genome”: ‘Mouse (*Mus Musculus*): mm10’

## Differential gene expression testing:

- Run DESeq2 with the following parameters:
  - “1: Factor”
    - “1: Factor level”: G1E
      - “Counts file(s)": featureCount files corresponding to the two G1E replicates
    - “2: Factor level": Mega
      - “Counts file(s)": featureCount files corresponding to the two Mega replicates

Along with the list of genes, DESEQ2 produces a graphical summary file that is useful to determine the quality of the experiment.

- **Visualization:** use the built-in genome browser in Galaxy, Trackster, or UCSC genome browser or IGV to view the BAM alignment file, Stringtie output files, GFFCompare file and the reference GTF file. The known and novel transcripts from G1E and Megakaryocytes can be viewed in the genome browser.

Below is the screenshot with the outputs from the above steps loaded in to IGV:



# Genomic Databases and Genome browsers

## EcoCyc – database for bacterium *Escherichia coli* K-12 MG1655

The screenshot shows the main page of the EcoCyc database. At the top, there's a navigation bar with links for Tools, Sites, Pathway Tools, Help, Subscribe to BioCyc, LOGIN, and Create Free Account. Below the header is the EcoCyc logo and a search bar. A banner at the top says "Change Current Database" and "Current Database: Escherichia coli K-12 substr MG1655 reference genome (EcoCyc)". The search bar contains the placeholder "Enter a gene, protein, metabolite or pathway". Below the search bar are links for "Genome Browser" and "Metabolic Map". A "Comparative Genome Analysis" section follows, featuring a grid of small genome maps for various *E. coli* strains. To the right, a large section is titled "EcoCyc *E. coli* Database", which includes a brief description of the database, a "Learn More" button, and a "Try Free" button. Below this are sections for "What people are saying" with quotes from researchers like Paul Babcock and Patricia Kew, and a "Regulation Diagrams" section.

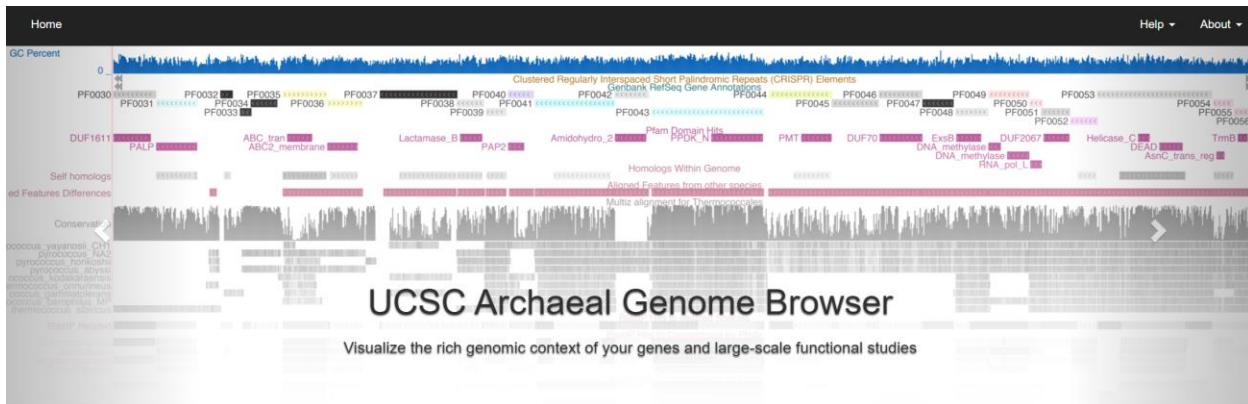
RegulonDB – is the primary database containing transcriptional regulation data in *Escherichia coli* k-12 manually curated from scientific publications.

The screenshot shows the main page of the RegulonDB database. At the top, there's a navigation bar with links for Home, Features, Integrated Views & Tools, Downloads, Doc & Help, and a search bar. Below the header is a large search box with placeholder text "Search in RegulonDB" and a "Search" button. To the right of the search box is a box for "Escherichia coli K-12 Transcriptional Regulatory Network" with a "Read more" link. The main content area is divided into several sections: "Downloads" (with links to "Experimental Datasets" and "Computational Predictions Datasets"), "RegulonDB Full Version" (with a link to download the database), and "RegulonDB Features" (with a bulleted list of features). At the bottom, there are logos for CCG, Centro de Ciencias Genómicas, and CONACYT, along with links for "How to cite", "Terms & conditions", "Funding", and "Contact us".

**PORTEco** – is a portal for *E. coli* research that enables viewing and downloading of datasets.

**PATRIC** (Pathosystems Resource Integration Center) – provides integrated data analysis tools that support biomedical research on infectious diseases caused by bacteria.

**UCSC archaeal genome browser** – a separate genome browser for archaea and certain bacterial genomes. It is linked to Galaxy that allows performing some genome analysis operations.



#### About the browser

The UCSC Archaeal Genome Browser is a window on the biology of more than 100 microbial species from the domain Archaea. Basic gene annotation is derived from NCBI Genbank/RefSeq entries, with overlays of sequence conservation across multiple species, nucleotide and protein motifs, non-coding RNA predictions, operon predictions, and other types of bioinformatic analyses. In addition, we display available gene expression data (microarray or high-throughput RNA sequencing). Direct contributions or notices of publication of functional genomic data or bioinformatic analyses from archaeal research labs are very welcome.

To get started, access the genome browsers using your favorite entry point:

1. Enter a partial genome name in Quick Search on this page, or
2. Search at the Genome Gateway pages

Quick Search

(Examples: "Pyro", "furi")

The Lowe Lab • Biomolecular Engineering • University of California Santa Cruz

**Greengenes** – a repository for 16S rRNA sequences and tools. These sequences are used in phylogeny studies to compare organisms and for species identification.

#### Index of /Download/

...	
FAQs/	
Microarray_Data/	
OTUs/	
Podcasts/	
Posters/	
Presentations/	
Protocols/	
Publications/	
Sequence_Data/	
Software/	
Taxonomic_Outlines/	
Tests/	
Trees/	
Tutorial/	
Crystal_Clear.tar	
crystal.tar	

**NCBI Genome Data Viewer (GDV)** – previously known as Map Viewer. GDV is a genome browser to view eukaryotic RefSeq genes.

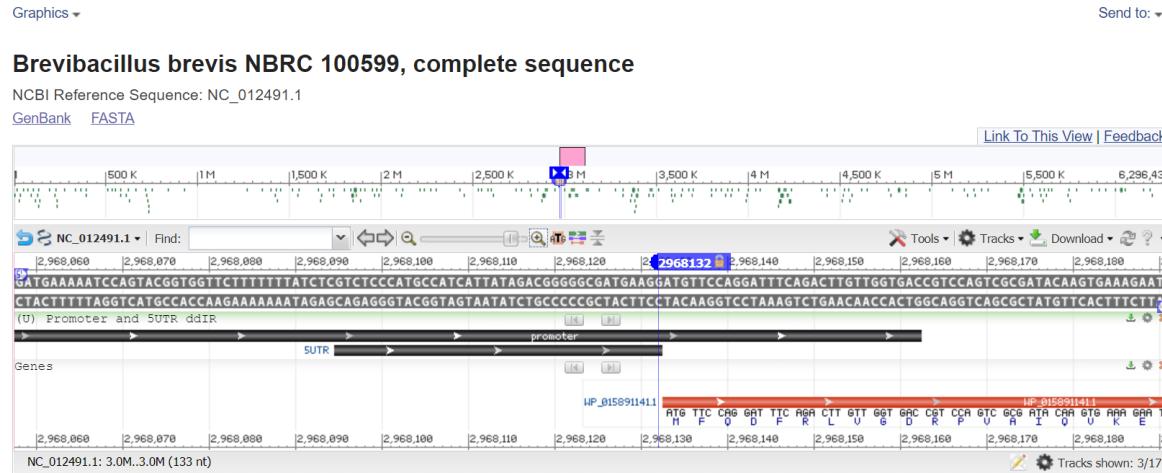
The screenshot shows the NCBI Genome Data Viewer interface. At the top, there's a navigation bar with the NIH logo, U.S. National Library of Medicine, NCBI, and a Log in button. Below the navigation bar is the title "Genome Data Viewer". A search bar labeled "Search organisms" contains "Homo sapiens (human)". A message box says: "To view more organisms in the tree, click on nodes that have '+' signs. Press and hold the '+' to expand and reveal all the subgroups. Or, search for an organism using the search box above." Another message box says: "New! Click on Switch view at the top to see another way of navigating genomes." To the right, there's a detailed view of the "Homo sapiens (human)" assembly. It includes a search bar for "Location, gene or phenotype", a dropdown for "Assembly" set to "GRCh38.p13", and buttons for "Browse genome", "BLAST genome", and "Download via NCBI Datasets". Below this are sections for "Assembly details" (including name, RefSeq accession, GenBank accession, submitter, level, category, and replaced by), "Annotation details" (annotation release 109, release date May 15, 2021), and a grid of chromosomes. A "Feedback" button is at the bottom right.

**Variation Viewer:** The NCBI variation viewer mostly focusses on human genome. Specific genes can be loaded on to the viewer. The many filters on the left panel helps the user choose the variation type to be displayed in the viewer. There is also an option to load user tracks with BED or WIG files. The NCBI genome browser can be used to view the contents of the BED files.

The screenshot shows the NCBI Variation Viewer interface. At the top, there's a navigation bar with the NIH logo, National Library of Medicine, National Center for Biotechnology Information, and a Log in button. A "COVID-19 Information" section is present. Below the navigation bar is the title "Variation Viewer". A message box says: "New to Variation Viewer? Read our quick overview! X". On the left, there's a sidebar with "Variation Viewer" and a "Variation Data" section. The main area shows a genomic track for "Chr 1 (NC\_000001.11)". The track displays various genomic features and variants. A message box in the center says: "Exon Navigator: There are too many (5089) genes in the region. Please narrow the region to enable exon navigation." The bottom of the screen shows a "Variation Data" table with columns like "Filter by", "Download", "edit columns", and "Items 1 - 20 of 41,882,009 <- First <- Prev Page 1 of 2,094,101 Next > Last >>".

## NCBI genome:

The [Genome](#) is a repository of information on genomes. It includes sequences, maps, annotations, assemblies, and chromosomes. We can browse the database by organisms and can even download the sequences of our interest. We can also upload custom data and view them as tracks.



[Ensembl genome browser](#): It is a genome browser primarily for sequenced genomes of chordates. It includes few genomes of non-chordate model organisms, however, not prokaryotes. The database is updated about every four months. At the time of preparation of this document (August 2021), the current release is Ensembl release 104. The database is free to use for all, and we can create login accounts. This would allow us to share details of our research/analysis with other members of the team. The homepage allows us to search species-specific information.

There are many videos on the usage of various Ensembl tool. The link can be found [here](#).

[Ensembl Genomes](#) is the database for non-vertebrates. It provides information on genomes of bacteria, protists, fungi, plants, and invertebrate metazoan. Using the bacterial portal, archaeal genomes can be accessed.

**Ensembl** BLAST/BLAT VEP Tools BioMart Downloads Help & Docs Blog Login/Register

Search all species

Tools BioMart > BLAST/BLAT > Variant Effect Predictor >

All tools Export custom datasets from Ensembl with this data-mining tool Search our genomes for your DNA or protein sequence Analyse your own variants and predict the functional consequences of known and unknown variants

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotates genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species

Ensembl Release 104 (May 2021)

- Update to the Ensembl Canonical transcript set.
- Human and mouse gene sets updated to GENCODE 38 and GENCODE M27, respectively.
- Retirement of gene names derived from BAC clones.

More release news  on our blog

Ensembl Rapid Release

New assemblies with gene and protein annotation every two weeks. Note: species that already exist on this site will continue to be updated with the full range of annotations.

The Ensembl Rapid Release website provides annotation for recently produced, publicly available vertebrate and non-vertebrate genomes from biodiversity initiatives such as Darwin Tree of Life, the Vertebrate Genomes Project and the Earth BioGenome Project.

Rapid Release news  on our blog

Other news from our blog

- 14 Jul 2021 Job Eukaryotic Annotation Team Leader
- 01 Jun 2021 Update Temporary disruption to Ensembl Tools and Archive services is now resolved
- 26 May 2021 Ensembl's efforts towards a greener future

Compare genes across species Find SNPs and other variants for my gene Gene expression in different tissues Retrieve gene sequence Find a Data Display Use my own data in Ensembl

EMBL-EBI Ensembl creates, integrates and distributes reference datasets and analysis tools that enable genomics. We are based at EMBL-EBI  and our software and data are freely available. Our acknowledgements page includes a list of current and previous funding bodies. How to cite Ensembl in your own publications.

Permanent link · View in archive site

About Us Get help Our sister sites Follow us

Searching for the Human TP53 in the homepage, leads us to the gene-specific page which provides information of the gene location on the chromosome, transcripts, orthologues, paralogues etc. On the left navigation panel, there are tools specific for gene data analysis. The region in detail view will allow us to view a region on the chromosome. We can see the coding and the non-coding transcripts, exons etc.

Clicking on a specific gene transcript, opens a new transcript tab that provides detailed information pertaining to the selected transcript. The exon data, variant, and protein information can be viewed by selecting the appropriate links from the left navigation panel. Below are the screenshots of the gene tab and the transcript tab from the Ensembl genome browser.

## Gene tab:

**Gene: TP53 ENSG00000141510**

Description: tumor protein p53 [Source:HGNC Symbol;Acc:HGNC:11998]

Gene Synonyms: LFS1, p53

Location: Chromosome 17: 7,661,779-7,687,538 reverse strand.

About this gene: This gene has 27 transcripts (splice variants), 253 orthologues, 2 paralogues and is associated with 398 phenotypes.

Transcripts: Show transcript table

**Summary**

Name: TP53 (HGNC Symbol)

CCDS: This gene is a member of the Human CCDS set: CCDS11118.1, CCDS45605.1, CCDS45606.1, CCDS73963.1, CCDS73964.1, CCDS73965.1, CCDS73966.1, CCDS73967.1, CCDS73968.1, CCDS73969.1, CCDS73970.1, CCDS73971.1.

UniProtKB: This gene has proteins that correspond to the following UniProtKB identifiers: P04637.

RefSeq: This Ensembl/Gencode gene contains transcript(s) for which we have selected identical RefSeq transcript(s). If there are other RefSeq transcripts available they will be in the External references table.

LRG: LRG\_321 provides a stable genomic reference framework for describing sequence variants for this gene

Ensembl version: ENSG00000141510.18

Other assemblies: This gene maps to 7,565,097-7,590,856 in GRCh37 coordinates. View this locus in the GRCh37 archive: ENSG00000141510.

Gene type: Protein coding

Annotation method: Annotation for this gene includes both automatic annotation from Ensembl and Havana manual curation, see article.

Annotation Attributes: overlapping locus [Definitions]

Go to Region in Detail for more tracks and navigation options (e.g. zooming)

Genomic predictions: Genscan predictions for GENSCAN00000035445 > GENSCAN00000035450 > Genscan predictions

## Transcript tab:

**Transcript: TP53-201 ENST00000269305.9**

Description: tumor protein p53 [Source:HGNC Symbol;Acc:HGNC:11998]

Gene Synonyms: LFS1, p53

Location: Chromosome 17: 7,668,421-7,687,490 reverse strand.

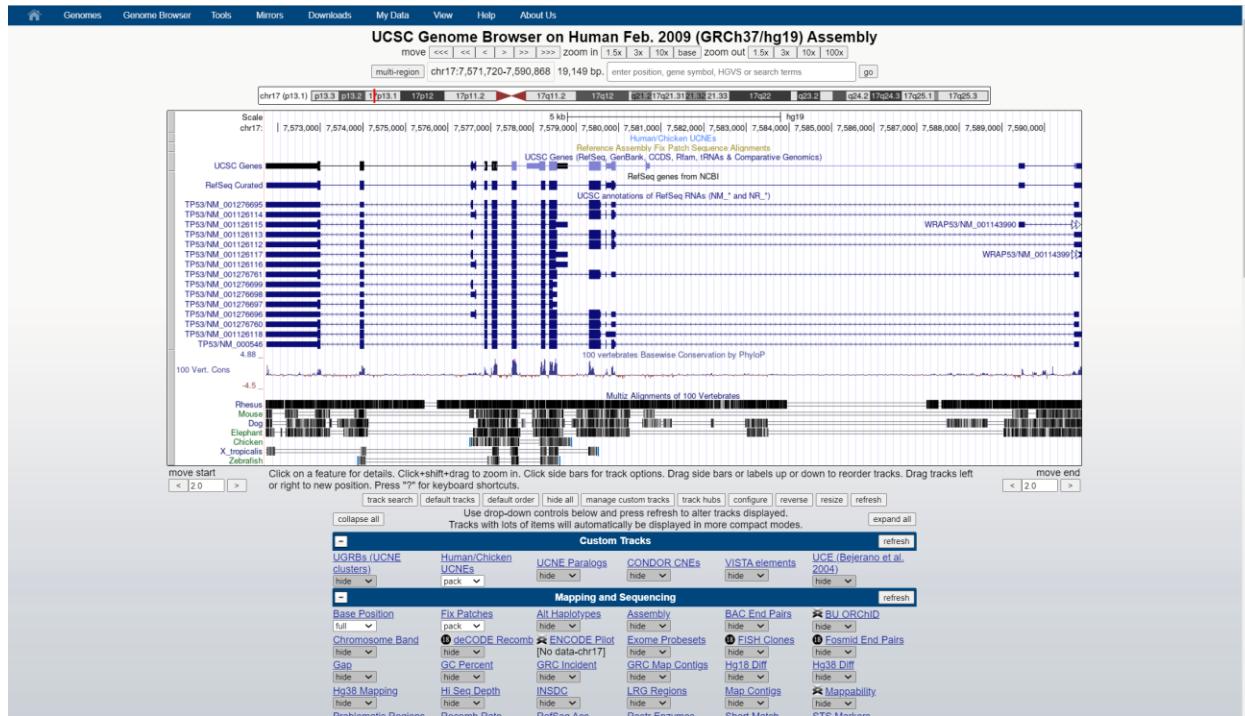
About this transcript: This transcript has 11 exons, is annotated with 404 domains and features, is associated with 11918 variant alleles and maps to 682 oligo probes.

Gene: This transcript is a product of gene ENSG00000141510.18 Hide transcript table

Show/hide columns (1 hidden)

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt Match	RefSeq Match	Flags
TP53-201	ENST00000269305.9	2512	393aa	Protein coding	CCDS11118	P04637-1	NM_000546.6	MANE Select v0.93 Ensembl Canonical GENCODE basic APPRI P1 TSL
TP53-204	ENST00000420246.6	2653	341aa	Protein coding	CCDS45606	P04637-2	-	GENCODE basic TSL1
TP53-226	ENST0000022645.4	2653	302aa	Protein coding	CCDS73971	P04637-5	-	GENCODE basic TSL1
TP53-219	ENST00000610292.4	2639	354aa	Protein coding	CCDS73969	P04637-3	-	GENCODE basic TSL1
TP53-206	ENST00000455263.6	2580	346aa	Protein coding	CCDS45605	P04637-3	-	GENCODE basic TSL1
TP53-220	ENST00000610538.4	2580	307aa	Protein coding	CCDS73970	P04637-6	-	GENCODE basic TSL1
TP53-225	ENST00000620730.4	2579	354aa	Protein coding	CCDS73969	P04637-4	-	GENCODE basic TSL1
TP53-205	ENST00000445886.6	2506	393aa	Protein coding	CCDS11118	P04637-1	-	GENCODE basic APPRI P1 TSL1
TP53-224	ENST00000619485.4	2506	354aa	Protein coding	CCDS73969	P04637-4	-	GENCODE basic TSL1
TP53-213	ENST00000510385.5	2404	209aa	Protein coding	CCDS73988	P04637-8	-	GENCODE basic TSL1
TP53-222	ENST00000618944.4	2404	182aa	Protein coding	CCDS73965	A0A087WXZ1	-	GENCODE basic TSL1
TP53-208	ENST00000504290.5	2331	214aa	Protein coding	CCDS73967	P04637-9	-	GENCODE basic TSL1
TP53-221	ENST00000610623.4	2331	187aa	Protein coding	CCDS73964	A0A087WTZ2	-	GENCODE basic TSL1
TP53-209	ENST00000504937.5	2271	281aa	Protein coding	CCDS73966	P04637-7	-	GENCODE basic TSL1
TP53-223	ENST00000619186.4	2271	234aa	Protein coding	CCDS73963	A0A087X1Q1	-	GENCODE basic TSL1
TP53-202	ENST00000359597.8	1152	343aa	Protein coding	-	J3KP33	-	GENCODE basic TSL1
TP53-203	ENST00000413465.6	1018	285aa	Protein coding	-	E7EOXT	-	GENCODE basic TSL1
TP53-212	ENST00000509690.5	729	199aa	Protein coding	-	E7ESS1	-	TSL4 CDS 3' incomplete
TP53-211	ENST00000508793.5	634	185aa	Protein coding	-	E7EMR6	-	TSL4 CDS 5' incomplete
TP53-218	ENST00000604346.5	568	143aa	Protein coding	-	SAR334	-	TSL4 CDS 3' incomplete

**UCSC Genome Browser:** It is a genome browser developed by the University of Santa Cruz and provides options to view genomic data by configuring specific tracks. It is also possible to upload custom data and create custom tracks to view alongside the existing data for comparison and analysis. It is primarily a eukaryotic genome browser, with 97 eukaryotic genomes, including 49 mammals. The database holds assembly data, mRNA, EST, RefSeq alignments. It provides links to other applications like Ensembl, NCBI Map viewer, Galaxy etc. Data from ENCODE project is also available to be viewed.

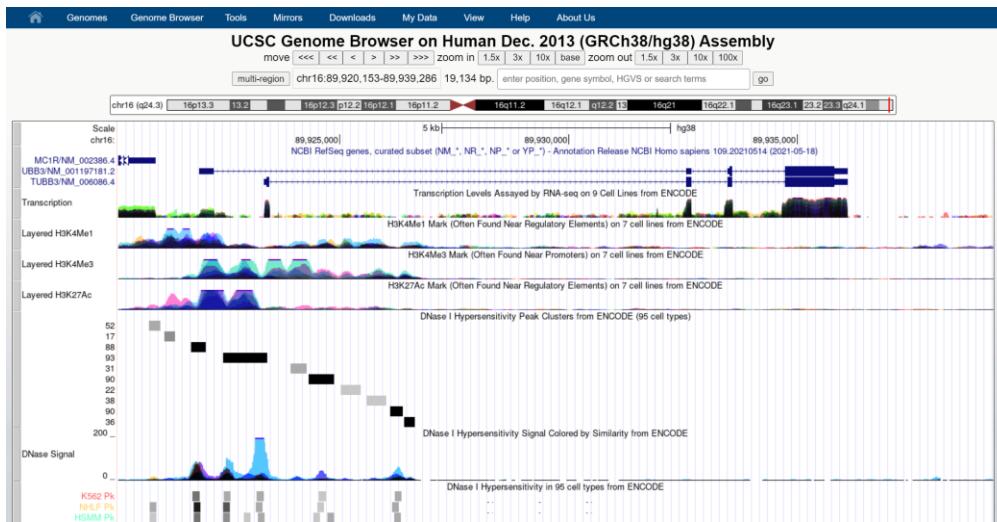


**Example:** Find the human TUBB3 gene using the UCSC Genome Browser (hg38). Turn on the Encode Regulation track (HINT: set display mode to “full” for these tracks) and NCBI RefSeq genes. In a few sentences, describe what you see at the TUBB3 locus in terms of the Encode Regulation tracks. Include in your answer what histone modification(s) appear(s) near the transcription start site of the TUBB3 gene. Submit a screenshot of this locus. (HINT: click View > PDF/EPS at the top of the browser page to export a PDF/EPS file.).

At the TUBB3 locus in the Encode Regulation track, we can see the following tracks- Transcription Levels Assayed by RNA-seq on 9 Cell Lines from ENCODE, Layered H3K4me1 track, Layered H3K27Ac track, DNase 1 hypersensitivity peak clusters track, DNase 1 hypersensitivity signal track, track with DNase 1 hypersensitivity in 95 cell types from ENCODE, TF clusters track and Transcription Factor ChIP-seq Peaks track. H3K4me1, H3k4me3 and H3k27Ac modifications appear near the TSS of TUBB3 gene. These are represented by peaks in the track that are observed close to the TSS or in the region of the

promoter of TUBB3 gene. The Layered H3K27Ac track show the modification of the histone proteins and is suggestive of enhancers. The peaks represent the histone modifications in HUVEC, NHEK, NHLF, K562 etc. cells from ENCODE.

Screen shot of the TUBB3 locus from UCSC genome browser:



Using the Table Browser, we can download data from UCSC. There is option to download data in many formats like BED or send the results directly to Galaxy. The user guide for Table Browser is available [here](#).

Another feature available at the UCSC Genome Browser is BLAT. This topic was covered in the previous section. The interface is shown below and can accessed by clicking on “Tools” from the top menu -> Blat

**dbSNP database:** The largest database of SNPs is hosted by NCBI. It holds submitted SNPs and curated reference SNPs. It also includes MNPs (multiple nucleotide polymorphisms) and short indels (insertions and deletions).

Web interface of dbSNP database:

The screenshot shows the dbSNP homepage. At the top, there's a navigation bar with links for NCBI, Resources, How To, Sign in to NCBI, and Help. Below the navigation is a search bar with "SNP" selected and an Advanced link. A "COVID-19 Information" banner is prominently displayed, featuring links to Public health information (CDC), Research information (NIH), SARS-CoV-2 data (NCBI), Prevention and treatment information (HHS), and Español. The main content area has a dark header "dbSNP". Below it, a text box states: "dbSNP contains human single nucleotide variations, microsatellites, and small-scale insertions and deletions along with publication, population frequency, molecular consequence, and genomic and RefSeq mapping information for both common variations and clinical mutations." On the left, there's a "Getting Started" sidebar with links to dbSNP 20th Anniversary, Overview of dbSNP, About Reference SNP (rs), Factsheet, and Entrez Updates (May 26, 2020). The center has sections for Submission (How to Submit, Hold Until Published (HUP) Policies, Submission Search) and Access Data (Web Search, eUtils API, Variation Services, FTP Download, Tutorials on GitHub). At the bottom, a callout box announces "ALFA Project Release 2 with over 900M variants from 192K subjects is now available (January 6, 2021)". It includes a video thumbnail for the ALFA CoLab Presentation at ASHG2020, with options to Watch later or Share.

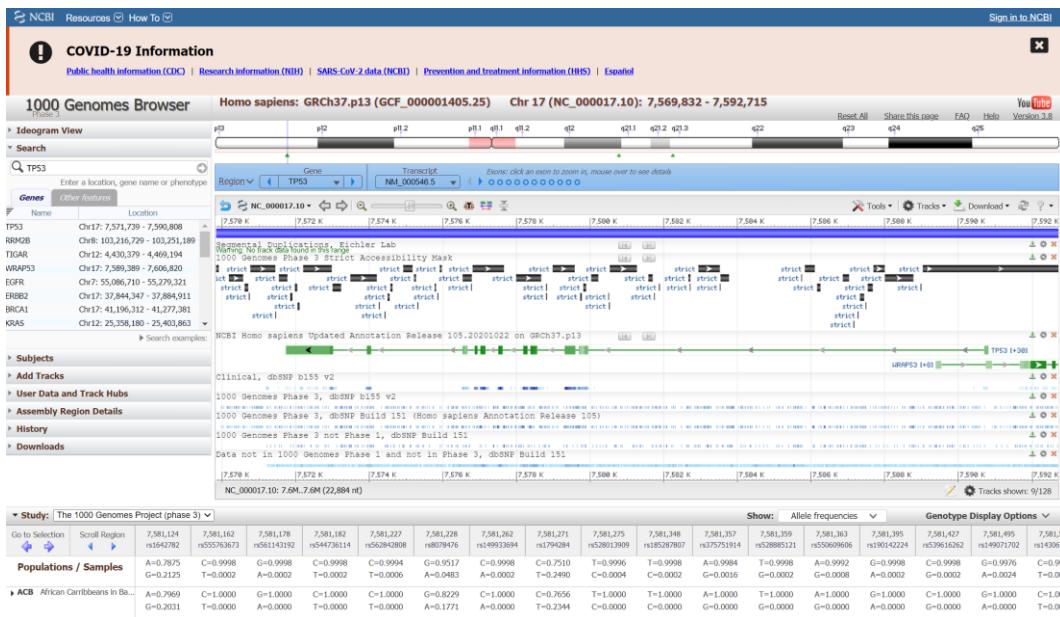
**SNPeffect database:** came into being in 2012, stores information pertaining to the effect of SNPs on proteins and their function. There are tools to predict protein stability and folding.

The screenshot shows the SNPeffect database homepage. At the top, there's a navigation bar with links for ABOUT, HELP, DATABASE (which is highlighted), META-ANALYSIS, and CONTACT. Below the navigation is a search bar labeled "Search and browse". Underneath the search bar, a breadcrumb navigation shows "Home > All mutations >". On the right side, there's a "Filter SNPeffect database" sidebar with various dropdown menus and input fields for filtering mutations based on Disease, Mutation Type, UniProt ID, Gene Name, dbSNP, Sequences (DIF\_TANGO, DIF\_WALTZ, DIF\_LIMBO), and FoldX: Free energy change. The main content area displays a table titled "63410 mutations listed" with columns for Variant, UniProt ID, Mutation, Disease, Mutation Type, dTANGO, dWALTZ, dLIMBO, and ddG. The table lists numerous rows of mutation data, such as VAR\_064762 (UniProt ID 1433B\_HUMAN, Mutation V99I, Disease Unclassified, dTANGO 0, dWALTZ -1, dLIMBO 0, ddG 0.55).

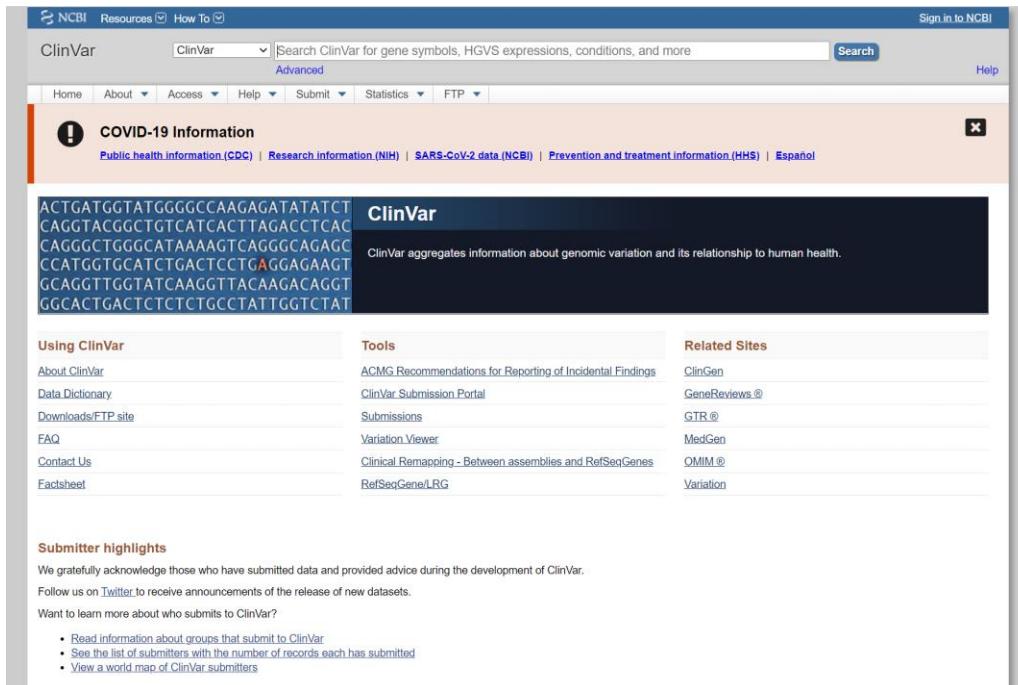
Variant	UniProt ID	Mutation	Disease	Mutation Type	dTANGO	dWALTZ	dLIMBO	ddG
VAR_064762	1433B_HUMAN	V99I	-	Unclassified	0	-1	0	0.55
VAR_048095	1433S_HUMAN	M155I	-	Polymorphism	0	37	0	0.52
VAR_056253	1A01_HUMAN	R205H	-	Polymorphism	-1	0	0	1.35
VAR_056250	1A01_HUMAN	N151K	-	Polymorphism	0	0	0	0.24
VAR_056251	1A01_HUMAN	I166T	-	Polymorphism	0	0	0	0.96
VAR_056252	1A01_HUMAN	R169H	-	Polymorphism	-1	0	0	0.31
VAR_016725	1A01_HUMAN	V182A	-	Polymorphism	0	0	0	-0.31
VAR_056247	1A01_HUMAN	R89G	-	Polymorphism	-1	0	0	1.57
VAR_056248	1A01_HUMAN	G131W	-	Polymorphism	0	0	9	3.3
VAR_056249	1A01_HUMAN	F133L	-	Polymorphism	0	0	0	1.23
VAR_016723	1A01_HUMAN	I121M	-	Polymorphism	-11	-1	3	0.79
VAR_016724	1A01_HUMAN	R180L	-	Polymorphism	0	0	1	-1.91
VAR_016721	1A01_HUMAN	A100E	-	Polymorphism	0	0	0	-0.04
VAR_004333	1A01_HUMAN	R41S	-	Polymorphism	-1	0	0	1.04
VAR_016720	1A01_HUMAN	M91V	-	Polymorphism	0	0	0	4.3
VAR_016719	1A01_HUMAN	G80R	-	Polymorphism	1	0	0	0.47
VAR_004332	1A01_HUMAN	F33S	-	Polymorphism	-34	-6	0	6.14
VAR_016722	1A01_HUMAN	D114A	-	Polymorphism	0	0	0	0.42
VAR_016730	1A02_HUMAN	W191G	-	Polymorphism	0	-1	8	1.99
VAR_004350	1A02_HUMAN	A260E	-	Polymorphism	0	0	0	1.31
VAR_004349	1A02_HUMAN	T187E	-	Polymorphism	0	-1	0	-1.03
VAR_016728	1A02_HUMAN	H98D	-	Polymorphism	0	0	0	1.96

**NCBI 1000 Genomes Browser:** This allows access to the data from the 1000 Genomes Project. We can analyze specific genes /regions/ chromosome positions, add ClinVar and dbSNP tracks by clicking on the 'Tracks' button. For each variation in the variation table, genotype information can be accessed pertaining to specific ethnicities. Custom data can also be uploaded. The help page for the browser can be found [here](#).

Below is a screen shot of the interface for the TP53 gene:



**ClinVar:** is a specialized database to hold a curated set of variations that have possible medical and clinical relevance. It focusses on human variations. It does not replace dbSNP. We can search the database using the gene symbol or disease. It is also available through UCSC Genome Browser under the group ‘Phenotype and Literature’ -> ‘ClinVar Variants’ and ‘ClinGen CNVs’ tracks.

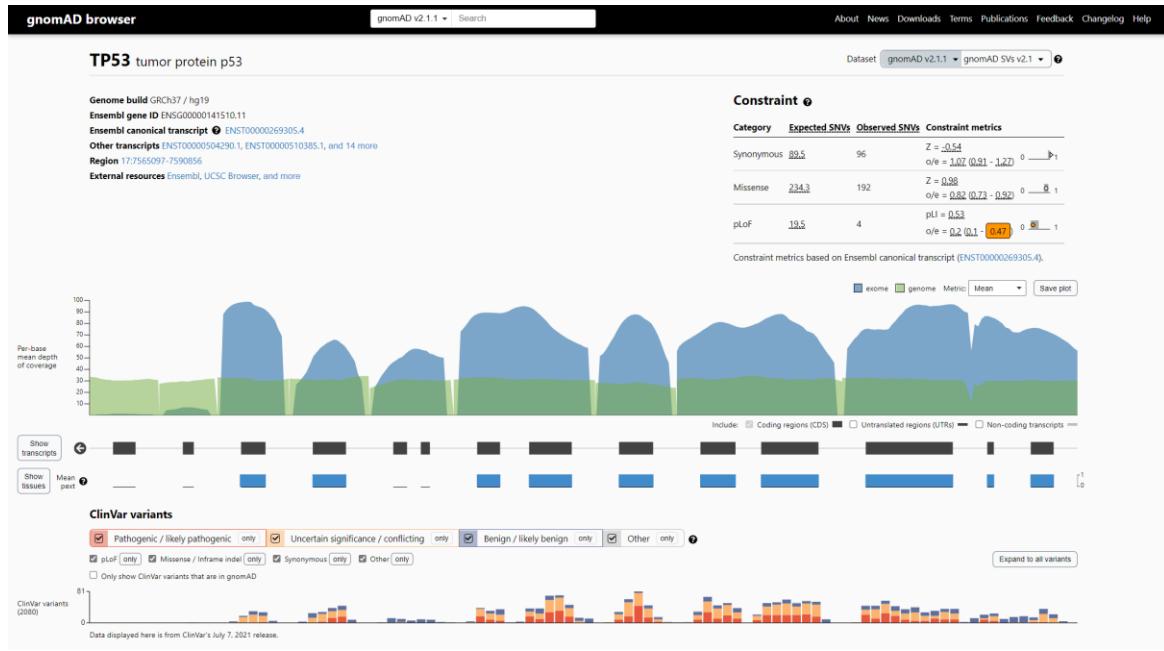


**DGV (Database of Genomic Variants):** is hosted by the Center for Applied Genomics, Toronto. By default, the hg19 assembly is mapped. In the UCSC Genome Browser, the same information can be accessed using the DGV Struct Var track, which may also be more user-friendly with gain in copy number indicated in blue and losses indicated in red.

DGV interface:

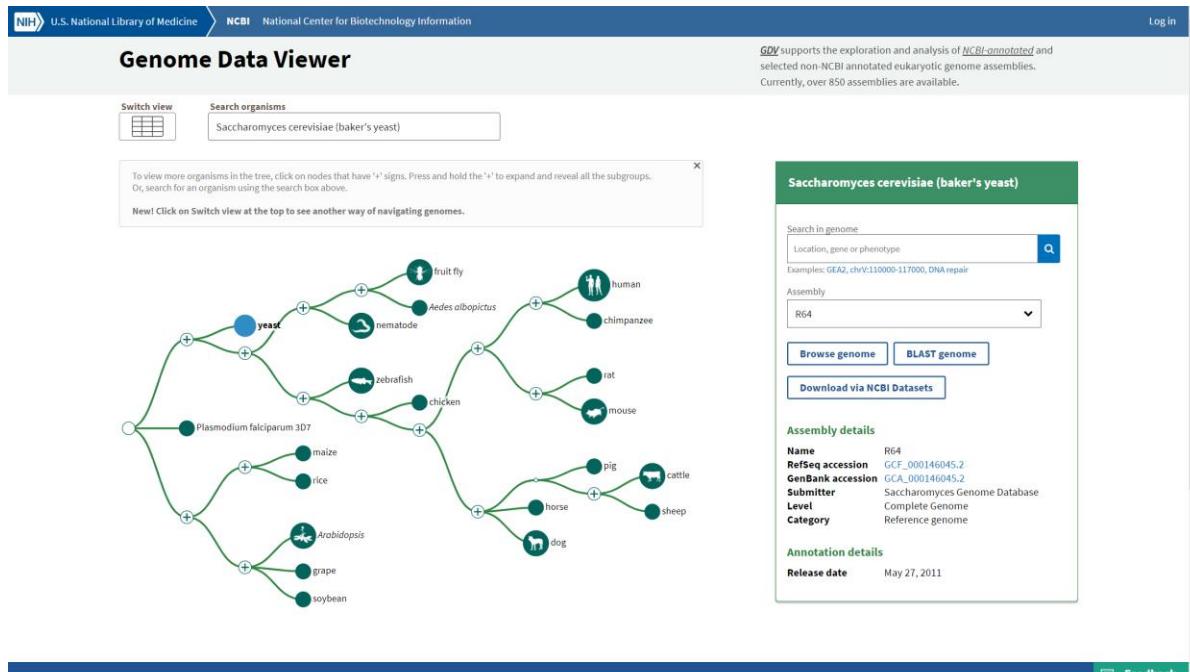
The screenshot shows the homepage of the Database of Genomic Variants (DGV). The header features the logo 'Database of Genomic Variants' and the subtitle 'A curated catalogue of human genomic structural variation'. Below the header is a navigation bar with links: About the Project, Downloads, Links, Statistics, Contact Us, and FAQ. The main search area includes a 'Keyword, Landmark or Region Search:' input field, a 'Search' button, and a dropdown menu set to 'GRCh37/hg19'. Below the search area, there's a section titled 'Find DGV Variants' with links for 'by Study', 'by Sample', 'by Method', 'by Variant', 'by Platform', and 'by Chromosome'. A 'Summary Statistics' section displays merged-level sample-level data: CNVs (983845), Inversions (4083), and Number of Studies (75). At the bottom, there's a news link 'News: February 2020 Update and Newsletter has been issued' and footer links for 'Hosted by The Centre for Applied Genomics', 'Grant support for DGV', and 'Please read the usage disclaimer'.

**GnomAD (Genome Aggregation Database):** It is a repository of exome and genome sequencing data from large-scale sequencing projects. It allows searches to be performed based on gene names, ensemble transcript ID, dbSNP ID, or chromosomal region. Below is the search result for TP53 gene. The peaks indicate protein-coding SNPs in the region. The CNV data below is represented by brown (gains) and red (losses). Below it is the table of SNPs (not shown in the screen shot).

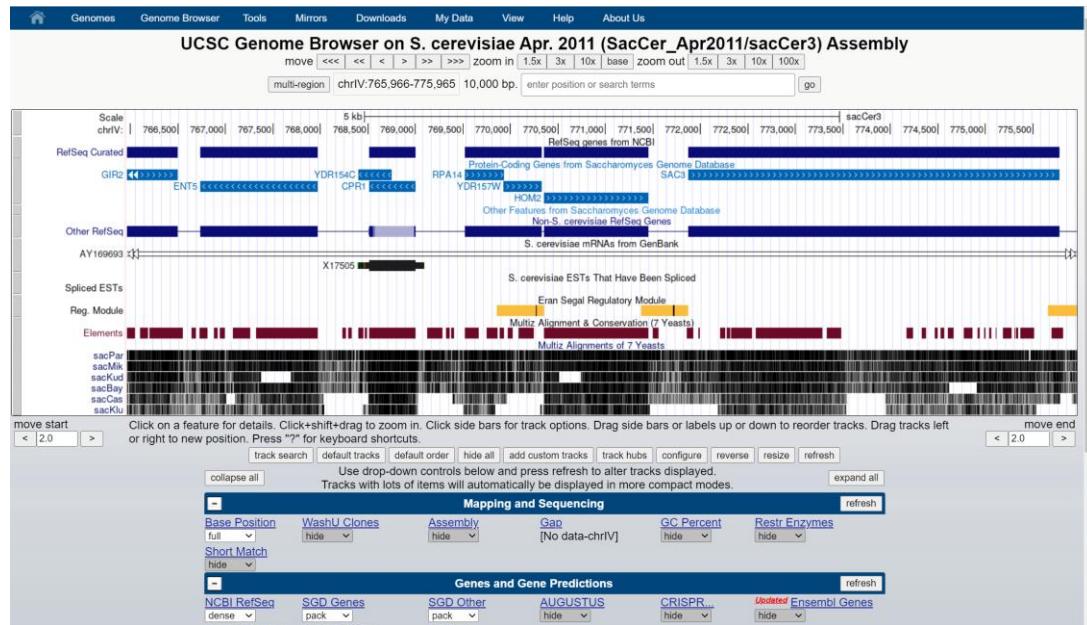


**Yeast Genome Browsers:** There are several browsers available for *Saccharomyces cerevisiae*. They are:

[NCBI Genome Data Browser](#) (NCBI Genomes) – Allows users to search by gene, location, or phenotype. It provides options to browse the genome, perform BLAST, and also download datasets.



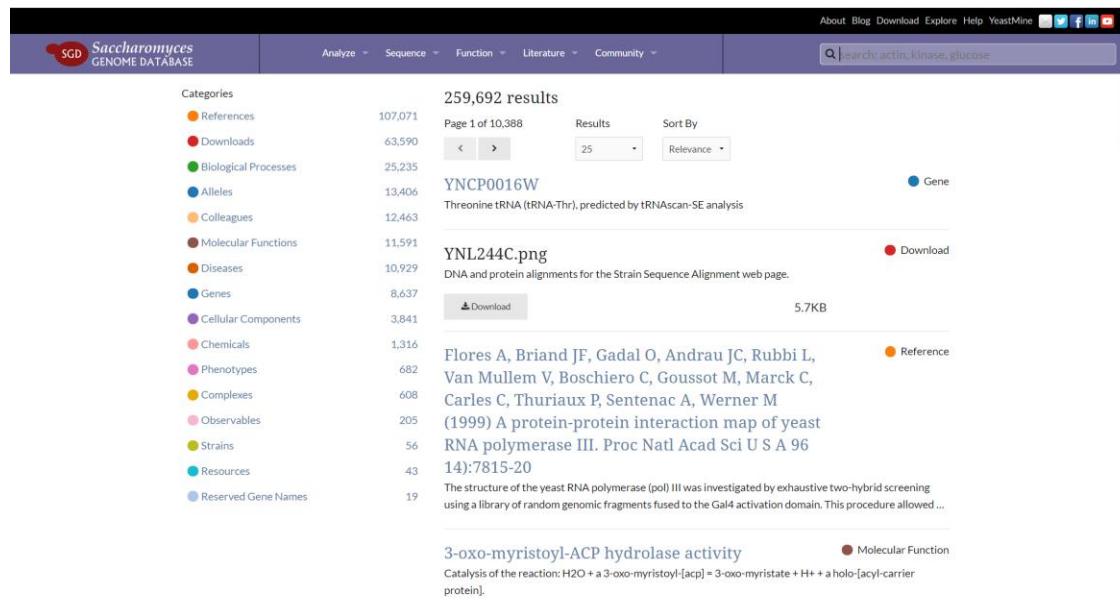
## S. cerevisiae Genome Browser - sacCer3 assembly at UCSC



**SGD (Saccharomyces Genome Database):** provides integrated biological information for budding yeast along with tools for data exploration to study the relationship between sequence and gene products in fungi and higher organisms.

Tutorials to help with the navigation of SGD is available [here](#).

Homepage of SGD:



**ScerTF** : is a database for yeast transcription factors (TFs). This website is useful to find genes regulated by a particular TF, analyzing the regulatory interactions between different TFs.

The screenshot shows the homepage of the ScerTF website. At the top, there is a navigation bar with links for Home, Browse, Download, About, and Data Sources. Below the navigation bar, the title "ScerTF" is displayed in a large, bold font. A descriptive text block follows, stating: "ScerTF catalogs over 1,200 position weight matrices (PWMs) for 196 different yeast transcription factors. We've curated 11 literature sources, benchmarked the published position-specific scoring matrices against in-vivo TF occupancy data and TF deletion experiments, and combined the most accurate models to produce a single collection of the best performing weight matrices for *Saccharomyces cerevisiae*." Another text block below explains the utility of the tool: "ScerTF is useful for a wide range of problems, such as linking regulatory sites with transcription factors, identifying a transcription factor based on a user-input matrix, finding the genes bound/regulated by a particular TF, and finding regulatory interactions between transcription factors." A search input field is present with the placeholder "Enter a TF name to find the recommended matrix for a particular TF, or enter a nucleotide sequence to identify all TFs that could bind a particular region". Below the search field, there are two sections: one for searching by TF name or DNA sequence, and another for pasting a matrix in consensus format. Both sections include a "Load example Sequence" or "Load example Matrix" link. At the bottom of the page, author information and citation details are provided: "Aaron T. Spivak and Gary D. Stormo" and "http://dx.doi.org/10.1093/nar/gkr1180". The citation itself is: "Citation: Spivak AT, Stormo GD (2012) ScerTF: a comprehensive database of benchmarked position weight matrices for *Saccharomyces* species. Nucleic Acids Res 40: D162.168".

**Yeast Gene Order Browser (YGOB):** helps with visualizing the syntenic context of a gene from several yeast genomes.

Welcome to the Yeast Gene Order Browser: YGOB is an online tool for visualising the synteny context of any gene from several yeast genomes. There are detailed user guidelines. Standard bioinformatics tools are available through the browser interface and the icons for each are explained on the help page.

You can start browsing immediately by typing the name of a yeast gene in the control console at the bottom of this page (e.g. ADH1).

The Yeast Gene Order Browser: combining curated homology and synteny context reveals gene fate in polyploid species  
Byrne KP and Wolfe KH  
Genome Research 2005 Oct;15(10):1456-61  
Supplemental: Methods Table 1 Table 2

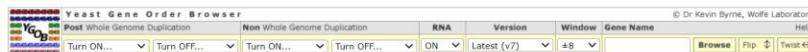
Version 7 (Aug2012) Latest Earlier versions... DATA

Added in August 2012 and including three new genomes (S.cerevisiae, S.mikatae, S.kudriavzevi and S.bayanus var. uvarum) from Scamell et al. G3 (2011) and the finalised versions of the three draft genomes in Version 6. We've updated the S. cerevisiae data in YGOB to the latest release (R64) from SGD. As a visual aid we have also added bars to the top of gene boxes to indicate the relative length of homologs.

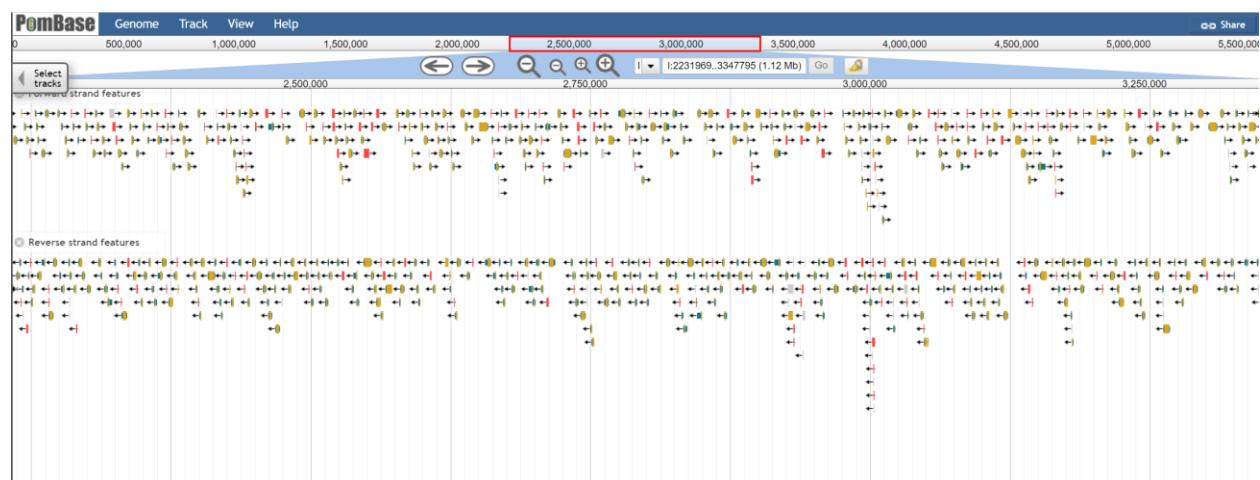
We expect this to be a 'benchmark' version of the browser, suitable for many types of analyses and likely to remain as the current YGOB version for the next few years. Researchers may find the associated datasets listed at the right of use to them.

New Server & Domain Name  
With this version we have also migrated YGOB to a new faster and more powerful server, which renders all aspects of YGOB much quicker than before and facilitates heavy traffic and large numbers of simultaneous users. The browser can now be found at <http://ygb.ued.ie>

© Dr Kevin Byrne - Wolfe Laboratory



**PomBase:** It is a database for the fission yeast *Schizosaccharomyces pombe*, providing access to large data sets, structural and functional annotation information.



**Ensembl Fungi:** is a browser for fungal genomes.

The screenshot shows the Ensembl Fungi homepage. At the top, there's a search bar with "All species" selected and a placeholder "e.g. NAT2 or alcohol". Below the search bar is a "Favourite genomes" section featuring *Saccharomyces cerevisiae* (R64-1-1). To the left, a sidebar lists "All genomes" with a dropdown menu set to "Select a species". A link "View full list of all species" is also present. The main content area has a header "What's New in Release 51" with sections for "Genomes" (1014 genomes total), "Updated data" (protein features, BioMarts, gene trees), and "Ensembl Rapid Release" (new assemblies every two weeks). It also features a "Join the *Zymoseptoria tritici* gene annotation team!" section and an "Archive sites" section with links to previous releases (49, 45, 40, 37). The footer includes a "Login/Register" link and a "Search Ensembl Fungi..." button.

**Aspergillus Genome Database (AspGD):** It has complete genome information of two species - *Aspergillus nidulans* and *Aspergillus fumigatus*. It is maintained by the Broad Institute.

The screenshot shows the Aspergillus Genome Database (AspGD) homepage. The header includes links for "About", "Site Map", "How to Cite", "Help", and social media icons. The main navigation bar has links for "Home", "Search", "GBrowse", "Sequence", "GO", "Tools", "Literature", "Download", and "Community". Below the navigation is a large image of *Aspergillus fumigatus* conidiophores stained with LPCB. A caption below the image reads: "Aspergillus fumigatus - conidiophores stained with LPCB. Courtesy of Yuri Amatnieks, Canada. (<http://thunderhouse4-yuri.blogspot.com/>)". The "About AspGD" section provides an overview of the database's purpose and content. The "Upcoming Meetings & Courses" section lists several events. The "New and Noteworthy" section highlights "Aspergillus annotation updates" from September 2015, mentioning changes in reference gene sets and new features added. The "AspGD Curation News" section lists recent additions like genome snapshots for *A. nidulans*, *A. fumigatus*, *A. niger*, and *A. oryzae*, along with new papers and analysis papers. The "AspGD and FungiDB integration" section discusses the integration of AspGD data into FungiDB. The footer includes a link to an "Open letter to the Aspergillus research community on genome database resources".

**Candida Genome Database (CGD)**: It primarily contains manually curated sequence information for *Candida albicans*. It also contains data pertaining to *Candida glabrata* and other species of *Candida*.

**New and Noteworthy**

**CANDIDA AND CANDIDIASIS 2021**

The Candida and Candidiasis 2021 meeting will take place online on March 21 - 27. Previously accepted presenters have been offered the opportunity to present their work at this meeting. New abstracts can also be submitted at this website. The deadline for new abstract submissions is now extended to Friday, 17 January 2021, 5 pm GMT.  
(Posted January 5, 2021)

**About CGD**

This is the home of the *Candida* Genome Database, a resource for genomic sequence data and gene and protein information for *Candida albicans* and related species. CGD is based on the *Saccharomyces* Genome Database and is funded by the National Institute of Dental & Craniofacial Research at the US National Institutes of Health.

**Meetings & Courses**

- Candida and Candidiasis 2021  
Online Meeting  
March 21 - 27, 2021  
Registration is now open at the [Meeting website](#). The registration deadline is Friday, March 12, 2021

**CGD Curation News**

- C. albicans* Genome Snapshot.
- C. glabrata* Genome Snapshot.
- C. parapsilosis* Genome Snapshot.
- C. dubliniensis* Genome Snapshot.
- C. auris* Genome Snapshot.
- New papers added to CGD this week.
- View Genome-wide Analysis papers in CGD.

**Frank Odds Obituary**

Read Frank Odds obituary written by Neil Gow, available [here](#).  
(Posted August 26, 2020)

**Candida and Candidiasis 2020 Conference postponed**

**FungiDB** : It is a data-mining interface focusing on comparative and functional genomics. It also allows for genomics comparison across species. The homepage provides many helpful tutorials.

**Search for...**

expand all | collapse all  
Filter the searches below...

- Genes
- Organisms
- Popset Isolate Sequences
- Genomic Sequences
- Genomic Segments
- SNPs
- ESTs
- Metabolic Pathways
- Compounds

**Overview of Resources and Tools**

Site search, e.g. NCU06658 or "reductase or "binding protein"

My Strategies Searches Tools My Workspace Data About Help Contact Us Guest

VEuPathDB Project

News and Tweets

Take a Tour Getting Started Search Strategies Genome Browser Transcriptomic Resources Phenotypic Data Analyze My Data Downloads How to Submit Data Curation and Annotation

**Getting Started**

VEuPathDB is packed with data, tools and visualizations that can help answer your research questions. We gather data from many sources, analyze according to standard workflows, and present the results for you to mine in a point and click interface. Here's how to get started:

**SITE SEARCH:** Explore the site; find what you need

Enter a term or ID in the site search box at the top of any page. The site search finds documents and records that contain your term and returns a summary of categorized matches. Its easy to find genes, pathways, searches, data sets and more with the site search.

VEuPathDB Site Search

**Tutorials and Exercises**

Grid view

**Lower eukaryotic model systems:** [Giardia DB](#), [Trypanocyc](#), [Sanger Protozoan Genome Projects](#).

Other single-celled organisms : [PlasmoDB](#), [T. gondii database](#), [Tetrahymena DB](#).

**Plants:**

[TAIR](#) – Arabidopsis database.

[PGSB](#) Plant Genome and Systems Biology - *Arabidopsis thaliana* project.

[IRGSP](#) – International Rice Genome Sequencing Project.

[MaizeGDB](#) – Maize genome database

[Dictyostelium](#) – Slime mold database.

[WormBase](#) – Database for *Caenorhabditis elegans*

[FlyBase](#) - Database of Drosophila Genes & Genomes

[MGI \(Mouse Genome Informatics\)](#) – is a database that stores information of laboratory mouse, and provides genetic, genomic, and biological data to study human diseases.

**TargetScan**: is a database for predicted miRNA targets in mammals. We can search the database by species and gene symbol.

The screenshot shows the TargetScanHuman homepage. At the top, it says "Prediction of microRNA targets" and "Release 7.2: March 2018 Agarwal et al., 2015". Below that, there's a search bar for "Search for predicted microRNA targets in mammals". It includes links to other TargetScan versions for Mouse, Worm, Fly, and Fish. The main search form has three sections:

1. Select a species: Human (dropdown)
2. Enter a human gene symbol (e.g. "Hmga2") or an Ensembl gene (ENSG00000149948) or transcript (ENST00000403681) ID
3. Do one of the following:
  - Select a broadly conserved\* microRNA family: Broadly conserved microRNA families (dropdown)
  - Select a conserved\* microRNA family: Conserved microRNA families (dropdown)
  - Select a poorly conserved but confidently annotated microRNA family: Poorly conserved microRNA families (dropdown)
  - Select another miRBase annotation: Other miRBase annotations (dropdown)
  - Enter a microRNA name (e.g. "miR-0-5p")

Below the form, there are notes about the terms "broadly conserved" and "conserved". There are "Submit" and "Reset" buttons at the bottom left.

**miRNEST 2.0** : is a collection of plant, animal, and virus miRNA information.

#### BROWSE PAGE

The miRNEST 2.0 browse page has several sections:

- Select species and source of miRNA sequences**: A section for "Plants" with a dropdown for "Select plant species", an "Animals" section with a dropdown for "Select animal species", and a "Viruses" section with a dropdown for "Select a virus".
- Choose database(s)**: A section with checkboxes for selecting databases. Options include "Select all", "miRNEST predictions", "microPC", "miRBase", "PMRD", "Huang et al.", and "Hao et al.". A note says you can also browse through miRNEST predictions in EST sequences using a [taxonomic tree](#).
- CLICK FOR MORE SEARCH OPTIONS**: A link to additional search parameters.
- Selected parametres**: Shows the selected organism ("none selected") and databases ("miRNEST, miRBase, PMRD, microPC, Hao et al., Huang et al. (by default))
- Go to another page**: A navigation bar with page numbers from 1 to 1957, a "Go to page" input field, and a "next >>" button. It also says "You are on page 1".
- Top species**: A list of species with their record counts:
  - Oryza sativa*, 4517 records
  - Populus trichocarpa*, 3079 records
  - Homo sapiens*, 2200 records
  - Arabidopsis thaliana*, 1938 records
  - Mus musculus*, 1184 records

**[PolymiRTS Database 3.0](#)** : has information of SNPs in miRNA target sites. It allows us to study how SNPs interfere with miRNA-based gene silencing.

**[miRDB](#)** : database for miRNA target predictions and functional annotations.

**[miRbase](#)** : repository of known miRNA sequences and annotations.

**[Rfam](#)** : resembles the Pfam database and is hosted by EBI. It provides information of ncRNA families.

**[NONCODE](#)** : database for ncRNAs (excluding tRNAs and rRNAs) especially lncRNAs.

**[ANCORA](#)** : database for highly conserved noncoding elements (HCNEs) in metazoan genomes.

**[cneViewer](#)** : displays conserved regions between human and zebra fish genomes. Users can search data based on gene name of CNE ID.

**[UCNEBase](#)** : repository of information pertaining to Ultra conserved non-coding elements (UCNEs) in 18 different vertebrate species.

**Tools for analysis of metagenomes:** [Real-Time metagenomics](#), Rapid Annotation using Subsystem Technology ([RAST](#)), metagenomics analysis tool in Galaxy.

## References

- Galaxy. (n.d.). De novo transcriptome reconstruction with RNA-Seq. Retrieved from  
<https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/de-novo/tutorial.html>
- Glimmer. (n.d.). ABOUT GLIMMER. Retrieved from  
<http://ccb.jhu.edu/software/glimmer/index.shtml>
- Mistry, M., & Khetani, R. (n.d.). Introduction to ChIP-Seq using high-performance computing. Retrieved from [https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05\\_peak\\_calling\\_macs.html](https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html)
- Yun, S., & Yun, S. (2021). AS.410.635.81.SU21 Bioinformatics: Tools for Genome Analysis-Course Materials. Retrieved from  
[https://blackboard.jhu.edu/webapps/blackboard/content/listContent.jsp?course\\_id=\\_2324\\_65\\_1&content\\_id=\\_9712340\\_1&mode=reset](https://blackboard.jhu.edu/webapps/blackboard/content/listContent.jsp?course_id=_2324_65_1&content_id=_9712340_1&mode=reset)