

# JAX Big Genomic Data Skills Training - 2018

## Human Disease Variant Discovery Module

### Introduction

This data analysis module introduces next generation sequence analysis to students. The sequence provided comes from human patients. The Jackson Laboratory (JAX) acquired the anonymous Bacterial Artificial Chromosomes (BACs) human DNA fragments in order to engineer a humanized mouse model. The human genomic regions (~250 kb in size) were used to create transgenic mice that carry a large fragment of human chromosome 12. In order to attempt to duplicate the human disease in mice JAX has made a wide range of transgenic mice carrying numerous mutations found in humans; additionally the normal (non-mutation) region of human chromosome 12 has been engineered (inserted via transgenesis) into mice as a control.

The goal of this exercise is identify the gene and potential pathogenic genotypes (variants) in a series of BAC sequences. The BACs were prepared by library preparation and sequenced at JAX on an Illumina miSeq platform. Sequencing was done as paired-end with 150 bp reads. Sequencing ~250 kb is a small project by modern high throughput standards and by aligning to a single human chromosome during data analysis this module is computationally less demanding. Specifics about how analyze data and target to a specific region of a genome are provided below.

### What do students need to know coming in (or cover before launching the analysis)?

Basic molecular biology- DNA, RNA, proteins. Transcription. Translation.

What is Next Generation or High Throughput DNA sequencing.

What does it mean to align sequence to a reference genome.

What is the difference between a variant, polymorphism, and mutation in genomic DNA, and what is a pathogenic mutation.

### Bonus Knowledge for students

Introductory knowledge of bioinformatics resources would be useful including NCBI gene resources and OMIM.

Know why different human genome sequence references exist, e.g. different 'releases' of the genome with different coordinates and annotations

Having experience with Ensembl genome database would be useful.

Having a basic understanding of genetic engineering technologies including creation of transgenic mice and logic behind humanized mice.

### Knowledge (concept) goals

What is a fastq file, and what information does it hold (in groups of four lines)?

What Quality Control (QC) steps are recommended for genomic sequencing

What is the difference between aligning directly to a genome or aligning to a target region.

How are duplicate short reads identified and removed to avoid bias.

**Practical skills goals**

Logging in to Galaxy

Uploading a working sequence set and a fasta file of a single human chromosome

Gaining familiarity with file types, fasta, fastq, bam, vcf

Perform QC steps and possibly process the sequence files to remove issues

Aligning individual files to a single human chromosome

Call sequence variants from aligned short reads

Interpreting types of variants reported in a vcf file

## Human Disease Variant Discovery Module

### 1. Getting Data

- a. Human Chromosome 12 file: In this module several Bacterial Artificial Chromosomes (BACs) from humans are sequenced. As we know the human sequence was cloned from Chromosome 12, therefore we target the whole module to this chromosome. By targeting to a single chromosome this will speed up the computational steps, especially alignment of sequence reads, as only a small portion of the human genome (just Chr12) is the target of the exercise.

A file named “chr12.fa” will be provided. This is a simple linear sequence file of human chromosome 12. This will become the ‘reference’ human sequence for the module.

- b. BAC Sequence Files

The data files are provided by ftp transfer, or shared via a galaxy shared data library. Files, by convention usually look like this:

1\_**S1**\_L001\_**R1**\_001.fastq.gz translated 1 = just #; S1 = experimental **Sample1**; L001 = Lane; R1 = **Read 1** “forward read”; 001 = just #; fastq = file type.

The key identifiers are the sample number **S1** (this usually represents the biological sample prepared for sequencing, (library prepped) and the **Read**, which will be **R1** or **R2**.

For this module all samples will come in paired reads or R1 and R2; each student or student group can or may be given one pair of Reads. For instance

Student group 1:      10\_S8\_R1\_001.fastq.gz  
                         10\_S8\_R2\_001.fastq.gz

This is genomic library preparation sample 8 (**S8**) reads 1 and 2.

Student group 2:      12\_S10\_R1\_001.fastq.gz  
                         12\_S10\_R2\_001.fastq.gz....and so on.....

Data in Galaxy will look like this



This dataset is large and only the first megabyte is shown below.

[Show all](#) | [Save](#)

```
@M00263:10:00000000-A44FG:1:1101:16361:1640 2:N:0:2
NAGTAGTCTGCTACATGATAAGGAAGCTGCTACAGGGAAGAGGCTATTTGGGATAAGGCTGGATGCATAAGTAAAGTTGAATATTGCAAGGAGTTGAAT
+
#,,55,/7<<@@@-</.8//8AA-,6CA>EDF-9-+++,7+7>-AFFG.+++8-88---,@,55CAE=CBCEFF-5CC=EEE-CC-@-+++8=--8A
@M00263:10:00000000-A44FG:1:1101:18918:1641 2:N:0:2
NCAAAGGGATACAAGTCAGTGTGACAATCAGAAGCAAAACCGTCCACCTTCTACTCACTGACAAGACCCAGAGAAGGCTCATAAGCAATTCCTCTGGAG
+
#5,555>+<-5@9@@C/.8/8ECAACEC;C-=@E=-9-77>@>C@FDEFFDFF@=-ACB@EF,ACEE+AEEEF=EEDDEE=->>--6A5A-55--6+
@M00263:10:00000000-A44FG:1:1101:15475:1641 2:N:0:2
NCCTATTGAACCTGGCAACCTCGGTTTTTATACTAGAGCTCAGGAGGCGCAGGACGACGGGCCACCCAGGATAAGAAAAAGGCCCCACTTTCGTCAT
+
#555,5>--5<@9-<--6ACE9+CEEEEC,C..9...9A-----+5**5+5+,5**5)<+5+++44=+4+=9++3;+4=@EEE**1;**
@M00263:10:00000000-A44FG:1:1101:13783:1642 2:N:0:2
NTATTTGATTCTCACTGAAATCTCACTGCCCTCTGGAAATATTACCGGATGTTTACTGTGCATAGCGAAGTGAGAGTAAGCTGCTCACTGAGGATCAAA
+
#5,5<=//<A<AA-<@E;.CCCEE>>EBEEFEEDD---9.9E.AA+++D+AE.C.8E8AE.A--5++*55-5A--5@-@C@-88C-AC---+88+5,
@M00263:10:00000000-A44FG:1:1101:13875:1642 2:N:0:2
NGAGATGGGTTATCTTGACATGTGCTTTTAGGAAAGAGACTCTTAGAGAGCAGTAGAGGACGCCAGAAAGGGCAGAGGAAAAATCTCAGCAGAAATGTG
```

## 2. Simple raw data visualization Line/Word/Character tool count

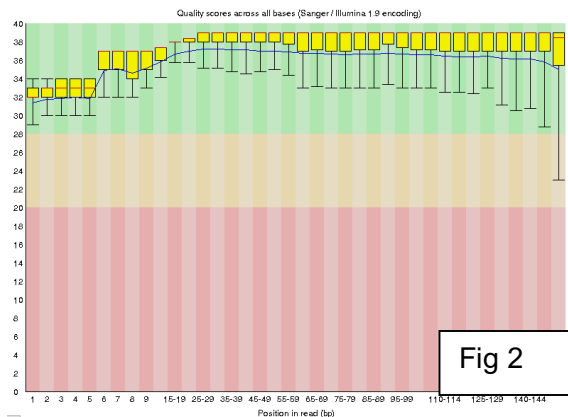
It may be valuable to have students run this very simple Galaxy tool. The tool simply tells you how many sequence reads are in your file; paired files should have equal numbers of reads. For instance after hitting the eye-ball symbol on the Line/Word/Character count tool students will see that the file includes, in this instance, over 20M reads (#lines):

1	2	3
#lines	words	characters
20241740	25302175	1773419817

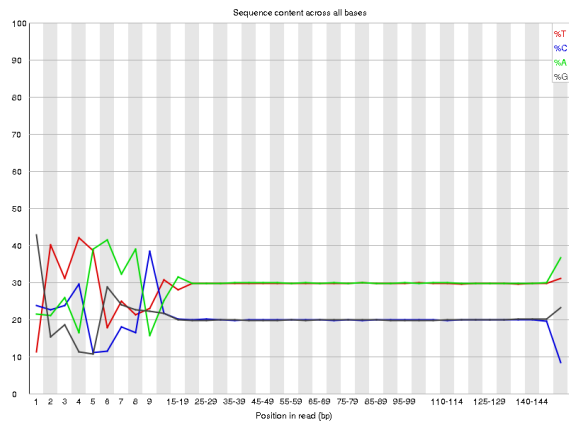
## 3. Performing Quality Control (QC) (fastqsanger)

The raw data is presented in a fastq file, which is specific for Illumina sequencing. This file is comprised not only of the nucleotide sequence, but also includes an ID number and quality score which is important for determining the integrity of the data obtained. A fastq file is obtained for both the forward and reverse reads (R1 and R2), and these typically range from 50-150 base pairs, in this module raw reads are 150bp. These files are stored separately and run through the FastQC tool on Galaxy in order perform quality control checks on raw sequence data. This tool is characterized by primarily the per base sequence quality (Fig.1), the per base sequence content (Fig.2), the adapter content (Fig.3) and the Kmer content (Fig.4). The per base sequence quality should be over 30 for it to be considered a high quality score for use. Quality scores tend to be lower near the beginning of the read and drop off near the end. The per base sequence content should be uniform, such that there are equal numbers of each base (~25%) over

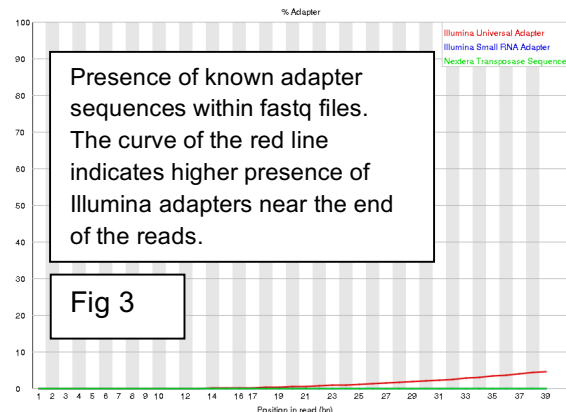
the whole read. The adapter content indicates the location and amount of the adapter sequence that is included in the read, which is important to note for trimming purposes. Finally, the Kmer content indicates sequences that are abnormally repeated. Based on the FastQC, the reads should be trimmed using the *Trimmomatic* tool in Galaxy in order to remove any low quality portions of the reads that would affect alignment in subsequent steps.



Quality scores of bases sequenced by illumina sequencing at each position. Scores are lower at the ends, however, all scores are high quality.



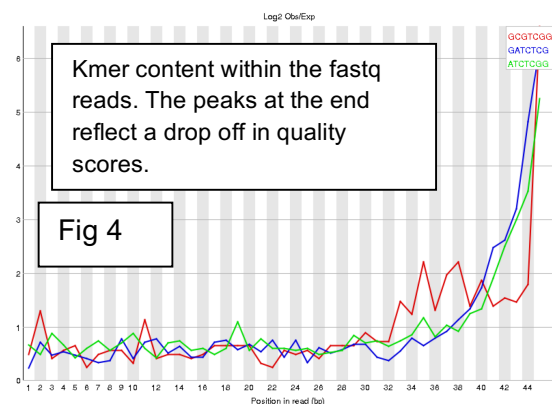
Relative content of each base In the sequences obtained by the illumina Sequencer. Noise at the beginning indicates the presence of a repeated sequence.



#### 4. QC processing with trimmomatic.

Trimmomatic is one popular tool for removing systematic problems from NGS data. (You are welcome to use others that are available and that you prefer).

A suggested set of settings for our data is shown below:



**Trimmomatic** flexible read trimming tool for Illumina NGS data (Galaxy Version 0.36.3)

**Single-end or paired-end reads?**

**Input FASTQ file (R1/first of pair)**

**Input FASTQ file (R2/second of pair)**

**Perform initial ILLUMINACLIP step?**

Cut adapter and other illumina-specific sequences from the read

**Trimmomatic Operation**

1: Trimmomatic Operation

**Select Trimmomatic operation to perform**

**Number of bases to average across**

**Average quality required**

Continued:

**Trimmomatic Operation**

1: Trimmomatic Operation

**Select Trimmomatic operation to perform**

**Number of bases to average across**

**Average quality required**


**What it does**

Trimmomatic performs a variety of useful trimming tasks for illumina paired-end and single ended data.

This tool allows the following trimming steps to be performed:

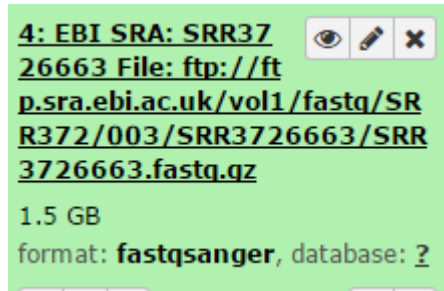
**ILLUMINACLIP:** Cut adapter and other illumina-specific sequences from the read  
**SLIDINGWINDOW:** Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold  
**MINLEN:** Drop the read if it is below a specified length  
**LEADING:** Cut bases off the start of a read, if below a threshold quality  
**TRAILING:** Cut bases off the end of a read, if below a threshold quality  
**CROP:** Cut the read to a specified length  
**HEADCROP:** Cut the specified number of bases from the start of the read

If ILLUMINACLIP is requested then it is always performed first; subsequent options can be mixed and matched and will be performed in the order that they have been specified.

 Note that trimming operation order is important.

- Note:** In some cases raw data files come in different formats. Usually fastq, fastqsanger or others. Data is essentially the same but certain analysis tools expect certain formats. If necessary sequence files can be converted:

- a. Load data in to Galaxy
- b. Highlight data file and click on ? after database (bottom right corner):



- c. Go to main screen and highlight “datatype” tab and use pull down to select type you need and hit save.

6. **Examine the updated (newly generated) fastq file** with FASTQC, using the same logic as for step 4 above.

Has Trimmomatic changed and/or improved following the QC processing?

7. **Alignment to Reference** using Map with BWA-MEM

Aligning a sequence to a reference is a critical and time consuming step in the process. This is the step where the short sequences are aligned back to a reference genome, human, mouse yeast etc. It is vital that you know what reference genome you are aligning to, specifically what version or release of an annotated genome. As described earlier, in this training module you will align to a built index file of human chromosome 12. In the window below notice that you will select: ‘use a genome from history and build index’ then ‘use the following dataset as a reference sequence’. This is where you specify chromosome 12 as ‘chr12.fa’. If you are using a cloud instance of galaxy and you want to map to a whole genome, you will need the reference genome installed within your cloud instance. In this (Human BAC Variant) module you should align to Human hg19 or “Human Feb. 2009 (GRCh37/hg19) (hg19)”. This will create a BAM file (Binary Alignment Mapping file). Analysis may take several hours (or days depending on activity on public galaxy instance).

**Map with BWA-MEM** - map medium and long reads (> 100 bp) against reference genome (Galaxy Version 0.2.2) Versions Options

**Will you select a reference genome from your history or use a built-in index?**

Use a genome from history and build index

Built-ins were indexed using default options. See `Indexes` section of help below

**Use the following dataset as the reference sequence**

34: chr12.fa

You can upload a FASTA sequence to the history and use it as reference

**Single or Paired-end reads**

Paired

Select between paired and single end data

**Select first set of reads**

25: Trimmomatic on 8\_S6\_L001\_R1\_001.fastq (R1 paired)

Specify dataset with forward reads

**Select second set of reads**

26: Trimmomatic on 8\_S6\_L001\_R2\_001.fastq (R2 paired)

Specify dataset with reverse reads

## 8. Post BWA-MEM

At this point it is possible to visualize all the reads aligned against the reference genome (human in this case). This can be done by using IGV which is linked to Galaxy.

**27: Map with BWA-MEM on data 4 and data 3 (mapped reads in BAM format)**

770.3 MB

format: **bam**, database: **hg19**

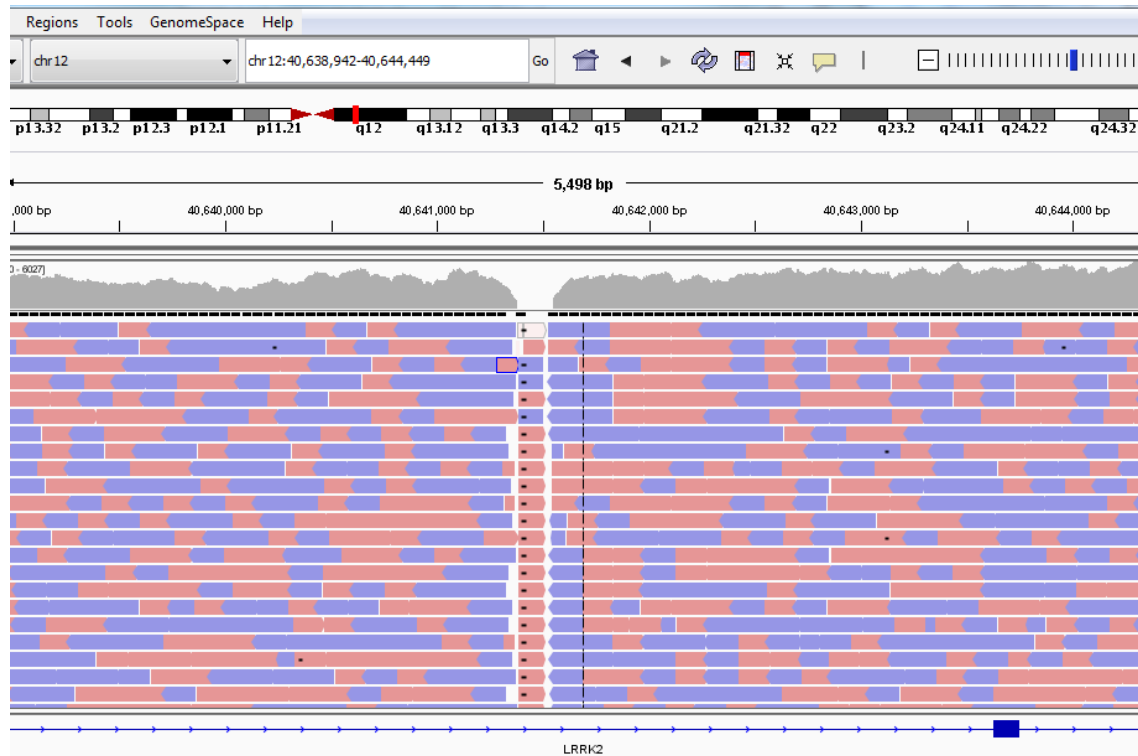
[main] Version: 0.7.10-r876-dirty  
[main] CMD: bwa mem -t 1 -v 1 -R @RG ID:M00263:10:000000000-A44FG:MJFF3 SM:MJFF3 PL:ILLUMINA LB:MJFF2 -I 250 /mnt/galaxyIndices/hg19/bwa\_mem /mnt/galaxy/files/000/dataset\_40 /mnt/galaxy/files/000/

display at UCSC [main](#)  
display at Ensembl [Current](#)  
display with IGV [local](#) [Human](#)  
[hg19](#)  
display in IGB [View](#)

Use the *display with IGV*. Depending on your computer platform you may need to install IGV on your machine. It will load the genome reference.



Be patient you will need to wait as it builds the graphical interface with all the reads. Also the level of zoom-in is important. For this module (which is targeted BAC sequencing of human chromosome 12) you may want to insert these chromosome coordinates into IGV: **chr12:40,596,343-40,684,481**. A likely duplicate sequence is seen here in IGV:



Note: IGV Visualization Tool




IGV can be used to visualize data with reference to a chromosome. The bottom track shows the genes on the chromosome including the introns and exons. Variants are visible in the reads shown above the gene. More information about each variant can be found by scrolling over the variant to see the cigar and phred scores. The total number of reads for each variant can be found including the number and percentage of reads for each type of base (A, T, C, G).

## 9. **Mark duplicates**

When looking for DNA sequence variants, using the Mark duplicates tool is important for weeding out duplicate (identical reads) that can introduce frequency and absolute number bias in variant calling. Duplicate short-sequence reads start and end on the exact base in a BAM file and can be easily identified in your BAM outputs from BWA-mem in step 7. Duplicates will also have identical base scores.


**MarkDuplicates** examine aligned records in BAM datasets to locate duplicate molecules (Galaxy Version 1.126.0) Options

**Select SAM/BAM dataset or dataset collection**

   34: (unavailable) Map with BWA-MEM on data 22, data 21, and data 33 (mapped reads in BAM form... ▼

If empty, upload or import a SAM/BAM dataset

**Comment**

 **Insert Comment**

You can provide multiple comments

**If true do not write duplicates to the output file instead of writing them with appropriate flags set**

☐ Yes ☒ No

REMOVE\_DUPLICATES; default=False

**Assume the input file is already sorted**

☐ Yes ☒ No

ASSUME\_SORTED; default=True

**The scoring strategy for choosing the non-duplicate among candidates**

SUM\_OF\_BASE\_QUALITIES ▼

Duplicate\_SCORING\_STRATEGY; default=SUM\_OF\_BASE\_QUALITIES

**Regular expression that can be used to parse read names in the incoming SAM/BAM dataset**

[a-zA-Z0-9]+:[0-9]:([0-9]+):([0-9]+):([0-9]+).\*

READ\_NAME\_REGEX; Read names are parsed to extract three variables: tile/region, x coordinate and y coordinate. These values are used to estimate the rate of optical duplication in order to give a more accurate estimated library size. See help below for more info; default=[a-zA-Z0-9]+:[0-9]:([0-9]+):([0-9]+):([0-9]+).\*

**The maximum offset between two duplicate clusters in order to consider them optical duplicates**

100

OPTICAL\_DUPLICATE\_PIXEL\_DISTANCE; default=100

## 10. Generating a Variant Call File

After aligning to a reference genome and removing duplicate reads, a VCF file is produced using FreeBayes. As with the aligning reads to a reference, in this module it is best to only search for sequence variants against the target region of human chromosome 12. This can be done by specifying FreeBayes as follows:

**FreeBayes - bayesian genetic variant detector** (Galaxy Version 0.4.1) Options




**Load reference genome from**


History ▼

**Sample BAM file**




1: Sample BAM file

**BAM file**

   38: MarkDuplicates on data 36: MarkDuplicates BAM output ▼

 **Insert Sample BAM file**

**Use the following dataset as the reference sequence**

   34: chr12.fa ▼

You can upload a FASTA sequence to the history and use it as reference

**Limit variant calling to a set of regions?**

Do not limit ▼

Sets --targets or --region options

**Choose parameter selection level**

1: Simple diploid calling ▼

Select how much control over the freebayes run you need

A VCF file should include all genetic changes at a scale smaller than the imputed read sizes (~50-150 bp in length). A sample VCF file is below.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP

**Info:** The VCF QUAL score is a Phred Quality score scaled to the probability that a base is incorrectly called.

Not all the variation calls in a VCF file are correct or worth further exploration. Galaxy's FreeBayes tool, which produces a VCF file from a BAM file, only declares variations that are corroborated by at least 2 reads or 20% of reads, a relatively low threshold that allows for extraneous variation calls. So, it is important to view a VCF file in some sort of visualization software, such as IGV, which aligns the reads and variation calls against a reference genome, making it easier to see which variation calls are strong and which are weak. Galaxy also has a tool (slice VCF) that limits VCF data to a specific part of the genome as specified by a bed file. Variations are then used to identify sequence variants which may be deleterious.

#### 11. Variant Identification (Ensembl Variant Effect predictor, VEP)

[http://useast.ensembl.org/Homo\\_sapiens/Tools/VEP?db=core](http://useast.ensembl.org/Homo_sapiens/Tools/VEP?db=core)

The first step is to download and save the VCF file produced by FreeBayes. The file should be rather small with 300-400 lines and a 100-200 KB in size. Upload the file through the 'Choose File' option on Ensembl VEP. **Important: for this exercise use Human reference GRCh37.p7. This will match the alignment to reference hg19; this is important.**

**e!GRCh37** BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog

Human (GRCh37.p13) VEP

**Web Tools**

- Web Tools
  - BLAST/BLAT
  - Variant Effect Predictor**
  - File Chameleon
  - Assembly Converter
  - ID History Converter
  - VCF to PED Converter
  - Allele Frequency Calculator
  - Data Slicer
  - Variation Pattern Finder

Configure this page

Custom tracks

Export data

Share this page

Bookmark this page

## Variant Effect Predictor ?

**i VEP for Human GRCh38**

If you are looking for VEP for Human GRCh38, please go to [GRCh38 website](#).

**i VEP for non-human species**

VEP is now only available on this site for Human (GRCh37). For other species, please visit our [main site](#).

Species: Human (Homo sapiens) Assembly: GRCh37.p13

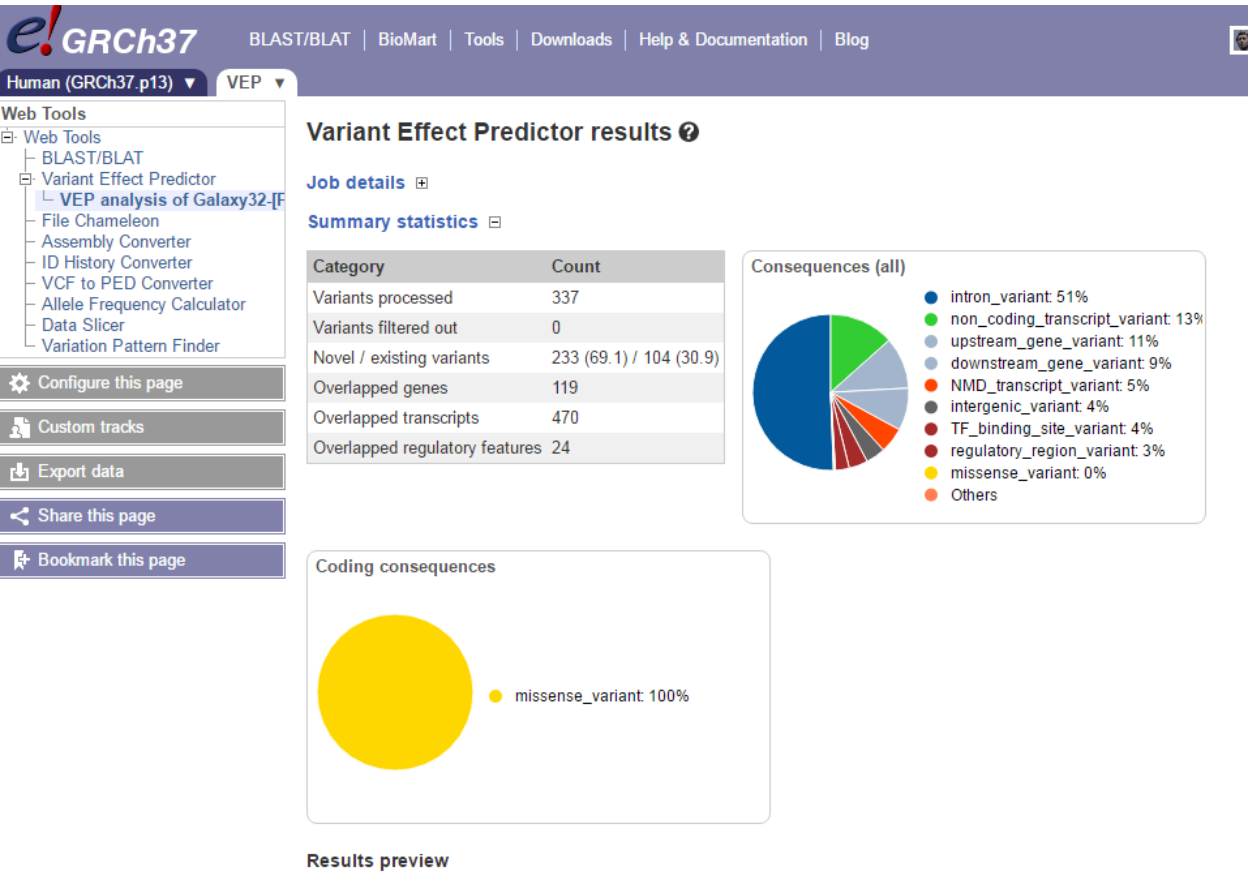
Name for this job (optional):

Either paste data:

Examples: [Ensembl default](#), [VCF](#), [Variant identifiers](#), [HGVS notations](#).  
NB: pileup format no longer supported

Or upload file: Choose File No file chosen

The VEP tool will look at the human genome in Ensembl and then compile a list of variants that are may be causative of a phenotype, in this case within the genetic interval entered, human chr. 12. Output in this module will look like this:



The results of VEP can be sorted consequence or gene symbol by toggling between options in the Ensembl table. A table view of VEP is appended under the summary. Consequence describes the change in gene function. IMPACT type is the proposed extent of the amino shift.

#Uploaded_variation	Location	Allele	Consequence	IMPACT	SYMBOL	Gene
.	15:4053869-4053871	A	"splice_region_variant,intron_variant,non_coding_transcript_variant"			
.	15:3277221-3277221	A	missense_variant	MODERATE	-	ENSMUSG00000064373
.	15:31273150-31273150	G	intron_variant	MODIFIER	1700001L05Rik	ENSMUSG00000039168
.	15:12368649-12368649	C	synonymous_variant	LOW	1700088E04Rik	ENSMUSG00000022197
.	15:10591621-10591630	AAAAAAAAG	intron_variant	MODIFIER	1700109K24Rik	ENSMUSG00000000000
.	15:33034074-33034074	C	3_prime_UTR_variant	MODIFIER	1810021B22Rik	ENSMUSG00000000000
.	15:31294473-31294473	A	"3_prime_UTR_variant,NMD_transcript_variant"	MODIFIER		1810021B22Rik
.	15:36092873-36092873	G	missense_variant	MODERATE	4930415020Rik	ENSMUSG00000000000
.	15:27856218-27856218	C	regulatory_region_variant	MODIFIER	4930483J18Rik	-
.	15:10334760-10334760	T	intron_variant	MODIFIER	4933427E11Rik	ENSMUSG000000005268
.	15:9100143-9100145	G	"splice_acceptor_variant,frameshift_variant"	HIGH	9930014A18Rik	ENSMUSG00000000000
.	15:9100143-9100145	G	"splice_acceptor_variant,frameshift_variant"	HIGH	Albg	ENSMUSG00000000000
.	15:35115675-35115675	G	intron_variant	MODIFIER	Acr	ENSMUSG00000022329
.	15:35671374-35671374	C	synonymous_variant	LOW	Adamts20	ENSMUSG000000037646
.	15:36092873-36092873	G	missense_variant	MODERATE	Adcy6	ENSMUSG000000037627
.	15:9100143-9100145	G	"non_coding_transcript_exon_variant,non_coding_transcript_variant"			
.	15:10334760-10334760	T	intron_variant	MODIFIER	Adgrb1	ENSMUSG000000005268
.	15:10334760-10334760	T	"intron_variant,NMD_transcript_variant"	MODIFIER	Ago2	ENSMUSG00000000000
.	15:35115675-35115675	G	"intron_variant,NMD_transcript_variant"	MODIFIER	Alg10b	ENSMUSG00000000000
.	15:38518935-38518935	G	"5_prime_UTR_variant,NMD_transcript_variant"	MODIFIER	Ankr	ENSMUSG00000000000
.	15:4053869-4053871	A	"splice_region_variant,intron_variant"	LOW	Ankrd33b	ENSMUSG00000000000
.	15:12185140-12185140	T	downstream_gene_variant	MODIFIER	Ankrd54	ENSMUSG0000000022201
.	15:35925930-35925930	G	intron_variant	MODIFIER	Ano6	ENSMUSG0000000037646
.	15:21586972-21586972	C	3_prime_UTR_variant	MODIFIER	Apobec3	ENSMUSG0000000040452
.	15:10591621-10591630	AAAAAAAAG	intron_variant	MODIFIER	Apo110a	ENSMUSG0000000022246
.	15:10591621-10591630	AAAAAAAAG	intron_variant	MODIFIER	Apo17c	ENSMUSG0000000022246

In this exercise there are numerous samples that have been provided as raw sequence. All samples are human. There are at least 3 different genotypes for the target gene: two contain likely pathogenic mutations causative of disease. One genotype represents normal, if there is such a thing as a normal genotype or phenotype.

**The end point goal for students is to discover the genotype, pathogenic-likely or normal for their samples. The end point may be in tying LRRK2 to Parkinson's if you have not told the students in advance what the gene of interest is.**



This is variant at human (hg38) at chr12:40340400 a Parkinsonian SNP rs34637584; this is for clone or sample #6

1\_S6\_L001\_R1\_001.fastq

1\_S6\_L001\_R2\_001.fastq

Display Settings: Summary

Send to: ▼

☐ rs34637584 [Homo sapiens]

1.

CATCATTGCAAAAGATTGCTGACTAC[A/G]GCATTGCTCAGTACTGCTGTAGAAT

Chromosome: 12:40340400

Gene: LRRK2 (GeneView)

Functional Consequence: missense

Allele Origin: G(germline)/A(germline)

Clinical significance: Pathogenic

Validated: by 1000G, by cluster, by frequency

Global MAF: A=0.0002/1

HGVs: NC\_000012.11:g.40734202G>A, NC\_000012.12:g.40340400G>A, NG\_011709.1:g.120390G>A, NM\_198578.3:c.6055G>A, NP\_940980.3:p.Gly2019Ser, XM\_005268629.1:c.6055G>A, XM\_005268629.3:c.6055G>A, XM\_005268630.1:c.6055G>A, XM\_005268631.1:c.5935G>A, XM\_005268632.1:c.6055G>A, XM\_011537877.2:c.6055G>A, XM\_017018787.1:c.2971G>A, XM\_017018788.1:c.2317G>A, XP\_005268686.1:p.Gly2019Ser, XP\_005268687.1:p.Gly2019Ser, XP\_005268688.1:p.Gly1979Ser, XP\_005268689.1:p.Gly2019Ser, XP\_011536179.1:p.Gly2019Ser, XP\_016874276.1:p.Gly991Ser, XP\_016874277.1:p.Gly773Ser

[PubMed](#) [Varview](#)

**\*609007**

## Table of Contents

## Title

Gene-Phenotype  
Relationships

## Text

## Description

Cloning and  
Expression

## Gene Structure

## Mapping

## Gene Function

Biochemical  
FeaturesMolecular  
Genetics

## Animal Model

## Allelic Variants

**.0006 PARKINSON DISEASE 8, AUTOSOMAL DOMINANT**

LRRK2, GLY2019SER

dbSNP:rs34637584

ExAC:rs34637584

RCV000325492...

In affected members of 4 of 61 (6.6%) unrelated families with autosomal dominant Parkinson disease (607060), [Di Fonzo et al. \(2005\)](#) identified a heterozygous 6055G>A transition in exon 41 of the LRRK2 gene, resulting in a gly2019-to-ser (G2019S) substitution. Two families were from Italy, and 1 each were from Portugal and Brazil. The gly2019 residue is highly conserved and is part of a 3-amino acid motif required by all human kinase proteins. [+](#)

[Gilks et al. \(2005\)](#) identified the G2019S mutation in 8 of 482 (1.6%) unrelated patients with Parkinson disease. Five of the patients had no family history of the disorder, suggesting either a de novo occurrence or reduced penetrance. [+](#)

[Nichols et al. \(2005\)](#) identified the G2019S mutation in 20 of 358 (6%) families with PD. In 1 family, 1 sib was heterozygous for the mutation and another was homozygous; the homozygous individual did not differ in clinical presentation.

▼ External  
Links

## ▶ Genome

## ▶ DNA

## ▶ Protein

## ▶ Gene Info

▶ Clinical  
Resources

## ▼ Variation

1000 Genome  
ClinVar  
ExAC  
gnomAD  
GWAS Catalog  
GWAS Central

## KEY to Samples:

**Samples S2, S4, & S6** contain a missense variant: chr12 g. 40734202G>A[1/1] (rs34637584) associated with Parkinson's disease, Gene LRRK2.

## Associated publication:

[Mutations in LRRK2 increase phosphorylation of peroxiredoxin 3 exacerbating oxidative stress-induced neuronal death.](#) Angeles DC, Gan BH, Onstead L, Zhao Y, Lim KL, Dachsel J, Melrose H, Farrer M, Wszolek ZK, Dickson DW, Tan EK. Hum Mutat. 2011 Dec;32(12):1390-7. doi: 10.1002/humu.21582. Epub 2011 Sep 12. PMID: 21850687

**Samples S8 & S10** contain a missense variant: chr12 g. 40704236C>G[1/1] (rs33939927) associated with Parkinson's disease, Gene LRRK2.

[Genetic etiology of Parkinson disease associated with mutations in the SNCA, PARK2, PINK1, PARK7, and LRRK2 genes: a mutation update.](#) Nuytemans K, Theuns J, Cruts M, Van Broeckhoven C. Hum Mutat. 2010 Jul;31(7):763-80. doi: 10.1002/humu.21277. Review. PMID:20506312

**Sample S5 is a normal sequence, no Parkinson's pathogenic variant**



**Variant Call Report: HUMAN BAC sequencing**

**BAC Library Identifier:** \_\_\_\_\_

**Galaxy Workflow used (cut and paste screen shot or otherwise share):**

**Outcome, Variants identified (give dbSNP rs### identifier code):**

**Please identify # of variants and whether the variants identified do or do not have human disease associations. What are the clinical associations? Provide one or two peer reviewed references that describe any clinically relevant SNPs.**