**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**
   1. 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.fasta” will be provided. This is a simple linear sequence file of human chromosome 12. This will become the ‘reference’ human sequence for the module.

* 1. 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 **S**ample**1**; L001 = Lane; R1 = **R**ead **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, file names are simplified. For instance

Student group 1: 10**S8R1**.fastq.gz

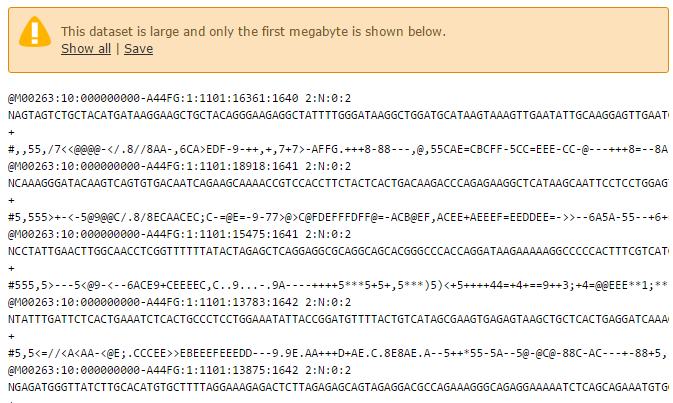
10**S8R2**.fastq.gz

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

Student group 2: 12**S10R1**.fastq.gz

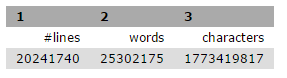
12**S10R2**.fastq.gz….and so on……..

Data in Galaxy will look like this



1. **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. **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.

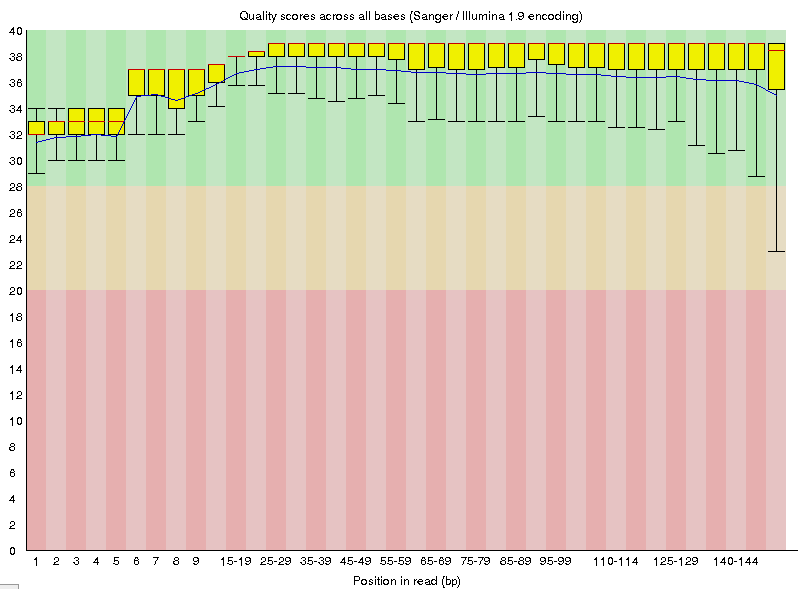
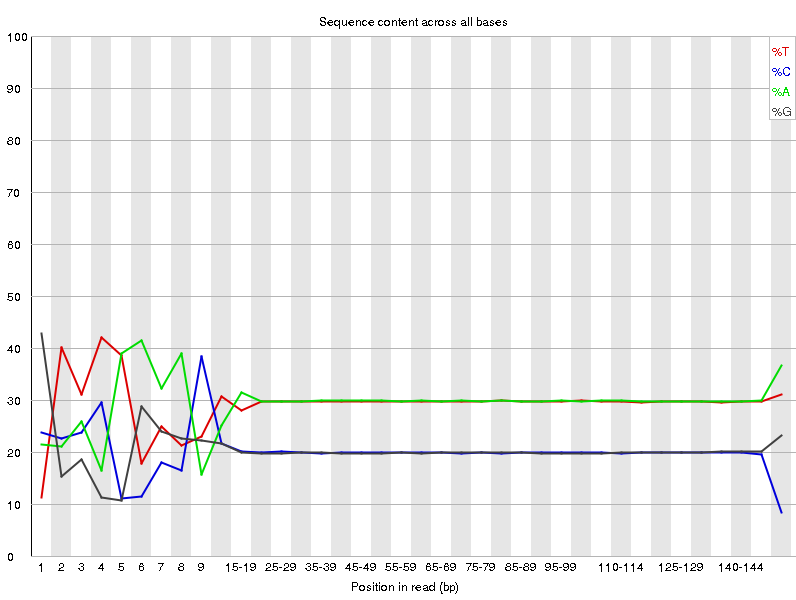


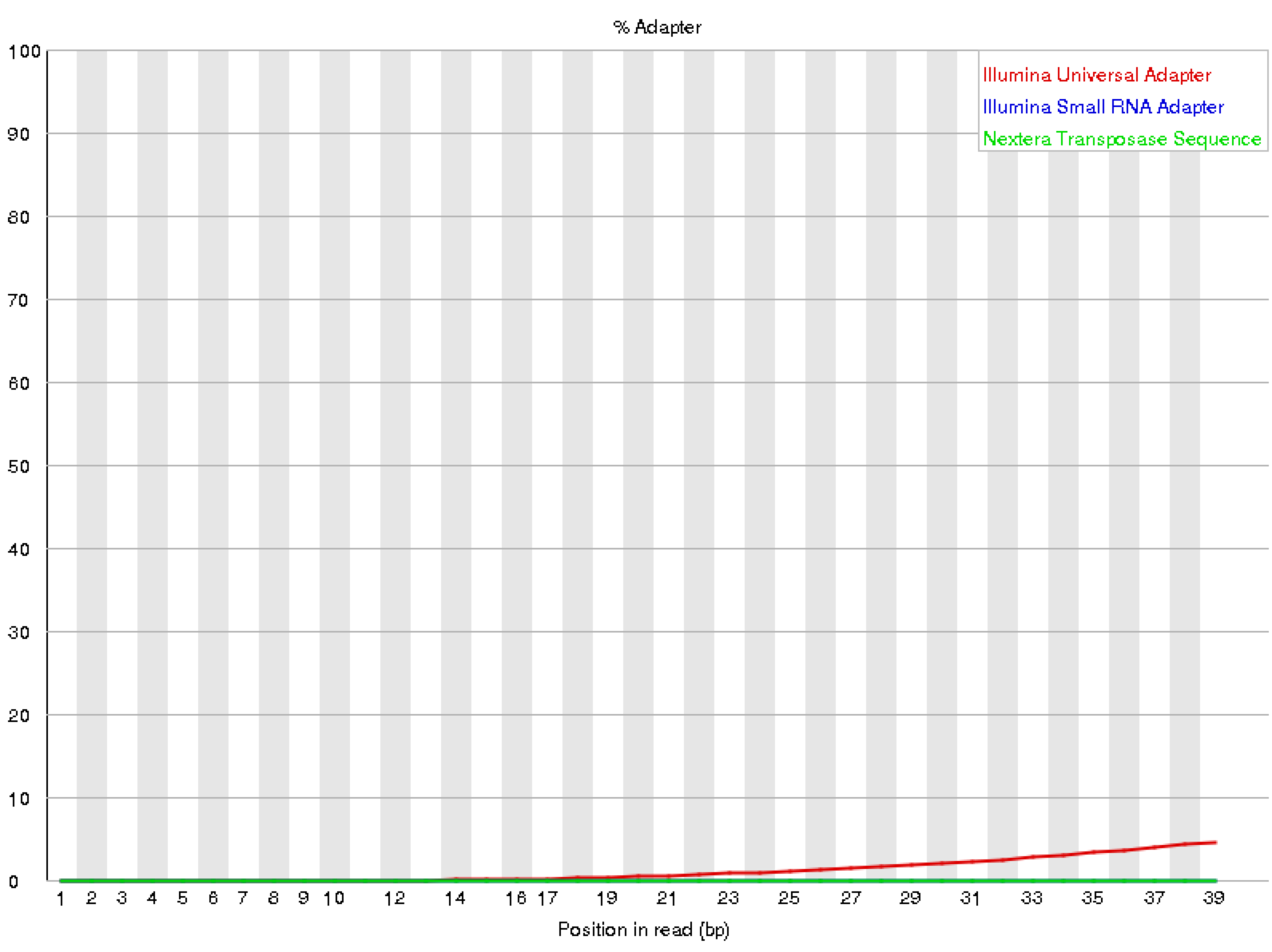
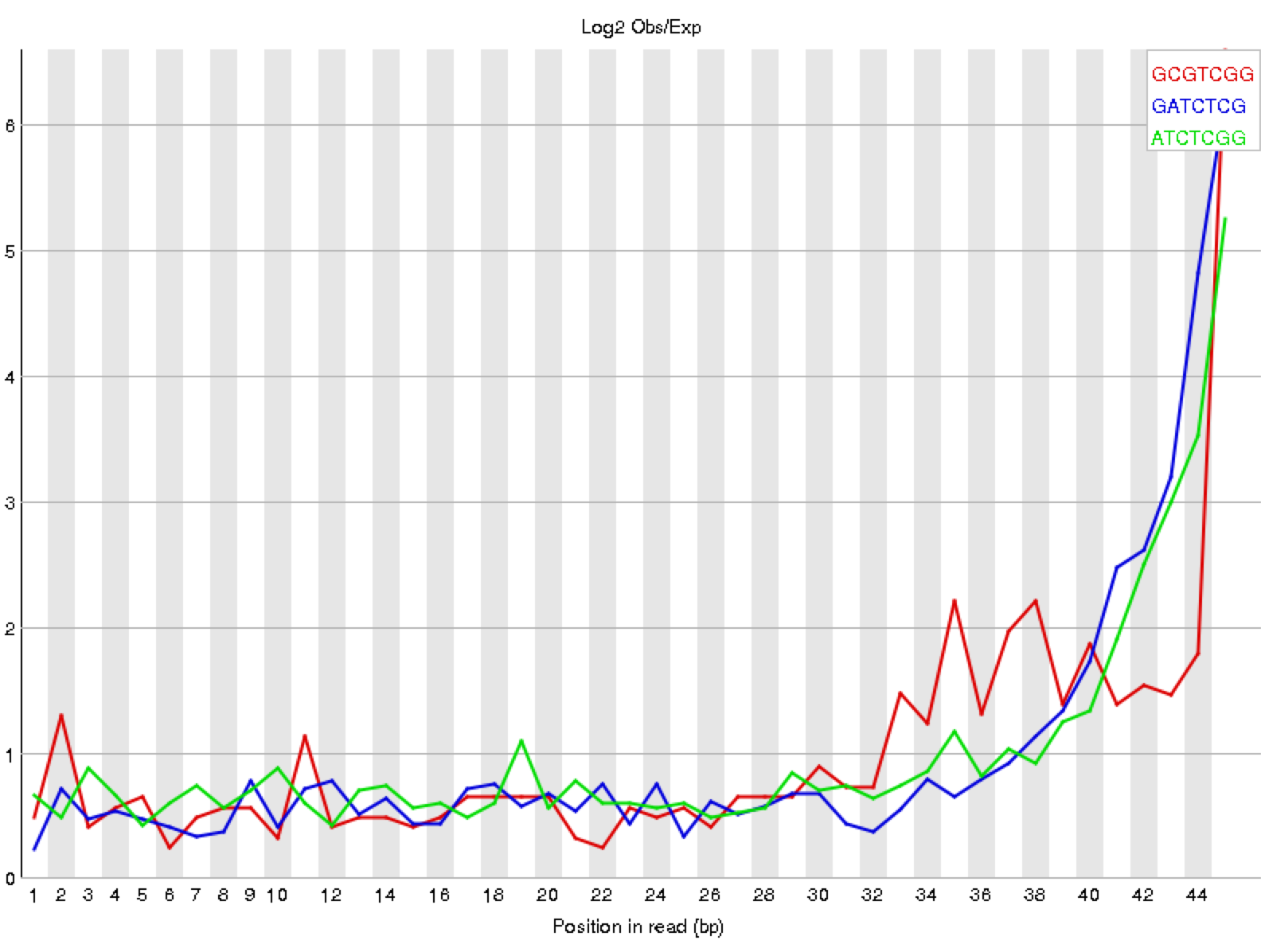
Fig 2

Fig 1

Quality scores of bases sequenced by Relative content of each base

illumina sequencing at each position. Scores are In the sequences obtained by the illumina

lower at the ends, however, all scores are Sequencer. Noise at the beginning indicates high quality. the presence of a repeated sequence.

Kmer content within the fastq reads. The peaks at the end reflect a drop off in quality scores.

Presence of known adapter sequences within fastq files. The curve of the red line indicates higher presence of Illumina adapters near the end of the reads.

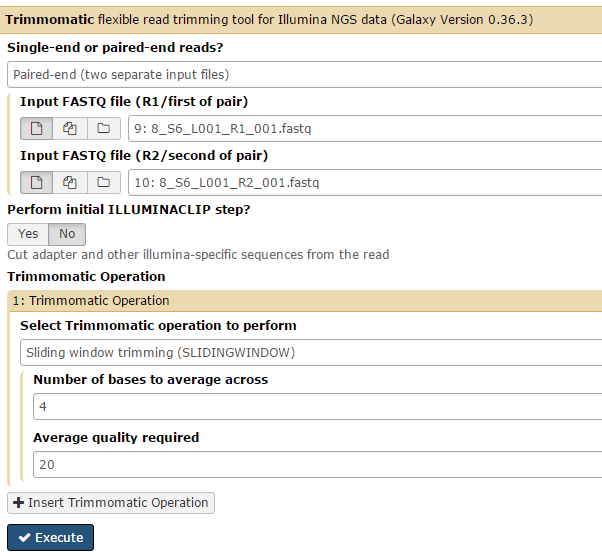
Fig 3

Fig 4

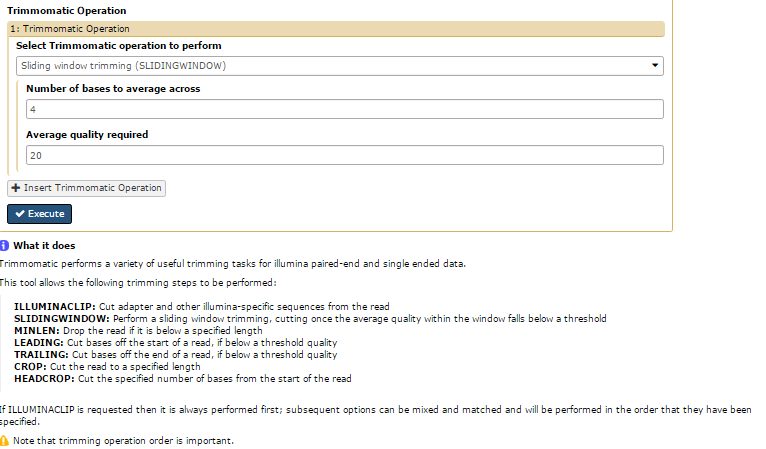
1. **QC processing with trimmomatic.**

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:



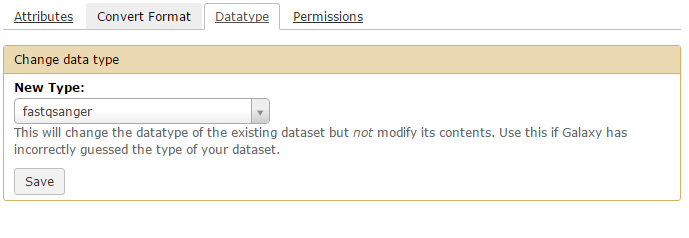
Continued:



1. **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:
   1. Load data in to Galaxy
   2. Highlight data file and click on **?** after database (bottom right corner):



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

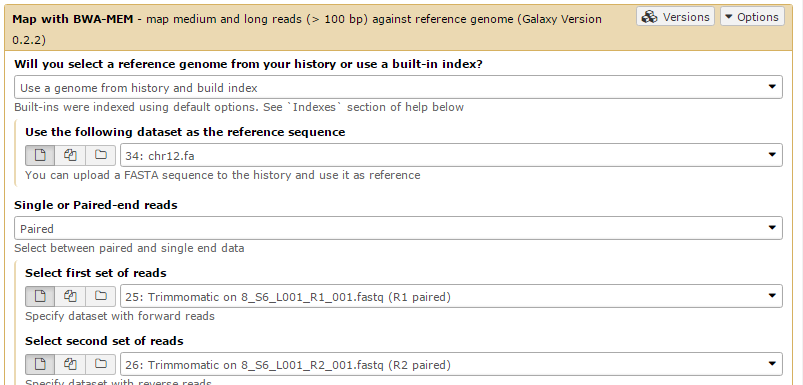


1. **Examine the updated (newly generated) fastq file** with FASTQC, using the same logic as for step 4 above.

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

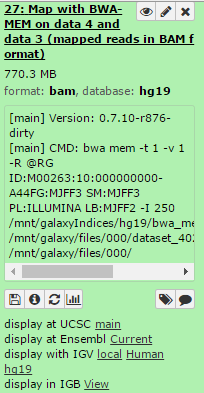
1. **Alignment to Reference** using Map with BWA-MEM

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).



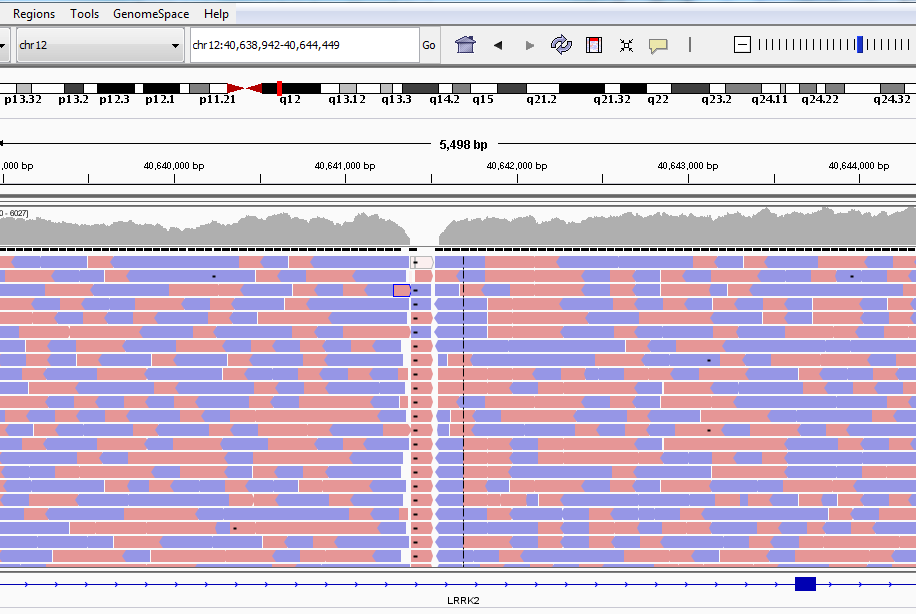
1. Post BWA-MEM

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.



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:

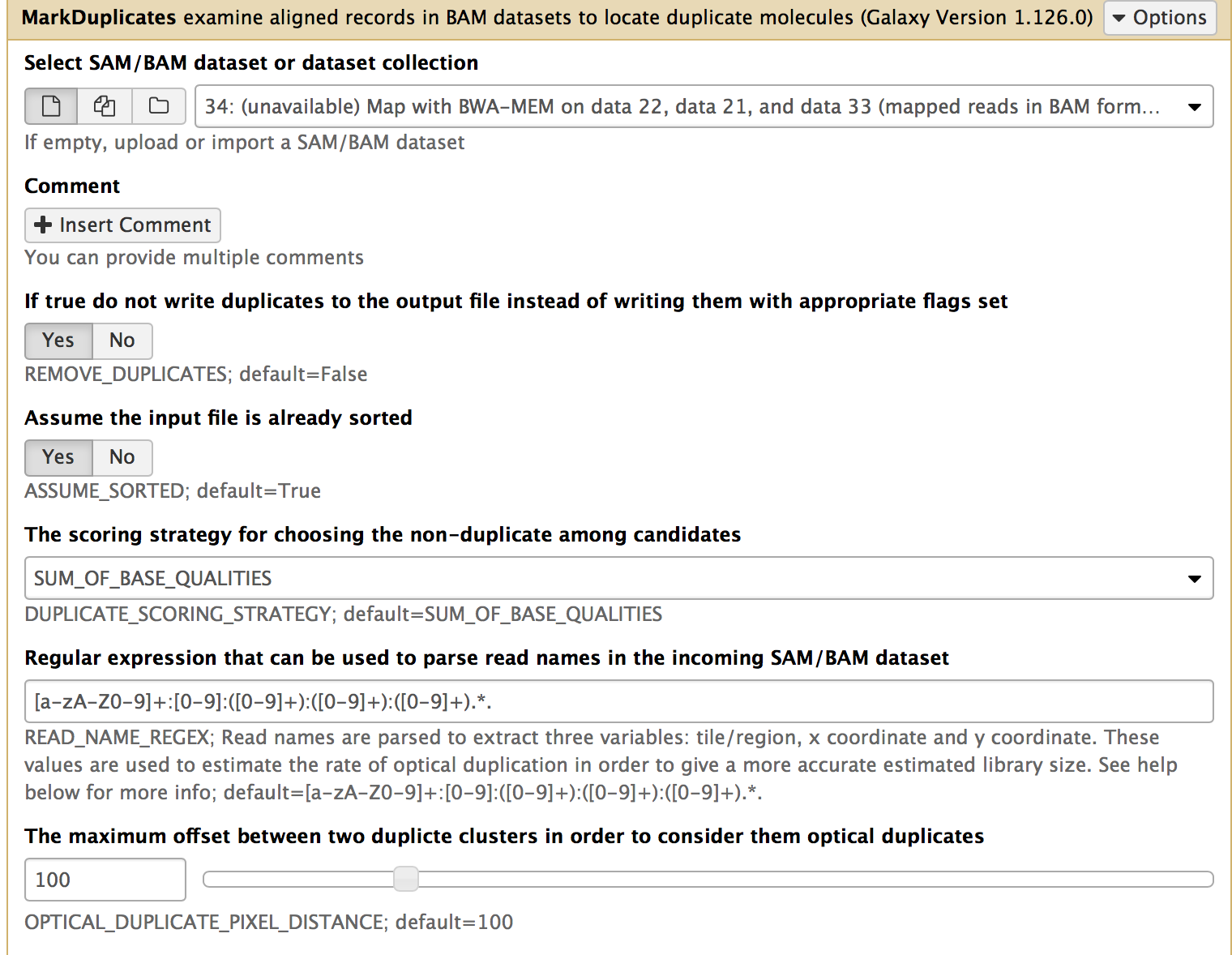


Note: IGV Visualization Tool

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).

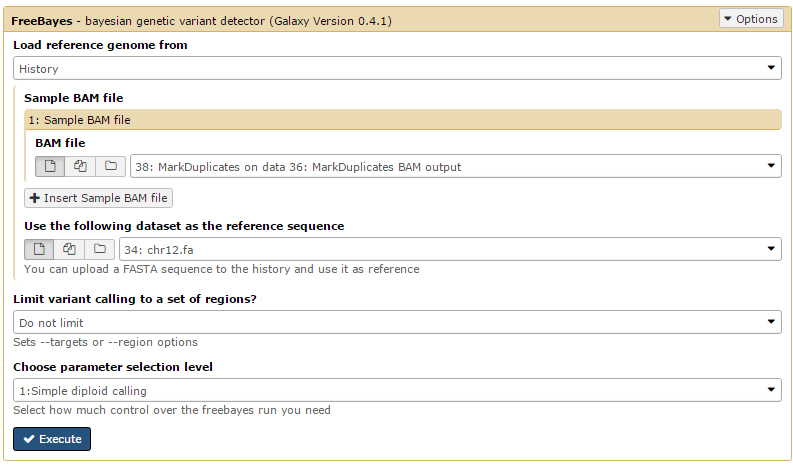
1. **Mark duplicates**

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.

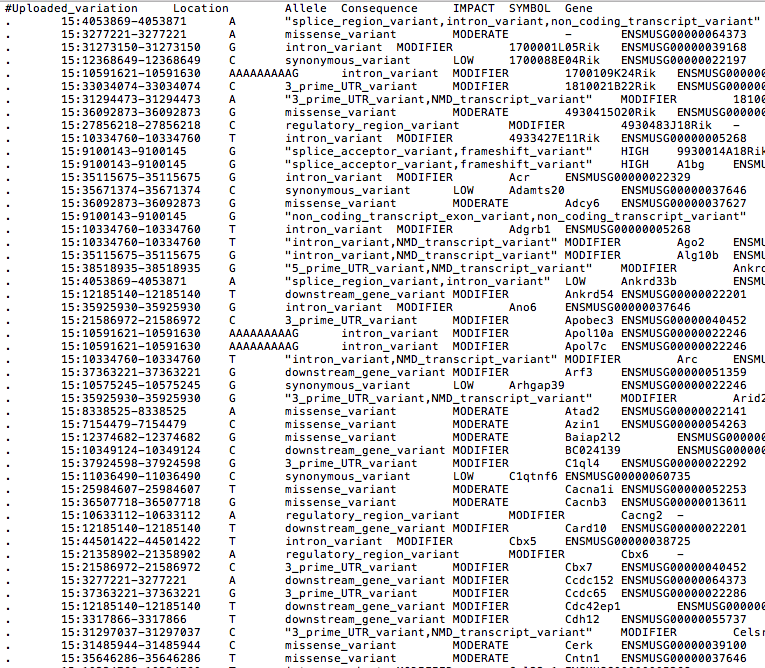


1. **Generating a Variant Call File**

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:



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.



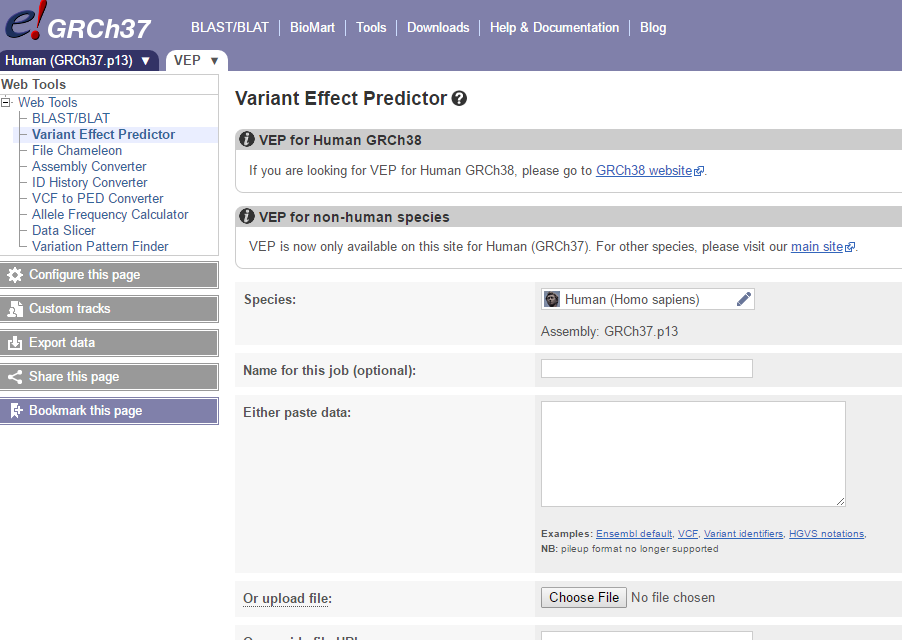
**Info:** The VCF QUAL score is a Phred Quality score scaled to the probability that a base is incorrectly called. Consequence describes the change in gene function. IMPACT type is the proposed extent of the amino shift.

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.

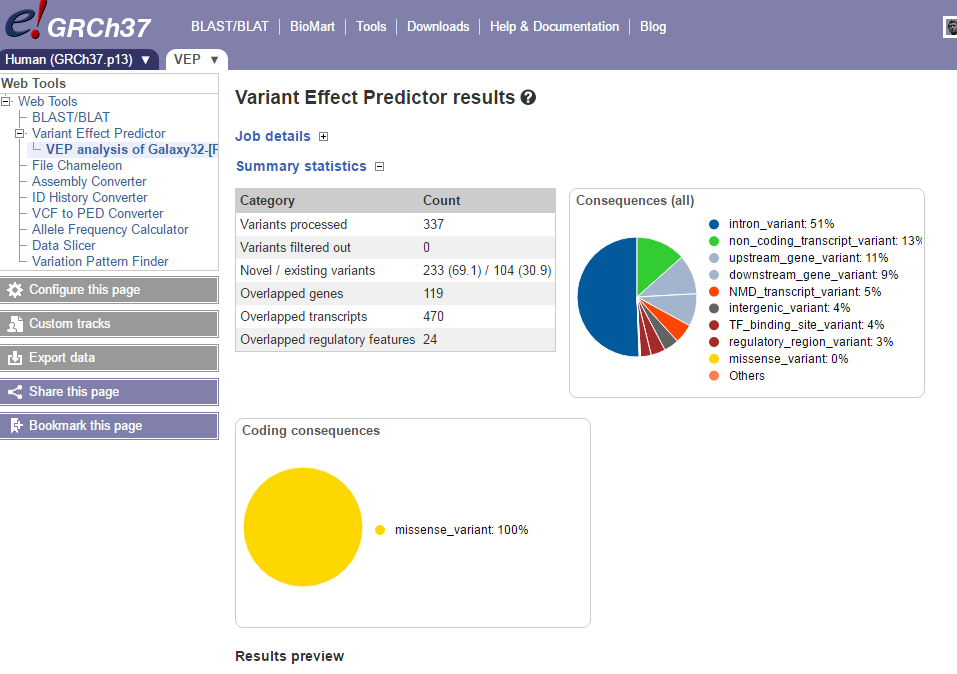
1. Variant Identification (Ensembl Variant Effect predictor, VEP)

<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.**



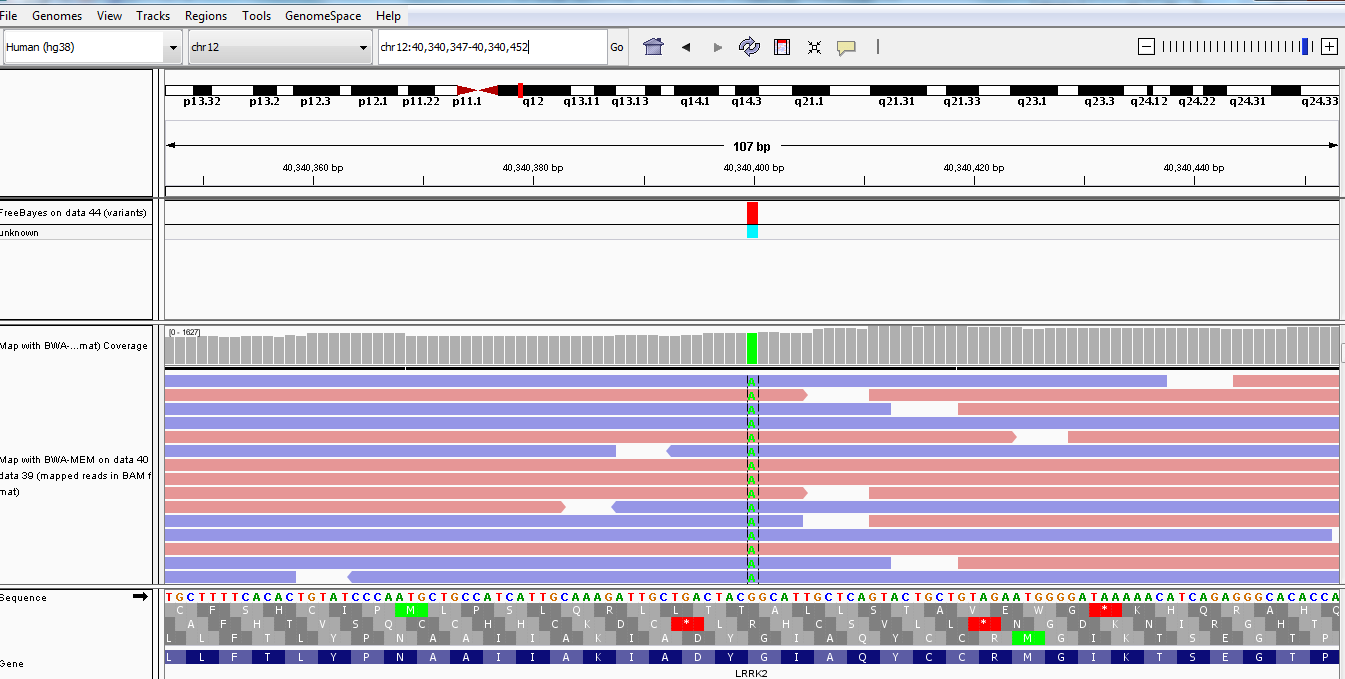
The VEP tool will look at the human genome in Ensembl and then compile a list of variants that are may be causative of a phenotype, in this case within the genetic interval entered, human chr. 12. Output in this module will look like this:



The results of VEP can be sorted consequence or gene symbol by toggling between options in the Ensembl table.

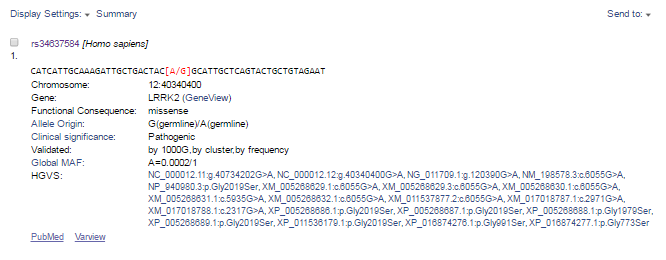
In this exercise there are numerous samples that have been provided as raw sequence. All samples are human. There are at least 3 different genotypes for the target gene: two contain likely pathogenic mutations causative of disease. One genotype represents normal, if there is such a thing as a normal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

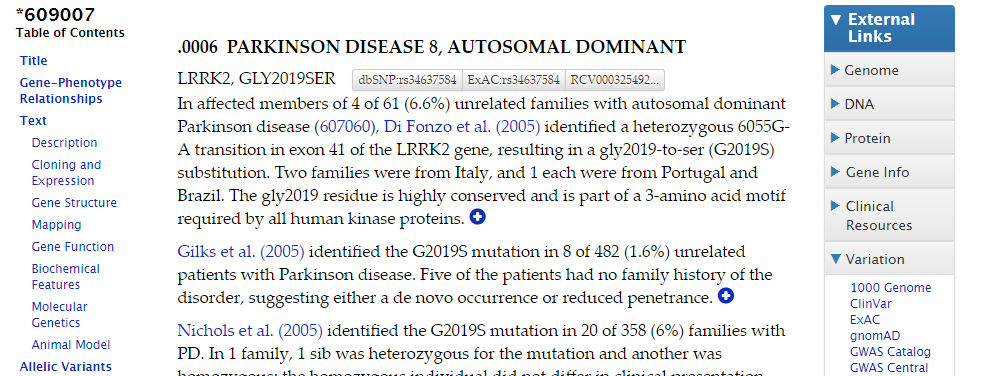
**The end point goal for students is to discover the genotype, pathogenic-likely or normal for their samples. The end point may be in tying LRRK2 to Parkinson’s if you have not told the students in advance what the gene of interest is.**



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





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.](https://www.ncbi.nlm.nih.gov/pubmed/21850687) 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.](https://www.ncbi.nlm.nih.gov/pubmed/20506312) 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.**