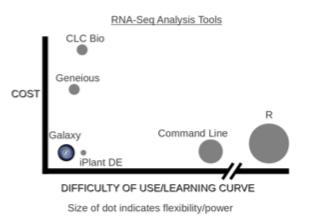
**Helen’s More Basic Instructions Based off Dr Barry Grant’s Genome Informatics Lecture.**

**Overview:** The purpose of this lab session is to introduce a set of tools used in high- throughput sequencing and the process of investigating interesting gene variance in Genomics. High-throughput sequencing is now routinely applied to gain insight into a wide range of important topics in biology and medicine [see: Soon et al. EMBO 2013].

In this lab, we will use the **Galaxy** web-based interface to a suite of bioinformatics tools for genomic sequence analysis. Galaxy is free and comparatively easy to use (see Figure 1 for a schematic comparison of some common bioinformatics RNA-Seq analysis methods).

Galaxy was originally written for genomic data analysis. However, the set of available tools has been greatly expanded over the years and Galaxy is now also used for gene expression, genome assembly, epigenomics, transcriptomics and host of other sub-disciplines in bioinformatics.

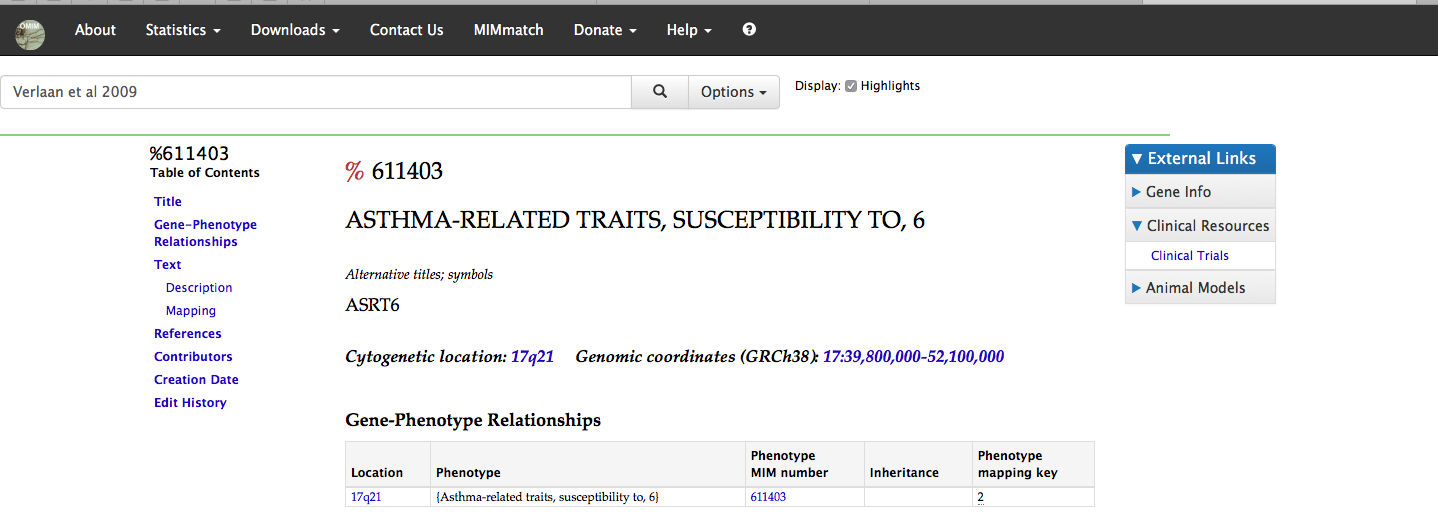
***Section 1- Identify genetic variants of interest***

1. Log In Details- Email- [Heking@uscd.edu](mailto:Heking@uscd.edu) Password- Hkdogs2\*
2. Type in h***ttps://github.com/helenistheking/RNAseq***
3. ***OMIM.org*** (an online database that focuses on the gene/phenotype relationship)

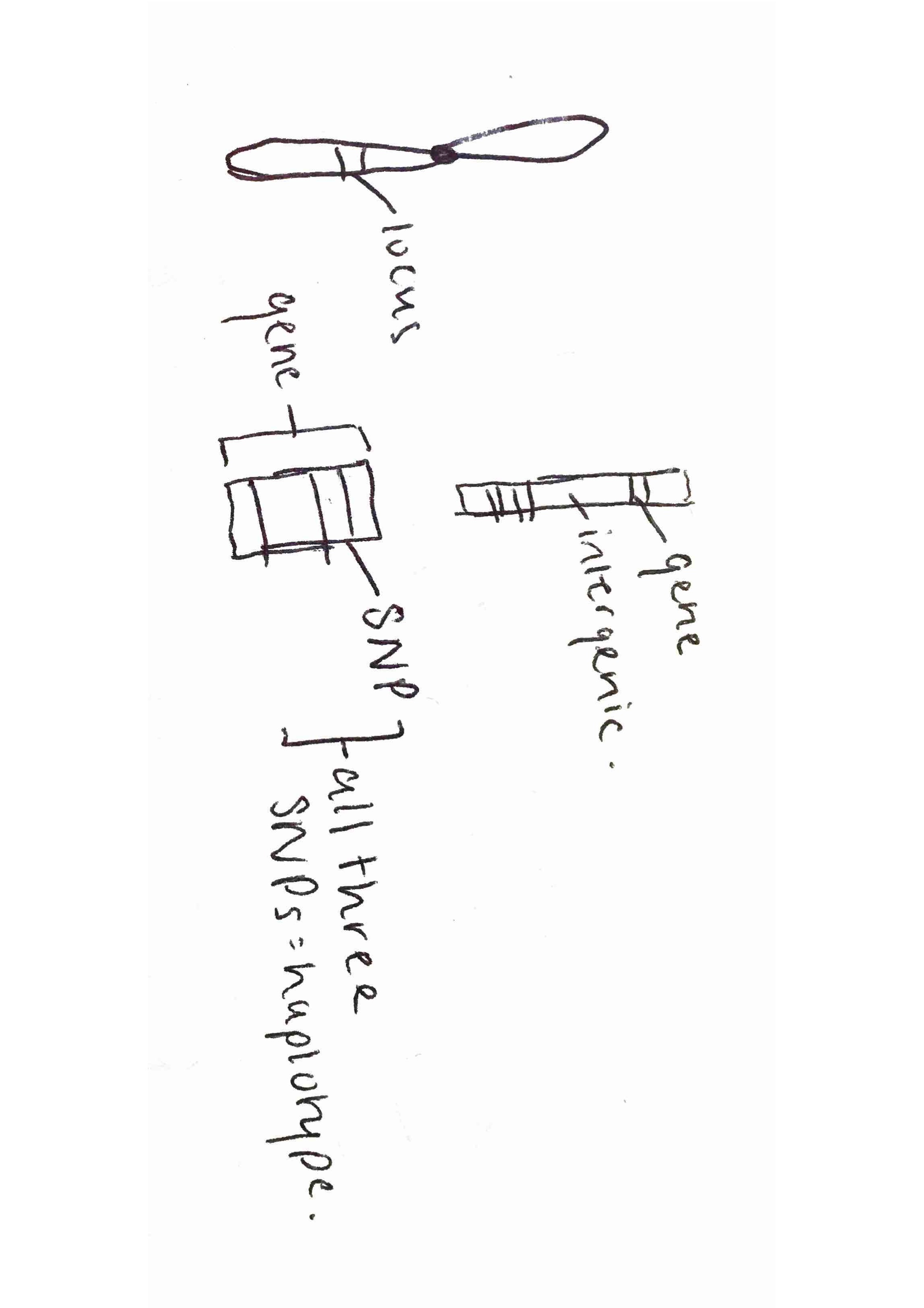
Type in Verlaan et al. 2009 into the search engine. Select the top hit.

We are researching this paper and we are trying to find genetic variants of interest.

***Q1****: What is the difference between a variant/SNP?*



1. Scroll down this page to see the papers related to this genomic locus- ASRT6. Diseases associated with this region: 17:39,800,000-52,100,000 include asthma related traits and asthma.

***Q2****: What are those 4 candidate SNPs?  
[HINT, you will may want to check the first few links of search result]* rs12936231, rs8067378, rs9303277, and rs7216389

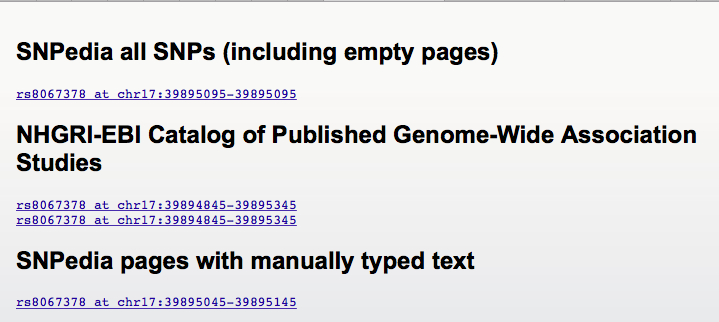
***Q3****: What three genes do these variants overlap or effect?  
[HINT, you can find the information from the ENSEMBLE page as shown in the image below with red rectangles]*ZPBP2, GSDMB, and ORMDL3

***Q4:*** *What is haplotype?*

1. Find the three genes quoted in the paper section. These are genes where the candidate

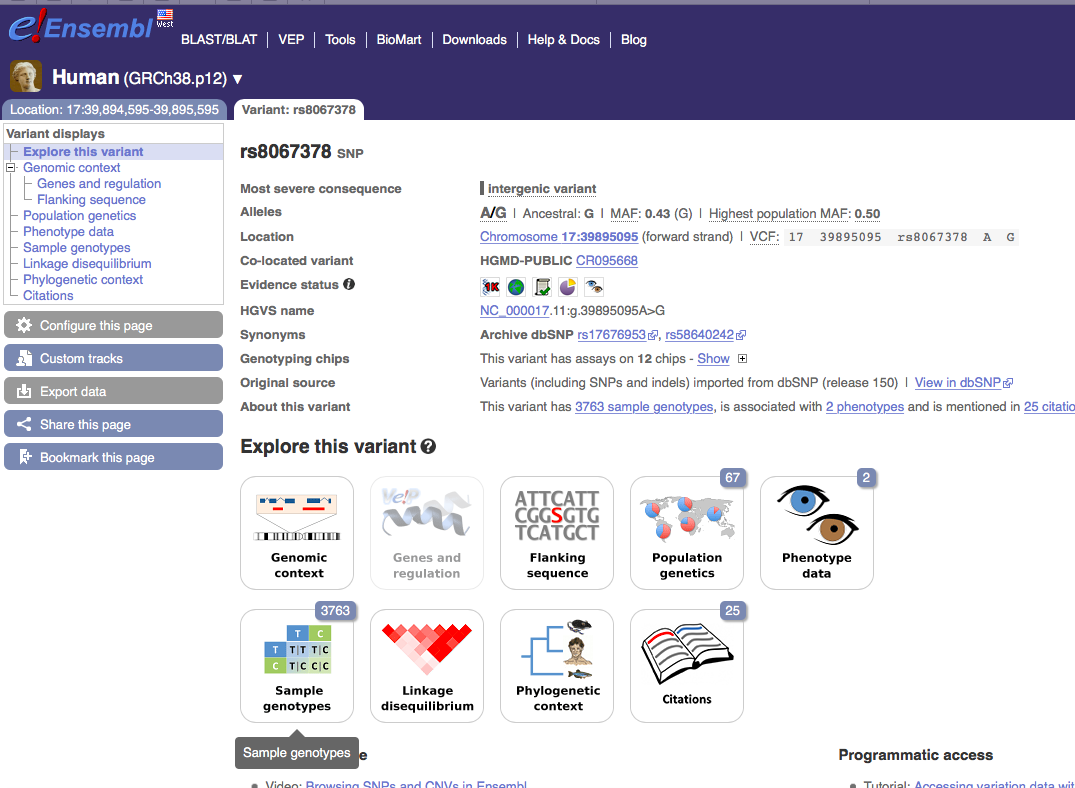
Haplotype is a set of DNA variants that tend to be inherited together.

1. ***Genome.ucsc.edu***. Choose genome browser in the top tool bar. Type in rs8067378 into the search engine on the home page.

Click on SNPedia pages with manually typed text.

Click zoom out 10x. Look at the genes around the highlighted SNP. Is there ZPBP2, GSDMB and ORMDL3?

1. Ensemble.org. Type in rs8067378. You are interested in the genotypes of the SNPs in a sample.



1. Look up on their genotypes, Mexican Ancestry in Los Angeles, California.

You should know that 14% of the population is homozygous for the G SNP. Which means it has G/G genotype. We will be using the genomic data from the individual HG00109 from this population background.

***Section 2: Initial RNA-seq Analysis***

1. Now you need to understand whether the SNP of this individual will affect gene expression. Find the raw RNAseq data of the sample above on the website found here. ***https://github.com/helenistheking/RNAseq***
2. You can see their format by clicking on the HG00109\_1.fastq and HG00109\_2.fastq files. You then need to click on Raw. Then go to the top tool bar, File, Save As and then page source. Click Don’t Append when this pops up.
3. To begin our analysis of this data we will use Galaxy on Jetstream. This is a web interface that goes through the Tuxedo protocol. Follow Barry’s instructions for accessing and logging into our very-own Galaxy Server.

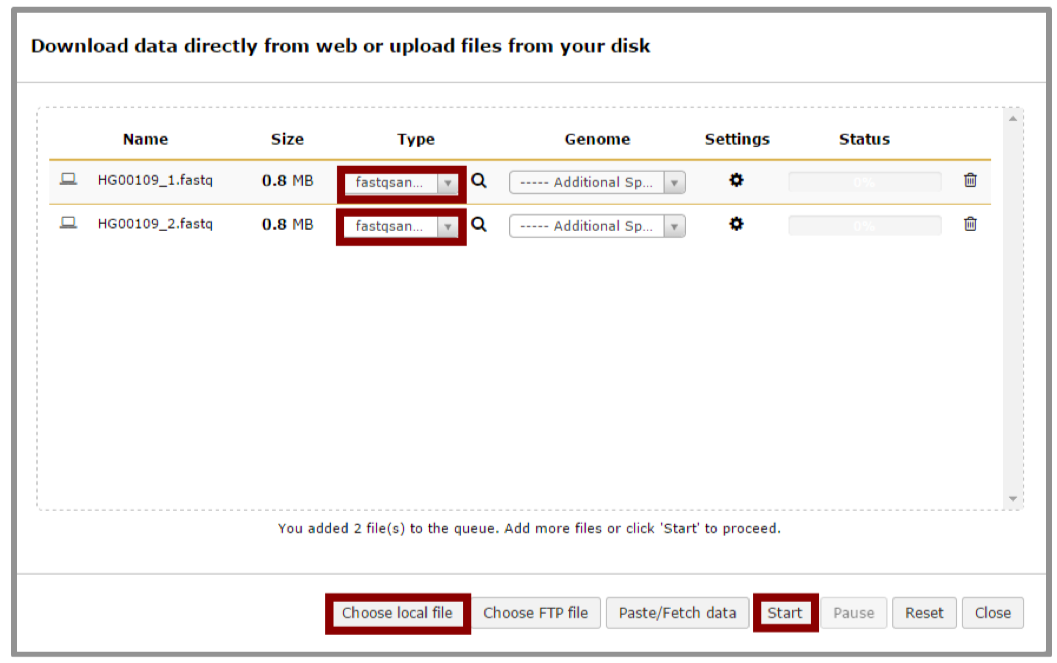
First we must create a “Jetstream service virtual machine instance. We will use these set of instructions.

[***https://bioboot.github.io/bggn213\_f17/jetstream/boot/***](https://bioboot.github.io/bggn213_f17/jetstream/boot/)

1. Once you are ready you should be able to type (or copy/paste) your assigned instance IP address into your web browser to see your very own Galaxy server. Under the User tab at the top of the page, select the Register link and follow the instructions on that page.



IPs

* 149.165.169.245
* 129.114.17.65
* 129.114.17.251
* 149.165.156.226
* 149.165.170.88
* 129.114.17.244

1. Upload our fast sequences  
   In the left side **Tools** list, click the **Get Data > Upload** File link to upload our sequence files for analysis. You can load them from your own local laptop (with chose local file option) or more simply upload them via the URL from above (with the paste/fetch data option i.e. No need to download them to your computer first - this is often useful when dealing with very large files). Be careful of the file type you upload. Tophat2 only takes **fastqsanger** file format. So, You need to choose fastqsanger for the upload Type. (http://en.wikipedia.org/ wiki/FASTQ\_format)



1. Now, you can check the data on the right panel. When they are colored gray they are still uploading and when they are green they are uploaded. Clicking in the name and various icons will provide more information to help you answer question 5 below.

***Q5***: How many sequences are there in the first file? What is the file size and format of the data? Make sure the format is fastqsanger here!  
[HINT, you can check the fastq format wiki for more information]

1. You should understand the reads a bit before analyzing them in detail. Run a quality control check with the FastQC tool on your data using the “NGS: QC and manipulation” > FastQC Read Quality reports.

FastQC performs several quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

Often, it is useful to trim reads to remove base positions that have a low median (or bottom quartile) score.

1. After running the FastQC program, you will get a FastQC Report both as a Webpage and Raw Data. Click on eye icon to view each version.

***Q6***: What is the GC content and sequence length of the second fastq file? [HINT, you may check “Basic Statistics”]

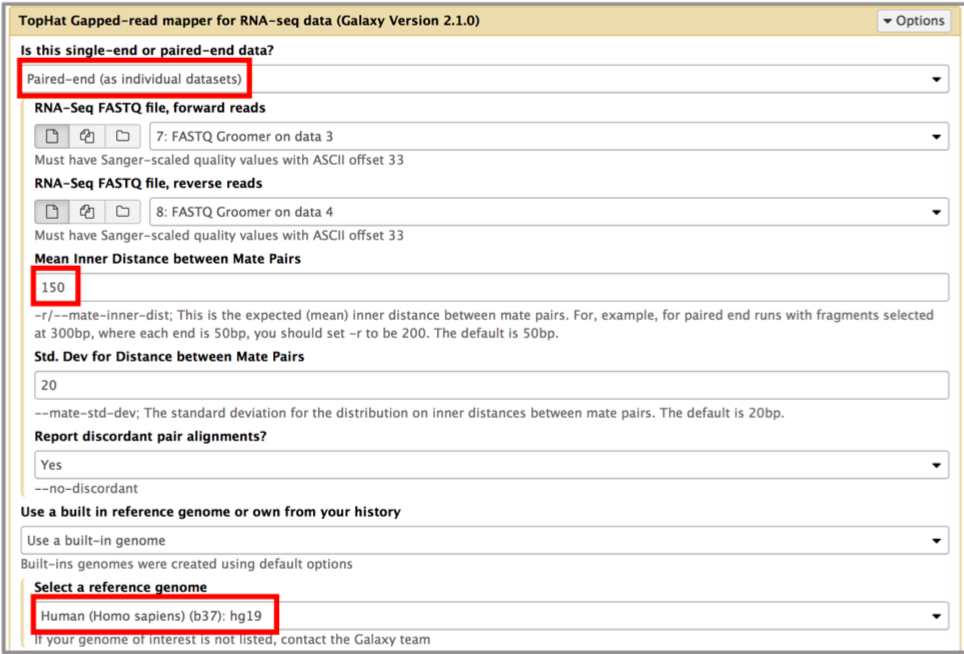
***Q7***: How about per base sequence quality? Does any base have a mean quality score below 20? [HINT, blue line is the mean quality score and for this exercise, assume a median quality score of below 20 to be unusable. Given this criterion, is trimming needed for the dataset?]

***Section 3: Mapping RNA-Seq reads to genome***

The next step is mapping the processed reads to the genome. The major challenge when mapping RNA-seq reads is that the reads, because they come from RNA, often cross splice junction boundaries; splice junctions are not present in a genome's sequence, and hence we use a mapper such as Tophat (http://ccb.jhu.edu/software/tophat) that is designed to map RNA-seq reads.

***Q8:*** What is splicing?

1. Use the NGS: RNA Analysis > Tophat tool to map RNA-seq reads to the hg19 build of the Human reference genome.

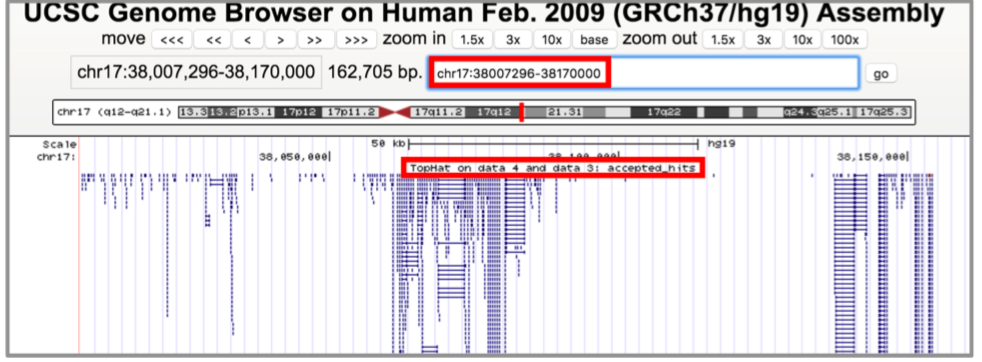


*Note: Our input data is pair-end data. For Tophat in Galaxy, you need to set paired-end as your input type and then provide the forward read file and reverse read file. Because the reads are paired, you'll also need to set mean inner distance between pairs; this is the average distance in basepairs between reads. Use a mean inner distance of 150 for our data as this was the fragment length from the experimental library preparation step. See the red rectangles in the image below for details of the settings to change.*

1. The calculation may take some time. There will eventually be five outputs: accepted\_hits, insertions, deletions, splice junctions and an alignment summary. We will focus only on the alignment **summary** and the **accepted hits** files for this exercise.

1. The accepted hits file is in BAM format, which is binary version of the human readable SAM format. To inspect these results we will convert the BAM file to SAM format using **NGS: SAMtools > BAM-to-SAM** tool. Once converted click the eye icon to view within galaxy. Note there is lots of metadata in the SAM file (lines beginning with @). After this is our alignment section, which includes details of the chromosome locations that our reads have been aligned to.
2. Once complete select and expand the accepted hits file in your history sidebar. Then Click on the “display at UCSC main” link.

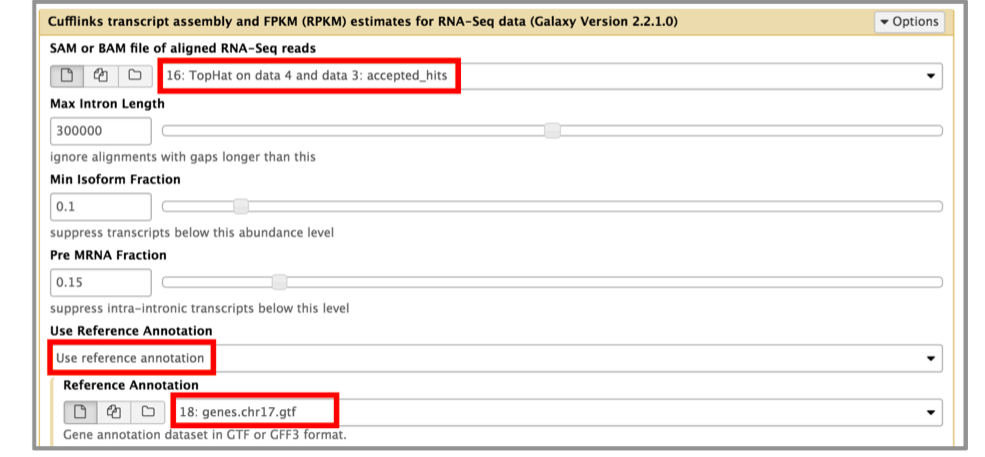
This will load your TopHat results as a custom track on the UCSC Genome Browser. You can then click on the custom track and change the display mode from Dense to Full and enter the region chr17:38007296-38170000 into the text box to see the pile-up of aligned sequence reads in this location.



***Q9:*** Where are most the accepted hits located?  
[HINT, you can view the SAM version of your accepted hits file in galaxy and also use the UCSC Genome Browser via following the galaxy provided link and focusing on particular regions as described above]

1. With alignment result from TopHat, we can now calculate gene expression with the NGS: RNA Analysis > Cufflinks tool. Before running Cufflinks, you should upload the reference annotation file “gene\_chr17.gtf” (available from into the workspace of Galaxy first. This is a tab-delimited text file obtained from UCSC describing  
   genomic features (locations of exons, stop\_codons, CDS, etc for our region of chromosome 17). Examine this file in galaxy before use).

The following figure shows the parameters you need to change when running cufflinks.



***Q10***: Is there any interesting gene around that area?  
[HINT, you can find genes around accepted hits in either the UCSC Genome Browser or IGV - depending on which browser you prefer]

Note that genes will have more reads mapped in a sample with high coverage than one with low read coverage – 2x depth ≈ 2x expression. Also longer genes will have more reads mapped than shorter genes – 2x length ≈ 2x more reads. Normalization allows us to compare across genes within a sample and between samples (e.g. WT and Mutant etc.)

***Q11***: Cufflinks again produces multiple output files that you can inspect from your right- hand-side galaxy history. From the “gene expression” output, what is the FPKM for the ORMDL3 gene? What are the other genes with above zero FPKM values?  
136853

Subsequent steps in a typical RNA-Seq analysis would use a tool such as DESeq2 (an R package) to set up a differential expression analysis to essentially compare the counts of each transcript/gene between different samples (including replicates) to assign a probability to the observed counts being generated if the gene is NOT differentially expressed between conditions. We will skip this step and move onto a population scale analysis to complete the circle back to our childhood asthma associated SNPs.

***Section 4: Population Scale Analysis***  
One sample is obviously not enough to know what is happening in a population. You are interested in assessing genetic differences on a population scale. So, you processed about ~230 samples and did the normalization on a genome level. Now, you want to find whether there is any association of the 4 asthma-associated SNPs (rs8067378...) on ORMDL3 expression. Find here (<https://github.com/helenistheking/RNAseq/blob/master/rs8067378_ENSG00000172057.6)>

The first column is sample name, the second column is genotype and the third column is the expression value.

1. Read this file into Excel, by doing copy and special paste (text). Then use text to columns, select delimitated and then space. This should separate data out into different columns.
2. Determine the sample size for each genotype and their corresponding median expression levels for each of these genotypes.
3. Generate a boxplot with a box per genotype, what could you infer from the relative expression value between A/A and G/G displayed in this plot? Does the SNP effect the expression of ORMDL3?

Reference:  
- All data files can also be found at: https://bioboot.github.io/bimm143\_W18/lectures/#13

Section 2 was adapted from https://usegalaxy.org/u/jeremy/p/galaxy-rna-seq-analysis- exercise.

- Verlaan, et al. Allele-specific chromatin remodeling in the ZPBP2/ GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease. Am. J. Hum. Genet. 85: 377-393, 2009.