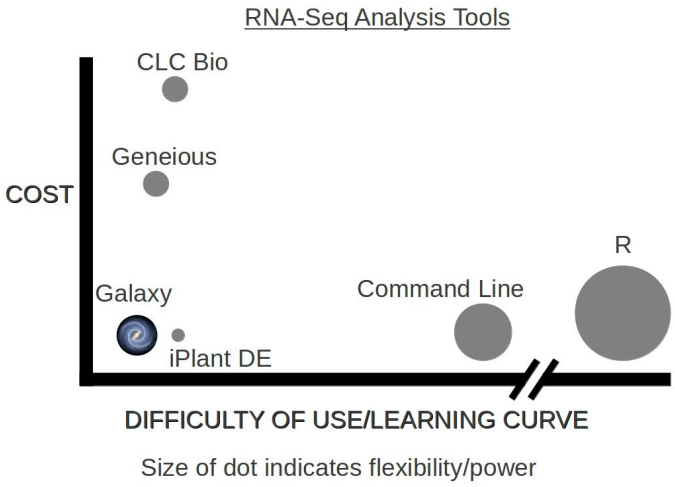
BIOINF525: INTRODUCTION TO BIOINFORMATICS LAB SESSION 4
Genome Informatics
Dr. Ryan E. Mills & Hongyang Li
Feb 2016

Overview: The purpose of this lab session is to cover a set of tools used in high-throughput sequencing and the process of investigating interesting gene variance in Genomics.

Introduction

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 on Ctools].

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).

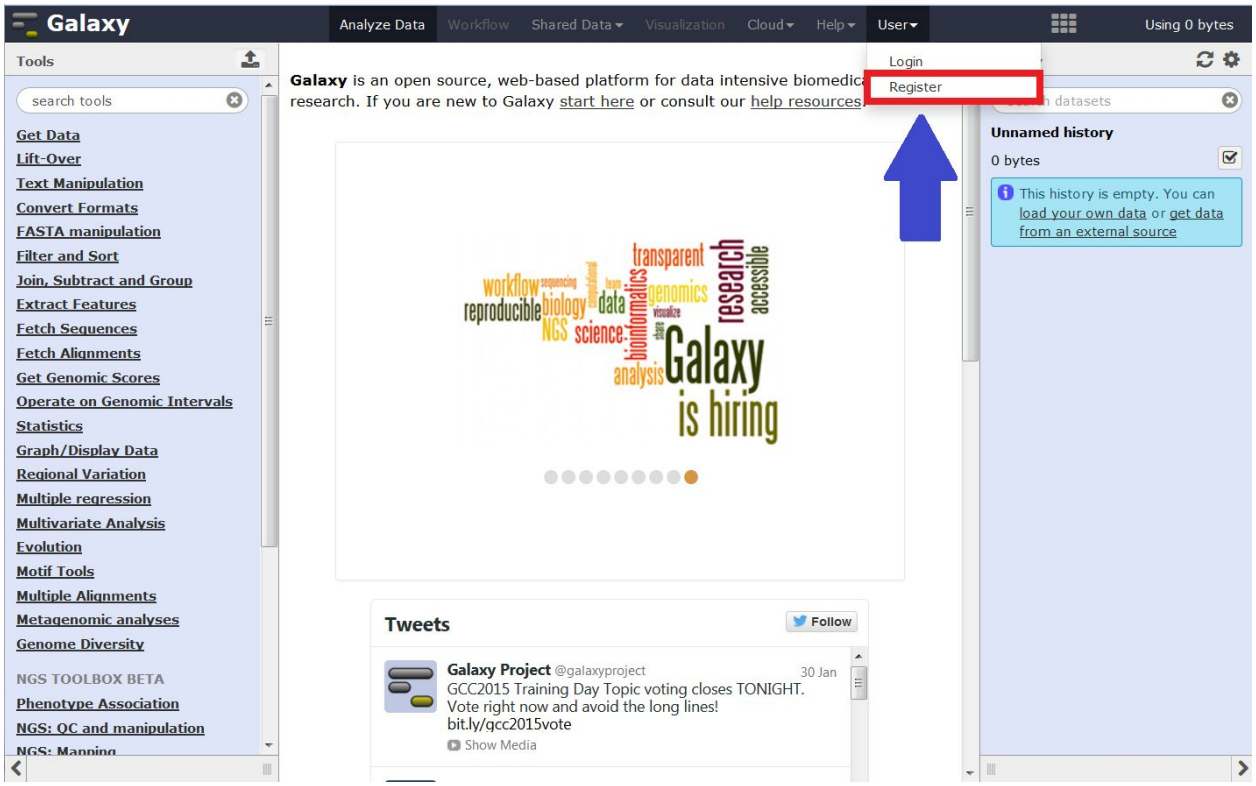


Galaxy was originally written for genomic data analysis. However, the set of available tools has been greatly expanded over the years and Galaxy is now also used for gene expression, genome assembly, proteomics, epigenomics, transcriptomics and host of other sub-disciplines in bioinformatics.

Registering for a Galaxy account

First create an account on the main public Galaxy portal @ https://main.g2.bx.psu.edu/

Under the **User tab** at the top of the page, select the **Register** link and follow the instructions on that page.



This will only take a moment, and will allow all the work that you do to persist between sessions and allow you to name, save, share, and publish Galaxy histories, workflows, datasets and pages.

***Section 1*: Find the interesting genome variance**

There are a number of gene variants associated with childhood asthma. A study from Verlaan et al. (2009) shows that 4 candidate SNPs demonstrate significant evidence for association. You want to find what they are in OMIM (http://www.omim.org)

Q1: What are those 4 candidate SNPs?

[HINT, you may want to check the first few links of search result]

rs12936231, rs8067378, rs9303277, and rs7216389

Q2: What are three genes be affected?

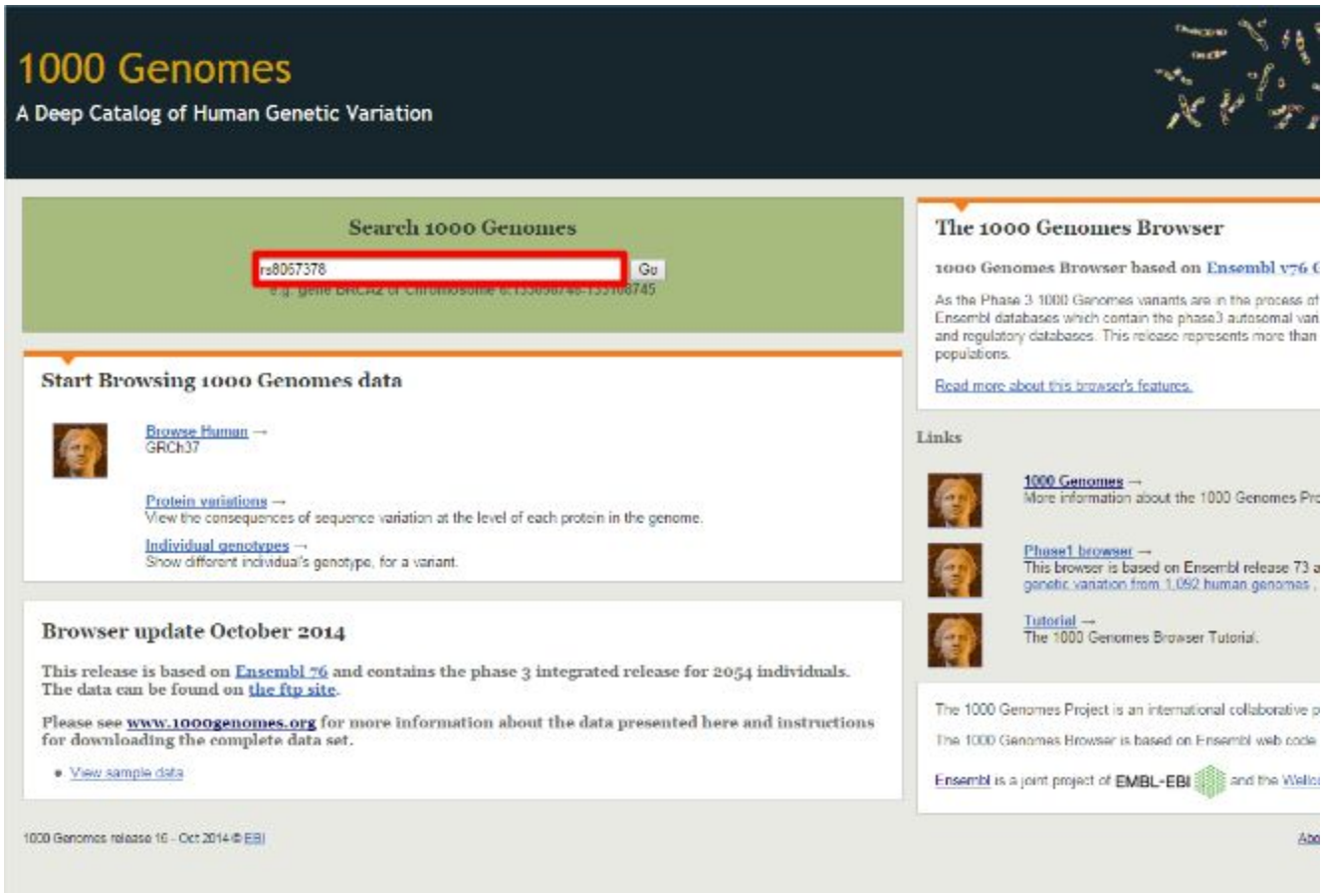
ZPBP2, GSDMB, and ORMDL3

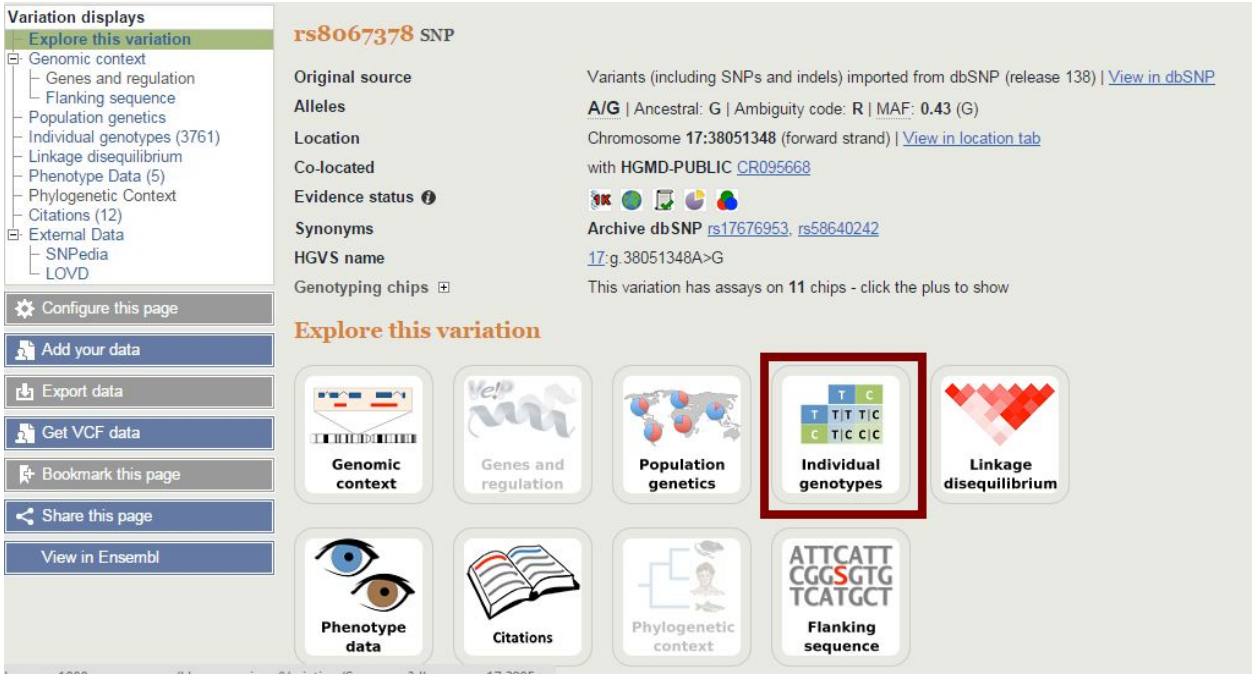
Now, you want to know the location of SNPs and genes on genome. You can find the information on UCSC genome browser (http://genome.ucsc.edu) or Ensembl genome browser (http://www.ensembl.org).

Q3: What is the location of rs8067378? What are the different alleles for rs8067378? [HINT, you may search in genome browser]

Q4: What are the downstream genes for rs8067378? Any genes named ZPBP2, GSDMB, and ORMDL3?

You are interested in the genotypes of these SNPs in a particular sample (**HG00109**). Go to the 1000 genomes browser (http://browser.1000genomes.org/) and look up their genotypes.



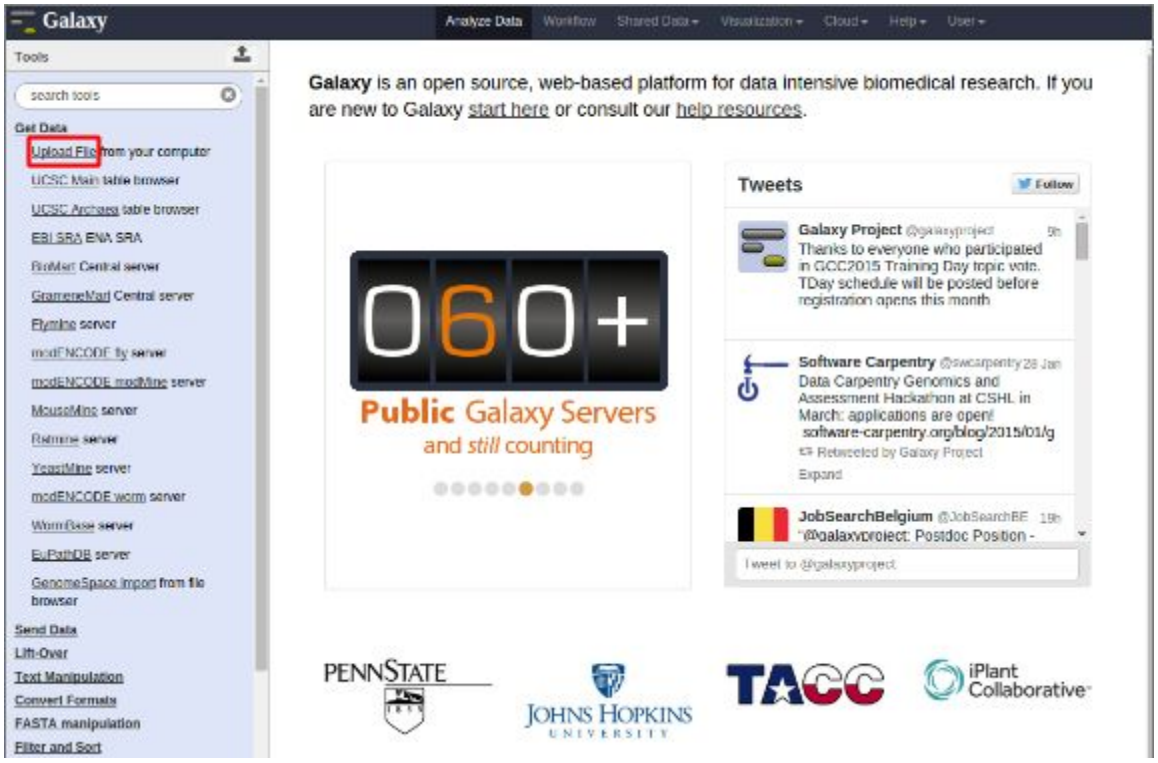


*Q5: What are the individual genotypes for the particular sample (**HG00109**)?*
[HINT: use 1000 genomes browser to look up genotypes]

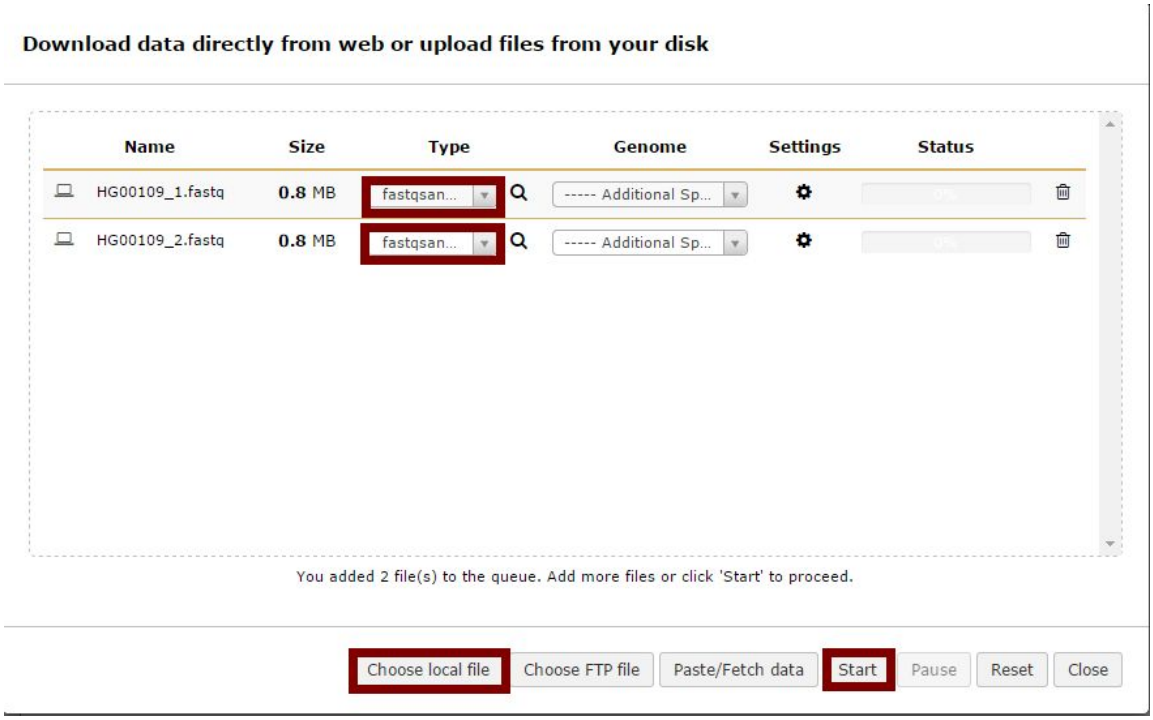
***Section 2*: RNA-Seq analysis**

Now, you want to understand whether the SNP will affect the expression of the gene.

You find the RNA-Seq data of one sample on CTools (HG00109_1.fastq, HG00109_2.fastq). However, this is the raw sequence fastq file. More detail about fastq format (http://en.wikipedia.org/wiki/FASTQ_format). To have a quick analysis of the data, you download and upload the file to Galaxy.



Be careful of the file type. Tophat2 only takes fastqsanger file format. So, You need to choose **fastqsanger** for the Type.



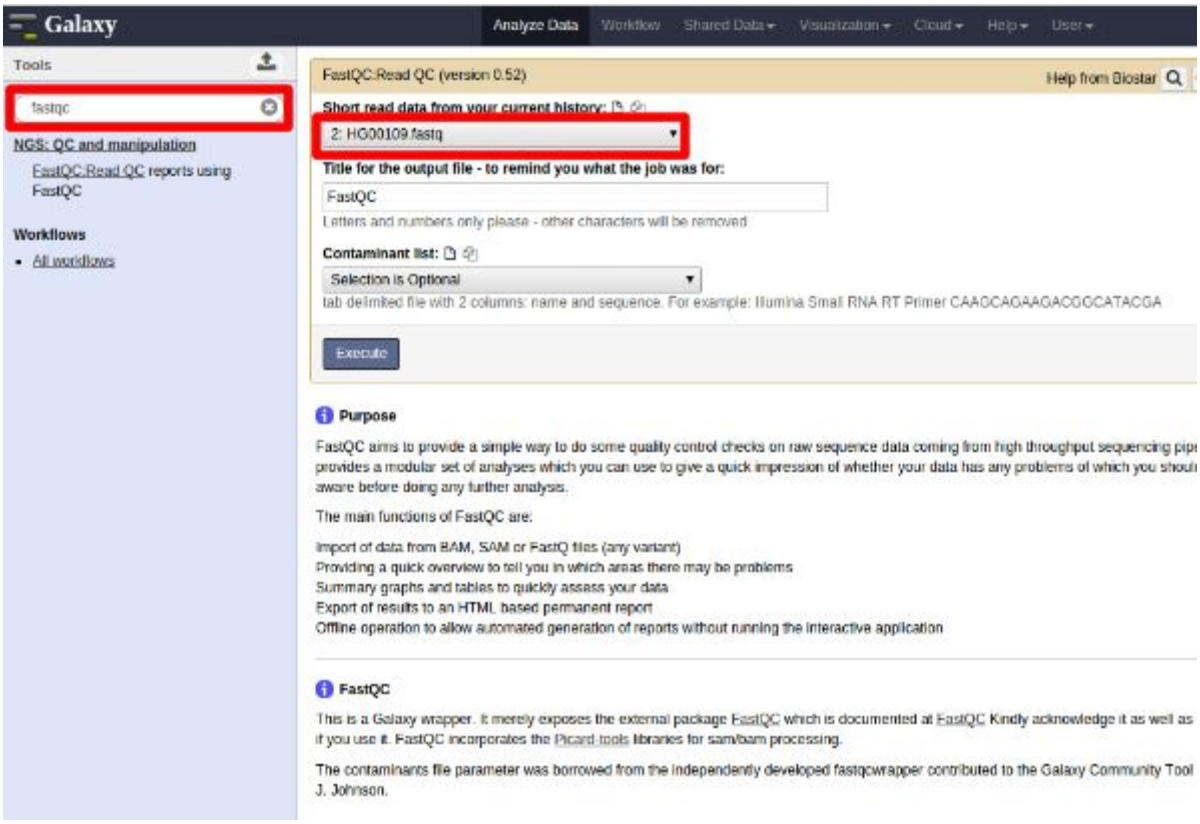
Now, you can check the data on the right panel. So, you will have better understanding about what each column/row represent.

Q6: What is the size and format of the data?

Q7: What does the first, second and fourth row represent?
[HINT, you can check the fastq format wiki for more information]

Q8: Does the first sequence have good quality?
[HINT, what is the quality score for each nucleotide?]

Quality Control
You should understand the reads a bit before analyzing them. Run a quality control check on your data using the [NGS: QC and manipulation >] FASTQC tool. Often, it is useful to trim reads to remove base positions that have a low median (or bottom quartile) score.



After running the FastQC program, you will get a FastQC Report. Click on the red box, the report will show in the center of data browser.

Q9: What is the GC content of and format of the fastq file?
[HINT, you may check "Basic Statistics"]

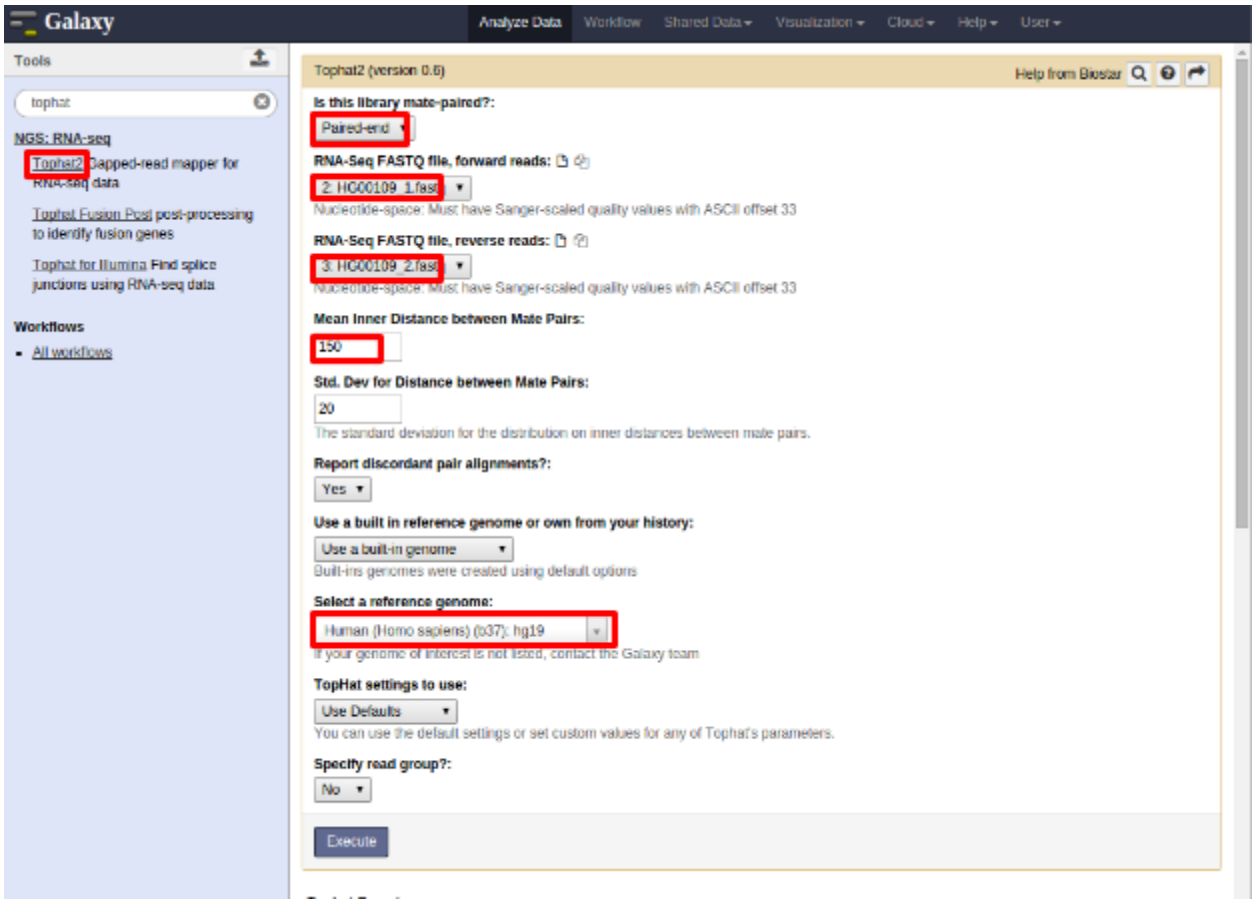
Q10: How about per base sequence quality? Does any base have a median quality score below 20?
[HINT, blue line is the median quality score.]

Q11: For this exercise, assume a median quality score of below 20 to be unusable. Given this criterion, is trimming needed for the datasets?

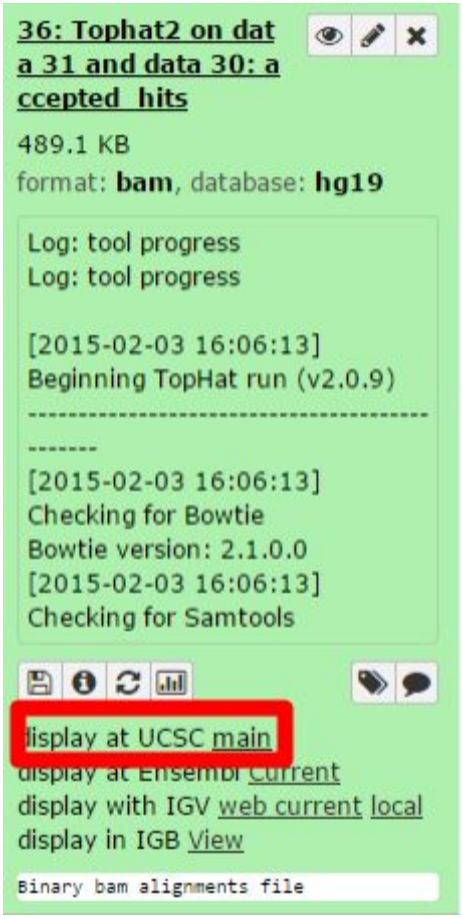
Map reads to genome

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 typical NGS mappers such as **Bowtie** (http://bowtie-bio.sourceforge.net/index.shtml) and **BWA** (http://bio-bwa.sourceforge.net/) are not ideal without modifying the genome sequence. Instead, it is better to use a mapper such as **Tophat** (http://ccb.jhu.edu/software/tophat) that is designed to map RNA-seq reads.

Use the [NGS: RNA Analysis >] Tophat tool to map RNA-seq reads to the hg19 build. The data you got is pair-end data. In Galaxy, you need to set forward read file and reverse read file. Because the reads are paired, you'll need to set mean inner distance between pairs; this is the average distance in basepairs between reads, not the total insert/fragment size. Use a mean inner distance of 150 for our data.



There will be four outputs: accepted_hits, insertions, deletions and splice junctions. You can visualize the accepted_hits on your favorite genome browser, like UCSC Genome Browser.

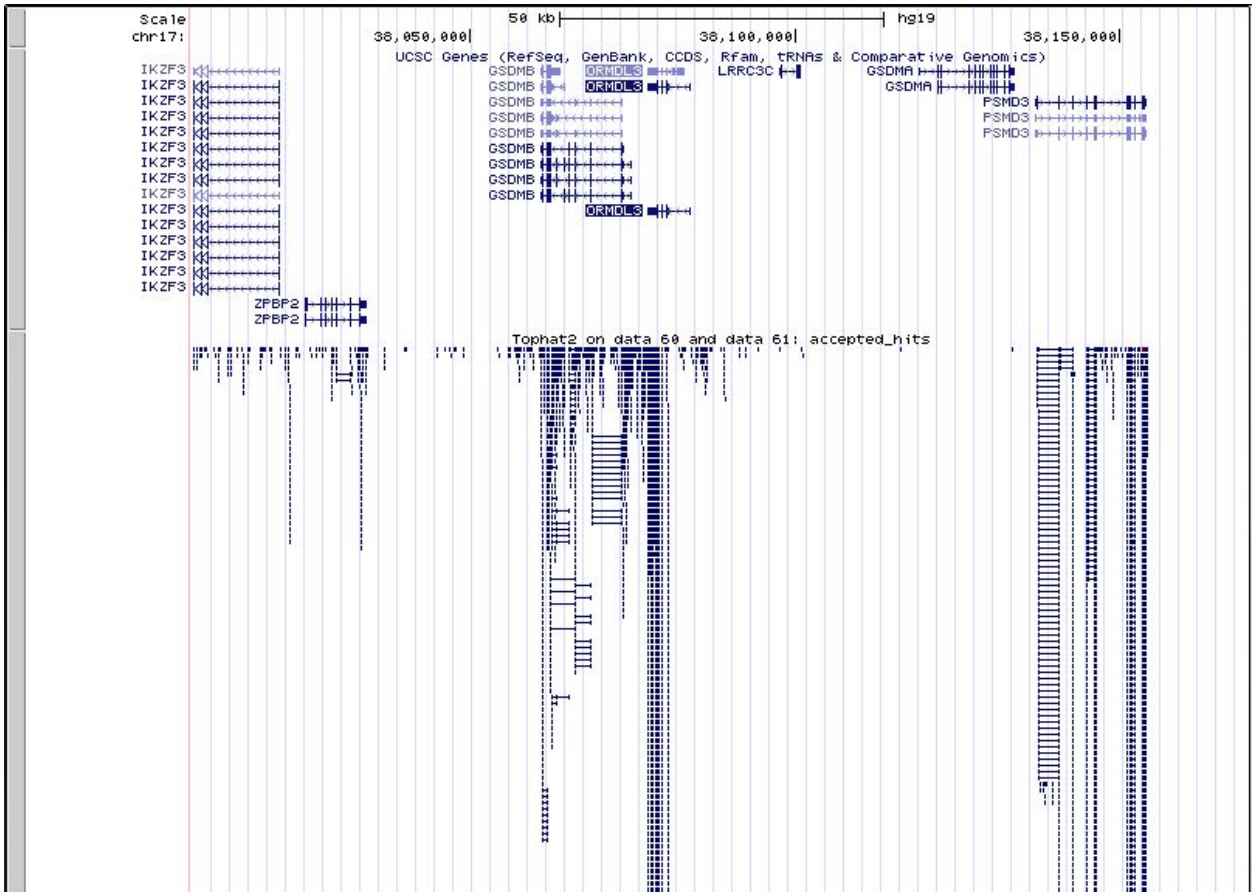


Q12: What is the first entry of splice junctions? Where is the junction located?
[HINT, check the output of Tophat "splice junctions"]

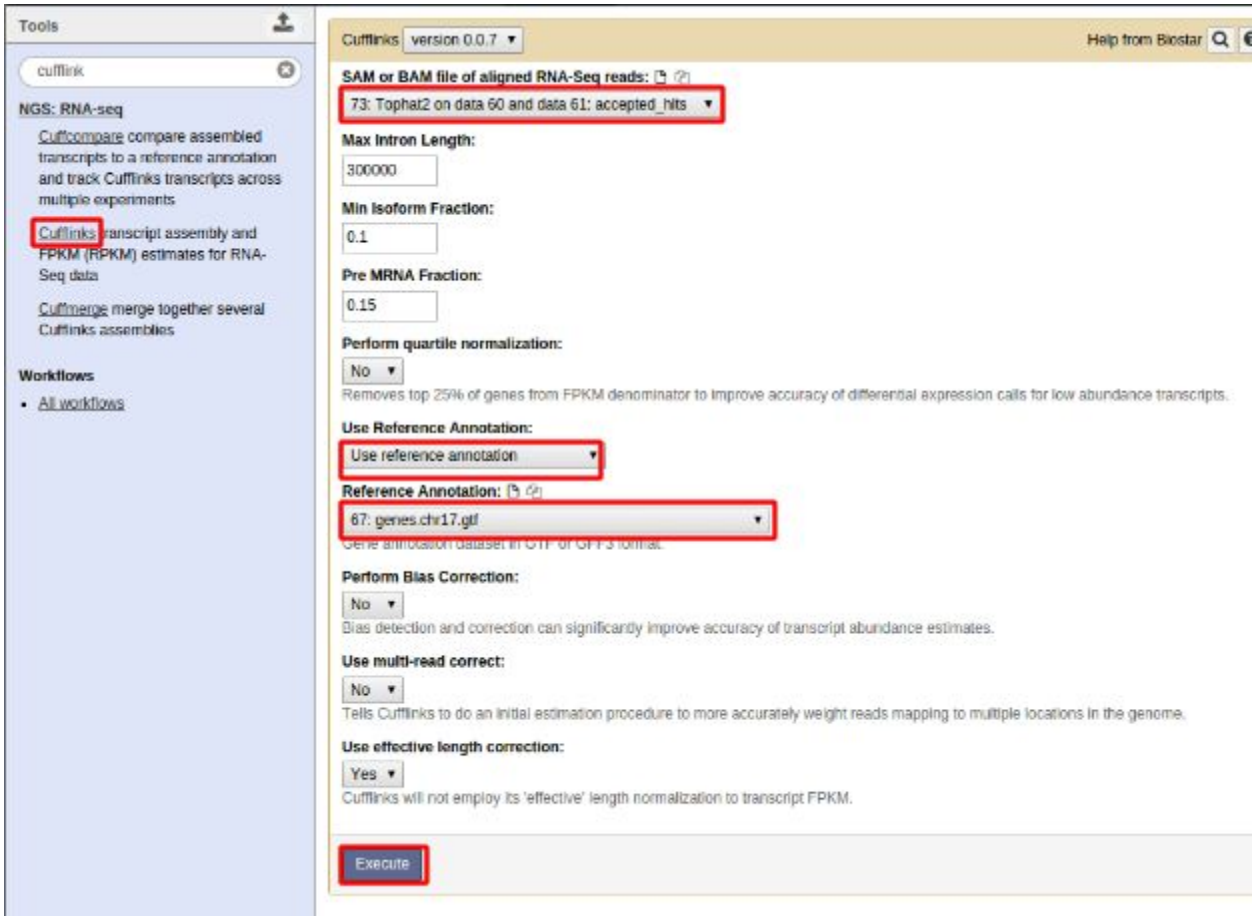
Q13: Where are most the hits located?
[HINT, you can view the accepted hits in UCSC Genome Browser, and search region: ***chr17:38007296-38170000****]*

Q14: Following Q13, is there any interesting gene around that area?
[HINT, you can find genes around accepted hits in UCSC Genome Browser]

The mapped reads on UCSC Genome Browser:



With alignment result from TopHat, you can calculate the gene expression by Cufflinks (http://cole-trapnell-lab.github.io/cufflinks/). Before running Cufflinks, you should upload the reference annotation file “gene_chr17.gtf” (on CTools also) into the workspace of Galaxy first. The following figure shows what parameters you need to change.



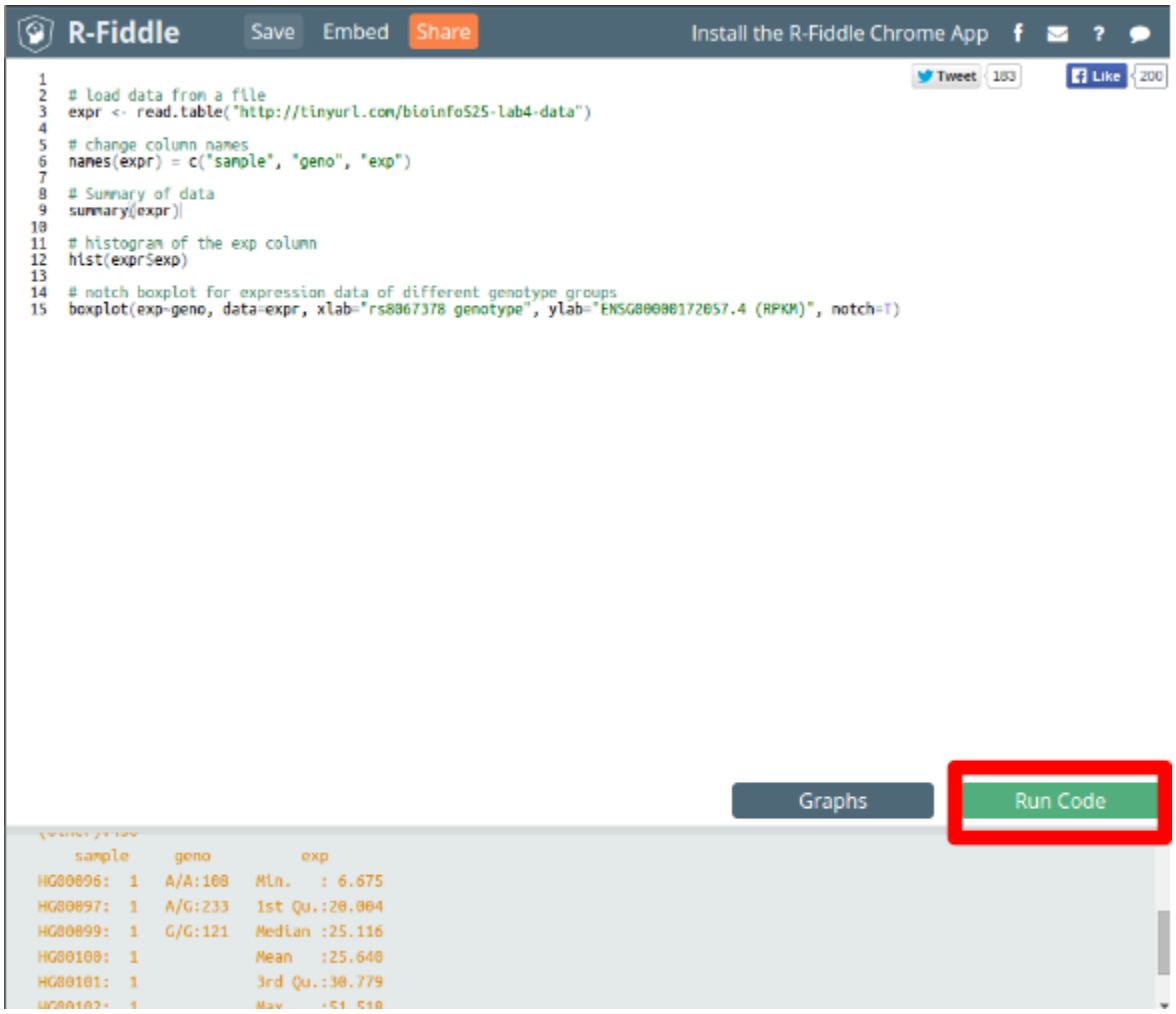
Q15: What is the FPKM for the gene from Q13?
136853

***Section 3*: Population Scale Analysis**

One sample is 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 genome level. Now, you want to find whether there is any association of the 4 asthma-associated SNPs (**rs8067378…)** on **ORMDL3** expression**.**

This is the final file you got (http://tinyurl.com/bioinfo525-lab4-data). The first column is sample name, the second column is genotype and the third column is the expression value.

You wrote some R code to get an overview about the data. The R code is displayed here (http://bit.ly/1wXl4Eo). (We will introduce R in the next lab)



Q16: What is the sample size for A/A?
[HINT, the lower section of the browser contains the output for your R code. "geno" is the column for genotype sample size]

Q17: What is the median expression value for A/A and G/G?
[HINT, you can find the value from the up right graphs. The graph is a boxplot, which you can learn more from here (http://en.wikipedia.org/wiki/Box_plot)]

Q18: What could you infer from the relative expression value between A/A and G/G? Does the SNP effect the expression of ORMDL3?

Q19: What one part of this lab or associated lecture material is still confusing? If appropriate please also indicate the question number from this lab instruction pdf and answer the question in the following anonymous form: http://tinyurl.com/bioinfo525-lab4

All data files can also be found at: https://github.com/ajing/Bioinfo525_lab4

You can also search in “Published Workflow” for “Bioinfo525_lab4”, which contains the second section of the lab.

Reference:

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.

The second section of the lab is adapted from https://usegalaxy.org/u/jeremy/p/galaxy-rna-seq-analysis-exercise .