Class12

James Woolley A16440072

```
expr <- read.table("genomicslab.txt")
head(expr)
```

```
   sample geno      exp
1 HG00367  A/G 28.96038
2 NA20768  A/G 20.24449
3 HG00361  A/A 31.32628
4 HG00135  A/A 34.11169
5 NA18870  G/G 18.25141
6 NA11993  A/A 32.89721
```

```
nrow(expr)
```

```
[1] 462
```

```
table(expr$geno)
```

```
A/A A/G G/G
108 233 121
```

```
summary(expr)
```

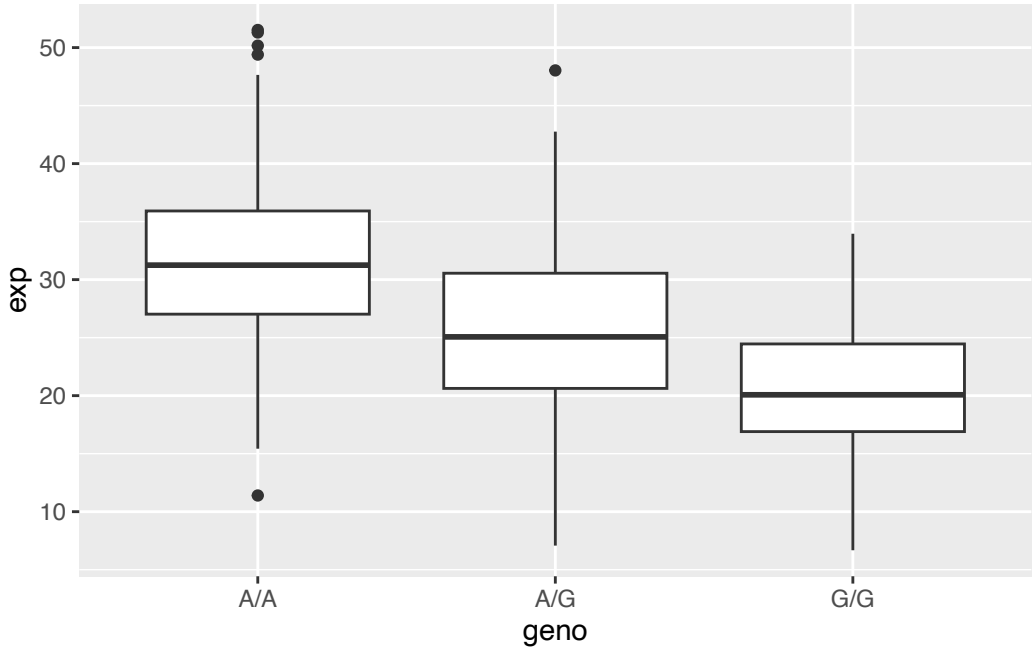
```
    sample              geno                exp        
 Length:462         Length:462         Min.   : 6.675  
 Class :character   Class :character   1st Qu.:20.004  
```

```
 Mode   :character   Mode   :character   Median :25.116
                                         Mean   :25.640
                                         3rd Qu.:30.779
                                         Max.   :51.518
```

Q13

There are 462 individuals in this data and 121 of them have the G|G genotype, 233 have the A/G genotype, and 108 have the A/A genotype. In order to find the median expression levels for all these samples, we can make a box plot and analyze the data. (See below). From the box plots we've made, we can see that A/A individuals have a median expression level of 31, A/G individuals have a median expression level of 25, and G/G individuals have a median expression level of 20.

```
library(ggplot2)
ggplot(expr) + aes(geno, exp) +
  geom_boxplot()
```



Q14

We can see that expression levels of ORMDL3 are noticeably different for the different genotypes, with A/A being the highest, A/G being a medium expression level, and G/G being the lowest expression level.

Introduction to Genome Informatics Lab

http://thegrantlab.org/bimm143

Barry J. Grant[1]

[1] Division of Biological Sciences, Section of Molecular Biology, University of California, San Diego, La Jolla, CA 92093, United States of America.

Abstract

High-throughput DNA sequencing has profoundly altered modern life science research. The decreasing cost and increasing accessibility of these "next-generation" methods is enabling new discoveries in diverse fields, from molecular, microbial and plant biology to disease diagnosis, cancer biology and beyond. While the importance of teaching these topics and their associated bioinformatics analysis skills is well-recognized, implementation of laboratory exercises is often beset by limited faculty expertise, dearth of computational resources and a lack of vetted teaching materials. Here we address these critical barriers with an accessible introduction to a set of freely available cloud-based genomics analysis tools and databases. In this lesson, students will learn to use the ENSEMBLE and OMIM databases, together with the Galaxy suite of bioinformatics tools, to investigate genomics, transcriptomics and population variability in the context of childhood asthma. These investigations are suitable for intermediate and upper division biology students who have previously taken at last one molecular biology class. No specific computational background is required beyond basic web browser usage. An optional extension exercise in section 4 delves into scripted data analysis with R.

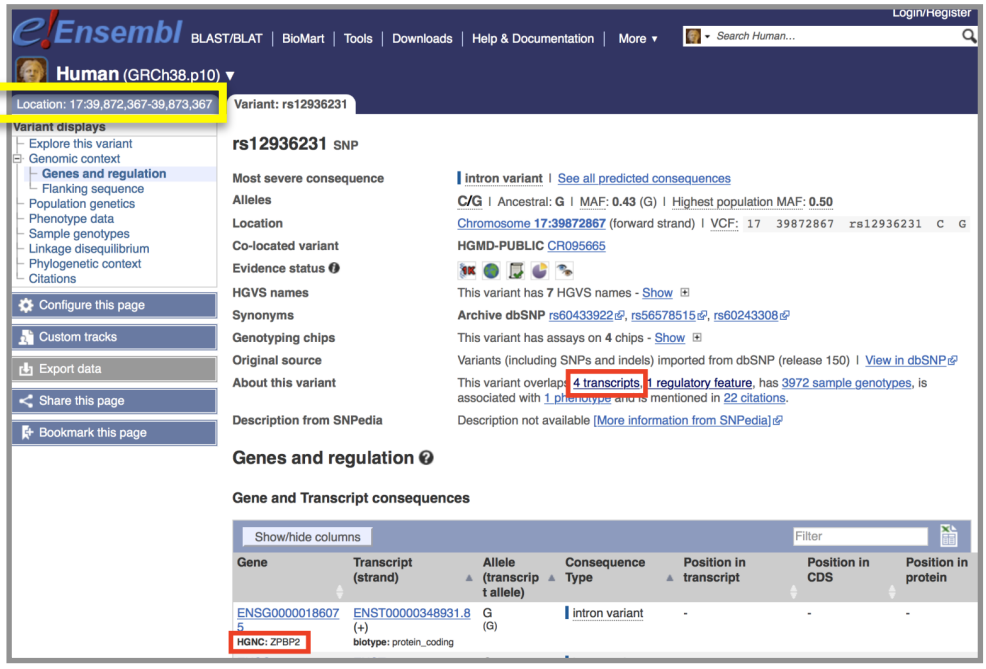
Student Laboratory Handout:

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

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 out what they are by visiting OMIM (http://www.omim.org) and locating the Verlaan et al. paper description.

***Q1**: What are those 4 candidate SNPs?*
*[HINT, you may want to check the first few links of search result and then record the **rs number** for these SNPs. The rs number is an accession number used by researchers and databases to refer to specific SNPs. It stands for Reference SNP cluster ID. A SNP is a location in the genome that is known to vary between individuals.]*

***Q2**: 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 indicating ZPBP2]

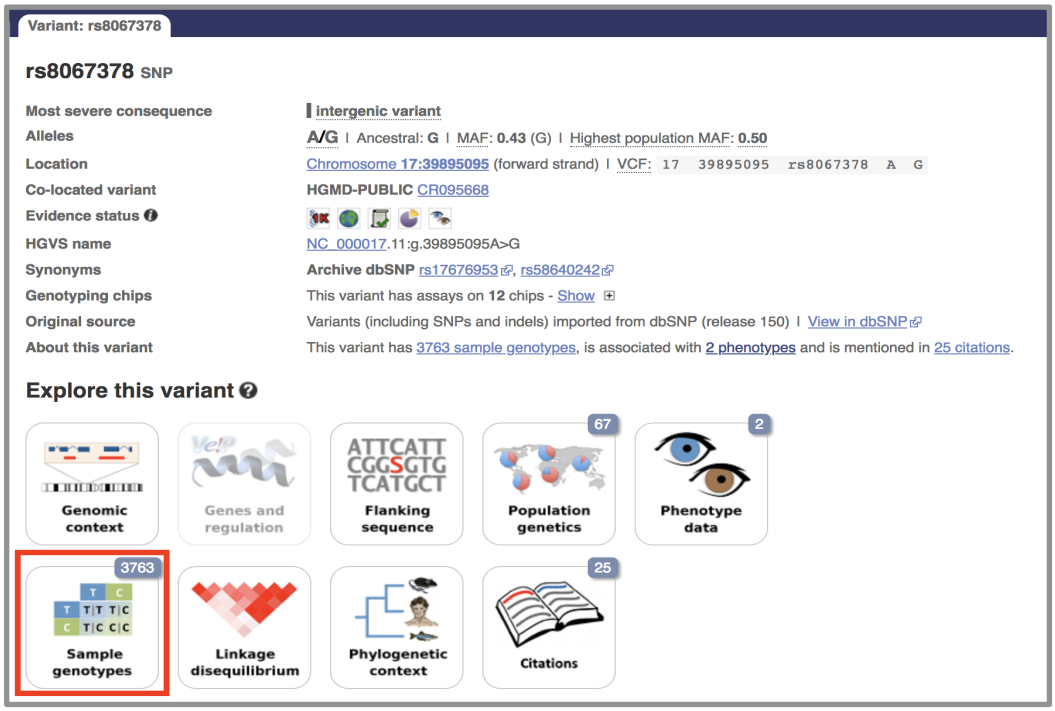


Now, you want to know the location of SNPs and genes in the genome. You can find the coordinates for the SNP itself on the Ensemble page along with overlapping genes or whether it is intergenic (i.e. between genes). However, to explore the surrounding regions and neighboring SNPs you will need to visit the linked Ensemble genome browser by clicking on the **Location** tab (highlighted with a yellow rectangle above).

***Q3**: What is the location of rs8067378 and what are the different alleles for rs8067378?*
[HINT, alleles and location are listed at the top of the Ensemble page as chromosome number and position. You may search in a genome browser to find this information]

***Q4**: Name at least 3 downstream genes for rs8067378?*

You are interested in the genotypes of these SNPs in a particular sample. Click on the "**Sample genotypes**" navigation link of of SNPs ensemble variant display page to look up their genotypes in the "Mexican Ancestry in Los Angeles, California" population.



***Q5**: What proportion of the Mexican Ancestry in Los Angeles sample population (MXL) are homozygous for the asthma associated SNP (G|G)?*
*[HINT: You can filter the displayed genotypes by entering the population code MXL. Then either count those of interest or download a CVS file for this population and use excel or the R functions **read.csv**(), and **table**() to answer this question]*

***Q6**. Back on the ENSEMBLE page, use the "search for a sample" field above to find the particular sample **HG00109**. This is a male from the GBR population group. What is the genotype for this sample?*

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

Now, you want to understand whether the SNP will affect gene expression. You can find the raw RNA-Seq data of this one sample on the class webpage:

https://bioboot.github.io/bggn213_W19/class-material/HG00109_1.fastq
https://bioboot.github.io/bggn213_W19/class-material/HG00109_2.fastq

Optional: Download and examine these files with your favorite UNIX utilities such as head, tail and less. You can use your RStudio Terminal tab to issue these commands.

Note: For more details about the ubiquitous **fastq** format see (http://en.wikipedia.org/wiki/FASTQ_format).

You can read about this while you are waiting for your **Galaxy server** to become available (see below). **Let Barry know when you are at this point so we can discuss common fastq formats further**.

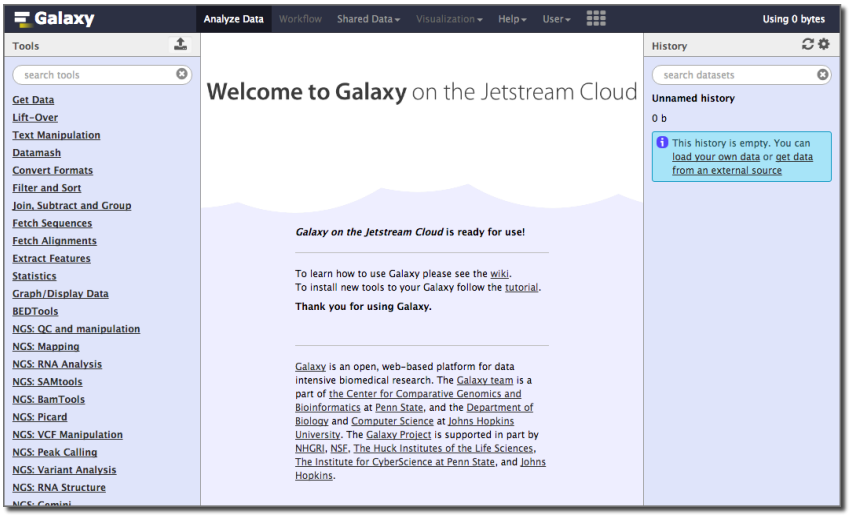
To begin our analysis of this data we will use **Galaxy** on either AWS or Jetstream cloud service providers.

Note: An alternative to Galaxy on AWS or Jetstream is to use the main **public Galaxy server** located at: https://usegalaxy.org/ .

Please see the *Instructor Notes* section below for an explanation of why we preferer a dedicated server for large class sizes.

Using Galaxy for NGS analyses

Follow Barry's instructions for accessing and logging into our very-own **Galaxy Server**. To find out more about Galaxy see: https://galaxyproject.org/tutorials/g101/



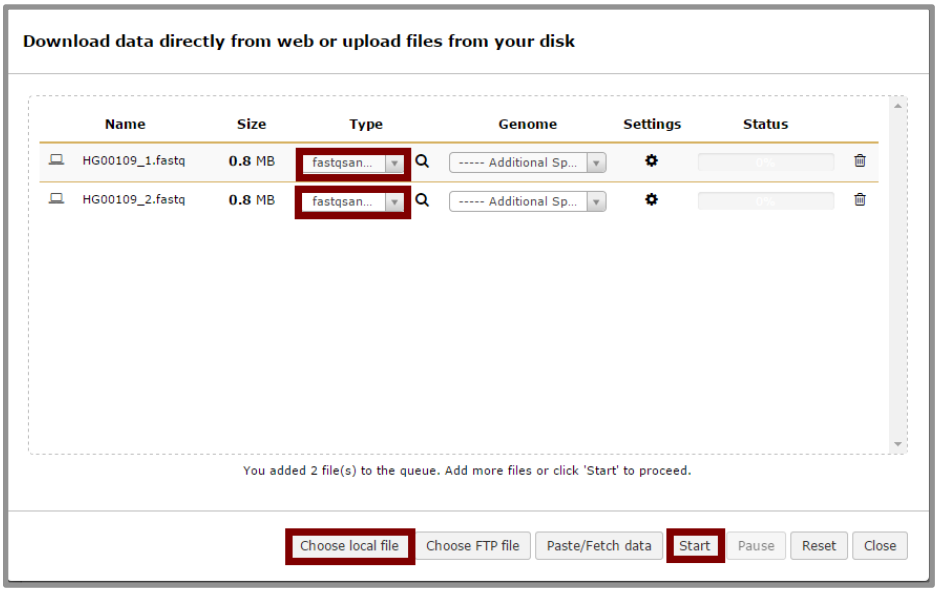
Once you are ready, you should be able to type (or copy/paste) your assigned instance IP address into your web browser to see your very own Galaxy server.

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

Upload our fastqsanger 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*.

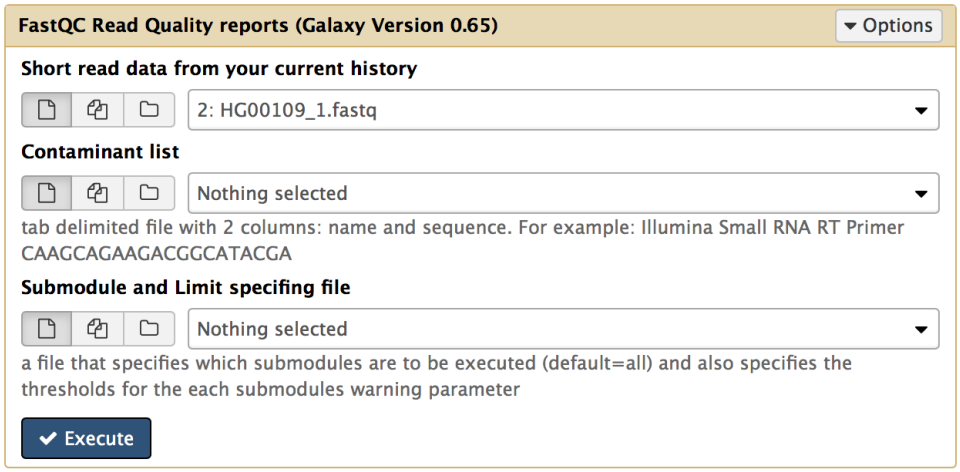


Now, you can check the data on the right panel. When they are colored gray they are still uploading and when they are green they are uploaded. Clicking in the name and various icons will provide more information to help you answer question 7 below.

***Q7**: 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]

Quality Control

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