Introduction to Genome Informatics Lab

http://thegrantlab.org/

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. An extension exercise in section 4 delves into scripted data analysis with R.

Student Laboratory Handout:

Section 1: Identify genetic variants of interest

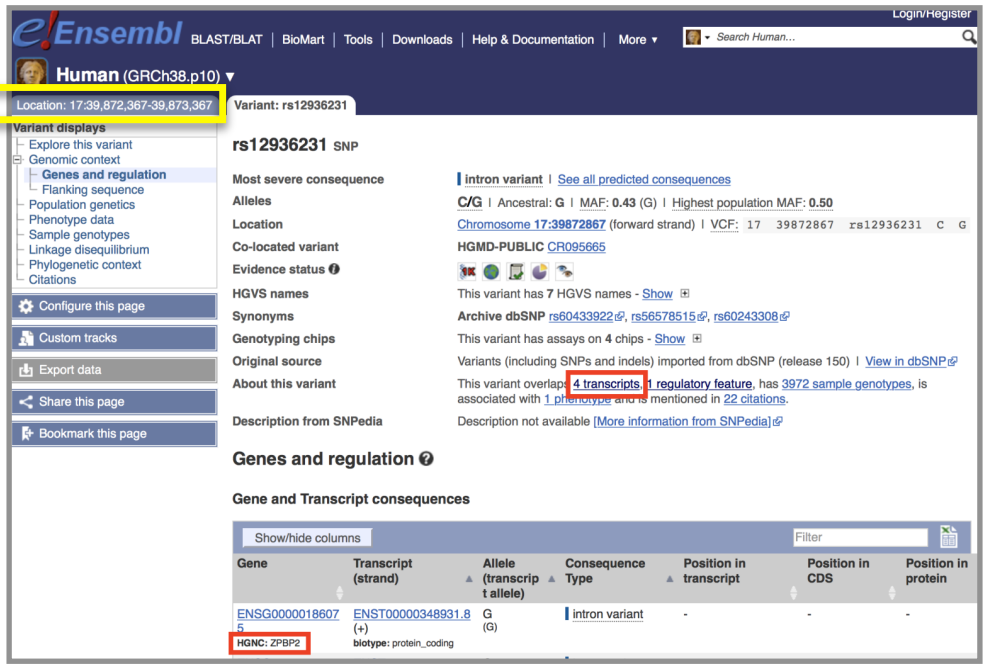
There are a number of gene variants associated with childhood asthma. A study from Verlaan *et al. (2009)* shows that 4 candidate SNPs demonstrate significant evidence for association. You want to find out what they are by visiting OMIM (http://www.omim.org) and locating the Verlaan et al. paper description.

***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.]*

4 candidate SNPs (rs12936231, rs8067378, rs9303277, and rs7216389) in 3 family-based childhood asthma cohorts, all 4 SNPs demonstrated significant evidence for association (p = 8.74 x 10(-7) to 4.95 x 10(-8)) in each of the cohorts, and haplotype association analysis confirmed a strong association of the overexpressed-associated haplotype with asthma risk (p = 8.62 x 10(-8)).

***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]
There are 3 genes ZPBP2, IKZF3 and GSDMB:
1. rs12936231 This variant overlaps 4 transcripts, 2 regulatory features, has 2504 sample genotypes, is associated with 1 phenotype and is mentioned in 58 citations. ZPBP2
2. rs8067378 This variant overlaps 1 regulatory feature, has 2504 sample genotypes, is associated with 5 phenotypes and is mentioned in 61 citations.

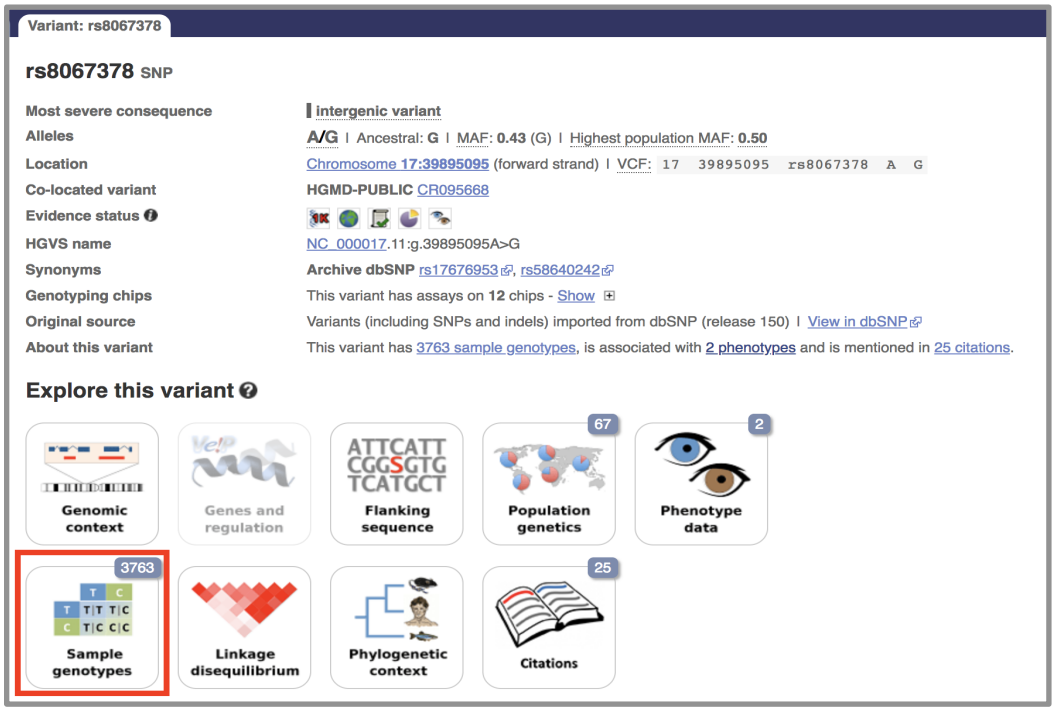


Now, you want to know the location of SNPs and genes in the genome. You can find the coordinates for the SNP itself on the Ensemble page along with overlapping genes or whether it is intergenic (i.e. between genes). However, to explore the surrounding regions and neighboring SNPs you will need to visit 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]
Variation ID: rs8067378
Variation Type: SNV, length 1
Location of rs8067378 is chromosome 17: 39,799,839-39,799,849.
Alleles: A/C/G
[Genomic locations]

***Q4**: Name at least 3 downstream genes for rs8067378?*
MIEN1, PGAP3, NEUROD2

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]*

Genotype..forward.strand.	ALL, AMR, MXL
A|A	22
A|G	21
G|A	12
G|G	9

***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?*

G|G

***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**.

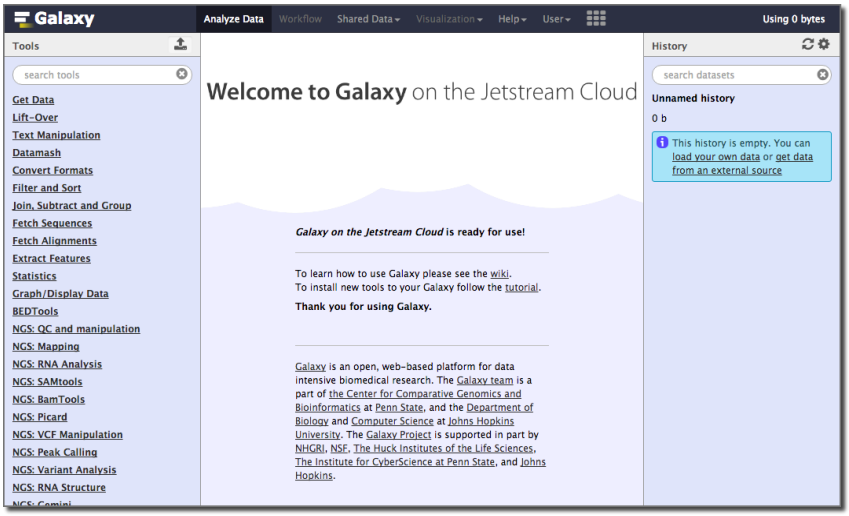
To begin our analysis of this data we will use **Galaxy** on either AWS or Jetstream cloud service providers.

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/



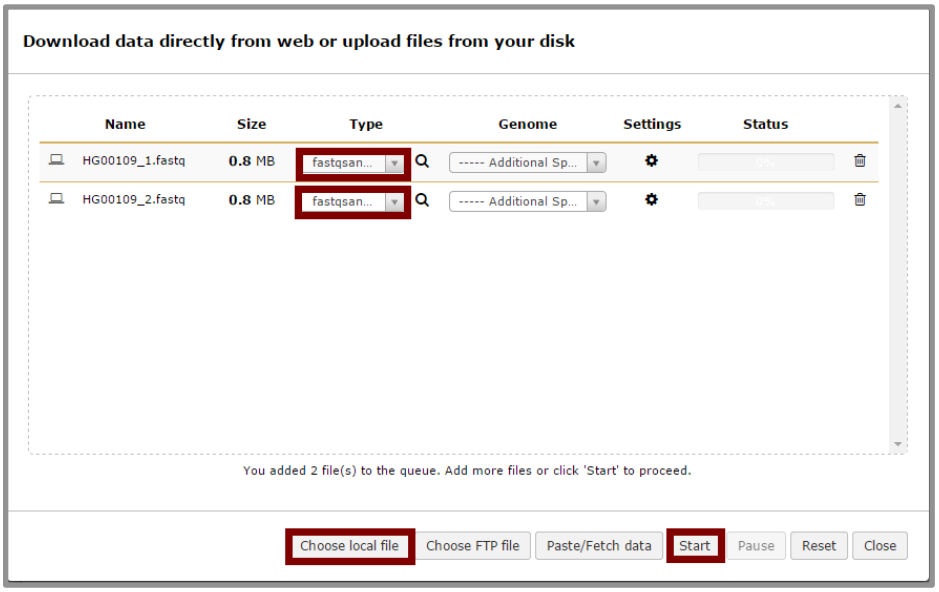
Once you are ready, you should be able to type (or copy/paste) your assigned instance IP address into your web browser to see your very own Galaxy server.

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*.

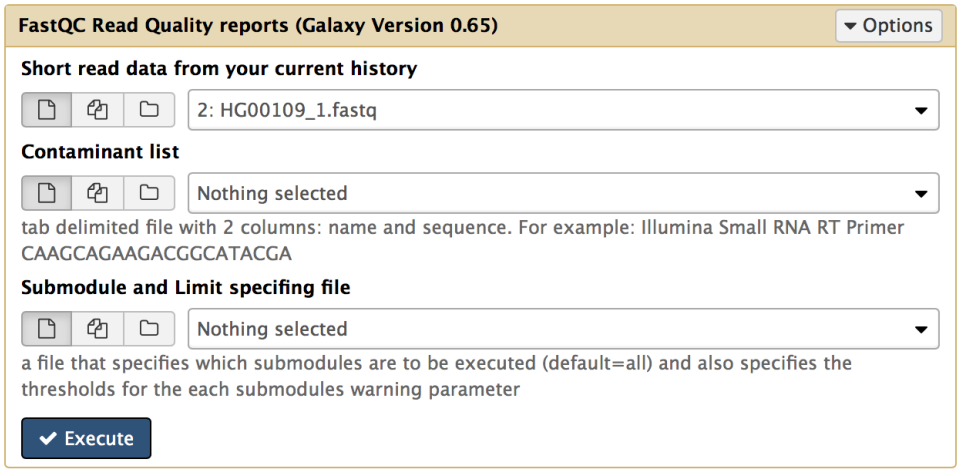


Now, you can check the data on the right panel. When they are colored gray they are still uploading and when they are green they are uploaded. Clicking in the name and various icons will provide more information to help you answer question 7 below.

***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]
File size: 742 kb
Format: fastqsanger
Sequence read consist of 4 lines:
Line 1: Header line (starts with "@")
Line 2: Sequence line

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.

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.

Note: You can find more about each analysis section (or module) here:
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/

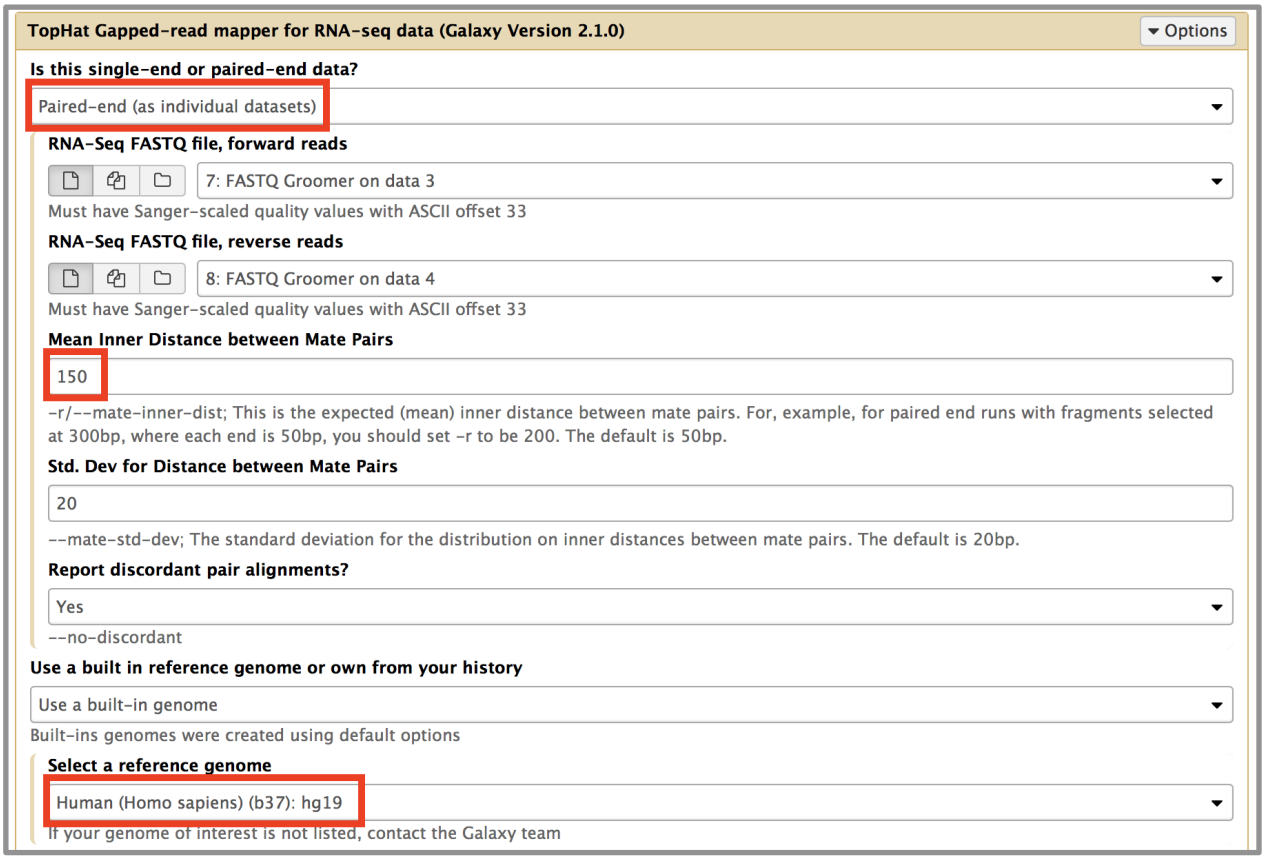
***Q8**: What is the GC content and sequence length of the second fastq file?*
[HINT, you may check "Basic Statistics"]
The second file's GC content is 54 %, and sequence length is 50-75. Total Sequences is 3863.

***Q9**: 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?]
According to the data " Per base sequence quality " , blue line indicates the mean quality score, and median quality score lower than 20 might be an indication that the data quality is compromised. Trimming involves removing low-quality bases from the ends of the sequences to improve overall data quality. In the case of the second sequence the median quality score '

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

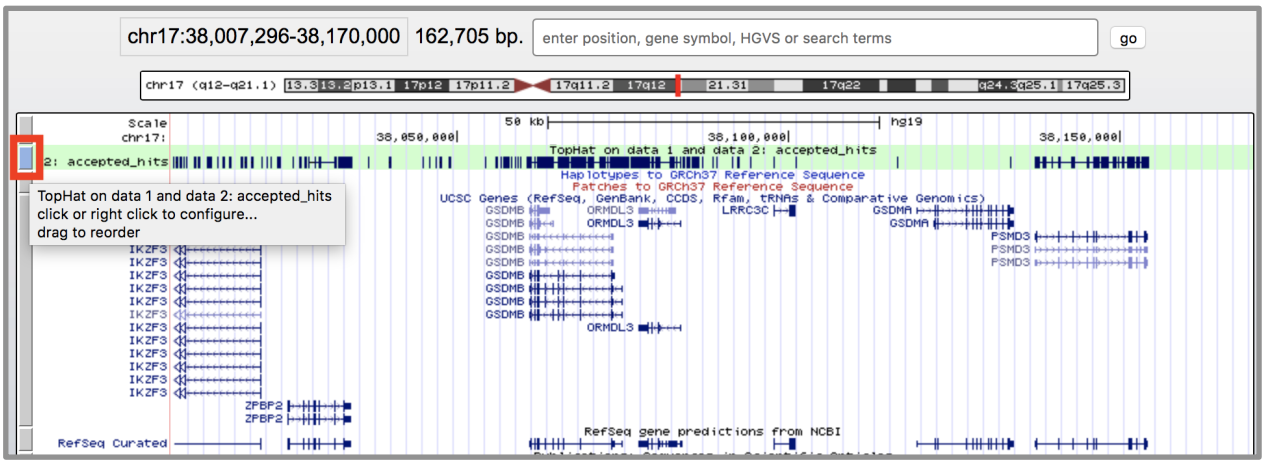
The next step is mapping the processed reads to the genome. The major challenge when mapping RNA-Seq reads is that the reads, because they come from RNA, often cross splice junction boundaries; splice junctions are not present in a genome's sequence, and hence typical NGS mappers such as **Bowtie** (http://bowtie-bio.sourceforge.net/index.shtml) 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.

Select **NGS: RNA Analysis > Tophat** tool to map RNA-seq reads to the **hg19** build of the Human reference genome. Note that on some Galaxy servers Tophat may be located under "Deprecated Tools" at the bottom of the left-hand menu.



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.

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, but the other files can be of interest depending upon the goal of any other analysis.

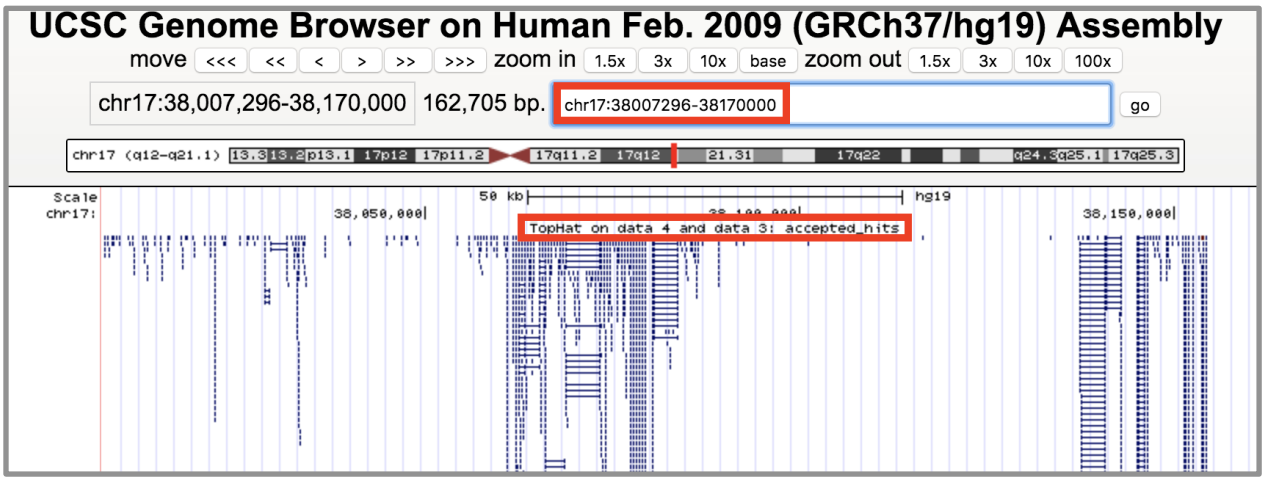
The accepted hits file is in BAM format, which is binary version of the human readable SAM format. To inspect these results we will convert the BAM file to SAM format using **NGS: SAMtools** > **BAM-to-SAM** tool. Once converted click the eye icon to view within galaxy. Note there is lots of metadata 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. See: https://bioboot.github.io/bggn213_W19/class-material/sam_format/

Display at UCSC

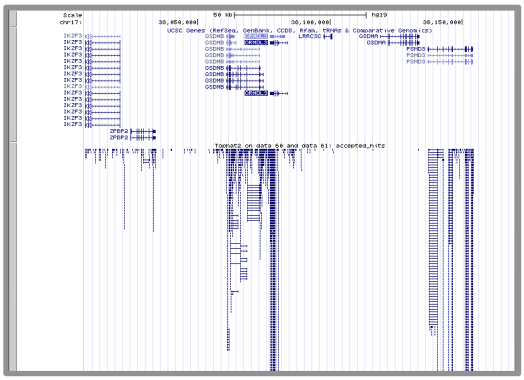
Once complete select and expand the **accepted hits** file in your history sidebar. Then Click on the "**display at UCSC main**" link.

This will load your TopHat results as a custom track on the **UCSC Genome Browser**. You can then click on the custom track (see above image) 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. See figure below for an example:

***Q10**: 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]*
hg19 chr17:37,913,968-38,020,441



You may also want to view your results in the stand-alone IGV browser available from:

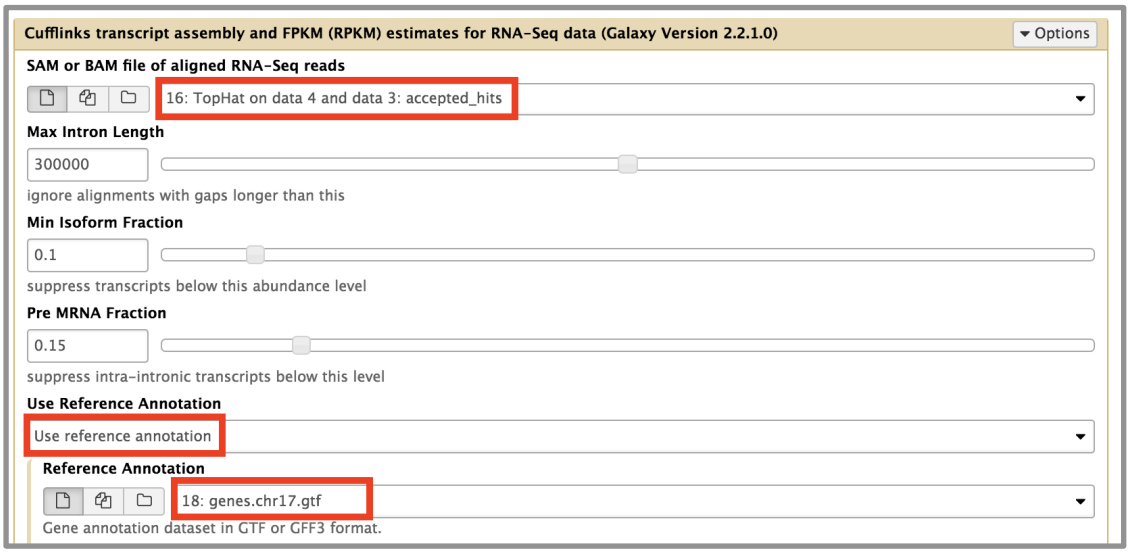


https://software.broadinstitute.org/software/igv/download You can download both your accepted hits bam file and the corresponding index file from clicking the disc save icon in galaxy.

Q11: *Following 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]
1)Homo sapiens IKAROS family zinc finger 3 (Aiolos) (IKZF3), transcript variant 1, 2) Homo sapiens gasdermin B (GSDMB), 3) Human Gene ORMDL3 (uc002htj.2), 4) Human Gene ZPBP2 (uc002hte.3), 5) Homo sapiens leucine rich repeat containing 3C (LRRC3C), 6) Homo sapiens proteasome (prosome, macropain) 26S subunit, non-ATPase, 3 (PSMD3)

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 form the course website:

https://bioboot.github.io/bimm143_W18/class-material/genes.chr17.gtf) 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.

***Q12**: 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?*

The FPKM for ORMDL3 gene is 136853

ZPBP2 - 4613.49
GSDMA - 133.634
GSDMB - 26366.3

Note. The FPKM metric attempts to normalize for sequencing depth and gene length. 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.)

If you have time you can run a separate **htseq-count** analysis for your "accepted hits" file. Open htseq-count, and set it up using default parameters other than "Aligned SAM/BAM File" and "GFF File" which you need to select from your history as in the above example.

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. The DESeq2 package will be the subject of a separate class. For now, 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 [HOMEWORK]**

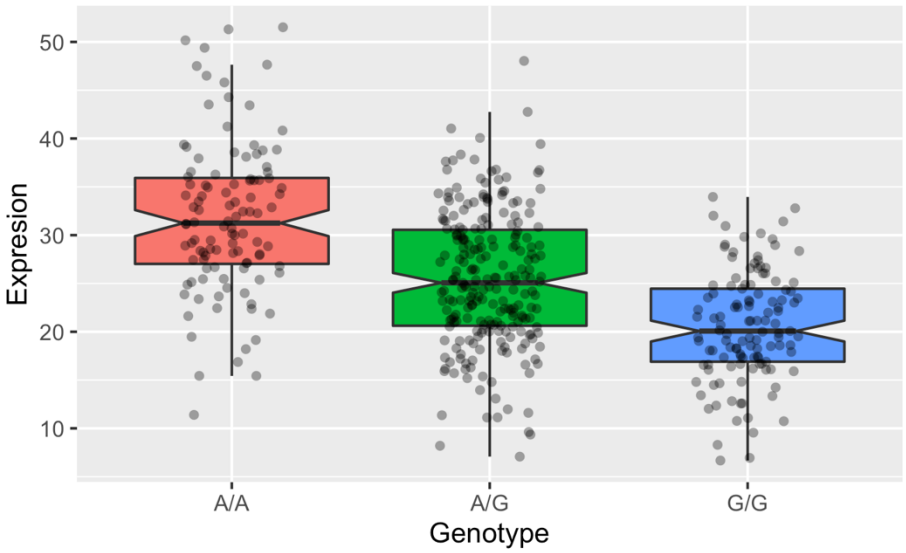
One sample is obviously not enough to know what is happening in a population. You are interested in assessing genetic differences on a population scale. So, you processed about ~230 samples and did the normalization on a genome level. Now, you want to find whether there is any association of the 4 asthma-associated SNPs (**rs8067378…)** on **ORMDL3** expression**.**

This is the final file you got (https://bioboot.github.io/bggn213_W19/class-material/rs8067378_ENSG00000172057.6.txt). The first column is sample name, the second column is genotype and the third column are the expression values.

Open a new RMarkdown document in RStudio to answer the following two questions. Submit your resulting PDF report with your working code, output and narrative text answering Q13 and Q14 to GradeScope.

Q13: *Read this file into R and determine the sample size for each genotype and their corresponding median expression levels for each of these genotypes.* **Hint**: The **read.table()**, **summary()** and **boxplot()** functions will likely be useful here. There is an example R script online to be used ONLY if you are struggling in vein. Note that you can find the medium value from saving the output of the **boxplot()** function to an R object and examining this object. There is also the **medium()** and **summary()** function that you can use to check your understanding.

***Q14**: 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?* **Hint**: An example boxplot is provided overleaf – yours does not need to be as polished as this one.


Expresion
50
40
30
20
10
A/A
A/G
G/G
Genotype

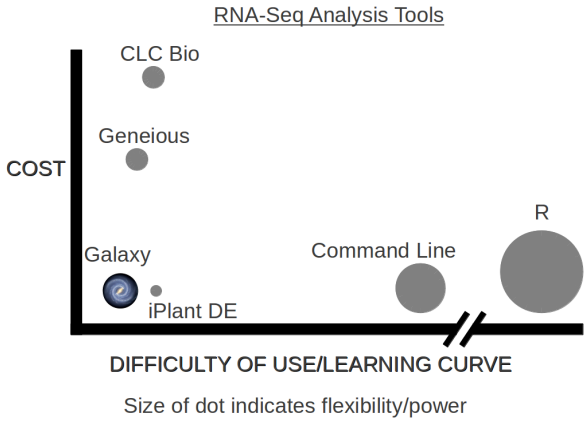
Equipment and Supplies:

This is a bioinformatics tutorial and requires no experimental laboratory equipment or supplies. The workshop could be done with participants bringing their own laptop computers (preferred) or hosted in a computer lab space accommodating 25 participants. Note that if laptops are used it is important that adequate Wi-Fi and power outlets are available. If hosted in a computer lab then guest login access to the computers will be required. No specific software is required beyond a recent web browser (Safari, Chrome or Firefox).

Instructor Notes:

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 (Soon *et al.* 2013).

In this lab we will use the **Galaxy** web-based interface to a suite of bioinformatics tools for genomic sequence analysis (Afgan *et al.* 2018) Galaxy is free and comparatively easy to use (see Figure opposite for a schematic comparison of some common bioinformatics RNA-Seq analysis methods). Using the public Galaxy server (found at https://usegalaxy.org/) simplifies instructor setup and minimizes the need for dedicated local computing infrastructure.



It is important to note however that using the public Galaxy server, as opposed to a local instance, can result in competition for resources and student jobs being "queued". This will result in variable wait times for job completion that depend on relative server load (i.e. other users demand for resources). This will be most notable during *Section 3 mapping of RNA-Seq reads*. Using the public server, we have observed completion times as little as 20 minutes and as long as one hour for this step. At UCSD we will utilize a custom Galaxy instance for each workshop participant to mitigate potential wait times.

Side-Note: 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.

Description of workshop presentation:

In addition to searching and exploring the major bioinformatics databases OMIM, ENSEMBLE and UCSC participants will use Galaxy's online interface to perform sequence quality assessment, alignment of sequence reads to a reference genome (a.k.a. mapping) and generate a counts table for expressed genes (see: Trapnell *et al.* 2012). All of these steps are performed "in the cloud" using offsite computing resources.

Reference:

- All data files can also be found at: https://bioboot.github.io/bggn213_W19/lectures/#13 Components of Section 2 were adapted from https://usegalaxy.org/u/jeremy/p/galaxy-rna-seq-analysis-exercise.

1. Verlaan DJ, Berlivet S, Hunninghake GM, Madore A-M, Larivière M, Moussette S, Grundberg E, Kwan T, Ouimet M, Ge B, Hoberman R, Swiatek M, Dias J, Lam KCL, Koka V, Harmsen E, Soto-Quiros M, Avila L, Celedón JC, Weiss ST, Dewar K, Sinnett D, Laprise C, Raby BA, Pastinen T, Naumova AK. Allele-specific chromatin remodeling in the ZPBP2/GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease. **Am J Hum Genet**. 2009 Sep;85(3):377–393. PMCID: PMC2771592

2. Soon WW, Hariharan M, Snyder MP. High-throughput sequencing for biology and medicine. **Mol Syst Biol**. 2013;9:640. PMCID: PMC3564260

3. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltemann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. **Nucleic Acids Res**. 2018 02;46(W1):W537–W544. PMCID: PMC6030816

4. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. **Nat Protoc**. 2012 Mar 1;7(3):562–578. PMCID: PMC3334321