BIMM-143: INTRODUCTION TO BIOINFORMATICS (Lecture 13)

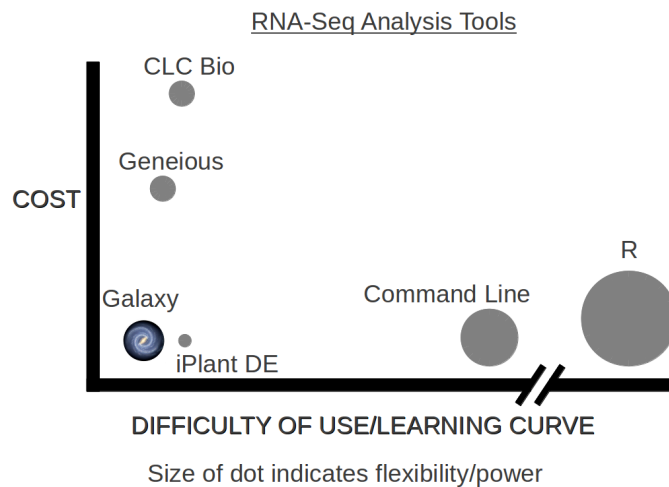Genome Informatics (Part 1)
http://thegrantlab.org/bimm143/
Dr. Barry Grant

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).

RNA-Seq Analysis Tools

CLC Bio

Geneious

COST

Galaxy

iPlant DE

R

Command Line

DIFFICULTY OF USE/LEARNING CURVE

Size of dot indicates flexibility/power

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

**Q1**: 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

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).

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 download a CVS file for this population from ENSEMBLE and use the R functions **read.csv**(), *and* **table**() *to answer this question]*
14%

Q6. *Back on the ENSEMBLE page, search for 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/bimm143_F18/class-material/HG00109_1.fastq
https://bioboot.github.io/bimm143_F18/class-material/HG00109_2.fastq

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 Jetstream.

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.
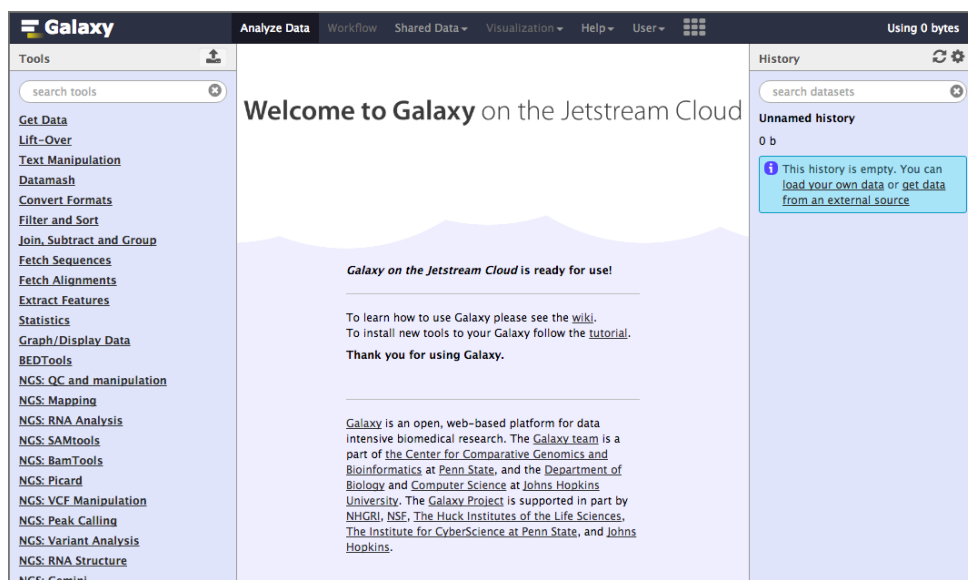
Upload our <mark>fastqsanger</mark> sequences

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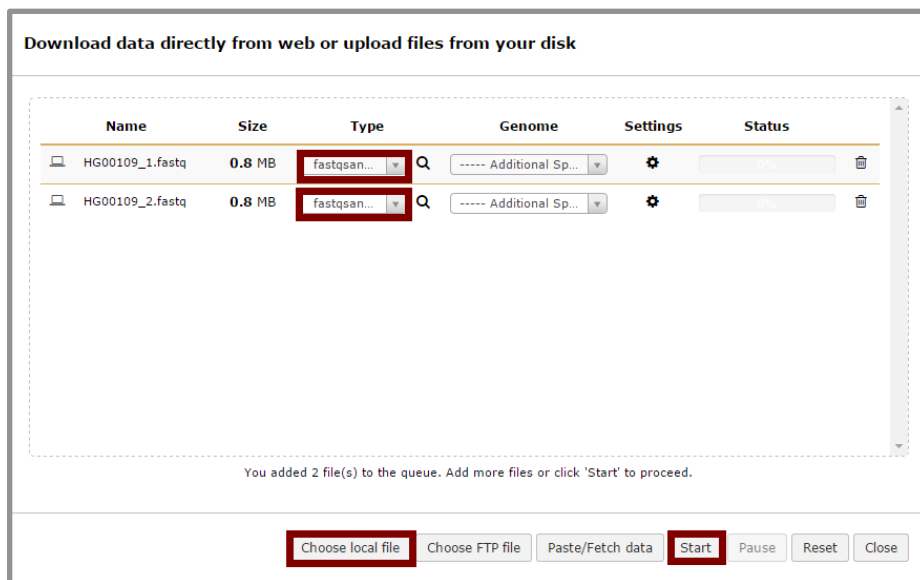
Now, you can check the data on the right panel. When they are colored gray they are still uploading and when they are green they are uploaded. Clicking in the name and various icons will provide more information to help you answer question 7 below.

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 <mark>fastqsanger</mark> here!
[HINT, you can check the fastq format wiki for more information]
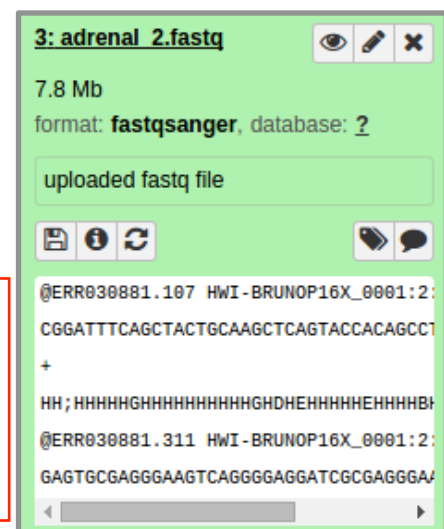
Q8: *Does the first sequence have good quality?*
[HINT, what is the quality score for each nucleotide? You can use the R functions gtools::asc() together with the seqinr::s2c() to return the values corresponding to the specified ASCII codes.
asc(s2c("DDDDCDEDCDDDDBBDDDCC@"))

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/

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

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.

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

TopHat Gapped-read mapper for RNA-seq data (Galaxy Version 2.1.0) ▼ Options

Is this single-end or paired-end data?

Paired-end (as individual datasets) ▼

RNA-Seq FASTQ file, forward reads

7: FASTQ Groomer on data 3 ▼

Must have Sanger-scaled quality values with ASCII offset 33

RNA-Seq FASTQ file, reverse reads

8: FASTQ Groomer on data 4 ▼

Must have Sanger-scaled quality values with ASCII offset 33

Mean Inner Distance between Mate Pairs

150

–r/--mate-inner-dist; This is the expected (mean) inner distance between mate pairs. For, example, for paired end runs with fragments selected at 300bp, where each end is 50bp, you should set –r to be 200. The default is 50bp.

Std. Dev for Distance between Mate Pairs

20

--mate-std-dev; The standard deviation for the distribution on inner distances between mate pairs. The default is 20bp.

Report discordant pair alignments?

Yes ▼

--no-discordant

Use a built in reference genome or own from your history

Use a built-in genome ▼

Built-ins genomes were created using default options

Select a reference genome

Human (Homo sapiens) (b37): hg19 ▼

If your genome of interest is not listed, contact the Galaxy team

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/bimm143_F18/class-material/sam_format/
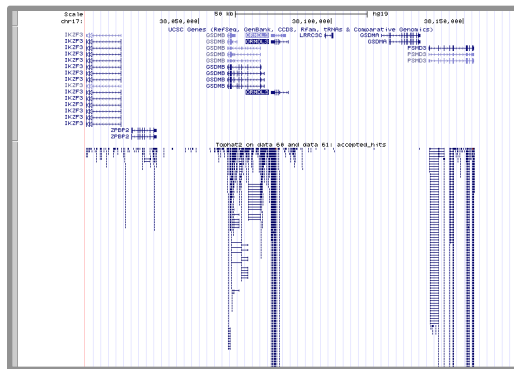
Display at UCSC
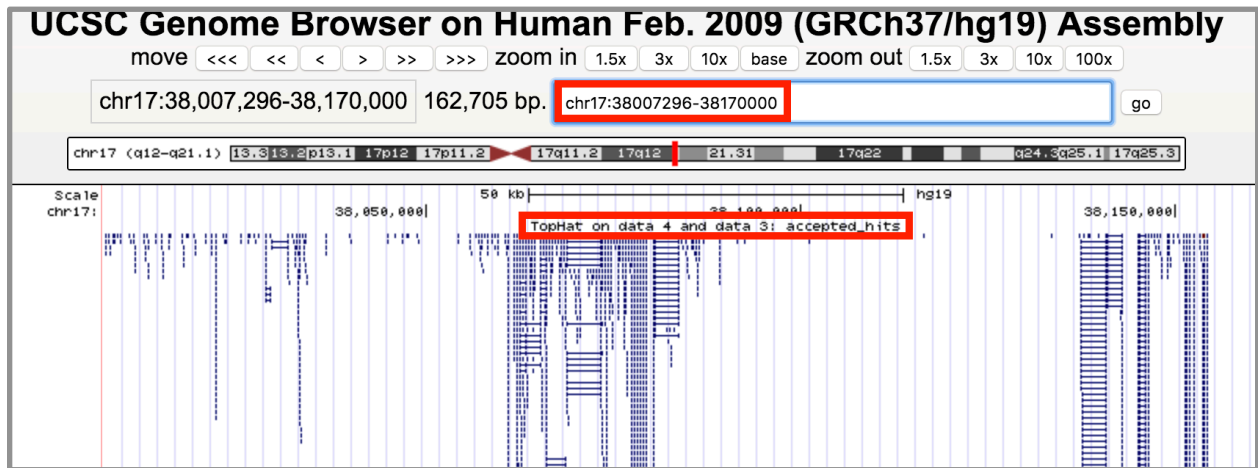Once complete select and expand the **accepted hits** file in your history sidebar. Then Click on the "**display at UCSC main**" link.
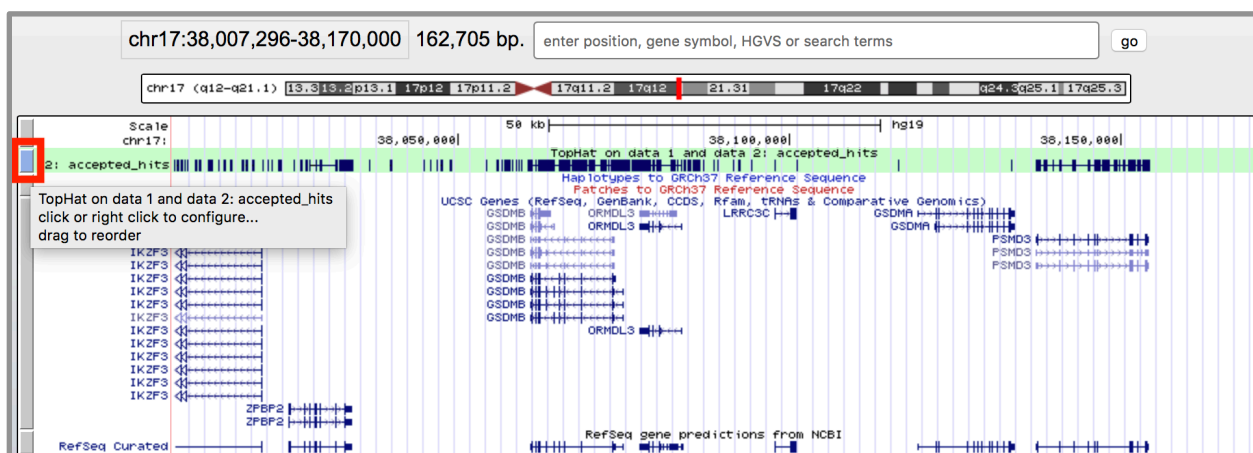
This will will load your TopHat results as a custom track on the **UCSC Genome Browser**. You can then click on the custom track (see above image) and change the display mode from **Dense** to **Full** and enter the region chr17:38007296-38170000 into

8

the text box to see the pile-up of aligned sequence reads in this location. See figure below for an example:

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.
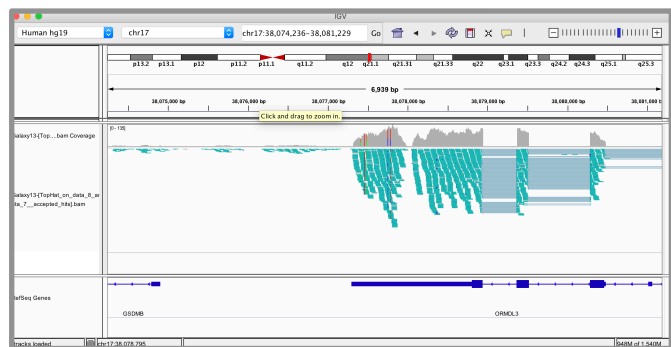
With alignment result from TopHat, we can now calculate gene expression with the **NGS: RNA Analysis** > **Cufflinks** tool. Before running Cufflinks, you should upload the reference annotation file "**gene.chr17.gtf**" (available form the course website: https://bioboot.github.io/bimm143_W18/class-material/genes.chr17.gtf) 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.

Note. In later classes we will discuss some shortfalls of the FPKM metric and why some more recent normalization schemes (that also attempt to normalize for sequencing depth and gene length) are preferred. 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.)

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**
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/bimm143_S18/class-material/rs8067378_ENSG00000172057.6.txt). The first column is sample name, the second column is genotype and the third column is the expression value.

Q14: *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.

Q15: *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?*

Q17: 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: https://forms.gle/Qx4UTSc4zkpafRvq9

Reference:

- All data files can also be found at: https://bioboot.github.io/bimm143_S18/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.