

## BGGN-213: FOUNDATIONS OF BIOINFORMATICS (Lecture 14)

Genome Informatics (Part 1)

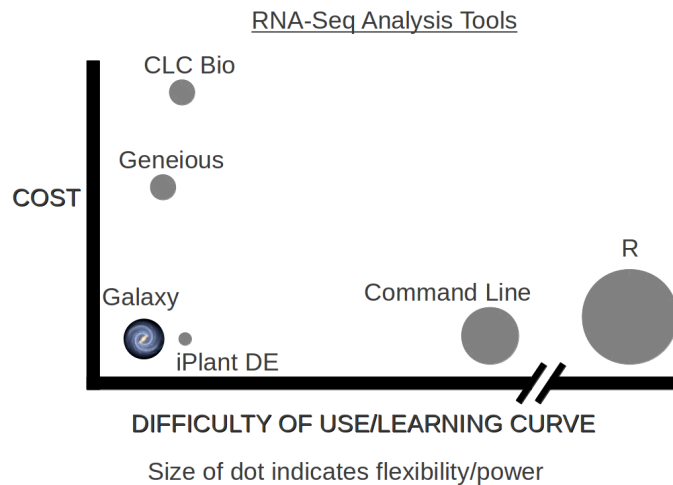
[https://bioboot.github.io/bggcn213\\_f17/lectures/#14](https://bioboot.github.io/bggcn213_f17/lectures/#14)

Dr. Barry Grant

Nov 2017

**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 related to sequencing.

### **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

**Q2: What three genes of these variants overlap?**

ZPBP2, GSDMB, and ORMDL3

**rs12936231** SNP

Most severe consequence: Intron variant | See all predicted consequences

Alleles: C/G | Ancestral: G | MAF: 0.43 (G) | Highest population MAF: 0.50

Location: Chromosome 17:39872867 (forward strand) | VCF: 17 39872867 rs12936231 C G

Co-located variant: HGMD-PUBLIC CR095665

Evidence status: [Icons]

HGVS names: This variant has 7 HGVS names - Show

Synonyms: Archive dbSNP rs60433922, rs56578515, rs60243308

Genotyping chips: This variant has assays on 4 chips - Show

Original source: Variants (including SNPs and indels) imported from dbSNP (release 150) | View in dbSNP

About this variant: This variant overlaps 4 transcripts, 1 regulatory feature, has 3972 sample genotypes, is associated with 1 phenotype and is mentioned in 22 citations.

Description from SNPPedia: Description not available | More information from SNPPedia

**Genes and regulation**

Gene and Transcript consequences

Gene	Transcript (strand)	Allele (transcript allele)	Consequence Type	Position in transcript	Position in CDS	Position in protein
ENSG00000186075	ENST00000348931.8 (+)	G (G)	Intron variant	-	-	-
HGNC: ZPBP2 biotype: protein_coding						

Now, you want to know the location of SNPs and genes in the 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 a genome browser]

Chromosome 17: 39,895,045-39,895,145. A/G variants (43% G)

**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. 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.

**Variation displays**  
**Explore this variation**  

- Genomic context
- Genes and regulation
- Flanking sequence
- Population genetics
- Individual genotypes (3761)
- Linkage disequilibrium
- Phenotype Data (5)
- Phylogenetic Context
- Citations (12)
- External Data
- SNPedia
- LOVD

**rs8067378 SNP**  
**Original source**  
**Alleles**  
**Location**  
**Co-located**  
**Evidence status**  
**Synonyms**  
**HGVS name**  
**Genotyping chips**

Variants (including SNPs and indels) imported from dbSNP (release 138) | [View in dbSNP](#)  
**A/G** | Ancestral: G | Ambiguity code: R | **MAF: 0.43 (G)**  
Chromosome **17:38051348** (forward strand) | [View in location tab](#)  
with HGMD-PUBLIC [CR095668](#)  
  
Archive dbSNP [rs17676953](#), [rs58640242](#)  
[17:g.38051348A>G](#)  
This variation has assays on **11** chips - click the plus to show


**Explore this variation**  

  
**Genomic context**

  
**Genes and regulation**

  
**Population genetics**

  
**Individual genotypes**

  
**Linkage disequilibrium**

  
**Phenotype data**

  
**Citations**

  
**Phylogenetic context**

  
**Flanking sequence**

**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 the expression of the gene.

You find the RNA-Seq data of one sample on the class webpage ([https://bioboot.github.io/bioinf525\\_w17/class-material/HG00109\\_1.fastq](https://bioboot.github.io/bioinf525_w17/class-material/HG00109_1.fastq), [https://bioboot.github.io/bioinf525\\_w17/class-material/HG00109\\_2.fastq](https://bioboot.github.io/bioinf525_w17/class-material/HG00109_2.fastq)). Note that this is the raw sequence fastq file.

Download and examine one of these files with your favorite UNIX utilities such as **head**, **tail** and **less**.

**Note:** For more details about the ubiquitous fastq format see ([http://en.wikipedia.org/wiki/FASTQ\\_format](http://en.wikipedia.org/wiki/FASTQ_format) ). You can read about this while you are waiting for your **jetstream** instance to become available (see below).

To begin our analysis of this data we will use Galaxy on Jetstream.

### ***Start a “galaxy standalone” jetstream computer instance***

Follow our previous instructions for booting and logging into a medium size **Galaxy Standalone** jetstream computer instance. See: [https://bioboot.github.io/bggn213\\_f17/jetstream/boot/](https://bioboot.github.io/bggn213_f17/jetstream/boot/)

Once your instance is available login via ssh with **your terminal application**. Your shell command should look something like the following:

```
ssh -i ~/bggn213_private_key tb170077@YOUR_IP_ADDRESS
```

**Note:** Your IP address in the above command is specific to your running computer instance (i.e. typically you will copy this from the jetstream website). The command above also assumes the *keyfile* resides in your home area. You may need to alter the file path to your keyfile if this is not the case.

Once it is up and running you should be able to type your instance IP address into your web browser to see your very own Galaxy server.

The screenshot shows the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', 'User', and a grid icon. The right corner shows 'Using 0 bytes'. The left sidebar is titled 'Tools' and contains a search bar and a list of tool categories: Get Data, Lift-Over, Text Manipulation, Datamash, Convert Formats, Filter and Sort, Join, Subtract and Group, Fetch Sequences, Fetch Alignments, Extract Features, Statistics, Graph/Display Data, BEDTools, NGS: QC and manipulation, NGS: Mapping, NGS: RNA Analysis, NGS: SAMtools, NGS: BamTools, NGS: Picard, NGS: VCF Manipulation, NGS: Peak Calling, NGS: Variant Analysis, NGS: RNA Structure, and NGS: Combi. The main content area has a header 'Welcome to Galaxy on the Jetstream Cloud' and a message: 'Galaxy on the Jetstream Cloud is ready for use!'. Below this, it says: 'To learn how to use Galaxy please see the [wiki](#). To install new tools to your Galaxy follow the [tutorial](#). Thank you for using Galaxy.' At the bottom, there is a paragraph about Galaxy being an open, web-based platform for data intensive biomedical research, supported by various institutions. The right sidebar is titled 'History' and shows 'Unnamed history' with '0 b'. A message box states: 'This history is empty. You can [load your own data](#) or [get data from an external source](#)'.

## 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 **choose 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*.

Download data directly from web or upload files from your disk

Name	Size	Type	Genome	Settings	Status
HG00109_1.fastq	0.8 MB	fastqsan...	----- Additional Sp...	⚙️	
HG00109_2.fastq	0.8 MB	fastqsan...	----- Additional Sp...	⚙️	

You added 2 file(s) to the queue. Add more files or click 'Start' to proceed.

**Choose local file** Choose FTP file Paste/Fetch data **Start** Pause Reset Close

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]

**3: adrenal 2.fastq**

7.8 Mb  
format: **fastqsanger**, database: ?

uploaded fastq file

```
@ERR030881.107 HWI-BRUNOP16X_0001:2:
CGGATTTTCAGCTACTGCAAGCTCAGTACCACAGCCT
+
HH;HHHHHHHHHHHHHHHHHHGHDEHHHHHHHHHHBH
@ERR030881.311 HWI-BRUNOP16X_0001:2:
GAGTGCGAGGGAAGTCAGGGGAGGATCGCGAGGGA/
```

**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 in detail. Run a quality control check with the FastQC tool on your data using the “**NGS: QC and manipulation**” > **FastQC Read Quality reports**.

The screenshot shows the 'FastQC Read Quality reports (Galaxy Version 0.65)' interface. It features a 'Short read data from your current history' section with a text input field containing two URLs: '4: https://bioboot.github.io/bioinf525\_w17/class-material/HG00109\_2.fastq' and '3: https://bioboot.github.io/bioinf525\_w17/class-material/HG00109\_1.fastq'. Below this is a 'Contaminant list' section with a dropdown menu set to 'Nothing selected' and a description: 'tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA'. The 'Submodule and Limit specifying file' section also has a dropdown menu set to 'Nothing selected' and a description: 'a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter'. At the bottom is a blue 'Execute' button with a checkmark icon.

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.

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”]

**Q10:** 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 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.

**Note:** Our input data is pair-end data. For Tophat in Galaxy, you need 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.

The calculation may take some time. There will eventually be five outputs: accepted\_hits, insertions, deletions, splice junctions and an alignment summary.

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

Is this single-end or paired-end data?

RNA-Seq FASTQ file, forward reads  
  
 Must have Sanger-scaled quality values with ASCII offset 33

RNA-Seq FASTQ file, reverse reads  
  
 Must have Sanger-scaled quality values with ASCII offset 33

Mean Inner Distance between Mate Pairs  
  
 -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.

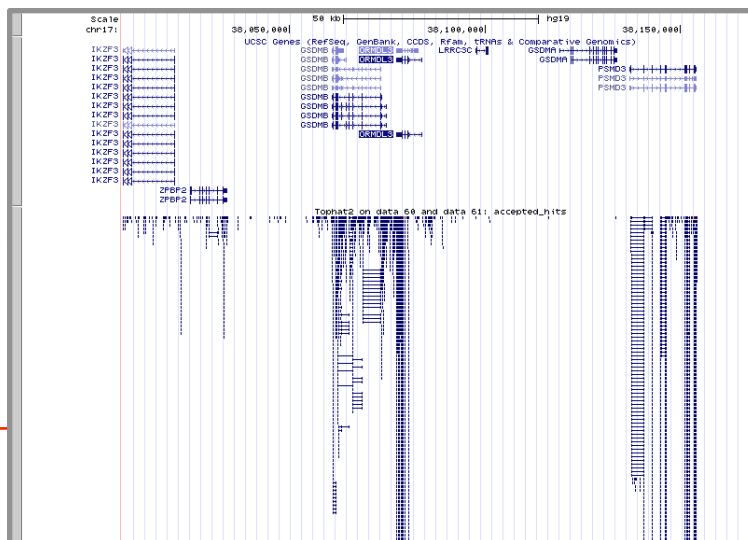
Std. Dev for Distance between Mate Pairs  
  
 --mate-std-dev; The standard deviation for the distribution on inner distances between mate pairs. The default is 20bp.

Report discordant pair alignments?  
  
 --no-discordant

Use a built in reference genome or own from your history  
  
 Built-ins genomes were created using default options

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

Clicking on the “**display at UCSC main**” link 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** to see the pile-up of aligned sequence reads. See figure below:



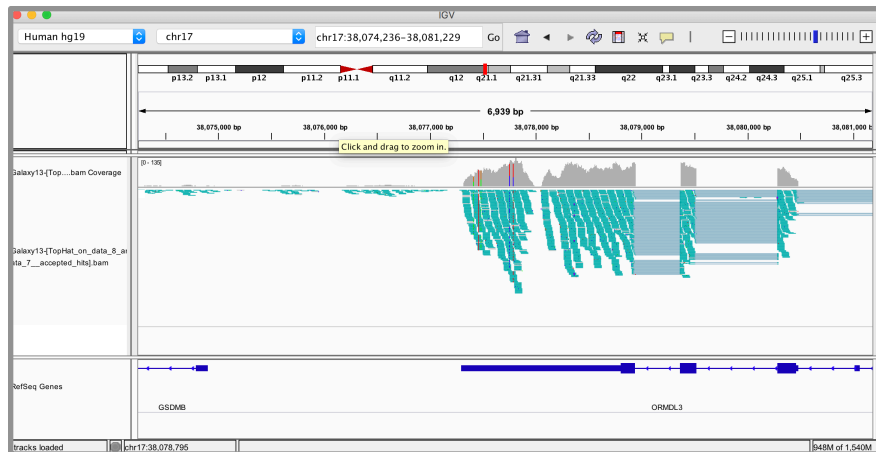
**Q11:** Where are most the accepted hits located?

[HINT, you can view the accepted

hits in **UCSC Genome Browser** via following the galaxy provided link. You can also search for a region, e.g. **chr17:38007296-38170000**]



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.



**Q12:** Following Q13, 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]

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” (available from the course website) into the workspace of Galaxy first. The following figure shows what parameters you need to change.

**Q13:** 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

#### **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/bgg213\\_f17/class-material/rs8067378\\_ENSG00000172057.6.txt](https://bioboot.github.io/bgg213_f17/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://goo.gl/forms/NXUnSuVTFvoU7WMD3>

#### **Reference:**

All data files can also be found at: [https://bioboot.github.io/bgg213\\_f17/lectures/#14](https://bioboot.github.io/bgg213_f17/lectures/#14) The second section of the lab is adapted from <https://usegalaxy.org/u/jeremy/p/galaxy-rna-seq-analysis-exercise>.

Verlaan, *et al.* Allele-specific chromatin remodeling in the ZBP2/ GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease. **Am. J. Hum. Genet.** 85: 377-393, 2009.