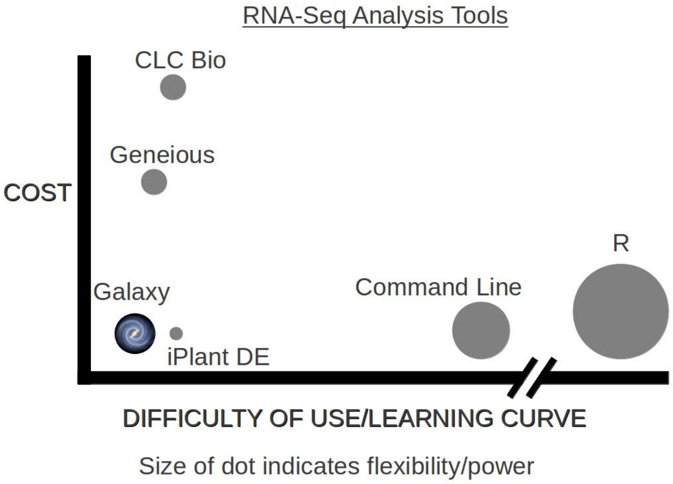
BIOINF525: INTRODUCTION TO BIOINFORMATICS LAB SESSION 4
Genome Informatics
Dr. Ryan E. Mills & Lauren Jepsen
Feb 2017

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].



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.

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 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

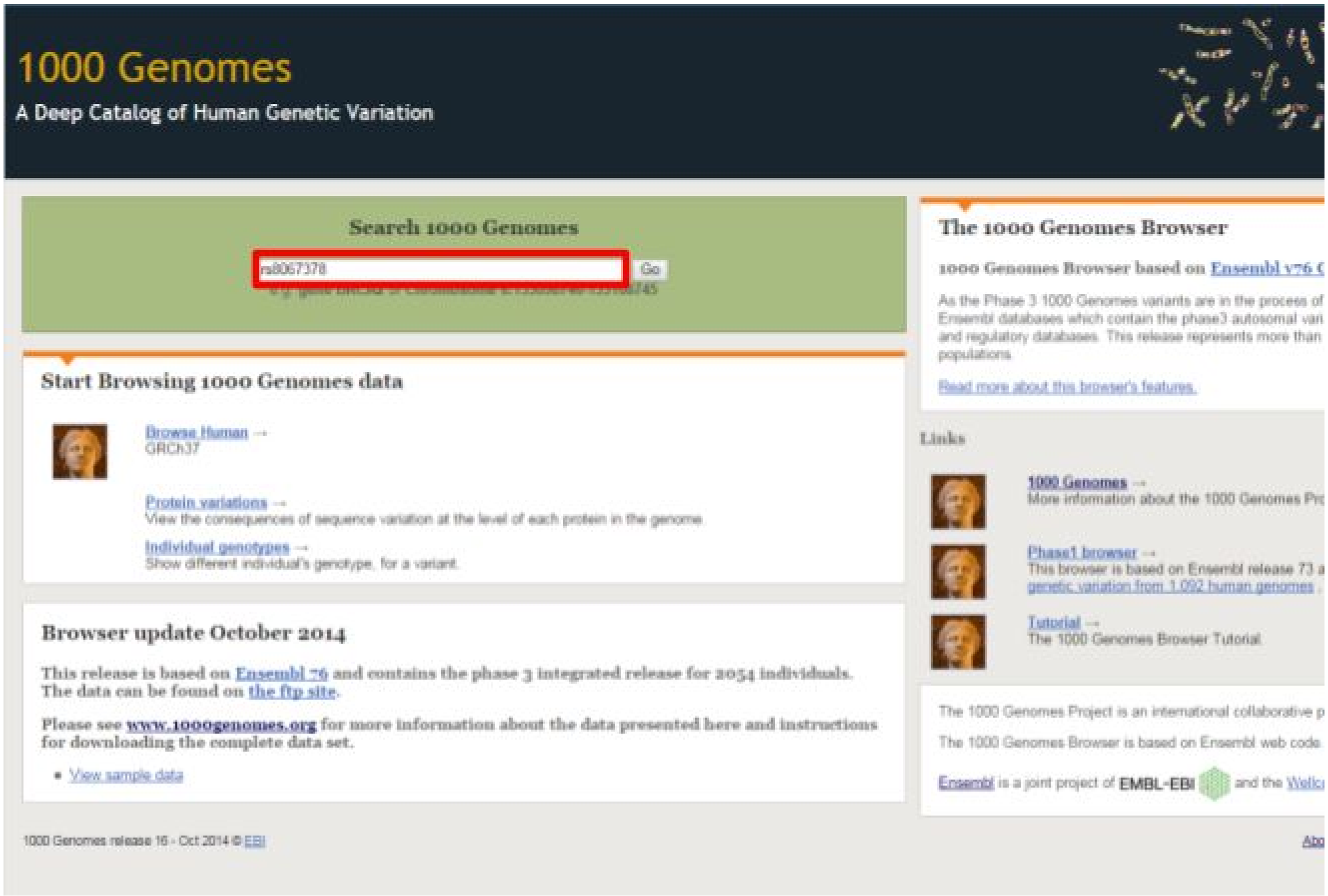
Q2: What are three genes be affected?
ZPBP2, GSDMB, and ORMDL3

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 a 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.



1000 Genomes

A Deep Catalog of Human Genetic Variation

Search 1000 Genomes

rs8067378

Start Browsing 1000 Genomes data

Browse Human →
GRCh37

Protein variations →
View the consequences of sequence variation at the level of each protein in the genome

Individual genotypes →
Show different individual's genotype, for a variant.

Browser update October 2014

This release is based on Ensembl 76 and contains the phase 3 integrated release for 2054 individuals. The data can be found on the ftp site.

Please see www.1000genomes.org for more information about the data presented here and instructions for downloading the complete data set.

- View sample data

1000 Genomes release 16 - Oct 2014 © EBI

The 1000 Genomes Browser

1000 Genomes Browser based on Ensembl v76 (

As the Phase 3 1000 Genomes variants are in the process of
Ensembl databases which contain the phase3 autosomal vari
and regulatory databases. This release represents more than
populations

Read more about this browser's features.

Links

1000 Genomes →
More information about the 1000 Genomes Pro

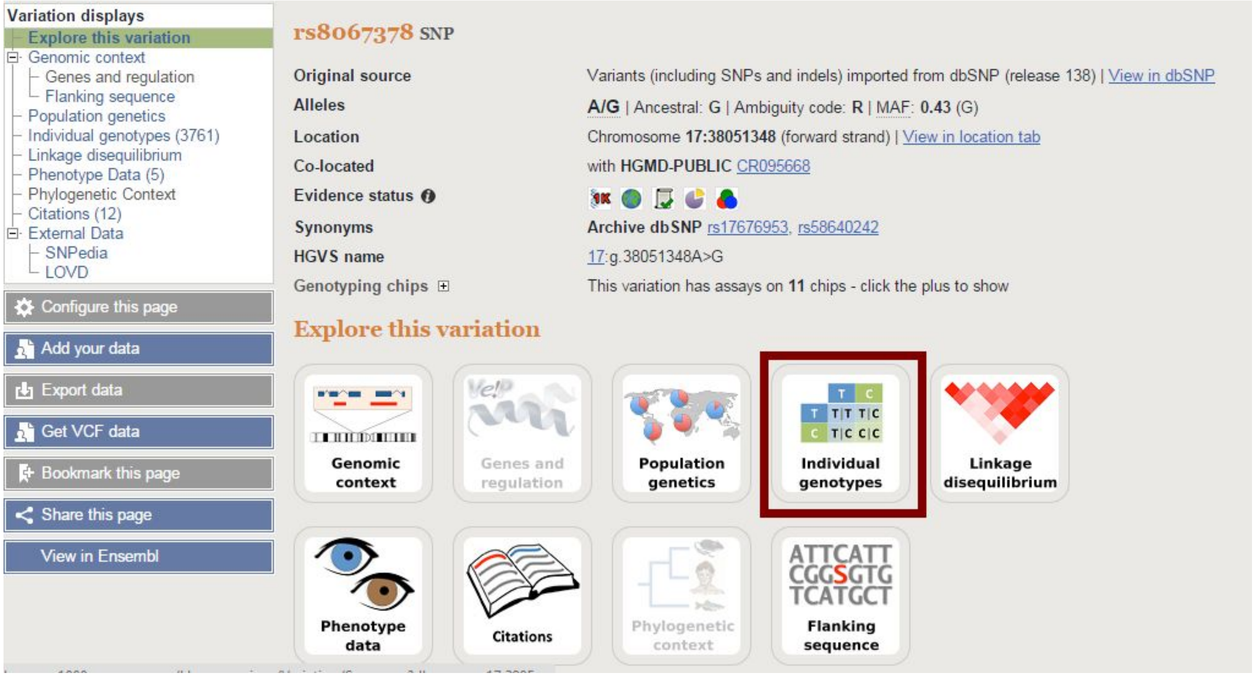
Phase1 browser →
This browser is based on Ensembl release 73 a
genetic variation from 1,092 human genomes .

Tutorial →
The 1000 Genomes Browser Tutorial

The 1000 Genomes Project is an international collaborative p

The 1000 Genomes Browser is based on Ensembl web code.

Ensembl is a joint project of EMBL-EBI and the Wellc



rs8067378 SNP

Original source	Variants (including SNPs and indels) imported from dbSNP (release 138) | View in dbSNP
Alleles	A/G | Ancestral: G | Ambiguity code: R | MAF: 0.43 (G)
Location	Chromosome 17:38051348 (forward strand) | View in location tab
Co-located	with HGMD-PUBLIC CR095668
Evidence status	
Synonyms	Archive dbSNP rs17676953, rs58640242
HGVS name	17:g.38051348A>G
Genotyping chips ⊞	This variation has assays on 11 chips - click the plus to show

Explore this variation

*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

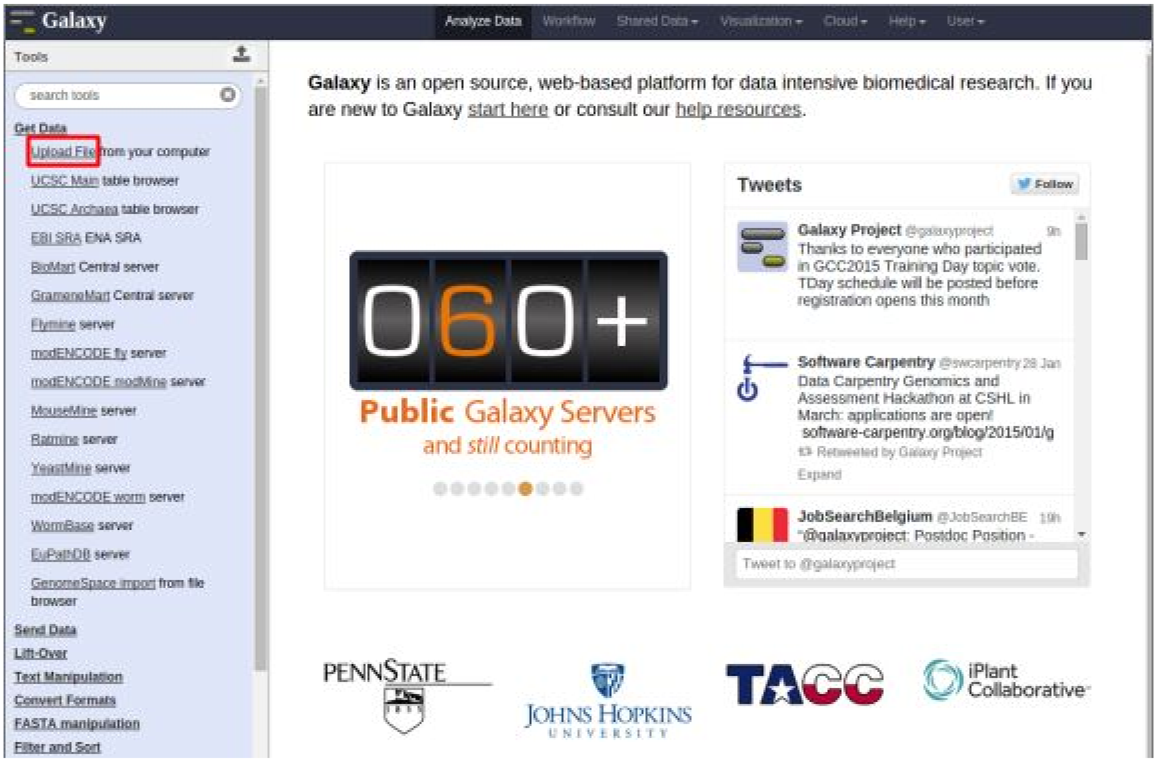
Now, you want to understand whether the SNP will affect the expression of the gene.

You find the RNA-Seq data of one sample on the class webpage (https://bioboot.github.io/bioinf525_w17/class-material/HG00109_1.fastq, https://bioboot.github.io/bioinf525_w17/class-material/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.

Accessing Galaxy

Please login to our local course instance of Galaxy at:
https://bcs2.bioinformatics.med.umich.edu:8080/

This will require you to login with your UM Level-1 (Kerberos) password 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.



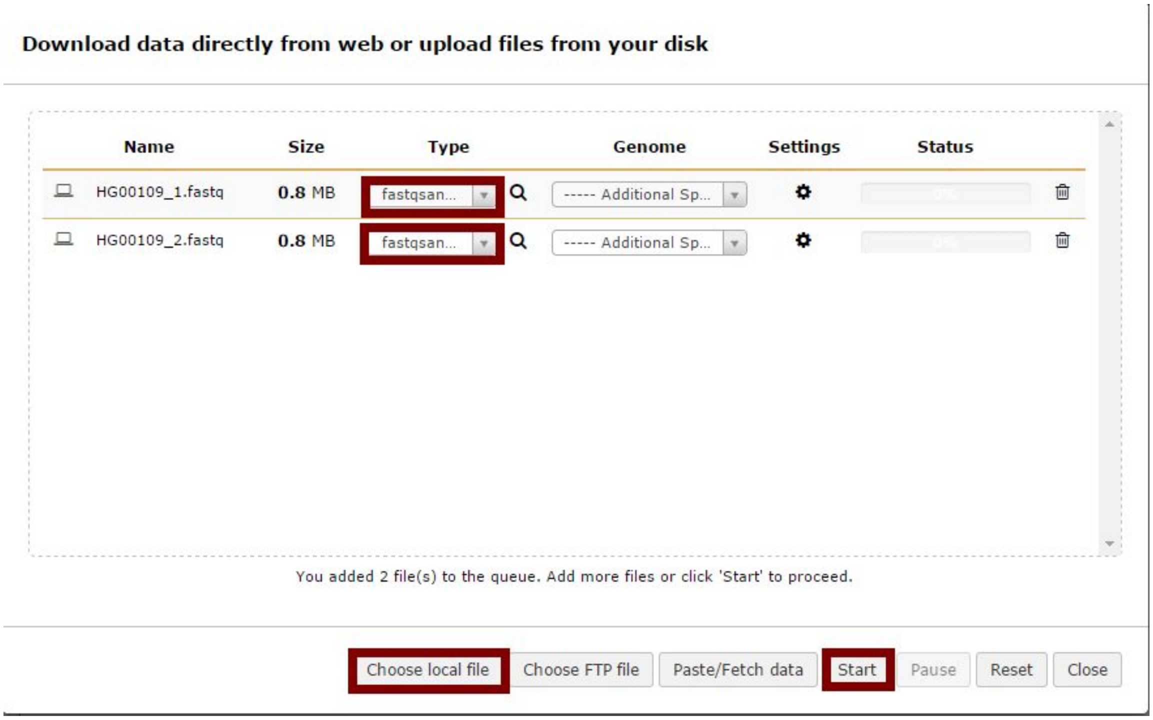
Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources.

Galaxy Project @galaxyproject 9h
Thanks to everyone who participated in GCC2015 Training Day topic vote. TDay schedule will be posted before registration opens this month

Software Carpentry @swcarpentry 28 Jan
Data Carpentry Genomics and Assessment Hackathon at CSHL in March: applications are open! software-carpentry.org/blog/2015/01/g
Retweeted by Galaxy Project
Expand

JobSearchBelgium @JobSearchBE 19h
"@galaxyproject: Postdoc Position -

Be careful of the file type. Tophat2 only takes fastqsanger file format. So, You need to choose **fastqsanger** for the 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...	⚙	0%	🗑
💻	HG00109_2.fastq	0.8 MB	fastqsan...	🔍	----- Additional Sp...	⚙	0%	🗑

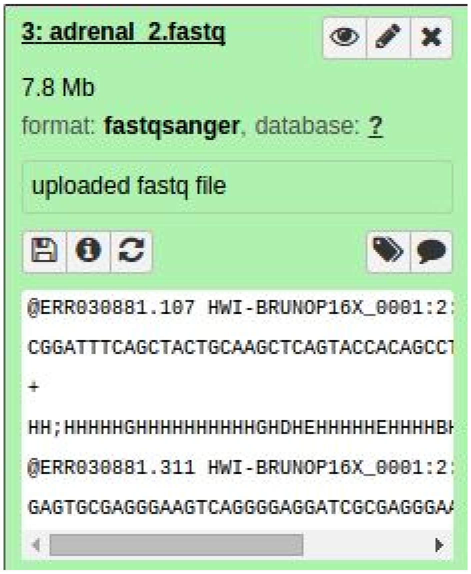
You added 2 file(s) to the queue. Add more files or click 'Start' to proceed.

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?]



3: adrenal_2.fastq

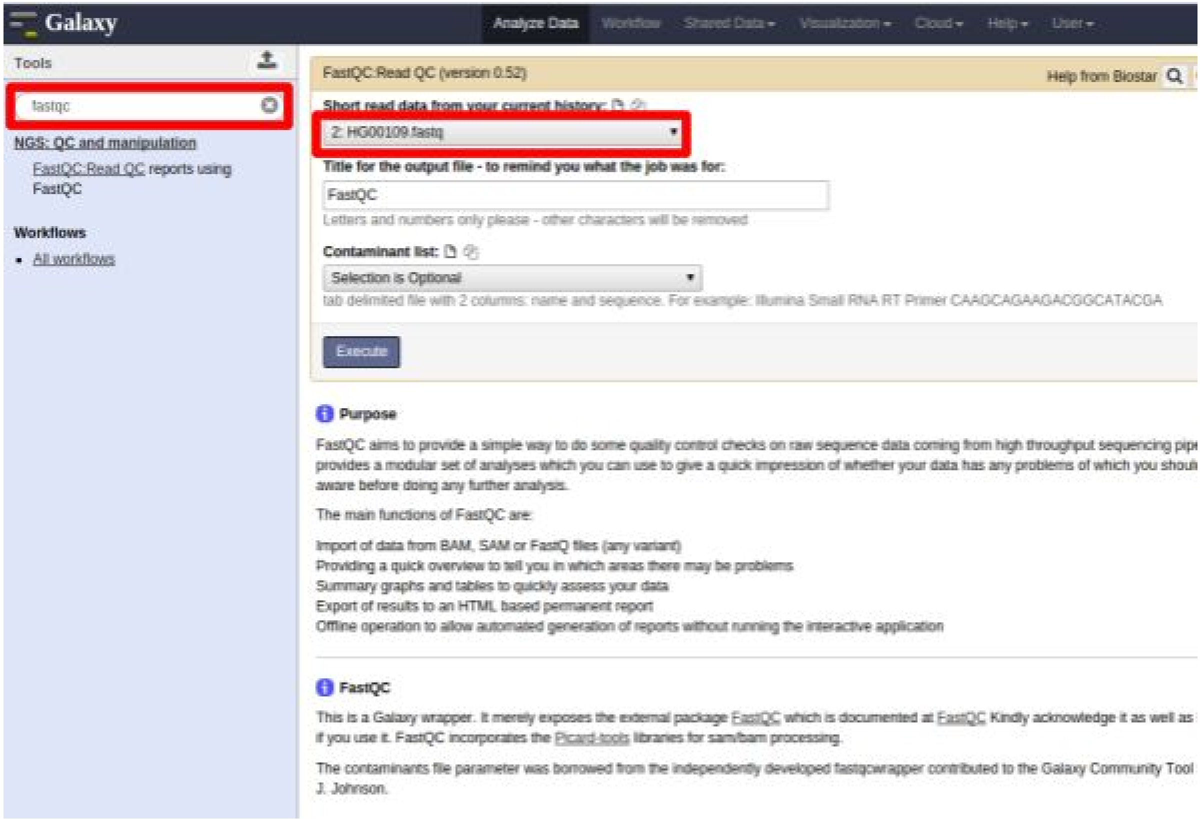
7.8 Mb

format: **fastqsanger**, database: ?

uploaded fastq file

```
@ERR030881.107 HWI-BRUNOP16X_0001:2:
CGGATTTCAGCTACTGCAAGCTCAGTACCACAGCCT
+
HH;HHHHHGHHHHHHHHHHGHDHEHHHHHHEHHHHBH
@ERR030881.311 HWI-BRUNOP16X_0001:2:
GAGTGCGAGGGAAGTCAGGGGAGGATCGCGAGGGAA
```

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.



fastqc

NGS: QC and manipulation

FastQC:Read QC reports using FastQC

Workflows

- All workflows

FastQC:Read QC (version 0.52)

Short read data from your current history:

2: HG00109.fastq

Title for the output file - to remind you what the job was for:

FastQC

Letters and numbers only please - other characters will be removed

Contaminant list:

Selection is Optional

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Purpose

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipe
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 shoul
aware before doing any further analysis.

The main functions of FastQC are:

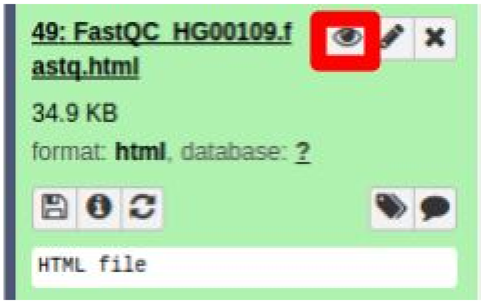
Import of data from BAM, SAM or FastQ files (any variant)
Providing a quick overview to tell you in which areas there may be problems
Summary graphs and tables to quickly assess your data
Export of results to an HTML based permanent report
Offline operation to allow automated generation of reports without running the interactive application

FastQC

This is a Galaxy wrapper. It merely exposes the external package FastQC which is documented at FastQC Kindly acknowledge it as well as
if you use it. FastQC incorporates the Picard-tools libraries for sam/bam processing.

The contaminants file parameter was borrowed from the independently developed fastqcwrapper contributed to the Galaxy Community Tool
J. Johnson.

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.



49: FastQC_HG00109.fastq.html

34.9 KB

format: **html**, database: ?

HTML file

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

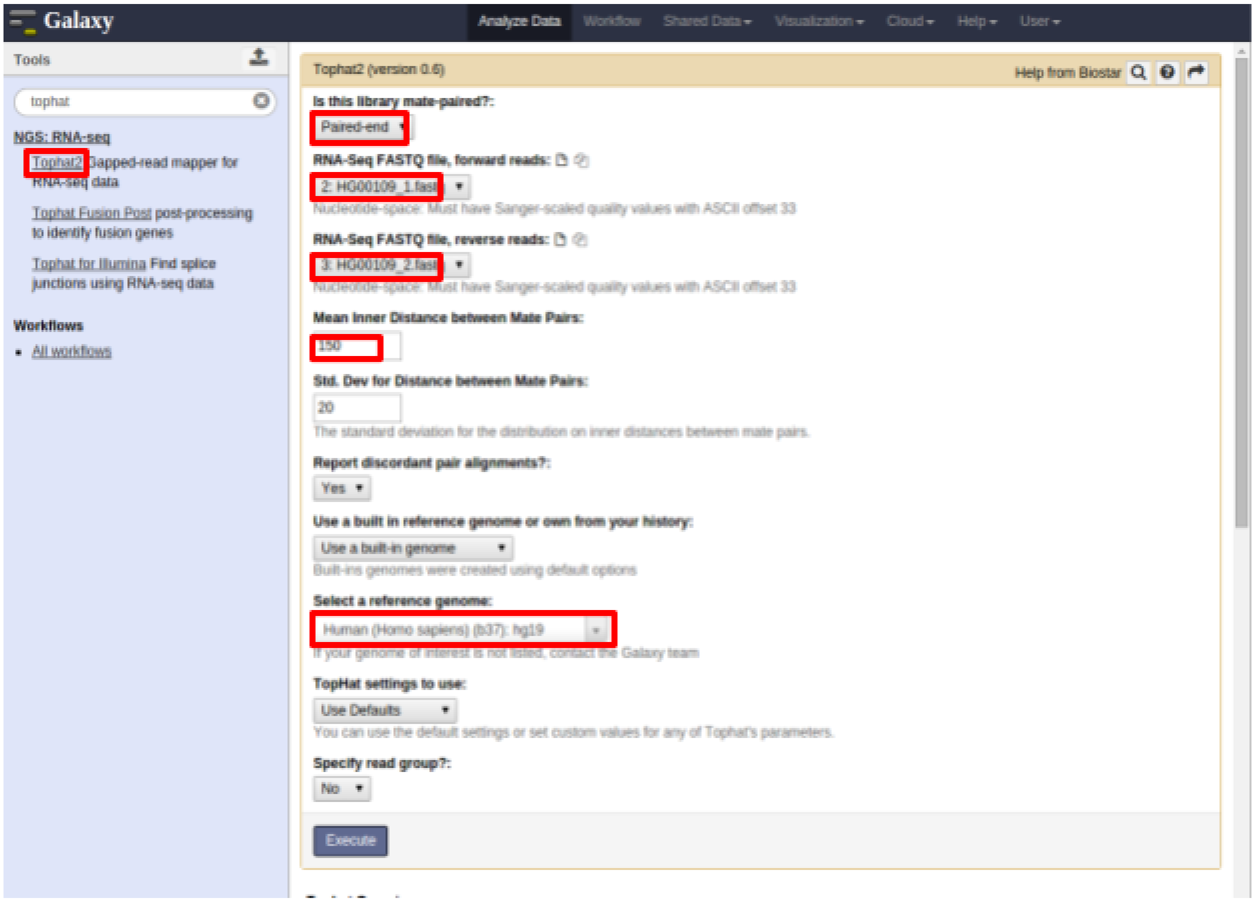
Q10: How about per base sequence quality? Does any base have a median quality score below 20?
[HINT, blue line is the median quality score.]

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.


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
Tools
tophat
NGS: RNA-seq
Tophat2 Gapped-read mapper for RNA-seq data
Tophat Fusion Post post-processing to identify fusion genes
Tophat for Illumina Find splice junctions using RNA-seq data
Workflows
All workflows
Help from Biostar
Execute

Tophat2 (version 0.6)

Is this library mate-paired?: Paired-end

RNA-Seq FASTQ file, forward reads: 2: HG00109_1.fastq

Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

RNA-Seq FASTQ file, reverse reads: 3: HG00109_2.fastq

Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Mean Inner Distance between Mate Pairs: 150

Std. Dev for Distance between Mate Pairs: 20

The standard deviation for the distribution on inner distances between mate pairs.

Report discordant pair alignments?: Yes

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

TopHat settings to use: Use Defaults

You can use the default settings or set custom values for any of Tophat's parameters.

Specify read group?: No

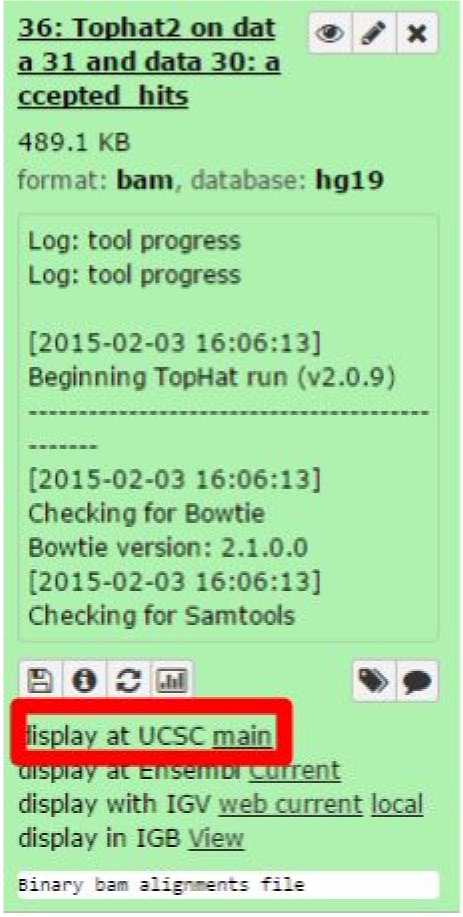
There will be four outputs: accepted_hits, insertions, deletions and splice junctions. You can visualize the accepted_hits on your favorite genome browser, like UCSC Genome Browser.

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:



36: Tophat2 on data 31 and data 30: accepted_hits

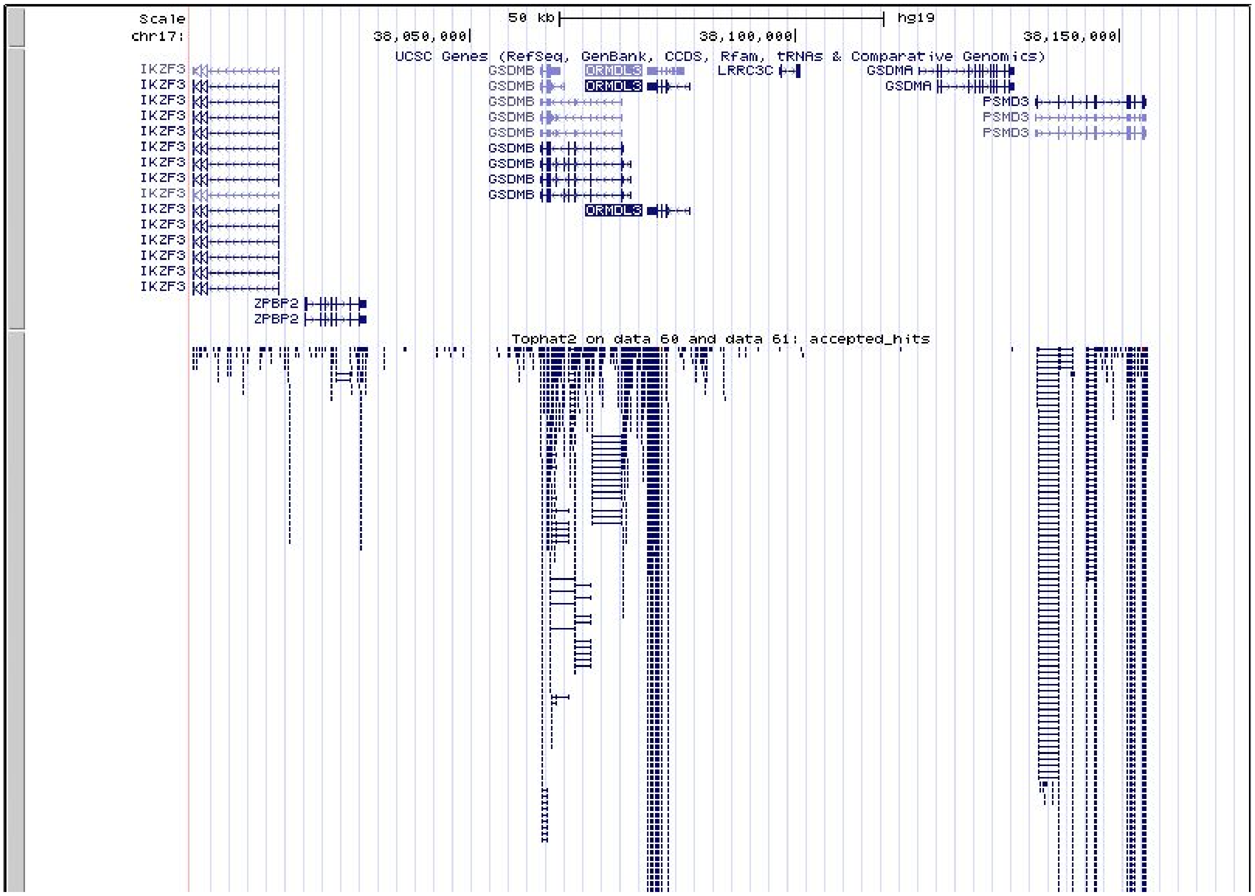
489.1 KB

format: **bam**, database: **hg19**

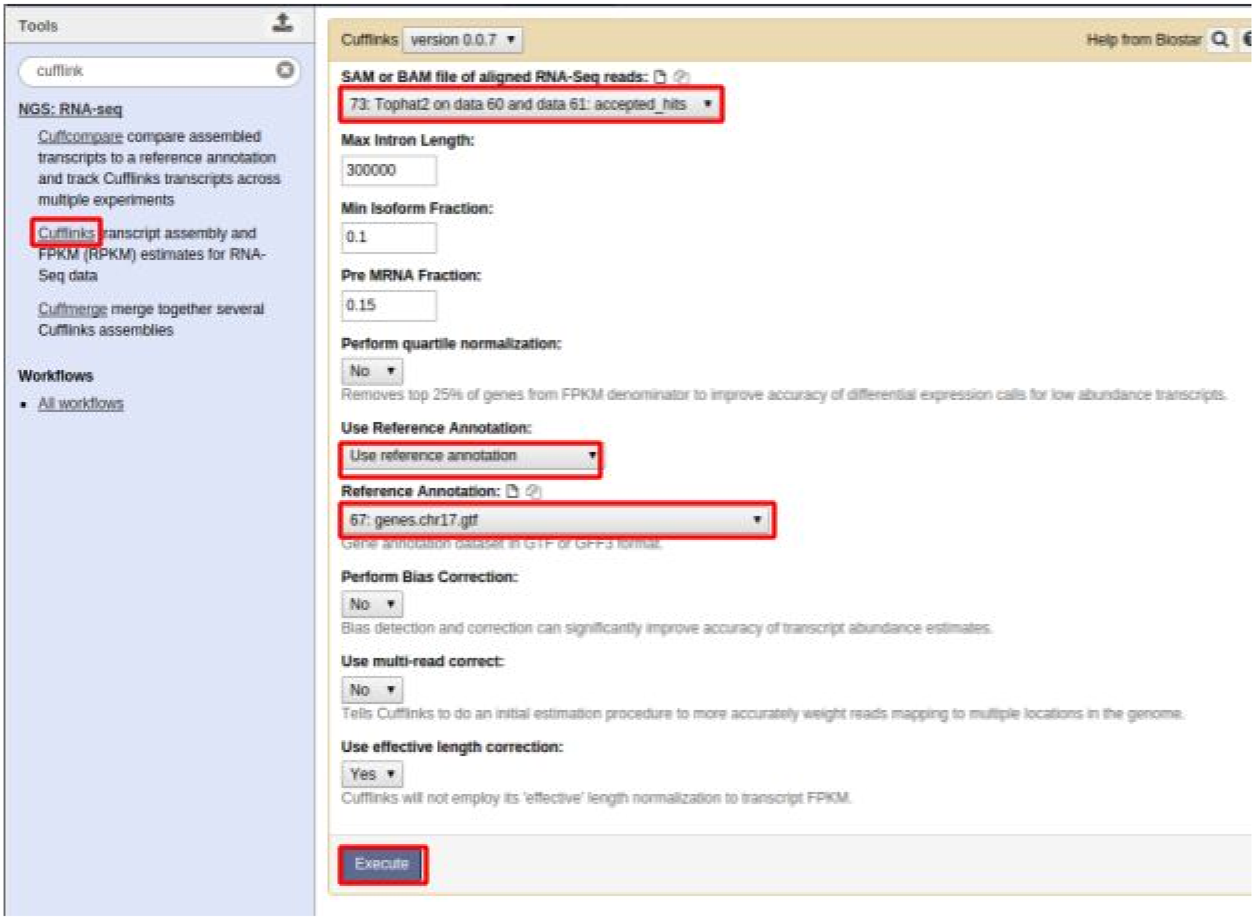
```
Log: tool progress
Log: tool progress

[2015-02-03 16:06:13]
Beginning TopHat run (v2.0.9)
-----------------------------------------
-------
[2015-02-03 16:06:13]
Checking for Bowtie
Bowtie version: 2.1.0.0
[2015-02-03 16:06:13]
Checking for Samtools
```

display at UCSC main
display at Ensembl Current
display with IGV web current local
display in IGB View

Binary bam alignments file


Scale
chr17:
50 kb
hg19
38,050,000
38,100,000
38,150,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
IKZF3
GSDMB
ORMDL3
LRRC3C
GSDMA
PSMD3
ZPBP2
Tophat2 on data 60 and data 61: accepted_hits

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.



NGS: RNA-seq

Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments

Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data

Cuffmerge merge together several Cufflinks assemblies

Workflows

- All workflows

Cufflinks version 0.0.7

SAM or BAM file of aligned RNA-Seq reads: 73: Tophat2 on data 60 and data 61: accepted_hits

Max Intron Length: 300000

Min Isoform Fraction: 0.1

Pre MRNA Fraction: 0.15

Perform quartile normalization: No

Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low abundance transcripts.

Use Reference Annotation: Use reference annotation

Reference Annotation: 67: genes.chr17.gtf

Gene annotation dataset in GTF or GFF3 format.

Perform Bias Correction: No

Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

Use multi-read correct: No

Tells Cufflinks to do an initial estimation procedure to more accurately weight reads mapping to multiple locations in the genome.

Use effective length correction: Yes

Cufflinks will not employ its 'effective' length normalization to transcript FPKM.

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

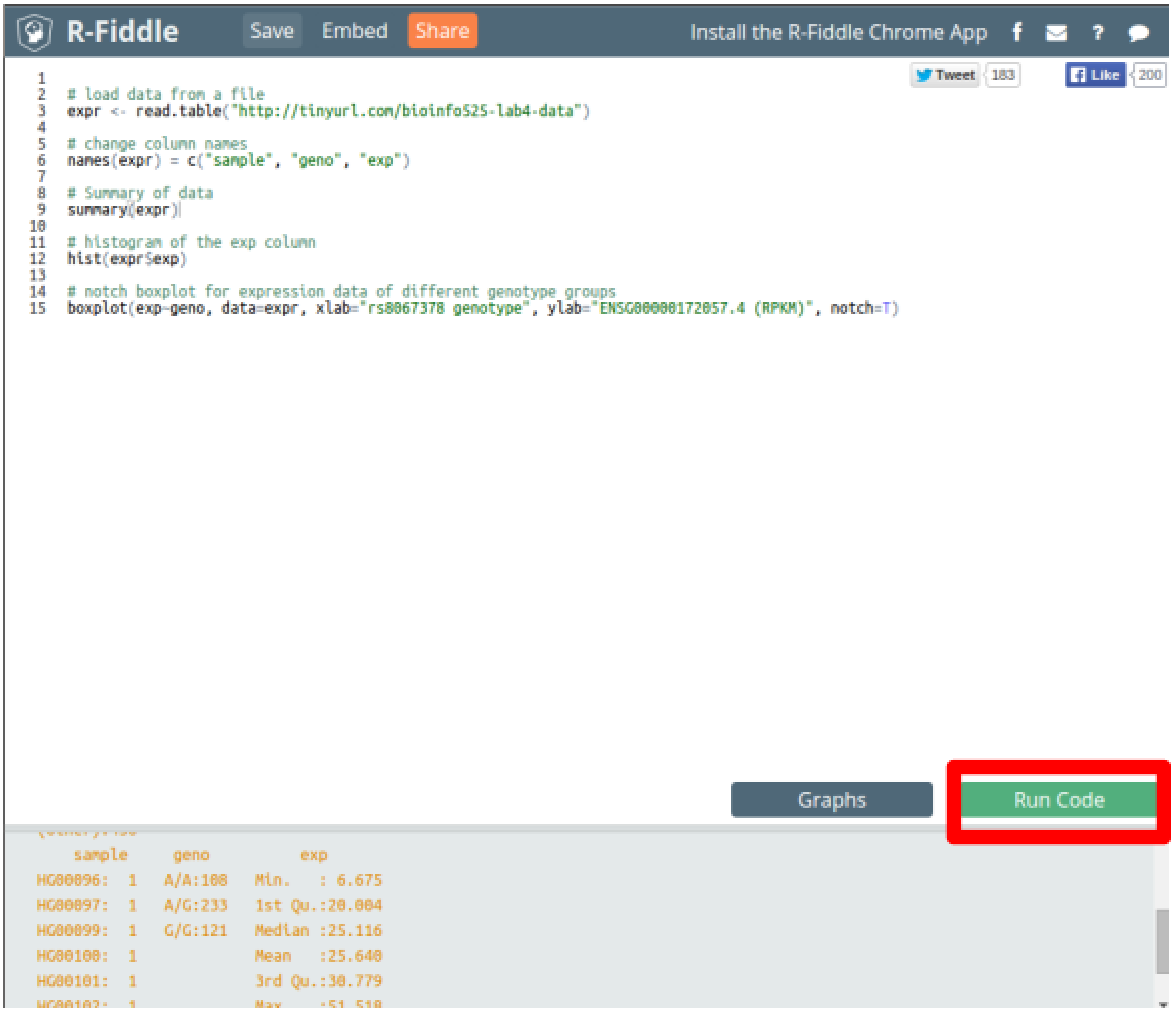
136853

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)



```r

# load data from a file
expr <- read.table("http://tinyurl.com/bioinfo525-lab4-data")

# change column names
names(expr) = c("sample", "geno", "exp")

# Summary of data
summary(expr)

# histogram of the exp column
hist(expr$exp)

# notch boxplot for expression data of different genotype groups
boxplot(exp~geno, data=expr, xlab="rs8067378 genotype", ylab="ENSG00000172057.4 (RPKM)", notch=T)
```

```
    sample      geno          exp        
 HG00096:  1   A/A:108   Min.   : 6.675  
 HG00097:  1   A/G:233   1st Qu.:20.004  
 HG00099:  1   G/G:121   Median :25.116  
 HG00100:  1             Mean   :25.640  
 HG00101:  1             3rd Qu.:30.779  
 HG00102:  1             Max.   :51.510  
```

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://bioboot.github.io/bioinf525_w17

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 .