

BIOINF525: INTRODUCTION TO BIOINFORMATICS LAB SESSION 4

Genome Informatics

Dr. Ryan E. Mills & Hongyang Li

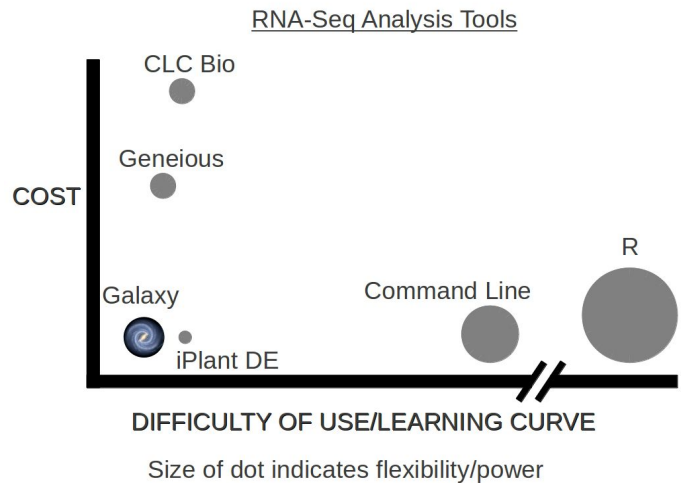
Feb 2016

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 on Ctools].

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.

Registering for a Galaxy account

First create an account on the main public Galaxy portal @ <https://main.g2.bx.psu.edu/>

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



This will only take a moment, 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.

Section 1: Find the interesting genome variance

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 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.

The screenshot displays the 1000 Genomes browser interface. At the top left, the text reads "1000 Genomes" and "A Deep Catalog of Human Genetic Variation". A search bar titled "Search 1000 Genomes" contains the text "rs8067378" and a "Go" button. Below the search bar, a snippet of text indicates the location: "Gene: BRCA2 of Chromosome 6: 13206746-132108745".

The main content area is divided into several sections:

- Start Browsing 1000 Genomes data:** This section includes a small portrait icon and three links: "Browse Human" (GRCh37), "Protein variations" (View the consequences of sequence variation at the level of each protein in the genome), and "Individual genotypes" (Show different individual's genotype, for a variant).
- Browser update October 2014:** This section states that the release is based on Ensembl 76 and contains the phase 3 integrated release for 2054 individuals. It provides a link to the ftp site and a link to view sample data.

On the right side of the page, there is a section titled "The 1000 Genomes Browser" which includes a description of the browser based on Ensembl v76 and a link to read more about its features. Below this is a "Links" section with three items: "1000 Genomes" (More information about the 1000 Genomes Project), "Phase1 browser" (This browser is based on Ensembl release 73 and genetic variation from 1,092 human genomes), and "Tutorial" (The 1000 Genomes Browser Tutorial).

At the bottom of the page, there is a footer that reads "1000 Genomes release 16 - Oct 2014 © EBI" and a small logo for the Wellcome Trust.

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

Configure this page

Add your data

Export data

Get VCF data

Bookmark this page

Share this page

View in Ensembl

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

- Genomic context
- Genes and regulation
- Population genetics
- Individual genotypes**
- Linkage disequilibrium
- Phenotype data
- Citations
- Phylogenetic context
- Flanking sequence

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 CTools (HG00109_1.fastq, 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.

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...		
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. So, you will have better understanding about what each column/row represent.

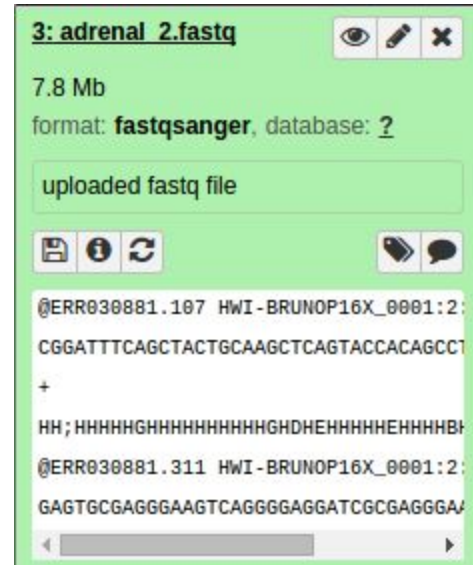
Q6: What is the size and format of the data?

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?]



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.

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User

Tools fastqc

NGS: QC and manipulation
FastQC:Read QC reports using FastQC

Workflows
All workflows

FastQC:Read QC (version 0.52) Help from Biostar

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 CAA0CAGAAAGACGGCATAACGA

Execute

Purpose
FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pip. 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.
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.

The screenshot shows the Galaxy web interface for the Tophat2 tool. The left sidebar contains a search bar with 'tophat' and a list of tools under 'NGS: RNA-seq', with 'Tophat2' selected. The main panel displays the tool configuration for 'Tophat2 (version 0.6)'. The configuration includes the following fields and options:

- Is this library mate-paired?:** Paired-end (selected)
- RNA-Seq FASTQ file, forward reads:** 2. HG00109_1.fast (selected)
- RNA-Seq FASTQ file, reverse reads:** 3. HG00109_2.fast (selected)
- Mean Inner Distance between Mate Pairs:** 150 (input field)
- Std. Dev for Distance between Mate Pairs:** 20 (input field)
- Report discordant pair alignments?:** Yes (selected)
- Use a built in reference genome or own from your history:** Use a built-in genome (selected)
- Select a reference genome:** Human (Homo sapiens) (b37): hg19 (selected)
- TopHat settings to use:** Use Defaults (selected)
- Specify read group?:** No (selected)

An 'Execute' button is located at the bottom of the configuration panel.

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

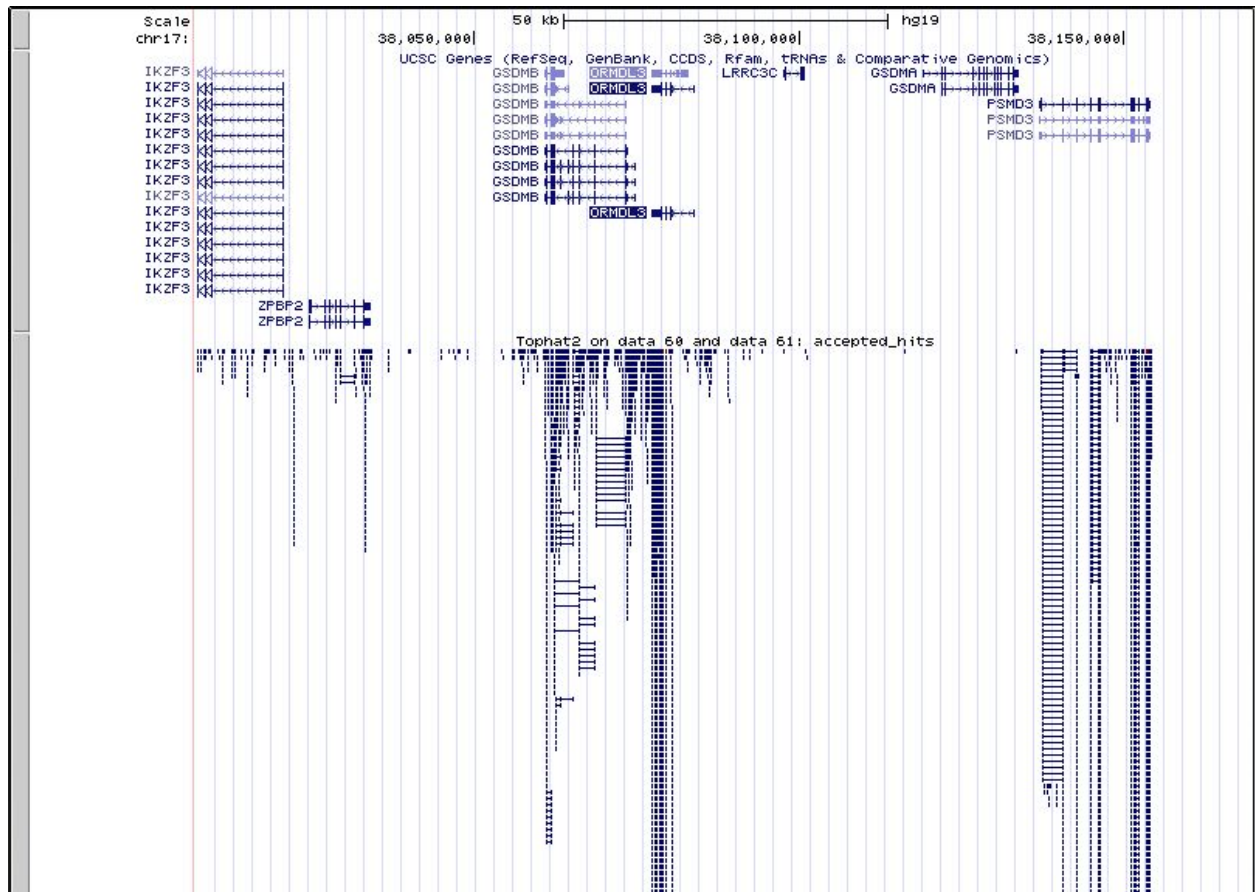
[2015-02-03 16:06:13]
Beginning TopHat run (v2.0.9)

[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



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.

The screenshot shows the Cufflinks web interface. On the left sidebar, under 'NGS: RNA-seq', the 'Cufflinks' option is highlighted with a red box. The main configuration panel on the right has several settings:

- SAM or BAM file of aligned RNA-Seq reads:** A dropdown menu with the value '73: Tophat2 on data 60 and data 61: accepted_hits' highlighted in red.
- Max Intron Length:** Input field with value '300000'.
- Min Isoform Fraction:** Input field with value '0.1'.
- Pre mRNA Fraction:** Input field with value '0.15'.
- Perform quartile normalization:** Dropdown menu with value 'No'.
- Use Reference Annotation:** Dropdown menu with value 'Use reference annotation' highlighted in red.
- Reference Annotation:** A dropdown menu with the value '67: genes.chr17.gtf' highlighted in red.
- Perform Bias Correction:** Dropdown menu with value 'No'.
- Use multi-read correct:** Dropdown menu with value 'No'.
- Use effective length correction:** Dropdown menu with value 'Yes'.

At the bottom of the configuration panel, there is a red 'Execute' button.

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)

The screenshot shows the R-Fiddle web interface. At the top, there are buttons for 'Save', 'Embed', and 'Share', along with an option to 'Install the R-Fiddle Chrome App'. Below the header, there is a code editor with the following R code:

```
1
2 # load data from a file
3 expr <- read.table("http://tinyurl.com/bioinfo525-lab4-data")
4
5 # change column names
6 names(expr) = c("sample", "geno", "exp")
7
8 # Summary of data
9 summary(expr)
10
11 # histogram of the exp column
12 hist(expr$exp)
13
14 # notch boxplot for expression data of different genotype groups
15 boxplot(exp~geno, data=expr, xlab="rs8867378 genotype", ylab="ENSG00000172057.4 (RPKM)", notch=T)
```

Below the code editor, there are two buttons: 'Graphs' and 'Run Code'. The 'Run Code' button is highlighted with a red box. Below the buttons, the output of the R code is displayed in a table format:

sample	geno	exp
HG00096: 1	A/A:108	Mean : 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://github.com/ajing/Bioinfo525_lab4

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> .