Class19: Genome Informatics

Chloe J. Welch

11/30/2021

Introduction to Genome Informatics Lab

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?

–> The 4 candidate SNPs are: rs12936231, rs8067378, rs9303277, and rs7216389.

Q2. What three genes do these variants overlap or effect?

–> The three genes are ZPBP2, IKZF3, and GSDMB.

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?

–> The location is Chromosome 17:39895095, and the different alleles are A/C/G|Ancestral: G|MAF: 0.43.

Q4. Name at least 3 downstream genes for rs8067378?

–> Three downstream genes are PSMD3, CSF3, and RARA.

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)?

Proportion of G/G in a Population

First, we began by downloading a CSV file from Ensembl: < https://uswest.ensembl.org/Homo_sapiens/Variation/Sample?db=core;r=17:39835097-39955098;v=rs8067378;vdb=variation;vf=105535077#373531_tablePanel >

We will now read the file:

```
mxl <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")
head(mxl)
```

```
##   Sample..Male.Female.Unknown. Genotype..forward.strand. Population.s. Father
## 1                  NA19648 (F)                       A|A ALL, AMR, MXL      -
## 2                  NA19649 (M)                       G|G ALL, AMR, MXL      -
## 3                  NA19651 (F)                       A|A ALL, AMR, MXL      -
## 4                  NA19652 (M)                       G|G ALL, AMR, MXL      -
## 5                  NA19654 (F)                       G|G ALL, AMR, MXL      -
## 6                  NA19655 (M)                       A|G ALL, AMR, MXL      -
##   Mother
## 1      -
## 2      -
## 3      -
## 4      -
## 5      -
## 6      -
```

What is the proportion of G/G?

```
table(mxl$Genotype..forward.strand.) / nrow(mxl) * 100
```

```
##
##     A|A     A|G     G|A     G|G
## 34.3750 32.8125 18.7500 14.0625
```

–> 14.06% of the Mexican Ancestry in Los Angeles sample population (MXL) are homozygous for the asthma-associated SNP (G/G).

Let's compare to another group. We will now download the data for the GBR (Great Britain) population: < https://uswest.ensembl.org/Homo_sapiens/Variation/Sample?db=core;r=17:39835097-39955098;v=rs8067378;vdb=variation;vf=105535077#373522_tablePanel >

```
gbr <- read.csv("373522-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")
head(gbr)
```

```
##   Sample..Male.Female.Unknown. Genotype..forward.strand. Population.s. Father
## 1                  HG00096 (M)                       A|A ALL, EUR, GBR      -
## 2                  HG00097 (F)                       G|A ALL, EUR, GBR      -
```

```
## 3                  HG00099 (F)                      G|G ALL, EUR, GBR      -
## 4                  HG00100 (F)                      A|A ALL, EUR, GBR      -
## 5                  HG00101 (M)                      A|A ALL, EUR, GBR      -
## 6                  HG00102 (F)                      A|A ALL, EUR, GBR      -
##   Mother
## 1      -
## 2      -
## 3      -
## 4      -
## 5      -
## 6      -
```

What is the proportion of G/G?

```
table(gbr$Genotype..forward.strand.) / nrow(gbr) * 100
```

```
##
##      A|A      A|G      G|A      G|G
## 25.27473 18.68132 26.37363 29.67033
```

The variant that is associated with childhood asthma is more frequent in the GBR population than in the MXL population. We will now explore this further.

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?

–> The genotype for this sample (a male) is 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.

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

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/

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).

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!

–> There are 3,863 sequences in the first file. The file size is 741.9 KB and the format of the data is "fastqsanger".

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.

Q8. What is the GC content and sequence length of the second fastq file?

–> The GC content of the second fastq file is 54%. The sequence length is 50-75.

Q9. How about per base sequence quality? Does any base have a mean quality score below 20?

–> The per base sequence quality (on average) is above 30 and in the green region. There are no bases with a mean quality score below 20, so trimming is not completely critical. This would be more helpful to improve low quality reads.

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.

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.

Q10. Where are most the accepted hits located?

–> Most of the accepted hits are located on chromosome 17 between positions 38,050,000 and 38,150,000.

Q11. Following **Q10**, is there any interesting gene around that area?

–> Yes— there are a few interesting genes around the area including IKZF3, ZPBP2, GSDMB and ORMDL3 just to name a few.

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).

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 the ORMDL3 gene is 136853. Other genes with above zero FPKM values are GSDMA, GSDMB, and ZPBP2.

NOTE: Sections 2 and 3 were completed using Galaxy and the UCSC genome browser.

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.

How many samples do we have?

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.

```
expr <- read.table("rs8067378_ENSG00000172057.6.txt")
head(expr)
```

```
##    sample geno      exp
## 1 HG00367  A/G 28.96038
## 2 NA20768  A/G 20.24449
## 3 HG00361  A/A 31.32628
## 4 HG00135  A/A 34.11169
## 5 NA18870  G/G 18.25141
## 6 NA11993  A/A 32.89721
```

```
nrow(expr)
```

```
## [1] 462
```

```
table(expr$geno)
```

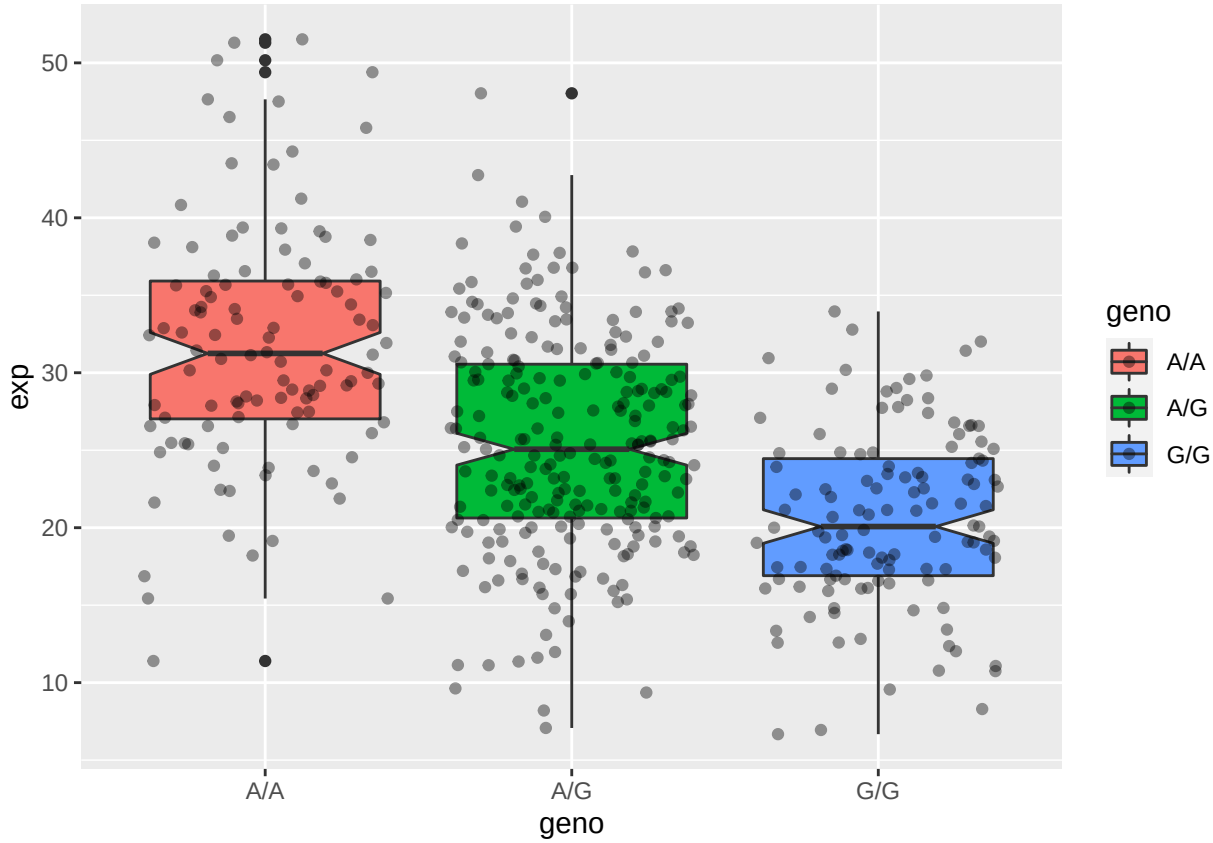
```
##
## A/A A/G G/G
## 108 233 121
```

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 affect the expression of ORMDL3?

Let's call `ggplot` so we can generate a summary figure.

```
library(ggplot2)
```

```
ggplot(expr) + aes(geno, exp, fill = geno) +
  geom_boxplot(notch = TRUE) +
  geom_jitter(alpha = 0.4)
```



Based on this plot, we can infer that the relative expression value of A/A is higher than that of G/G. Because we see variation in the expression levels between the three different genotypes, there could be association between the asthma-related SNPs and the ORMDL3 gene due to the genetic changes resulting in these different genotypes.