Assignment 3

Introduction

We have taken a tissue sample from a genetically modified mouse and performed RNA sequencing with the intention of exploring the expression levels. Attached to this assignment is a BAM file for one of those individual cells (in '.sam' format to be readable). This file contains reads which have been aligned to the mouse genome.

A BAM file is the standard data template for all of the genetic information in a library. As such, it is the starting point for many analyses approaches. We would like you to become familiar with the contents of a BAM file and their biological implications by investigating the contents of this file.

Use the tables linked to in the appendix to help you interpret the contents of the file.

In your answers, please list all relevant code used and results found.

Data and Requirements

Data for Q1-Q2: single_cell_RNA_seq_bam.sam

Data for Q3: RNA_seq_annotated_variants.vcf

Questions

Q1. Sequencing technologies

Why areas of the genome with high GC content are hard to sequence?

Q2. Global alignment exercise

Similar to the approach for Needleman–Wunsch algorithm, find the best global alignment between the two following sequences:

ATTCGAC

ATCAC

Use a gap penalty of -2 and the following scoring matrix:

Transition
Transversion
Matrix

	A	T	C	G
A	1	–5	–5	–1
T	–5	1	–1	–5
C	–5	–1	1	–5
G	–1	–5	–5	1

In your answer, please include the grid table (used for storing the scores and traceback) and also include how you calculated the first top-left 9 elements of the table.

Q3. Looking at the Metadata of an alignment (SAM) file

Q3.1. Use `read.csv("single_cell_RNA_seq_bam.sam", nrows=73, sep="\t", header=FALSE, fill=TRUE)` to load the first 73 lines of the header of the file and print the contents. These lines contain tabulated information about the BAM file and the circumstances of its data collection. According to the header table in section 1.3 of the BAM/SAM document in the appendix, what do the `SN` and `LN` tags indicate?

Q3.2. A sequence is any template string of bases to which we can align a read. This includes chromosomes (which are continuous sequences of bases) and new strings resulting from genetic modifications. What is the length of the X chromosome, in bp, for our alignment?

Fun fact (not tested): One of the sequences in this BAM file titled Cre_ERT2 is a Cre-recombinase variant. Cre recombinase, when combined with the loxP sequence (see cre-lox recombination) is the primary element used in many experiments (such as this one) to induce a genetic modification to certain cells in vivo. This allows us to study the effect of changing a gene at any time during the life cycle of an organism, and has allowed us to make discoveries in many areas including stem cell research.

Q4. Looking at the Reads of an alignment (SAM) file

Q4.1. Use the code below to load the reads into an R dataframe. Each row contains one read. How many reads are there in this BAM file?

```
sam <- read.csv("single_cell_RNA_seq_bam.sam", sep="\t", header=FALSE,
comment.char="@", col.names = paste0("V",seq_len(30)), fill=TRUE)
sam <- sam[paste0("V",seq_len(11))]
```

Q4.2. Print out the 10th row of the dataframe to look at the format of a read. Compare it to the mandatory BAM fields table in section 1.4 of the SAM/BAM documentation in the appendix. **The order of columns in the bam file have been preserved in the dataframe**. Which column of your dataframe should you look at to find the chromosome to which the read was aligned? To which BAM data field does the dataframe column "V11" correspond?

Q4.3. How many reads in this file align to chromosome X?

Hint: You can compare a column vector to a constant using logical symbols (==, <, >, etc.) to get a column vector of TRUE or FALSE. Remember, when summing, a true symbol is worth "1" while a false symbol is worth "0".

Q4.4. What is the mean base quality (BQ) for reads aligning to chromosome X?

Q4.5. Plot the distribution of BQs across all bases and reads as a boxplot. Comment on your observation.

Hint: This is similar to a boxplots that was provided in the lecture related to primary analysis.

Q4.6. Referring to section 1.4 of the SAM/BAM documentation, what column contains the leftmost mapping position of the reads?

Q4.7. In order to transform a BAM file into expression levels for each gene, we need to count the number of reads covering a particular location or gene. The protein Hspa8 is located on chromosome 9 at bases 40801273 - 40805199. How many reads have their leftmost mapping position aligned within these coordinates?

Hint: you can implement AND logic on two column vectors with "&".

Q4.8. Mapping quality is an indication of how well a read aligned to the reference genome during the alignment step of processing our library data. It is reported as an integer between 0 and 255. How many reads have mapping quality less than 50?

Q4.9. What is the mean mapping quality of the reads which have mapping quality less than 50?

Hint: you can obtain a subset of a dataframe by using df[bool_vec,] where bool_vec contains TRUE/FALSE elements and bool_vec and df have the same number of rows.

Q4.10. (bonus): The genome of the mouse used in this experiment has been edited to include the DNA sequence for the protein 'tdTomato', which is a fluorophore. Count the number of reads which align to the tdTomato sequence. Assuming that these reads are accurate, would you expect this cell to emit fluorescently? What might be the purpose of modifying a genome to include a fluorophore?

Hint: Think about studying cell populations under a microscope.

Q5. Investigating the Variants

We have used Strelka, which is a variant-calling tool, to find all of the SNPs and short indels in the genome of this cell using the BAM file. The variants were then annotated using snpEff to label them with information such as which gene they affect and the type of modification they result in once the RNA undergoes translation to a protein. The results are in a VCF file (extension '.vcf') which is attached.

Q5.1. Use the following lines of code to obtain the header of the file and a dataframe where each row is a variant. As you can see, information in the VCF file is organised by multiple levels of character-separated data, so it will take multiple rounds of parsing to extract relevant information. For the first variant (row) in the dataframe, what is the reference allele base at the

site, and what is the alternative allele called by Strelka?

Hint: Take a look at the VCF Variant Call Format document in the appendix for details on each column name.

```
vcf_con <- file("RNA_seq_annotated_variants.vcf", open="r")
vcf_file <- readLines(vcf_con)
close(vcf_con)
vcf <- data.frame(vcf_file)
header <- vcf[grepl("##", vcf$vcf_file), ]
factor(header)
variants <- read.csv("RNA_seq_annotated_variants.vcf", skip=length(header),
header=TRUE, sep="\t")
```

Q5.2. The INFO field is organised into variables by the form 'TAG=value' (see the VCF Variant Call Format document). Write code to obtain the entirety of the ANN info value contents from the INFO field for the first variant.

Hint: You will need strsplit() and grep()/grepl() to accomplish this. Take a look at https://www.math.ucla.edu/~anderson/rw1001/library/base/html/strsplit.html and https://stackoverflow.com/questions/21311386/using-grep-to-help-subset-a-data-frame-in-r for how to make use of them. With which character should you split the string?

Hint: Make sure to convert the INFO field entry to string format using as.character() so that it can be passed into strsplit().

Q5.3. Each INFO tag-value pair is detailed in a line of the header, beginning with the tag '##INFO=<ID=VARIABLE, ...'. Look for the header entry starting with '##INFO=<ID=ANN, ...' which details the format of the ANN value contents. This tag-value pair contains the results of the annotations found by snpEff. Based on the ANN value of the first variant, what does the 'Annotation' field tell us about this variant?

Hint: snpEff can return multiple annotation entries for the same variant because some variants may have multiple possible effects. The first annotation entry is the most confident/important and, resultantly, you should only look at the first entry to answer this and all subsequent question. You can use strsplit() again with ',' separation character if you wish to look at each of the ANN entries separately.

Hint: Refer to the snpEff documentation in the appendix for a list of snpEff annotation label names and summaries of their effects.

Q5.4. Perform the parsing done in Q5.1-3 again on variant line 683. What gene would this variant affect?

Q5.5. Within the entire VCF file, how many variants (in total) do we have per type (synonymous/nonsynonymous SNVs, frameshift indels, etc.)?

Q5.6. What is a frameshift variant? Does it have a greater or lesser effect on the resultant protein than a missense variant? Why?

Q5.7. We can divide variants into two broad categories: intronic/intergenic and exonic. Count the number of potential intronic variants. What do you notice about the number of intronic variants (compared to overall number of variants)?

Hint: Use grepl() on the INFO field to look for tell-tale tags.

Hint: assume no overlap between exonic and intronic tags within a variant entry.

Q5.8. List all the genes that have been affected by coding mutations and have high impact. What do you find that is interesting?

Hint: You can use SNPeff HIGH/MODERATE impact field to help you finding those genes.

Q5.9. (bonus): Using Strelka on our data, we can detect indels, but only to a limited extent. Most of the reads in our BAM file have read lengths around 60bp long. Why might this have consequences for the detection of insertions that are longer than 60bp?

Q5.10. Variant Allele Frequency (VAF) is an important metric that helps us to measure how many DNA molecules in a given sample are carrying a given variant. It also helps to identify potential false-positive situations caused by incorrect base calls or alignment. VAF is calculated by

The number of variant alleles /(The number of Variant alleles + The number of Reference alleles)

In the form of a boxplot, plot the distribution of the VAFs across all the variants. How many variants have VAF > 5%? How many of those variants (the ones with >5% VAF) are in coding regions?

Hint: You will need to parse the genotype encoding field (GT:GQ:GQX:DP:DPF...) to get allele counts and then get VAF. To understand that column, look at the VCF Variant Call Format Document (GATK) section 5.

APPENDIX

SAM/BAM Format Specification Document: https://samtools.github.io/hts-specs/SAMv1.pdf

VCF Variant Call Format Document (GATK): https://gatk.broadinstitute.org/hc/en-us/articles/360035531692-VCF-Variant-Call-Format

snpEff Annotations Document: http://snpeff.sourceforge.net/VCFannotationformat_v1.0.pdf