Khan Inan
NGS 7653
Assignment week 6

Q1.1a Report the coverage depth for sample CR2342. The genome size for Chlamydomonas is approximately 120 Mb. Explain how you arrived at this answer [1 points]

The coverage depth that I arrived to for CR2342 was about 82.691X. I used the equations and method shown in the lecture

My work is shown here below:

```
# Coverage = LN/G where C=coverage, L= read length, N = number of reads and G= haploid size of genome
# L=51
# N=194567996
# G=120,000,000 (or 120 Mb)
# C = (51*194567996)/120,000,000 which is equal to ~82.691x coverage
```

Q1.1b. When reporting coverage depth genome wide for a BAM, what are two reasons why it might it not be accurate to simply count the number of reads in the alignment (e.g., samtools view -c), multiply by the read length from the sequencing (e.g., 100 in a 2 X 100 PE run), and divide by the genome size? [1 point]

There are a few reasons that it might not be accurate and one is that the depth is reported based the length of the reads and the total #, and it does not map the reads according to the genome. This is not accurate because unmapped reads are included in this calculation, so the formula is purely for hypothetical calculations. The second reason is that the length of the reads can vary in length based on hard or soft clipping or trimming because of various adapters.

Q1.1c Explain what MQ0 (=mapping quality of zero) represents in the stats output for reads mapped with BWA. Then answer the following multiple-choice question. Which of the following situations would you expect to find MQ0 reads mapped to gene A (or A') in the reference? Choose the single best answer. (a) gene A is duplicated in the reference to form identical copies gene A and A', but is single copy gene A in the sequenced sample (b) gene A is single copy in the reference but duplicated to form identical copies gene A and gene A' in the sequenced sample (c) both situations should produce MQ0 reads mapped to the reference [1 point].

MQ0 with BWA mapped reads essentially means that they are mapped to multiple locations on the reference genome. I would choose answer choice C because the reads in MQ0 would have mapped GENE A -> A' in the reference genome

Q1.1d. The lines in the samtools stat output beginning with “IS” contain the insert size and the corresponding number of pairs falling into each insert size category. Use these data to devise a crude method to predict deletions using this empirical insert size distribution [2 points].

First I would need to extract the size of the inserts, and the # of pairs in each category of the inserts and I would extract this information from the output file in samtools. After obtained this info, I would need to use some kind of data visualization like a graph to plot the size of the inserts against the # of pairs. The obtained graph would display the # of reads for each size.
The tail that should be enriched for deletions in the sample would be the right end of the tail. The majority of the data seems to be less than 400 while being greater than 300. A safe threshold to choose in regards to this spread of data would be 450 since there are few outliers greater than this value, and anything that is greater can be considered a deletion for the sample genome

Q2.1a What is the read depth at position 10,001 on chromosome_1 for sample CR2342? Please show your command with your answer.

The depth at this position seems to be 241

```
samtools depth -r chromosome_1:10001-10001 /scratch/work/courses/BI7653/hw6.2023/CR2342.bam
# This identifies read depth for only position 10,001 on chromosome_1 for sample CR2342.
```

Q2.2a. Paste the contents of your job submission script into your assignment document [1 point].

```
#!/bin/bash
#
#SBATCH --nodes=1
#SBATCH --tasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --time=8:00:00
#SBATCH --mem=8GB
#SBATCH --job-name=bedcov_coverage
#SBATCH --mail-type=FAIL,END
#SBATCH --mail-user=ki2100@nyu.edu

module purge
module load samtools/intel/1.14

samtools bedcov /scratch/work/courses/BI7653/hw6.2023/chromosome_1.500bp_intervals.bed /scratch/work/courses/BI7653/hw6.202
3/CR2342.bam /scratch/work/courses/BI7653/hw6.2023/CR407.bam
```

Q2.2b. What is the coverage of the last 10 intervals CR407 and CR2342 in the output file: (tail -n 10) in your answers file [1 point].

```
chromosome_1    8029000 8029500 13544   21495
chromosome_1    8029500 8030000 3065    17839
chromosome_1    8030000 8030500 2383    19434
chromosome_1    8030500 8031000 7392    28084
chromosome_1    8031000 8031500 11040   38217
chromosome_1    8031500 8032000 9891    26758
chromosome_1    8032000 8032500 10937   26281
chromosome_1    8032500 8033000 19142   31353
chromosome_1    8033000 8033500 17844   28520
chromosome_1    8033500 8033585 1323    2457
```

Q2.3 Include your plot in your MarkDown report or use the example code to create a pdf (which you must submit with your answer) [1 point]

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(dplyr)
library(tidyverse)
bedcov.tbl_df <- read_tsv("slurm-15920505.out",col_names=F)
```

```
## Rows: 16068 Columns: 5
```

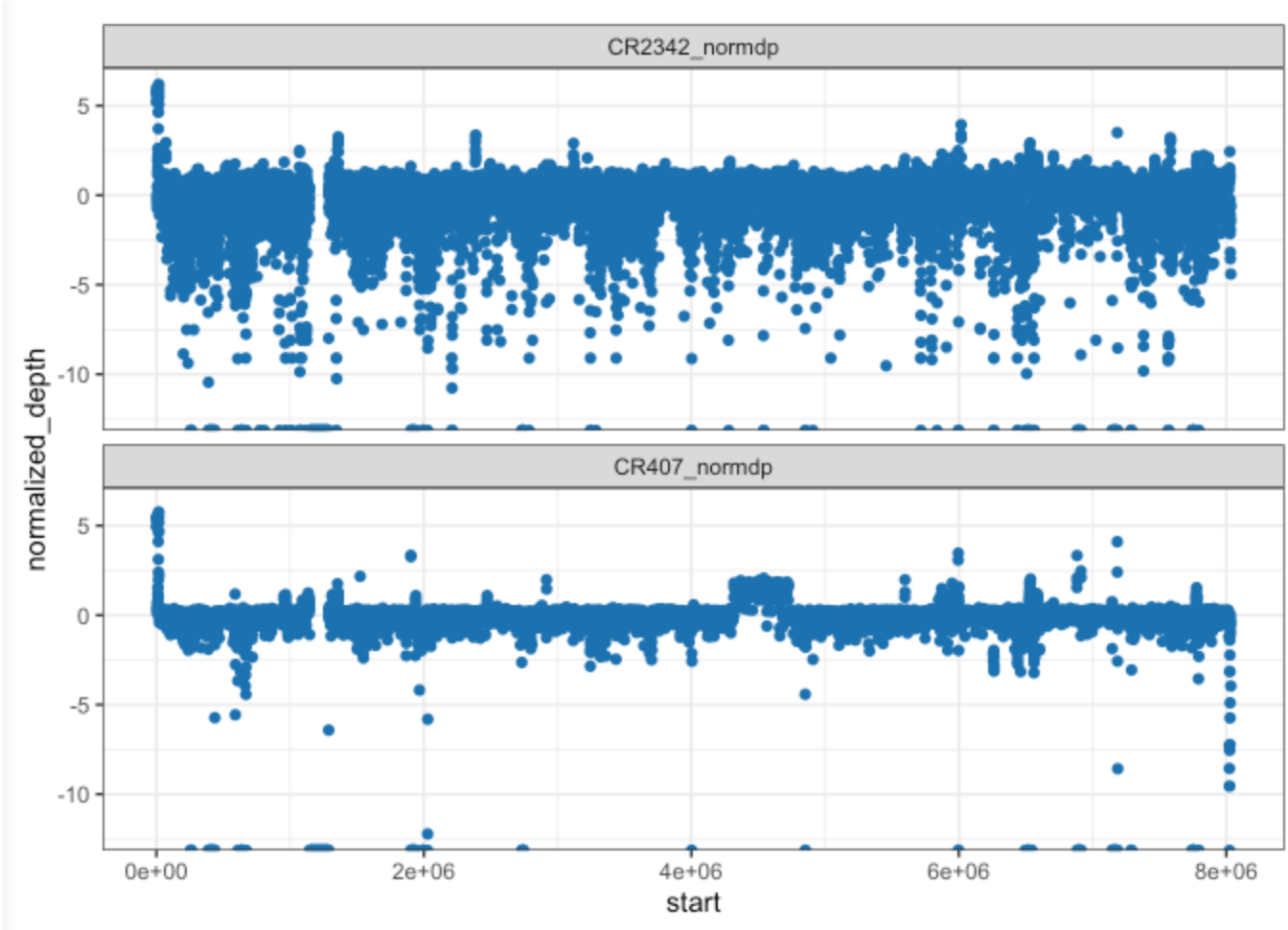
```
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: "\t"
## chr (1): X1
## dbl (4): X2, X3, X4, X5
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
names(bedcov.tbl_df) <- c('chr','start','end','CR2342_dp','CR407_dp')

bedcov.tbl_df <- bedcov.tbl_df %>%
  mutate(CR2342_normdp = log2( CR2342_dp / median(CR2342_dp,na.rm=T)))
bedcov.tbl_df <- bedcov.tbl_df %>%
  mutate(CR407_normdp = log2( CR407_dp / median(CR407_dp,na.rm=T)))

bedcov_pivoted.tbl_df <- bedcov.tbl_df %>%
  select(-CR2342_dp,-CR407_dp) %>%
  pivot_longer(cols = c(-chr,-start,-end),
               names_to = 'sample',
               values_to = 'normalized_depth')

bedcov_pivoted.tbl_df %>%
  ggplot(aes(x = start,y = normalized_depth)) + geom_point(color="#0072B2") +
  facet_wrap(~ sample,nrow=2) + theme_bw()
```



Q2.4a Which sample has a large (~ 400 kb) duplication on chromosome 1? Approximately what position on the chromosome is the duplication?

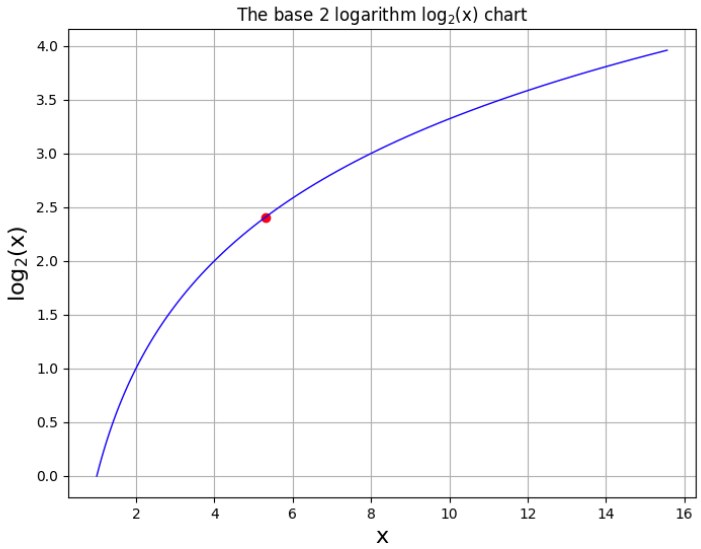
The sample that has a large duplication on chromosome 1 seems to be CR407 and at around the starting position of 4e+06 on the x-acis

Q2.4b What is the approximate log2 value in this duplicated region? Based on this log2 value, what do you think the copy number of this duplication might be given that Chlamydomonas are haploid?

The log2 value in the specified region is around 2.4-2.5. The copy number of this duplication give that Chlamydomonas are haploid should be about 5.3-5.6 based on the log2 curves shown on the next page

$\log_2 5.3$ =

2.4059923596758366



$\log_2 5.6$ =

2.485426827170242