Homework 6 - Xiaoyan Wen

XiaoyanWen

3/2/2022

Next Generation Sequence Analysis Homework Week 6

Many types of NGS analyses use sequencing coverage and the insert size distribution from BAM alignments of paired-end data to predict structural variants. There are hundreds of software applications designed for this purpose. In this assignment, you will learn to extract relevant information from BAM files and conduct a coverage depth exercise. You will use the outputs to devise a crude method to detect insertion/deletion mutations and a large copy number variant.

About the data

This week you will work with short read data from the haploid alga Chlamydomonas reinhardtii that were aligned to the Chlamydomonas reference genome using BWA aln.

These data were published in Flowers JM, Hazzouri KM, Pham GM, et al. Whole-Genome Resequencing Reveals Extensive Natural Variation in the Model Green Alga Chlamydomonas reinhardtii. Plant Cell. 2015;27(9):2353-2369. doi:10.1105/tpc.15.00492

Task 1: Summarize coverage and insert size distribution from BAM alignments

Summarizing genomewide coverage information in BAM alignment(s) is a fundamental task. There are many packages that provide functions for this purpose including **Samtools, Bamtools, Picard-tools, and Deeptools** to name a few.

Here you will use the **Samtools stats** program to determine coverage (average coverage genomewide) for paired-end (2 x 51 PE) reads from a Chlamydomonas strain (CC-2342):

```
`/scratch/work/courses/BI7653/hw6.2022/CR2342.bam`
```

The samtools documentation is here: http://www.htslib.org/doc/samtools-1.6.html (http://www.htslib.org/doc/samtools-1.6.html)

Begin by logging in to Greene and creating a directory for Week 6, Task 1.

```
srun --time=4:00:00 --mem=4GB --pty /bin/bash

# Create directories for ngs.week6 and task1 in your /scratch
cd $SCRATCH
mkdir ngs.week6
cd ngs.week6
mkdir task1
cd task1
```

Now create a slurm job submission script with a command line to run `samtools stats` on the CR2342 BAM. Capture the STDOUT using Unix redirection (“>”).

```
#!/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=slurm_wk6_task1
#SBATCH --mail-type=FAIL
#SBATCH --mail-user=xw2470@nyu.edu

#SBATCH -o slurm.%N.%j.out # STDOUT
#SBATCH -e slurm.%N.%j.err # STDERR

module purge


echo script begin: $(date)

module load samtools/intel/1.14
samtools stats /scratch/work/courses/BI7653/hw6.2022/CR2342.bam > CR2342_coverage.text

echo script completed: $(date)
```

When the job is complete, use the output to answer the following:

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

$$1 points$$

.

```
# Answer:
# coverage depth = total length  / genomeSize
depth = 9922967796 / 120000000
depth
```

```
## [1] 82.6914
```

Q1.1b. When reporting coverage depth genomewide 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?

$$1point$$

```
# Answer: because read depth is not evenly distributed for all positions. Read depth differences among positions depending on multiple influence factors including GC content and PCR preferences & error.
```

Q1.1c Explain what MQ0 (=mapping quality of zero) represents in the stats output for reads mapped with BWA.

```
# Answer: Zero mapping quality indicates that the read maps to multiple locations.
```

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 identifical copies gene A and gene A' in the sequenced sample

c. both situations should produce MQ0 reads mapped to the reference

$$1point$$

.

```
# Answer:c
```

Q1.1d. An important property in some structural variant analyses is the insert size distribution in paired-end data. For example, so-called "read pair" structural variant detection methods use the insert size distribution to identify regions with deletions in the sample genome relative to the reference.

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

$$2points$$

.

Include in your answer:

- A description of the rational for your test.
- Which tail of the read length distribution should be enriched for deletions in the sample and why
- Define a reasonable threshold insert size below or above which are likely to be enriched for deletions in the sample genome

```
# Answer:to use the insert size distribution to identify insertion/deletion, we first plot a count histogram for <distance between mapped read pairs>.Clusters of multiple read pairs with unusual distances between them represent read pairs that span probable insertions or deletions.

# The the upper tail of the distribution (paired end read with larger distance) are enriched with deletions in the sample.

# a reasonable threshold could be set as: mapped read pair distance > {mean + 2*standard_deviation}.
# Given:
        ##SN      insert size average:    435.0
        ##SN      insert size standard deviation: 820.3

# samples with mapped read pair distance > 2076 are likely to be enriched for deletions.
```

Task 2: Coverage depth in genomic regions and copy number variant discovery

Samtools has a program called "depth" which calculates the number of reads covering each position in a SAM or BAM alignment.

If you are not already at a compute node, check out one for interactive use now.

```
    srun --time=4:00:00 --mem=4GB --pty /bin/bash
```

$ Now create a directory for Week 6, Task 2 in your `/scratch`

$ Load the module samtools on Greene and read the documentation by entering `samtools depth` at the command line.

$ Calculate depth per base position using a single samtools depth command for the region chromosome_1:10001-10020 from the sample alignment for Chlamydomonas strain CR2342 in Task 1. Note the following:

- You can limit the depth calculation to a region, or interval, with -r.
- You must use "chromosome_1" as this is the syntax for chromosome 1 in the reference genome to which the reads were aligned
- The necessary command can be run interactively (no need for a slurm job)

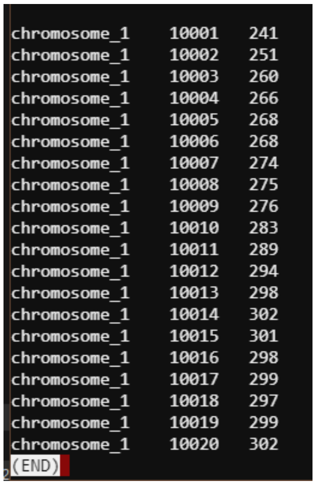
Q2.1 Answer the following

$$1point$$

:

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

```
# Answer: read depth at position 10,001 is 241.
# command
samtools depth -a -r chromosome_1:10001-10020 /scratch/work/courses/BI7653/hw6.2022/CR2342.bam | less
```



```
chromosome_1    10001   241
chromosome_1    10002   251
chromosome_1    10003   260
chromosome_1    10004   266
chromosome_1    10005   268
chromosome_1    10006   268
chromosome_1    10007   274
chromosome_1    10008   275
chromosome_1    10009   276
chromosome_1    10010   283
chromosome_1    10011   289
chromosome_1    10012   294
chromosome_1    10013   298
chromosome_1    10014   302
chromosome_1    10015   301
chromosome_1    10016   298
chromosome_1    10017   299
chromosome_1    10018   297
chromosome_1    10019   299
chromosome_1    10020   302
(END)
```

Q2.1b What is the coverage in the interval chromosome_1:10001-10020 for CR2342?

```
# Answer:
coverage = sum(241,251,260,266,268,268,274,275,276,283,289,294,298,302,301,298,299,297,299,302)/20
coverage
```

```
## [1] 282.05
```

Now you will calculate coverage depth in genomic intervals to identify a copy number variant by running `samtools bedcov` from a slurm job submission script. You will use `samtools bedcov` to generate a covereage depth value for all non-overlapping 500 bp intervals on chromosome_1 for the both samples CR407 and CR2342.

The bam and their index files are here:

```
    /scratch/work/courses/BI7653/hw6.2022/CR2342.bam
    /scratch/work/courses/BI7653/hw6.2022/CR407.bam
```

The intervals we wish to calculate coverage are in BED format here:

```
    /scratch/work/courses/BI7653/hw6.2022/chromosome_1.500bp_intervals.bed
```

BED is a standard format for storing genomic intervals NGS analysis. Typically BED is tab-delimited plain text with three columns chromosome,start and end position. The start position in BED is “zero-based” (the genomic interval start position minus one) while the end position of each interval is “one-based” (the genomic interval end position).

```
    # review the file showing hidden characters (^I = tab, $ = \n end of line character)
    cat -et /scratch/work/courses/BI7653/hw6.2022/chromosome_1.500bp_intervals.bed | less # q to exit
```

Now review the documentation for samtools bedcov and construct a single command line to calculate coverage depth for both samples in the chromosome_1 intervals. Notice that you can provide two or more BAM alignments to the bedcov command to produce a single output file with coverage values for samples (columns) for each interval in the input BED file (rows).

Now submit a slurm job submission script with your bedcov command line. The column order of samples is the same order of the input BAMs in your bedcov command line.

When you are ready, submit the job using sbatch.

Q2.2a. Paste the contents of your job submission script into your assignment document

$$1point$$

.

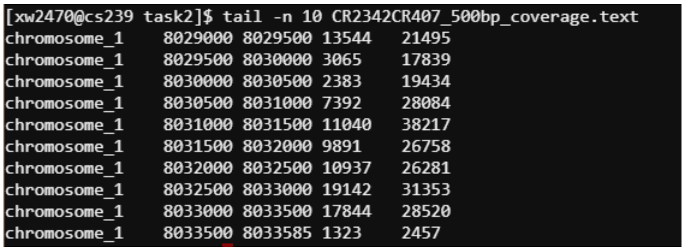
```
#!/bin/bash
#
#SBATCH --nodes=1
#SBATCH --tasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --time=8:00:00
#SBATCH --mem=8GB
#SBATCH --job-name=slurm_wk6_task2
#SBATCH --mail-type=FAIL
#SBATCH --mail-user=xw2470@nyu.edu

#SBATCH -o slurm.%N.%j.out # STDOUT
#SBATCH -e slurm.%N.%j.err # STDERR

module purge


echo script begin: $(date)

module load samtools/intel/1.14

samtools bedcov /scratch/work/courses/BI7653/hw6.2022/chromosome_1.500bp_intervals.bed \
/scratch/work/courses/BI7653/hw6.2022/CR2342.bam \
/scratch/work/courses/BI7653/hw6.2022/CR407.bam \
> CR2342CR407_500bp_coverage.text


echo script completed: $(date)
```

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

$$1point$$

.

```
tail -n 10 CR2342CR407_500bp_coverage.text
```



```
[xw2470@cs239 task2]$ tail -n 10 CR2342CR407_500bp_coverage.text
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
```

$ Now we will plot a normalized coverage depth value from all chromosome_1 intervals in the two algae strains and try to visualize a copy number variant. If you wish, you may want to transfer the bedcov output to your personal computer to normalize the data and plot in R.

- We will normalize the coverage in each interval using a simple transformation so that intervals with a typical depth value will be centered on zero.
- log2(interval depth / median(depth))
- We will do this transformation separately for each sample.

Please review the following code which you may use to generate a plot of normalized coverage depth in 500 bp intervals on chromosome 1 for each sample. You will need to write your own command to create a column of normalized depth for sample CR2342, using the example given for CR407.

You need to modify the input filename argument to the read.table function and make sure the "names" function correctly specifies the order of the columns in your bedcov output file. That is, you may need to reverse the order of the CR407 and CR2342 column names depending on the order you specified these samples in your bedcov command line.

Feel free to customize the appearance of the figure using the ggplot function if you would like such as by changing the size of the points by adding a "size" argument to the geom_point function, or perhaps better, by altering the dimensions of the pdf to increase or decrease the plot proportions and improve readability.

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)

$$1point$$

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

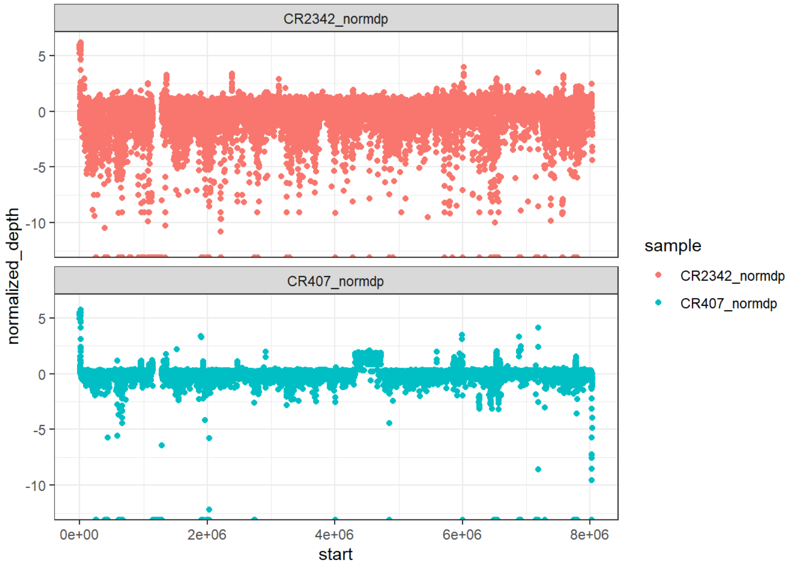
```
# read data
bedcov.tbl_df <- read_tsv("CR2342CR407_500bp_coverage.text",col_names=F)
```

```
## 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.
```

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

# Add columns to data frame with normalized depth values for each strain
bedcov.tbl_df <- bedcov.tbl_df %>%
  mutate(CR2342_normdp = log2( CR2342_dp / median(CR2342_dp,na.rm=T)),
         CR407_normdp = log2( CR407_dp / median(CR407_dp,na.rm=T)))

# transform data frame into long format for ggplot
bedcov_pivoted.tbl_df <- bedcov.tbl_df %>%
  select(-CR407_dp,-CR2342_dp) %>%
  pivot_longer(cols = c(-chr,-start,-end),
               names_to = 'sample',
               values_to = 'normalized_depth')

# pdf('normalized_depth.pdf',width=8,height=6)
bedcov_pivoted.tbl_df %>%
  ggplot(aes(x = start,y = normalized_depth, col=sample)) +
  geom_point() +
  facet_wrap(~ sample,nrow=2) +
  theme_bw()
```



```
#dev.off()
```

Q2.4 Now answer the following

$$1point$$

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

```
# Answer: CR407 has a large (\~ 400 kb) duplication on chromosome 1, approximately position 4.3e06 ~ 4.8e06.
```

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?

```
# Answer: the approximate log2 value is around 2;
# the copy number is linearly correlated with interval coverage and might be around 4 in this case given that Chlamydomonas
 are haploid.
```