Epigenetics Practical

Task 2

Guangyi Chen (2581241)

March 18, 2020

1

To create symbolic links to all comma-separated value files (.csv) in a folder named files, do such:

```
ln -s ../files/*.csv
```

where . is the target directory

2

data frame is a widely-used data storing format which serves as data tables. It contains data rows and columns, whose names represent different variables. The data in each member of the row and column is called cell. Typically in Biological data, data columns indicate samples and rows suggest predictors like genes.

Matrix is an other important 2-dimensional data format in R. They look similar however:
1). In *matrix*, all the data elements should be of the same data types, but in *data frame*, each column can contain variable types of data;
2). *Data frame* has row names and column names, nevertheless *matrix* doesn't.
3). *Data frame* can be regarded as the more generalized form of *matrix*.

3

To specify codes not being excuted in R-Markdown, we could use back-ticks around the according codes. For example:

```
‘‘‘
python
target codes
print s
‘‘‘
```

4

An example of annotation sheet (sample metadata sheet) is:

sample1	test1	data/test1.fastq.gz	fastq	FqRd1
sample1	test2	data/test2.fastq.gz	fastq	FqRd2
sample1	test3	data/test3.fastq.gz	fastq	FqRd3

where each of the column represents **sampleID**, **runID**, **path**, **type**, and **view** in order.

5

The outcome results generated from grape-nf are in pipeline.db file, which is TSV formatted.
The accepted formats of inout data of IGV are: BAM, BED, BEDPE, BedGraph, bigBed, bigWig, Birdsuite Files, broadPeak, CBS Chemical Reactivity Probing Profiles, chrom.sizes, CN, Custom, File Formats, Cytoband, FASTA, GCT, CRAM, genePred, GFF/GTF, GISTIC, Goby, GWAS, IGV, LOH, MAF (Multiple Alignment Format), MAF (Mutation Annotation Format), Merged BAM File, MUT, narrowPeak, PSL, RES, RNA Secondary Structure Formats, SAM, Sample Info (Attributes) file, SEG, TDF, Track Line, Type Line, VCF, WIG.
Among which the BAM and FASTA can be generated by grape-nf (at least).

6

To create an R data object for storing data (DGEList as in the example), first to obtain the scaling factors for length per observation and to adjust accordingly in order to avoid changing the magnitude of the counts. Secondly, we compute the effective library sizes from scaled counts and account for the composition biases between samples. Then combining the effective library sizes with length factors and calculating the offsets for a log-link GLM. After than applying *DGEList()* function to create an object. At the end we could do data filtering on this object.