RNA-seq Analysis Practice

Huang Ao 17338053

This practice is a simple recast of a group of data obtained from `https://www.jianshu.com/p/0ab0e2aeca14`, as to familiarize the rudimentary pipelines for RNA-seq data processing.

1 Reference Genome Preparation

1.1 Download Reference Genome

1. Download the whole genome sequences from UCSC.

```
# byronadams @ ByrondeMacBook-Pro in ~/HuangAo_dir/genome
$ wget http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.gz

# byronadams @ ByrondeMacBook-Pro in ~/HuangAo_dir/genome
$ gzip -d hg19.fa.gz

# byronadams @ ByrondeMacBook-Pro in ~/HuangAo_dir/genome
$ cp ./hg19.fa ~/HuangAo_dir/RNAseq_test_homo/bt2index_result
```

2. Establish indexes for the genome using bowtie2.

```
# byronadams @ ByrondeMacBook-Pro in ~/HuangAo_dir/RNAseq_test_homo/bt2index_result
$ bowtie2-build ./hg19.fa ./hg19_bt2_index

# byronadams @ ByrondeMacBook-Pro in ~/HuangAo_dir/RNAseq_test_homo/bt2index_result
$ ls
hg19.fa                  hg19_bt2_index.3.bt2           hg19_bt2_index.rev.2.bt2
hg19_bt2_index.1.bt2     hg19_bt2_index.4.bt2           hg19_bt2_index.2.bt2
hg19_bt2_index.rev.1.bt2
```

2 RNA-seq Files Processing

1. Obtain the SRA files.

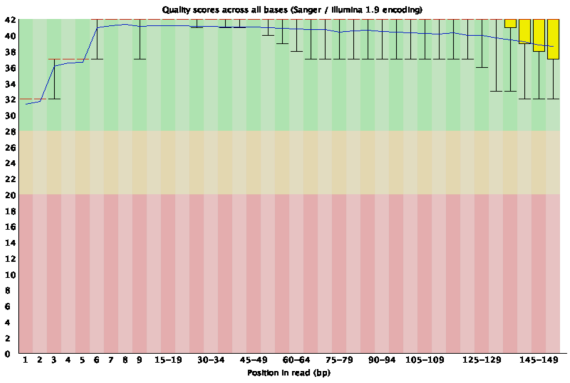
The raw data of RNA-seq were obtained from `https://www.jianshu.com/p/0ab0e2aeca14`, which had already been transformed to .fa.gz files. We can see that the original files were organized from paired-end sequencing results and can be decompressed with sra-tools. Both treatment group and the control group had two repeats.

```
# byronadams @ ByrondeMacBook-Pro in ~/HuangAo_dir/RNAseq_test_homo/raw_data
$ ls
hela_ctrl_rep1_R1.fq.gz  hela_ctrl_rep2_R2.fq.gz  hela_treat_rep2_R1.fq.gz
hela_ctrl_rep1_R2.fq.gz  hela_treat_rep1_R1.fq.gz hela_treat_rep2_R2.fq.gz
hela_ctrl_rep2_R1.fq.gz  hela_treat_rep1_R2.fq.gz
```

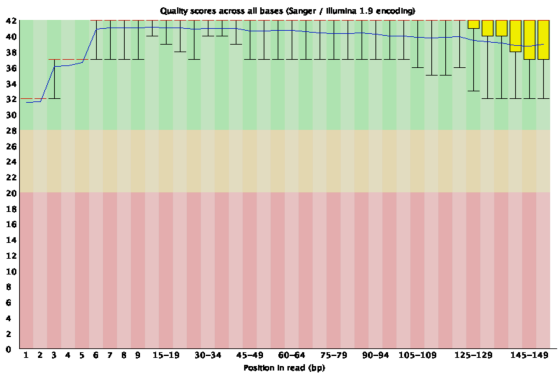
2. Generate FastQC report for sequence quality.

```
# byronadams @ ByrondeMacBook-Pro in ~/HuangAo_dir/RNAseq_test_homo
$ mkdir fastqc_analysis

# byronadams @ ByrondeMacBook-Pro in ~/HuangAo_dir/RNAseq_test_mucus
$ fastqc -o ./fastqc_analysis ./*.gz

# byronadams @ ByrondeMacBook-Pro in ~/HuangAo_dir/RNAseq_test_mucus
$ cd fastqc_analysis

# byronadams @ ByrondeMacBook-Pro in ~/HuangAo_dir/RNAseq_test_homo/fastqc_analysis
$ ls
hela_ctrl_rep1_R1_fastqc.html    hela_ctrl_rep1_R2_fastqc.zip
hela_ctrl_rep2_R2_fastqc.html    hela_treat_rep1_R1_fastqc.zip
hela_treat_rep2_R1_fastqc.html   hela_treat_rep2_R2_fastqc.zip
hela_ctrl_rep1_R1_fastqc.zip     hela_ctrl_rep2_R1_fastqc.html
hela_ctrl_rep2_R2_fastqc.zip     hela_treat_rep1_R2_fastqc.html
hela_treat_rep2_R1_fastqc.zip    hela_ctrl_rep1_R2_fastqc.html
hela_ctrl_rep2_R1_fastqc.zip     hela_treat_rep1_R1_fastqc.html
hela_treat_rep1_R2_fastqc.zip    hela_treat_rep2_R2_fastqc.html
```

Open the .html files, we can see some examples of fastqc reports as in Figure 1. The sequencing quality of the presented samples were relatively good.



ctrl-1 rep-1 forward



test-1 rep-1 forward

Figure 1: Sequence Quality Per Base

3. Cut adaptors from reads.

```
# byronadams @ ByrondeMacBook-Pro in ~/HuangAo_dir/RNAseq_test_homo/fastqc_analysis
$ cd ..
```

Establish a shell script, cutadapt.sh to efficiently cut adaptors from reads.

```
mkdir -p ./cutadapt_result

for case_name in hela_ctrl_rep1 hela_ctrl_rep2 hela_treat_rep1 hela_treat_rep2

do
        fq_1=./raw_data/${case_name}_R1.fq.gz
        fq_2=./raw_data/${case_name}_R2.fq.gz

        out_1=./cutadapt_result/${case_name}_R1_cut.fq.gz
        out_2=./cutadapt_result/${case_name}_R2_cut.fq.gz

        nohup cutadapt --times 1 -e 0.1 -O 3 --quality-cutoff 6 -m 75 -a AGATCGGAAGAGC -A
            AGATCGGAAGAGC -o $out_1 -p $out_2 $fq_1 $fq_2 > cutadapt.log

done
```

4. Map reads back to the reference genome.

Establish another shell script, tophat2_map.sh to map reads back to the reference genome.

```
mkdir -p tophat_result

index=/Users/byronadams/HuangAo_dir/RNAseq_test_homo/bt2index_result/hg19_bt2_index

for case_name in hela_treat_rep1

do
        in_1=./cutadapt_result/${case_name}_R1_cut.fq.gz
        in_2=./cutadapt_result/${case_name}_R2_cut.fq.gz

        out_tophat=./tophat_result/${case_name}_result

        nohup tophat2 -p 32 -G ./RefSeq_genes_hg19.gtf -o $out_tophat $index $in_1 $in_2 >
            tophat_2.log

done
```

5. Calculate the expression difference using cuffdiff.

Of note, the .gtf annotation file were downloaded from UCSC.

```
mkdir -p ./cuffdiff_result

ctrl_group=./tophat_result/hela_ctrl_rep1_result/accepted_hits.bam,./tophat_result/
    hela_ctrl_rep2_result/accepted_hits.bam
test_group=./tophat_result/hela_treat_rep1_result/accepted_hits.bam,./tophat_result/
    hela_treat_rep2_result/accepted_hits.bam

```

```
nohup cuffdiff -o ./cuffdiff_result -p 32 --min-reps-for-js-test 2 ./RefSeq_genes_hg19.gtf
    $ctrl_group $test_group > cuffdiff.log
```

3 Visualization

The cuffdiff result, stored in gene_exp.diff, was visualized in volcano plot with R, cf. Figure 2.

```
### inputing data
gene_diff = read.table("/Users/byronadams/HuangAo_dir/RNAseq_test_homo/cuffdiff_result/
gene_exp.diff",header = T)

test_fpkm = gene_diff$value_2
ctrl_fpkm = gene_diff$value_1

lg2_FC = log2(test_fpkm / ctrl_fpkm)
lg10_pvalue = - log10(gene_diff$p_value)

### eliminating unqualified data
lg10_pvalue_mod = lg10_pvalue[lg10_pvalue > 0]  # eliminate "p-value = 1" data

lg2_FC[ctrl_fpkm == 0|test_fpkm == 0 ] = 0        # eliminate -inf inf data
lg2_FC_mod = lg2_FC[lg10_pvalue > 0]            # filter x to obtain equal length with y

### selecting data of significance
color_vec = rep(rgb(0,0,0,0.2),length(lg2_FC_mod))

sig_filter =  (test_fpkm > 0) & (ctrl_fpkm > 0) & (test_fpkm>=1 | ctrl_fpkm>=1) &
        (lg2_FC >= 1) & (gene_diff$p_value < 0.05)
sig_filter2 = (test_fpkm > 0) & (ctrl_fpkm > 0) & (test_fpkm>=1 | ctrl_fpkm>=1) &
        (lg2_FC <= -1) & (gene_diff$p_value < 0.05)
sig_filter = sig_filter[lg10_pvalue > 0]          # obtain equal length with x and y
sig_filter2 = sig_filter2[lg10_pvalue > 0]

color_vec[sig_filter] = rgb(1,0,0)
color_vec[sig_filter2] = rgb(0,1,0)

###ploting volcano map
plot(x = lg2_FC_mod, y = lg10_pvalue_mod,
xlim = c(-5,5), ylim = c(0,4),
col = color_vec, pch = 16, cex = 0.7,
 xlab = "log_2 FoldChange", ylab = "log_10 p-value"
 )

abline(h = - log10(0.05), lty = 3, col = "black")
abline(v = log2(2), lty = 3, col = "black")
abline(v = - log2(2), lty = 3, col = "black")
```

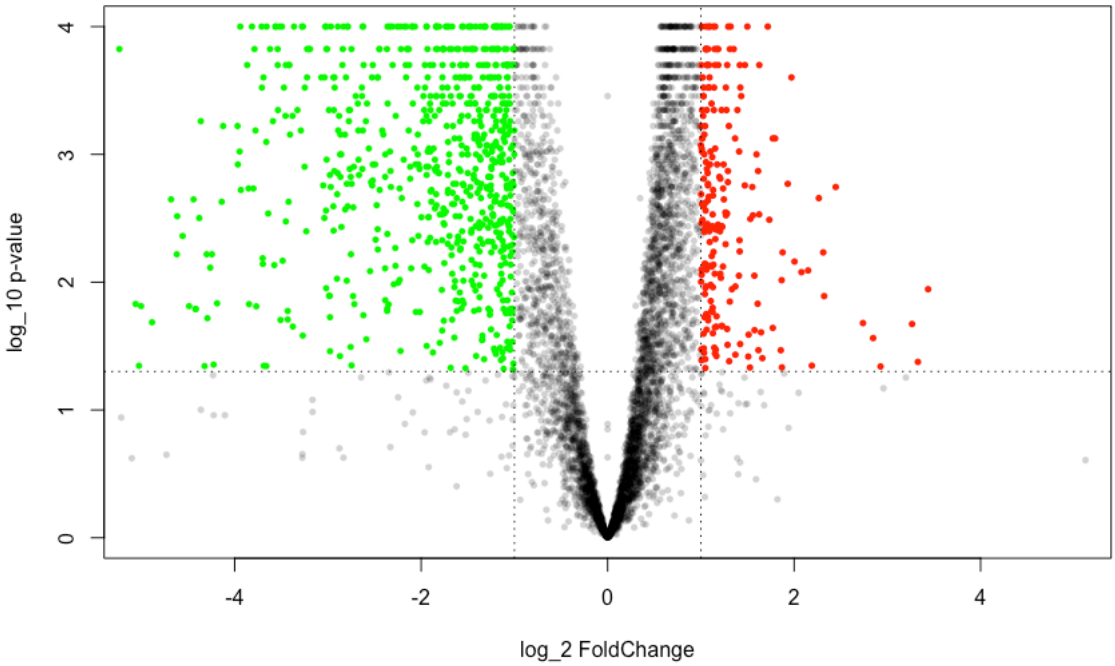


图 1: Volcano Plot Visualization of Expression Difference

4 Reference

Data used in this study were obtained from `https://www.jianshu.com/p/0ab0e2aeca14`. The working pipelines and methods in the process of RNA-seq data analysis were mostly invoked from a *Zhihu Live* by *Meng Haowei*, cf. `https://www.zhihu.com/lives/865204175334670336`. All methods introduced were meticulously studied and some were put into practice.