Gulishana Adilijiang (adiliji2)

STAT 530 Bioinformatics: Homework 6

Problem1:

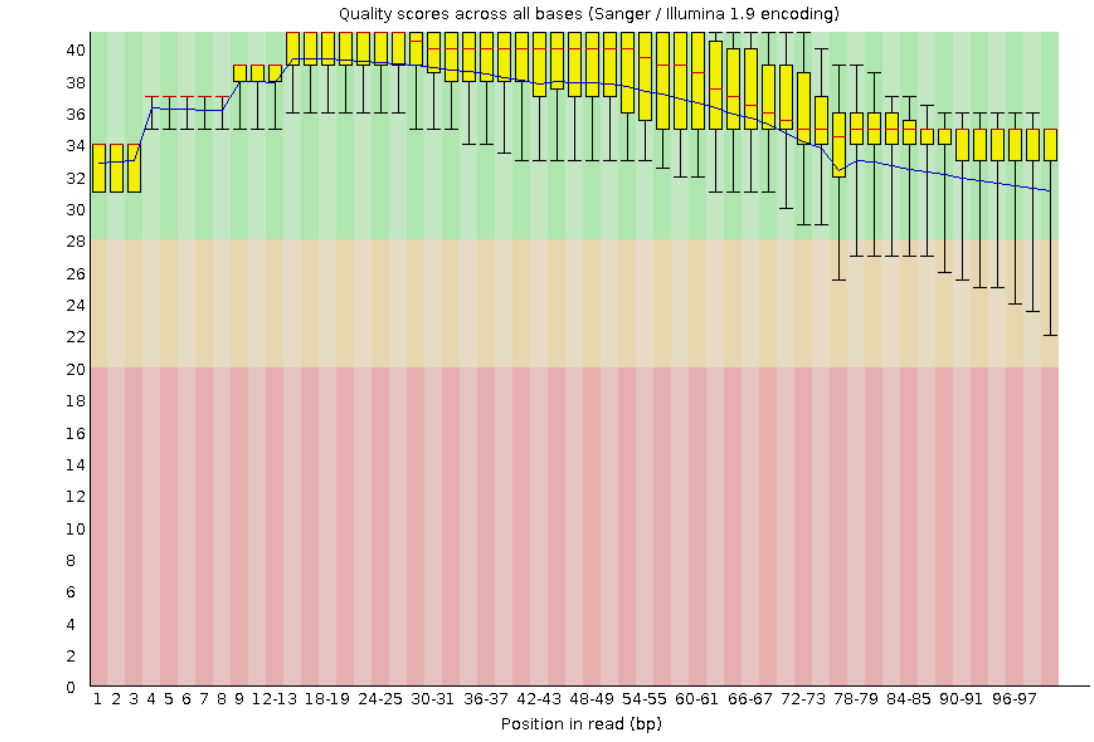
Check the RNA-seq data qualities of the two .fastq.gz files that were converted by SRA from the .sra file in Homework 5 problem 4, which are from a single honeybee.

$ fastqc HW5_SRR4017758_1.fastq.gz

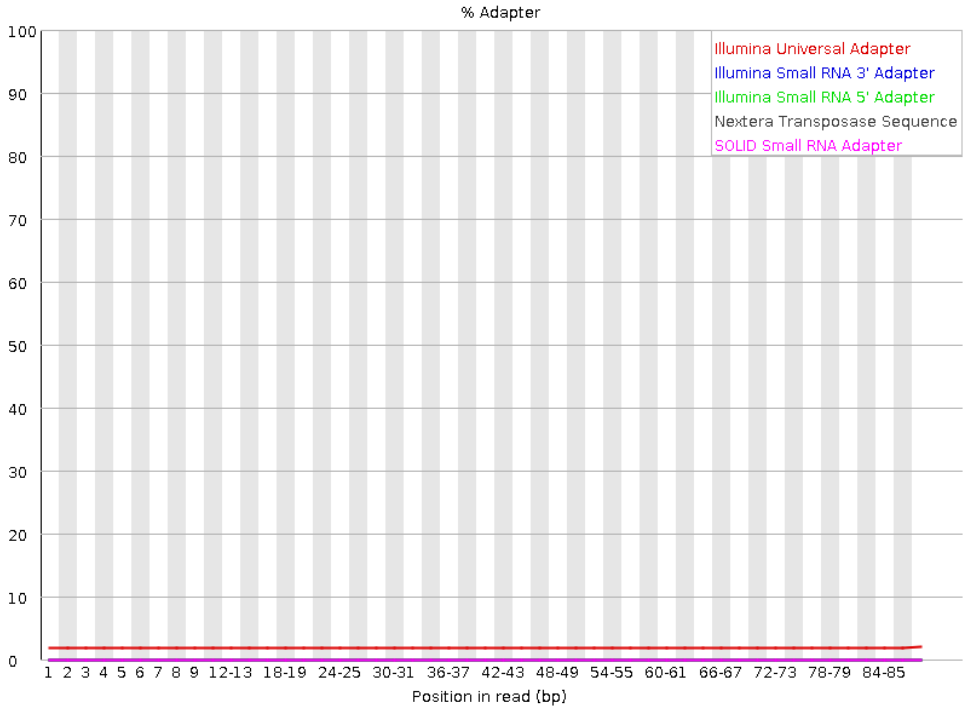
Basic Statistics

Measure	Value
Filename	HW5_SRR4017758_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	28821216
Sequences flagged as poor quality	0
Sequence length	100
%GC	41

Per base sequence quality



Adapter Content



Overrepresented sequences

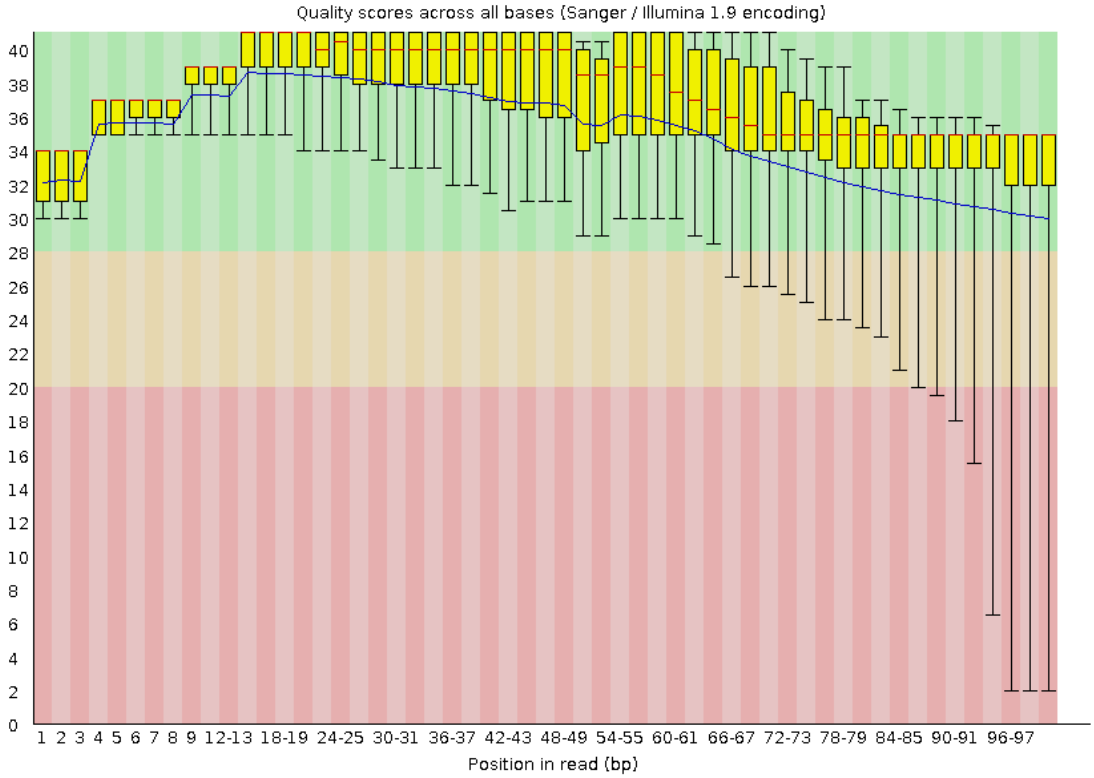
Sequence	Count	Percentage	Possible Source
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATG	532838	1.8487700171984416	TruSeq Adapter, Index 5 (100% over 49bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGC	80591	0.2796238715257538	TruSeq Adapter, Index 5 (100% over 50bp)
GCTAATTCCAGATGTTTCCTATTAAACGGTGAACAAGGCATGCCATGCTT	35506	0.12319396933148136	No Hit

$ fastqc HW5_SRR4017758_2.fastq.gz

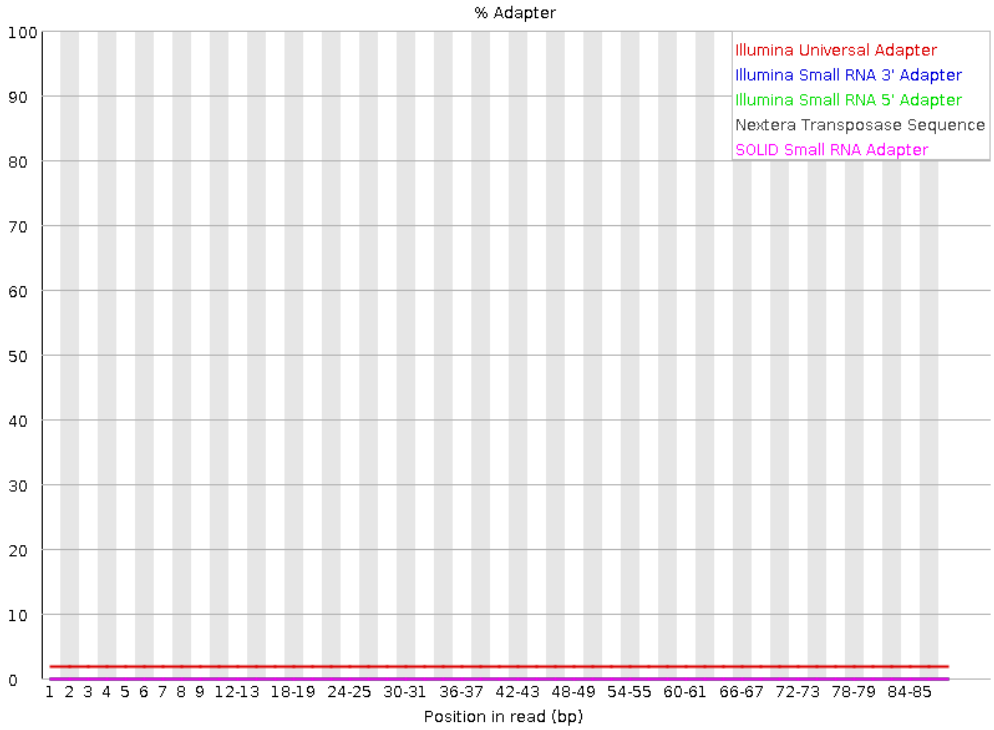
Basic Statistics

Measure	Value
Filename	HW5_SRR4017758_2.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	28821216
Sequences flagged as poor quality	0
Sequence length	100
%GC	41

Per base sequence quality



Adapter Content



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCC	546741	1.8970087868603462	Illumina Single End PCR Primer 1 (100% over 50bp)

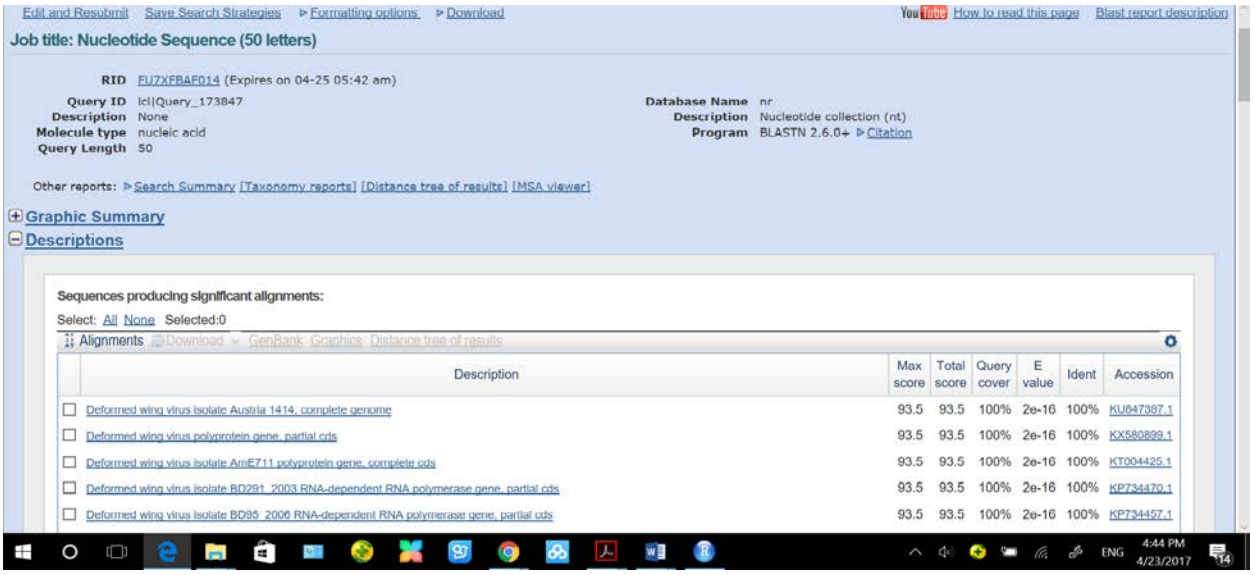
Quality Control Analysis:

In the HW5_SRR4017748_1.fastq.gz file, the overall sequence qualities are pretty good. However, the reads still have adapters not trimmed yet, which is the TruSeq Adapter Index 5.

In the HW5_SRR4017748_2.fastq.gz file, the overall sequence qualities are also fine. However, the reads still have adapters not trimmed yet as well, which is Illumina Single-End PCR Primer 1.

Hence these untrimmed adaptors cause the low alignment rate.

However, this is only the technical problem. There is another overexpressed sequence in the HW5_SRR4017748_1.fastq.gz file. Then I blast the sequence in the NCBI.



The result shows that the sequence comes from a deformed wing virus. This indicates that the sample was contaminated by a kind of virus, which also causes the low alignment rate.

Problem2:

Part (a)

R Notebook

```
## Part(a)
setwd("C:/Users/Gulishana/Desktop/HW6_dm_counts")
library(edgeR);

## Loading required package: limma

## Warning: package 'limma' was built under R version 3.3.3

targets=readTargets(file="dm_targets.txt");
raw_counts=readDGE(targets$file,comment.char="_",header=F);
dim(raw_counts$counts)

## [1] 14869    12

## filter out genes with low expression
keep=rowSums(cpm(raw_counts)>1)>=3;
counts=raw_counts[keep,keep.lib.sizes=F];

## recalculate the library size
names(counts);

## [1] "samples" "counts"

dim(counts$counts);

## [1] 9679   12

## calculate TMM normalization factors
counts=calcNormFactors(counts);

## fit a two-way ANOVA
design=model.matrix(~sex+mating+sex*mating,data=targets);
print(design)

##    (Intercept) sexmale matingpolygamous sexmale:matingpolygamous
## 1            1       0                0                        0
## 2            1       0                0                        0
## 3            1       0                0                        0
## 4            1       0                1                        0
## 5            1       0                1                        0
## 6            1       0                1                        0
## 7            1       1                0                        0
## 8            1       1                0                        0
## 9            1       1                0                        0
## 10           1       1                1                        1
## 11           1       1                1                        1
```

```
## 12           1       1                1                          1
## attr(,"assign")
## [1] 0 1 2 3
## attr(,"contrasts")
## attr(,"contrasts")$sex
## [1] "contr.treatment"
##
## attr(,"contrasts")$mating
## [1] "contr.treatment"
```

```
## use tagwise dispersion
counts=estimateGLMCommonDisp(counts,design);
counts=estimateGLMTrendedDisp(counts,design);
counts=estimateGLMTagwiseDisp(counts,design);

## fit the GLM regression
fit=glmFit(counts,design,dispersion=counts$tagwise.dispersion);

head(fit$coefficients);
```

```
##             (Intercept)     sexmale matingpolygamous
## FBgn0000008  -10.262992  0.19059518       0.19258156
## FBgn0000017   -8.620820  0.28048687       0.04814777
## FBgn0000018  -11.655784 -0.09425485      -0.14625286
## FBgn0000024   -9.668338  0.78003183       0.18595410
## FBgn0000028  -11.425994  0.10294410       0.38386437
## FBgn0000032  -10.940798  0.12306910       0.08148035
##             sexmale:matingpolygamous
## FBgn0000008              -0.25786451
## FBgn0000017              -0.16424834
## FBgn0000018               0.00849485
## FBgn0000024              -0.28405913
## FBgn0000028              -0.20539584
## FBgn0000032              -0.11189656
```

```
genes=c("FBgn0000018","FBgn0000042","FBgn0025712")
fit$coefficients[genes,]
```

```
##             (Intercept)     sexmale matingpolygamous
## FBgn0000018  -11.655784 -0.09425485      -0.14625286
## FBgn0000042   -6.497025 -0.01334732       0.07286698
## FBgn0025712   -9.763452  0.04473060       0.04472839
##             sexmale:matingpolygamous
## FBgn0000018              0.008494850
## FBgn0000042              0.008304273
## FBgn0025712             -0.045479982
```

In the above coefficients, “Intercept” is $\beta_0 hat$, “sexmale” is $\beta_1 hat$, “matingpolygamous” is $\beta_2 hat$, “sexmale:matingpolygamous” is $\beta_3 hat$.

Part (b)

In the above saturated two-way ANOVA regression model:

$$\log E(y_{gi} | sex_i, mating_i) = ß_{g0} + ß_{g1}\, sex_i + ß_{g2}\, mating_i + ß_{g3}\, sex_i * mating_i$$

We can get the counts of genes of males and females among the polygamous flies, respectively:

$$\log E(y_{gi} | sexmale=1, matingpolygamous=1) = ß_{g0} + ß_{g1} + ß_{g2} + ß_{g3}$$

$$\log E(y_{gi} | sexmale=0, matingpolygamous=1) = ß_{g0} + ß_{g2}$$

Then, the differential expression between males and females among polygamous flies is:

$$\log E(y_{gi} | sexmale=1, matingpolygamous=1) - \log E(y_{gi} | sexmale=0, matingpolygamous=1)$$

$$= ß_{g1} + ß_{g3}$$

Hence, here we test the null hypothesis $H_0: ß_{g1} + ß_{g3} = 0$

R Notebook

```
## Part(b)
## make contrast to test Ho: beta0+beta3=0
colnames(design)=c("intercept","male","polygamy","male_polygamy")
contrast=makeContrasts(male+male_polygamy,levels=make.names(colnames(desig
n)));
test=glmLRT(fit,contrast = contrast);
de_poly=decideTestsDGE(test,p.value=0.1)
table(de_poly)
```

```
## de_poly
##   -1    0    1
##  393 9175  111
```

Therefore,

there are **111** male-biased genes among polygamous flies;

there are **393** female-biased genes among polygamous flies;

and there are **9175** unbiased genes among polygamous flies.

We can see that most of the genes are sexually unbiased among polygamous flies.

Part (c)

Hence, we want to test the differential expression between monogamy and polygamy among male flies. The counts of genes of monogamy and polygamy among male flies are, respectively:

$$\log E(y_{gi} | sexmale=1, matingpolygamous=0) = ß_{g0} + ß_{g1}$$

$$\log E(y_{gi} | sexmale=1, matingpolygamous=1) = ß_{g0} + ß_{g1} + ß_{g2} + ß_{g3}$$

Then, the differential expression between monogamy and polygamy among male flies is:

$$\log E(y_{gi} | sexmale=1, matingpolygamous=0) - \log E(y_{gi} | sexmale=1, matingpolygamous=1)$$

$$= - (ß_{g2} + ß_{g3})$$

That is, here we test the null hypothesis H_0: $- (ß_{g2} + ß_{g3}) = 0$

R Notebook

```
## Part(c)
## use the male- and female-biased genes in polygamous flies from Part(b)
male_biased=which(de_poly==1)
female_biased=which(de_poly==-1)
unbiased=which(de_poly==0)

## make contrast to test Ho: -(beta2+beta3)=0
contrast=makeContrasts(-(polygamy+male_polygamy),levels=make.names(colnames(d
esign)));

test_male_biased=glmLRT(fit[male_biased,],contrast = contrast);
male=decideTestsDGE(test_male_biased,p.value=0.1)
table(male);

## male
## -1  0  1
## 12 92  7

test_female_biased=glmLRT(fit[female_biased,],contrast = contrast);
female=decideTestsDGE(test_female_biased,p.value=0.1)
table(female)

## female
##  -1   0   1
##   4 343  46

test_unbiased=glmLRT(fit[unbiased,],contrast = contrast);
un=decideTestsDGE(test_unbiased,p.value=0.1)
table(un)
```

```
## un
##   -1    0    1
##   21 9090   64
```

```
## test the difference significance between samples
t.test(male,un,alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  male and un
## t = -1.2678, df = 110.14, p-value = 0.1038
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf 0.01533628
## sample estimates:
##    mean of x    mean of y
## -0.045045045  0.004686649
```

```
t.test(female,un,alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  female and un
## t = 5.935, df = 394.68, p-value = 3.219e-09
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.07379732        Inf
## sample estimates:
##   mean of x   mean of y
## 0.106870229 0.004686649
```

Conclusions of the paper:

In the paper, figure 2 (b) showed the distributions of male- and female-biased genes in the male whole-fly transcriptome:

(1) The female-biased genes (n=4305) in male flies showed increased expression (relative to unbiased genes) under monogamy (relative to polygamy), $p<0.001$.

(2) The male-biased genes (n=3118) in male flies exhibited decreased expression (relative to unbiased genes) under monogamy (relative to polygamy), $p<0.001$.

Result of my analysis:

For male-biased genes in male flies, they show a decreased expression pattern (relative to unbiased genes), p=0.1038 (nearly significant at 0.1 significance level).

For female-biased genes in male flies, they show a significant increased expression pattern (relative to unbiased genes), p<0.001.

Hence, the behavior of the male-biased genes in male flies is basically consistent with what the paper concluded, with a significance level around 0.1. And the behavior of the female-biased genes in male flies indeed support the conclusions of the paper, with a significance level lower than 0.001.

Problem3:

R Notebook

```
n=1000
sims=200
mu_rep=rep(0,length=n)
mu_seq=seq(0,5,length=n)
MLE_vs_JS=matrix(NA,nrow=sims,ncol=4)
colnames(MLE_vs_JS)=c("MLE_rep","JS_rep","MLE_seq","JS_seq")

for(i in 1:sims){
  X_rep=rnorm(n,mean=mu_rep,sd=1);
  X_seq=rnorm(n,mean=mu_seq,sd=1);
  MLE_rep=X_rep
  MLE_seq=X_seq
  JS_rep =X_rep*( 1 - (n-2)/(t(X_rep)%*%X_rep) );
  JS_seq =X_seq*( 1 - (n-2)/(t(X_seq)%*%X_seq) );

  MLE_vs_JS[i,1]=sum( (MLE_rep-mu_rep)^2 )/n;
  MLE_vs_JS[i,2]=sum( ( JS_rep-mu_rep)^2 )/n;
  MLE_vs_JS[i,3]=sum( (MLE_seq-mu_seq)^2 )/n;
  MLE_vs_JS[i,4]=sum( ( JS_seq-mu_seq)^2 )/n;
}

colSums(MLE_vs_JS)/sims

##     MLE_rep      JS_rep     MLE_seq      JS_seq
## 1.003513335 0.001914169 1.004846212 0.897409822
```

In both cases, the average error over 200 simulations is smaller for **James-Stein estimator**. And the MLE estimator will always be close to 1.

Hence, the **James-Stein method** is more accurate.