Summary of the FastQC Analysis

Siddhant Srivastava

September 18, 2016

Summary

This report describes the brief descriptive data analysis of the quality of the mapped reads performed with TopHat for the three fetus and three adults sampels. Both the alignment of the RNA-seq of the samples and the quality of the reads was performed Galaxy. The FastQC files, both the html and txt files were downloaded to performed the descriptive analysis of quality of the reads.

Note This report was prepared with R (https://www.r-project.org/) and knitr (http://yihui.name/knitr/) Find below the dataset containing the the following information:

1. run = RNA-seq run ID of the samples
2. sample = Sample ID
3. sra = SRA of the sample
4. age = age of the individuals from where the samples were collected
5. ageg = age group (fetus, adults)
6. sex = gender of individuals from where the samples were collected
7. race = race of the individuals
8. rin = RNA-seq RIN for each sample
9. fraction = type of sample (cytosol, nucleus, or total; in our case all are total RNA)
10. input = sequence input for TopHat alignment
11. lrmapped = left mapped reads
12. lrxalgn = left mapped reads with multiple alignments
13. rrmapped = right mapped reads
14. rrxalgn = right mapped reads with multiple alignments
15. pairsalgn = pairs alignments
16. pairsxalgn = pairs with multiple alignments
17. pairsdic = pairs with discordant alignment
18. concrdate = percentage of concordant alignment

```
##          run Sample       sra   age  ageg    sex race rin fraction
## 1 SRR1554538  R3462 SRS686966 -0.40 fetus female   AA 6.4    total
## 2 SRR1554541  R3485 SRS686969 -0.38 fetus   male   AA 5.7    total
## 3 SRR1554566  R4706 SRS686994 -0.50 fetus   male HISP 8.3    total
## 4 SRR1554535  R2869 SRS686963 41.58 adult   male   AA 8.7    total
## 5 SRR1554534  R2857 SRS686962 40.42 adult   male   AA 8.4    total
## 6 SRR1554561  R4166 SRS686989 43.88 adult   male   AA 8.7    total
##      input lrmapped lrxalgn rrmapped rrxalgn pairsalgn pairsxalgn
## 1 68026190 67088091 3738770 67075490 3735467  66250920    3662859
## 2 69278357 68265941 3636033 67776469 3597945  66908646    3524849
## 3 53161501 52335170 3270989 52169854 3256793  51483760    3196366
## 4 38063721 37316159 1387617 37210955 1381431  36539910    1343138
## 5 28181772 27608601 1407014 27497548 1400106  27003303    1368095
## 6 39272751 38337793 1841343 38304718 1838653  37448178    1789054
##   pairsdisc concrdate
## 1   1585480     0.951
## 2   1369026     0.946
## 3   1063015     0.948
```

```
## 4    650680     0.943
## 5    441005     0.943
## 6    710131     0.935
```

Find below the new dataset with these three additional variable: **percent mapped (pct_mapped), average of per sequence quality (psqallmean), quality scores (psqscallmean), and percentageGC (gc_concent)**.

```
## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion
```

```
##          run Sample       sra   age  ageg    sex rin pct_mapped psqallmean
## 1 SRR1554538  R3462 SRS686966 -0.40 fetus female 6.4  0.9861171   34.80674
## 2 SRR1554541  R3485 SRS686969 -0.38 fetus   male 5.7  0.9818536   33.63508
## 3 SRR1554566  R4706 SRS686994 -0.50 fetus   male 8.3  0.9829014   34.00784
## 4 SRR1554535  R2869 SRS686963 41.58 adult   male 8.7  0.9789783   33.46695
## 5 SRR1554534  R2857 SRS686962 40.42 adult   male 8.4  0.9776913   35.81316
## 6 SRR1554561  R4166 SRS686989 43.88 adult   male 8.7  0.9757721   35.01644
##   psqscallmean gc_content
## 1      3965141       0.47
## 2      3960700       0.46
## 3      3128929       0.49
## 4      2027034       0.47
## 5      1571383       0.51
## 6      2182210       0.52
```

To evaluate differences in reads quality between age groups, the dataset was subset by column 6 above, and selected the last four variables added to the dataset. The summary statistics for each age group-subset was then calculated to see differences between the age groups.

Find below the descriptive statistics for the **fetus** age group.

```
##    pct_mapped       psqallmean     psqscallmean       gc_content
##  Min.   :0.9819   Min.   :33.64   Min.   :3128929   Min.   :0.4600
##  1st Qu.:0.9824   1st Qu.:33.82   1st Qu.:3544814   1st Qu.:0.4650
##  Median :0.9829   Median :34.01   Median :3960700   Median :0.4700
##  Mean   :0.9836   Mean   :34.15   Mean   :3684923   Mean   :0.4733
##  3rd Qu.:0.9845   3rd Qu.:34.41   3rd Qu.:3962921   3rd Qu.:0.4800
##  Max.   :0.9861   Max.   :34.81   Max.   :3965141   Max.   :0.4900
```

Find below the descriptive statistics for the **adult** age group.

```
##    pct_mapped       psqallmean     psqscallmean       gc_content
##  Min.   :0.9758   Min.   :33.47   Min.   :1571383   Min.   :0.470
```

```
##  1st Qu.:0.9767   1st Qu.:34.24   1st Qu.:1799208   1st Qu.:0.490
##  Median :0.9777   Median :35.02   Median :2027034   Median :0.510
##  Mean   :0.9775   Mean   :34.77   Mean   :1926876   Mean   :0.500
##  3rd Qu.:0.9783   3rd Qu.:35.41   3rd Qu.:2104622   3rd Qu.:0.515
##  Max.   :0.9790   Max.   :35.81   Max.   :2182210   Max.   :0.520
```

From the summary statistics above, we can see that the **mean of the percentage mapped reads (see pct_mapped columns)** in the fetus group is **1% higher** than the adult group. In addition, there is a **similar per sequence quality** between both groups, but a **higher per sequence quality score in the fetus group**. Lastly, there seems to be more GC concent for the adult group.