Task 3: QC the Alignment

Dongyuan Wu

2020-06-07

In task 2, I downloaded FASTQ data using EBI SRA on Galaxy, and then used HISAT2 for alignment. There are totally 12 samples (6 fetal vs. 6 adult) available. However, due to the resource limitation, I only chose half of them (3 fetal vs. 3 adult) to analyze. In addition, there are 2 runs (SRR15545xx and SRR20713xx) for each sample. Only SRR15545xx runs were used.

In this task, I used FastQC on Galaxy to do quality control. All the parameters were default. In this way, I got two kinds of results for each sample: a webpage report with summary graphs and a text file with statistical results. Due to page limitation, I directly summarize information from the text file using R (without showing any graphs).

I am not sure what kind of number from report represents *"average quality score of mapped reads"*, so I calculated two numbers:

1. **mean_perbase**: I extracted the `Mean` column from `Per base sequence quality` table, and calculated the avarage of this column.
2. **mean_perseq**: This one is computed from `Per sequence quality scores` table. The rule is: Calculated the sum of `Quality` × `Count`, and then divided by the sum of `Count`.

run	isolate	age	sex	RIN	sample	nreads	perc_read	mean_perbase	mean_perseq
SRR1554537	R3452	-0.3836	female	9.6	fetal	55133946	100%	34.09632	33.72967
SRR1554538	R3462	-0.4027	female	6.4	fetal	68026190	100%	34.79535	34.44037
SRR1554541	R3485	-0.3836	male	5.7	fetal	69278357	100%	33.61498	33.26189
SRR1554535	R2869	41.5800	male	8.7	adult	38063721	100%	33.43514	33.02989
SRR1554536	R3098	44.1700	female	5.3	adult	21450348	100%	34.22886	33.87449
SRR1554539	R3467	36.5000	female	9.0	adult	33742728	100%	35.84102	35.49709

Question: Is the mapping rates similar for fetal and adult samples?

Yes. As we can see, both fetal group and adult group have 100% mapping rates.

Question: Is there a trend in the average quality score of mapped reads?

No. I used two-sided t-test to compare fetal group and adult group. The results below show that neither **mean_perbase** nor **mean_perseq** has difference between two groups under the 0.05 significance level (their p-values are all larger than 0.05).

```
t.test(dat$mean_perbase[dat$sample == "fetal"], dat$mean_perbase[dat$sample == "adult"])
```

```
##
##  Welch Two Sample t-test
##
## data:  dat$mean_perbase[dat$sample == "fetal"] and dat$mean_perbase[dat$sample == "adult"]
## t = -0.42319, df = 2.8888, p-value = 0.7017
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.890798  2.225219
## sample estimates:
## mean of x mean of y
##  34.16888  34.50167
```

```
t.test(dat$mean_perseq[dat$sample == "fetal"], dat$mean_perseq[dat$sample == "adult"])
```

```
##
##  Welch Two Sample t-test
##
## data:  dat$mean_perseq[dat$sample == "fetal"] and dat$mean_perseq[dat$sample == "adult"]
## t = -0.40352, df = 2.8531, p-value = 0.7149
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.947890  2.301529
## sample estimates:
## mean of x mean of y
##  33.81064  34.13382
```