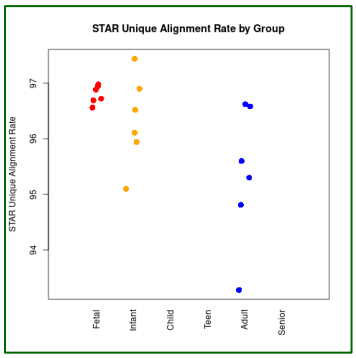
In preparation for Week 10, I have also started to upload reproducibility materials on GitHub:

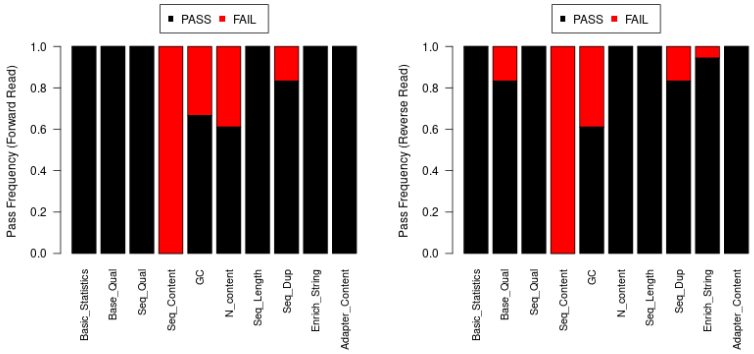
https://github.com/cwarden45/JHU_Coursera_GDS_Capstone/blob/main/Week5/README.md

For example, questions about the alignment rate were discussed in Week 4 (whose report can be seen here). Briefly, **the alignment rate looked good for most samples, with the possible exception of SRR1554536.** If using the **STAR unique alignment rate**, it would be **greater than 90%** for all samples.



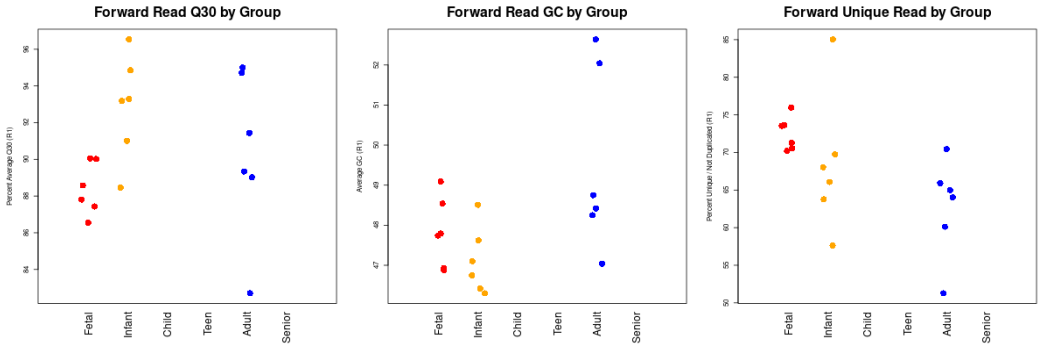
This week, I also added a plot by group for the STAR unique alignment rate (**shown right**), which significantly varies between Adult and Fetal samples. However, this is less significant than the difference in aligned counts from last week's report.

If I run FastQC on the starting reads, then I can check for flags defined by FastQC:



The "*Per base sequence content*" (**Seq_Content**) always has a "FAIL" status, so I did not focus on that as a confounding factor between the groups. The variation for the other variables is somewhat different for the forward (R1) and reverse (R2) read. However, "*Per sequence GC content*" (**GC**) and "*Sequence Duplication Levels*" (**Seq_Dup**) varies for both R1 and R2. So, I defined possible QC variables per sample for comparison based upon R1.

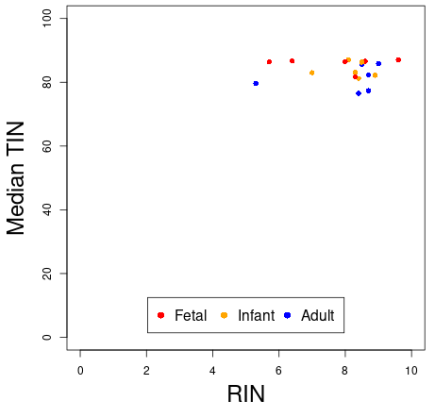
Because it was requested to be checked for this assignment, I have also extracted **percent average Q30** scores (the fraction of reads with quality scores greater than or equal to 30). That is plotted by group below, along with a selected summarization of the 2 variables mentioned above (average GC and percent unique, for **Seq_Content** and **Seq_Dup** respectively).



The unique read percent significantly varies. However, the FASTQ FastQC summary doesn't directly relate to the alignment, so I also ran **RSeQC** for RNA-Seq specific alignment quality control metrics.

First, in order to verify the library strand for the next step, I used the "*infer_experiment.py*" function to show that the library was **unstranded** (approximately 50% of reads on each strand within the housekeeping gene regions).

Second, RSeQC can calculate TIN (Transcript Integrity Number) scores that are based upon the aligned RNA-Seq reads to housekeeping genes, with the intention of reflecting RNA quality similar to RIN (RNA Integrity Number) scores calculated from ribosomal RNA profiles. This TIN calculation is accomplished using "**tin.py**". You can see the correlation between median TIN score and RIN scores to the right:



Surprisingly, there are not as many lower TIN scores (below 80 for TIN, versus below 8 for RIN, for example). There is also a subtle (but significant) difference between groups where the TIN scores are higher for the Fetal samples, although I would usually not consider this amount of difference sufficient to flag a sample as poor quality (if it was in the other direction).

An alternative view more similar to the presentation slides is also shown on the GitHub page (for reasons of space). If the scale of the TIN score plot was different, then this TIN plot might look somewhat like the STAR alignment rate opening the report (although perhaps surprising if those were expected to be lower for Fetal samples), but with a slightly higher ANOVA p-value.

In short, downstream analysis may help provide additional understanding about impact of these QC metrics:

1) RIN / TIN difference
2) Qualitatively Lower Q30 in Fetal Samples
3) Qualitatively Higher Unique Sequence Rate in Fetal Samples (ANOVA p-value < 0.01)
4) 2 GC Outliers (within Adult Samples)

I have included a summary of p-values (**Adult-versus-Fetal**) below:

Comparison	Method	P-value
Group (Adult/Fetal) vs STAR unique alignment rate	ANOVA	0.019
Group (Adult/Fetal) vs STAR unique alignment count	ANOVA	0.00932
Group (Adult/Fetal) vs percent average Q30	ANOVA	0.34
Group (Adult/Fetal) vs average GC	ANOVA	0.12
Group (Adult/Fetal) vs percent unique	ANOVA	0.0063
Group (Adult/Fetal) vs TIN	ANOVA	0.033
Group (Adult/Fetal) vs RIN	ANOVA	0.69
RIN vs TIN	Linear Regression	0.94

You can also see a plot of correlation between various variables in this heatmap from Week 7.