We are only required to compare the "**Adult**" and "**Fetal**" group samples. However, given the possibility of confounding factors among the Fetal samples, I also downloaded and aligned the "*Infant*" samples (the next closest set of samples that I am capable of downloading).

Reads were re-aligned using STAR, which is specifically designed with RNA-Seq data in mind (in order to handle splitting reads across splice junctions).

I will discuss this more in future reports, but I noticed that the ranking of samples changed when different methods. This is most clear for sample SRR1554536 :

Sample	**Group**	***Unique* Aligned Percent (STAR Log)**	**Aligned Percent (samtools flagstat)**	**Aligned Percent (GenomicAlignments)**
SRR1554537	**Fetal**	96.69%	99.67%	98.21%
SRR1554538	**Fetal**	96.72%	99.66%	98.33%
SRR1554541	**Fetal**	96.95%	99.75%	98.13%
SRR1554566	**Fetal**	96.56%	99.72%	99.45%
SRR1554567	**Fetal**	96.98%	99.71%	97.91%
SRR1554568	**Fetal**	96.88%	99.69%	98.47%
SRR1554544	**Infant**	96.11%	99.68%	78.54%
SRR1554546	**Infant**	95.10%	99.73%	74.07%
SRR1554549	**Infant**	96.52%	99.53%	87.51%
SRR1554551	**Infant**	95.94%	99.62%	78.82%
SRR1554553	**Infant**	97.44%	99.74%	91.62%
SRR1554554	**Infant**	96.90%	99.62%	90.91%
SRR1554534	**Adult**	94.81%	99.24%	84.51%
SRR1554535	**Adult**	95.30%	99.31%	83.33%
SRR1554536	**Adult**	93.28%	99.78%	42.90%
SRR1554539	**Adult**	96.62%	99.10%	90.00%
SRR1554556	**Adult**	96.58%	99.70%	87.89%
SRR1554561	**Adult**	95.60%	99.15%	89.07%

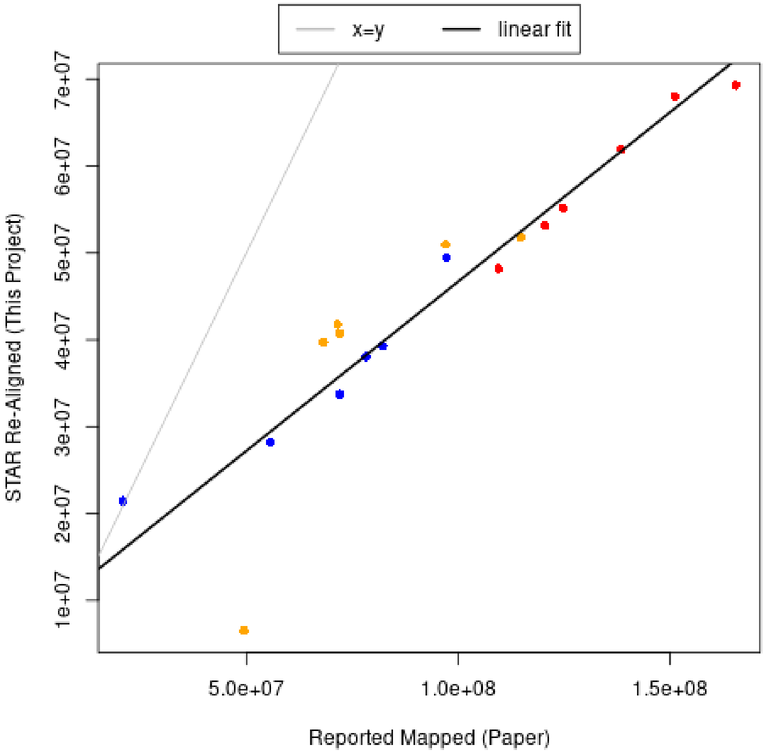
I will focus on annotated gene read counts for downstream analysis. The GenomicAlignments alignment rate is correlated with a relatively more similar absolute value to the quantification rate in Week 6. However, if you normalize to the read counts, then that may adjust for most of those differences.

I also downloaded the full metadata, including NCBI Biosample accessions that were defined without depositing the associated data. The deposited data is unbalanced with respect to RIN score, but SRR1554537 is a Fetal sample that was deposited with a RIN score of 9.6. There is also an Adult sample (SRR1554536) with a lower RIN score of 5.3. So, perhaps that is OK for an Adult-vs-Fetal comparison.

Most samples used for the direct comparison are African American. If you only count each of the cytosol / nucleus samples once (per donor), then there were a total of 30 African American (AA), 12 Caucasian (CAUS), 1 Asian (AS), and 1 Hispanic (HISP) samples. The only publicly deposited sample for the Adult-vs-Fetal comparison that is not African-American is SRR1554566. So, if we wanted to be extra safe, we could remove that fetal Hispanic sample. There are 4 African-American and 2 Caucasian "Infant" samples, which could also be filtered to only use Afrcian-American samples (if needed).

There are 30 male and 14 female samples. Each of the 3 downloaded groups have exactly 4 male and 2 female samples.

Each Biosample has 2 runs. Only 1 run per sample was listed in the project description, so that is all that I downloaded and re-aligned. So, the absolute counts are noticeably different than reported in the paper, but the aligned fragments are well correlated with the mapping count reported in the paper:



Please note that the plot above uses absolute counts, rather than the percentages shown in the earlier table. **However, the point is that I believe this is a good match for what was expected.**

So, if the proportion of total reads varied between runs, then that could explain why the sample with the lowest absolute unique STAR alignment count (Infant sample SRR1554553) is not the sample with the lowest alignment percentage (Adult sample SRR1554536). Normalization should correct for varying total / aligned reads, but we can try to assess if that was effective or not during downstream analysis.

Based upon the paper, I am hypothesizing that the Fetal samples should look more different than the Infant and Adult samples. However, exactly how similarity is defined can be important.

In preparation for Week 10, I have also started to upload reproducibility materials on GitHub:
https://github.com/cwarden45/JHU_Coursera_GDS_Capstone/tree/main/Week4

While currently incomplete, there is Week 5 content on GitHub (since some of that is used in the code for this week's assignment).