Task 3. Genomic Data Science Capstone.

For this task, six RNASeq samples were send from NCBI Short Read Archieve to Galaxy version 2.10.8 using the tool “Faster Download and Extract Reads in FASTQ from NCBI SRA” using its defaults. All additional analyses were performed in Galaxy. Of the six samples, three represented fetal brain samples (SRR1554537, SRR1554538, SRR1554541), and three represented adult brain samples (SRR1554534, SRR1554535, SRR1554539). These sequences were mapped to the human genome build GRCh37/hg19 from February 2009 using HISAT 2.1 with all options set for the default Galaxy parameters, except from output summaries to file and in machine friendly format (both set to ‘yes’), and the spliced alignment option to output a bam file compatible with StringTie for further analysis. In Table 1 and Fig. 1 are presented the quality control check performed with FastQC 0.11.8 and HISAT and summarized utilizing MultiQC 1.8. The alignments for fetal and adult samples passed most quality check parameters, except for per base sequence content, which presented an uneven patters in the first ~10 bases, that can be related to “random” primers used for Illumina RNA-Seq library preparation, and said to not influence final results. Additionally, SRR1554534 presented low quality base pairs in the end of part of the sequences, failing FastQC mean quality scores control. Even so, all samples presented a trend of diminished quality toward the end of the reads (Fig. 2). As expected, there were more sequences in the fetal samples, and only the fetal samples presented samples that failed MultiQC quality control. The percentage of aligned samples was similar between fetal and adult samples. Most samples presented similar GC content and number of duplicates, and only the fetal samples presented failed samples. The percentage of mapped reads seems appropriate, with similar mapping rates for fetal and adult samples. Most of the reads presented an phread quality above 28, with median values ranging between 36-38.

Table 1. Summary statistics from short read alignment to the reference genome GRCh37/hg19 using HISAT 2.10 in Galaxy 2.10.8. The first three samples represents the fetal samples, the three last are the adult samples.

Sample Name	% Aligned	% Dups	% GC	Length	% Failed	M Seqs
SRR1554537	98.5%	15.0%	48%	100 bp	10%	117.9
SRR1554538	98.8%	17.8%	47%	100 bp	10%	145.6
SRR1554541	98.7%	19.1%	46%	100 bp	20%	147.3
SRR1554534	98.6%	17.0%	51%	100 bp	0%	59.9
SRR1554535	98.8%	16.4%	47%	100 bp	0%	80.0
SRR1554539	98.7%	17.0%	48%	100 bp	0%	70.4

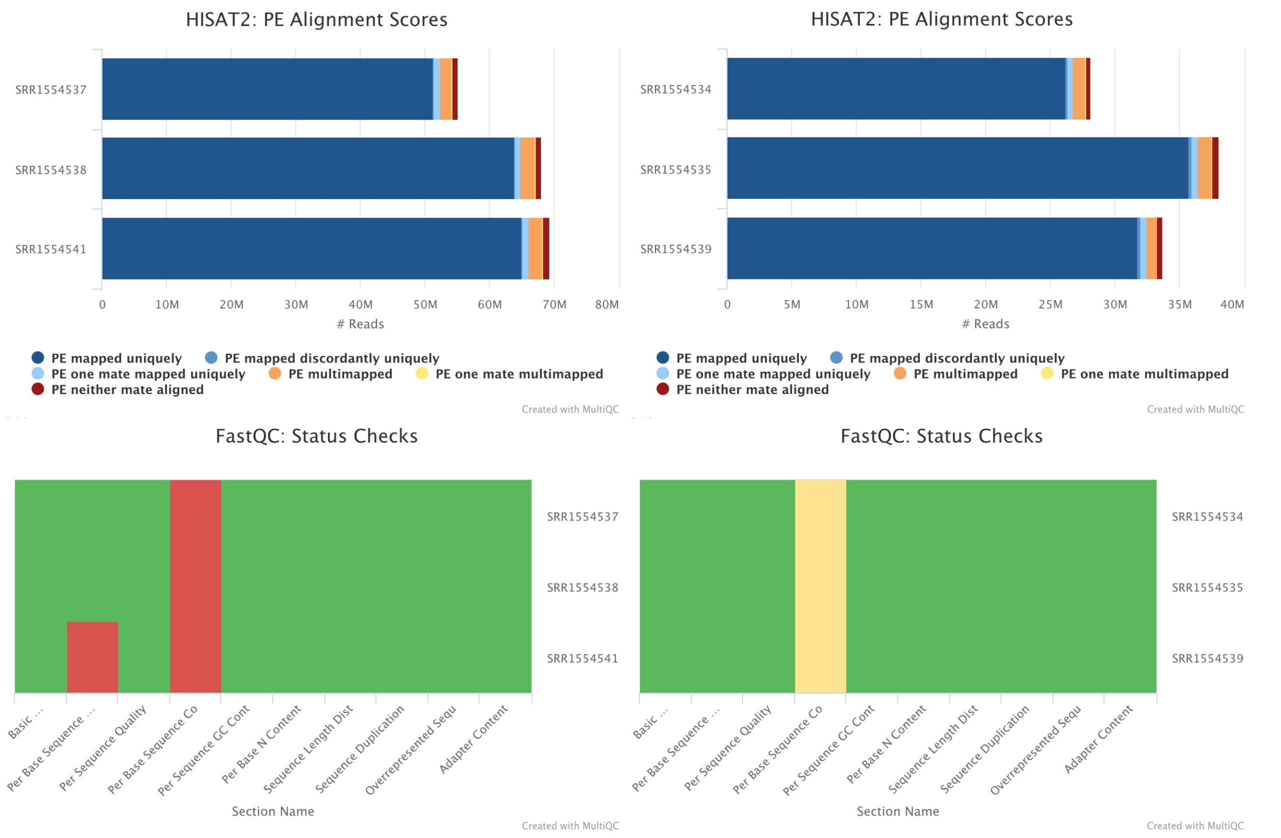

Fig. 1. On the left side are the HISAT2 alignment scores (upper) and FastQC status checks (lower) for the fetal samples. On the right side are the HISAT2 alignment scores (upper) and FastQC status checks (lower) for the adult samples. Graphs generated in MultiQC.

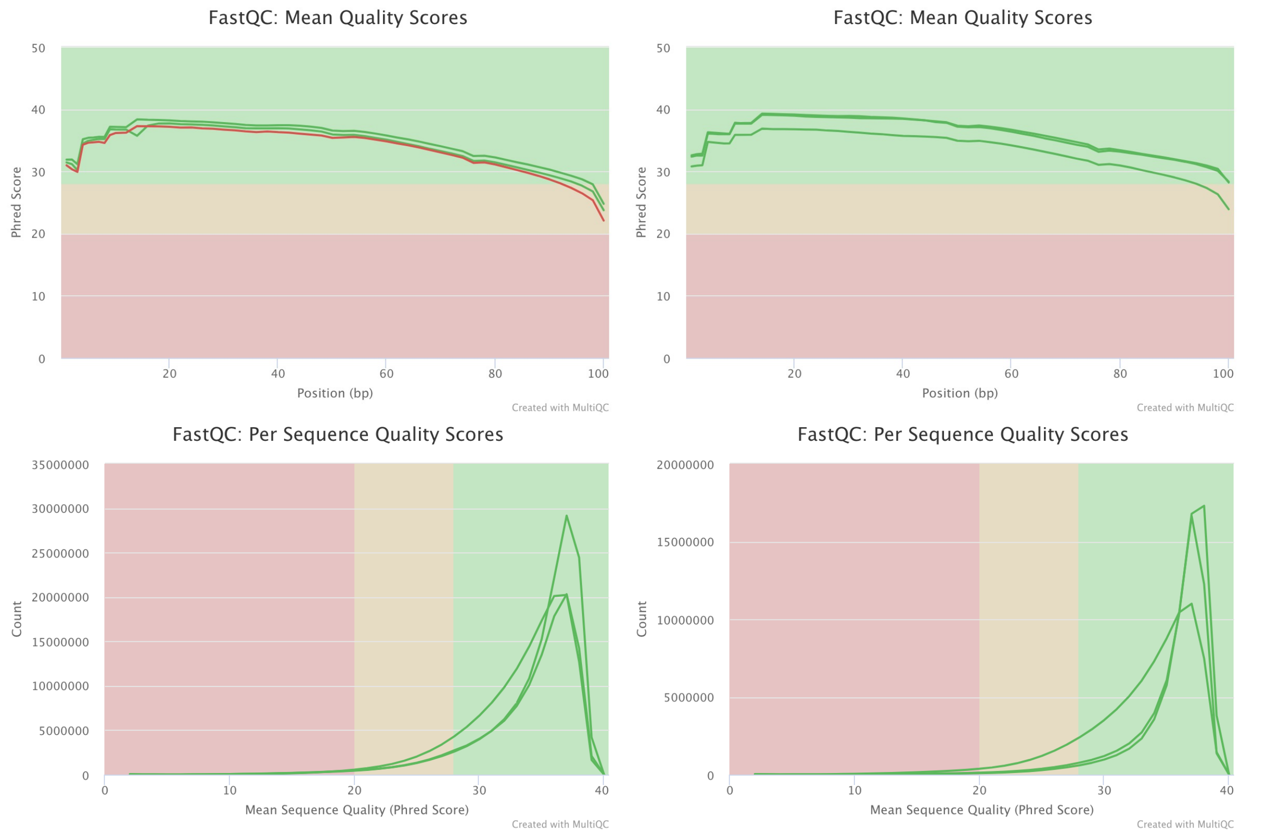

Fig. 2. Mean quality scores of fetal (left) and adult (right) samples. Quality scores analyzed with FastQC and plots generated with MultiQC.