Task 2. 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. 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 are presented the alignment summaries for each of the six samples. Checking for quality control using FastQC 0.11.8 in the obtained alignments showed that all samples have about 13 nucleotides in the 5’ side that presented very poor quality and needed to be trimmed. Additional quality control measures obtained suggests that the reads can be filtered to keep only reads with quality of 30 or above. Given these considerations, the reads from all samples were filtered using Trimmomatic 0.38, using sliding window trimming with average quality 30 (averaged across four bases), a minimum length of 75 bases, and with the first 13 bases from the start of each sequence being cropped. The paired reads returned after filtering with Trimmomatic were then aligned using HISAT with the same parameters specified above. The summaries for the alignments of the filtered sequences are presented at Table 2.

For both alignment strategies used, the overall alignment rates was larger than 96% using the hg19 genome, which was expected as all data come from human samples, and are of good quality. This suggests that the splice aligner used was appropriated for the task.

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	Total pairs	Not aligned concordantly or discordantly	Aligned concordantly once	Aligned concordantly more than once	Aligned discordantly	Total unpaired reads	Not aligned	Aligned once	Aligned more than once	Overall alignment rate
SRR1554537	55,133,946	1,772,449 (3.21%)	51,343,735 (93.13%)	1,839,854 (3.34%)	177,908 (0.32%)	3,544,898	1,661,016 (46.86%)	1,691,201 (47.71%)	192,681 (5.44%)	98.49%
SRR1554538	68,026,190	1,716,294 (2.52%)	63,816,033 (93.81%)	2,258,216 (3.32%)	235,647 (0.35%)	3,432,588	1,603,700 (46.72%)	1,610,868 (46.93%)	218,020 (6.35%)	98.82%
SRR1554541	69,278,357	1,964,936	64,946,039	2,170,100	197,282	3,929,872	1,877,084	1,854,233	198,555	98.65%

		(2.84%)	(93.75%)	(3.13%)	(0.28%)		(47.76%)	(47.18%)	(5.05%)	
SRR1554534	28,181,772	848,926 (3.01%)	26,194,327 (92.95%)	951,201 (3.38%)	187,318 (0.66%)	1,697,852	791,153 (46.60%)	813,395 (47.91%)	93,304 (5.50%)	98.60%
SRR1554535	38,063,721	1,050,105 (2.76%)	35,708,040 (93.81%)	1,066,736 (2.80%)	238,840 (0.63%)	2,100,210	951,095 (45.29%)	1,038,266 (49.44%)	110,849 (5.28%)	98.75%
SRR1554539	33,742,728	932,218 (2.76%)	31,785,139 (94.20%)	762,344 (2.26%)	263,027 (0.78%)	1,864,436	858,995 (46.07%)	914,569 (49.05%)	90,872 (4.87%)	98.73%

Table 2. Summary statistics from short read sequences filtered using Trimmomatic and aligned 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	Total pairs	Not aligned concordantly or discordantly	Aligned concordantly once	Aligned concordantly more than once	Aligned discordantly	Total unpaired reads	Not aligned	Aligned once	Aligned more than once	Overall alignment rate
SRR1554537	17,059,825	1,014,901 (5.95%)	15,133,599 (88.71%)	634,949 (3.72%)	276,376 (1.62%)	2,029,802	984,438 (48.50%)	953,715 (46.99%)	91,649 (4.52%)	97.11%
SRR1554538	31,732,959	1,782,583 (5.62%)	28,378,607 (89.43%)	1,159,398 (3.65%)	412,371 (1.30%)	3,565,166	1,746,021 (48.97%)	1,672,999 (46.93%)	146,146 (4.10%)	97.25%
SRR1554541	19,743,031	1,090,321 (5.52%)	17,721,195 (89.76%)	662,212 (3.35%)	269,303 (1.36%)	2,180,642	1,073,199 (49.21%)	1,022,919 (46.91%)	84,524 (3.88%)	97.28%
SRR1554534	17,147,070	1,185,710 (6.91%)	14,726,316 (85.88%)	588,957 (3.43%)	646,087 (3.77%)	2,371,420	1,108,297 (46.74%)	1,139,286 (48.04%)	123837 (5.22%)	96.77%
SRR1554535	12,241,098	612,012 (5.00%)	11,012,042 (89.96%)	356,685 (2.91%)	260,359 (2.13%)	1,224,024	577,471 (47.18%)	588,253 (48.06%)	58300 (4.76%)	97.64%
SRR1554539	20,447,157	1,239,215 (6.06%)	18,248,932 (89.25%)	500,212 (2.45%)	458,798 (2.24%)	2,478,430	1,208,867 (48.78%)	1,181,399 (47.67%)	88164 (3.56%)	97.04%