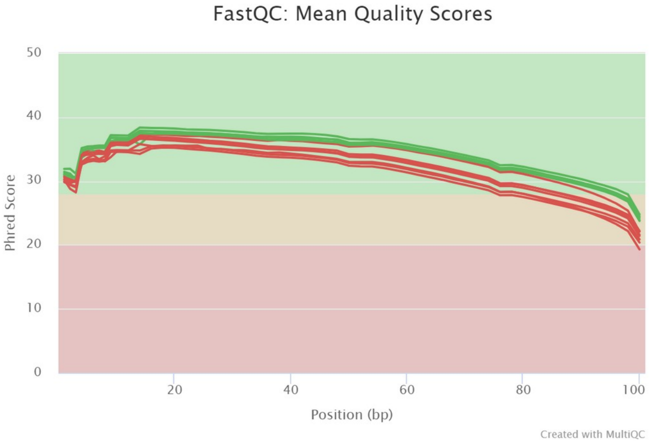
Re-analysis of fetal and adult brain raw RNA-seq data from the study “Developmental regulation of human cortex transcription and its clinical relevance at base resolution” (Jaffe et al, 2015 Jan PMID:25501035)

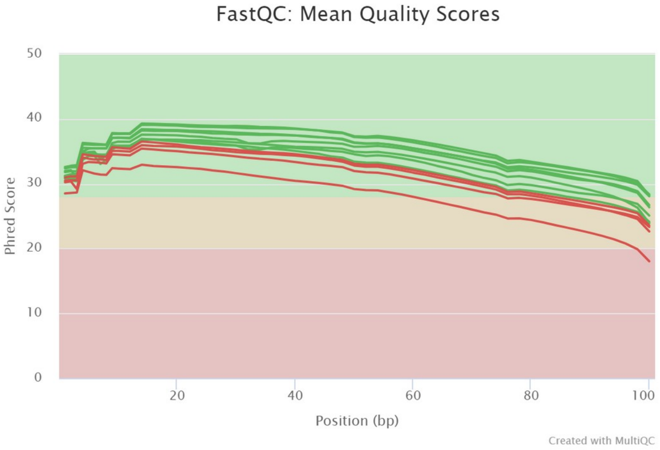
Task 2: Quality control of the alignment

For this task, I first ran on the Galaxy Main Server the FastQC Read Quality reports (Galaxy Version 0.72+galaxy1) tool with default parameters on all the 24 Dataset collections (containing BAM files) produced by HISAT2. Then, I used MultiQC aggregate results from bioinformatics analyses into a single report (Galaxy Version 1.8+galaxy1) tool to aggregate quality data from each of the 2 groups of samples (Adults and Fetals), setting “FastQC” as the tool used to generate logs and “Raw data” as Type of FastQC output.

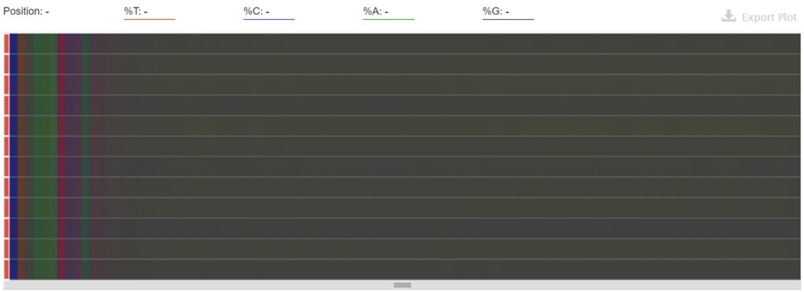
Looking at the various plots in the MultiQC Webpage output, the major problem affecting some of the samples appears to be the mean per base phred-scaled quality score, which tend (not surprisingly) to reduce from the beginning to the end of the reads, and in some case falls below the good-quality threshold of 28. As shown in the following images (red lines), this issue affects especially the fetal samples (7/12) but also adult samples (4/12).



Fetal samples

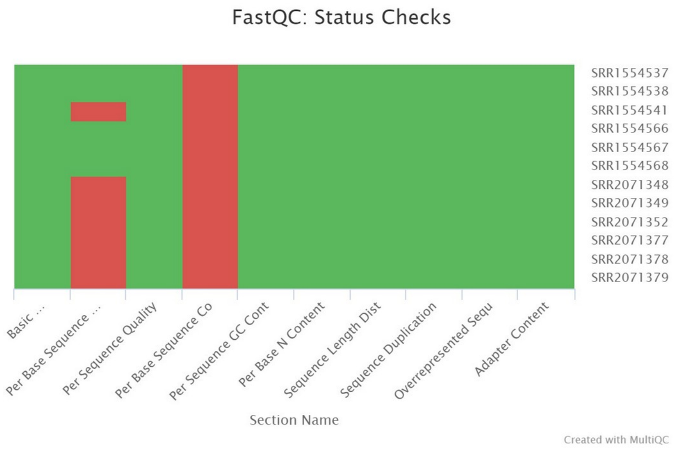


Adult samples

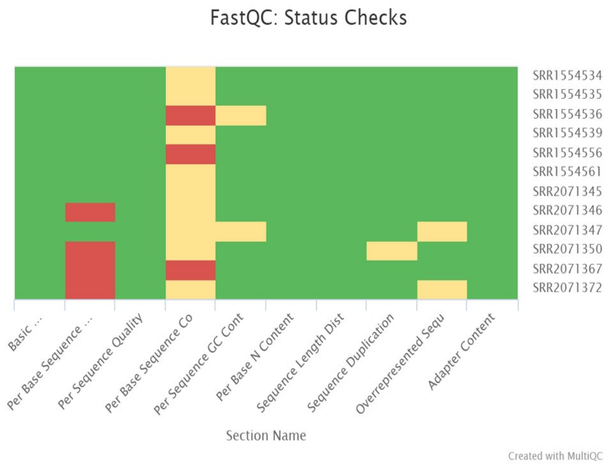


Another potential issue could be the Per Base Sequence Content, which is very unbalanced for the first 10 bp (see plot on the right, that refers to fetal samples). Anyway, considering as low-quality all the samples affected by this problem would leave us with very few good quality samples. Furthermore, as stated in 9th Discussion-28 October 2010 - BioWiki (bioinfo-core.org): “*A lot of groups have found that RNA-Seq libraries created with Illumina kits show this odd bias in the first ~10 bases of the run. This seems to be due to the 'random' primers which are used in the library generation, which may not be quite as random as you'd hope. We've not removed this biased sequence and the results seem to be OK.*” For those reasons, I chose to ignore this parameter in my quality evaluation, though I think it could be interesting to follow an analysis pipeline a bit different from the one suggested in the Capstone project, trimming the first 10bp from the reads before alignment.

So, I decided to flag as “low quality” (see Table 1) only the samples with problems in the Per base mean quality scores, according to the following summarizing plots:



Fetal samples



Adult samples

In order to respect the assignment request, I then retreived the Data from the “Per Sequence Quality Scores” plot as a .csv file, loaded it into R and calculated an average quality score for each sample, that I reported in Table 1. Here is the code I used for fetal samples:

```
fetal_means←read.csv("FetalQCmeans.csv")
fetal_run_means←as.numeric()
for(i in 2:13) {
   fetal_run_means[i]<-weighted.mean(fetal_means[,1],fetal_means[,i]/1000)
}
```

The mapping rate informations shown in Table 1 for each sample come, meanwhile, from the previous assignment (summing up the % of reads mapped exactly 1 time and the % of reads mapped >1 times)

Statistical tests

In order to answer the questions asked in the assignment and to better understand the data, I performed some two sample t-test using the R function `t.test`.

- A one-sided test on fetal and adult mapping rate (alternative Hp: fetal mean>adult mean), giving the following result:
  Mean of Fetal samples= 93.91%   Mean of Adult samples= 88.94%   p-value:0.095
  Therefore, there is no statistical evidence that fetal samples have a better mapping rate than adult ones.
- To explore the potential trend in the average quality score, I tried a two-sided test on fetal and adult groups for this parameter, but the results weren’t significant at all (respective means of 32.51 and 32.58, p-value=0.88)
- Visual inspection of Table 1 led me to suppose that a difference in average quality scores could exist between samples from the first series of reads (SRR155…) and the second one (SRR207…), possibly due to different sequencing strategies or machine. So I performed a one-sided test to see if the mean of the first series average quality score is greater than the second series one. I used my phenotype table and the code: `t.test(qcmeans[pdata$Series==1],qcmeans[pdata$Series==2],alternative="g")`.
  The results were highly significant:
  Series 1 mean =34.13   Series 2 mean=31.16   p-value= 9,41*10^-7
  So, probably for technical reasons (but I didn’t find precise informations about that), the first replicate of each sample as, on average, a significantly higher quality score than the second replicate.

Sample	Group	Run	Mapping Rates (%)	Average Quality Score
R3452_DLPFC_polyA_RNAseq_total	Fetal	SRR1554537	99.14	33.73
		SRR2071348	86.76	30.50
R3462_DLPFC_polyA_RNAseq_total	Fetal	SRR1554538	99.19	34.44
		SRR2071349	89.21	31.82
R3485_DLPFC_polyA_RNAseq_total	Fetal	SRR1554541	99.19	33.26
		SRR2071352	86.15	30.89
R4706_DLPFC_polyA_RNAseq_total;	Fetal	SRR1554566	99.27	33.67
		SRR2071377	91.47	32.02
R4707_DLPFC_polyA_RNAseq_total;	Fetal	SRR1554567	99.22	33.93
		SRR2071378	90.48	32.20
R4708_DLPFC_polyA_RNAseq_total	Fetal	SRR1554568	99.19	33.64
		SRR2071379	87.69	30.92
R2869_DLPFC_polyA_RNAseq_total	Adult	SRR1554535	98.6	33.03
		SRR2071346	71.31	27.68
R3098_DLPFC_polyA_RNAseq_total	Adult	SRR1554536	99.34	33.87
		SRR2071347	88.74	32.52
R3467_DLPFC_polyA_RNAseq_total	Adult	SRR1554539	98.51	35.50
		SRR2071350	74.37	31.61
R3969_DLPFC_polyA_RNAseq_total	Adult	SRR1554556	99.13	34.41
		SRR2071367	85.92	31.20
R4166_DLPFC_polyA_RNAseq_total	Adult	SRR1554561	98.48	34.64
		SRR2071372	77.07	30.58
R2857 DLPFC polyA+ transcriptome	Adult	SRR1554534	98.53	35.44
		SRR2071345	77.22	31.99

Table 1 (In red the samples flagged as poor-quality in the phenotype data table)