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: Alignment

For this step, I worked on the Galaxy Main Server (Galaxy (usegalaxy.org)) and I ran 2 separate workflows, one for Fetal samples and one for Adult samples (this splitting choice due to practical reasons). I aligned both the technical replicates available from all the 12 experiments (6 from Fetal brain, 6 from Adult brain).

First, I uploaded the data using the “Faster Download and Extract Reads in FASTQ format from NCBI SRA (Galaxy Version 2.10.8+galaxy0)” tool, using “SRR accession” as Input type (SRR accession numbers are available at BioSample Links for BioProject (Select 245228) - BioSample - NCBI (nih.gov)). I ran the jobs with the default options, that produce 4 outputs for each run, the relevant one for my purpose being the collection named “Pair-end Data (fasterq-dump)”.

I aligned each of those Paired-end Dataset Collections to the built-in human genome hg19 assembly using the spliced alignment tool “HISAT2 A fast and sensitive alignment program (Galaxy Version 2.1.0+galaxy5)” , with default parameters.

The following table summarizes the alignment results, as shown in the Galaxy preview of each of the BAM files generated by HISAT2.

Note: running “Samtools flagstat tabulate descriptive stats for BAM datset (Galaxy Version 2.0.3)” tool on the same BAM files gives very similar results (not reported here) on the alignment rates, but with the number of “ paired in sequencing” reads which is the exactly twice as the reads number appearing in the preview, and all the stats approximatively doubled too. That’s probably because Samtools flagstat handles the forward and the reverse strand separately.

Sample	Group	Run	Number of reads	Reads aligned 1 time		Reads aligned >1 times	
R3452_DLPFC_polyA_RNAseq_total	Fetal	SRR1554537	55133946	**52688941**	(95.57%)	**1968339**	(3.57%)
		SRR2071348	125129957	**103438455**	(82.66%)	**5135199**	(4.10%)
R3462_DLPFC_polyA_RNAseq_total	Fetal	SRR1554538	68026190	**65058884**	(95.64%)	**2418001**	(3.55%)
		SRR2071349	231868929	**197514368**	(85.18%)	**9349945**	(4.03%)
R3485_DLPFC_polyA_RNAseq_total	Fetal	SRR1554541	69278357	**66402692**	(95.85%)	**2314177**	(3.34%)
		SRR2071352	98565417	**81230039**	(82.41%)	**3689326**	(3.74%)
R4706_DLPFC_polyA_RNAseq_total;	Fetal	SRR1554566	53161501	**50650544**	(95.28%)	**2122867**	(3.99%)
		SRR2071377	66177040	**57455637**	(86.82%)	**3078685**	(4.65%)
R4707_DLPFC_polyA_RNAseq_total;	Fetal	SRR1554567	61922935	**59358909**	(95.86%)	**2079334**	(3.36%)
		SRR2071378	77670609	**67264077**	(86.60%)	**3014568**	(3.88%)
R4708_DLPFC_polyA_RNAseq_total	Fetal	SRR1554568	48184702	**46139292**	(95.76%)	**1654068**	(3.43%)
		SRR2071379	119229759	**99907401**	(83.79%)	**4655699**	(3.90%)
R2869_DLPFC_polyA_RNAseq_total	Adult	SRR1554535	38063721	**36356538**	(95.51%)	**1176452**	(3.09%)
		SRR2071346	66455655	**44561572**	(67.05%)	**2834071**	(4.26%)
R3098_DLPFC_polyA_RNAseq_total	Adult	SRR1554536	21450348	**20545796**	(95.78%)	**762692**	(3.56%)
		SRR2071347	37025651	**29994524**	(81.01%)	**2862415**	(7.73%)
R3467_DLPFC_polyA_RNAseq_total	Adult	SRR1554539	33742728	**32420443**	(96.08%)	**818997**	(2.43%)
		SRR2071350	49834056	**36065378**	(72.37%)	**994647**	(2.00%)
R3969_DLPFC_polyA_RNAseq_total	Adult	SRR1554556	49480779	**47652724**	(96.31%)	**1394503**	(2.82%)
		SRR2071367	64284397	**53110871**	(82.62%)	**2118531**	(3.30%)
R4166_DLPFC_polyA_RNAseq_total	Adult	SRR1554561	39272751	**37498673**	(95.48%)	**1176778**	(3.00%)
		SRR2071372	58125226	**42329486**	(72.82%)	**2468926**	(4.25%)
R2857 DLPFC polyA+ transcriptome	Adult	SRR1554534	28181772	**26727332**	(94.84%)	**1039669**	(3.69%)
		SRR2071345	41133326	**29437553**	(71.57%)	**2325133**	(5.65%)

As shown, the alignment rate obtained is in general good. Note that the percentage of aligned reads is always lower in the second replicate, but the total number of aligned reads is nonetheless higher, as the total number of reads produced is bigger.

It could be interesting to investigate if there is a reason why the reads from adult samples appear to align a bit worse than the fetal ones.