

Genomic Data Science Capstone: Week 5 (QC the Alignment)

For this assignment I will briefly describe the data analysis of the quality of mapped reads performed with FastQC for three fetus and adult brain samples. Both the alignment of the RNA-seq of the samples and the quality of the reads was performed using Galaxy website.

Table 1. Sample data summary and alignment analysis.

##	ID	Sample	SRA	Age	Type	Sex	Race	RIN	Fraction
##1	SRR1554538	R3462	SRS686966	- 0.40	Fetus	Female	AA	6.4	Total
##2	SRR1554541	R3485	SRS686969	- 0.38	Fetus	Male	AA	5.7	Total
##3	SRR1554566	R4706	SRS686994	- 0.50	Fetus	Male	AA	8.3	Total
##4	SRR1554535	R2869	SRS686963	41.58	Adult	Male	AA	8.7	Total
##5	SRR1554534	R2857	SRS686962	40.42	Adult	Male	AA	8.4	Total
##6	SRR1554561	R4166	SRS686989	43.88	Adult	Male	AA	8.7	Total

##	input	lrmapped	lrxalgn	rrmapped	rrxalgn	pairsalgn	pairsxalgn
##1	68026190	67088091	3738770	67075490	3735467	66250920	3662859
##2	69278357	68265941	3636033	67776469	3597945	66908646	3524849
##3	53161501	52335170	3270989	52169854	3256793	51483760	3196366
##4	38063721	37316159	1387617	37210955	1381431	36539910	1343138
##5	28181772	27608601	1407014	27497548	1400106	27003303	1368095
##6	39272751	38337793	1841343	38304718	1838653	37448178	1789054

lrmapped = left mapped reads; **lrxalgn** = left mapped reads with multiple alignments; **rrmapped** = right mapped reads; **rrxalgn** = right mapped reads with multiple alignments; **pairsalgn** = pairs alignments; **pairsxalgn** = pairs with multiple alignments;

##	pairsdisc	concrdate
##1	1585480	0.951
##2	1369026	0.946
##3	1063015	0.948
##4	650680	0.943
##5	441005	0.943
##6	710131	0.935

pairsdic = pairs with discordant alignment;
concrdate = percentage of concordant alignment.

Table 2. Data analysis of percent mapped, average of per sequence quality, quality scores and GC percentage of fetus and adult brain samples.

##	Pct_mapped	psqallmean	psqscallmean	gc_content
##1	0.9861171	34.80674	3965141	0.47
##2	0.9818536	33.63508	3960700	0.46
##3	0.9829014	34.00784	3128929	0.49
##4	0.9789783	33.46695	2027034	0.47
##5	0.9776913	35.81316	1571383	0.51
##6	0.9757721	35.01644	2182210	0.52
Pct_mapped = percent mapped; psqallmean = average of per sequence quality; psqscallmean = quality scores; and gc_content = GC percentage				

Table 3. Summary statistics calculation for each age group-subset fetus vs adult brains.

##	Pct_mapped	Psqallmean	psqscallmean	gc_content
##1	Min.: 0.9819	Min.: 33.64	Min.: 3128929	Min.: 0.4600
##2	1 st Qu.: 0.9824	1 st Qu.: 33.82	1 st Qu.: 3544814	1 st Qu.: 0.4650
##3	Median: 0.9829	Median: 34.01	Median: 3960700	Median: 0.4700
##4	Mean: 0.9836	Mean: 34.15	Mean: 3684923	Mean: 0.4733
##5	3 rd Qu.: 0.9845	3 rd Qu.: 34.41	3 rd Qu.: 3962921	3 rd Qu.: 0.4800
##6	Max. : 0.9861	Max. : 34.41	Max. : 3965141	Max. : 0.4900
Table content: Fetus samples				
##	Pct_mapped	psqallmean	psqscallmean	gc_content
##1	Min.: 0.9758	Min.: 33.47	Min.: 1571383	Min.: 0.470
##2	1 st Qu.: 0.9767	1 st Qu.: 34.24	1 st Qu.: 1799208	1 st Qu.: 0.490
##3	Median: 0.9777	Median: 35.02	Median: 2027034	Median: 0.510
##4	Mean: 0.9775	Mean: 34.77	Mean: 1926876	Mean: 0.500
##5	3 rd Qu.: 0.9783	3 rd Qu.: 35.41	3 rd Qu.: 2104622	3 rd Qu.: 0.515
##6	Max. : 0.9790	Max. : 35.81	Max. : 2182210	Max. : 0.520
Table content: Adult samples				

In summary, this data suggests that the mean of the percentage of mapped reads in fetus group is 1% higher than adult group. Besides, there is a similar per sequence quality between both groups, but with a higher per sequence quality score in fetus group. Lastly, there seems to be more GC content for adult group.