Genomic Data Science Capstone Report

Ly Le

September 2021

1 Introduction

1.1 Background

Transcriptome analysis of human brain provides fundamental insight about development and disease, but largely relies on existing annotation. We sequenced transcriptomes of 72 prefrontal cortex samples across six life stages, and identified 50,650 differentially expression regions (DERs) associated with developmental and aging, agnostic of annotation. While many DERs annotated to non-exonic sequence (41.1%), most were similarly regulated in cytosolic mRNA extracted from independent samples. The DERs were developmentally conserved across 16 brain regions and within the developing mouse cortex, and were expressed in diverse cell and tissue types. The DERs were further enriched for active chromatin marks and clinical risk for neurodevelopmental disorders like schizophrenia. Lastly, we demonstrate quantitatively that these DERs associate with a changing neuronal phenotype related to differentiation and maturation. These data highlight conserved molecular signatures of transcriptional dynamics across brain development, some potential clinical relevance and the incomplete annotation of the human brain transcriptome.

1.2 Problem

In this project, I determine whether or not the gene expression level is different between fetals and adults. To address this problem, I will re-conduct the analysis which described in this paper.

2 Data

Although there are 48 different samples belonging to 6 different age groups from fetal to old in the original research, I only randomly select 3 samples for each fetal (¡0 years) and adult (20-50 years) group. This is because the data is significantly large and some following steps are extremely time-consuming.
I downloaded these 6 samples from European Nucleotide Archive. Three of them were fetal: SRR1554538, SRR1554566, SRR1554568. The rest three datasets

were adult: SRR1554536, SRR1554556, SRR1554561. Each file was paired-end library, and there were two fastq files for each sample. For example, SRR1554538 contains SRR1554538_1 and SRR1554538_2, each file was in fastq.gz format.

3 Experiment workflow

3.1 Alignment

I made use of galaxy to align all samples to the suitable reference genome. Firstly, all files were uploaded to the server with the following settings: file type was fastqs.gz; reference genome was hg19 (b37). Then, HISAT2 (Version 2.1.0+galaxy5) is chosen to run alignment: the reference genome was built-in genome Human (Homo Sapiens)(b37) hg19; Paired-end two files for each sample. There were two output files: a BAM file contained alignment results; an alignment quality summary file.

3.2 Quality control on the alignments

FastQC (Version 0.72+galaxy1) on galaxy server was applied to perform the quality control. Number of reads were in the range of 21,450,348 to 68,026,190. Number of sequences were in the range of 45,322,851 to 147,128,996. Percentage of GC were in range 46 to 52. All the 6 alignment rates were close to 99.8%, average quality per read was 37, which indicated the alignment results were good and the quality of reads were good.
After, I want to compare the mapping rates between fetal and adult. Firstly, I gathered information for each sample from `https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=245228` by clicking "Send to - File - Download Full XML". Then, I converted the xml file into a csv file named "Week5QCdata.csv" by Python programming.

4 Reference

`http://binf.gmu.edu/swang36/NGS/Genomic_Capstone_Report.html`