BMEG3105 Data Analytics for Personalized Genomics and Precision Medicine

Lecture 16: Genomic Analysis
Lecturer: Yu LI (李煜)
Scriber: Kenny TANJAYA (1155198328)

16.0.1 - Review:

- How to detect the short variant from the reads in genetic variant calling
- The pipeline is kind of tedious, but we mentioned the important steps in last lecture

16.0.2 - Learning outcomes:

- The pipeline (Is a concrete tool you can use in the future, ensuring you know what you are expecting from each step and which file you are looking for)
- The file format (In the real analysis, reads are used for practice. You will know what to input to a specific step and in cases where you get an error, you know what to change)
- Trouble-shooting
- What will we learn today? (Genomic Analysis -> GWAS [**16.1**], RNA-seq Analysis -> Gene fusion for structural variants [**16.2**], Epigenome Analysis -> Peak calling [**16.3**])

16.1 - Genomic Analysis

- **Why do we want to study the variants?** This is to check for abnormal gene expression in the genome, epigenome or transcriptome.
- One example of further downstream analysis is **Genome-Wide Association Studies (GWAS)**, in which we can use it to try to determine whether specific variant(s) in many individuals can be associated with a trait (disease) or if it is just due to random noise. We can do this by spotting the variant that is common amongst everyone affected but absent in all those unaffected. However, only one gene affects a disease very rarely, only in ideal cases. In reality, the diseases can be related to lots of different genomic variants/SNPs, with around 3.5 million SNPs.
- **Bonferroni correction** can be used to combat this, setting a p-value at 0.05 as per the norm. What is the probability that we are making at least one error whilst processing these 3.5 million SNPs? According to Bonferroni correction, **Adjusted p-value = (p-value) / (number of SNPs to test)**.

16.2 - RNA-seq Analysis

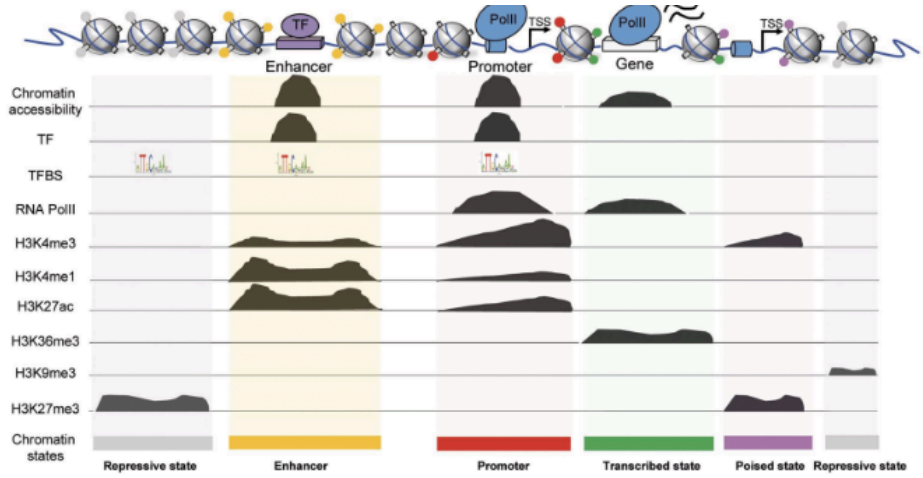
- **Process of data analysis:** Exploration, Choose informative genes, Choose and tune classification algorithm, Validation, Implementation
- **If there are two same mappings of the short reads to the genome sequence, how can we decide which section of the genome should it map to?** In mature mRNA, introns are spliced out of the pre-mRNA and are not at all present in the final protein structure. Because of this, some long reads might cover multiple exons from the

genomic DNA template, so it might be disjointed (for areas spanning over one cut-out intron). Thus, the mapping algorithm should be slightly modified to identify gene fusion.

- **What is gene fusion?** It is a specific kind of structural variant related to cancer, produced by somatic genome rearrangements. For chromosomal translocation, the exon from one gene might transfer to the other, causing a gene fusion in one of the chromosomes. In interstitial deletion, introns between two exons might get deleted so gene fusion occurs as well. Another way gene fusion can happen is through chromosomal inversion.
- **How do we detect gene fusion?** By using RNA-seq. Break points are in introns, and we need whole genome sequencing. Whole exome sequencing is not enough. Detecting fusion in RNA-seq requires much less sequencing than WGS, especially with long reads.

16.3 - Epigenome Analysis

- In epigenomic variant sequencing, we want to isolate DNA methylation, Chromatin modifications, DNase I hypersensitive sites and Transcription factors in order to monitor the sequencing protocols.
- **Process of data analysis:** Sample preparation of the epigenome through ChIP into DNA fragments, Sample sequencing into single-end reads, Read mapping, Downstream analysis, Peak calling for Motif and GO analyses.
- **Peak calling** allows us to compare the peak shape against a random background to show correlation.
- Nowadays, the entire pipeline is very tedious.
- **Histone marks and chromatin accessibility:**



- Using different protocols, we can detect different histones and chromatin accessabilities.

16.4 - Take-home Message of the Previous 2 Lectures:

- The variant-calling pipeline (Reasons for the steps, File interpretation, Factors affecting variant calling)

- GWAS (p-value correction)
- Gene fusion (Definition, How it can be detected by RNA-seq)
- Epigenomics (Gene expression regulation of structure and environment, Data analytics pipeline)