BMEG 3105 - Fall 2024

Lecture 15: Cancer genomics overview & genomics analysis

Lecture Date: 30 October 2024 Deadline: 6 November 11:59 p.m.

Lecturer: Prof. LI Yu

Scribe: Wong Kin Hang

Recap from last lecture

1. Underfitting
    a. Simple linear combination, the relationship among different variables within the image may be much more complicated than that
    b. High training error and test error
2. Overfitting
    a. Statistically: the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably
    b. Machine learning: the method is more complex than the problem, such that it can perform well on the training dataset but does not perform well on the testing dataset
    c. Low training error but high test error
3. Multi-omics
    a. Genome
    b. Epigenome
    c. Transcriptome
    d. Proteome
    e. Metabolome
    f. Phenome
4. Differential gene expression analysis
    a. Statistical analysis to discover quantitative changes in expression levels between experimental groups
    b. For a given gene, whether the gene expression difference is significant, other than due to natural random variation
5. T-test
    a. To check is there a significant difference between two sets of data
    b. General idea
        i. Calculate a test statistic based on the mean and variance of the data
        ii. Test statistic follows t-distribution
        iii. P-value: the probability that the result from the data occurred by chance

1. Along with test statistic, t-value
2. The smaller p-value is, the more confident we are

6. Gene enrichment analysis
    a. A biological pathway is a series of interactions among molecules in a cell that leads to a certain product or a change in a cell. Such a pathway can trigger the assembly of new molecules, such as a fat or protein. Pathways can also turn genes on and off, or spur a cell to move
        i. KEGG pathway database
        ii. Each pathway contains a set of genes
    b. 213 genes associated with type-II diabetes are identified by experiments
7. Testing association (How to identify pathways related with type-II diabetes?)

	In gene set	Not in gene set	Total
In pathway	(a)	(b)	
Not in pathway	(c)	(d)	
Total			

    a. If the pathway is related to type-II diabetes
        i. The number of genes (not) related to both should be high
        ii. The number of genes related to just one should be low
    b. If there are related
        i. (a), (d) should be large
        ii. (b), (c) should be small
8. Fisher's exact test
    a. A statistical significance test used in the analysis of contingency tables
    b. Suppose pathway and type-II diabetes are independent
        i. $p = \frac{[(a+b)!(c+d)!(a+c)!(b+d)!]}{a!b!c!d!(a+b+c+d)!}$
        ii. p-value is much easier to be calculated with computer

Cancer genomics overview

1. What is cancer?
    a. Cancer is a disease in which some of the body's cells grow uncontrollably and spread to other parts of the body
2. Why do we want to study cancer?
    a. Cancer was world's second leading cause of death in 2016
3. How do we study cancer?
    a. We will use genomics/multi-omics methods to study it
        i. Genome/Epigenome/Transcriptome/Proteome/Metabolome

Genome

Variant calling

1. Why do we care about variants?

    a. We can efficiently describe a genome with relation to a reference

    b. Genetic differences among people lead to differences in disease risk and response to treatment

    c. Genetic variation is used to find genes and variants that contribute to disease

    d. Cancers are genetic variants at multiple levels

2. Different types of genomic variants

    a. Short variant (<50bp)

    b. CNV

    c. SV

    d. PathSeq

3. How to discover the genetic variants?

    a. Library preparation

    b. Sequencing

4. Sequence mapping recap

    a. Slide each read along the genome, calculate the difference

5. Variants VS errors

    a. Must distinguish between actual variation (real change) and errors (artifacts)

    b. Errors can creep in on various levels:

        i. PCR artifacts (amplification of errors)

        ii. Sequencing (errors in base calling)

        iii. Alignment (misalignment, mis-gapped alignments)

        iv. Variant calling (low depth of coverage, few samples)

        v. Genotyping (poor annotation)

6. Data pre-processing step

    a. Mapping

        i. Map the reads produced by the sequence to the reference

ii. Mapping and alignment algorithms: BWA for DNA, STAR for RNAseq

iii. Input format: FASTQ

iv. Output format: Sequence/ Binary Alignment Map (SAM/ BAM)

v. CIGAR (Concise Idiosyncratic Gapped Alignment Report)

b. Marking duplicates

i. Duplicates: non-independent measurements of a sequence fragment

ii. Must be removed to assess support for alleles correctly

c. Base recalibration

GWAS

1. Genome-wide association studies
2. Try to determine whether specific variant(s) in many individuals can be associated with a trait (disease)
3. Bonferroni correction
    a. Adjusted p-value = p-value/ number of tests
    b. Suppose we have 1 million SNPs to test:
        i. Adjusted p-value = $\frac{0.05}{1000000}$
        ii. Adjusted p-value = 5 * 10-8

RNA-seq data analysis

1. Exploration
2. Choose informative genes
3. Choose and tune classification algorithm
4. Validation
5. Implementation

Gene fusion---structural variant

1. The first fusion gene was described in cancer cells in the early 1980s

2. Novel gene formed by fusion of two distinct wild type genes

3. In cancer: produced by somatic genome rearrangements

    a. Chromosomal Translocation

    b. Interstitial Deletion

    c. Chromosomal Inversion

4. RNA-seq for gene fusion detection

    a. Detecting fusion in RNA-seq requires much less sequencing than WGS, especially with long reads

5. Studying cancer at multiple levels

    a. Genetic variants

        i. Genome

        ii. Gene fusion (RNA-seq)

    b. Abnormal gene expression

        i. Genome (genetic information)

        ii. Epigenome (environment)

        iii. Transcriptome (direct measurement)

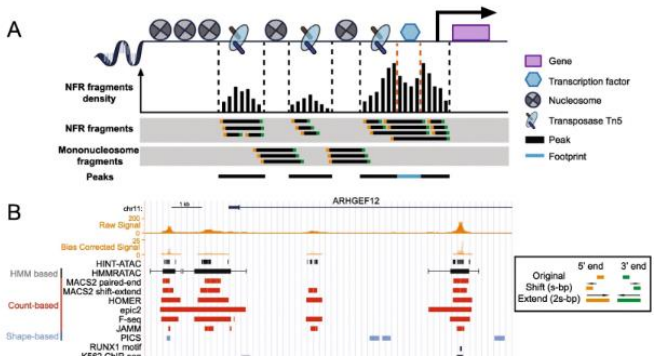
Epigenome

1. The overall data analytics pipeline for epigenomics

    a. Sample preparation and sequencing

    b. Computational analysis

Peak calling



[1]

Reference

[1] Li, Yu (2024). *BMEG3105: Data Analytics for Personalized Genomics and Precision Medicine - Cancer Genomics Overview & Genomics Analysis.*