CS-E5870 High-Throughput Bioinformatics

Exam, December 16, 2016

You are NOT allowed to use calculators or any other additional equipments/material in the exam. Please write your answers in English. Please write carefully. To help explain your answers better, you can also draw small diagrams/other pictures.

1. **Briefly** describe the following terms/concepts
   a) The power of a statistical hypothesis test (1 point)
   b) Genotype calling (1 point)
   c) Fastq quality score (also called Phred score) (1 point)
   d) Explain the reason why alignment is generally more challenging for human RNA-seq read data than for human ChIP-seq data. (1 point)
   e) DNA methylation (1 point)
   f) Describe your favourite term/concept you learned in this course (your chosen concept must be different from the ones listed above) (1 point)

2. Answer/Describe the following:
   a) List possible reasons why a mismatch can happen in an aligned sequence read. (2 points)
   b) Explain the RPKM quantification and normalization method for gene expression (assuming RNA-seq data). (2 points)
   c) Explain the following three concepts that are important for experimental design: replicates (technical and biological), blocking, and randomization. You can explain the concepts qualitatively, i.e., you do not need to formulate your answer in terms of mathematical equation. (2 points)

3. Answer/Describe the following:
   a) Describe the gene set enrichment analysis (GSEA) method. You can assume that you have a gene list ordered based on differential expression analysis and you also have a pre-defined gene set (e.g. genes belonging to a biological process, KEGG pathway, etc.). (3 points)
   b) Explain the TopHat method for aligning RNA-seq read data when you are given a reference genome but you do not have a reference transcriptome. (3 points)

4. Bisulfite sequencing (BS-seq) is the gold standard method for quantifying DNA methylation. Describe the experimental bisulfite treatment step in bisulfite sequencing (BS-seq) experiment. Also explain how BS-seq data can be aligned to a reference genome and how methylation level can be estimated for a single cytosine. (6 points)

5. A standard (state-of-the-art) approach to identify protein-DNA interactions for a selected protein is to carry out chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq). Describe the MACS method for identifying protein-DNA binding sites from ChIP-seq data, assuming a control input-DNA sequencing data is also available from the same biological sample. (6 points)