CS-E5870 High-Throughput Bioinformatics

Exam, November 21, 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) A perfect match probe in Affymetrix microarray (1 point)
   b) Genotype calling (1 point)
   c) Footprint in DNase-seq data (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 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 quantile normalization method between two or more microarray data sets. (2 points)
   c) Explain the RPKM quantification and normalization method for gene expression (assuming RNA-seq data). (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) Describe the statistically motivated DNASE2TF method for finding footprints from DNase-seq data. (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. The so-called multiple testing issue can severely impact most of the bioinformatics analysis. Explain the concept of multiple testing and explain the Bonferroni method (or any other method) for correcting statistical tests for multiple testing. Also, discuss the multiple testing correction in two different bioinformatics applications (e.g. detection of differentially expressed genes, detection of differentially methylated cytosines in DNA, detection of protein-DNA interaction sites, etc.). How severe the multiple testing problem is in these different applications (when compared to each other)? (6 points)