CS-E5875 High-Throughput Bioinformatics

Exam, February 21, 2020

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 (6 points in total)
   a) Significance level of a statistical hypothesis test (1 point)
   b) *De novo* genome assembly (1 point)
   c) Fastq quality score (also called Phred score) (1 point)
   d) DNA methylation (1 point)
   e) Somatic mutation (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. Describe the GATK's simple Bayesian genotype calling method for identifying single nucleotide polymorphisms (SNPs) from aligned DNA sequencing reads of a single individual. (6 points)

3. Answer the following: (6 points in total)
   a) Describe possible reasons why a mismatch can happen in an aligned sequence read. (3 points)
   b) Explain the RPKM quantification and normalization method for RNA-seq based gene expression data. (3 points)

4. 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)

5. The so-called multiple testing issue can severely impact most of the bioinformatics analysis. Explain the concepts of multiple testing, family-wise error rate (FWER), and false discovery rate (FDR). Explain also the Bonferroni method (or any other method) for correcting statistical tests for multiple testing. (6 points)