CS-E5875 High-Throughput Bioinformatics

Exam, December 11, 2023

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, the use of mathematical notation and equations is encouraged. The use of diagrams and drawings is also encouraged. You can get full points from the exam also by preparing "compact" answers (e.g. max. 1 page/question). Unnecessarily long answers (e.g. more than 2 or 3 pages/question, depending on writing style of course) may even negatively affect the evaluation.

Questions:

1. The so-called multiple testing issue can severely impact many types of bioinformatics analysis. (6 points in total)
   a) Explain the underlying reason for the multiple testing problem. (2 points)
   b) The Bonferroni multiple testing method controls the family-wise error rate (FWER). whereas the Benjamini-Hochberg multiple testing method controls the false discovery rate (FDR). Why is it typically more convenient to control FDR than FWER in bioinformatics applications? (1 point)
   c b) Explain the Benjamini-Hochberg method for correcting hypothesis tests for multiple testing. (3 points)
2. Briefly describe the DROP-seq single cell sequencing protocol: (6 points in total)
   - Describe the experimental aspects, including especially the cell barcode and the UMI barcode (4 points)
   - Describe how the cell and the UMI barcodes can be used to quantify single cell data at the level of individual cells and unique molecules. (2 points)
3. Briefly describe the supervised method called ACTINN for cell type identification from single-cell RNA sequencing (scRNA) data. (5 points)
4. Describe the GATK's simple Bayesian genotyper for calling single nucleotide polymorphisms (SNPs) from DNA sequencing data of a single individual. (5 points)