# CS-E5870 High-Throughput Bioinformatics

## Exam, December 18, 2017

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) Fastq quality score (also called Phred score) (1 point)
   c) Volcano plot (1 point)
   d) Somatic mutation (1 point)
   e) False discovery rate (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) Describe possible reasons why a mismatch can happen in an aligned sequence read. (2 points)
   b) Assuming you have aligned RNA-seq reads, explain the exon intersection method for gene expression quantification. (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 standard quality control methods for high-throughput sequencing data. In the lectures we discussed FastQC tool and covered at least 5 different types of quality measures. Try to remember and briefly describe at least four. (3 points)

4. Describe the beta-binomial generalized linear model method, called RadMeth, for identifying differentially methylated cytosines from bisulfite sequencing data. In addition to describing the statistical method, here are some discussion points that, among others, you can describe in your answer: why beta-binomial compound distribution is used instead of binomial (or any other) distribution, why linear model is included into the model, etc. You do not need to describe alignment of bisulphite sequencing read data, and you can ignore the part of the RadMeth method where evidence from nearby cytosines is merged. (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 (e.g., when compared to each other)? (6 points)