T-61.5050 High-throughput bioinformatics **Exam 21.05.2013, 09:00-12:00**
Dr. Elisabeth Georgii, Suleiman Khan

There are 6 questions (remember to turn the page).
You can reach 30 points in total, 5 points for each question.

Question 1 (5p): Definitions

Define the following concepts briefly *(1-2 sentences per concept)*:

a) Epigenomics (1p)

b) Metagenomics (1p)

c) Affymetrix probe set (1p)

d) Genome-wide association analysis (1p)

e) False discovery rate vs. false positive rate (1p)

Question 2 (5p): Sequencing, Part one

a) Describe the principle of the color space in the SOLiD technology. What is the rationale behind it? What is required for decoding? Which representation of reads is used in alignments and why? *(Write approximately 4 sentences.)* (2p)

b) Construct the suffix array and the Burrows-Wheeler transform for the string "GAGTAC$". Describe the construction procedure in 1-2 sentences. For which bioinformatics task is the Burrows-Wheeler transform used and what is the benefit from using it? (3p)

Question 3 (5p): Sequencing, Part two

a) What is re-sequencing? (1p)

b) Briefly define the following terms: contig, scaffold, coverage. (1p)

c) What are "de Bruijn graphs" and how are they used in the context of sequencing analysis? (3p)

Question 4 (5p): Transcriptomics

You would like to analyze differential expression between two subtypes of blood cancer. You have the opportunity to collaborate with a hospital where blood cancer patients are treated. Shortly describe the main steps you need to do in a microarray-based approach and the main steps you need to do in a RNA-seq-based approach. Summarize similarities and differences between the two approaches and discuss their advantages and disadvantages. *Write approximately 0.5 to 1 page (try to focus on the main aspects and use relevant keywords, rather than describing details).*

Questions 5 and 6 are on the next page!

Question 5 (5p): Unsupervised learning

a) How does unsupervised learning differ from supervised learning? (1p)

b) Given a gene expression dataset, describe a possible approach to obtain and visualize a clustering of genes. How do you evaluate the clusters of genes? What are advantages and possible limitations of your approach? *(Explain how you apply a certain method and what it yields; do not explain how the method works.)* (2p)

c) Your friend has computed a pairwise distance matrix between genes by integrating multiple data types in a sophisticated way. He asks you to do a clustering of genes based on this distance matrix. Name one clustering method that you can use for this task and one clustering method that you cannot use for this task! (1p)

d) What is the motivation for biclustering; how is a bicluster defined? (1p)

Question 6 (5p): Networks

a) How are co-expression networks constructed? What is the drawback of co-expression networks? (1p)

b) What is sparse linear regression? How is it used in gene network reconstruction? (1p)

c) Can one use the same gene network learning methods for RNA-seq data and microarray data? (1p)

d) What is a module? How can predicted modules be evaluated? (1p)

e) What is the idea behind the EM-algorithm? (1p)