UNIVERSITETET I OSLO
Det matematisk-naturvitenskapelige fakultet

Exam in INFBIO9121/INFBIO5121
High Throughput Sequencing technologies and bioinformatics analysis
Day of exam: 06.11.2015
Exam time: 10.00-12.00
The exam set consists of 2 pages
No attachments
Allowed materials: none
Teacher Lex Nederbragt can be reached on 480 28 722

Ensure that the exam set is complete before you start answering questions.

Please use separate sheets to answer each question.

Note: PhD students should answer all questions, while master students should not answer those marked as PhD students only. You need at least 50 points (MSc student) or 70 points (PhD student) to pass this written exam.

Question 1 - Assembly - 20 or 30 points

a) What is the conceptual difference between contigs and scaffolds? (5 p)
b) Given even, high coverage error-free reads, describe what can cause assemblies to become fragmented (5 p).
c) Explain why it is difficult to say, given a set of assemblies, which one is “best”. (10 p)
d) *PhD students only:* What is the conceptual difference between how the de Bruijn and Overlap (overlap layout consensus) graphs are used in the assembly process. (10 p)

Question 2 - Variant calling - 25 points

a) Describe at least 2 factors that introduce uncertainty into the variant calling process, and explain how and why they introduce uncertainty. (10 p)
b) What is re-alignment, and why is it done? (5 p)
c) Describe briefly, in general terms, what the Burrows-Wheeler transform is, where it is used in the field of high throughput sequencing, and why it is needed. (10 p)

Question 3 - RNAseq - 15 or 25 p

a) What is the definition of a trancriptome? How would you create a reference transcriptome for a non-model organism? (5 p)
b) Briefly describe the pros and cons of *ab initio* (reference based) and *de novo* transcriptome strategies. (10 p)
c) *PhD students only:* Briefly describe the main differences between DESeq and EdgeR. Briefly explain the underlying assumptions behind a Differential Gene Expression analysis when comparing two groups with replicates. (10p)

Question 4 - Statistical genomics and reproducibility - 10 or 20p

a) Briefly describe why reproducibility is important for computer-based analyses? Explain some of the main problems and how they may be overcome? (10 p)

b) *PhD students only:* Briefly explain the key components of Hypothesis testing. (10 p)