UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Exam in INF-BIO9121/INF-BIO5121 - High Throughput Sequencing technologies and bioinformatics analysis
Day of exam: 03.11.2024
Exam time: 9.00-11.00
The exam set consists of 2 pages
No attachments
Allowed materials: none

Teachers: Karin Lagesen (915 75 916), and Lex Nederbragt (480 28 722)

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

Please use separate sheets to answer each questions.

Note: You need at least 40 points (MSc student) or 60 points (PhD student) to pass this written exam.

1: Phred quality scores: (10p) (12 mins) - Tim5

a) Explain in words what a phred quality score is? (2p)
b) Provide the specific formula. (2p)
c) Give three examples of where it is used in high-throughput sequencing and variant calling (be brief). (2 point per example)

2: Trimming of reads: (15p) (18 mins) - Lex

a) Describe what read trimming is, and describe a way in which it can be done (5p)
b) Describe why and how trimming of low quality nucleotides in Illumina reads improves de novo genome assembly (5p)
c) Describe why and how trimming of low quality nucleotides in Illumina reads improves RNAseq analysis (5p) **- Rebekah**

3: Assembly and de Bruijn graphs (15p): (18 mins) - Lex

Although de Bruijn graph assembly was developed for capillary-based sequencing data, it did not reach prominence until the emergence of high-throughput short-read sequencing methods.

a) Give a brief explanation of what a de Bruijn graph is (5p)
b) Explain what particular advantages the de Bruijn graph approach has that made it popular for the assembly of early Solexa/Illumina short reads (5p)
c) Explain the major disadvantage(s) of the de Bruijn graph approach over other approaches (5p)

4: Read mapping : (15p) (12 mins) - Tim4a

a) Explain briefly what read mapping is (4p)

b) Describe how and why using paired end reads can increase the quality of the mapping results. (5p)5

c) We are given a fastq file which should contain reads from a DNA sample of one specific species. We are also given a draft assembly of the genome of that species. When mapping the reads to the draft sequence, it becomes apparent that a large fraction of the reads do not map to the draft assembly. Provide 3 causes for the unmapped reads. (6p)

5: Coverage and variant calling (10p): - Tim

a) What do we mean by "site or base coverage" in the context of variant calling? (3p)

b) Why is high coverage better than low coverage when performing variant calling? (4p)

c) Provide at least one concrete example of a case where higher coverage would enable you to make a correct inference about the state of a site in a sample, but low coverage would have lead to an error. (3p)

6: Statistical models and RNA-seq data (20p) - Rebeka/Monica

a) What is Poisson counting error? (4p)

b) How does the Poisson counting error affect low counts and high counts in RNAseq data? (4p)

c) List two ways you could alter your experimental design to reduce the impact of Poisson counting error. (4p)

d) How does the Negative binomial model differ from the standard Poisson model? (4p)

e) In the context of differential expression analysis, what types of variance does the standard Poisson model and the Negative Bionomial model explain? (4p)

7: Hypothesis testing: (15p) (18 minu4tes) - Boris

a) Briefly describe what hypothesis testing is (5p)

b) Briefly explain the key components of Hypothesis testing. (10p)