Exam IN-BIOS 5000 / 9000 - December 2022

Welcome to the exam in IN-BIOS 5000/9000, *Genome Sequencing Technologies, Assembly, Variant Calling and Statistical Genomics*, at the University of Oslo, December 2022

Date and time: Thursday 1 December 2022 from 17.00 to 19.00 (2 hours)

Permitted materials: None

The exam text is available in English only. Answers may be given in English or Norwegian.

The maximum number of points for each problem is indicated. The maximum number of points in total for all problems is 100.

Professor Torbjørn Rognes (phone 907 55 587) will be present at the exam room approximately 30 minutes after the start of the exam.

Problem 1 - True or false? (20p)

(Scoring: +1 for each correct answer, -1 for wrong answer, and 0 for no answer; minimum 0, maximum 20)

Indicate whether the following statements are true or false:

(a) BAM files are binary files representing mostly the same information as VCF files in a more compact form.

(b) BWA and Bowtie are tools for *de novo* genome assembly.

(c) Cross-validation is a technique used to estimate the performance of a classifier.

(d) FastQC is a tool to perform fast quality control of a genome assembly.

(e) Filtering called variants is only required in somatic variant calling.

(f) GATK and FreeBayes are tools for variant calling.

(g) Genome assembly is easier when the sequencing reads are shorter.

(h) If two genomic features A and B are associated, we can assume that A causes B.

(i) Illumina sequencing technology usually produces reads of length below 300 bp.

(j) In FASTA files each entry starts with an at sign (@).

(k) In genome assembly, contigs are formed from overlapping sequencing reads.

(l) Incorrect repeats in assemblies can mostly be eliminated by increased coverage during sequencing (deeper sequencing).

(m) Insertions and deletions are the most common types of sequencing errors with Illumina technology.

(n) Mutect2 is a tool to detect SNPs in the germline.

(o) Overfitting is a problem in machine learning when the model is too simple.

(p) Paired-end reads may be used to link contigs into scaffolds during genome assembly.

(q) Somatic variant calling typically involves comparing genetic variants found in tumor cells against germline cells.

(r) The base quality of Illumina reads is generally lowest in the 3' end.

(s) To perform clustering, it is necessary to have labeled data.

(t) Velvet is a genome assembly program based on the overlap-layout-consensus (OLC) approach.

Problem 2 - Sequencing (25p)

(a) Describe briefly the differences between the Illumina, PacBio and Nanopore sequencing technologies, their advantages and disadvantages, and for which main applications they are suitable. (10p)

(b) Raw reads from PacBio or Nanopore sequencing generally have a higher amount of errors than raw reads from Illumina sequencing. However, enhancements like PacBio HiFi reads, or Nanopore 2D or $1D^2$ reads are used to substantially decrease the error rate. Describe how these techniques work and how the sequence quality improvement is achieved. (10p)

(c) Describe the FASTQ file format in detail and give an example of an entry with a short sequence in this format. Explain how the characters in the last line of each entry are determined. (5p)

Problem 3 - *De novo* genome assembly (5p)

(a) The N_{50} is often used to measure assembly quality. Given eight small contigs of length 10, 3, 7, 15, 5, 40, 7, and 13. What is the N_{50} of this assembly? Explain how you found the value. (5p)

Problem 4 - Variant calling (20p)

(a) Name and describe two different types of genetic variation in a genome that involves more than just one or a few bases. (5p)

(b) An important early step in variant calling involves FASTQ files and a reference genome. What is done in this step? Mention a bioinformatics tool that can be used for this purpose. What kind of file do we end up with after this step? (5p)

(c) A few lines of a VCF file are shown below. Name and describe the types of variants illustrated on each line and explain the contents of the first 7 columns. (10p)

```
#fileformat=VCFv4.0
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  S1      S2
chr1    100     rs7     T       C       123     PASS    .       .       .       .
chr2    200     .       A       AC      73      .       .       .       .       .
chrX    300     .       GC      G       198     PASS    .       .       .       .
chrM    400     .       C       T,CT    999     PASS    .       .       .       .
```

Problem 5 - Statistical genomics (15p)

Assume we are interested in investigating whether a given set of SNPs occur within (inside) a given set of genes more than expected by "chance".

(a) Which track types can we use to represent the SNPs and genes? (5p)

(b) Describe a possible null model for this analysis. (5p)

(c) Describe a suitable test statistic for this analysis. (5p)

Problem 6 - Machine learning (15p)

(a) Given a dataset of DNA sequences each labelled as 0 or 1 indicating whether they include a particular binding site or not. Describe in general terms how you would approach training and evaluating a machine learning classifier that will predict whether new sequences contains a binding site. Assume the new sequences coming from the same distribution. You have two different machine learning methods you can choose from (for instance, ML1: k-mer frequency encoding with logistic regression, and ML2: one-hot encoding with a neural network). (10p)

(b) Describe the differences between parameters and hyper-parameters of an ML algorithm. (5p)