Examination in MVE510 - Introduction to bioinformatics
Time: Wednesday April 8 2020, 14.00-18.00
Examiner: Erik Kristiansson
On call: Erik Kristiansson, phone 070-5259751
Aids: All aids are allowed, you can use any software and retrieve information online. It is not allowed, in any way, to communicate with anyone other than exam guard and examiner. It is, furthermore, not allowed to communicate any information that can be accessed by another persons.

Maximum is 40 points. Preliminary, grade 3 requires 20 points, grade 4 requires 28 points and grade 5 requires 35 points.

1. Data from next-generation DNA sequencing typically contains a wide range of errors which needs to be removed or corrected. This process is called pre-processing.

   (a) Describe two ways of removing errors from short read data from the Illumina platform. Discuss also their strengths and weaknesses. (3p)

   (b) What are site specific errors and why are they problematic? Do the methods you mentioned in (a) address the issues associated with site specific errors? Why? Why not? (3p)

2. The Needleman-Wunsch algorithm is used to perform global alginments between two DNA sequences.

   (a) Explain the Needleman-Wunsch algorithm. What steps are necessary to identify the optimal global alignment between two sequences? Assume a linear gap penalty. (3p)

   (b) Demonstrate the algorithm by calculating the global alignment between

   ```
   Sequence 1: AGCTGATT
   Sequence 2: ACCGTT
   ```

   Use the scoring matrix

   $$S(a,b) = \begin{cases} 5 & \text{if } a = b \\ -3 & \text{otherwise} \end{cases}$$

   and a linear gap penalty with $d = -4$. Report all optimal alignments and their corresponding scores. (3p)

3. In a research project, whole-genome sequencing was done to identify single nucleotide polymorphisms (SNPs) associated with an inherited human disease.

   (a) Describe and discuss the main bioinformatical steps necessary to identify SNPs from sequence reads generated by short read next generation DNA sequencing. (3p)

   (b) Describe one statistical approach to identify SNPs and discuss its advantages and weaknesses. (4p)

4. In transcriptomics, RNA is converted to cDNA which is then sequenced. This results in reads that are randomly selected from the RNA molecules present in the samples.

   (a) Normalization is an important step in the analysis of transcriptomic data. Name two different normalization methods and describe their advantages and disadvantages. (3p)

   (b) Differentially expressed genes, i.e. genes that differ in expression between group of samples, are identified using statistical methods. Explain one method that is based on a normal assumption and one method that takes advantage of the discrete structure of RNA-seq data. Discuss the strengths and weaknesses of each method. (4p)

5. Assume that the nucleotide sequence AGTTCT$ represent a part of a bacterial gene (where $ is the termination character).

   (a) Calculate the Burrows-Wheeler transform (BWT) of the sequence. (3p)

   (b) Demonstrate how the original sequence can be reconstructed from BWT calculated in (a) (3p)

6. In shotgun metagenomics, DNA fragments are randomly selected from a microbial community and then sequenced. The sequencing is done with either short or long read sequencing techniques.

   (a) Do you see any advantages of using long reads instead of short reads? Do you see any disadvantages? (2p)

   (b) Would you analyze shotgun metagenomics data from short and long reads the same way? If yes, describe why. If no, specify how you would make the analysis different. (2p)

   (c) It is possible to generate both short and long read sequencing data from the same sample. Discuss how shotgun metagenomics can be improved by combining short and long read data. (4p)

GOOD LUCK!!