Examination in MVE510 - Introduction to bioinformatics
Time: Wednesday August 29 2018, 08.30-12.30
Examiner: Erik Kristiansson
On call: Erik Kristiansson, phone 070-5259751
Aids: None

Maximum is 40 points. Grade 3 requires 18 points, grade 4 requires 26 points and grade 5 requires 34 points.

1. Illumina is one of the most commonly used platforms for next-generation DNA sequencing.

   (a) Describe briefly how Illumina sequencing works. (3p)

   (b) Describe the characteristics and error patters of sequence reads generated by the Illumina platform. How does Illumina compare against traditional Sanger sequencing? (3p)

2. Use the Smith-Waterman algorithm to perform a local alignment between

   ```
   Sequence 1: AGGTTA
   Sequence 2: GGCTT
   ```

   Use the scoring matrix

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

   and a linear gap with the penalty $d = -4$. Report the best alignment and its corresponding score. (5p)

3. In a research project, scientists are investigating mutations in the human genome associated with a rare genetic disease.

   (a) The scientists want to achieve an average genome coverage of 10 times. The sequencing is performed on the Illumina platform using 100 bases long paired-end reads. How many reads and how many bases are necessary to achieve an average coverage of 10? (2p)

   (b) Describe the main steps in bioinformatics analysis of finding ('calling') mutations in a genome. (3p)

   (c) What is your view on a coverage of 10 times? Is it high or low? Do you see any risks with using a coverage of 10 times? (2p)

4. Normalization is the process used to remove systematic effects in transcriptomic data.

   (a) Describe the difference between 'within sample normalization' and 'between sample normalization'. What kind of systematic effects do they remove? (3p)

   (b) Describe how the normalization method RPKM ('read per kilobase per million') work. (2p)

   (c) Name one advantage of using median normalization instead of RPKM. (1p)

5. Studies in transcriptomics, metagenomics and other areas of genomics are often based on a large number of statistical tests. This can give rise to false positives.

   (a) Describe the concept of false discovery rate (FDR) and explain how it can be used to control for the number of false positives. What does a FDR cut-off of 0.01 mean? (3p)

   (b) In a transcriptomics study, 20,000 genes were measured in a number of samples and assessed for differential expression. The top 10 most significant genes can be seen in the table below. Use the Benjamini-Hochberg algorithm to calculate the FDR for the p-values below. How many of the tests are significant if the FDR significance cut-off is set to 0.01? (2p)

Gene	P-value
Gene 7321	6.5×10^{-10}
Gene 6236	2.3×10^{-9}
Gene 18345	7.8×10^{-9}
Gene 64	1.1×10^{-8}
Gene 1098	9.3×10^{-7}
Gene 14037	4.5×10^{-6}
Gene 2254	2.3×10^{-6}
Gene 845	1.2×10^{-5}
Gene 16346	5.6×10^{-5}
Gene 1658	8.5×10^{-5}

6. Gene set enrichment analysis is a bioinformatical concept to improve the biological interpretation of gene lists generated in transcriptomics and metagenomics.

   (a) Explain the concept of integrative analysis. What is require to perform integrative analysis and what insights can it provide? (2p)

   (b) Assume that you have performed a transcriptomic analysis of 20,000 genes of which 1,000 were found to be significant (had a false discovery rate less than 0.05). Assume further that you are interested to assess the overrepresentation of the Gene Ontology-term 'response to stress' (GO:0006950). Of the 20,000 measured genes, 750 are annotated with the 'response to stress' GO-term, of which 250 were significant. Describe how you can use a hypergeometric test (also known as the Fisher's exact test) to statistically assess whether 'response to stress' is overrepresented among the significant genes. Note: You do not need to calculate the p-value, only describe how you do it. (3p)

7. In metagenomics two main techniques are used: amplicon sequenicing of marker genes and shotgun sequencing of random fragments DNA.

   (a) How does amplicon sequencing work? Describe the process, the necessary bioinformatics analysis and what biological insights amplicon sequencing can provide (3p)

   (b) Relate amplicon sequencing to shotgun sequencing. What kind of information can you extract from data generated by shotgun sequencing that is not present in data generated by amplicon sequencing? (3p)

GOOD LUCK!!