2018 written exam in IN-BIOS9000/IN-BIOS5000

Genome Sequencing Technologies, Assembly, Variant Calling and Statistical Genomics

Day of exam: 30.11.2018

Exam time: 14.30-16.30

The exam set consists of 2 pages

No attachments

Allowed materials: none

Teachers: Arvind Sundaram (90873284)

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.

Questions

High throughput sequencing (10 points)

1. Which sequencing platform(s) is(are) best suited for generating de novo genome assemblies? Explain how sequencing platforms influence the final draft assembly. **(5 pts)**

2. Describe why pre-processing of sequenced reads (ex. Illumina) is important. **(5 pts)**

De novo assembly (30 points)

1. Regarding measuring assembly quality:
   a. Given five contigs of length 832,368,173,98 and 29, what is the N50 for this assembly? **(3 pts)**
   b. Give one reason why the N50 is not a reliable quality measure. **(3 pts)**

2. Explain how reads are used when building an assembly graph using the Overlap-Layout-Consensus method, vs de Bruijn based assembly methods. **(6 pts)**

3. Let's say your read set consists of these two reads: TGACTGA and ATGACCT. We are using de Bruijn based assembly methods.
   a. Draw the de Bruijn graph for these reads using a k-mer size of 4. **(8 pts)**
   b. Can you, by examining the reads, explain why a k-mer size of 3 would be a bad idea for these reads? Do not draw the graph for this k-mer size. **(4 pts)**

4. Long read data usually creates more contiguous assemblies than short read data, despite usually having a higher error rate
   a. Mention why long read data often gives longer contigs **(2 pts)**

b. Describe one method of how we can error-correct long read data to get more correct assemblies **(4 pts)**

Variant calling - including mapping and alignment (30 points)

1. Mapping and alignment:
   a. What is the conceptual difference between mapping and alignment? **(3 pts)**
   b. What is the name of the algorithm that is central to mapping sequence reads to a reference, and why is it needed? **(3 pts)**
   c. Name one of the sequence alignment algorithms we studied and explain the two main "phases/parts" of the algorithm. **(3 pts)**

2. Bases in a fastq file have a base quality:
   a. What is the base quality a measure of? **(2 pts)**
   b. Why do all bases not have the same quality? **(2 pts)**
   c. How does base quality influence the variant calling process (you may illustrate with an example)? **(4 pts)**

3. Variant calling
   a. What is the conceptual difference between a variant site and a genotype? **(3 pts)**
   b. Describe 2 factors (other than sequencing base quality) that introduce uncertainty into the variant calling process and explain how and why they introduce uncertainty. **(10 pts in total; 5 pts for each factor)**

Statistical genomics (30 points)

1. Question 1
   a. There are many different track types (ways to represent genomic tracks). Mention two such track types and for each of them give an example of a genomic/epigenomic data set that can be represented by that track type. **(5 pts)**
   b. Assume we are interested in measuring the co-localization of two genomic tracks (to what degree elements in the track occur at the same locations in the genome). Suggest a possible test statistic that can be used to measure this. **(5 pts)**
   c. An example of a descriptive statistic is "number of elements in a genomic track". Mention two other examples of descriptive statistics for genomic tracks and mention one reason why it might be useful to compute descriptive statistics before doing an analysis. **(7 pts)**

2. Question 2
   a. Explain how we can use Monte Carlo simulation to investigate whether two genomic tracks cover the same locations in the genome more than expected by chance. **(7 pts)**
   b. When making a randomized version of a genomic track, why do we typically want to preserve some of the structure of the original track? **(6 pts)**