Bioinformatik für Biologen SS 2019

Complementary Homework 11

This homework will not be collected or graded, but its contents and the papers to read may be relevant in an upcoming lecture, test or assignment.

1. Describe the two phylogenomic approaches you have heard about to reconstruct species-trees. Give reasons why it is preferable to use more than one gene/protein to reconstruct species-trees.

2. Explain in your own words, how Sanger sequencing, Illumina sequencing, Ion Torrent, PacBio, and Nanopore sequencing works.
Why does a capillary sequencer does not need four capillaries but just one compared to four lanes on a gel in the past?

3. Read and understand the following paper:
T.C. Glenn (2011) Field guide to next-generation DNA sequencers.
Mol. Ecol. Res. 11, 759-769.
DOI: 10.1111/j.1755-0998.2011.03024.x

The numbers for the sequencing technologies from this article are updated regularly. So please use these values below:
http://www.molecularecologist.com/next-gen-fieldguide-2016/

The following has even more updated numbers and also describes some additional and upcoming technologies:

S. Goodwin, J.D. McPherson and W.R. McCombie (2016) Coming of age: ten years of nextgeneration sequencing technologies.
Nat. Rev. Genet. 17, 333-351.
DOI: 10.1038/nrg.2016.49

Answer the following questions

4. What are the main points from the lecture and the Glenn (2011) paper distinguishing 1st, 2nd and 3rd generation sequencing?

5. (a) What are paired end reads?
(b) How are long-jump libraries constructed? When sequencing a long-jump library with paired-end sequencing, how are the read pairs oriented on the genomic sequence? Compare this to 'normal' paired-end reads and describe why they are the same or different.

6. Extract the below information about the typical sequencers of the following types (a) capillary Sanger sequencer, (b) Ion Torrent PGM/S5, (c) Illumina MiSeq v3/HiSeq400, (d) PacBio RS II/Sequel and (e) Oxford Nanopore MinION/PromethION discussed in the above papers:

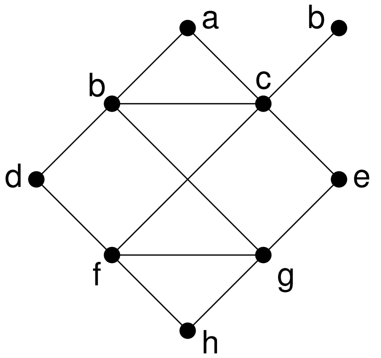
- How many reads does the machine produce in a single run?
- How long are these reads?
- How long does a single run take?
- How many bases does one get from a single run?
- Can the sequencer produce paired-end reads?
- What is the primary rate and type of error?

7. Assume the 5 sequencing methods introduced during the lecture (Sanger capillary sequencing, Illumina, Ion Torrent, PacBio, Nanopore sequencing). Review the sequencing process. Assume 12 cycles of sequencing (i.e. adding the chemistry like nucleotides, etc for the elongation by the polymerase up to a possible subsequent washing step and measuring the signal). What are the possible lengths of reads obtained within the 12 cycles? State the minimum and maximum lengths for each method, or whether the method does not have such cycles.

8. Define what the terms read, contig and scaffold describe.
Explain briefly how one constructs contigs from reads and scaffolds from reads in the case of de novo assembly following the overlap-layout-consensus (OLC) approach (like with CAP3).
What information can be used to guide the construction of contigs from reads and scaffolds from contigs, respectively?

How can the optical mapping technology mentioned in the Goodwin et al. (2016) also help to order scaffolds according to a reference. How does it differ from FISH mentioned in the lecture?

1. Analyze the following Graph:



- does an Euler Path exist for this graph?
- does an Euler Cycle exist for this graph?
- write down the sequence of nodes for an Euler Path and/or Euler Cycle, if it exists, respectively.
- write down degrees of the nodes.

Add-on: Does an Euler Path and/or Euler Cycle exist for the Königsberg-Brigde-Problem (i.e. the original map of the city)?

2. Two research groups study flies with a genome size of about 150mio nucleotides and both had the genome sequenced.
One group got sequencing data from a PacBio machine that produced 60000 reads with average length 10000nt.
The other group obtained sequencing data from an Illumina HiSeq machine. This sequencer sequenced 500mio fragments and for each fragment it obtained two paired-end reads of length 100nt each.

- compute the coverage for each of the sequencing data.
- discuss briefly which are the advantages and disadvantages of each dataset for the assembly process**.**
- which obstacles can be overcome if a reference genome exists for the assembly of the single-end data?

3. You have sequenced 16 times the read ACGTATCGA and once ACGTTTCGA.

- generate a de Bruijn Graph for the reads with k=4.
- what structure does that graph contain?
- what would happen to that structure during the subsequent steps in the assembly, and why?
- how do you distinguish ‘real’ paths in a de Bruijn Graph from those caused by sequencing errors?

4. You are analyzing a hypothetical genome with 7 regions:
region1-regionR-region2-regionR-region3-regionR-region4
regions with the same name are exact repeats.
- draw the sketch of an condensed de-Bruijn-Graph which would result from this genome.
- in how many separate contigs would one reconstruct?
- how many paths through the graph exist to resolve the genomic structure?
- are there possible ways to figure out which path is the correct one?
- is there a way to infer how often a repeat existed in the originally sequenced template?

5. You have done a de-novo assembly and you have got 14 contigs with the following lengths:
100, 1000, 200, 2000, 2500, 300, 3000,
500, 4000, 600, 700, 800, 8000, 900
What is the N50 value?
What is the difference to an NG50 value?