ITO Integrative Transcriptomics
Prof. K. Nieselt,
Institute for Bioinformatics and Medical Informatics Tübingen
Prof. S. Nahnsen,
Institute for Bioinformatics and Medical Informatics Tübingen

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Lecture: Grundlagen der Bioinformatik — SoSe 2022

Assignment 7 (20 points)

Hand out: Thursday, June 16
Hand in due : Thursday, June 23, 18:00
Direct inquiries via the ILIAS forum or to your respective tutor at:
Mathias Witte Paz: iizwi01@uni-tuebingen.de
theresa-anisja.harbig@uni-tuebingen.de
meret.haeusler@student.uni-tuebingen.de
jules.kreuer@student.uni-tuebingen.de
simon.heumos@qbic.uni-tuebingen.de

Theoretical Assignments

1. **Assembly of the largest eukaryotic genome** (3P)

   Identify the tool used for the reconstruction of the largest sequenced eukaryotic genome until date. Do not forget to correctly cite the tool! Furthermore, answer the following questions:

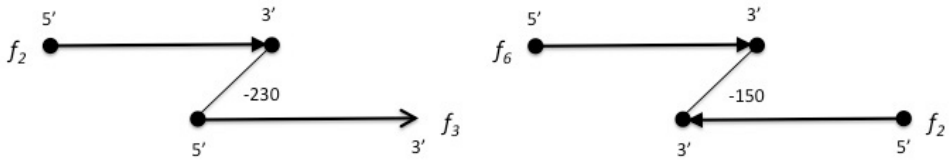
   (a) Besides its size, what is the most striking feature of this genome?
   (b) What is the assembly approach followed by the tool used to reconstruct this genome?
   (c) What other large eukaryotic genome was assembled using this tool?

2. **Greedy assembly with OLC** (5P)

   Given reads $\mathcal{F} = \{f_1, f_2, \ldots, f_6\}$, each of approximate length 500, that overlap as follows:

   (a) $o(5'f_3 - 3'f_2) = 230$
   (b) $o(3'f_1 - 3'f_2) = 150$;
   (c) $o(5'f_5 - 3'f_4) = 125$;
   (d) $o(3'f_1 - 3'f_4) = 100$;
   (e) $o(5'f_4 - 5'f_6) = 400$;
   (f) $o(3'f_1 - 3'f_3) = 375$;
   (g) $o(5'f_5 - 5'f_6) = 50$;
   (h) $o(3'f_6 - 5'f_1) = 220$;
   (i) $o(3'f_6 - 3'f_3) = 110$;

   where e.g. $o(5'f_3 - 3'f_2) = 230$ means that the $5'$-end of f_3 overlaps the $3'$-end of f_2 by 230 bases and $o(3'f_6 - 3'f_2) = 150$ means that the $3'$-ends of f_6 and f_2 overlap by 150 bases (see below).



(a) **The overlap graph:** Draw the overlap graph OG, labeling all edges as discussed in the lecture.

(b) **A minimal spanning tree:** Draw a minimal spanning tree in OG, containing all read edges.

(c) **The layout:** Draw the layout of the reads as given by the minimal spanning tree, indicating the approximate coordinates of the start and end of each read. What is the length of the final assembly?

(d) **Consistent overlaps:** Are all overlaps consistent with the computed layout? If not, which overlaps are not consistent with the layout, and why?

Hand in the overlap graph, the resulting layout as well as the consensus sequence, you can hand in a handwritten and **scanned** solution. **Note:** If you choose to hand in a handwritten solution please provide clear and legible solutions. If you do not have the possibility of scanning, you can use mobile apps that return good quality scans, instead of only including a picture of your notes.

Practical Assignments

3. Contigs versus coverage (5P)

When using the shotgun method, short (partial) sequences (reads) randomly distributed over the genome are determined in the sequencing phase. They are combined into larger contigs in the assembly phase. The following equation states the relationship between the number of reads and the expected number of contigs.

For this let

G	=	genome length in base pairs
L	=	mean length of a read
N	=	number of sequenced reads
c	=	LN/G = expected coverage
T	=	length of the overlap between two reads
θ	=	T/L, the proportion that two reads must overlap to be concatenated

Then according to the Lander-Waterman model (1988) the expected number of contigs is

$$Ne^{-c(1-\theta)}$$

For each of the five different theta, $\theta \in \{0.2, 0.3, 0.4, 0.5, 0.6\}$, generate a plot of this function against the coverage $0 \leq c \leq 20$. Scale the y-axis in units G/L, for the expected number of contigs to be independent of the genome size itself.

Explain the resulting plot. You do not need to explain the equation or the Lander-Waterman model.

Citation: Lander ES, Waterman MS (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics. 2 (3): 231–239. doi:10.1016/0888-7543(88)90007-9

4. Sequencing Read Quality Control (7P)

We have simulated 3 sets of short paired-end reads (see material folder) for the genome of the pathogen *Mycobacterium leprae* (NCBI accession code: NC_002677) using the tool ART[1]. Your task is to perform a quality control on the reads and compare the results of the different sequencing runs. For this:

[1] `https://www.niehs.nih.gov/research/resources/software/biostatistics/art/index.cfm`

(a) Shortly explain the term paired-end sequencing and its difference to a single-end sequencing approach.

(b) Run FastQC[2] on the six compressed FastQ files. **Hint:** You do not need to gunzip the files to run FastQC.

(c) Summarize all FastQC results into one common HTML report using MultiQC[3]. Hand in the created report as an additional file.

(d) Find out the length of the input genome and compute the mean coverage for each sequencing run. Remember: we have reads from paired-end sequencing approach!

(e) Discuss the quality of the three different paired-end sequencing experiments. Include high-quality and meaningful figures from the MultiQC report in your discussion. In your text, make sure that you correctly refer to the included figures. How certain would be the results of an assembly for each of the different sequencing runs?

Please read the questions carefully. If there are any questions, you may ask them during the tutorial session or in the forum of ILIAS. You will usually get an answer in time, but late e-mails (e.g. the evening of the hand-in) might not be answered in time. Please upload all your solutions to ILIAS. Don't forget to put your names on every sheet **and** in your source code files. Please pack both your source code as well as the theoretical part into one single archive file and give it a name using this scheme: `<name1>_<name2>_<Assignment>_<#>.zip`. The program should run without any modification needed.

[2] `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`

[3] `https://multiqc.info`