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 15
Hand in due : Thursday, June 22, 18:00
Direct inquiries via the ILIAS forum or to your respective tutor at:
caroline.jachmann@uni-tuebingen.de
meret.haeusler@student.uni-tuebingen.de
simon.hackl@uni-tuebingen.de
carolin.schwitalla@qbic.uni-tuebingen.de

Theoretical Assignments

1. **Branching structure of genealogies** (6P)

   There are two different unlabelled coalescent trees on 4 haplotypes:

   (a) Sketch them and work out the respective probabilities of their occurrence.
   (b) How about labelled coalescent trees of 4 haplotypes? Can you draw all of them?
   (c) Explain, why are there more labelled coalescent trees on 4 haplotypes than rooted phylogenetic trees on 4 taxa. Remember there are 15 rooted phylogenetic trees on 4 taxa

2. **Sequencing technologies** (4P)

   Describe the approach used by ULTIMA sequencing.

   Write 200 words. Remember to cite your sources.

Practical Assignments

3. **Sequencing Read Quality Control** (10P)

   In this exercise you will simulate 3 sets of reads, perform a quality control on the reads, and compare the results of the different simulations.

   (a) Make yourself familiar with the ART[1] tool for read simulation. Based on the provided fasta file, generate 3 sets of paired-end reads that have length of 150bp and are generated with an Illumina HiSeq 2500. To run the tool properly you need to provide the uncompressed fasta file as input. For the remainder of the tasks you will only need the FastQ files. Change the coverage and quality for each run:

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

Run 1: Coverage: 30X, shift the quality score by -20 for the first read and -35 for the second read.

Run 2: Coverage: 20X, shift the quality score by 7 for the first read and 5 for the second read.

Run 3: Coverage: 10X, shift the quality score by 2 for the first read and 3 for the second read.

In addition, set `-m 300 -s 20`.

(b) Compress your fastq files with bgzip[2]. Hand-in the compressed files. Also, report the full command you used to generate each read set, as well as explain each parameter.

(c) Investigate the quality of your reads using FastQC[3] on each file.

(d) Summarize all FastQC results into one common HTML report using MultiQC[4]. Hand in the created HTML report as an additional file.

(e) Discuss the quality of the three different paired-end sequencing experiments. What is the impact of the quality scores of each run? Include 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?
Write about 350-400 words.

Hint: Software installation can be simplified by using conda or docker. However, it is not required for you to use this.

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] `http://www.htslib.org/doc/bgzip.html`

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

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