Prof. Dr. Daniel Huson
Zentrum für Bioinformatik
Fachbereich Informatik
Mathematisch-Naturwissenschaftliche Fakultät

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Grundlagen der Bioinformatik **SoSe 2018**

Assignment 07 **Submit electronically in Ilias by 18.6.2018, 10h**

1 Illumina sequencing instruments (3 points)

List and briefly describe three different sequencing instruments currently offered by Illumina Inc. What are the typical run-times and output characteristics? What is an intended application domain for each?

For each, name a published paper that uses the instrument and briefly describe what was sequenced and why.

2 Long read sequencing (2 points)

List and briefly describe the sequencing devices currently offered by PacBio and Oxford Nanopore. What are the typical run-times and output characteristics? What is an intended application domain for each?

For each, name a published paper that uses the instrument and briefly describe what was sequenced and why.

3 Mystery dataset (5 points)

Download the file `reads-5000.fasta` from Ilias. This file contains 5,000 sequencing reads.

Using your own code or existing tools or web resources (please state exactly which approach you used in each case), do the following.

First, summarize this data: What is the distribution of read lengths? What is the frequency of all four nucleotides in these reads, that is, what is the GC content?

Second, all reads were obtained from an environmental sample and are presumed to come from the same prokaryotic organism. Determine which organism these reads were obtained from using BLASTN on NCBI. Based on the identity of the organism, from which environment do you think the sample came from?

Third, based on some of the highest scoring alignments, estimate the sequencing error rate, under the assumption that the reference genome for such a highest-scoring alignment is the true "source genome" for the read).

Finally, which sequencing technology do you think produced these reads (and why do you think so?)