
DTU

DTU Health Technology
Bioinformatics

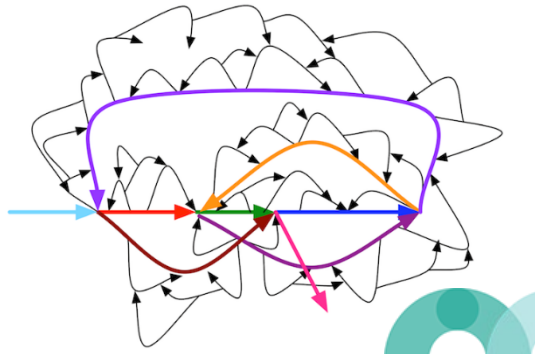
Week 6 Recap

Gisle Vestergaard
Associate Professor
Section of Bioinformatics
Technical University of Denmark
gisves@dtu.dk

Last Week

- Quiz I or Deliverable V
- Sequence alignment
- 16s rRNA amplicon analysis
- Quantitative metagenomics
- Exercises…

Today



- Lessons from Quiz I
- Quiz II
- Metagenomics *de novo* assembly
- Talk by Henrik Bjørn from Clinical Microbiomics
- More exercises

clinical microbiomics

Last weeks quiz

- How many lines is in a fastq file? What does each of the four lines contain? (2 point)

Four lines...I literally gave you the answer in the next line

Line 1: @sequence identifier

Line2: raw sequence

Line3: + (seldomly also the sequence identifier

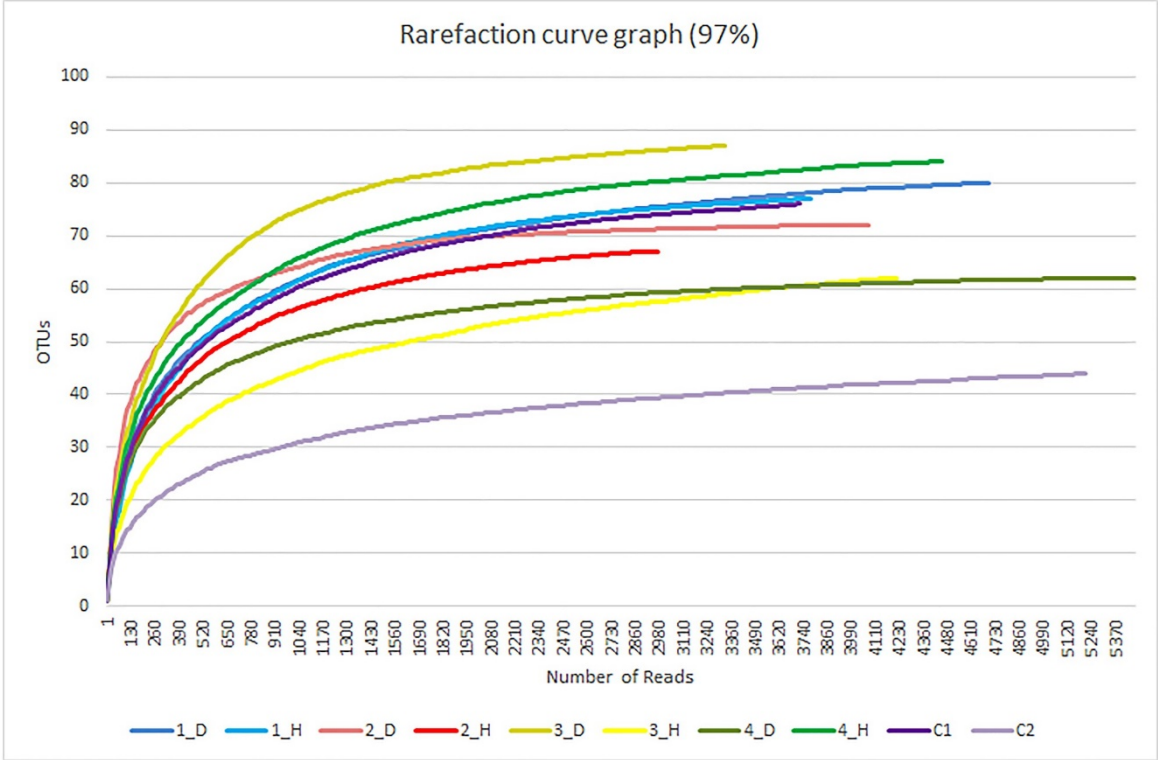
Line4: Sequence quality score. Must (obviously) contain the same number of scores as letters in the raw sequence

1. What is the difference between the SAM and BAM file format standard? Which one should we generally use and why?

- Sequence Alignment Map (SAM) is a text-based format originally for storing biological sequences aligned to a reference sequence. The binary equivalent of a SAM file is a Binary Alignment Map (BAM) file, which stores the same data in a compressed binary representation.
- We should use the BAM file format when possible.
- Often the SAM/BAM files contain a lot of information and end up as quite large files. Since the BAM file is compressed it takes up less storage.

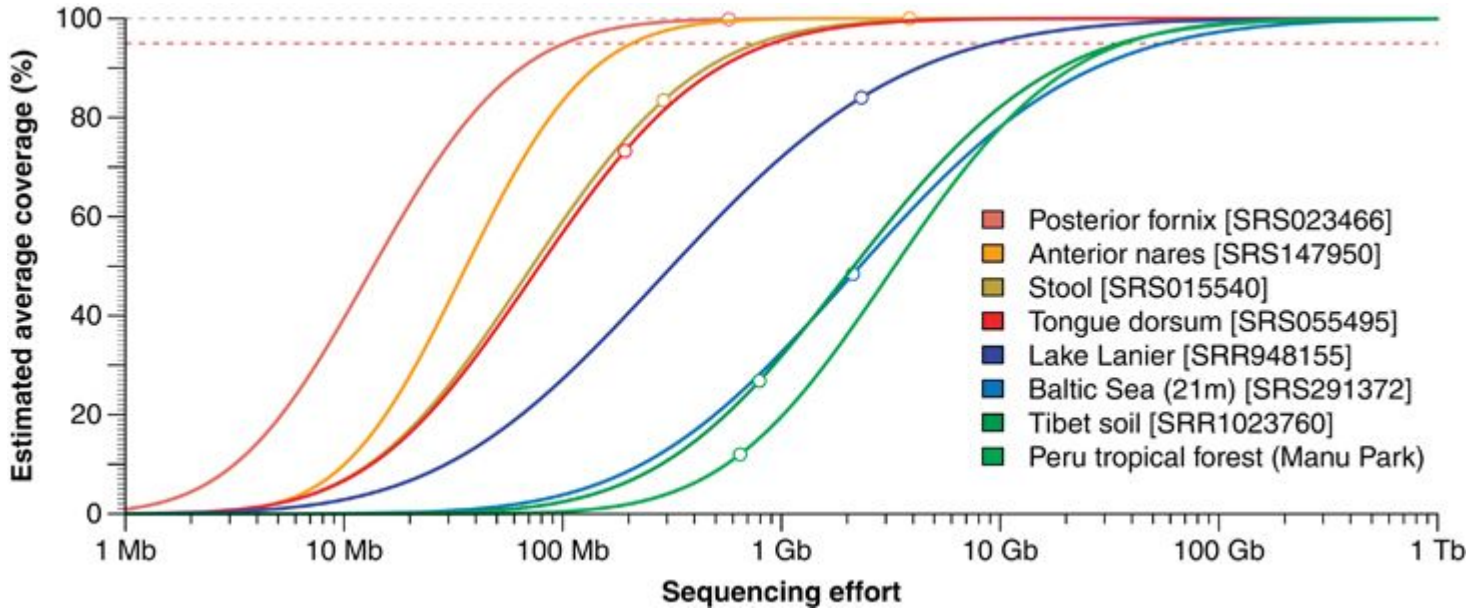
5. How does one check a 16s rRNA amplicon sample for adequate sequencing depth?

- Rarefaction plot



6. How does one check a shotgun metagenome sample for sequencing depth?

- Nonpareil curves. Nonpareil uses the redundancy of the reads in a metagenomic dataset to estimate the average coverage and predict the ammount of sequences that will be required to achieve "nearly complete coverage".



Why should we check for sequencing depth in a metagenomic study?

- We can see how much sequence is needed to describe an entire microbiome, thus avoiding over-sequencing. We can also give an honest estimate for how descriptive our dataset really is.