DTU

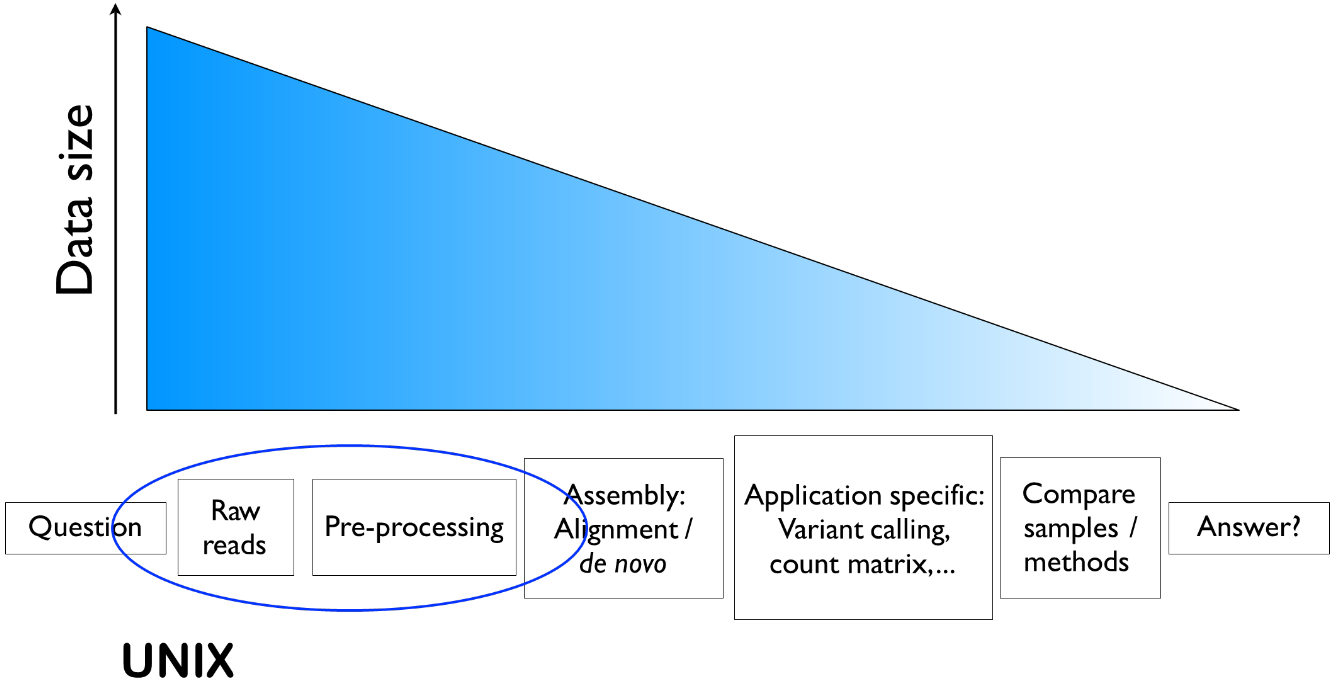
DTU Health Technology
Bioinformatics

Week 5 Recap

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

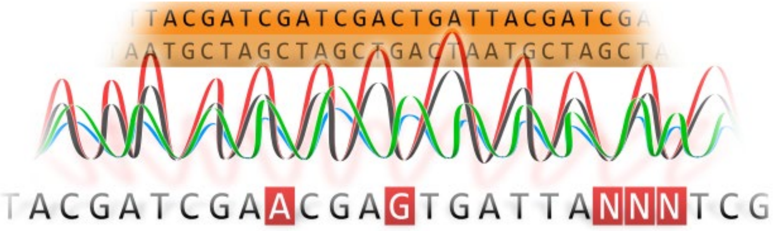
Last Week

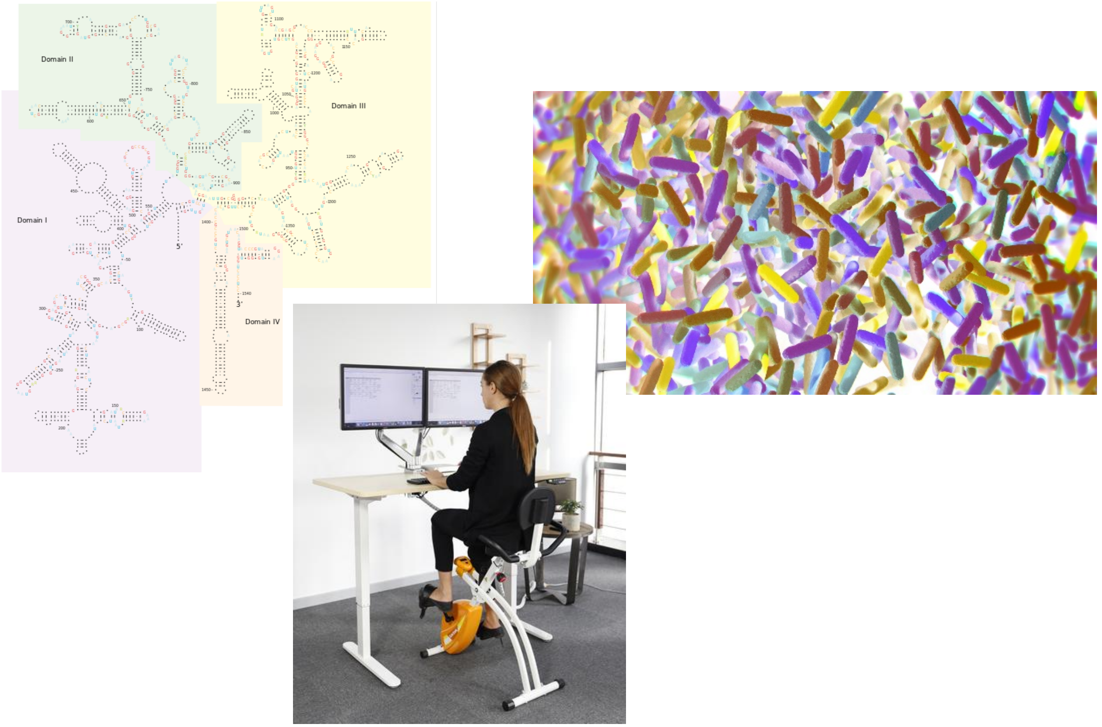
- Getting on the cloud
- Sequencing technology
- Raw sequencing reads
- Preprocessing reads


Data size
Question
Raw reads
Pre-processing
Assembly: Alignment / *de novo*
Application specific: Variant calling, count matrix,…
Compare samples / methods
Answer?
UNIX

Today

- Last Wednesday
- NGS Quiz in Discord groups
- Sequence alignment
- Presentations
- 16s rRNA Amplicon sequencing
- Metagenomics and microbial ecology





Please check webpage for times!

Technological advances continues to advance metagenomic possibilities

- New Technologies means new possibilities
- …also means new types of errors
- Illumina is the current workhorse
  - Great for many applications
- Long read technology
  - Adding information
  - Resolves difficult regions during genome assembly

Reminder: things ~~can~~ will go wrong

```
template
read
```

mismatch

```
AGCAATCTCAATTACAAATATACACCAACAAA
AGCAATCTCAATTACAGATATACACCAACAAA
```

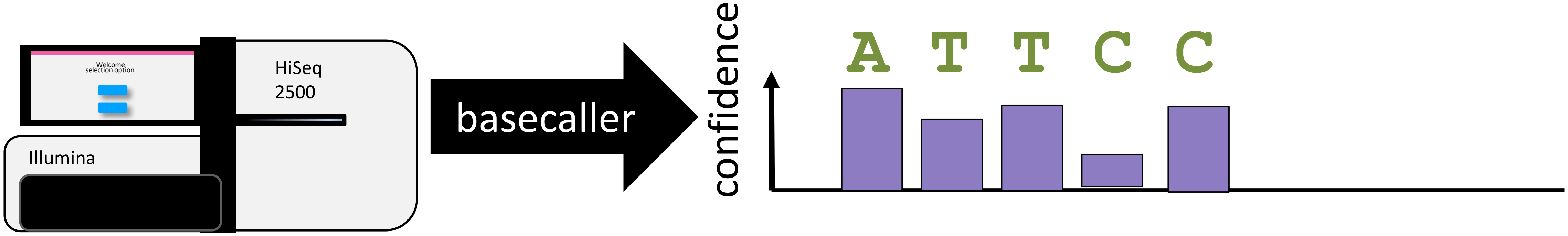
insertion

```
AGCAATCTCAATTACA-AATATACACCAACAA
AGCAATCTCAATTACACGATATACACCAACAA
```

deletion

```
AGCAATCTCAATTACAAATATACACCAACAAA
AGCAATCTCAATTACA-ATATACACCAACAAA
```

Basecalling

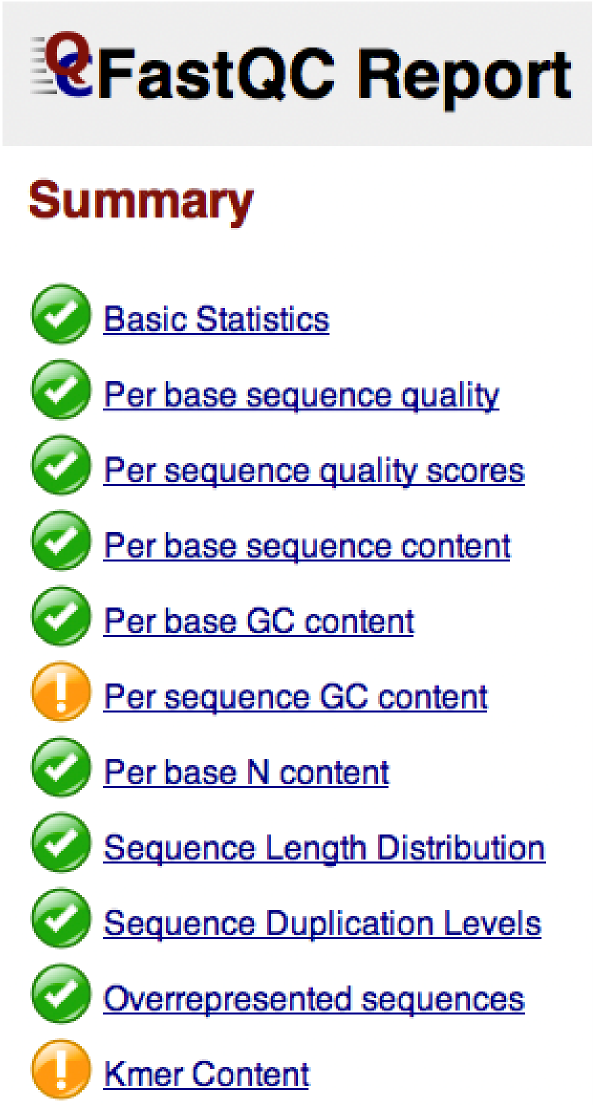


```
@ILLUMINA-C90280_0030_FC:5:1:2675:1090#NNNNNN/1
A    T    T    C    C    C    G    ...
```

```
0.6837722   0.9369043  0.9920567   0.9996019 0.9997488  0.9993690  0.9999000 ...
```

Fastqc reports

- Report basic statistics on your data
- Identify issues with your data
- **Use at each step of preprocessing to check progress**



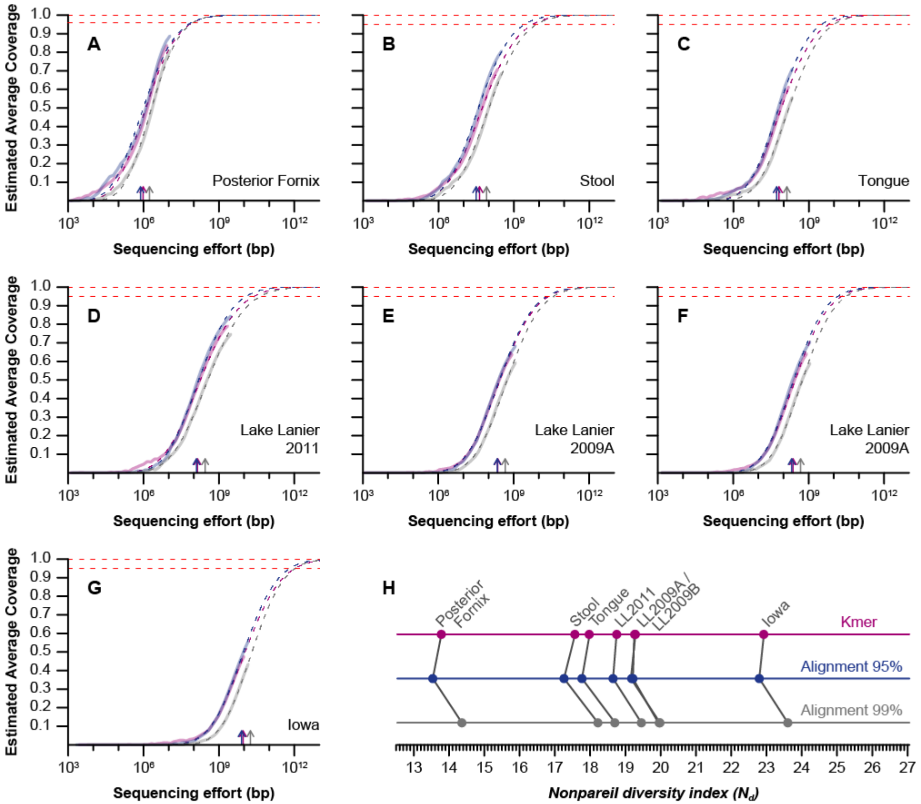
FastQC Report

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ✓ Per base GC content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ! Kmer Content

Sequencing depth for shotgun metagenomics

- No reference database like 16s, therefore we cannot use rarefaction
- Nonpareil: How often do I find the same read in a dataset?



QUIZ TIME!