DTU

DTU Health Technology
Bioinformatics

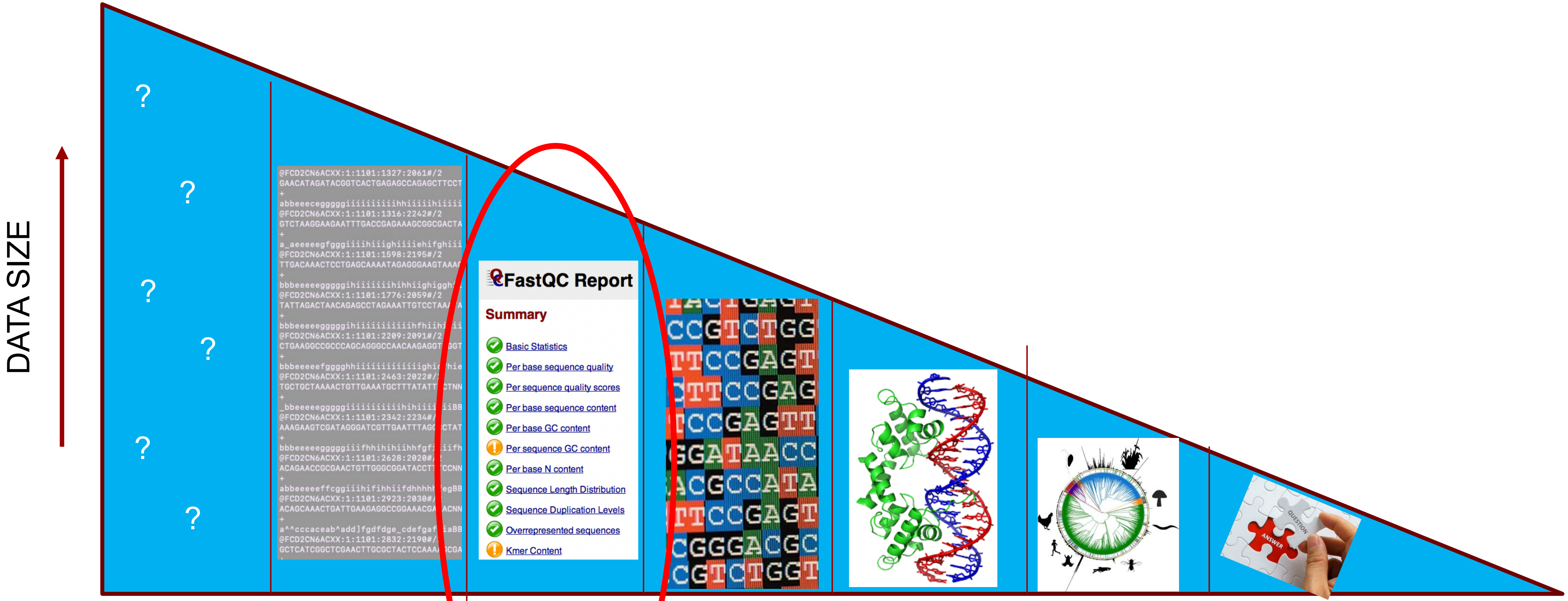
Data Preprocessing

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

Menu

- The main steps in NGS analysis
- Why is preprocessing important?
- Preprocessing
  - Fastqc reports
  - Adapters
  - K-mers
  - Depth of coverage vs Breadth of coverage
  - Merge paired end reads
  - Ion Torrent data
- Exercises

Generalized NGS analysis



```
@FCD2CN6ACXX:1:1101:1327:2061#/2
GAACATAGATACGGTCACTGAGAGCCAGAGCTTCCT
+
abbeeeceggggggiiiiiiiiiihhiiiiihiiiii
@FCD2CN6ACXX:1:1101:1316:2242#/2
GTCTAAGGAAGAATTTGACCGAGAAAGCGGCGACTA
+
a_aeeeeegfgggiiiihiiighiiiiehifghiii
@FCD2CN6ACXX:1:1101:1598:2195#/2
TTGACAAACTCCTGAGCAAAATAGAGGGAAGTAAAG
+
bbbeeeeeegggggihiiiiiiiihihhiighiggh
@FCD2CN6ACXX:1:1101:1776:2059#/2
TATTAGACTAACAGAGCCTAGAAATTGTCCTAA A
+
bbbeeeeeeggggihiiiiiiiiiiiihfhiihi ii
@FCD2CN6ACXX:1:1101:2209:2091#/2
CTGAAGGCCGCCCAGCAGGGCCAACAAGAGGT GGT
+
bbbeeeeeefggggghhiiiiiiiiiiiiiighic hie
@FCD2CN6ACXX:1:1101:2463:2022#/
TGCTGCTAAAACTGTTGAAATGCTTTATATT CTNN
+
_bbeeeeeegggggiiiiiiiiiiiihihiiii iiBB
@FCD2CN6ACXX:1:1101:2342:2234#/
AAAGAAGTCGATAGGGATCGTTGAATTTAGG CTAT
+
bbbeeeeeeggggggiiifhhihihiihhfgf iifh
@FCD2CN6ACXX:1:1101:2628:2020#/
ACAGAACCGCGAACTGTTGGGCGGATACCTT CCNN
+
abbeeeeeffcggiiihifihhiifdhhhhh egBB
@FCD2CN6ACXX:1:1101:2923:2030#/
ACAGCAAACTGATTGAAGAGGCCGGAAACGA ACNN
+
a^^cccaceab^add]fgdfdge_cdefgaf iaBB
@FCD2CN6ACXX:1:1101:2832:2190#/
GCTCATCGGCTCGAACTTGCGCTACTCCAAA GCGA
```

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

Scientific question

Raw reads

Pre-processing

Assembly: Alignment / *de novo*

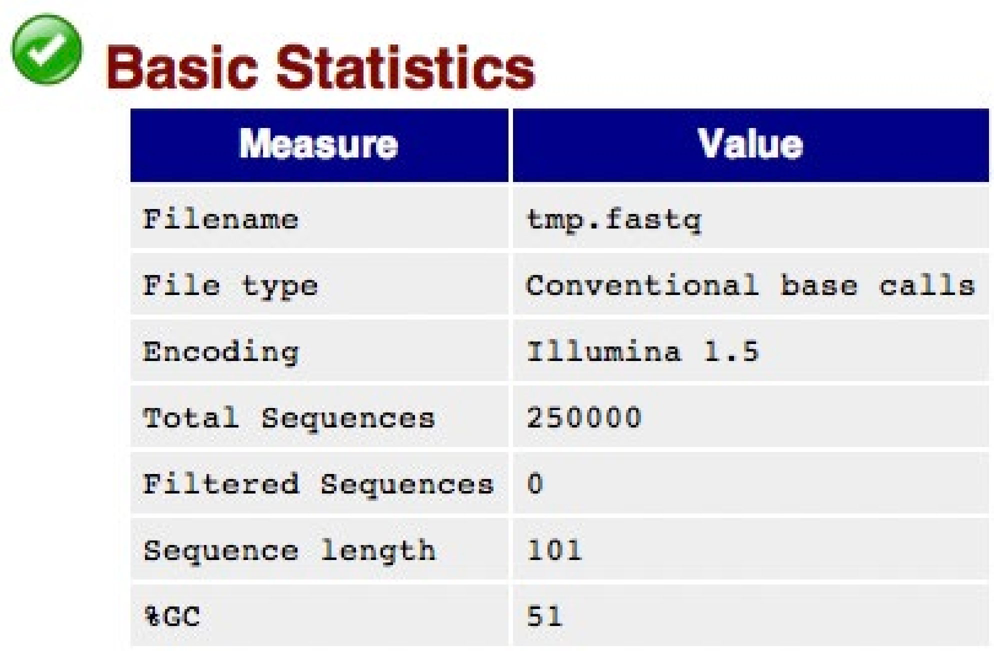
Analysis

Compare samples / methods

Answer

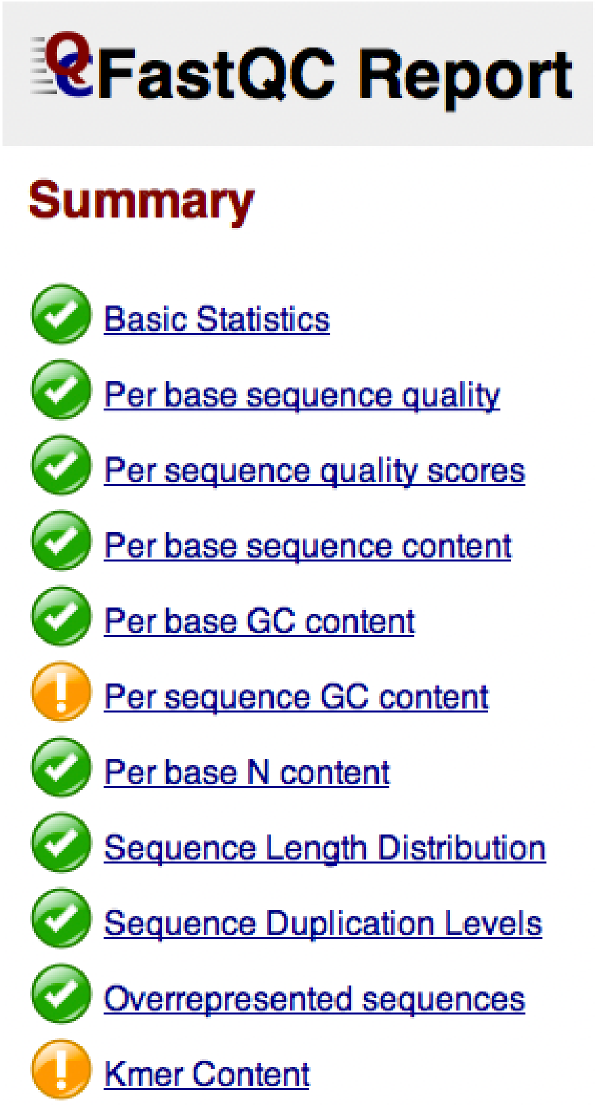
Fastqc reports

- Report basic statistics on your data
- Identify issues with your data



✓ Basic Statistics

Measure	Value
Filename	tmp.fastq
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Filtered Sequences	0
Sequence length	101
%GC	51

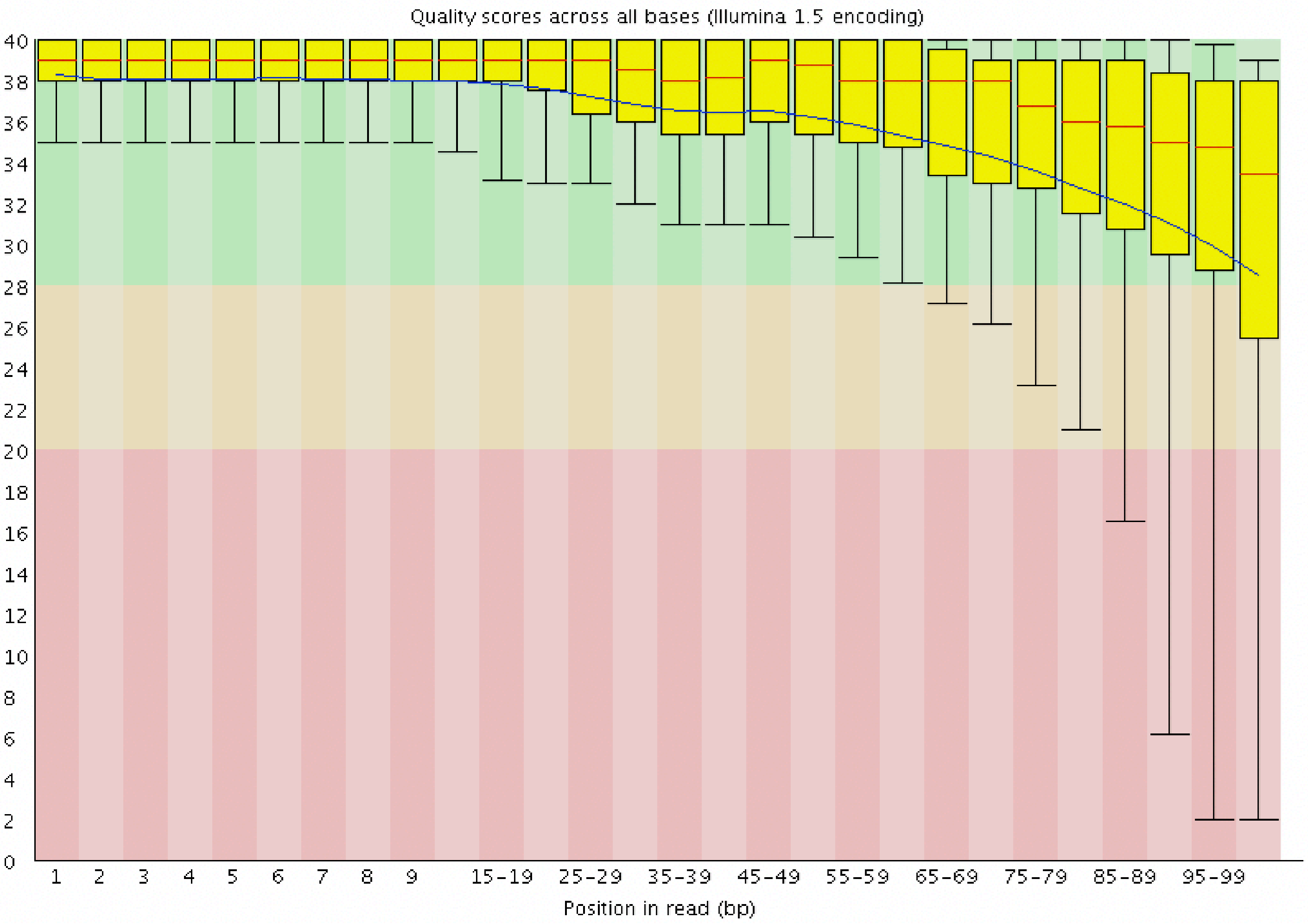


FastQC Report

Summary

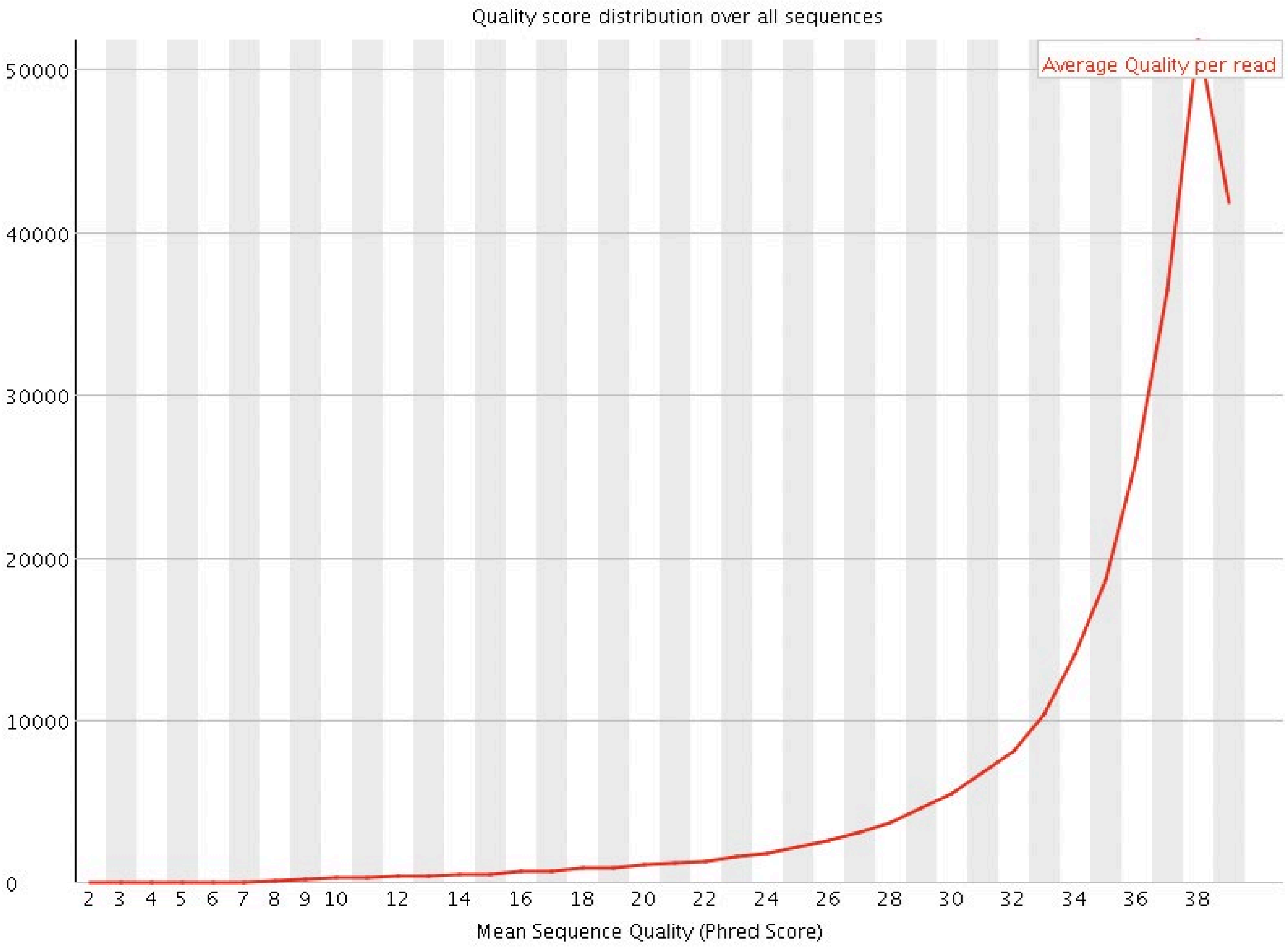
- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ✓ Per base GC content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ! Kmer Content

Per base sequence quality



Quality often decreases over the read.

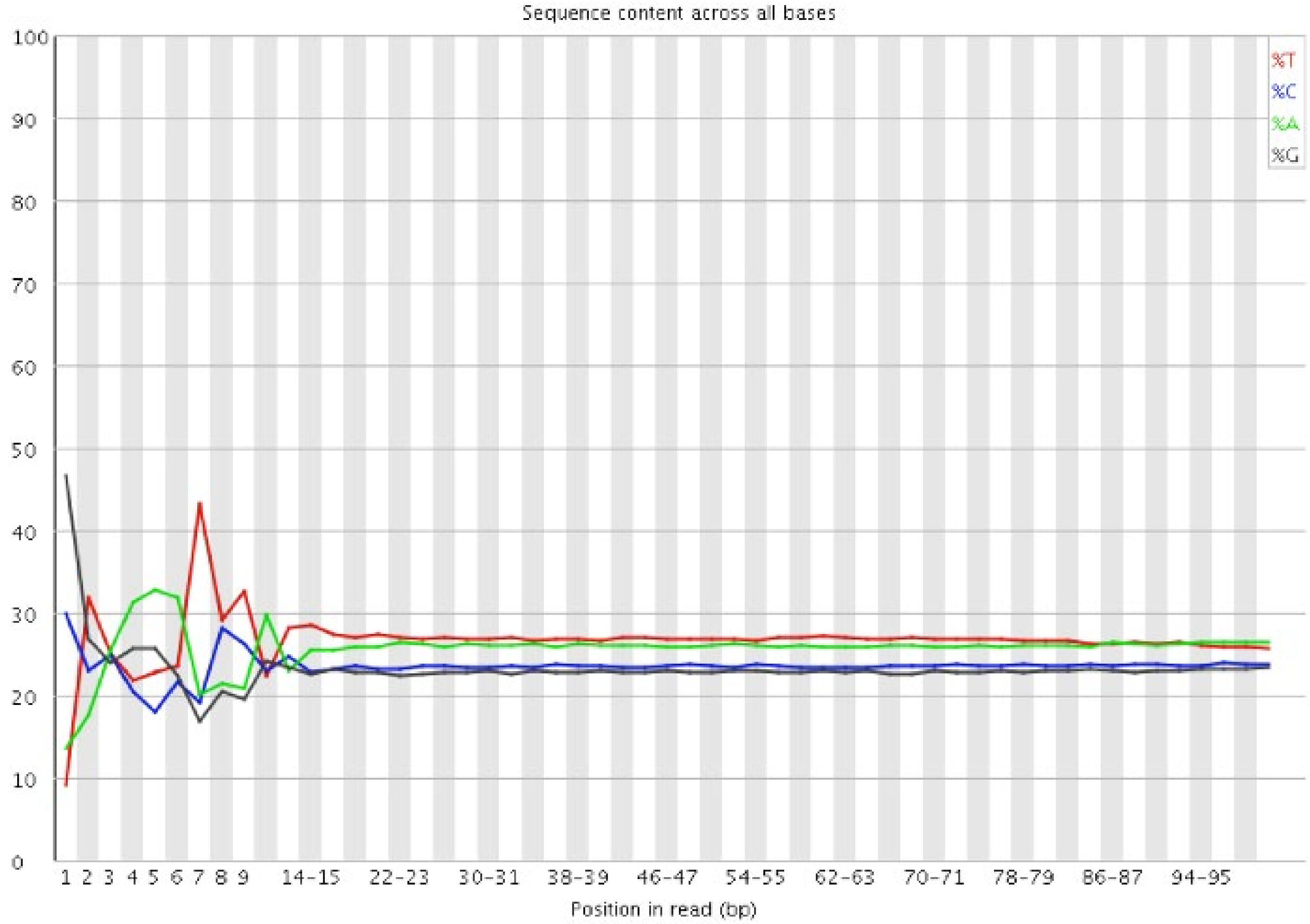
Average quality



Remove reads with a quality below 20.

Remove reads with ‘N’ base calls.

Trim from 5’



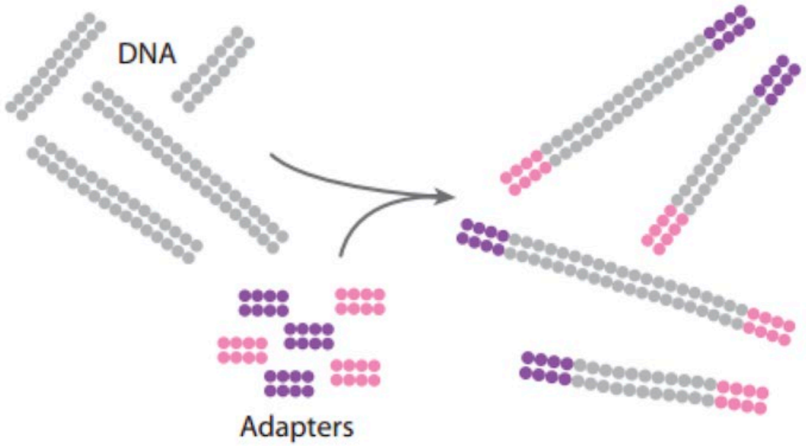
Sometimes something is fishy in the beginning of the read.

It is recommended to remove the first number of bases from the 5’.

How many bases would you remove in this case?

Adapters

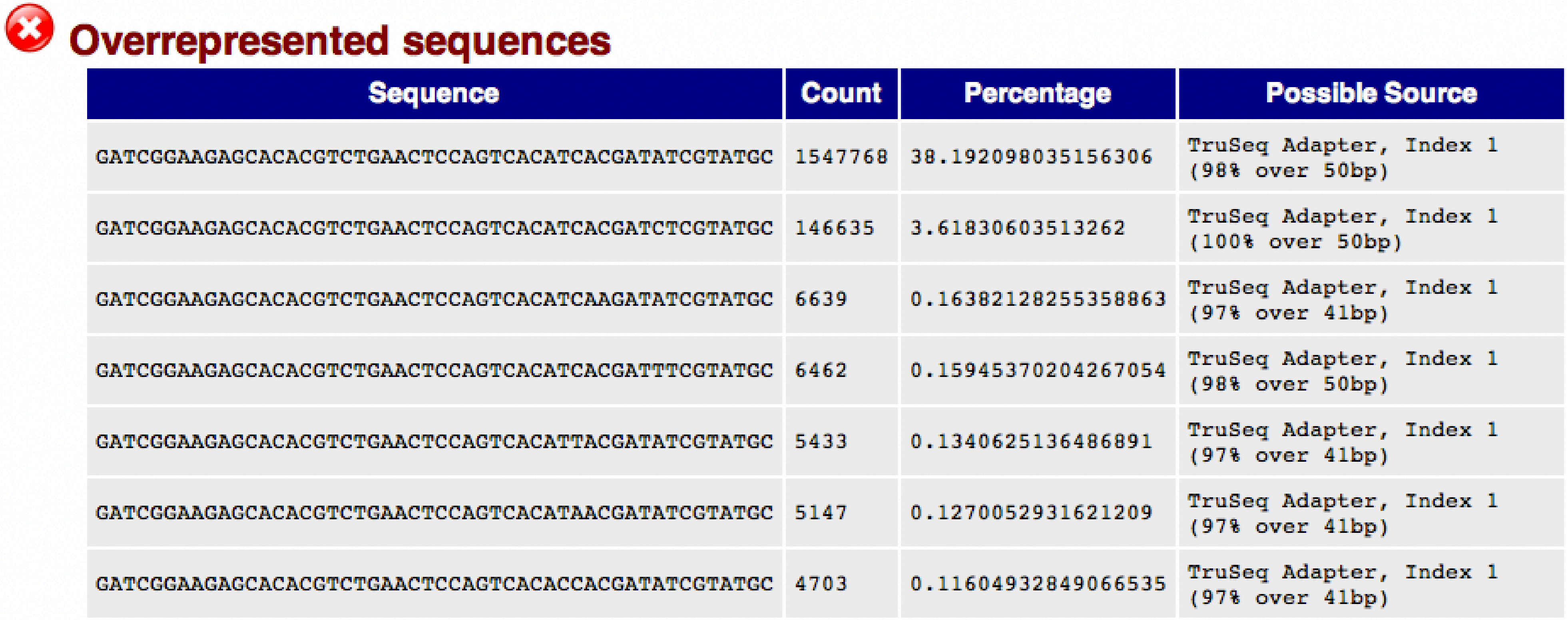
- Sometimes adapters / primers are also part of the read
- Adapter / primers are non-biological sequences
- The artificial repeats will disturb alignments and *de novo* assembly
- The sequence is often known, if not, FastQC may find them



Prepare genomic DNA sample

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Adapters

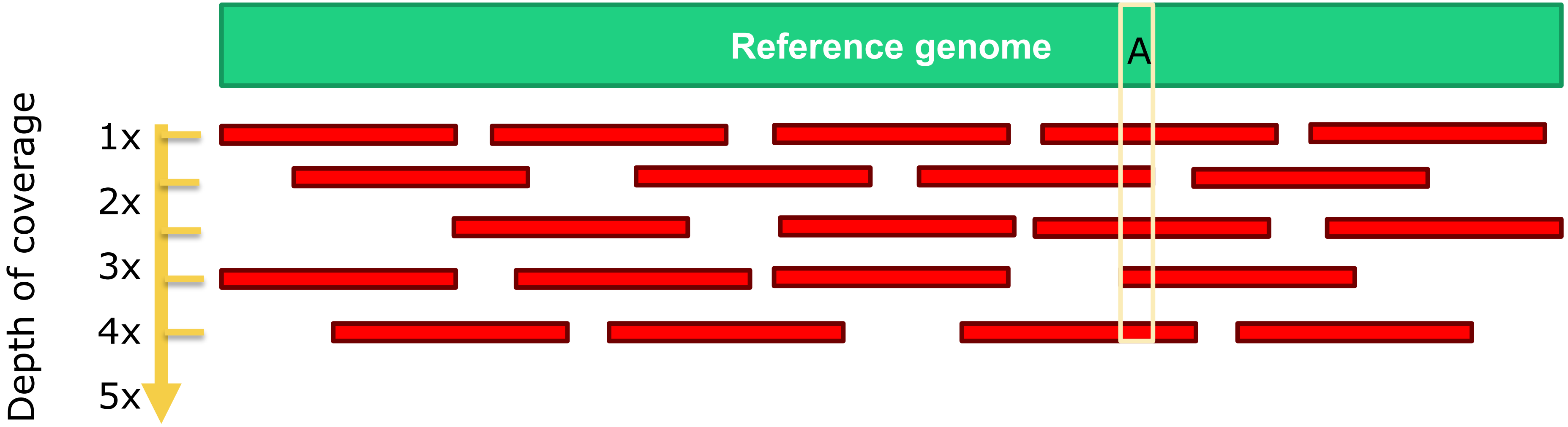


Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATATCGTATGC	1547768	38.192098035156306	TruSeq Adapter, Index 1 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGC	146635	3.61830603513262	TruSeq Adapter, Index 1 (100% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCAAGATATCGTATGC	6639	0.1638212825535886 3	TruSeq Adapter, Index 1 (97% over 41bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATTTCGTATGC	6462	0.15945370204267054	TruSeq Adapter, Index 1 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATTACGATATCGTATGC	5433	0.134062513648689 1	TruSeq Adapter, Index 1 (97% over 41bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATAACGATATCGTATGC	5147	0.127005293162120 9	TruSeq Adapter, Index 1 (97% over 41bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACCACGATATCGTATGC	4703	0.11604932849066535	TruSeq Adapter, Index 1 (97% over 41bp)

We will use “Cutadapt” and “AdapterRemoval”, but other programs can also do the job.

Sequencing Depth



How many times that your data covers the genome (average).

Sequencing depth

$$C = N \times \frac{L}{G}$$

N: Number of reads

L: Read length

G: Genome size

C: Sequencing depth

Example:

N = 5 mill

L: 100 bases

G: 5 mill bases

$$C = 5.000.000 \times \frac{100}{5.000.000}$$

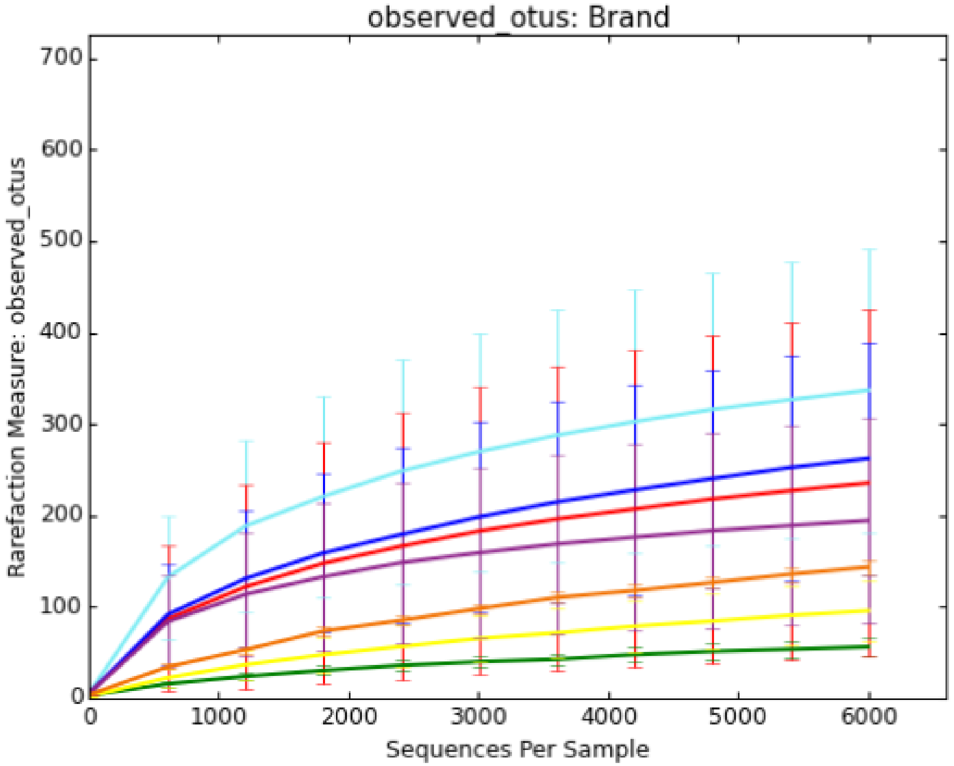
$$C = 5 \times \frac{100}{5}$$

$$C = 100X$$

On average there are 100 reads covering each position in the genome

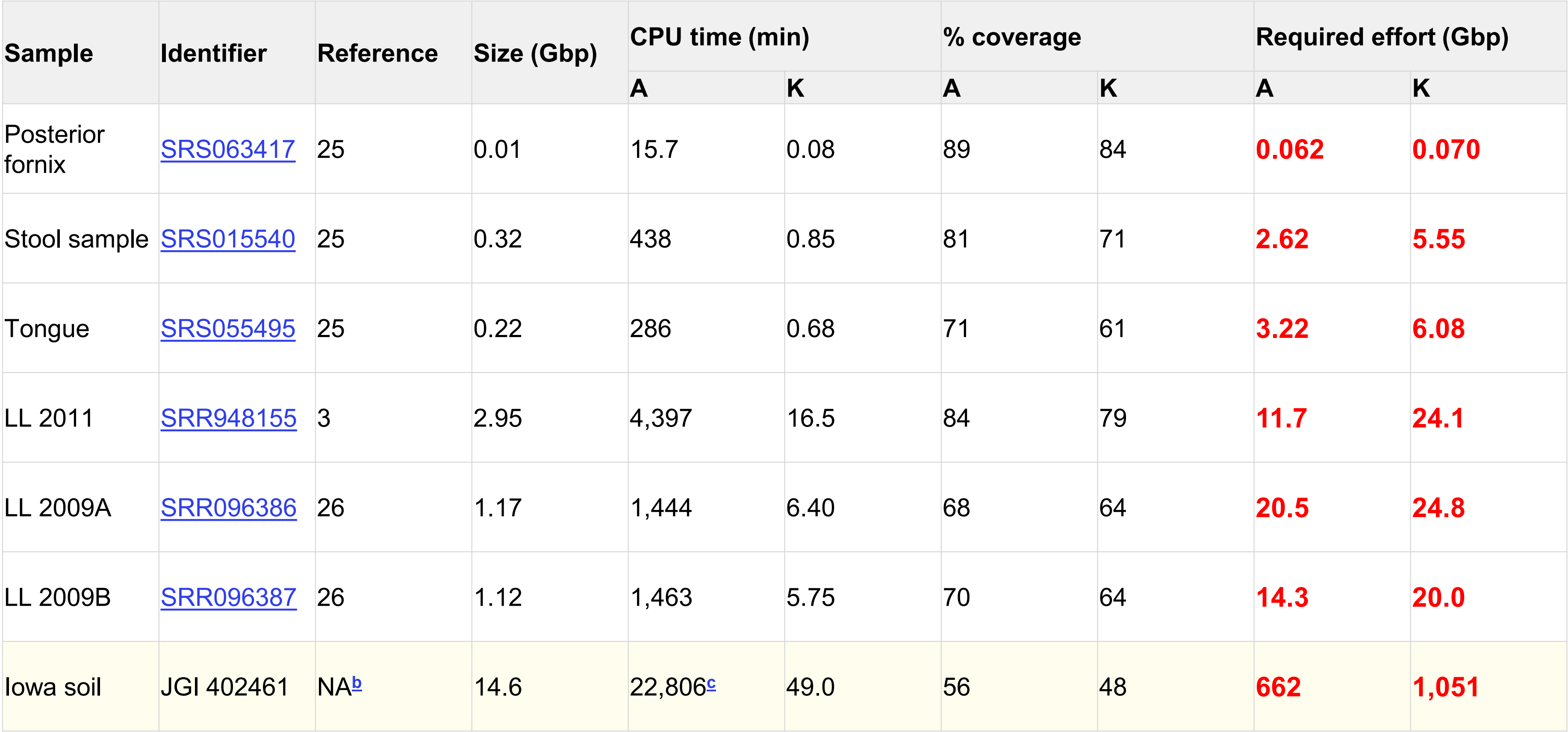
16s rRNA amplicon sequencing depth

- Rarefaction plots!
- Shows types of bacteria as a function of amount of reads



Sequence needed to describe a microbiome I

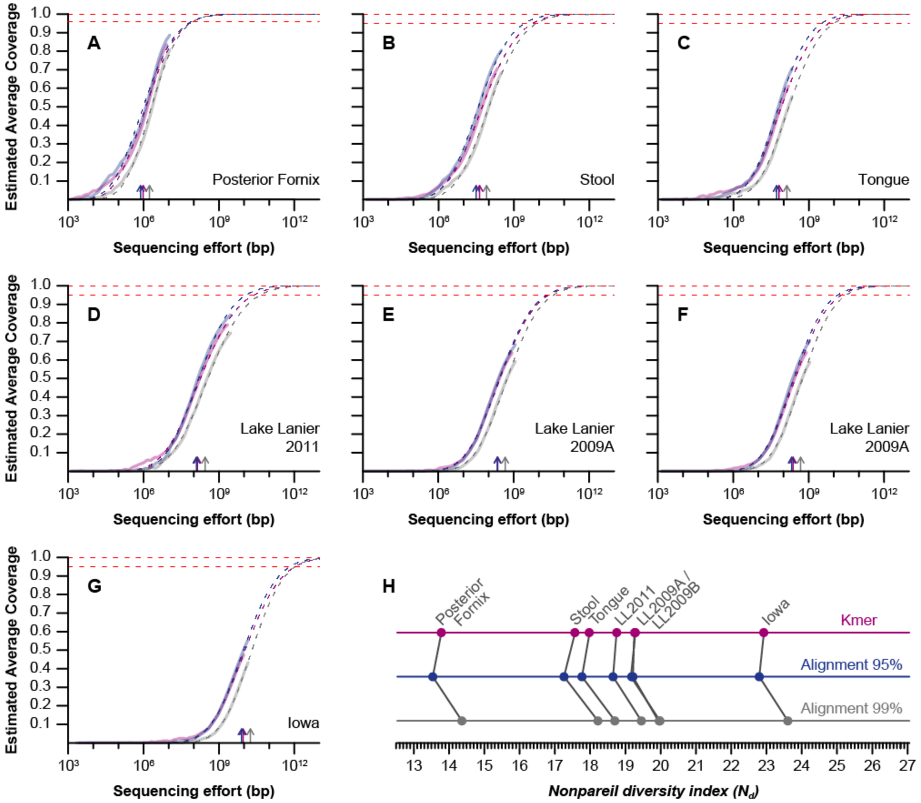
- Huge difference in microbiome diversity



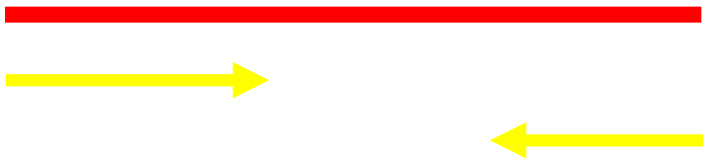
Sample	Identifier	Reference	Size (Gbp)	CPU time (min)		% coverage		Required effort (Gbp)	
				A	K	A	K	A	K
Posterior fornix	SRS063417	25	0.01	15.7	0.08	89	84	**0.062**	**0.070**
Stool sample	SRS015540	25	0.32	438	0.85	81	71	**2.62**	**5.55**
Tongue	SRS055495	25	0.22	286	0.68	71	61	**3.22**	**6.08**
LL 2011	SRR948155	3	2.95	4,397	16.5	84	79	**11.7**	**24.1**
LL 2009A	SRR096386	26	1.17	1,444	6.40	68	64	**20.5**	**24.8**
LL 2009B	SRR096387	26	1.12	1,463	5.75	70	64	**14.3**	**20.0**
Iowa soil	JGI 402461	NA[b]	14.6	22,806[c]	49.0	56	48	**662**	**1,051**

Sequence needed to describe a microbiome

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



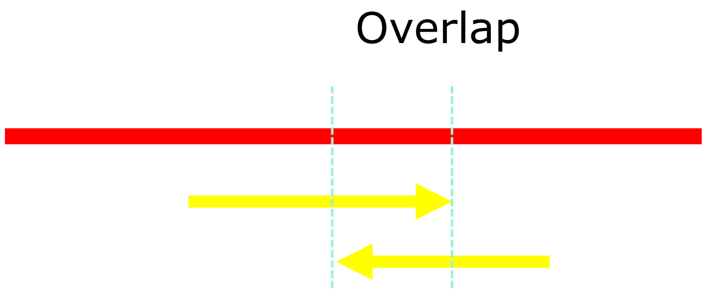
Merge paired end reads



Insert size: 500nt

Reads: 100nt

Middle: 300nt



Insert size: 180nt

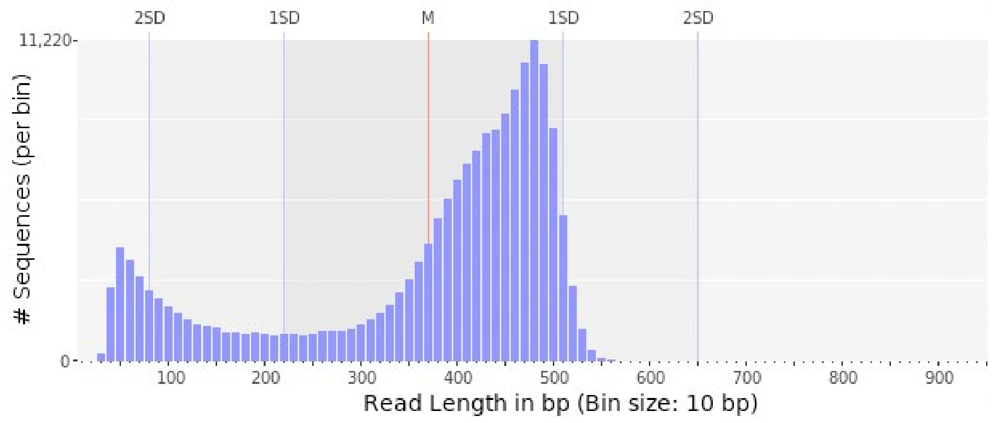
Reads: 100nt

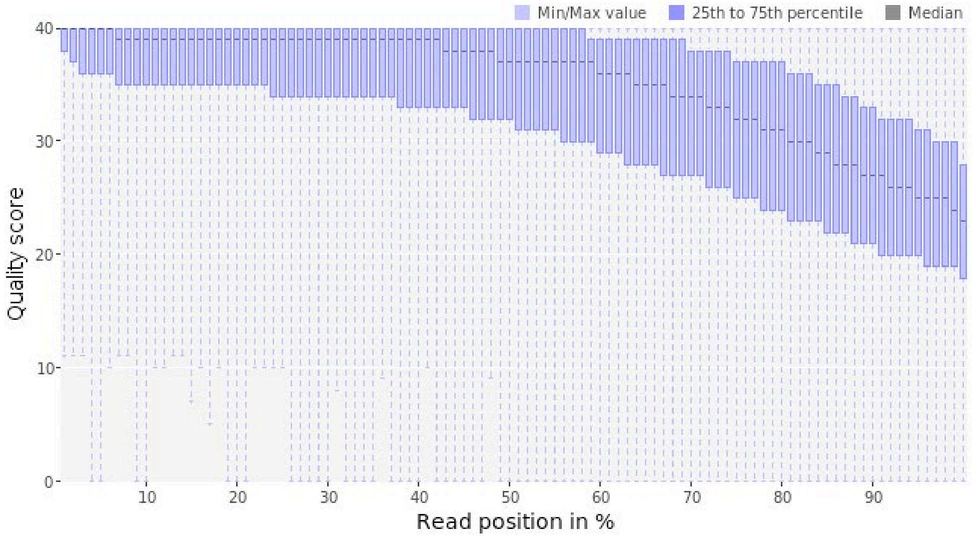
Middle: -20nt

- Merge overlapping pairs into single longer read
- Smart because Illumina reads have low quality in the 3’
- Very useful for *de novo* assembly

454 / Ion torrent data

- Main problem is indels at homopolymer runs
- (Trim homopolymers), trim trailing poor quality bases
- Remove very short reads
- For *de novo* adapters should be removed (prinseq)
- For alignment we use Smith- Waterman (local) so less important





Final – but important note

- Lots of data - storage is expensive!
- Keep data compressed whenever possible (gzip, bzip, bam)
- Remove intermediate files and files that can easily be re-created