DATA PREPROCESSING

NEXT GENERATION SEQUENCING ANALYSIS

BY KATRINE HØJHOLT IVERSEN

OUTLINE

- 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

MAIN STEPS IN NGS ANALYSIS

DATA SIZE

?

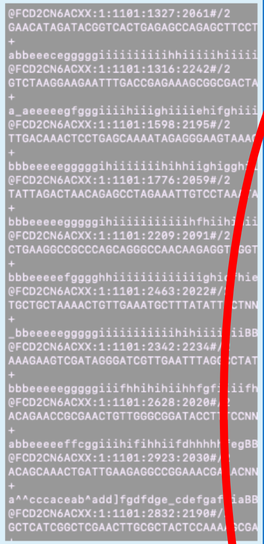
?

?

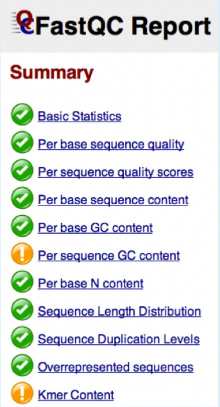
?

?

?



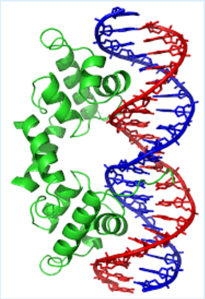
```
@FCD2CN6ACXX:1:1101:1327:2061#/2
GAACATAGATACGGTCACTGAGAGCCAGAGCTTCCT
+
abbeeeceggggggiiiiiiiiiiihhiiiiiihiiiii
@FCD2CN6ACXX:1:1101:1316:2242#/2
GTCTAAGGAAGAATTTGACCGAGAAAGCGGCGACTA
+
a_aeeeeeegfgggiiiihiiighiiiiehifghiii
@FCD2CN6ACXX:1:1101:1598:2195#/2
TTGACAAACTCCTGAGCAAAATAGAGGGAAGTAAAG
+
bbbeeeeeeggggggihiiiiiiiihihhiighigghi
@FCD2CN6ACXX:1:1101:1776:2059#/2
TATTAGACTAACAGAGCCTAGAAATTGTCCTAAATA
+
bbbeeeeeeggggggihiiiiiiiiiiiihfhiihiiii
@FCD2CN6ACXX:1:1101:2209:2091#/2
CTGAAGGCCGCCCAGCAGGGCCAACAAGAGGTGGGT
+
bbbeeeeeefggggghhiiiiiiiiiiiiiighicfhie
@FCD2CN6ACXX:1:1101:2463:2022#/2
TGCTGCTAAAACTGTTGAAATGCTTTATATTCCTNN
+
_bbeeeeeegggggiiiiiiiiiiiihihiiiiiiiBB
@FCD2CN6ACXX:1:1101:2342:2234#/2
AAAGAAGTCGATAGGGATCGTTGAATTTAGCCTAT
+
bbbeeeeeeggggggiiifhhihihiihhfgfiiiifh
@FCD2CN6ACXX:1:1101:2628:2020#/2
ACAGAACCGCGAACTGTTGGGCGGATACCTTCCNN
+
abbeeeeeffcggiiihifihhiifdhhhhhfegBB
@FCD2CN6ACXX:1:1101:2923:2030#/2
ACAGCAAACTGATTGAAGAGGCCGGAAACGAAACNN
+
a^^cccaceab^add]fgdfdge_cdefgafiiaBB
@FCD2CN6ACXX:1:1101:2832:2190#/2
GCTCATCGGCTCGAACTTGCGCTACTCCAAAAGCGA
+
```

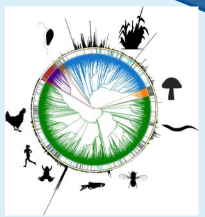


FastQC Report

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

WHY IS PREPROCESSING IMPORTANT?

Quality?

Every base in a read have a quality score
Note: bases are not always correct!

Errors?

Different sequencing technologies has different error profiles.

Do we trust our data?

Adapters?

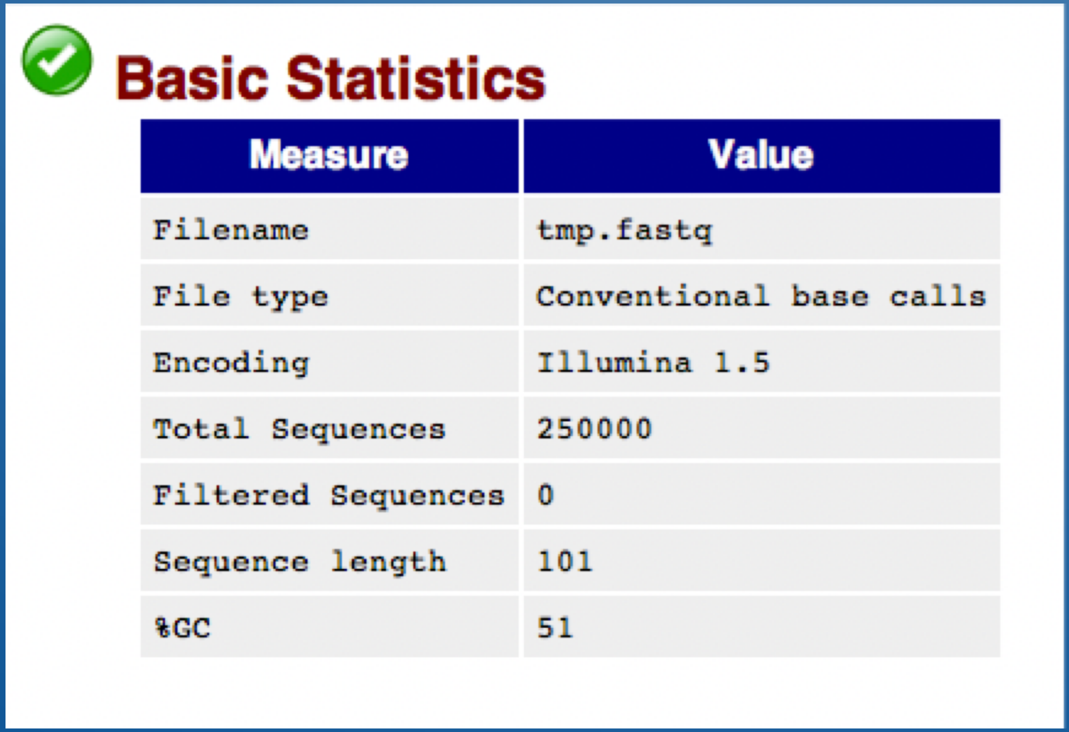
Adapters/primers are non-biological sequences that can be a part of the raw data.

Sequencing depth?

How deep is the sample sequenced. How many times that your data covers the genome.

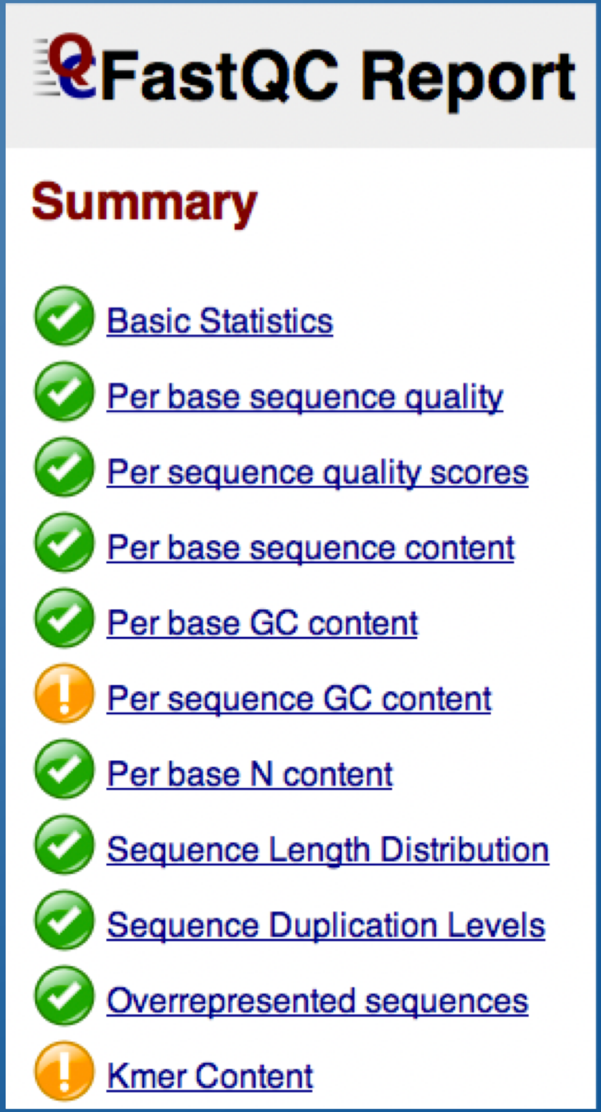
FASTQC REPORTS

- 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

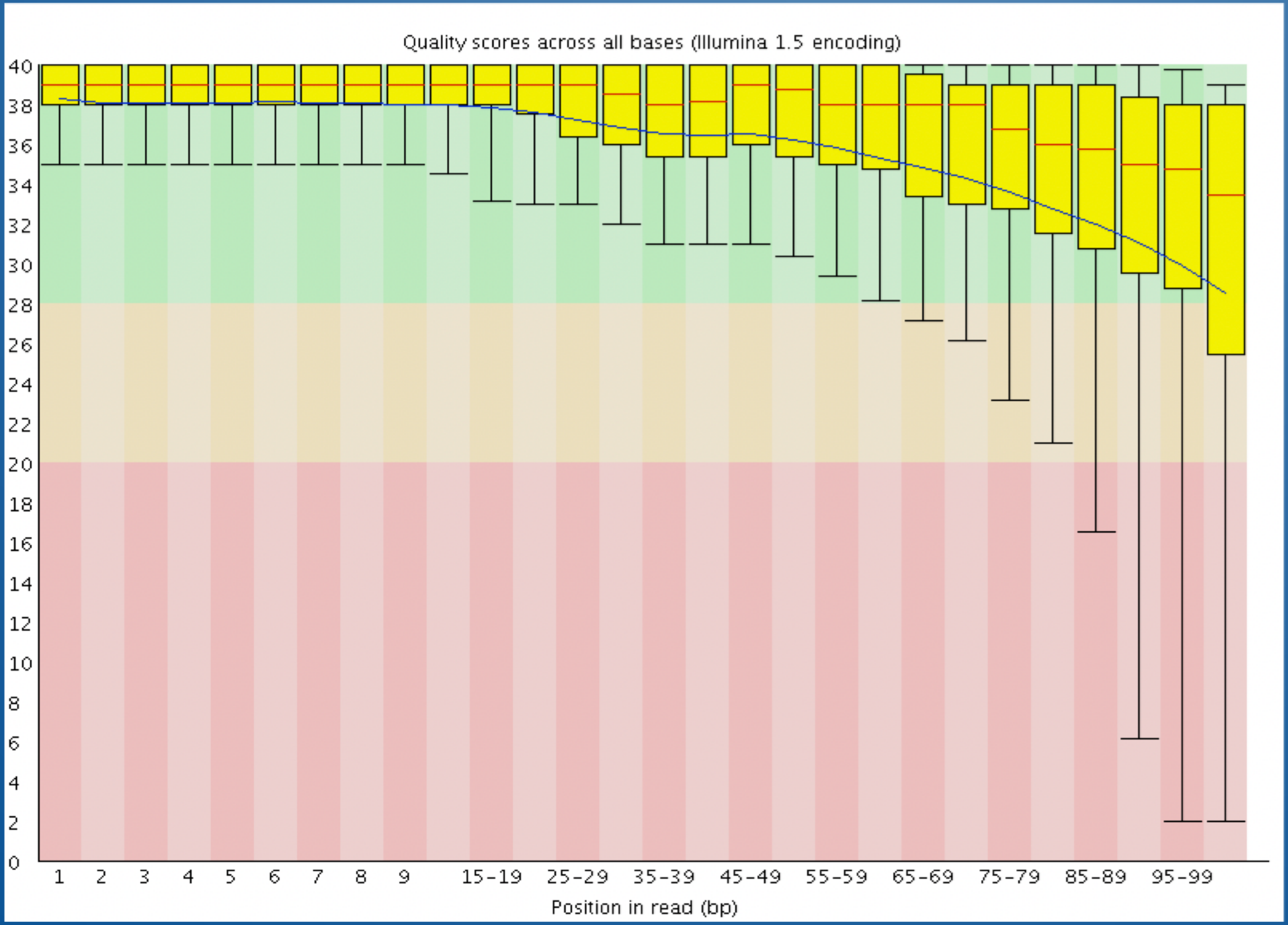


FastQC Report

Summary

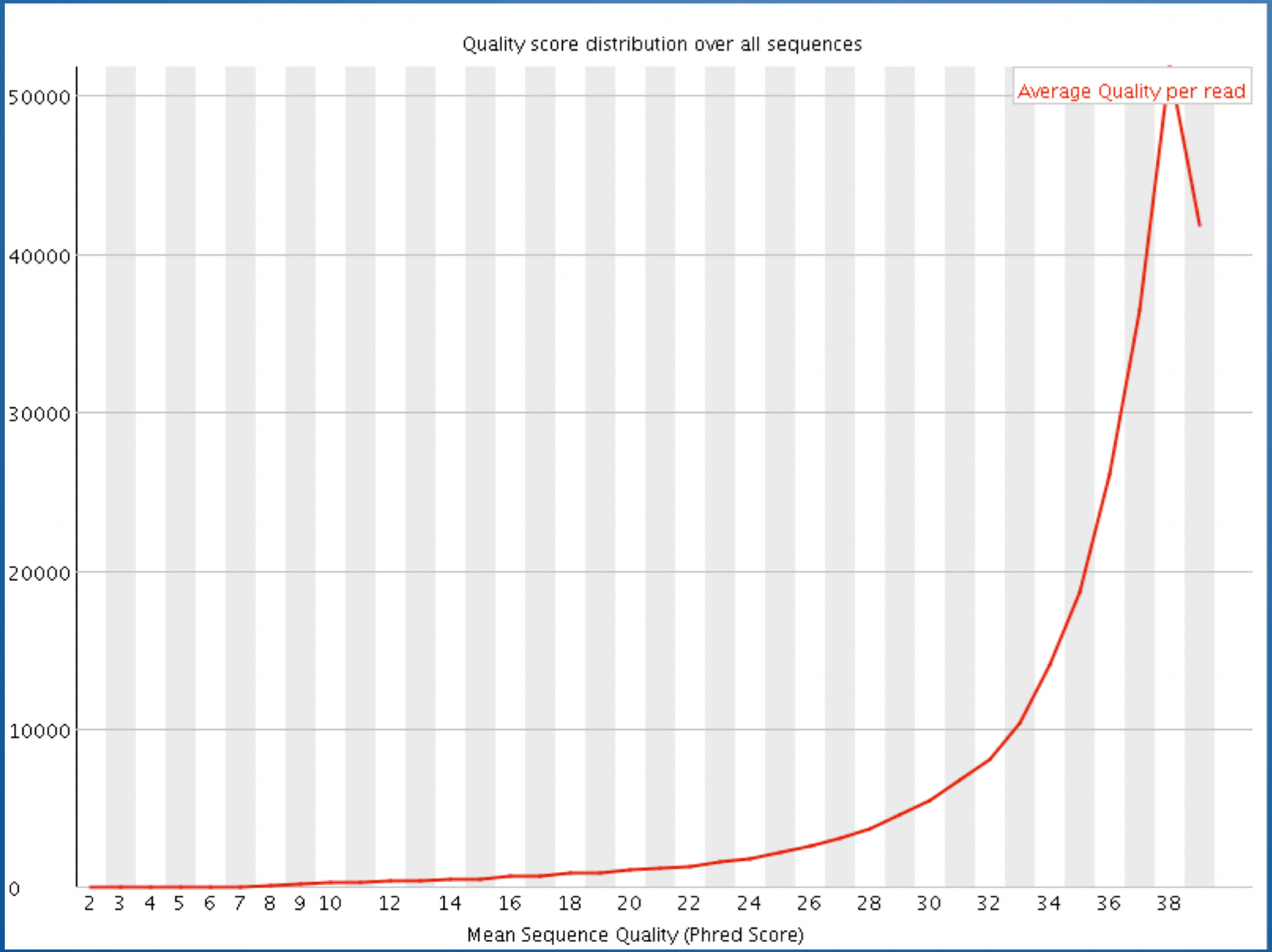
- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ✓ Per base GC content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ! Kmer Content

PER BASE SEQUENCE QUALITY



Quality often decreases over the read.

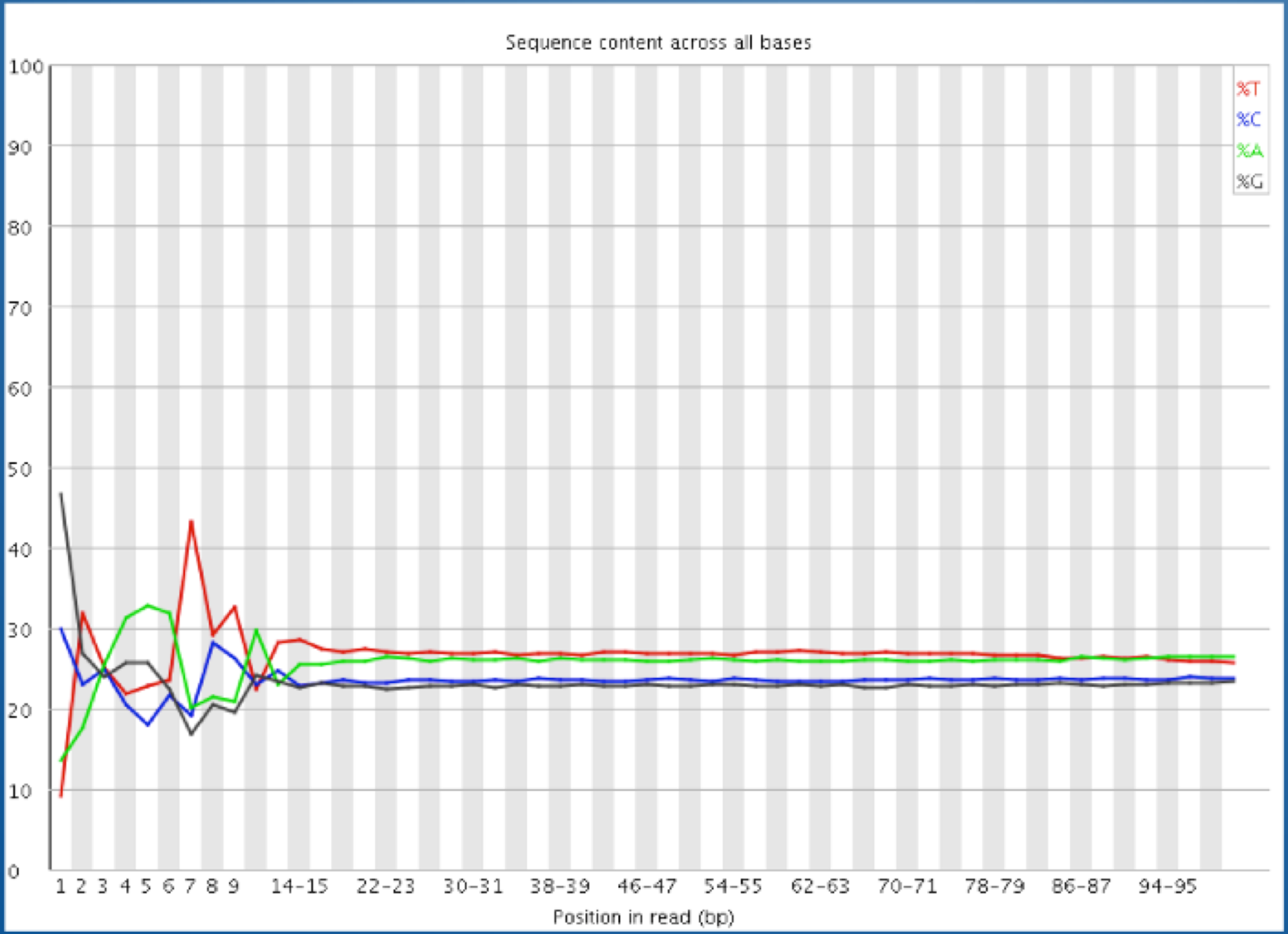
AVERAGE QUALITY



Remove reads with a quality below 20.

Remove reads with ‘N’ base calls.

TRIM FROM 5’



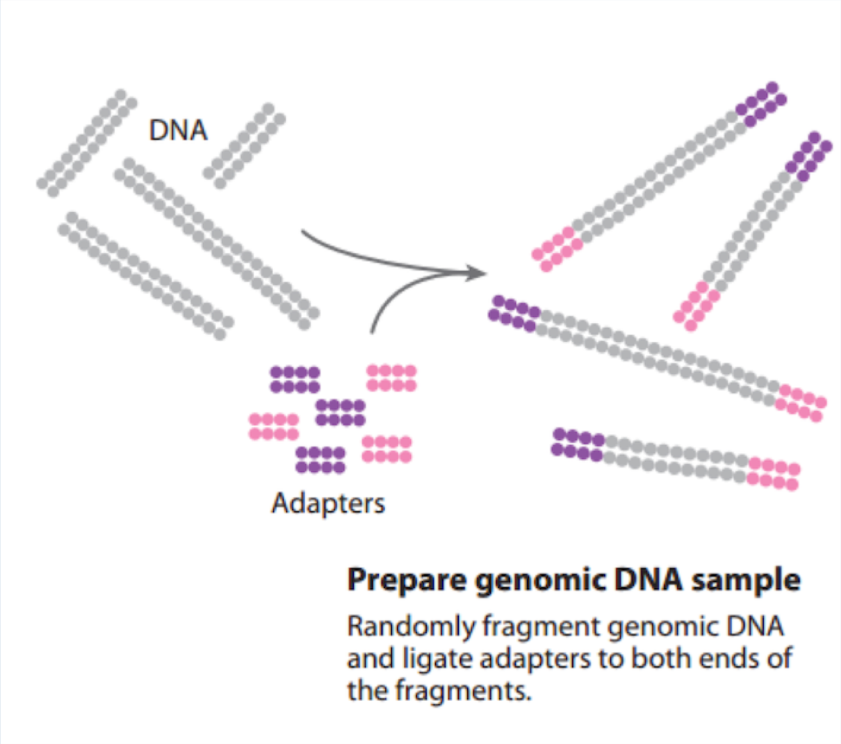
Sometimes something is fishy in the beginning of the read.

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

How many bases would you remove in this case?

ADAPTERS

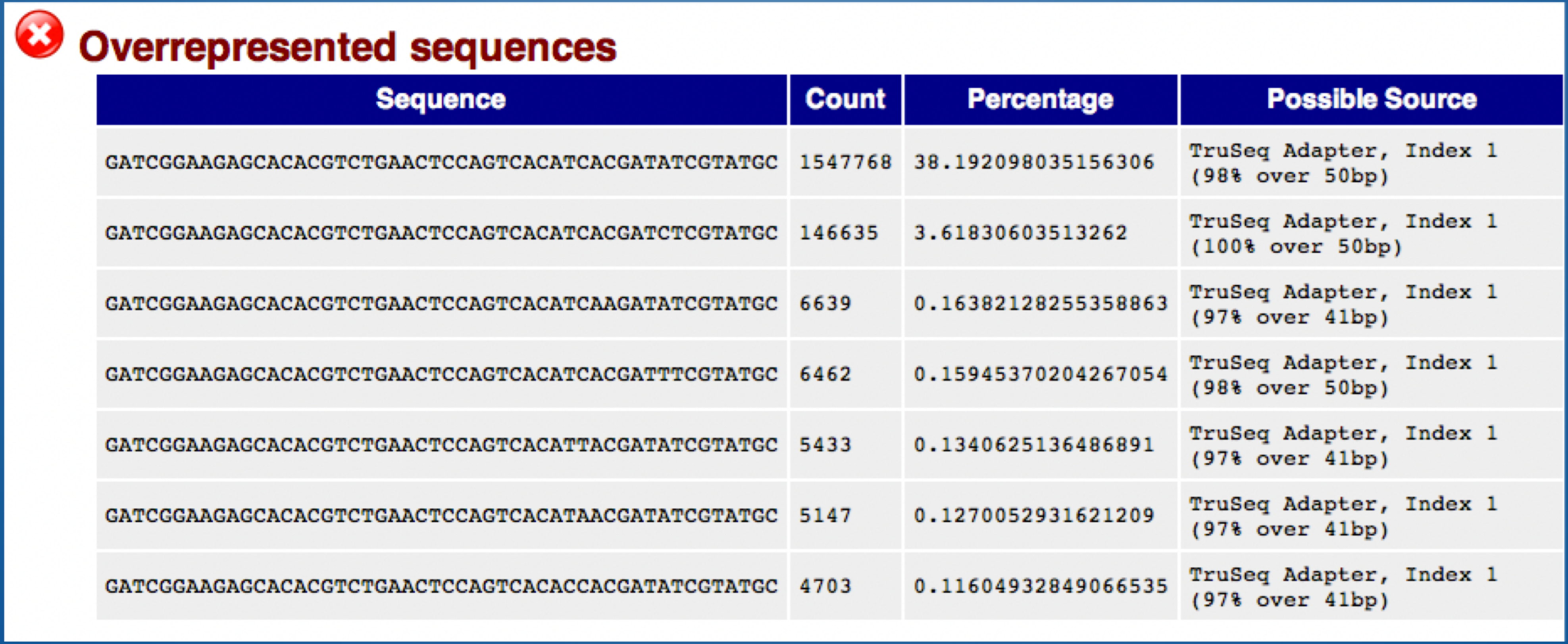
- Sometimes adapters / primers are also part of the read
- Adapter / primers are non-biological sequences
- The artificial repeats will disturb alignments and *de novo* assembly
- The sequence is often known, if not, FastQC may find them



Prepare genomic DNA sample

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

ADAPTERS



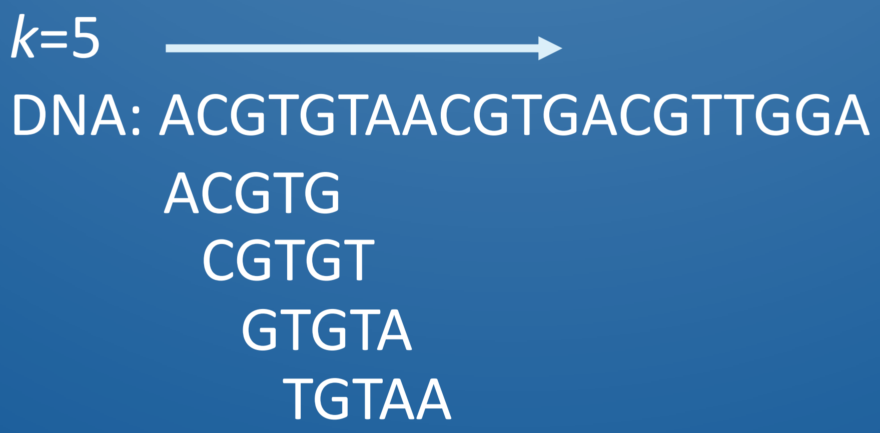
Overrepresented sequences

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	0.16382128255358863	TruSeq Adapter, Index 1 (97% over 41bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATTTCGTATGC	6462	0.15945370204267054	TruSeq Adapter, Index 1 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATTACGATATCGTATGC	5433	0.1340625136486891	TruSeq Adapter, Index 1 (97% over 41bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATAACGATATCGTATGC	5147	0.1270052931621209	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.

K-MER CORRECTION

- Create a sliding window of size *k*, move it over all your reads and count occurrence of *k*-mers
- We can use this to correct sequencing errors!

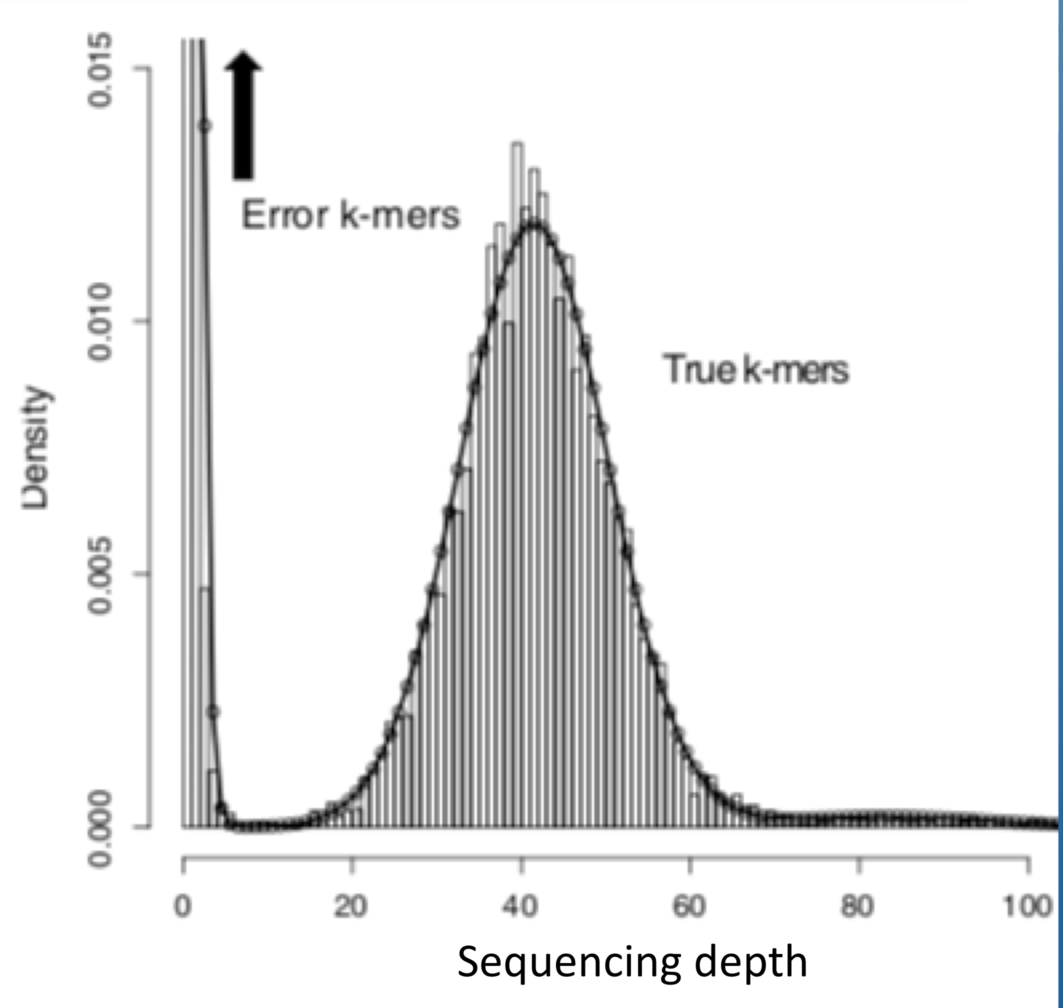


k=5

DNA: ACGTGTAACGTGACGTTGGA

ACGTG
CGTGT
GTGTA
TGTAA

K-MER CORRECTION



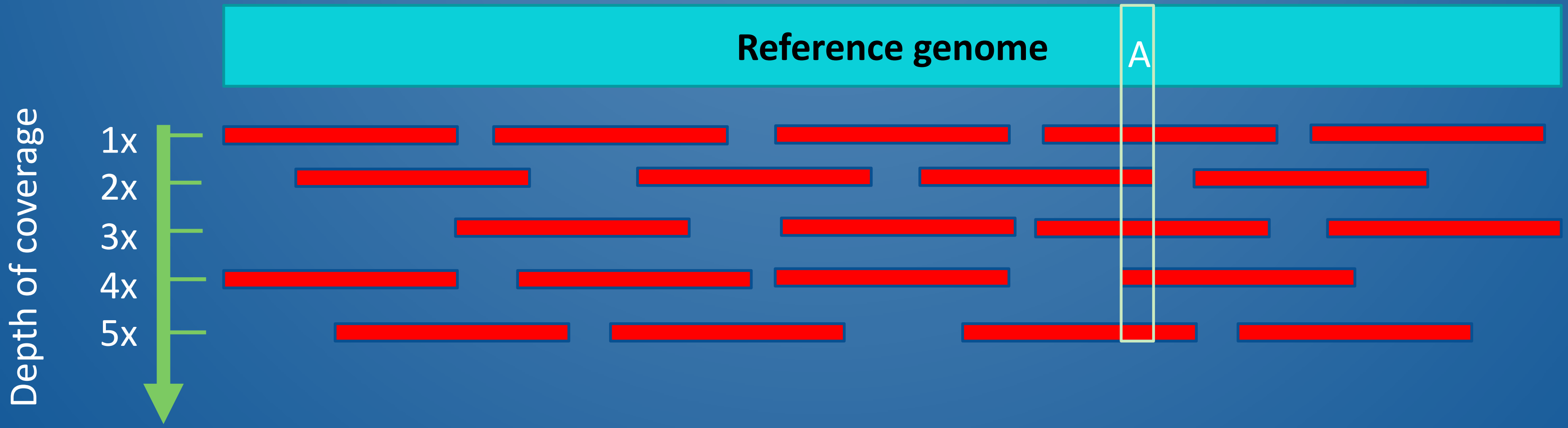
Concept: rare *k*-mers are sequencing errors.

In general we need a > 15x sequencing depth

ACGTGGTTGCCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA

Kelley *et al.*, 2010

SEQUENCING DEPTH



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

SEQUENCING DEPTH

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

N: Number of reads

L: Read length

G: Genome size

C: Sequencing depth

Example:

N = 5 mill

L: 100 bases

G: 5 mill bases

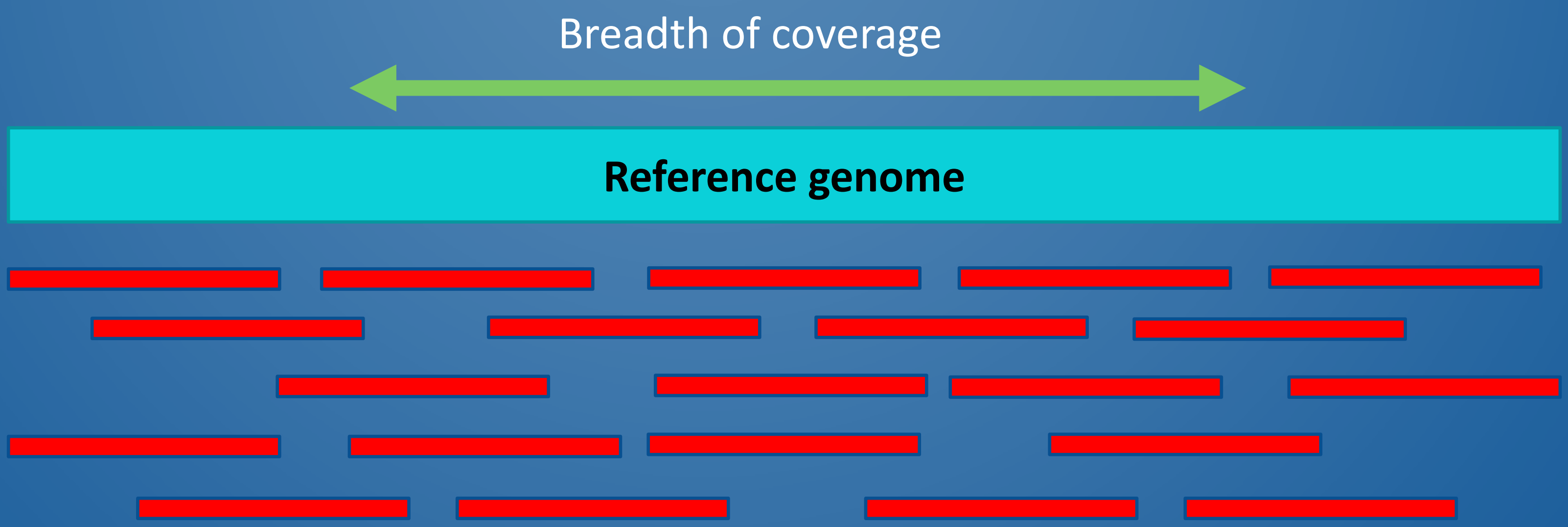
$$C = 5.000.000\times\frac{100}{5.000.000}$$

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

$$C = 100X$$

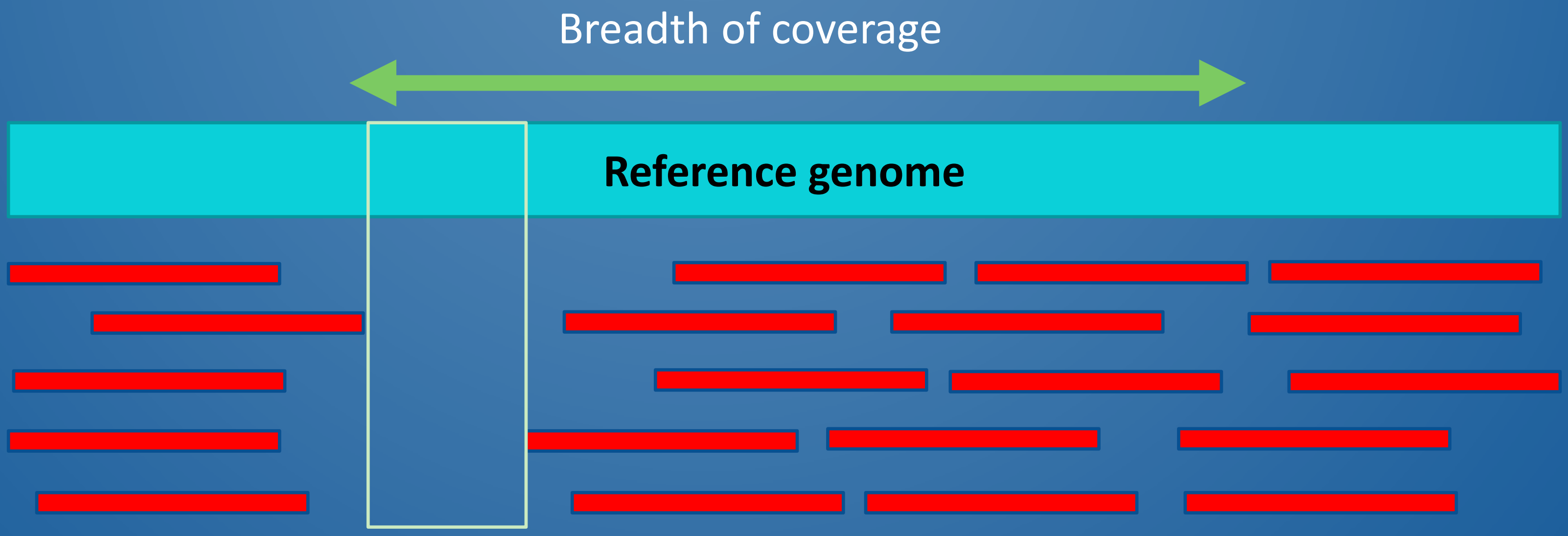
On average there are 100 reads covering each position in the genome

GENOME COVERAGE



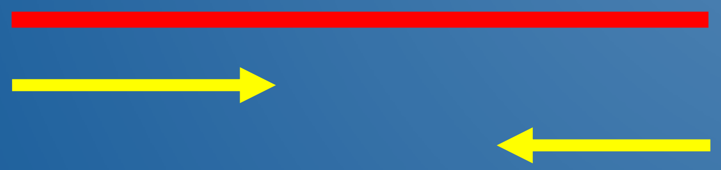
How much of the reference genome is covered by your data

GENOME COVERAGE

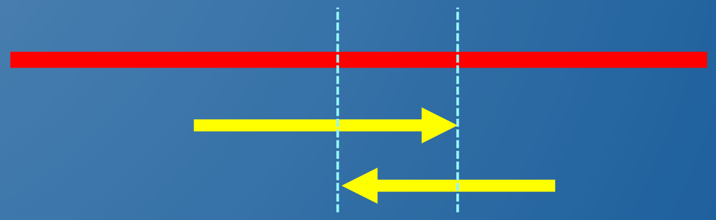


Uncovered part of the genome

MERGE PAIRED END READS



Insert size: 500nt
Reads: 100nt
Middle: 300nt

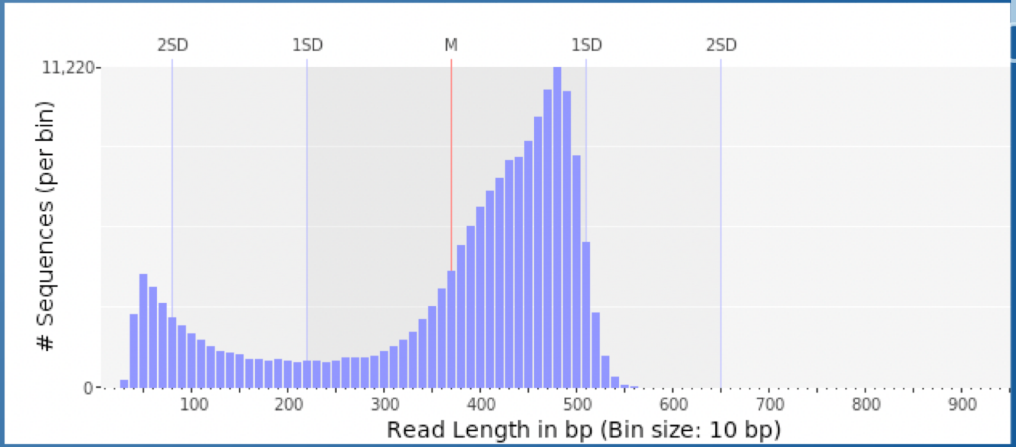


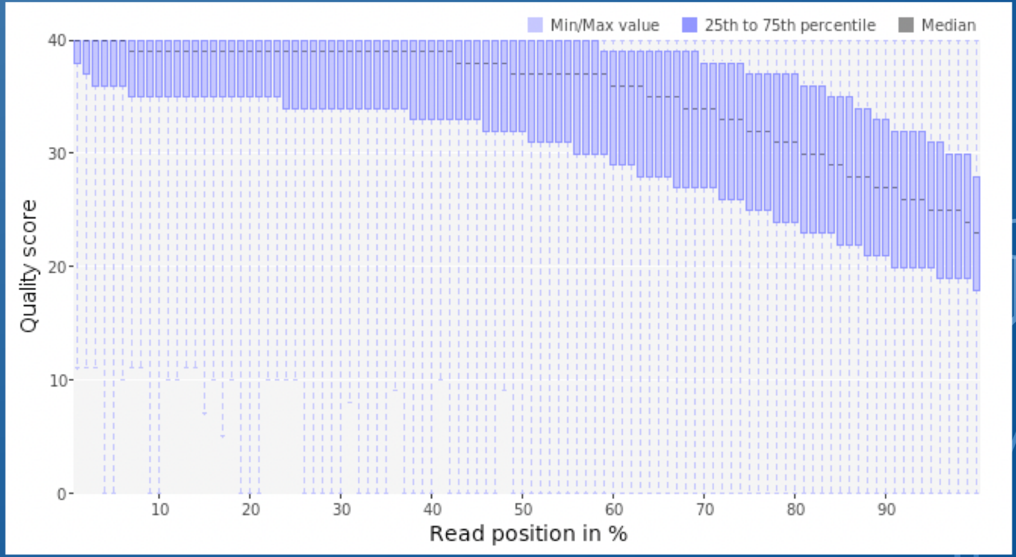
Insert size: 180nt
Reads: 100nt
Middle: -20nt

- 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