Quality control, filtering, assembly & validation

Workshop in marine metagenomics

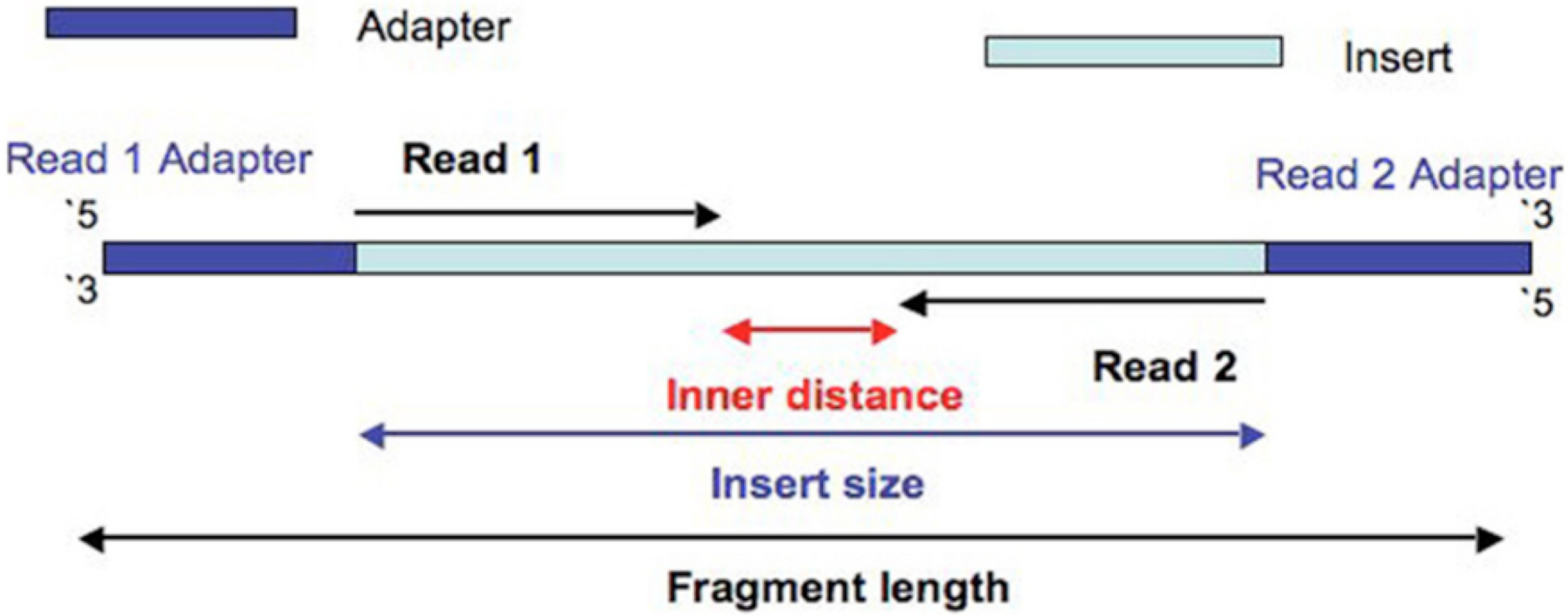
Tromsø November 2018

www.elixir-europe.org

https://jobsite.procore.com/this-friday-s-funny-the-history-of-quality-control

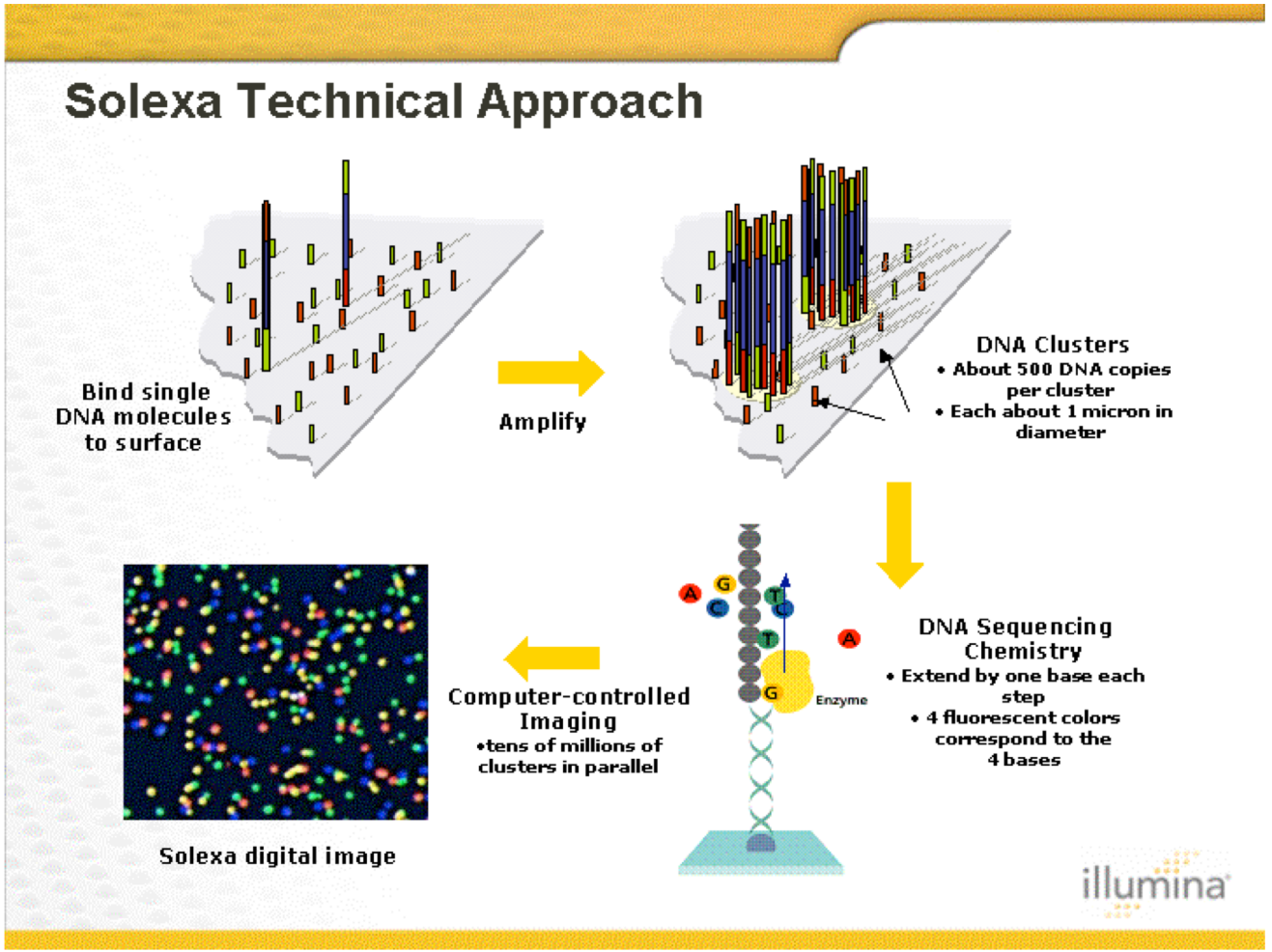
DNA sequencing recap – Illumina technology

Paired-end (PE) Illumina data



Sequencing quality scores measure the probability that a base is called incorrectly

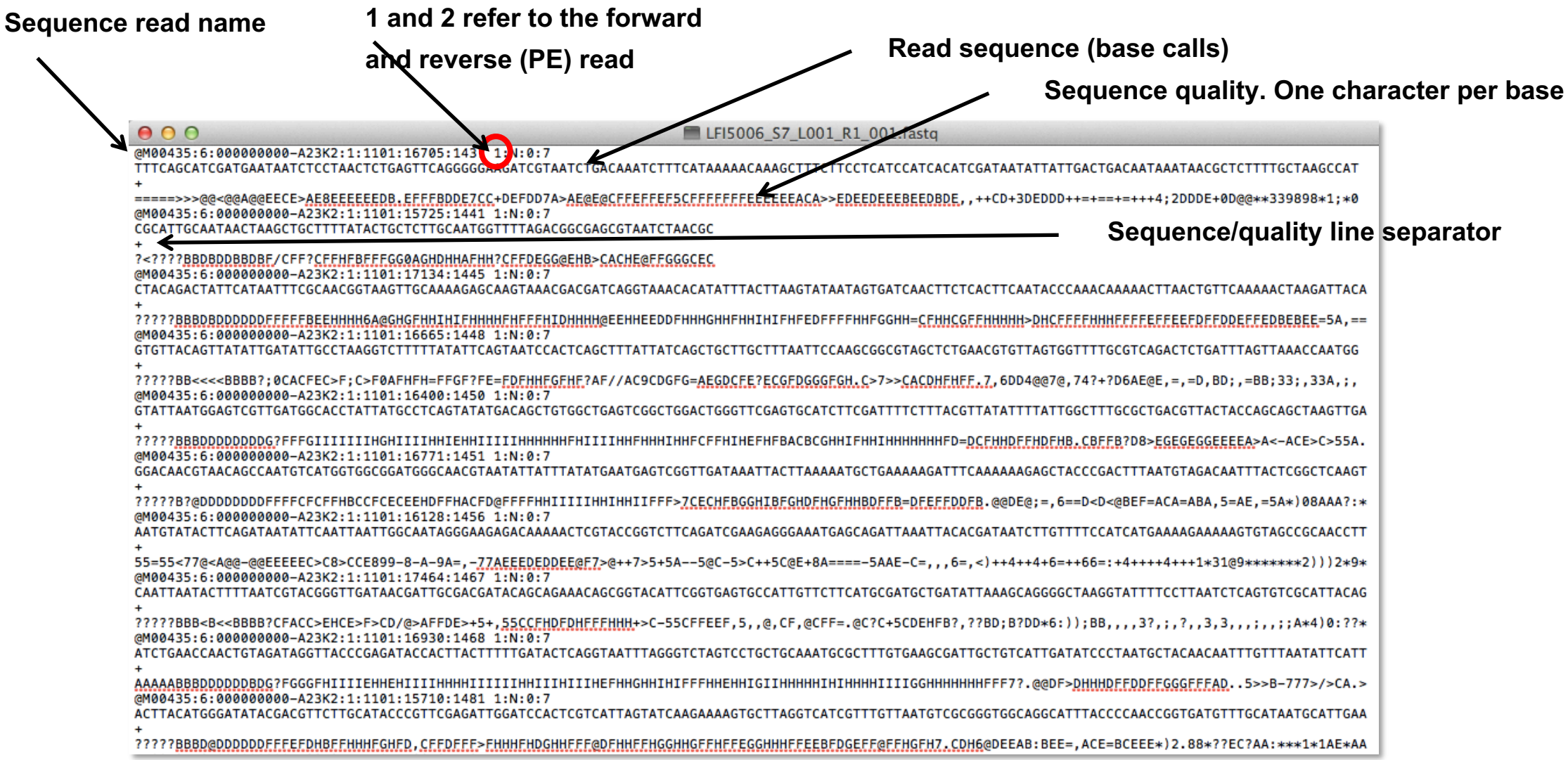
Base calls are made directly from signal intensity measurements



Sequencing technologies are not perfect and do produce errors

QC ensures that the data used for downstream analysis does not contain (too many) errors and poor quality sequences

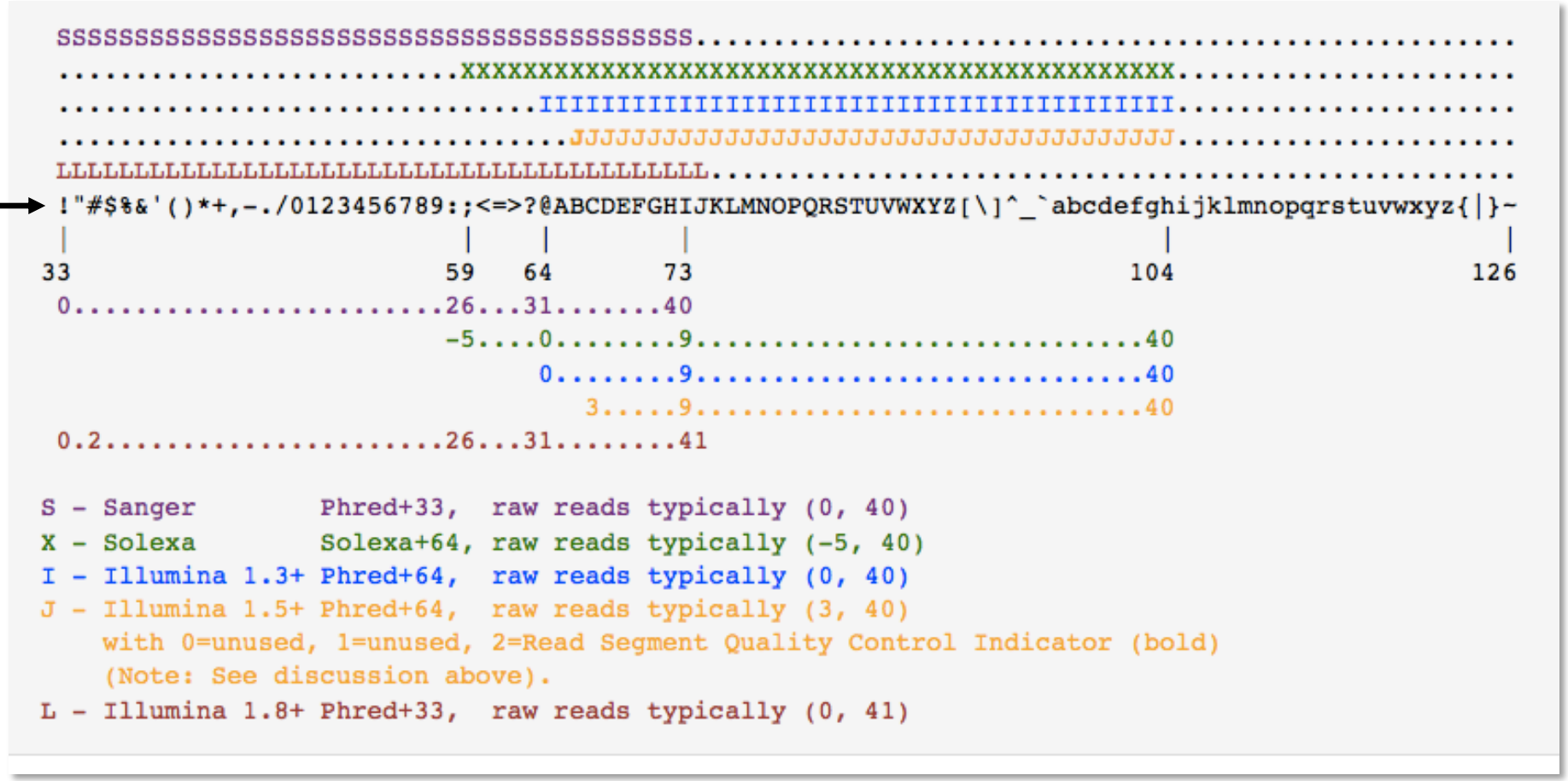
The Illumina output are FASTQ files



A range of sequence quality scores depending on technology and the base caller

Most modern sequencing machine, such as MiSeq use Illumina 1.8+

ASCII codes →



Q30 is considered a benchmark for sequence quality in next-generation sequencing

A quality score of 99,9% (Q30) will have an incorrect base call probability of 1 in 1000

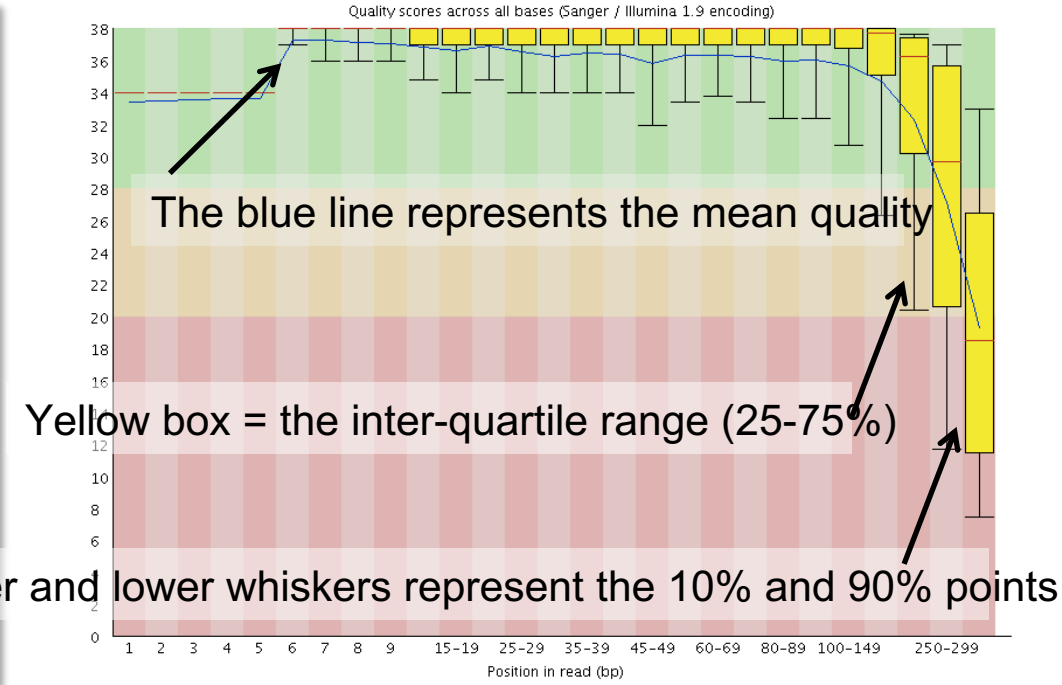
Relationship Between Sequencing Quality Score and Base Call Accuracy:

Quality Score	Probability of Incorrect Base Call	Inferred Base Call Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%

When performing Quality Control (QC) you generate a general summary of the input data

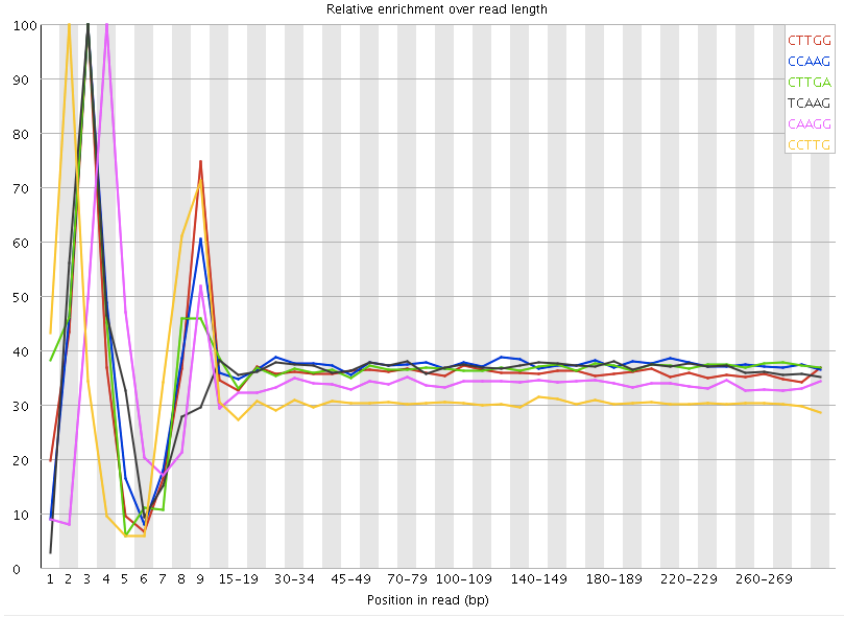
The range of quality values across all bases at each position in the FastQ file

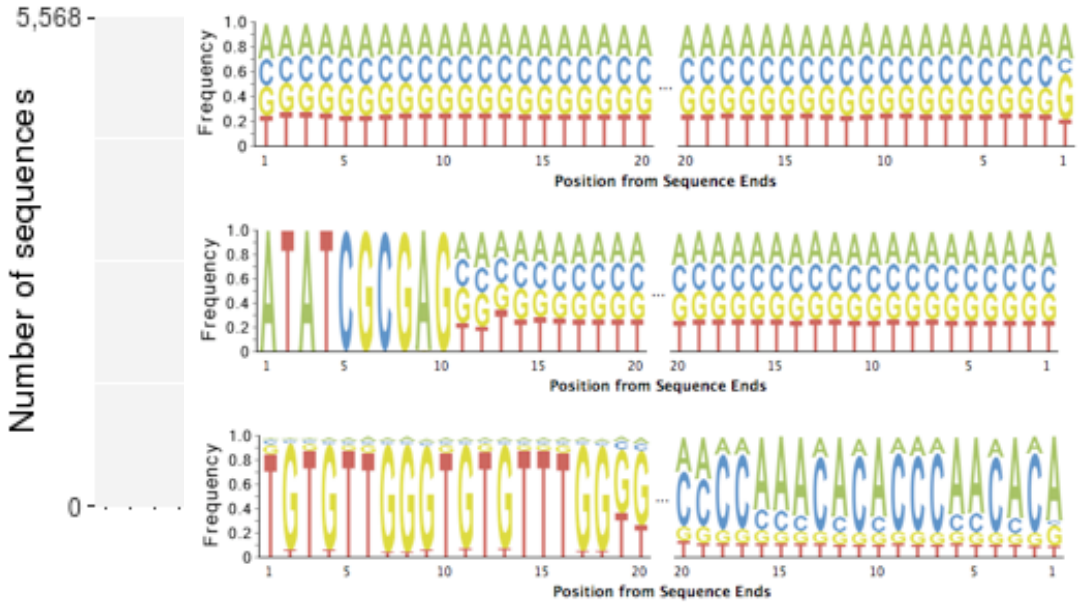
Measure	Value
Filename	CAV3_S1_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2172538
Filtered Sequences	0
Sequence length	35-301
%GC	41



Low quality sequences can cause problems during downstream analysis

Analysing K-mer content may identify artefacts in sequence reads, eg. multiplex identifiers, adapters, and primer sequences

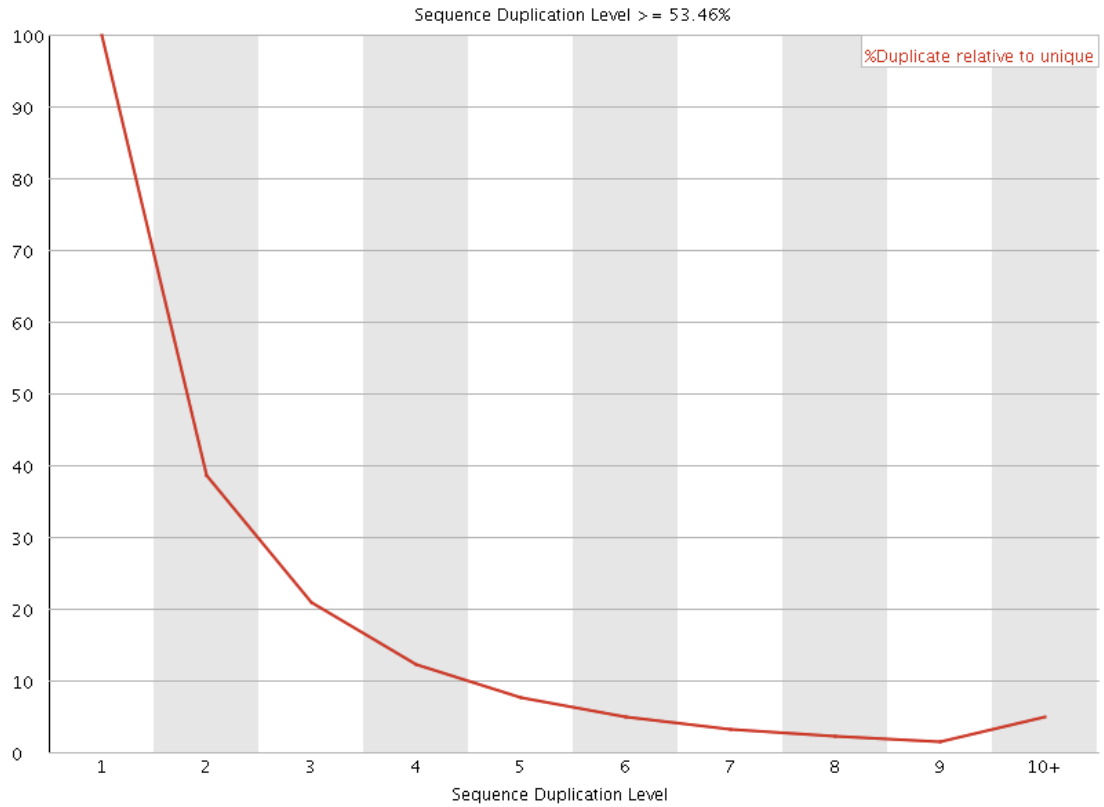




Identification of sequence duplication since ideally, no reads should not start at the same position and have the same errors

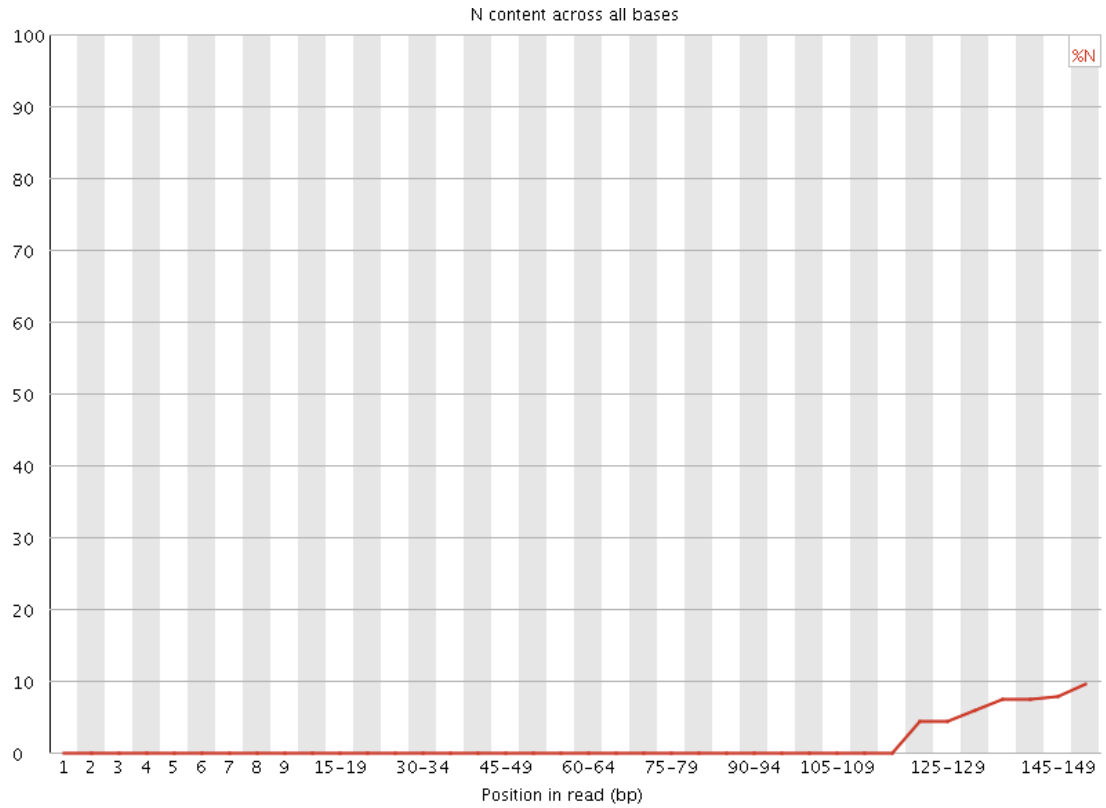
Most likely to indicate some kind of enrichment bias

E.g. duplications can arise during library preparation if there are too few fragments to PCR



Ambiguous bases (Ns) arises when a sequencer is unable to make a base call with sufficient confidence

A high number of Ns can be a sign for a low quality sequence and can cause problems during downstream analysis



Popular QC tools

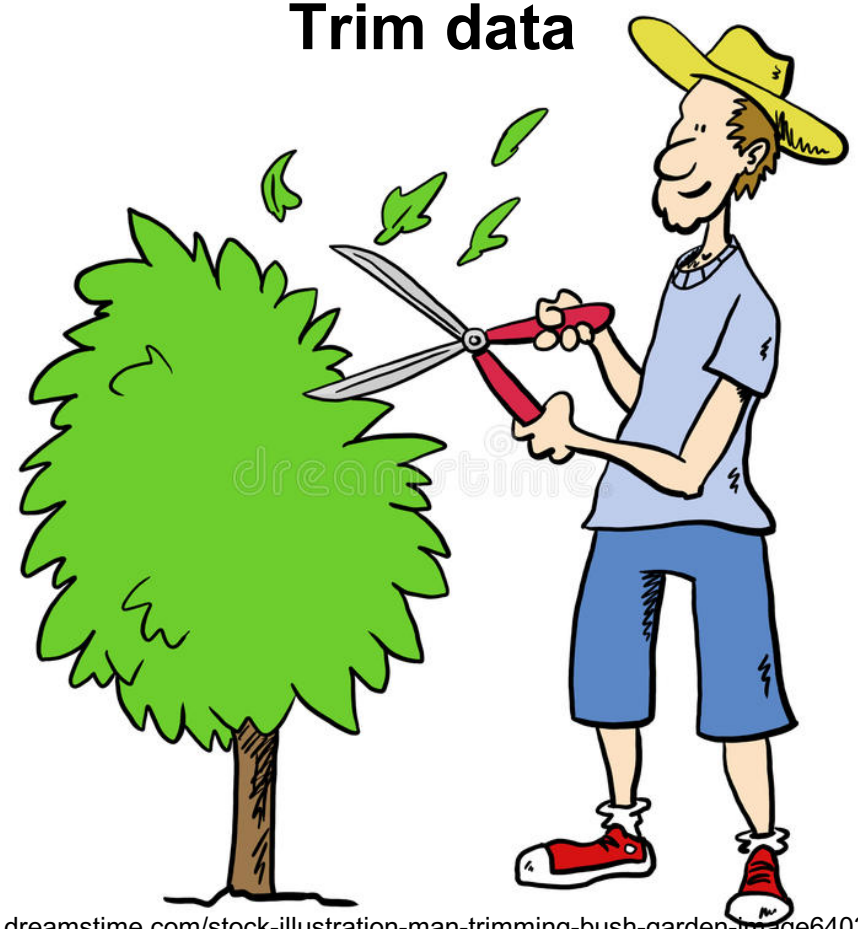
	FastQC	Trimmomatic	Prinseq	Kraken
Standalone	+	+	+	+
Web tool	-	-	+	-
GUI	+	-	+	-
Summary statistics	+	-	+	+
Adaptor trimming	-	+	+	+
Quality trimming	-	+	+	+
Format conversion	-	-	+	
Sequencing technologies	Illumina, PacBio	Illumina	Illumina, 454	Illumina

Why is it important to perform QC and filtering/trimming?

Data analysis also costs money and time

Filter data

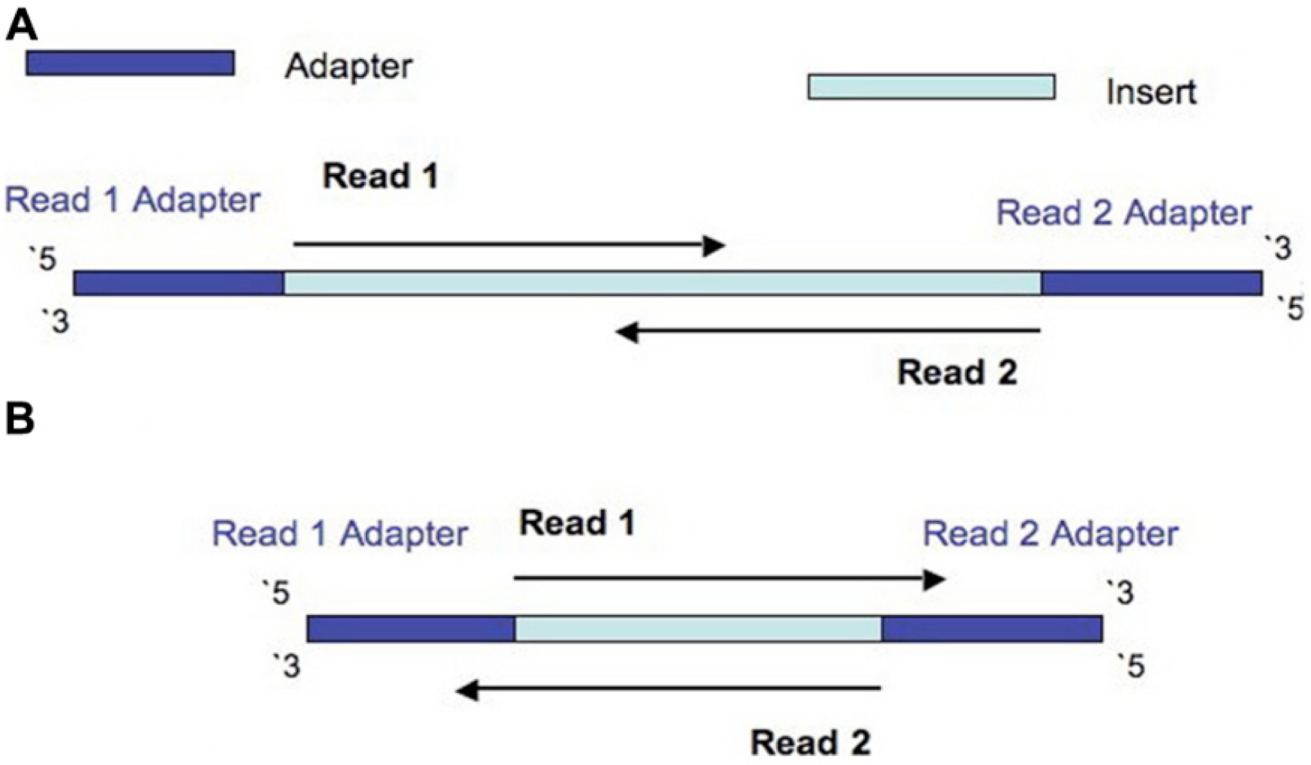
Trim data

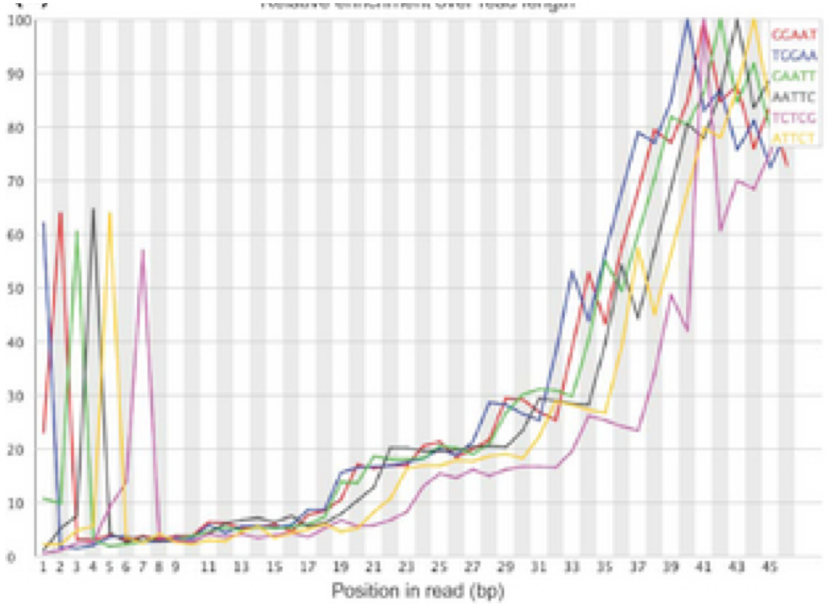


https://www.dreamstime.com/stock-illustration-man-trimming-bush-garden-image64030511

We trim the start and end of reads to remove poor quality data or adapter sequences

If the DNA fragment is shorter than the read length, the sequence reaction will go through the read and into the adapter



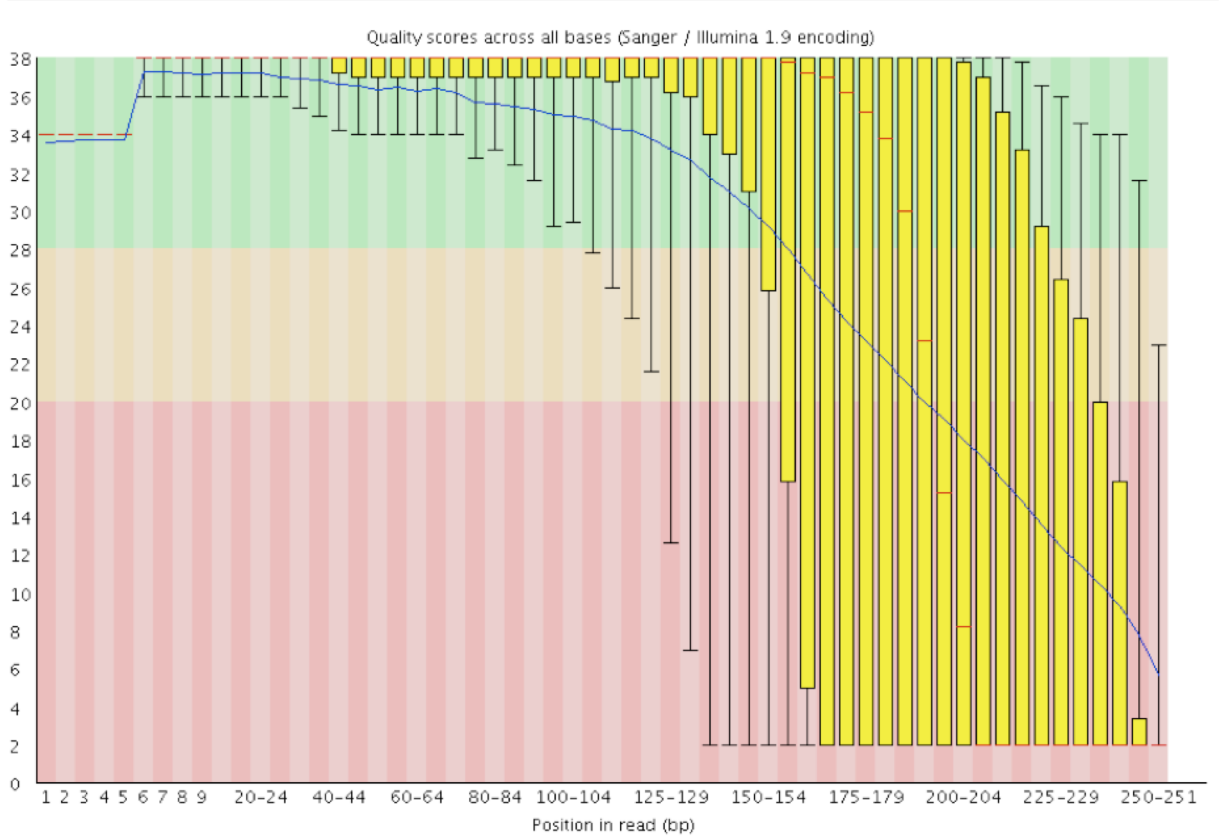


We filter reads to remove poor quality data

Common to set quality thresholds and remove all reads that does no comply

Eg. remove all reads with a lower average Q score than 20, or reads with more than 5 Ns

For example:
Remove all reads with a lower average Q score than 20



For example:
Remove all reads with more than 5 Ns

Popular trimming tools

Many tools available – here are some :

Trimmomatic

CutAdapt

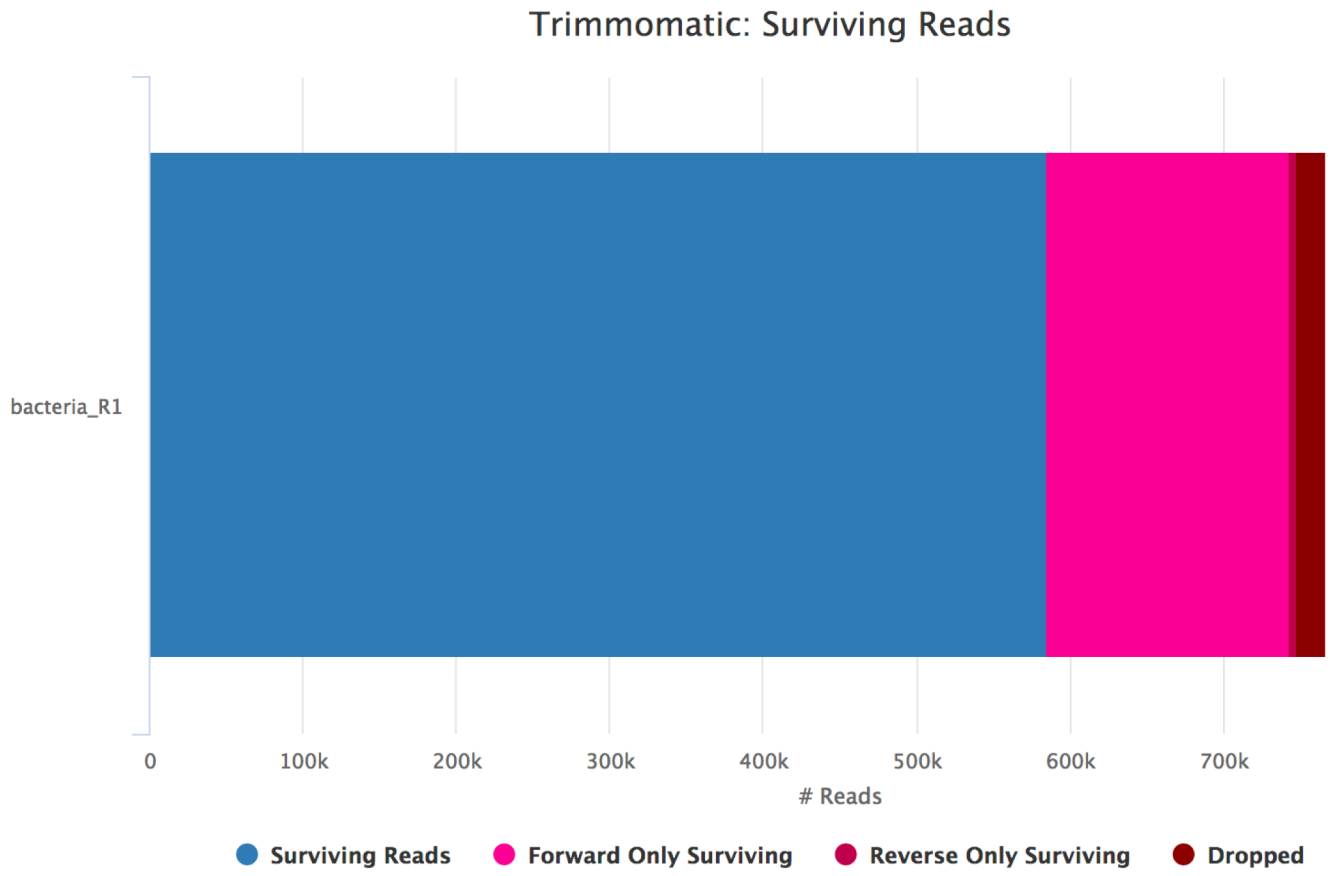
AlienTrimmer

Sickle

Trim Galore

Sycthe

Prinseq



It is important to remove sequence contaminations as early as possible

There can be many sources of contamination in the final sequence library

For example:
PhiX sequences from the sequencing kit

http://losnuevosguerreros.org

For example:
Metagenomic samples may often contain host DNA

Popular decontamination tools

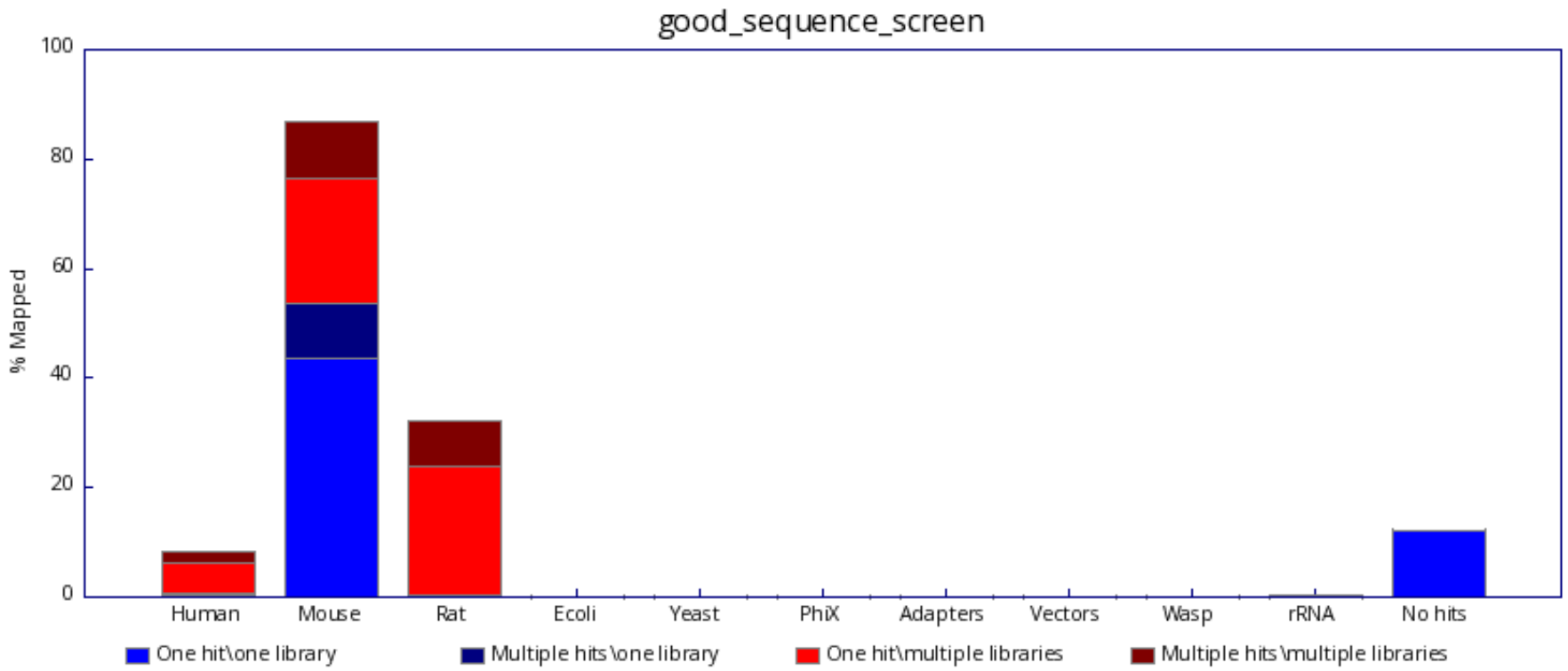
Removal of host contamination:

Fastq Screen

DeconSeq

CS-SCORE

VecScreen



https://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/

Some other typical pre-assembly steps

Merge overlapping paired-end reads - BBmerge