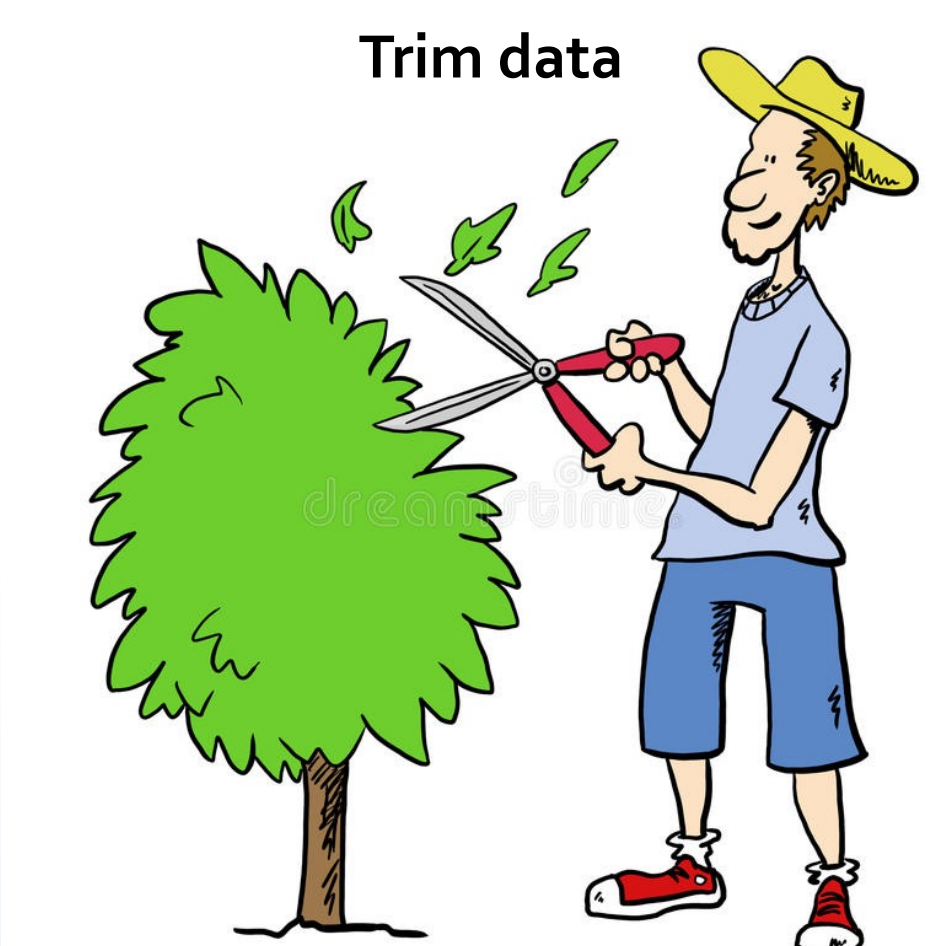
Trimming and filtering

Filter data

© Can Stock Photo - csp34642023

Trim data



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

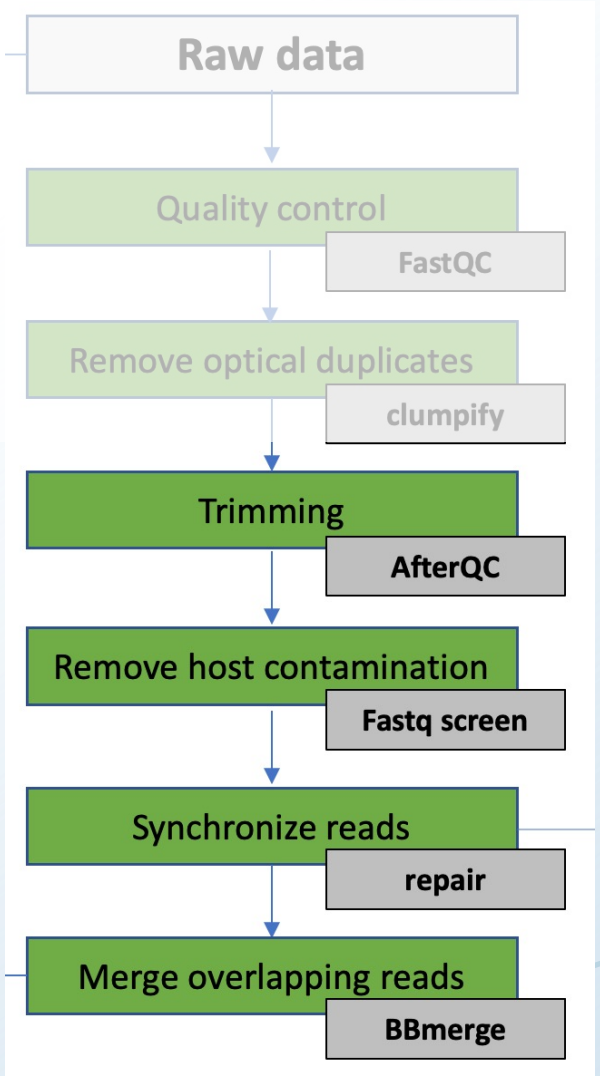
Overview of this talk

Removing poor quality data

Data decontamination

Synchronization of paired-end reads

Merging overlapping reads pairs



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

Data analysis also costs money and time

Filtering data

Length related

Quality score related

GC content related

Ambiguity code related

Sequence complexity related

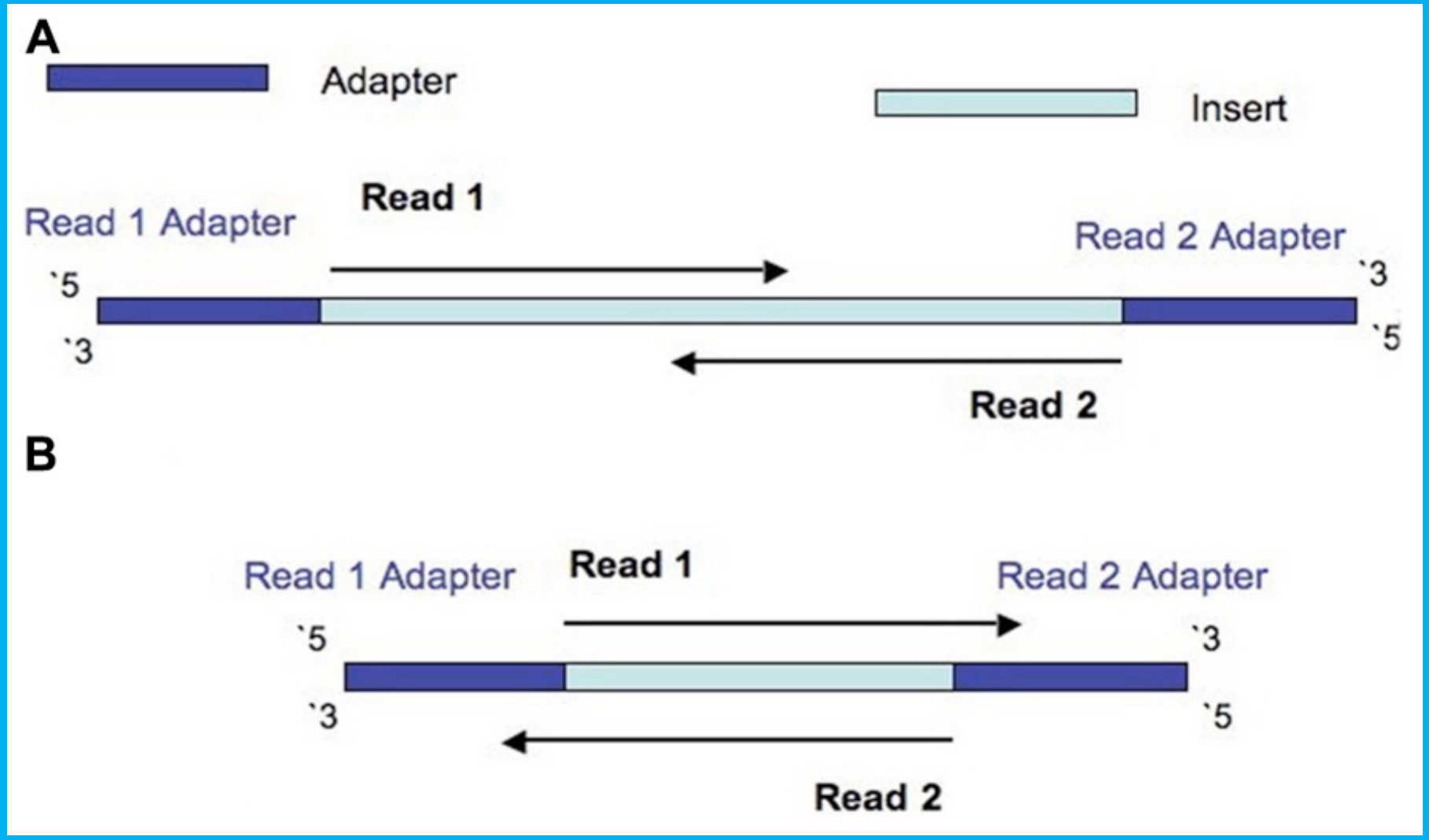
Trimming data

Trim by length/position – fixed, e.g. 20 bp

Trim tails

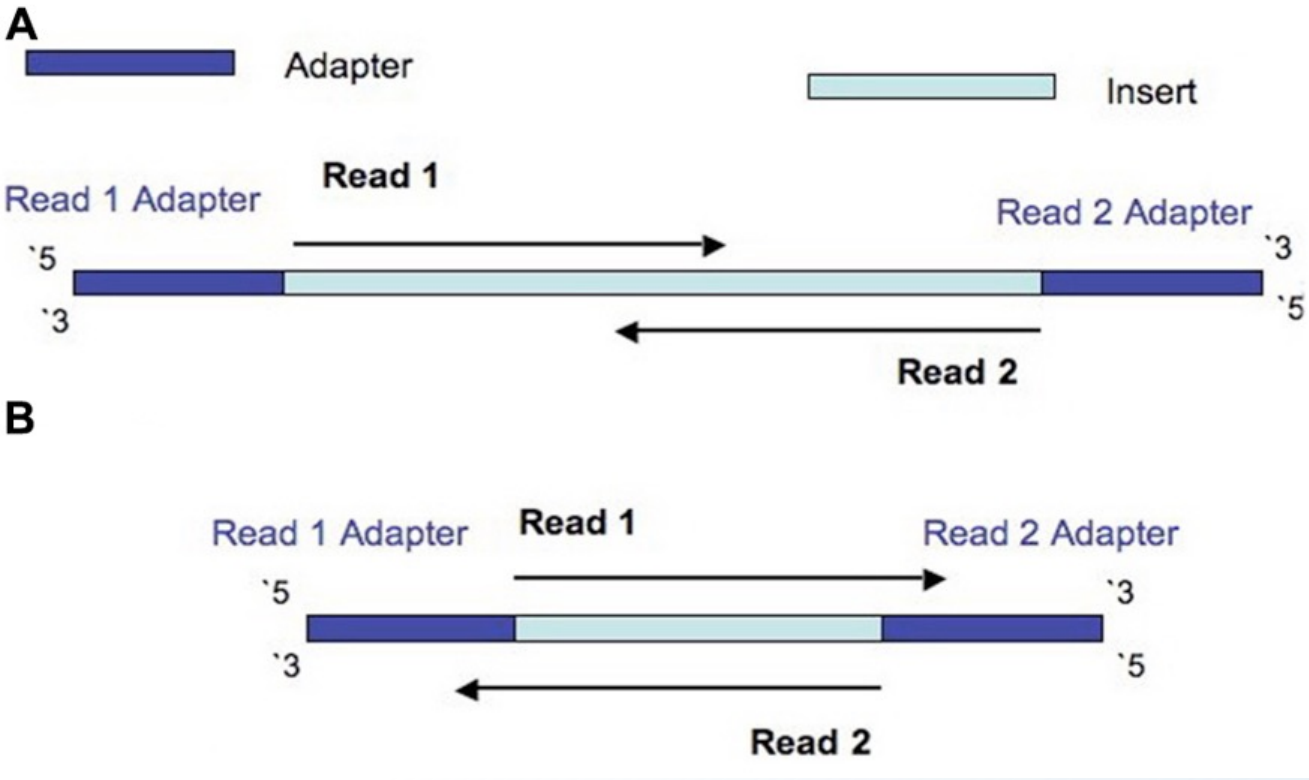
Trim ends by quality scores

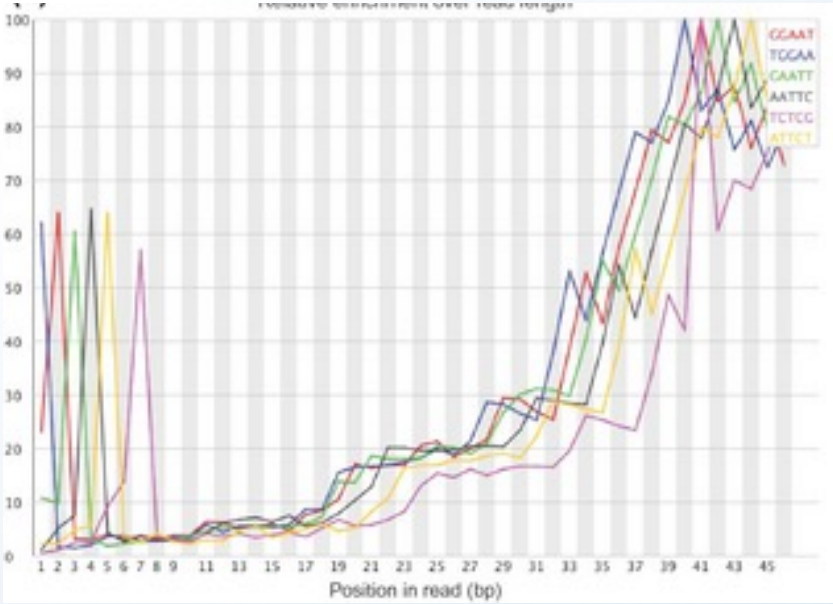
Trimming - Discuss two and two: what is this figure showing?



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

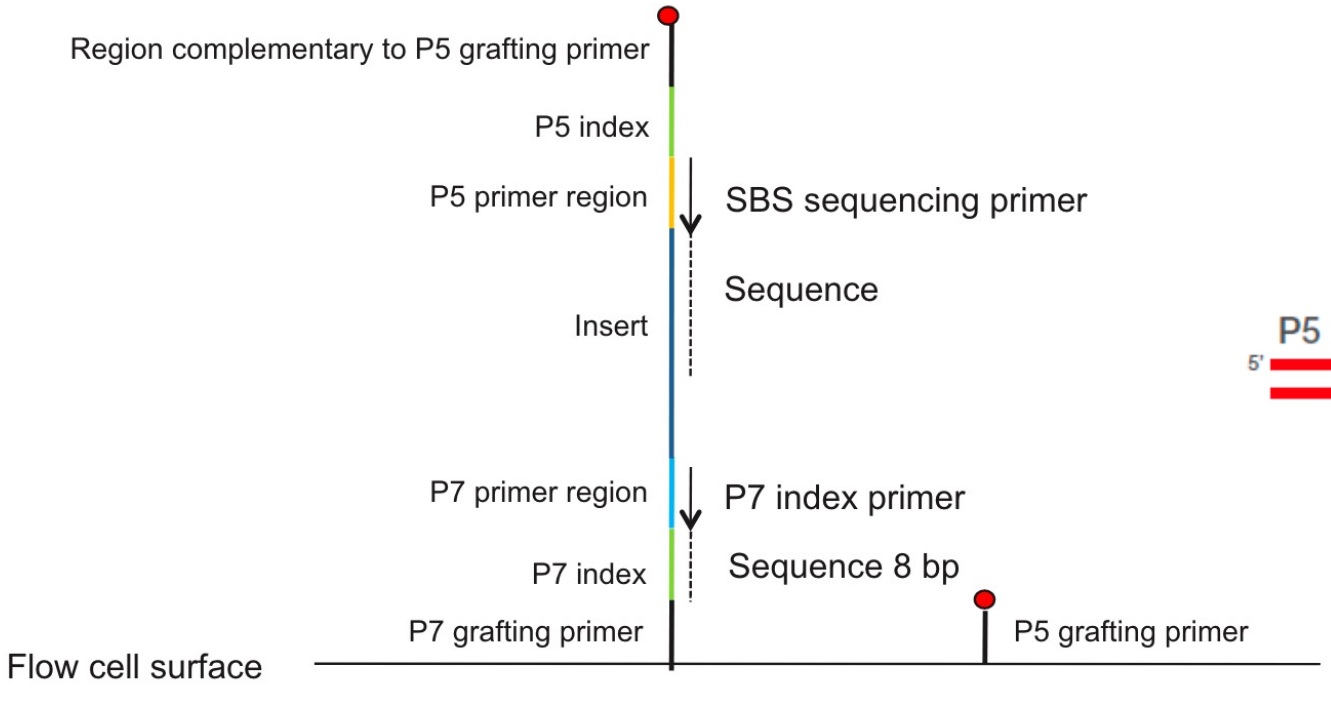
If the DNA fragment is shorter than the read length, the sequence reaction will go through the read and into the adapter

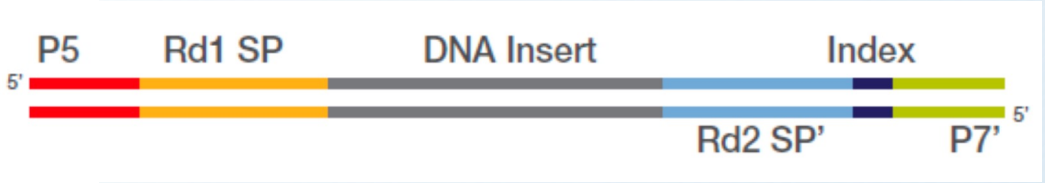




Illumina adapter sequences

When preparing a (TrueSeq) library, adaptors are ligated to the DNA of interest



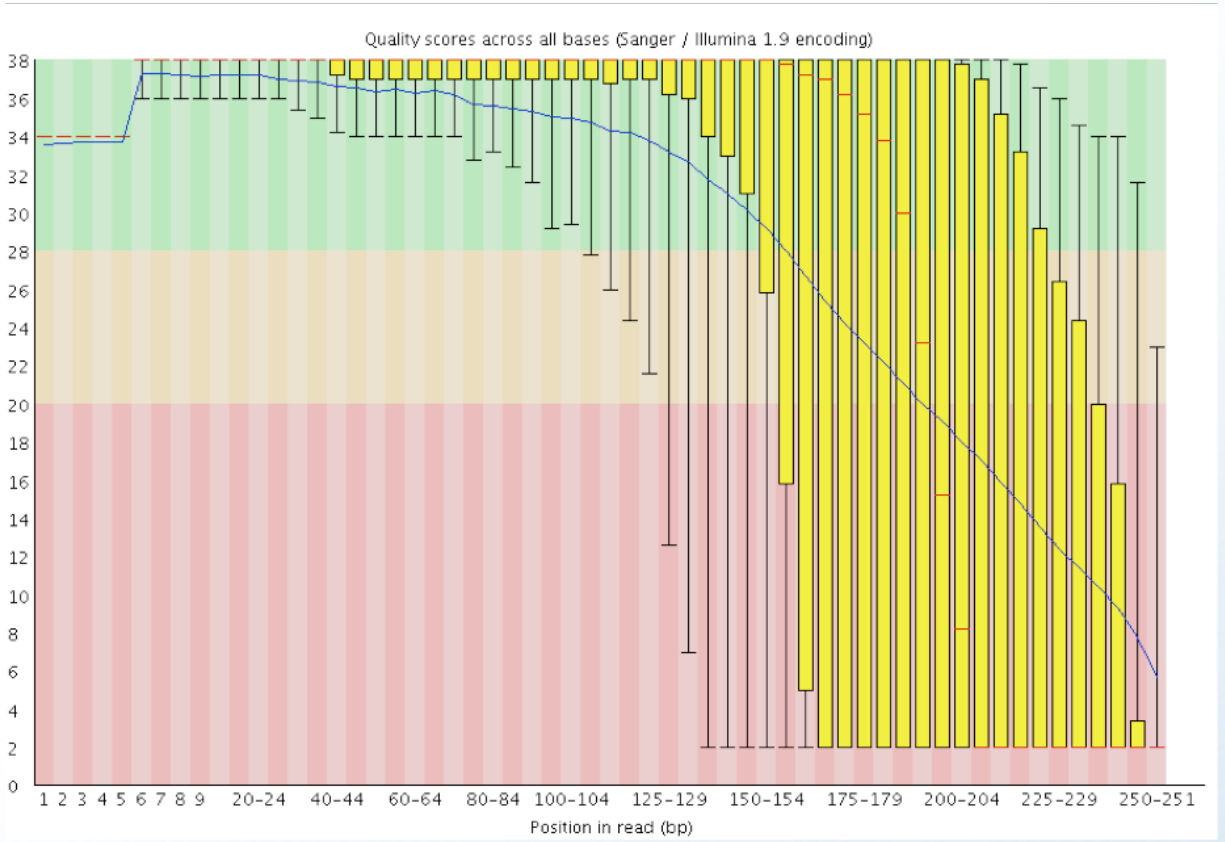


We filter reads to remove poor quality data

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

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

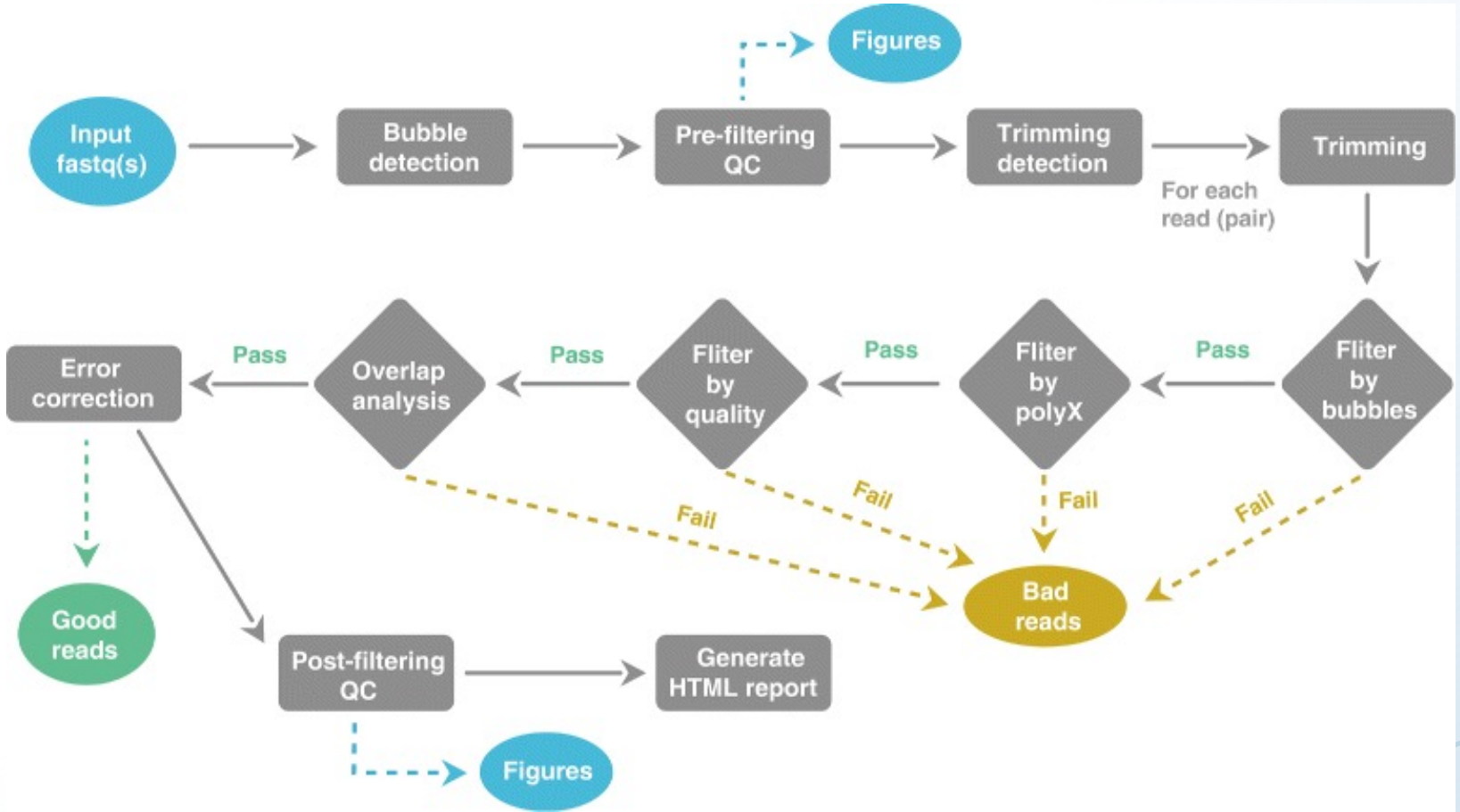
For example:
Remove all reads with a lower average Q score than 20



For example:
Remove all reads with more than 5 Ns

AfterQC - Automatic Filtering, Trimming, Error Removing and Quality Control for FASTQ data

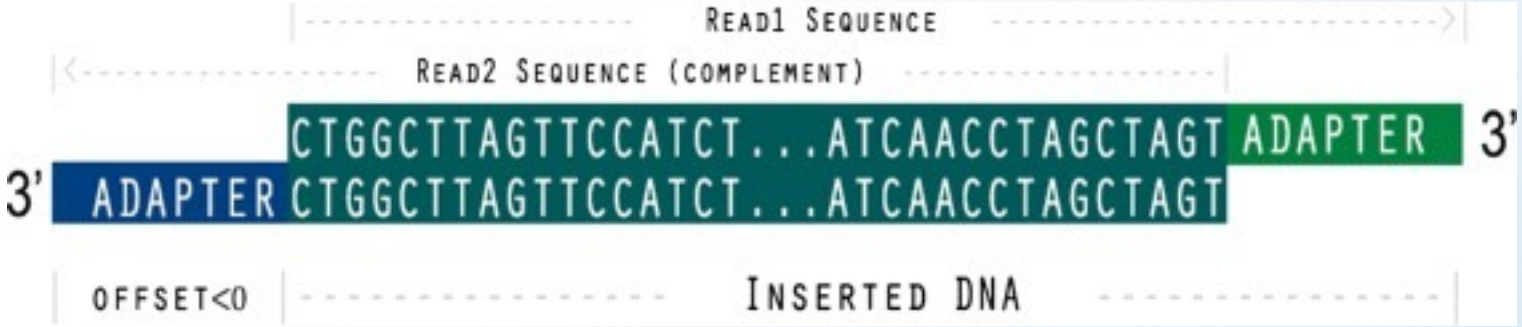
Performs quality control and filtering/trimming of the sequence reads



AfterQC analyses the overlap of paired sequences for pair-end sequencing data

AfterQC will correct the low quality base according to its high quality mate

Removal of 3' adapter in the tail

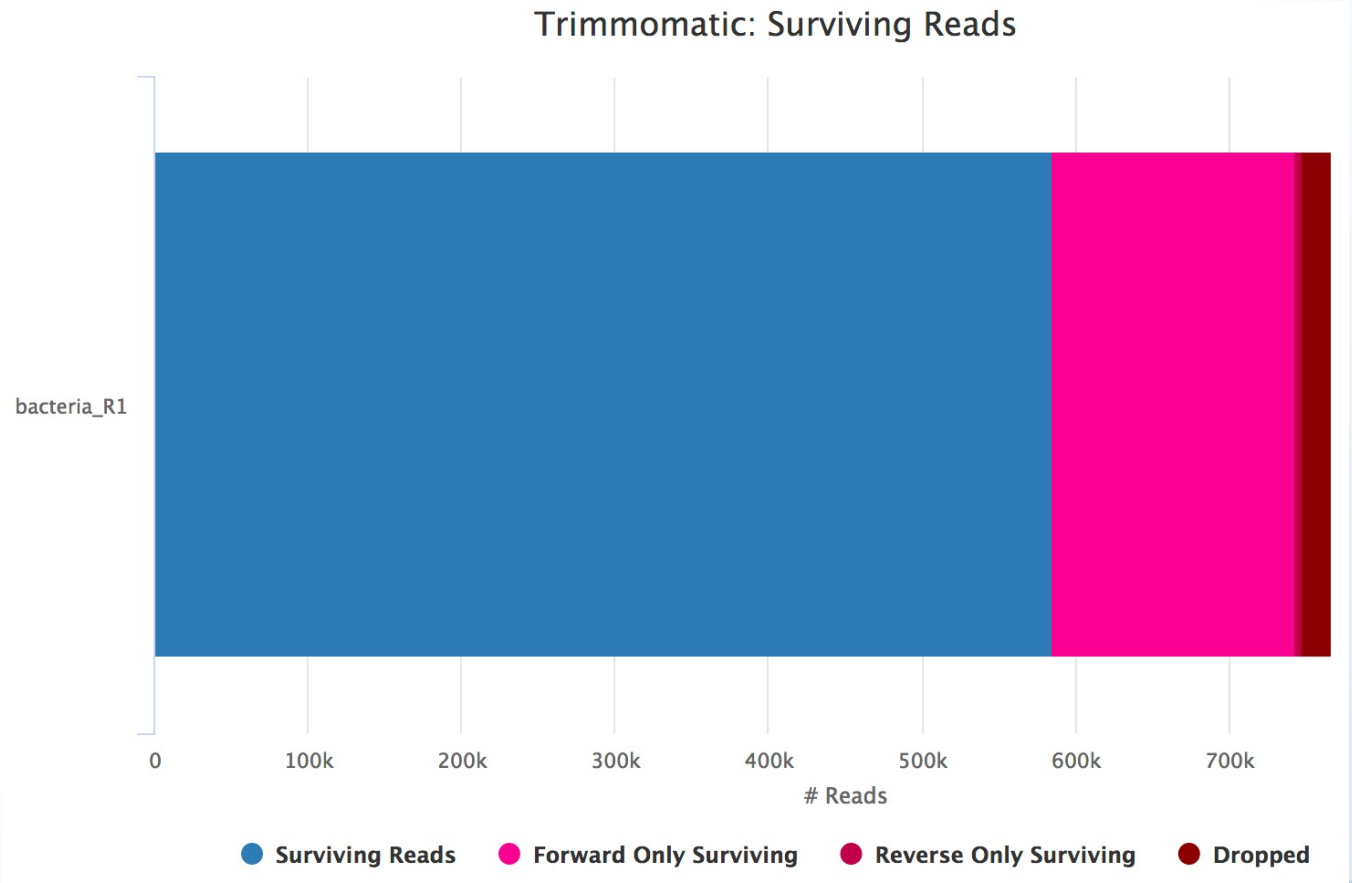


Chen, et al. AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. 2017

Popular trimming tools

Many tools available – here are some :

- AfterQC
- Fastp
- 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 contain sequences from the host

PhiX control

PhiX is used as a quality and calibration control for Illumina sequencing runs

10% of the genomes that are published in literature are contaminated with PhiX

https://www.illumina.com

Commentary | Open Access | Published: 30 March 2015

Large-scale contamination of microbial isolate genomes by Illumina PhiX control

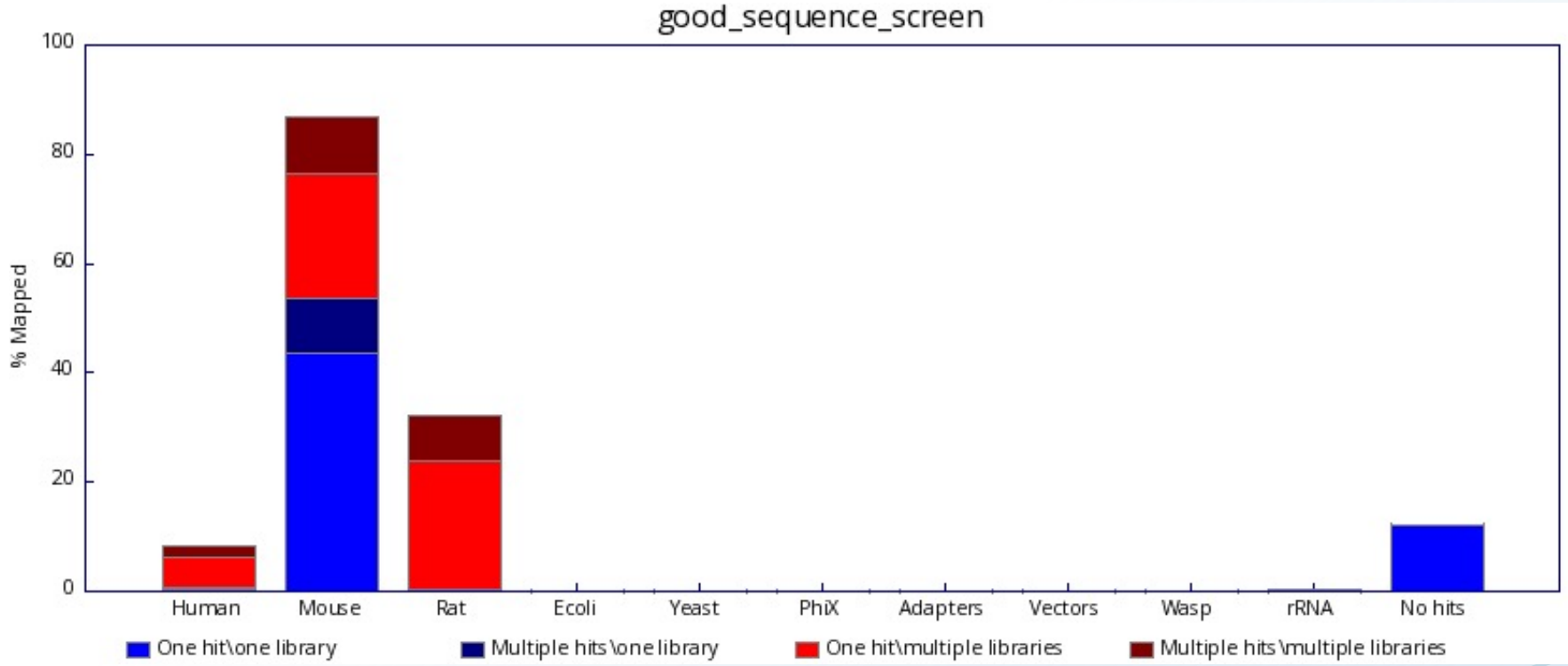
Supratim Mukherjee, Marcel Huntemann, Natalia Ivanova, Nikos C Kyrpides & Amrita Pati

Standards in Genomic Sciences **10**, Article number: 18 (2015) | Cite this article

FastQ Screen allows you to screen sequences against a set of sequence databases

Normally used to screen for host contaminations

Label sequences that match sequences in the database you provide

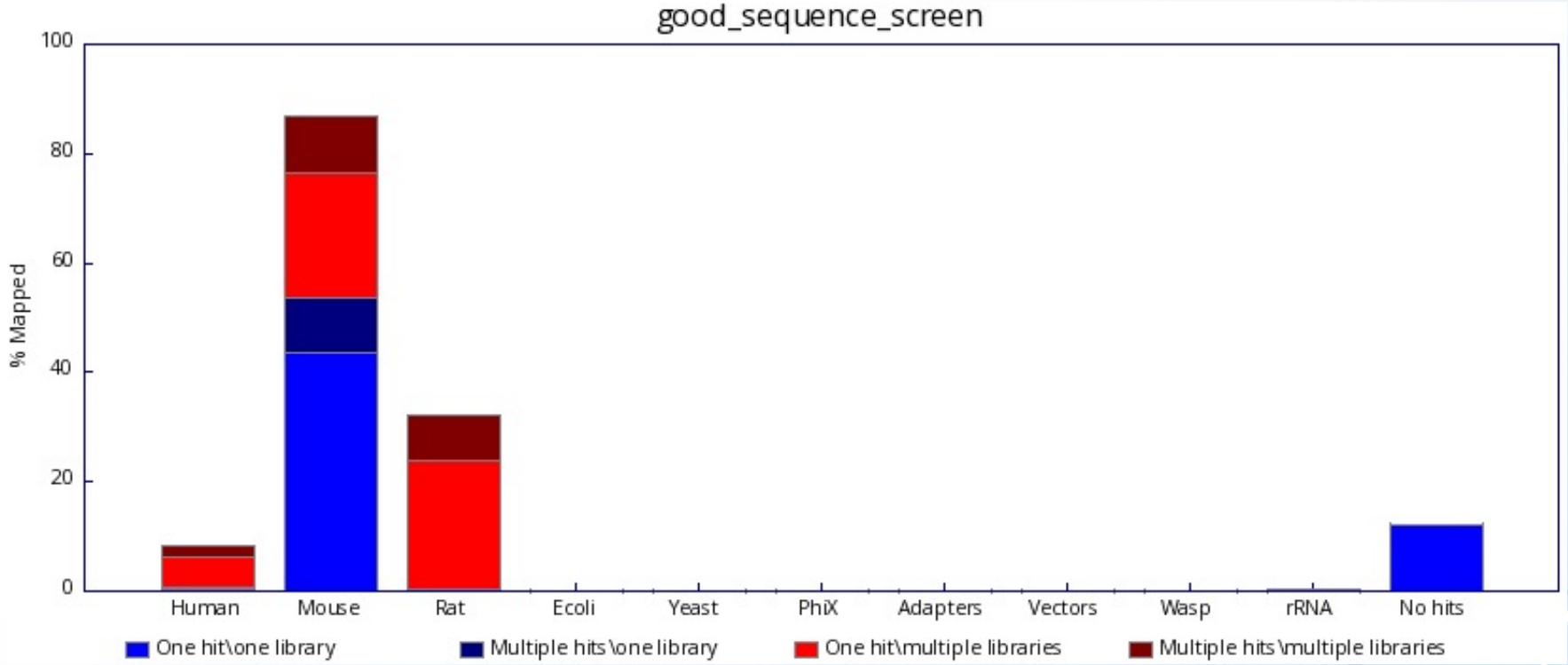


https://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/

Popular decontamination tools

Removal of host contamination

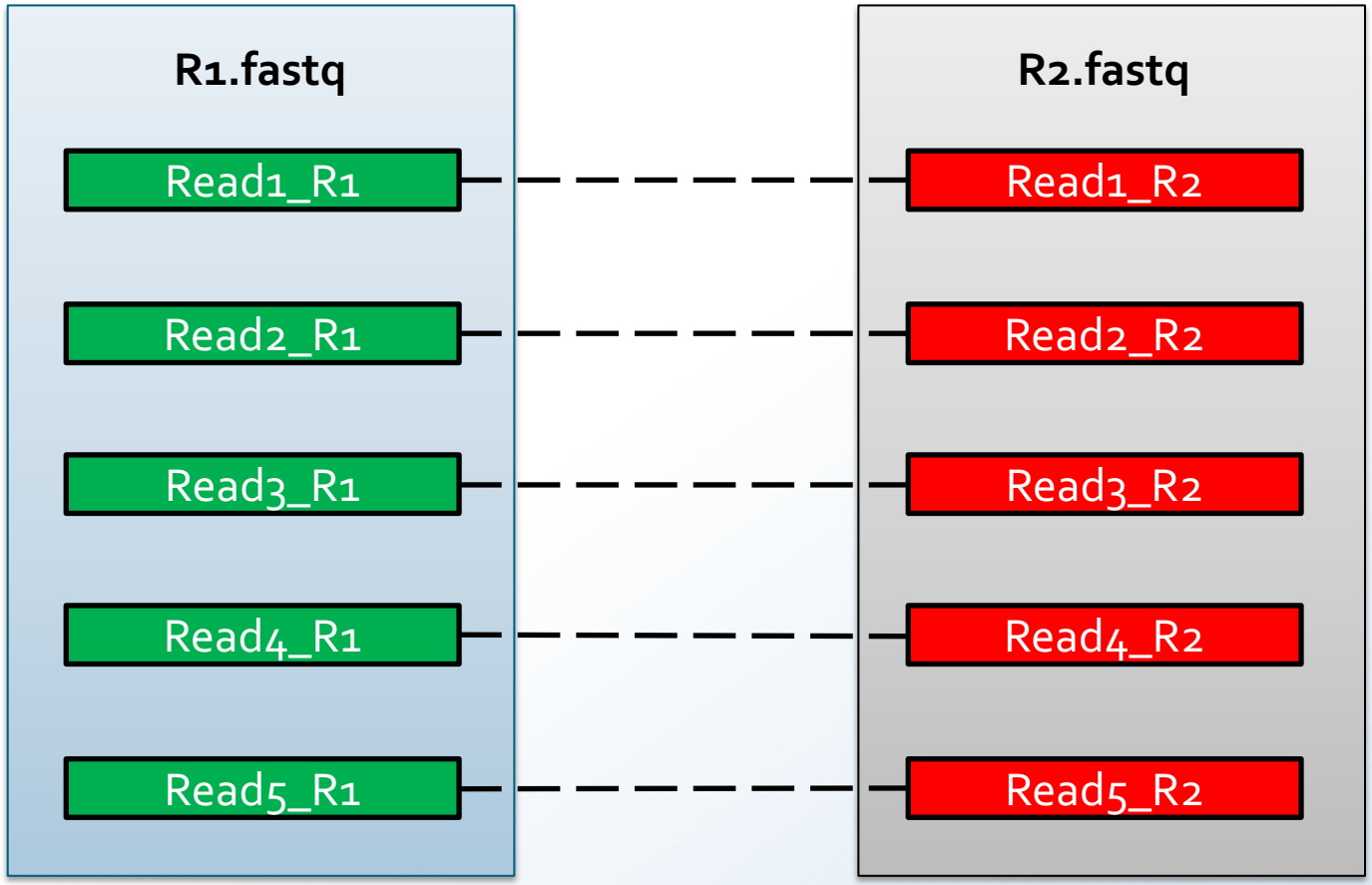
- Fastq Screen
- DeconSeq
- CS-SCORE
- VecScreen
- Kraken



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

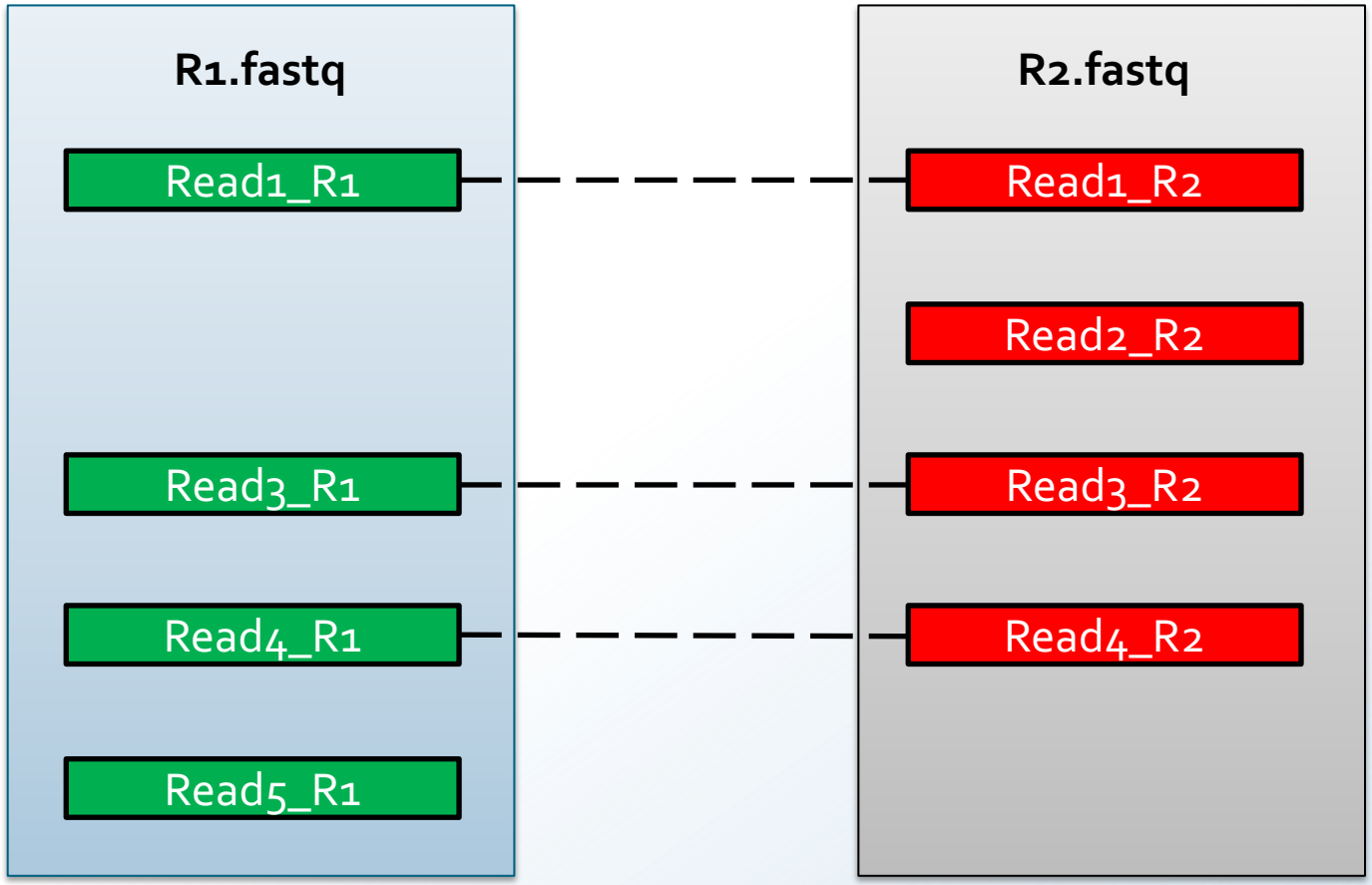
Synchronized FASTQ files are important for many tools

Some tools may remove one of the reads in a read-pair, hence the FASTQ files gets out of sync



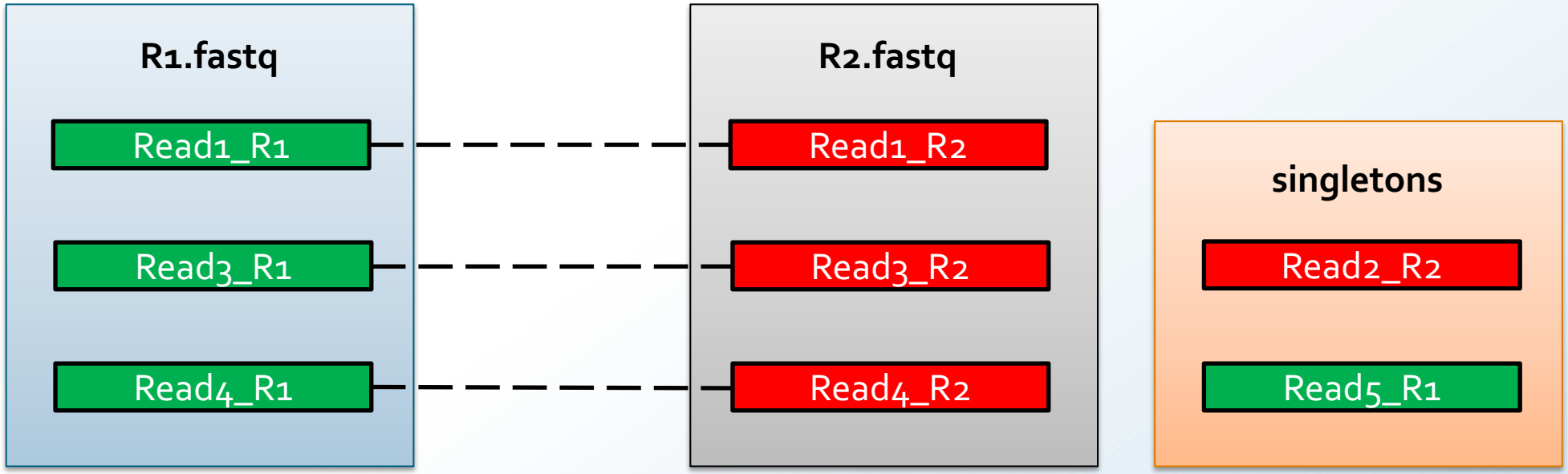
Synchronized FASTQ files are important for many tools

Some tools may remove one of the reads in a read-pair, hence the FASTQ files gets out of sync

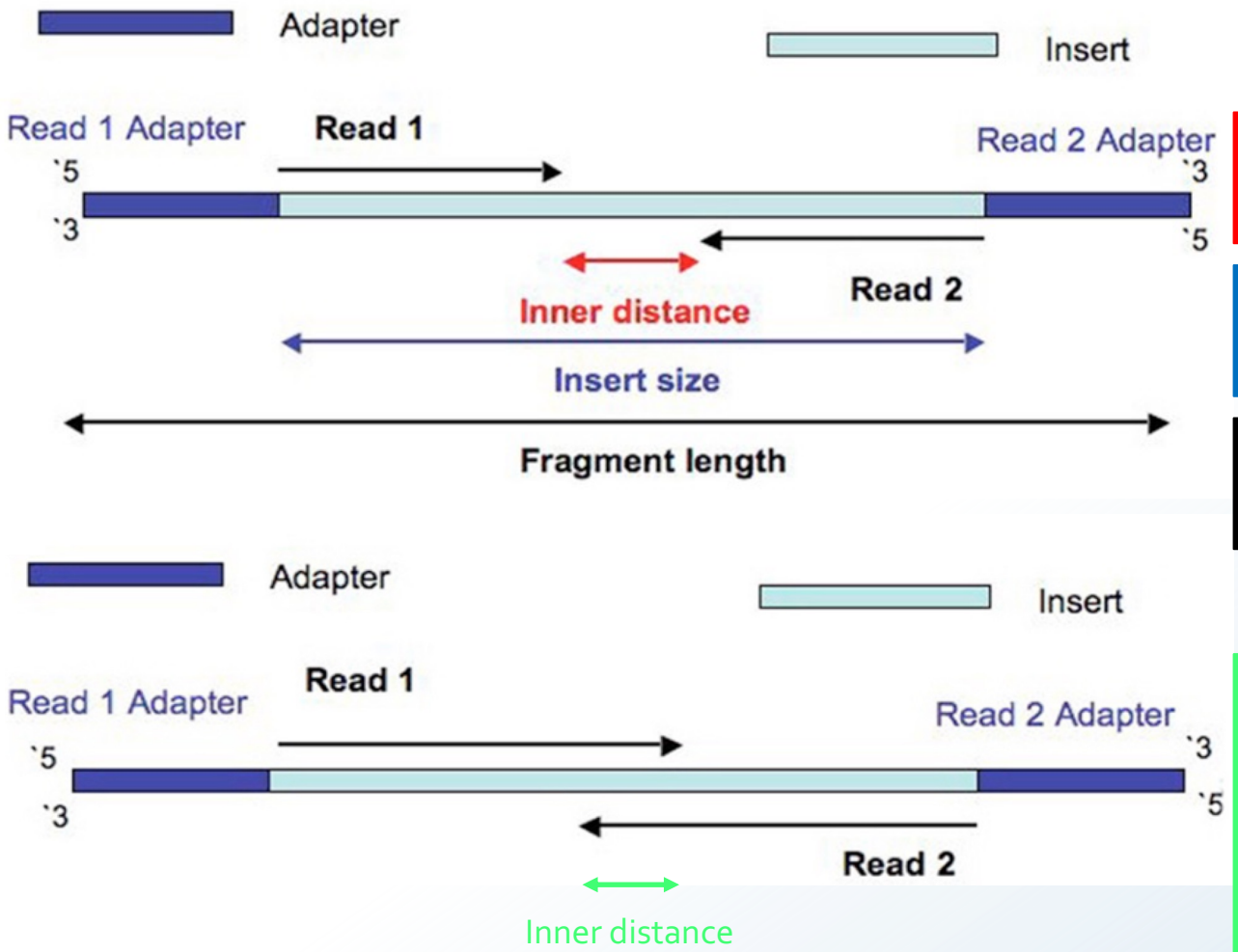


Synchronize FASTQ files with Repair

Repair pulls out the "singletons" and produce synchronized paired-end FASTQ files



DNA insert sizes and overlapping read pairs



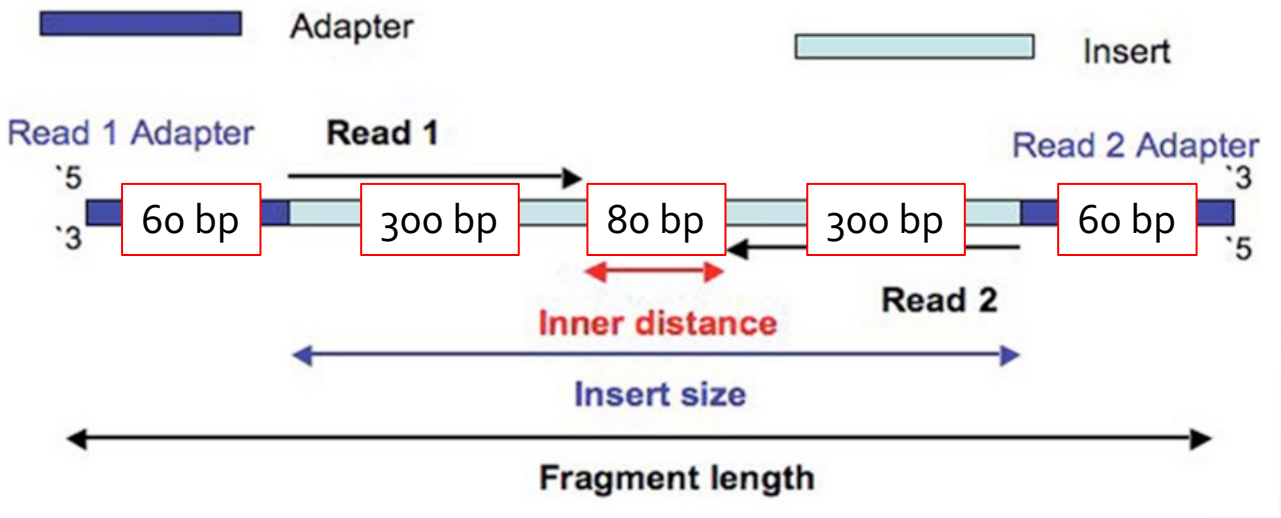
Inner distance is the unsequenced middle piece

Insert size is stretch of sequence between the paired-end adapters

Fragment size includes the ligated Illumina adapters

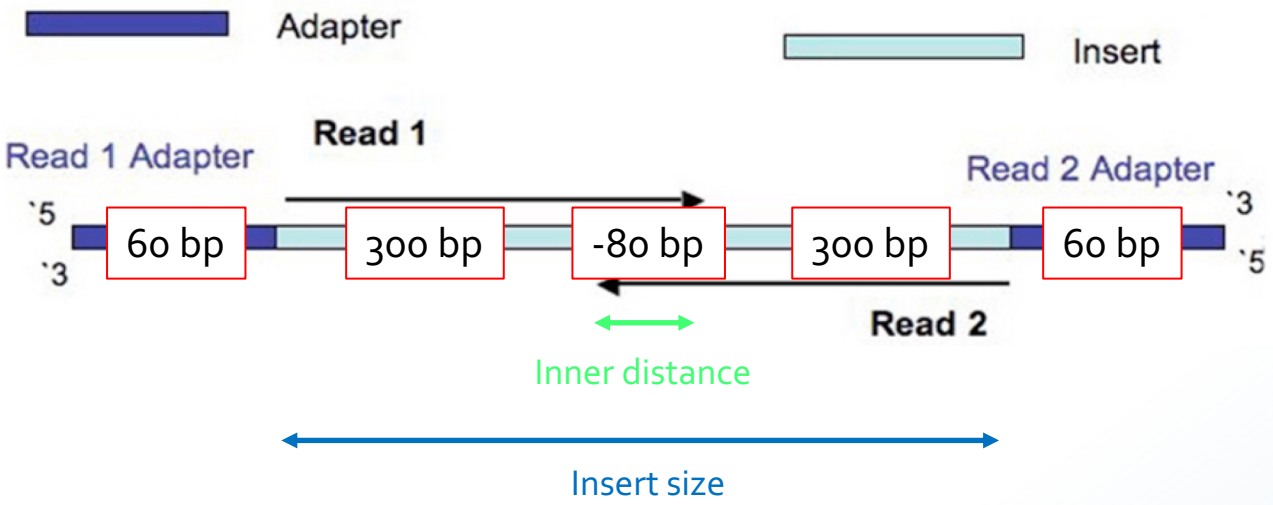
When read pairs overlap, the inner distance is a negative number
or
a positive number describing the overlap

DNA insert sizes and overlapping read pairs



Inner distance = 80 bp
Insert size = R1 + R2 + inner distance = 680 bp
Fragment size = R1 + R2 + inner distance + adapters (x2) = 800 bp

Insert size in BBMerge is a bit confusing since the inner distance for overlapping reads is negative

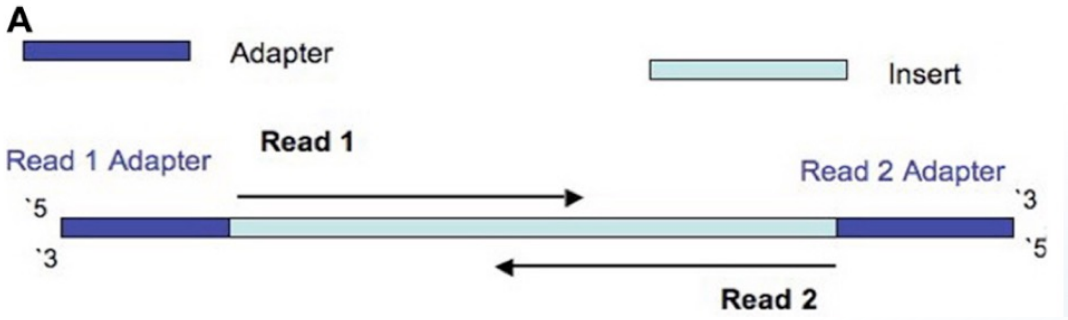


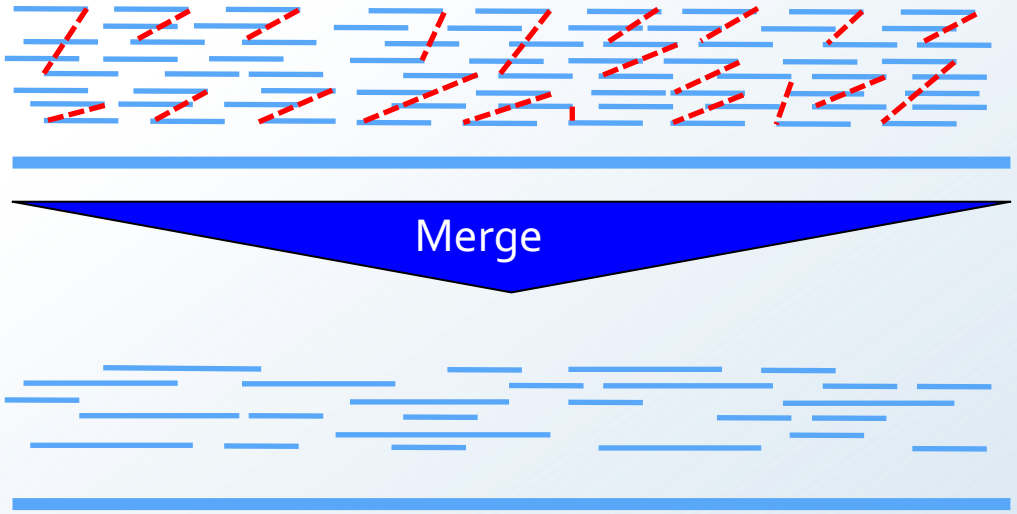
#InsertSize	Count (reads)
…	…
101	10248
102	10619
103	10397
104	10357
105	10218
106	10435
107	10288
108	9973
…	…

Inner distance = BBMerge Insert size = -80
Insert size = R1 + R2 + BBMerge Insert size = 520 bp
Fragment size = R1 + R2 + inner distance + adapters (x2) = 640 bp

Generate longer reads by overlapping and merging read pairs before assembling a genome using BBmerge

Merging reads will reduce computational costs and improve the assembly





Generate longer reads by overlapping and merging read pairs before assembling a genome using BBmerge

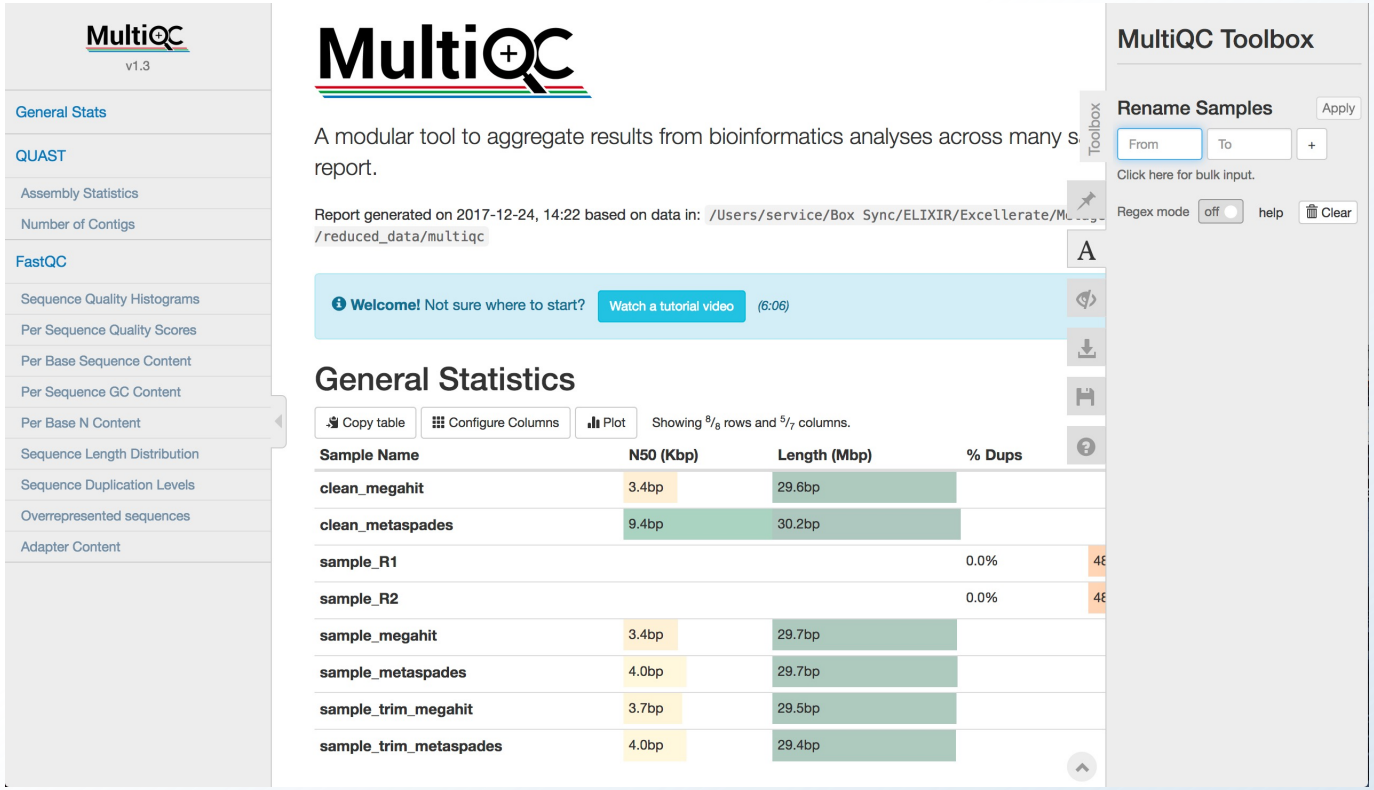
Merging reads will improve the quality of the reads and generate longer reads

Longer reads allow the use of longer k-mers or fewer comparisons

Program	NA50 (bp)	Total Misassemblies	Indels/ 100 kbp	Genome Completeness (%)
Raw Data	60007	119	1.13	84.5
BBMerge	102577	127	0.84	84.88
BBMerge-REM	119328	117	0.81	85.18
BBMerge-RSEM	104441	115	0.84	84.88
COPE	89603	294	1.52	85.17
COPE-M3	98240	227	1.24	83.92
fastq-join	80672	183	1.17	84.74
FLASH	94846	282	1.41	85.20
leeHom	101992	290	1.1	84.91
PEAR	60937	660	1.46	84.28
Stitch	5623	20986	47.78	68.38
USEARCH	102156	131	0.88	84.77
XORRO	97403	158	1.08	84.85

https://doi.org/10.1371/journal.pone.0185056.t003

Generate report and show to your boss☺

MultiQC is a reporting tool that parses summary statistics from results and log files generated by other bioinformatics tools

Parses relevant information from log files to a HTML report file