Quality Filtering

Niklaus Zemp
20 June 2025

Genetic Diversity Centre (GDC)
Bioinformatics
ETH Zurich

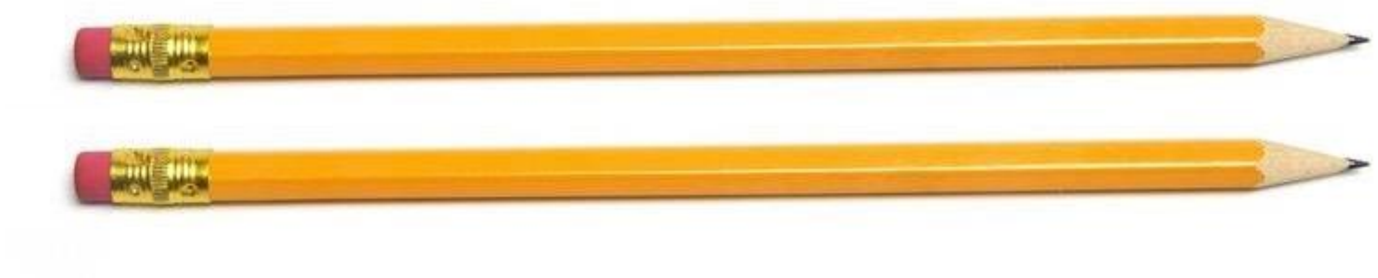
Sequencing technologies

Short read- (Ilumina, Aviti)

Low error rate

Quality filtering

Long read – (PacBio, ONT)



High error rate

Error correction

Check your data

Quality control

FastQC High Throughput Sequence QC Report

Version: 0.10.1

www.bioinformatics.babraham.ac.uk/projects/

© Simon Andrews, Babraham Bioinformatics, 2011

Picard BAM/SAM reader ©The Broad Institute, 2009

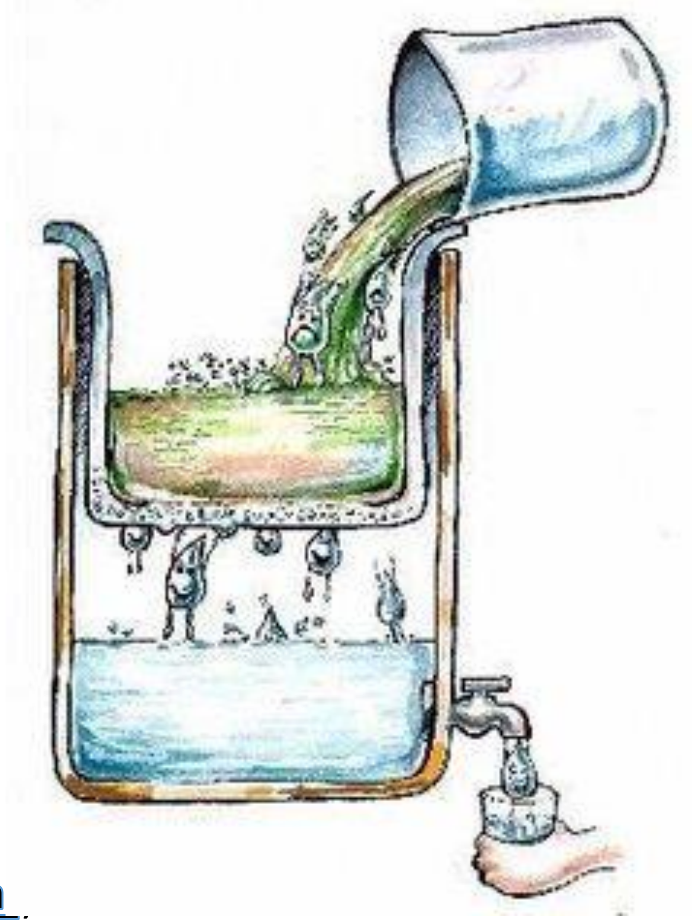
BZip decompression ©Matthew J. Francis, 2011

Use File > Open to select the sequence file you want to check

FastQ Screen

Contamination screening for NGS data

Tools for quality filtering



FASTX-toolkits (http://hannonlab.cshl.edu/fastx_toolkit)

PRINSEQ (http://prinseq.sourceforge.net/)

Cutadapt (http://cutadapt.readthedocs.io/en/stable/guide.html)

Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic)

Adapterremoval (https://github.com/MikkelSchubert/adapterremova

Fastp (https://github.com/OpenGene/fastp)

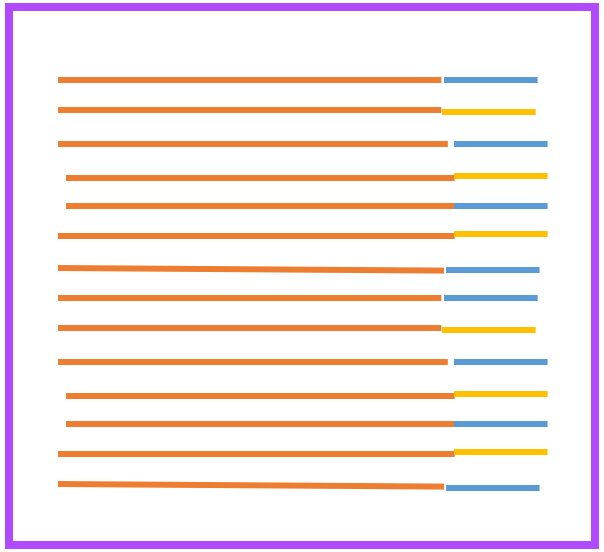
bbmap (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmap-guide/)

…

Demultiplexing

Normally done by the Illumina software

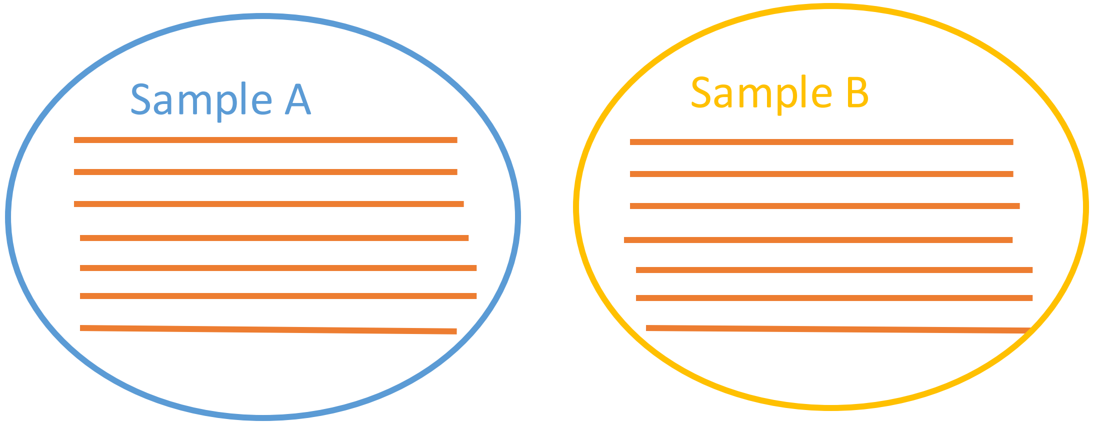
- A low number of reads is always wrongly inferred



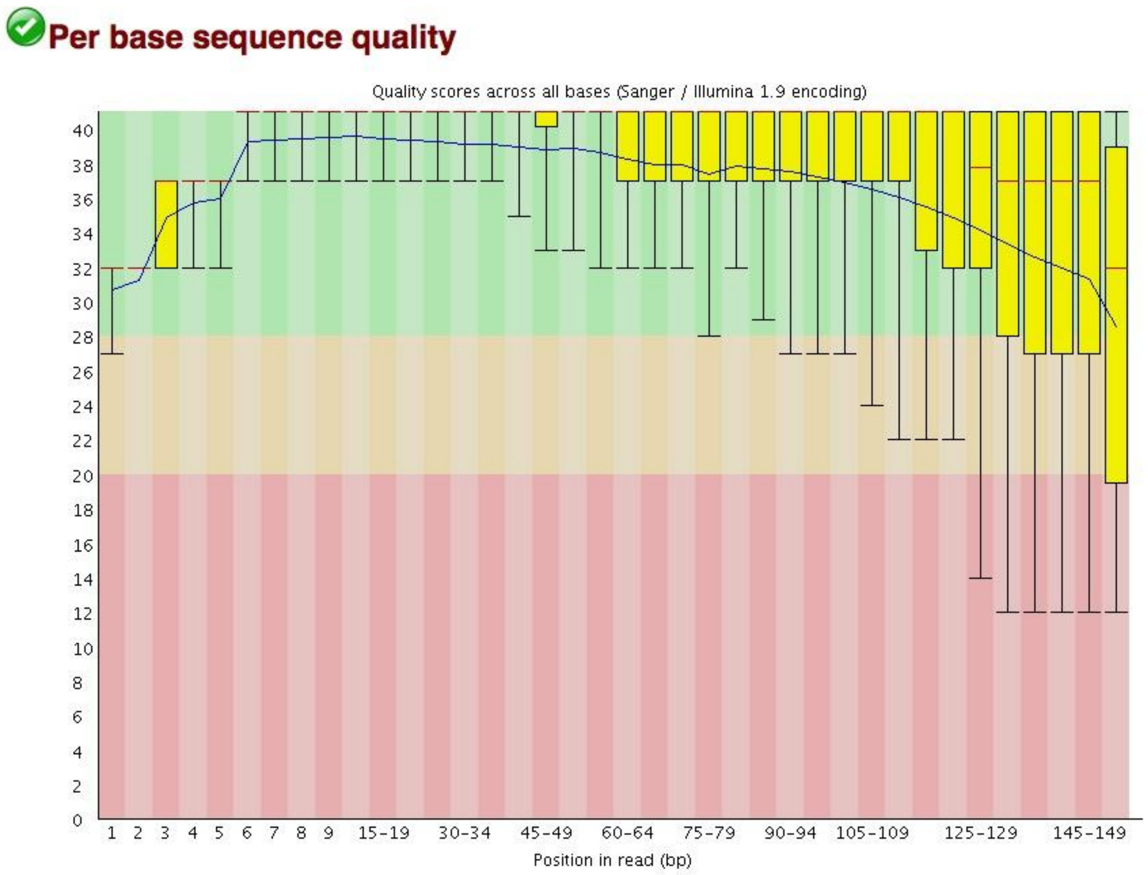
-> Rare events are more affected

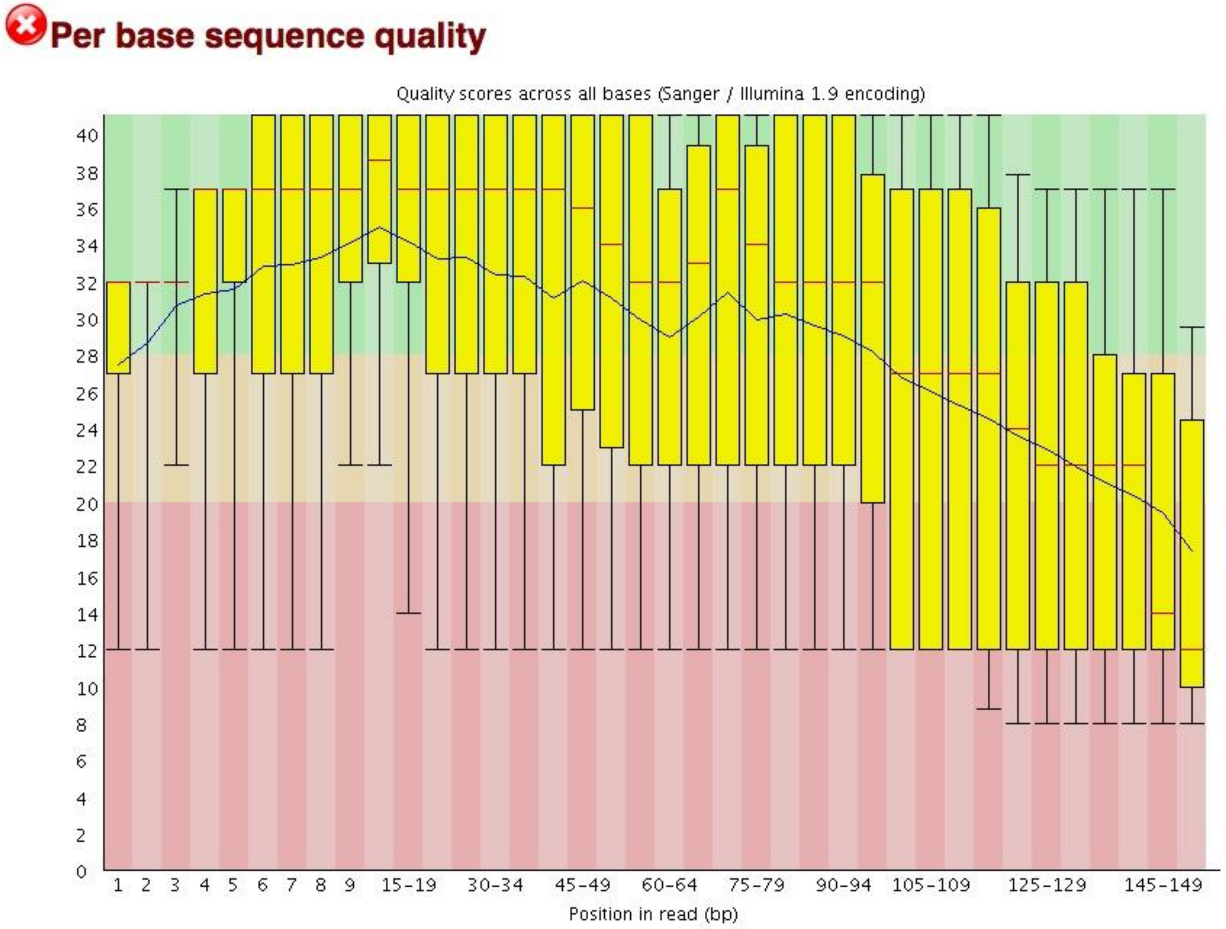
-> Use replicates

-> Use unique dual barcodes



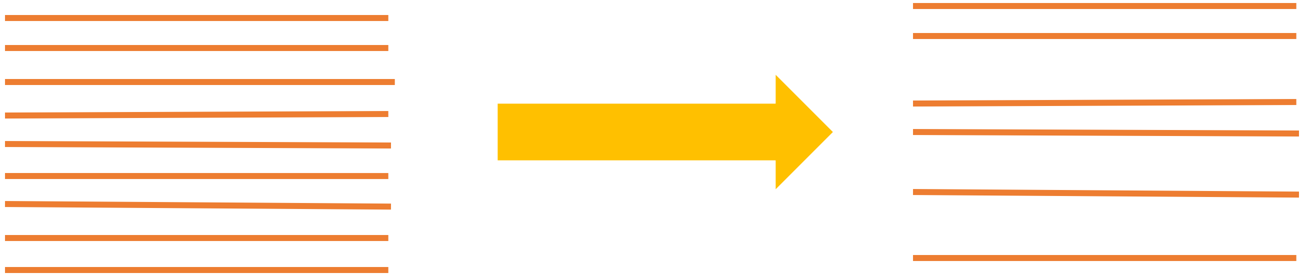
Per bases sequence quality


✓ Per base sequence quality

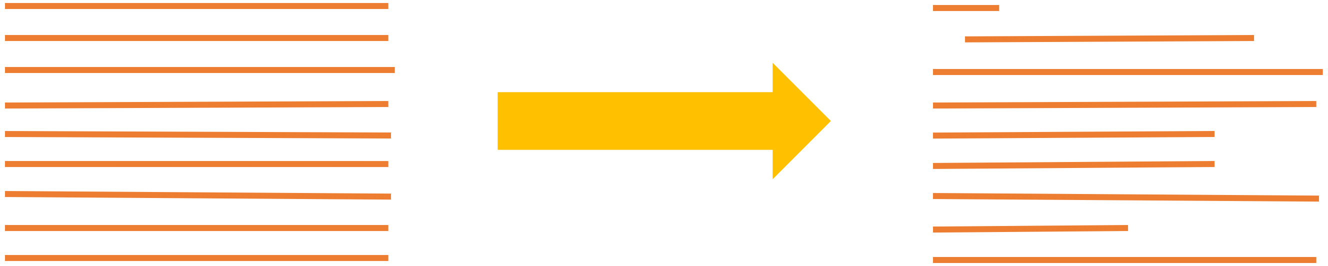

✗ Per base sequence quality

Filtering and/or trimming

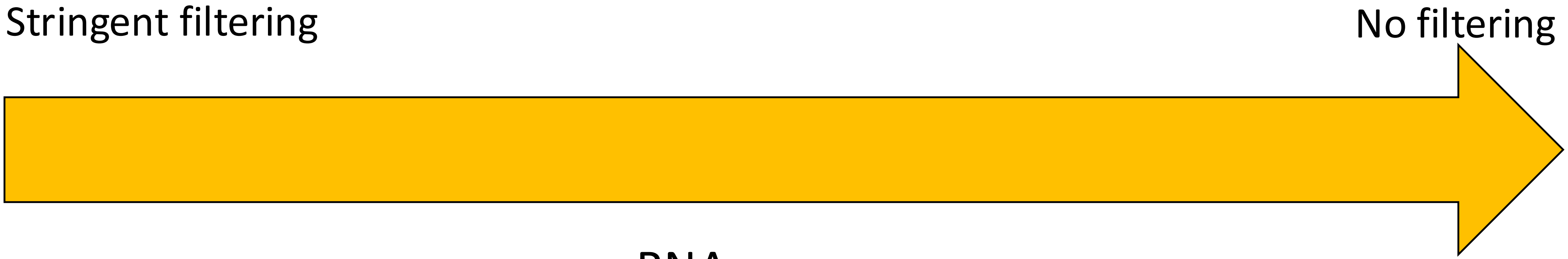
Filtering



Trimming



How stringent do we need to be?



Stringent filtering

No filtering

RNA-seq

de novo assembly

AmpSeq

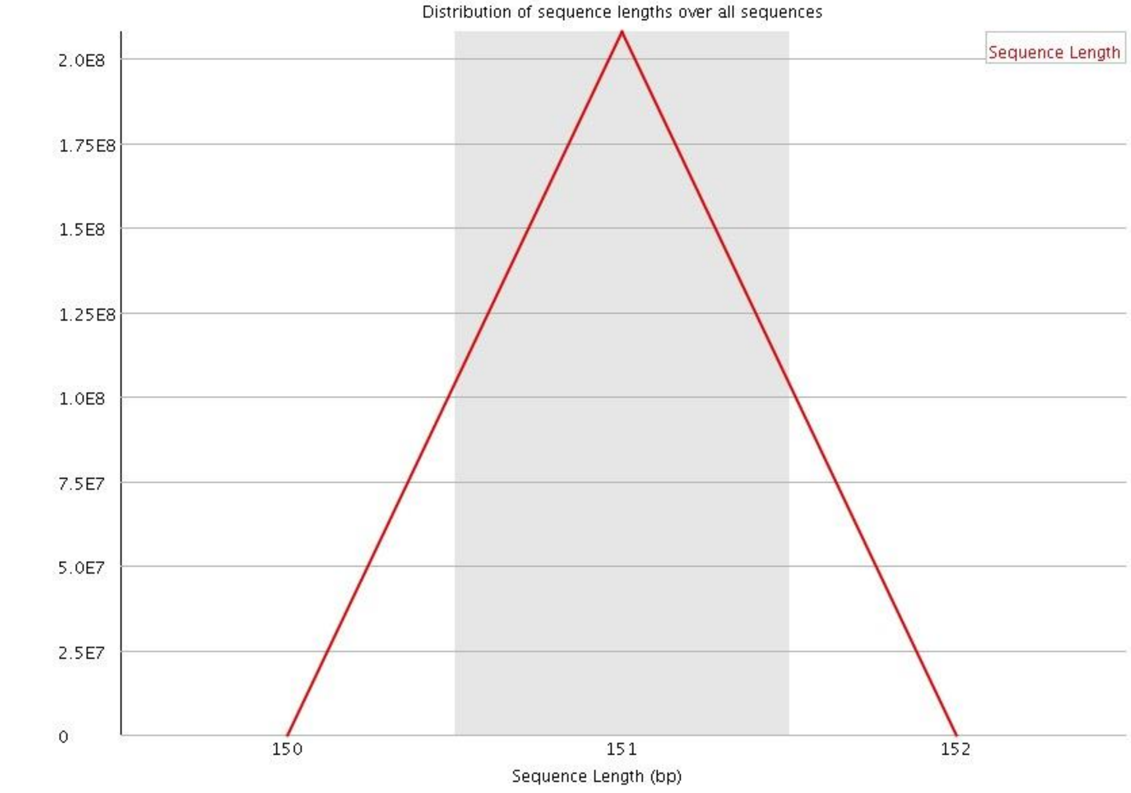
Ancient DNA

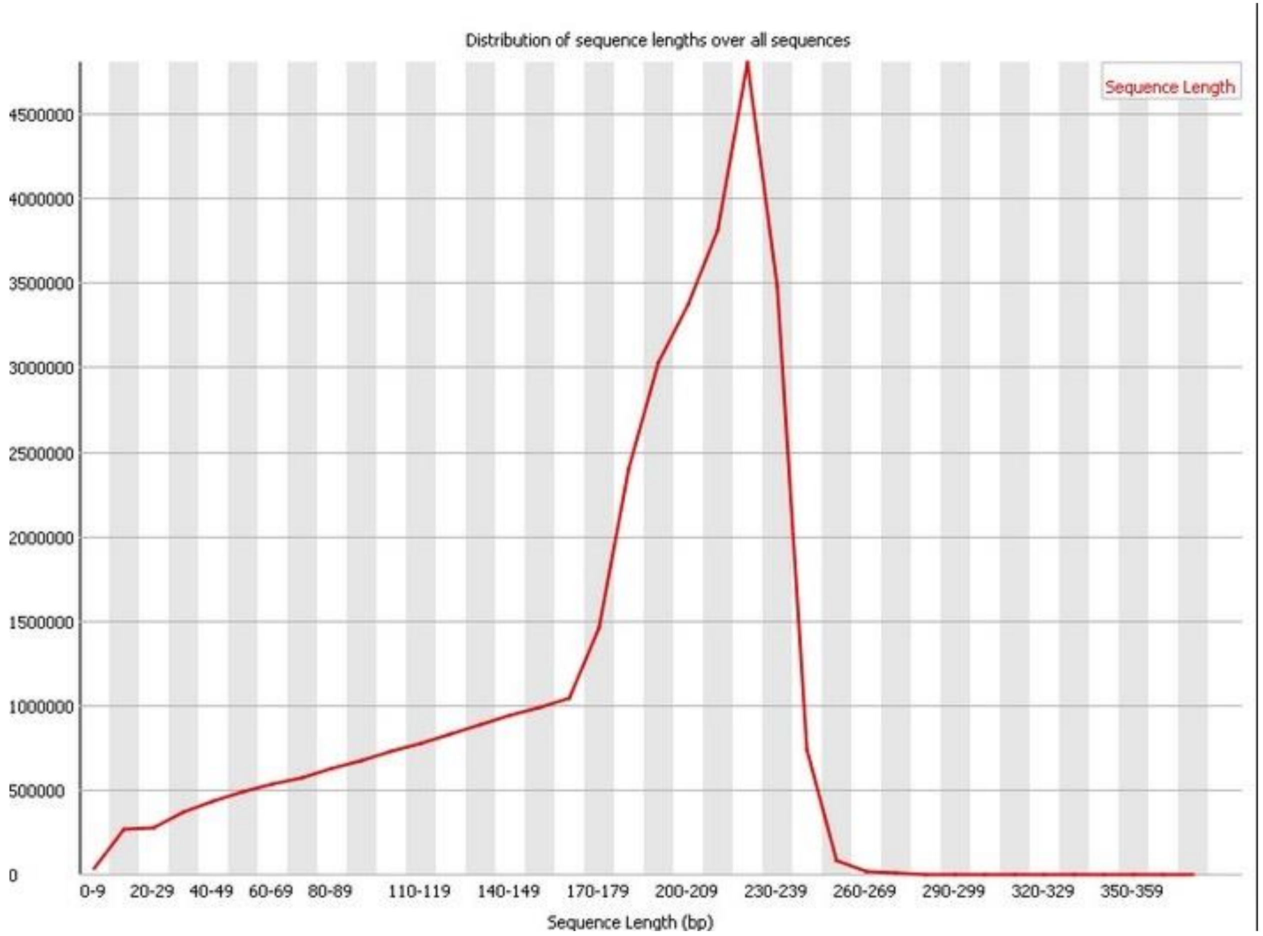
Re-sequencing

RAD-seq

Sequence length

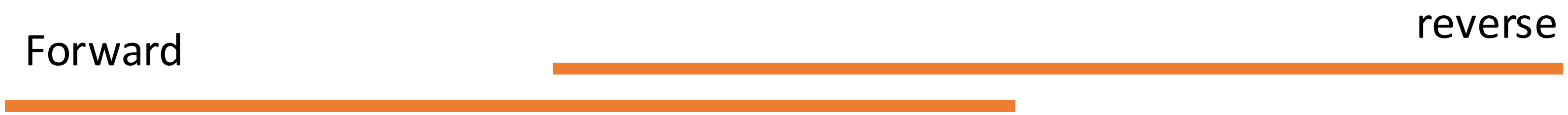
✓ **Sequence Length Distribution**





-> remove too short reads

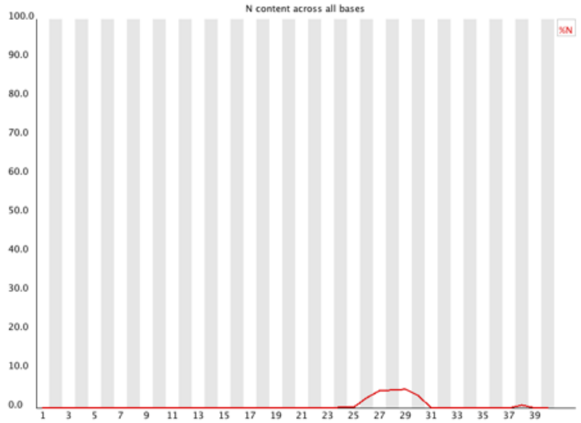
Merge forward and revers reads



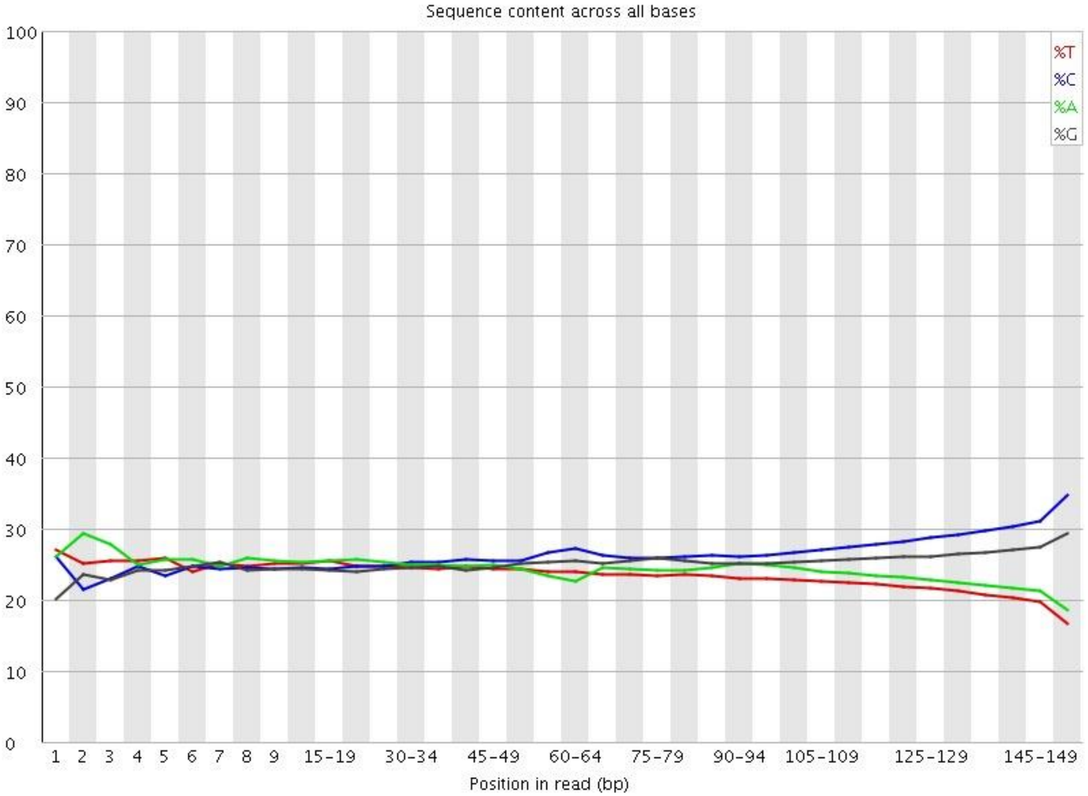
Forward

reverse

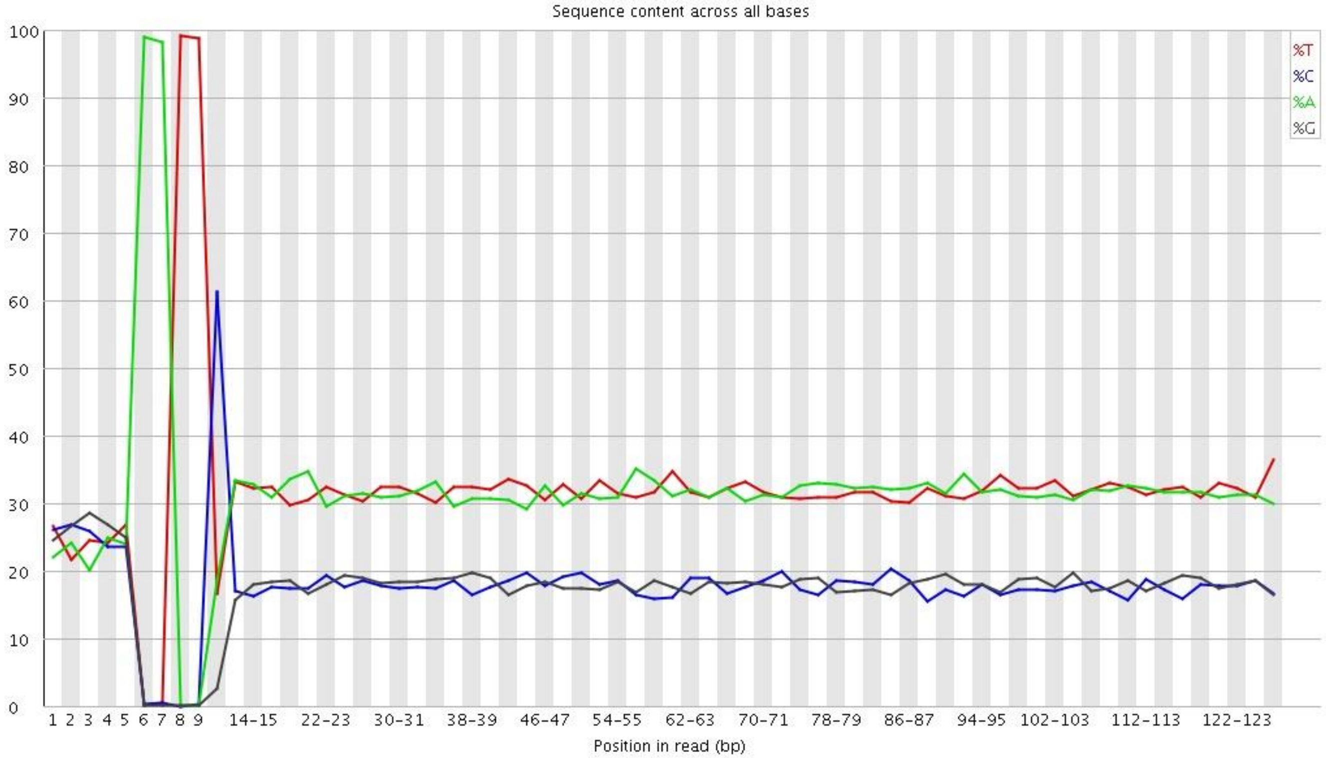
Stretches of Ns, Poly-A or Poly-G



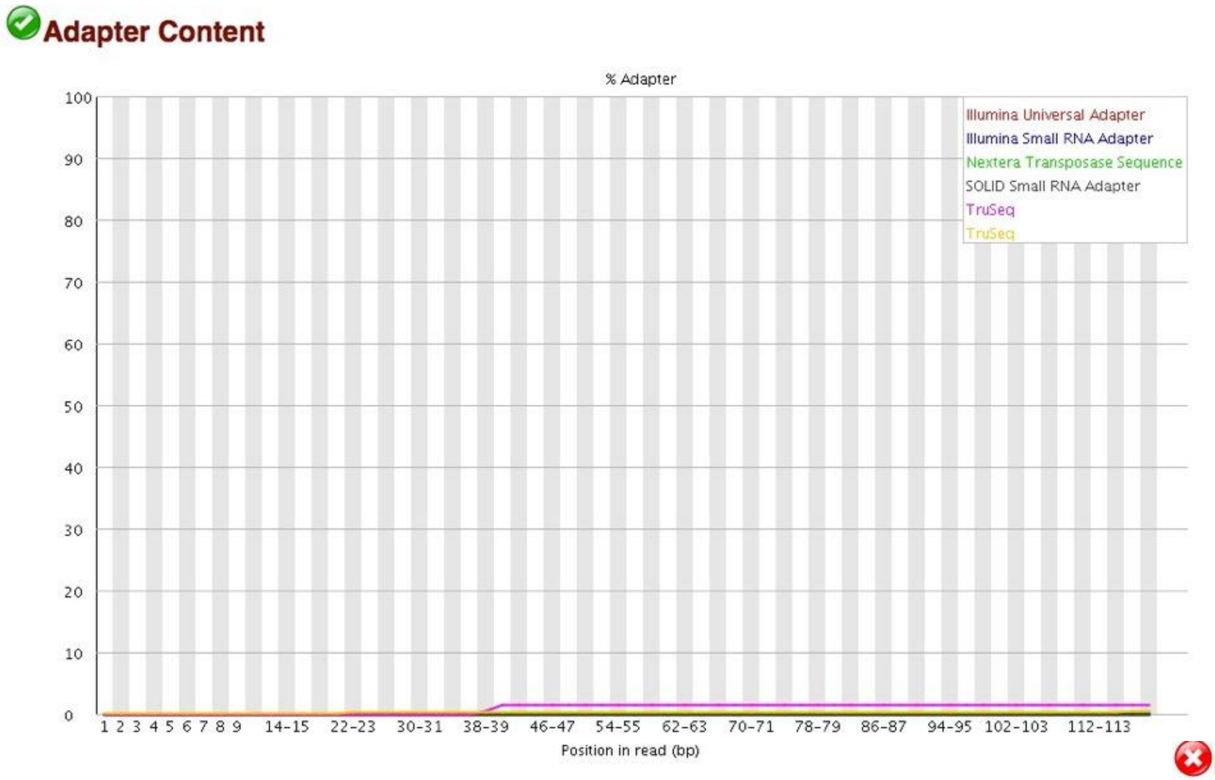
Per base sequence content



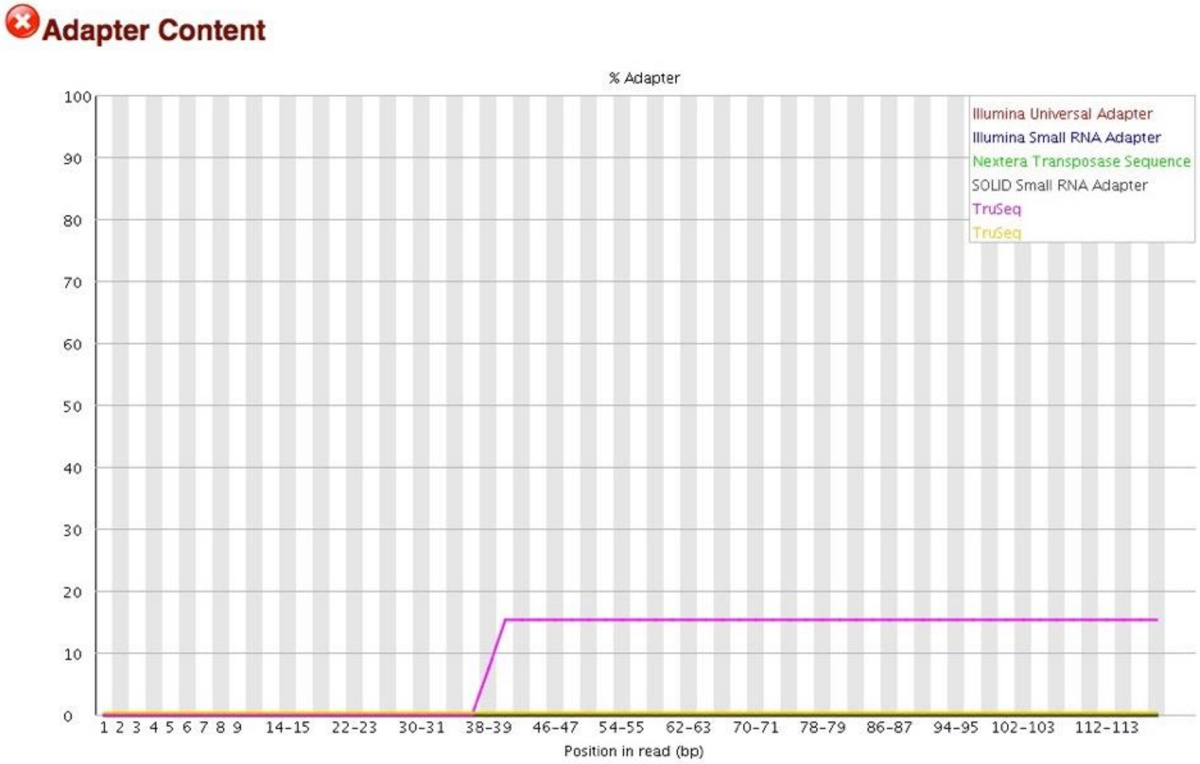
Per base sequence content



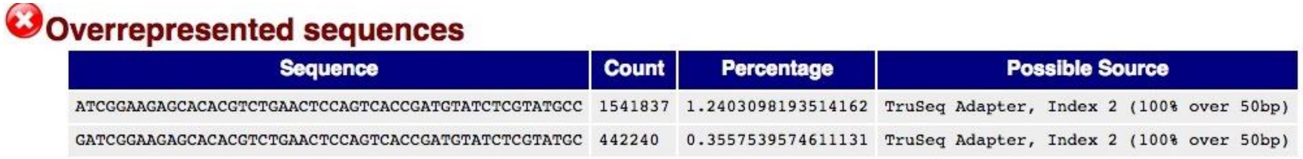
Adapter, primers or indexes



Adapter Content



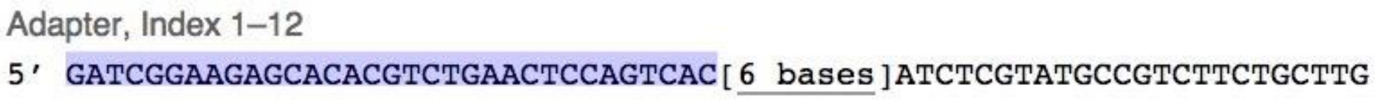
Adapter Content



Overrepresented sequences

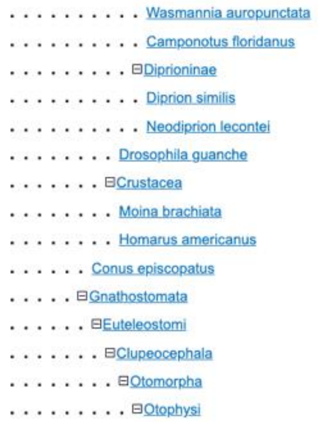
Sequence	Count	Percentage	Possible Source
ATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCC	1541837	1.2403098193514162	TruSeq Adapter, Index 2 (100% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGC	442240	0.3557539574611131	TruSeq Adapter, Index 2 (100% over 50bp)

Illumina adapters in many published genomes



Adapter, Index 1–12

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCAC[6 bases]ATCTCGTATGCCGTCTTCTGCTTG



. . . . Wasmannia auropunctata

. . . . Camponotus floridanus

. . . . ⊟Diprioninae

. . . . Diprion similis

. . . . Neodiprion lecontei

. . . . Drosophila guanche

. . . . ⊟Crustacea

. . . . Moina brachiata

. . . . Homarus americanus

. . . . . . Conus episcopatus

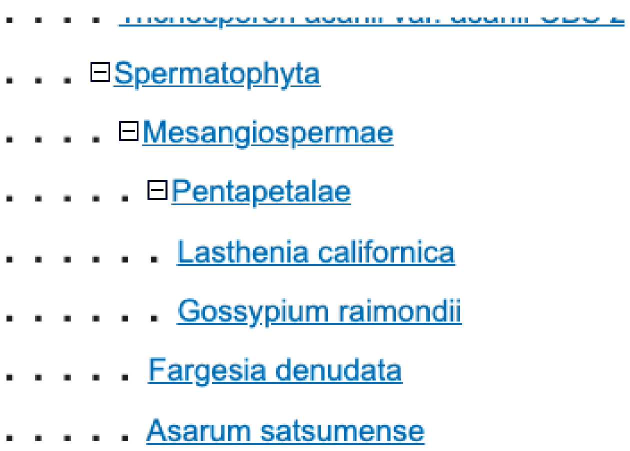
. . . . . ⊟Gnathostomata

. . . . . . ⊟Euteleostomi

. . . . ⊟Clupeocephala

. . . . ⊟Otomorpha

. . . . ⊟Otophysi



. . . . . Trichosporon asahii var. asahii CBS 2

. . . ⊟Spermatophyta

. . . . ⊟Mesangiospermae

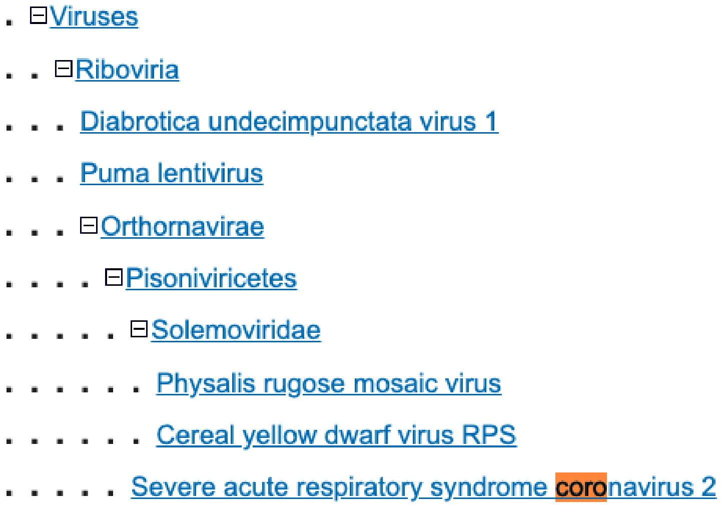
. . . . . ⊟Pentapetalae

. . . . Lasthenia californica

. . . . Gossypium raimondii

. . . . . . Fargesia denudata

. . . . . . Asarum satsumense



. ⊟Viruses

. . ⊟Riboviria

. . . Diabrotica undecimpunctata virus 1

. . . Puma lentivirus

. . . ⊟Orthornavirae

. . . . ⊟Pisoniviricetes

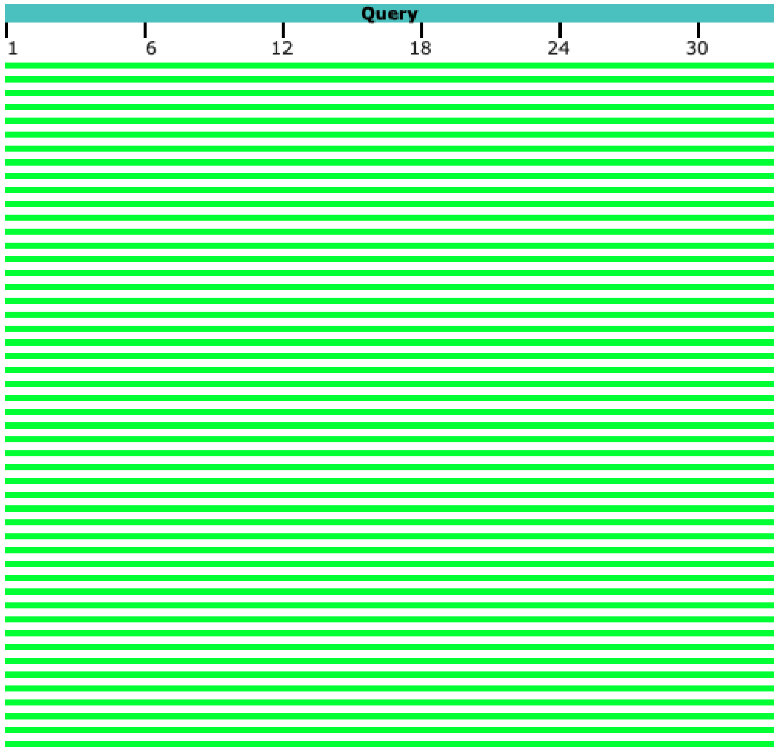
. . . . . ⊟Solemoviridae

. . . . . . Physalis rugose mosaic virus

. . . . . . Cereal yellow dwarf virus RPS

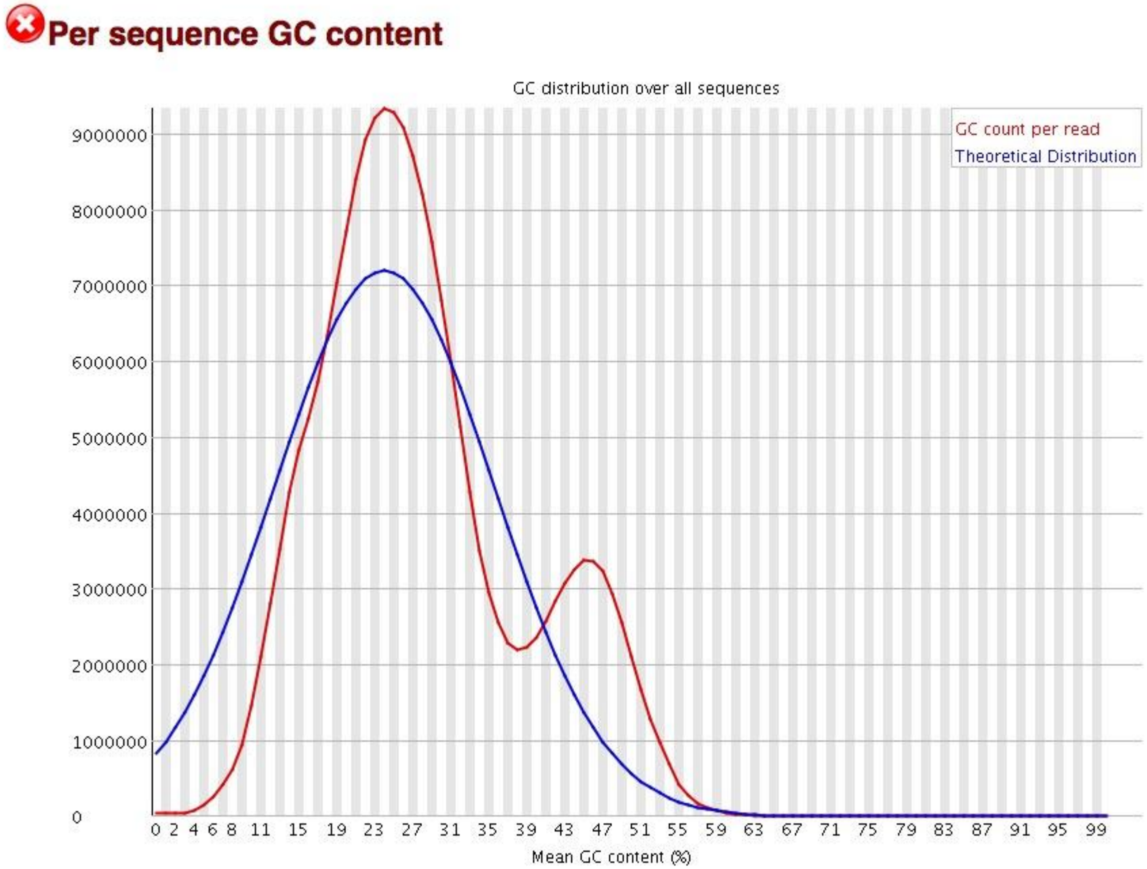
. . . . . Severe acute respiratory syndrome coronavirus 2

Distribution of the top 535 Blast Hits on 405 subject sequenc



I have cloudiness,
you have this weird smell.
We are both CONTAMINATED!
What will happen to us?
www.biocomicals.com
Aura unim

Contaminants

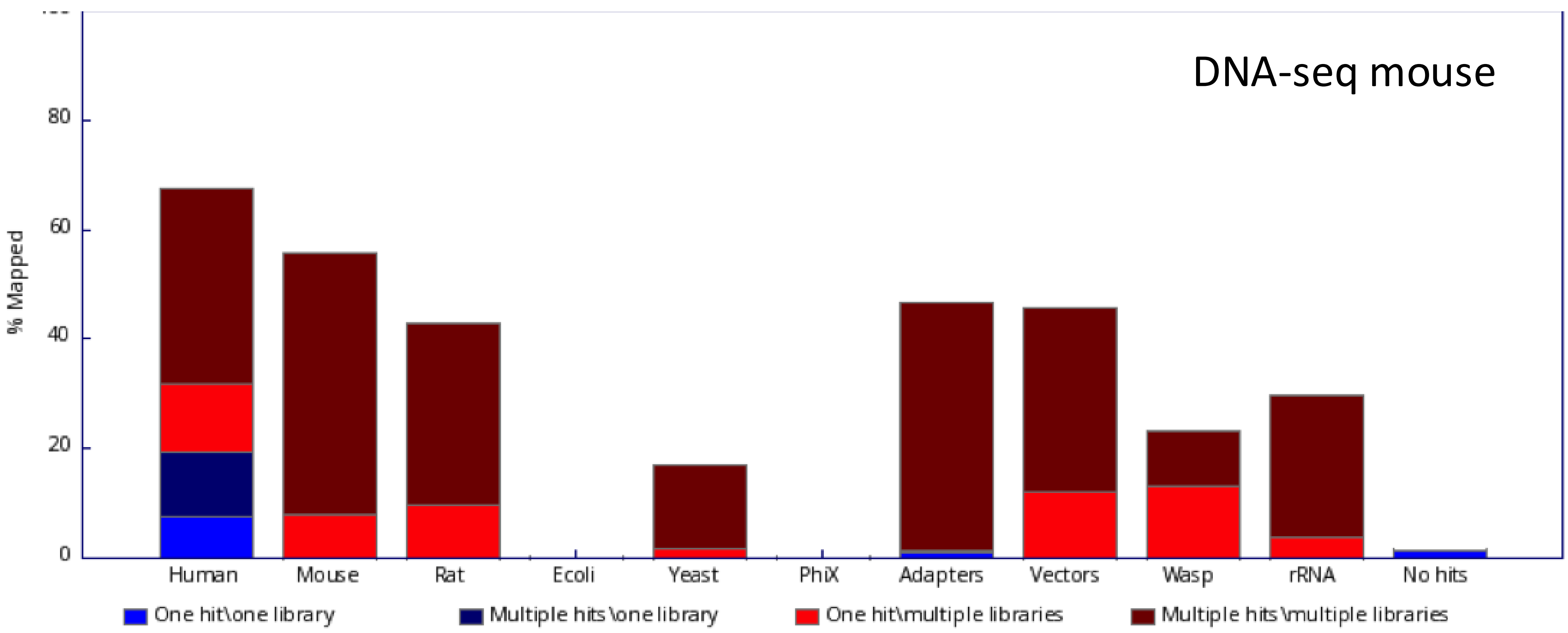


Per sequence GC content

Contamination

FastQ Screen

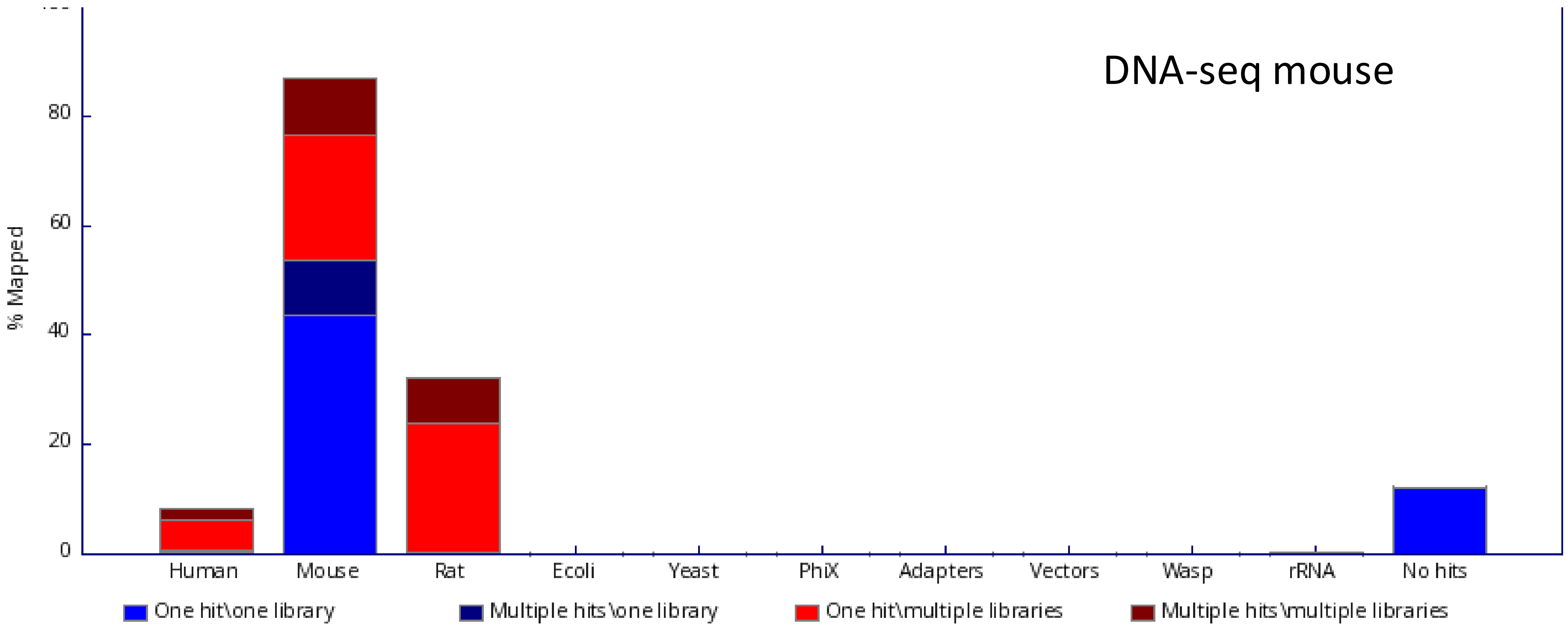
Contamination screening for NGS data



Contamination

FastQ Screen

Contamination screening for NGS data



DNA-seq mouse

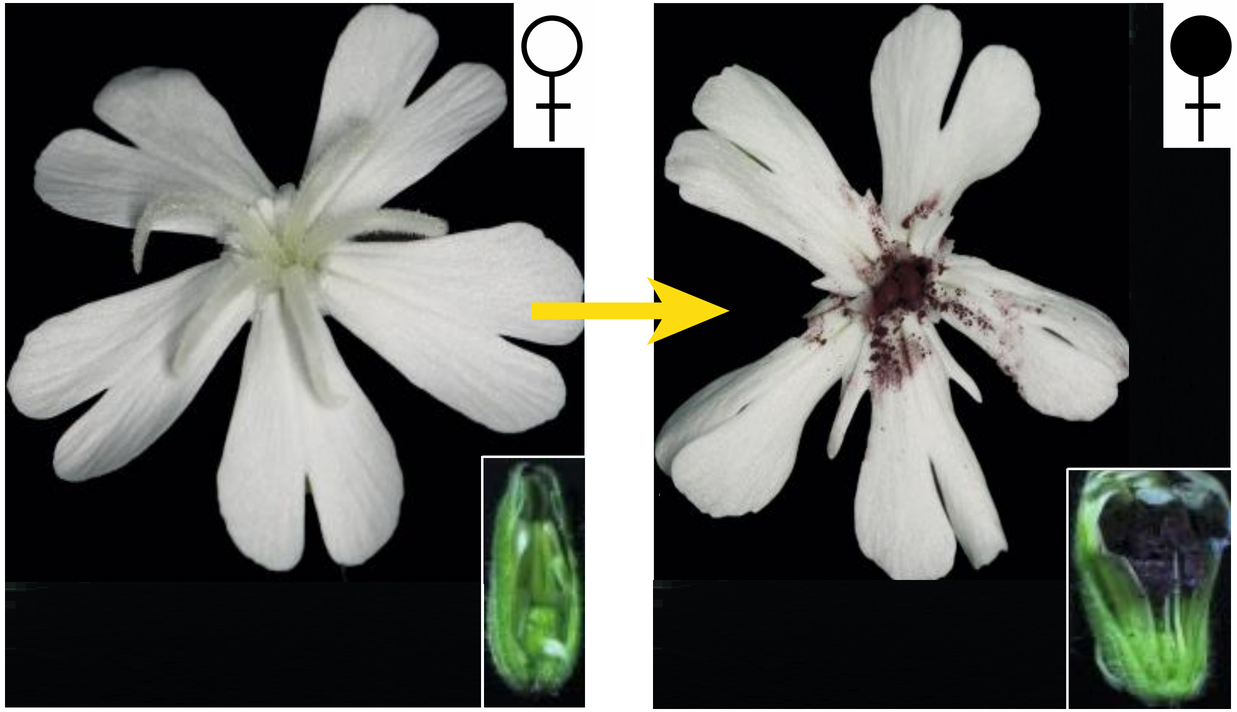
Tools for removing contaminants

Often not needed since they occur randomly

-> replicates

-> sufficient DNA input

Dual RNA-seq approache



- Healthy plant transcriptome
- Fungal reads (less than 5 % of all reads)

Zemp et al. (2015)

Tools for removing contaminants

Random contaminants

Often not needed since they occur randomly -> replicates

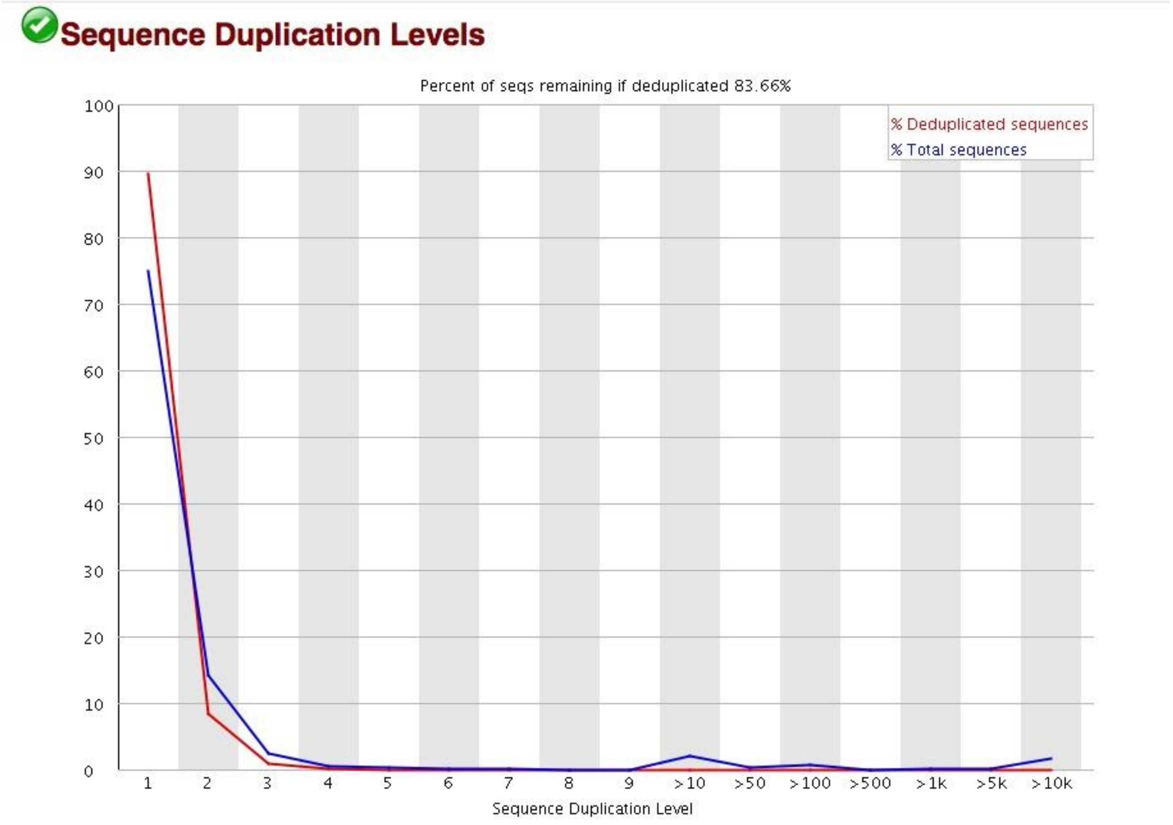
de novo **assembly in Host-pathogen Systems:**

Blast assembled contigs against databases/genome

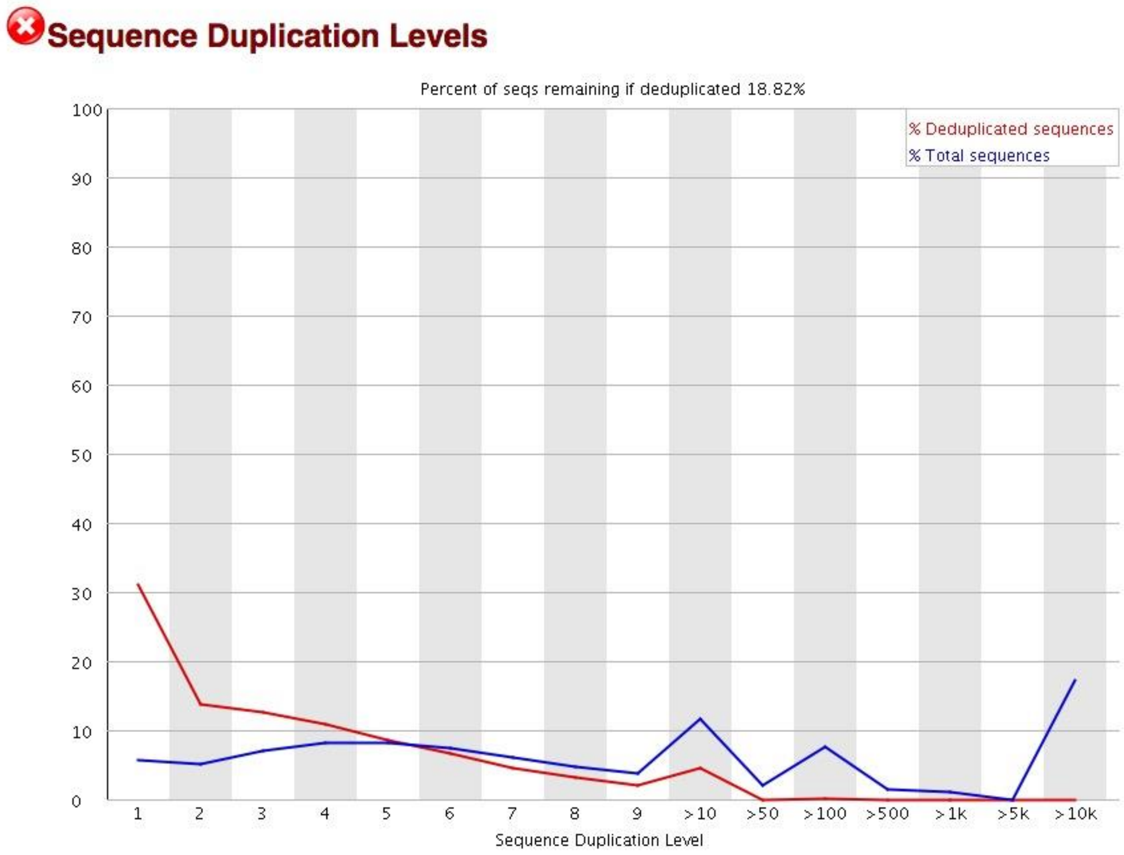
“blast” raw-reads against databases (Kraken, Kaiju)

Filter based on GC content

Duplication levels

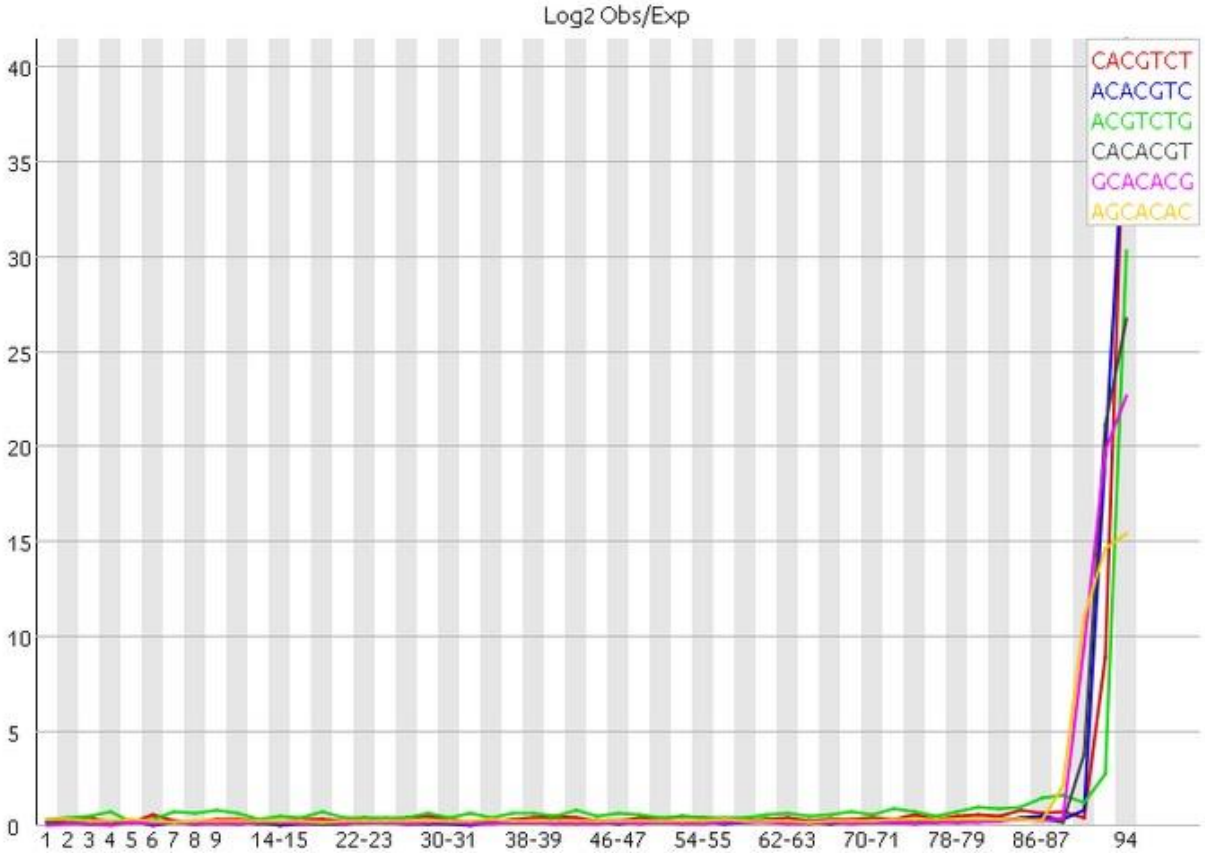


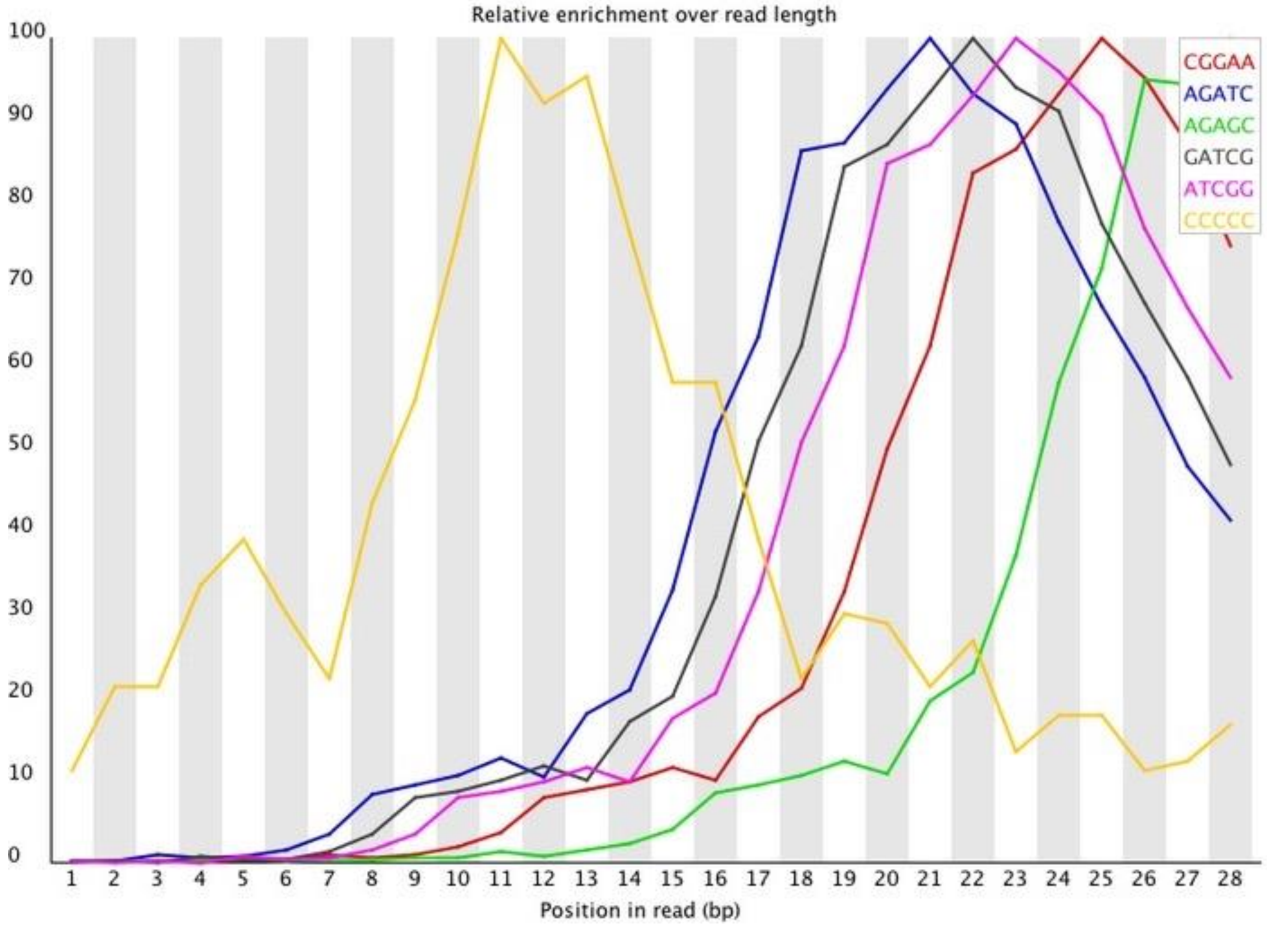
✓ Sequence Duplication Levels



✗ Sequence Duplication Levels

Kmer content

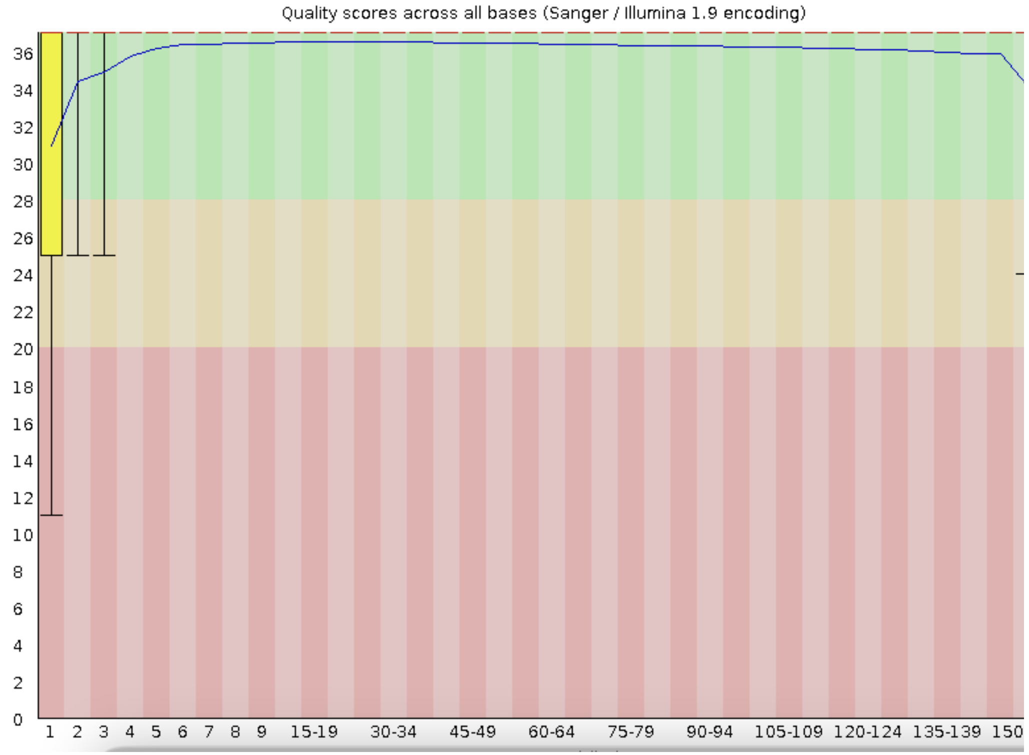




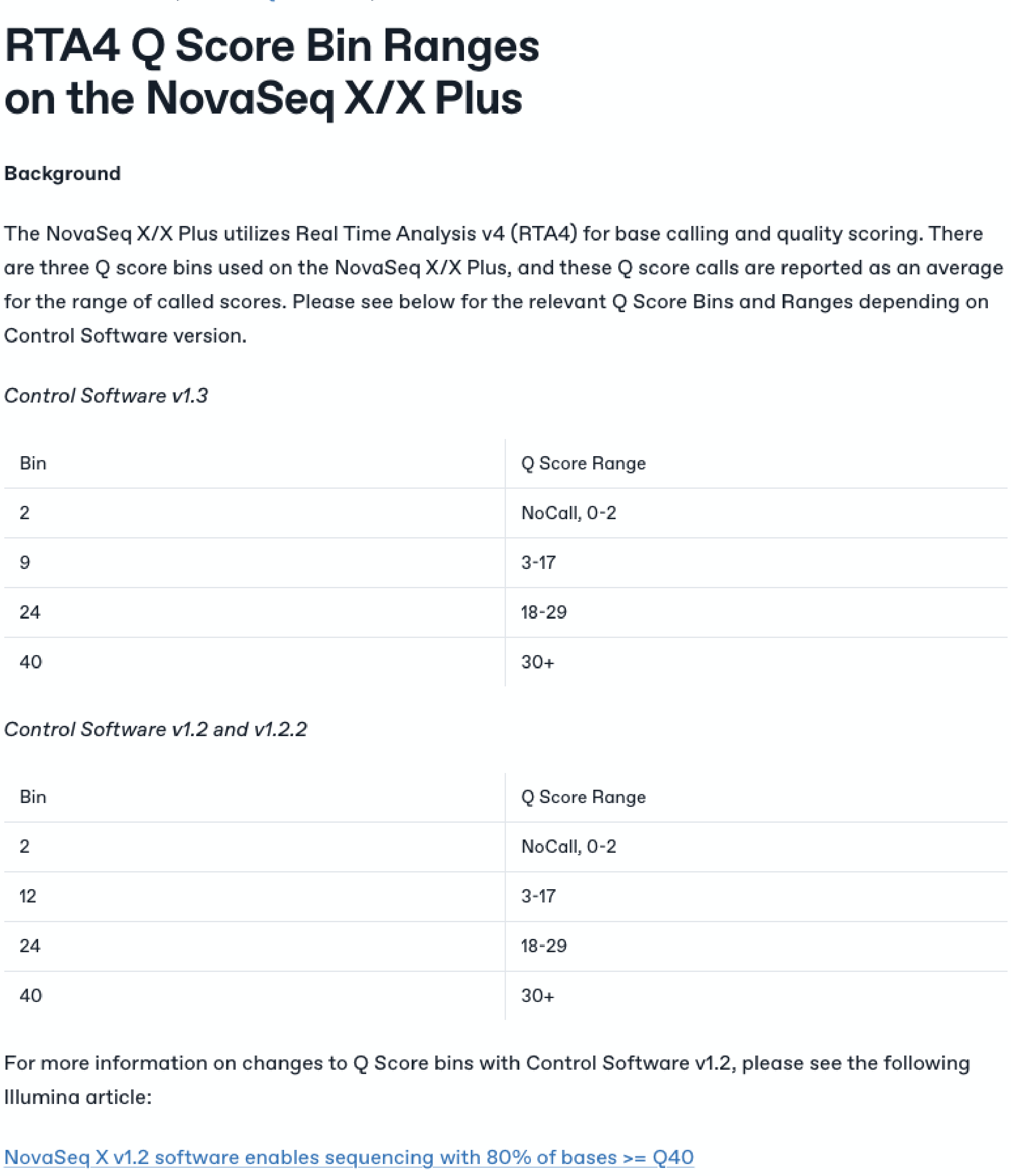
This module will issue a warning if any k-mer is enriched more than 3 fold overall, or more than 5 fold at any individual position.

Newer is better!

✓ **Per base sequence quality**



NovaSeq X, RNAseq



RTA4 Q Score Bin Ranges on the NovaSeq X/X Plus

Background

The NovaSeq X/X Plus utilizes Real Time Analysis v4 (RTA4) for base calling and quality scoring. There are three Q score bins used on the NovaSeq X/X Plus, and these Q score calls are reported as an average for the range of called scores. Please see below for the relevant Q Score Bins and Ranges depending on Control Software version.

Control Software v1.3

Bin	Q Score Range
2	NoCall, 0-2
9	3-17
24	18-29
40	30+

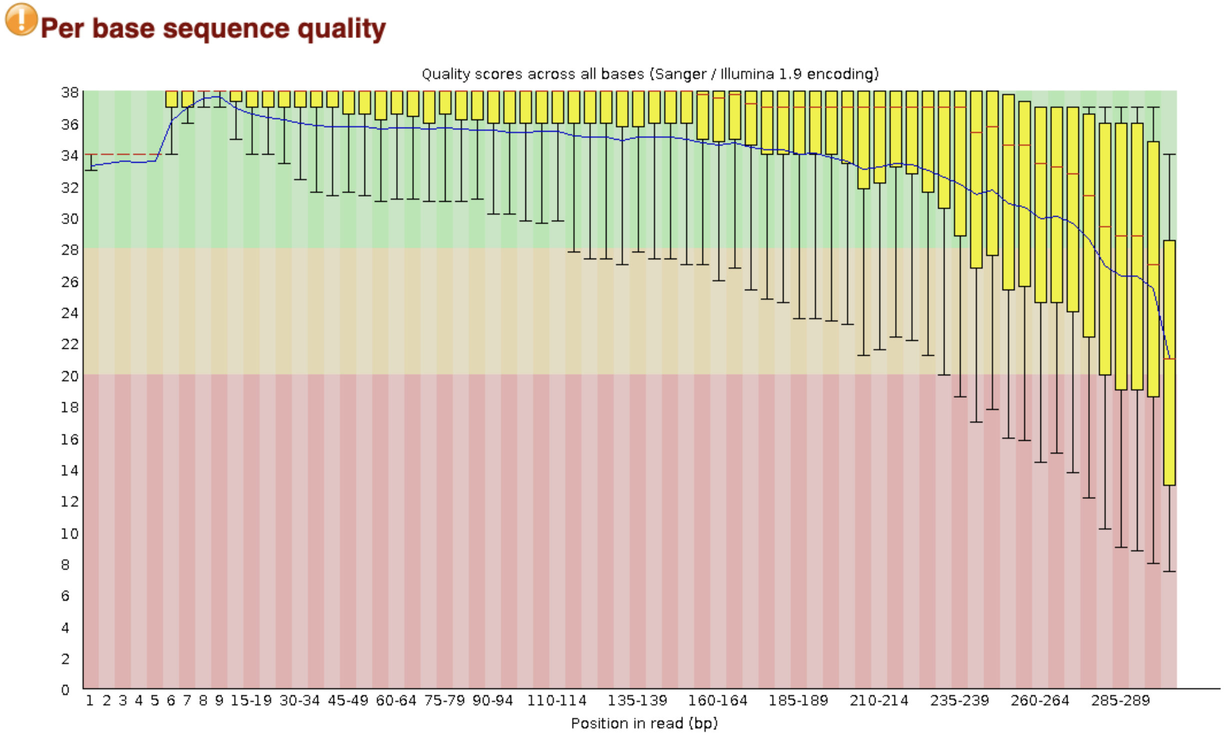
Control Software v1.2 and v1.2.2

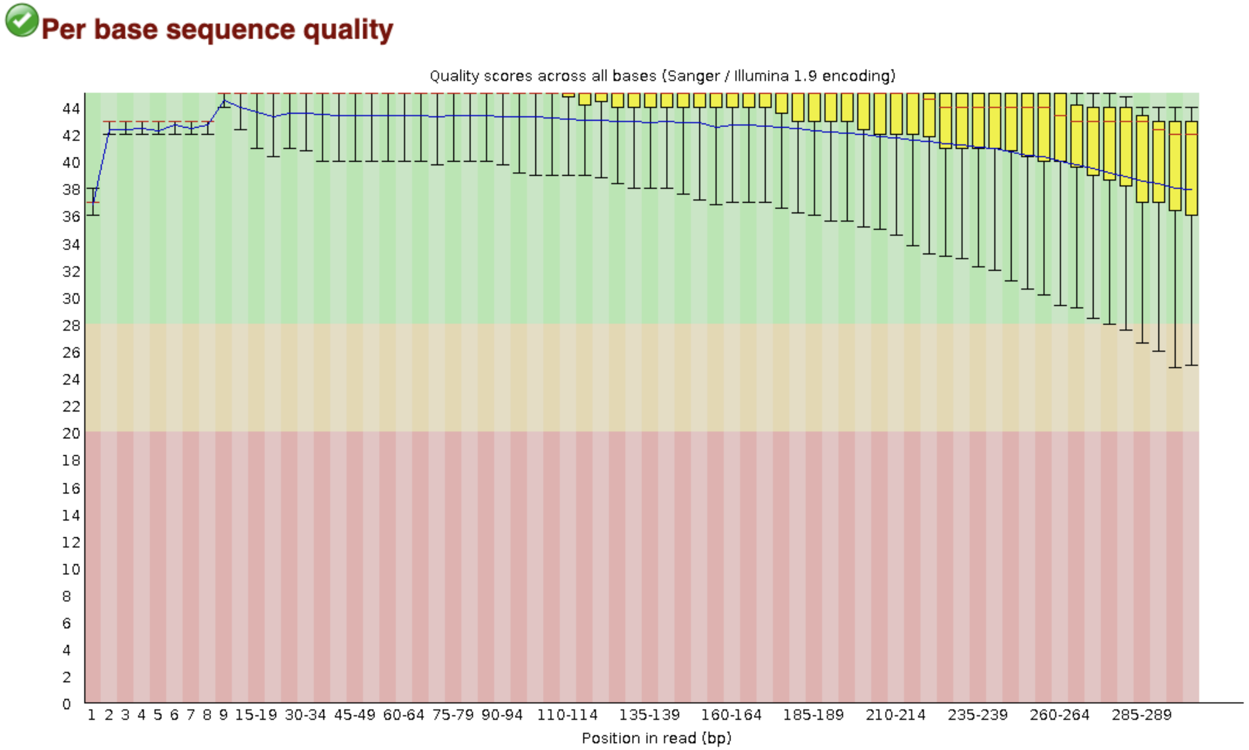
Bin	Q Score Range
2	NoCall, 0-2
12	3-17
24	18-29
40	30+

For more information on changes to Q Score bins with Control Software v1.2, please see the following Illumina article:

NovaSeq X v1.2 software enables sequencing with 80% of bases >= Q40

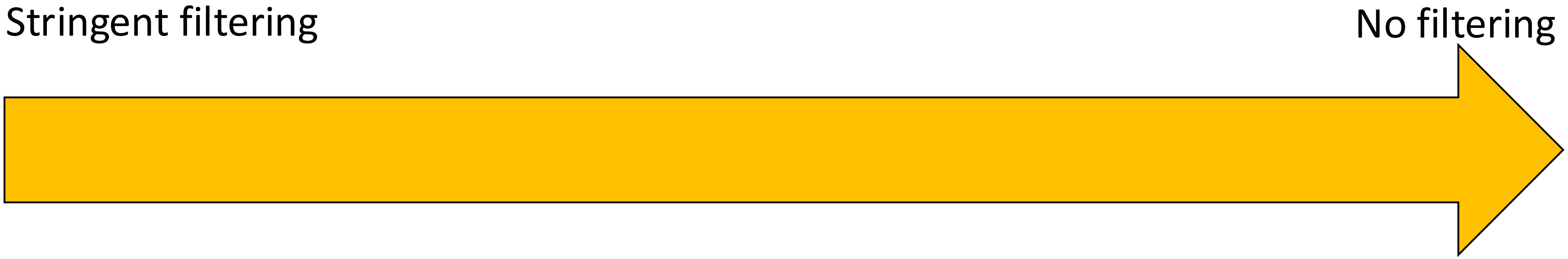
Newer is better!





RADseq, R1

How stringent do we need to be?



Stringent filtering

No filtering

de novo assembly

AmpSeq

Ancient DNA

Re-sequencing

RNA-seq

RAD-seq

Take home message

- Fastqc has been developed for DNAseq
- Check your raw data
- Stringent filtering/trimming is often not needed

Quality filtering is a process that is used to remove low-quality reads or bases from next-generation sequencing data in the FASTQ format. The goal of quality filtering is to improve the overall accuracy and reliability of the downstream analyses that will be performed on the data, such as alignment, variant calling, and gene expression analysis.

There are several methods for quality filtering FASTQ files, and the specific approach that is used will depend on the specific needs and goals of the analysis. Some common methods for quality filtering include:

- Trimming: This involves removing low-quality bases from the ends of reads. Trimming is often used to remove adapter sequences or other contaminants that may have been introduced during the sequencing process.
- Filtering by Phred quality score: This involves identifying reads or bases that have a Phred quality score below a certain threshold, and removing them from the data. The Phred quality score is a measure of the accuracy of a base call, with higher scores indicating higher confidence in the accuracy of the call.
- Filtering by length: This involves removing reads that are shorter than a certain length threshold. Shorter reads are often of lower quality and may not be useful for downstream analyses.
- Overall, quality filtering is an important step in the analysis of FASTQ data, as it helps to ensure that the downstream analyses are as accurate and reliable as possible.