Quality Filtering

Niklaus Zemp

25 June 2021

Genetic Diversity Centre (GDC)
Bioinformatics
ETH Zurich

Check your data

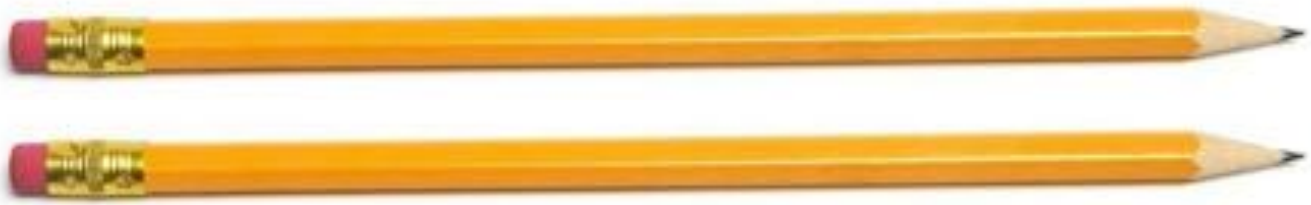
Sequencing technologies

Short read- (Ilumina)

Low error rate

Quality filtering

Long read – (PacBio, ONT)



High error rate

Error correction

Quality control

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)

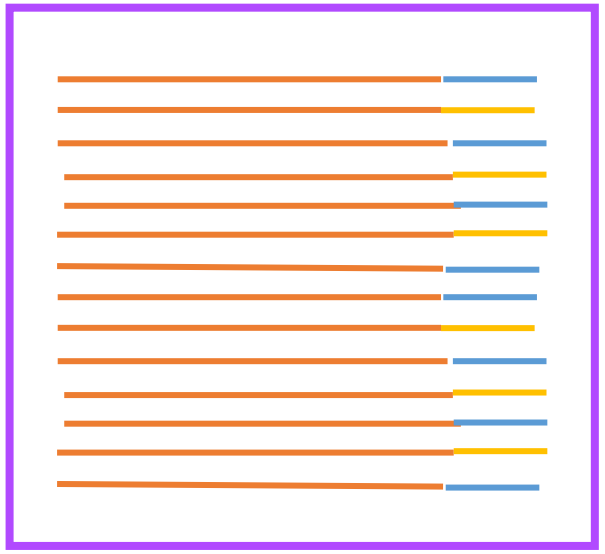
bbmap (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmap-guide/)

…

Demultiplexing

Normally done by the Illumina software

- A low number of reads is always wrongly inferred



New Results

Index Switching Causes "Spreading-Of-Signal" Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing

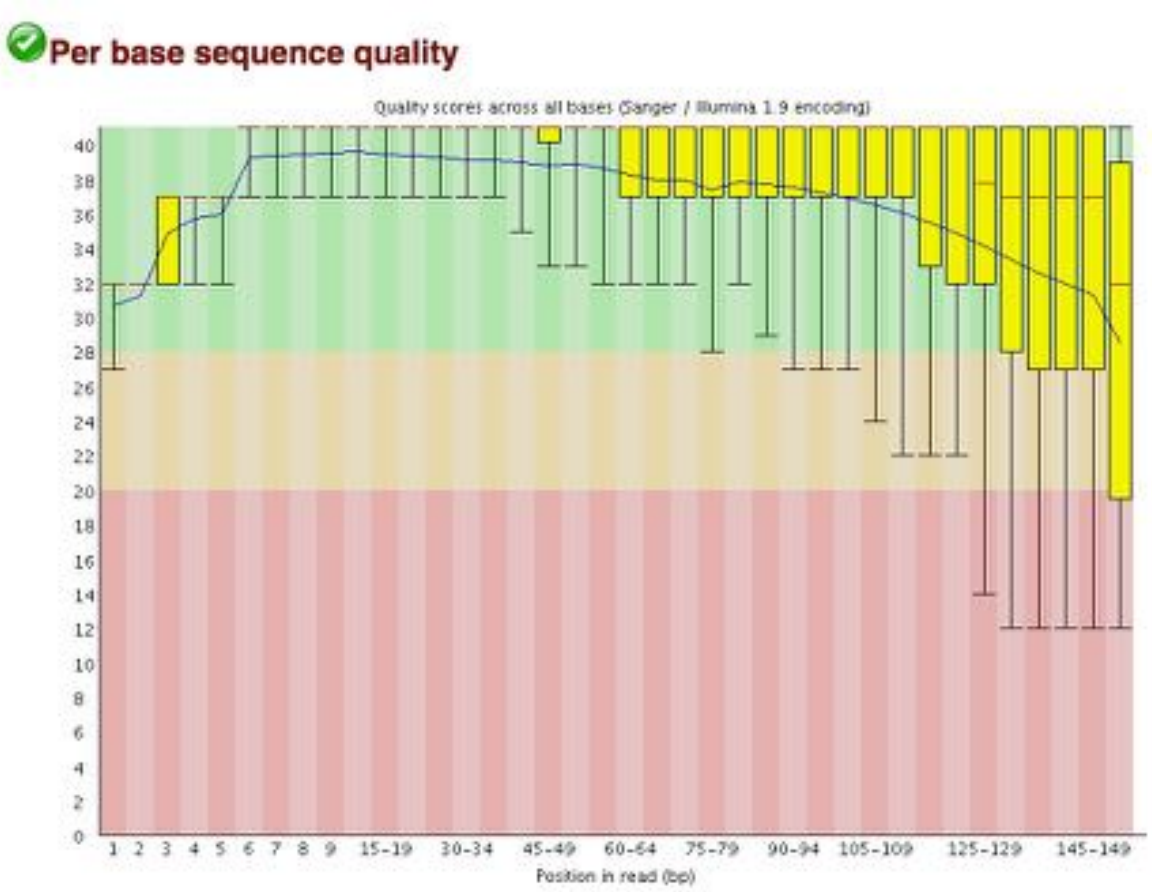
Index hopping is normally less then 1%

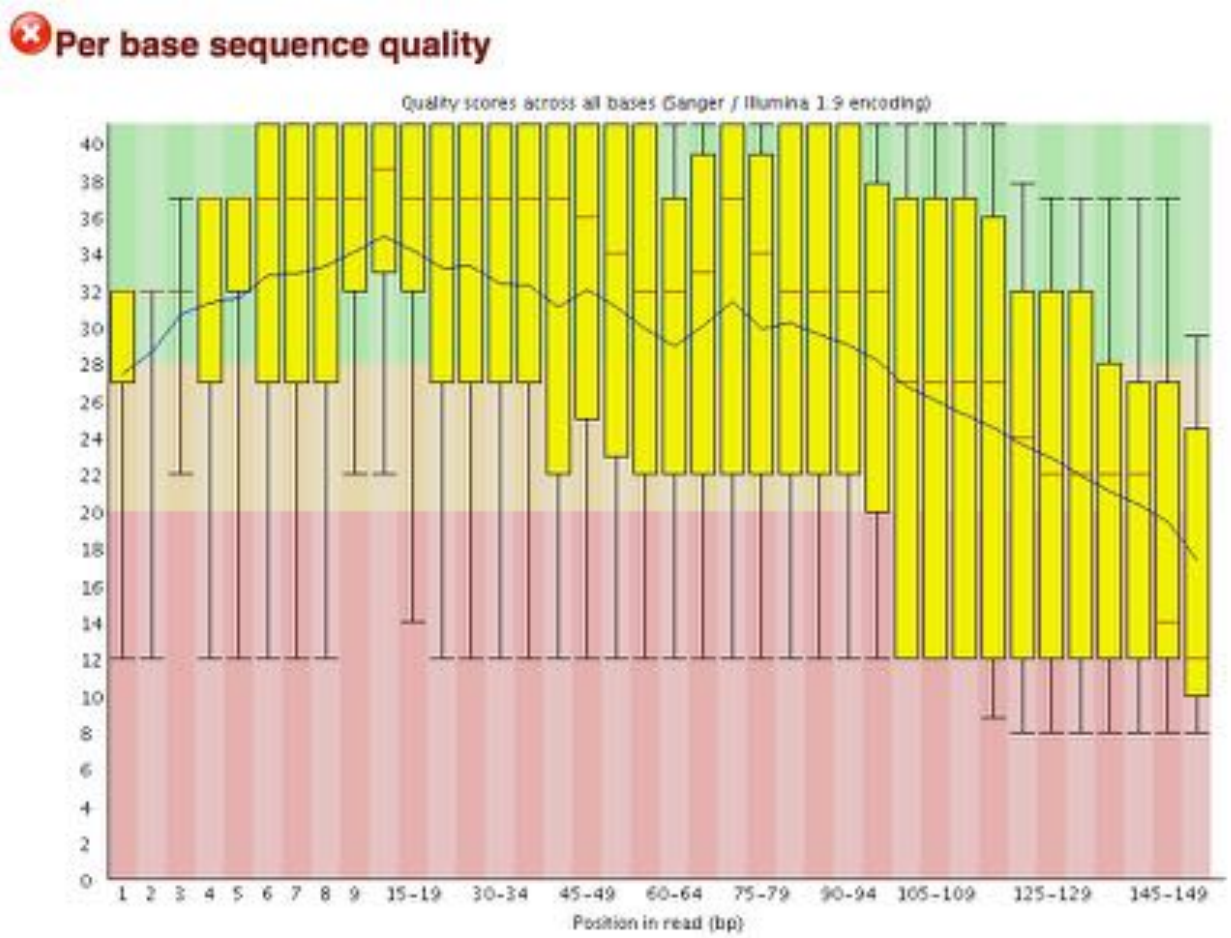
-> Rare events are more affected

-> Use replicates

-> Use unique dual barcodes

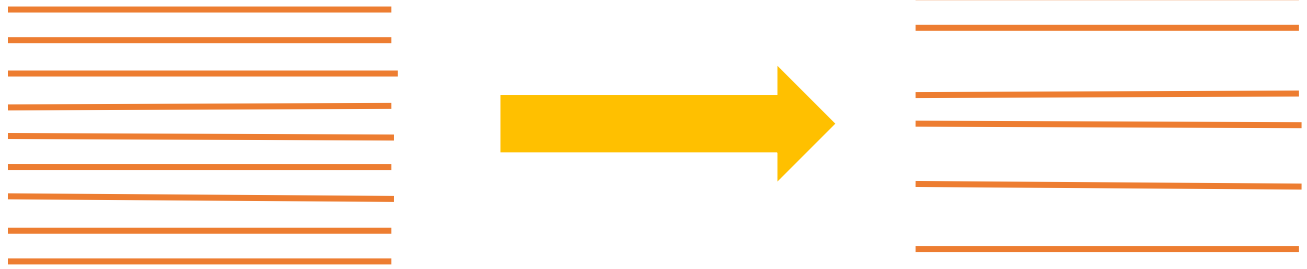
Per bases sequence quality



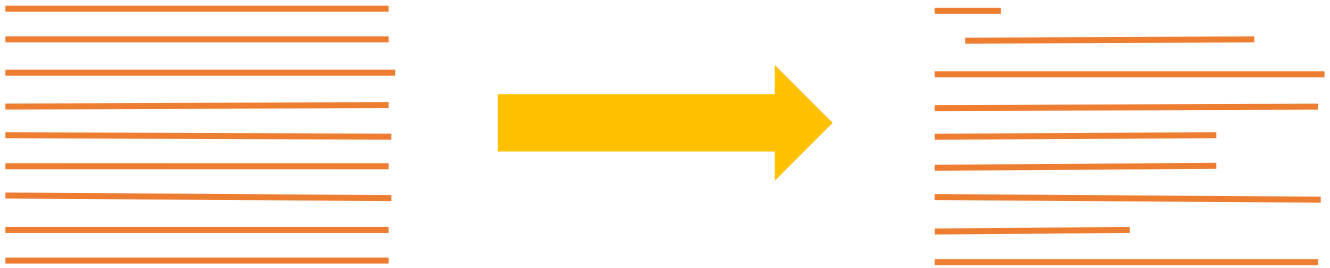


Filtering and/or trimming

Filtering



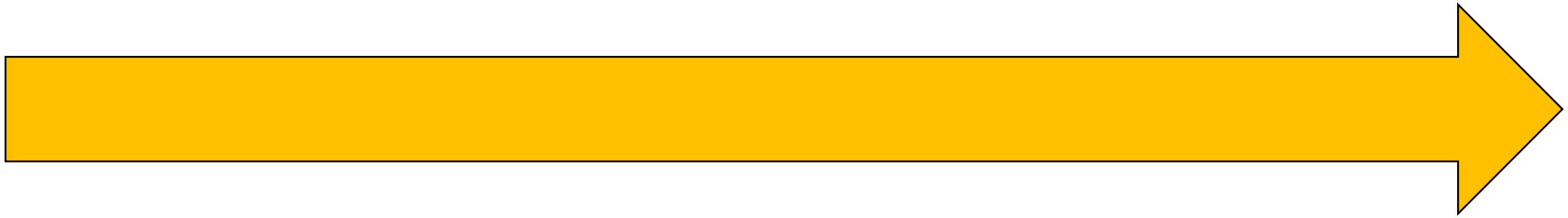
Trimming



How stringent do we need to be?

Stringent

No filtering



de novo assembly

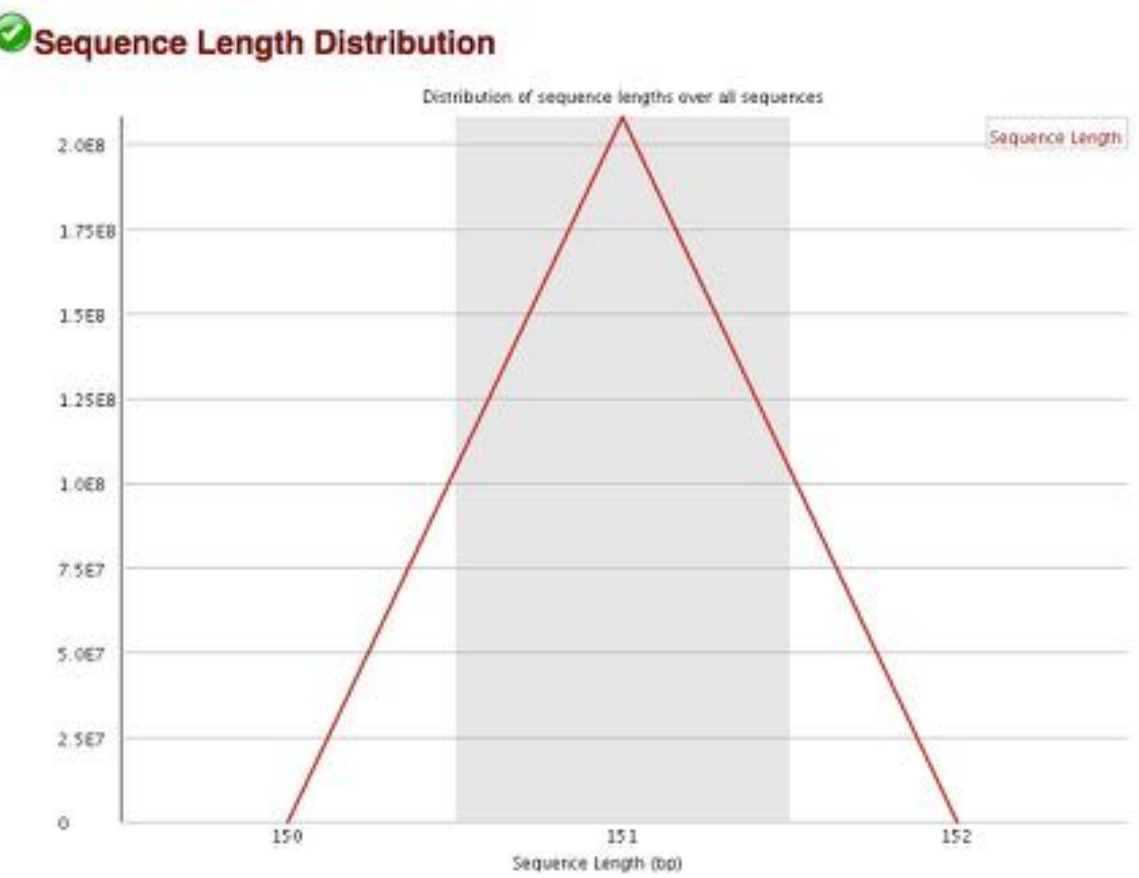
Low coverage sequencing

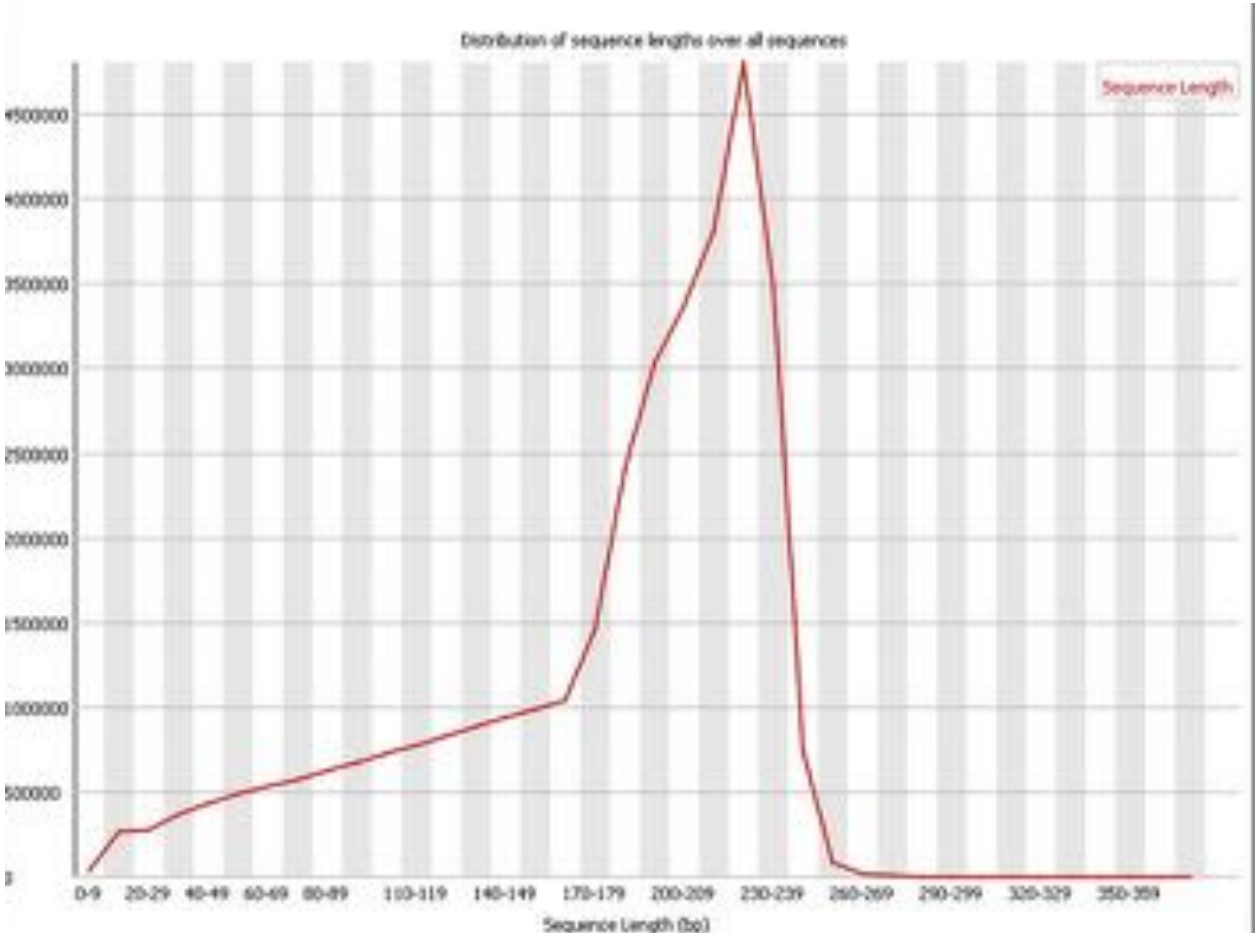
Amplicon sequencing

Re-sequencing

RNA-seq

Sequence length

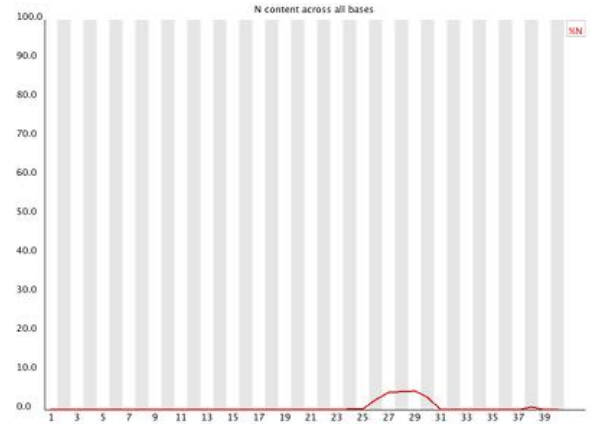


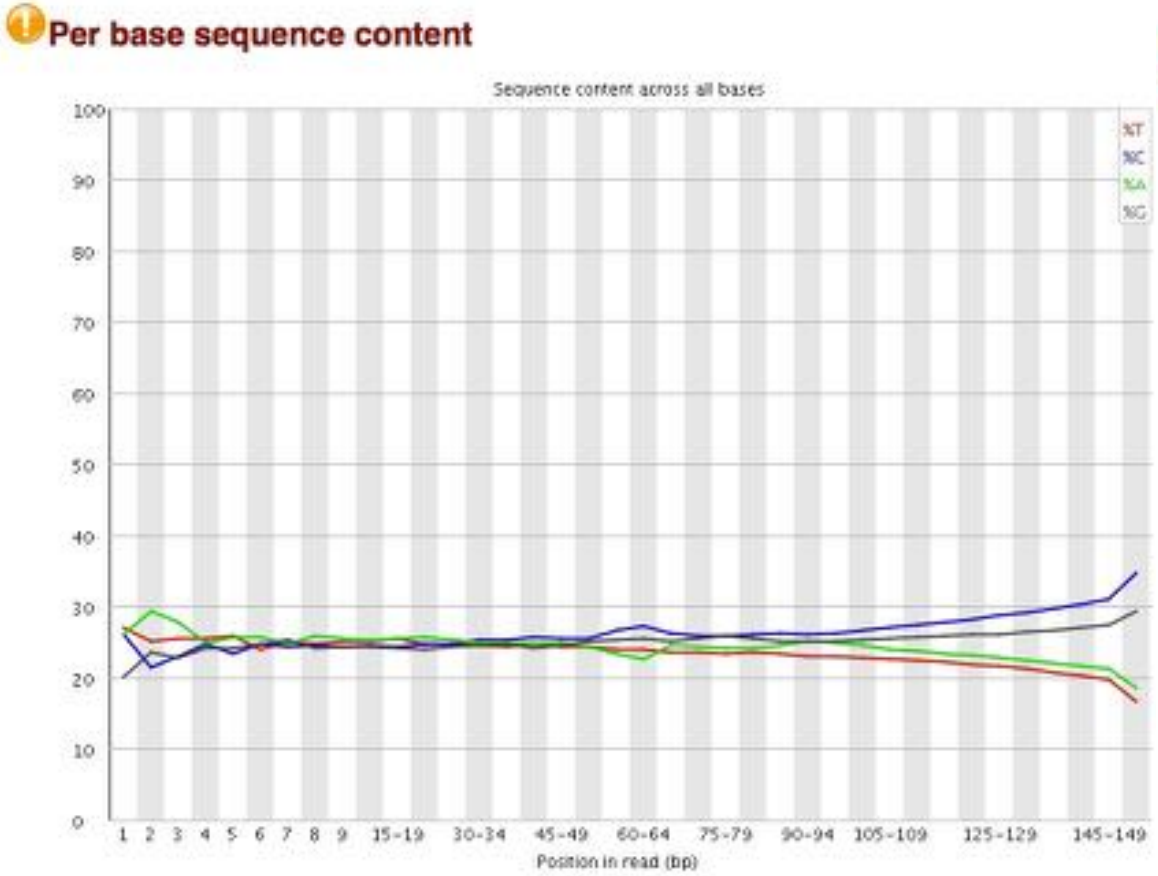


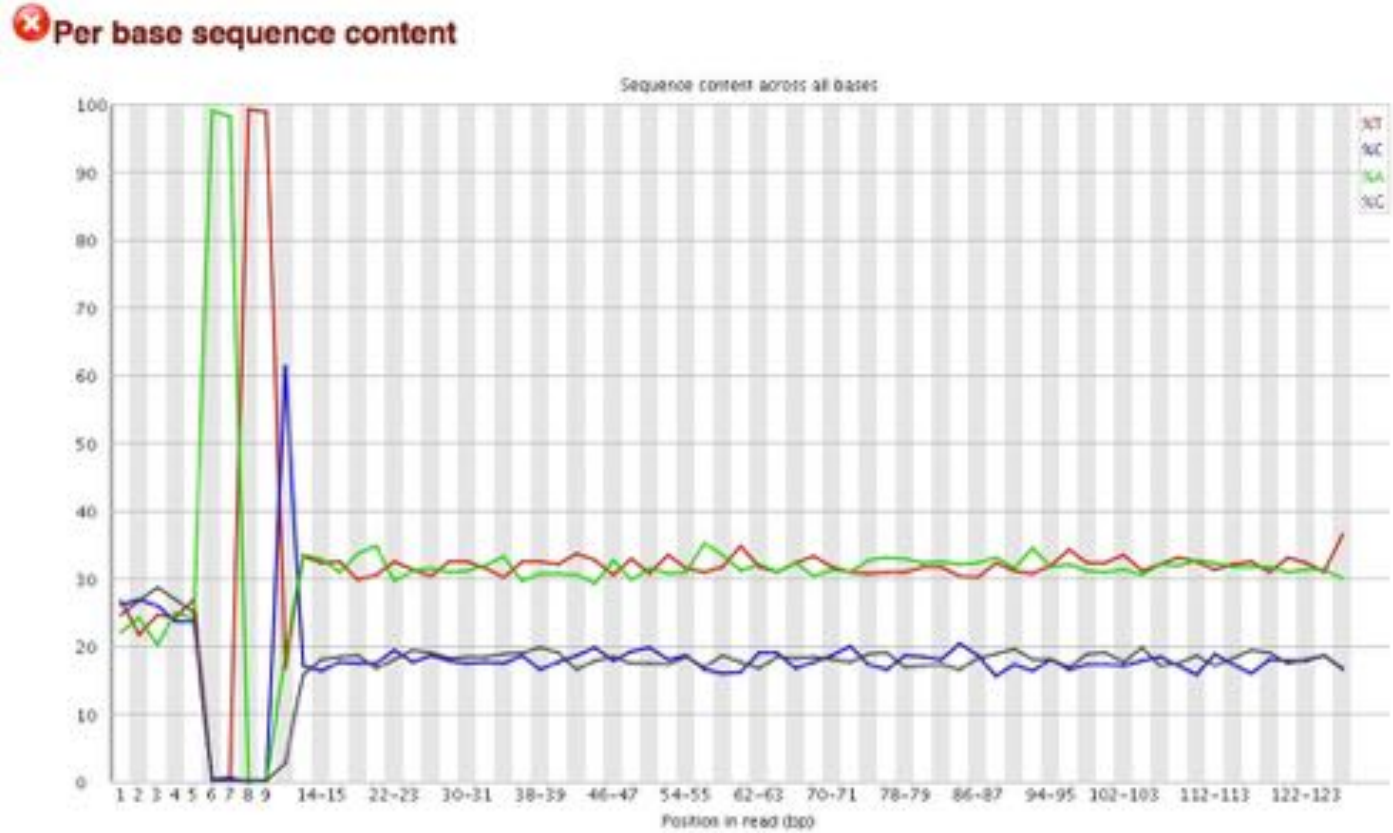
-> remove too short reads

Merge forward and revers reads

Stretches of Ns, Poly-A or Poly-G

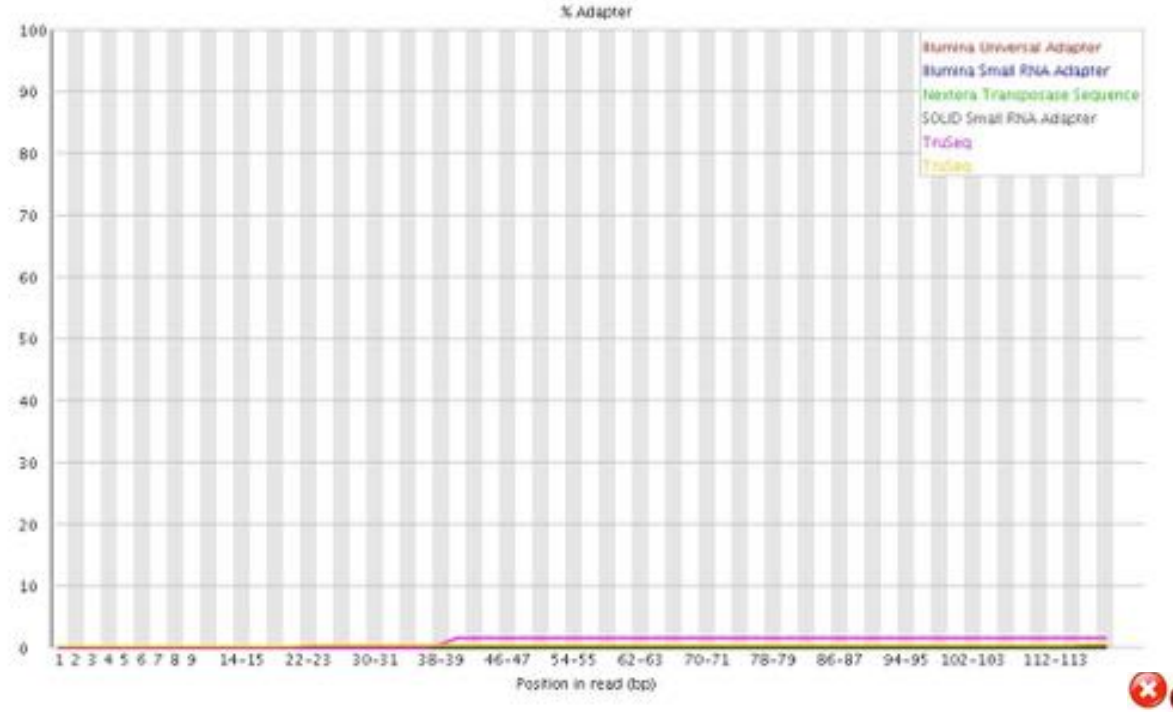




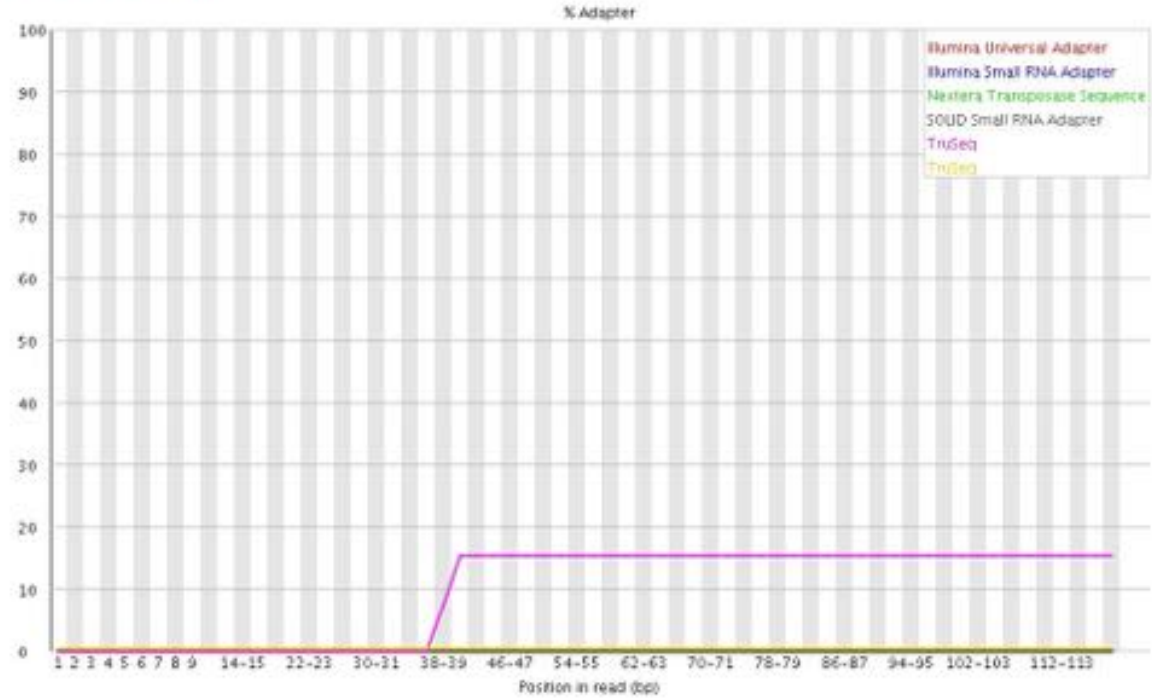


Adapter, primers or indexes

Adapter Content



Adapter Content



Overrepresented sequences

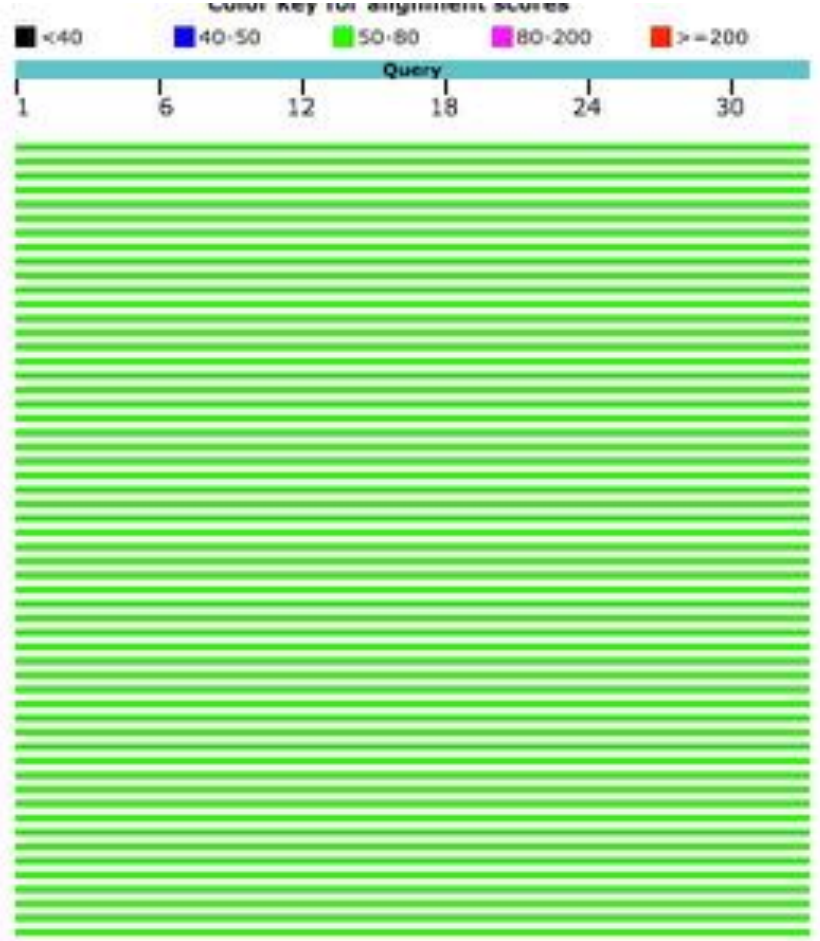
Sequence	Count	Percentage	Possible Source
ATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCC	1541837	1.2403098193514162	TruSeq Adapter, Index 2 (100% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGC	442240	0.3557539574611131	TruSeq Adapter, Index 2 (100% over 50bp)

Illumina adapters in many published genomes

Adapter, Index 1–12

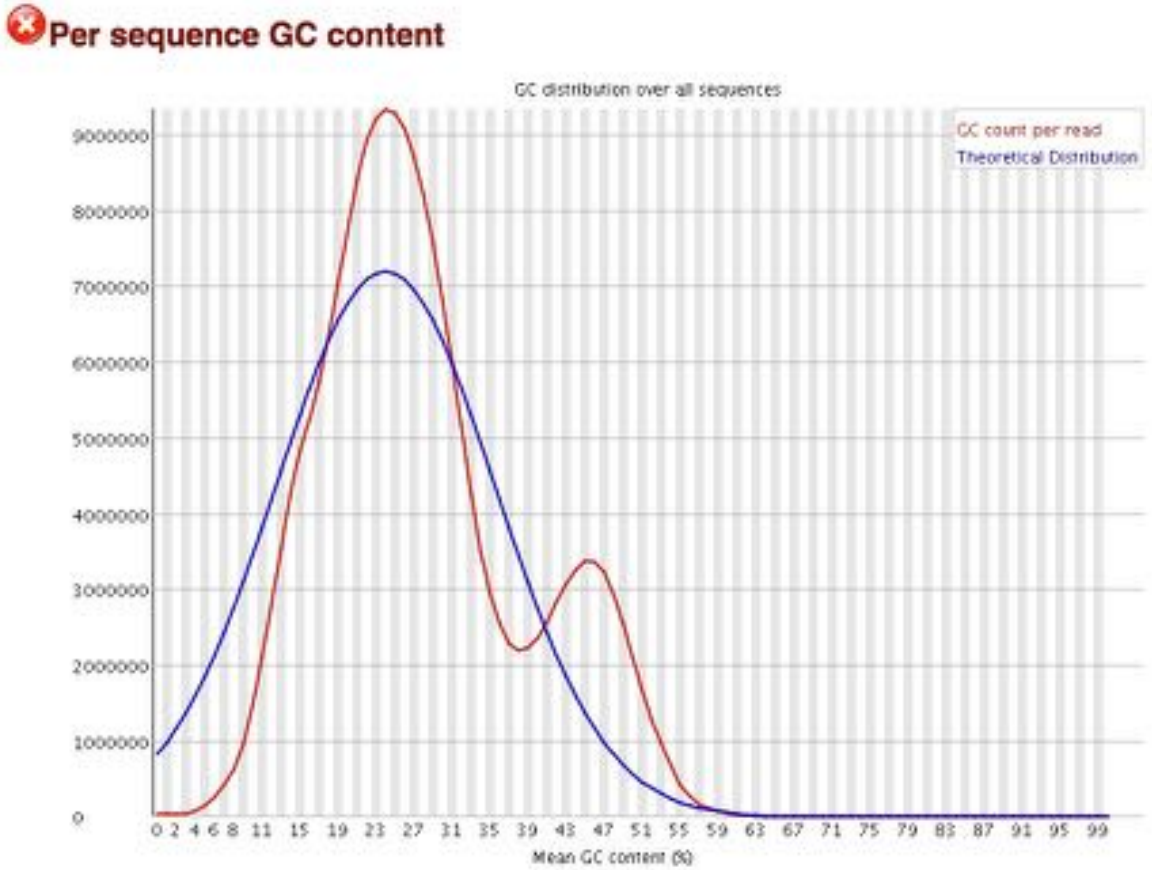
5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCAC[6 bases]ATCTCGTATGCCGTCTTCTGCTTG

Clupea	harengus
Wasmannia	auropunctata
Mesorhizobium	sp.
Cephus	cinctus
Streptomyces	griseorubens
Escherichia	coli
Pediococcus	acidilactici
Trichosporon	asahii
Camelus	ferus
Pseudomonas	tolaasii
Sarcophilus	harrisii
Halomonas	sp.
Fusarium	pseudograminearum
Cyprinus	carpio
Corynebacterium	provencense
Eggerthellaceae	bacterium
Mycobacterium	bovis
Lepisosteus	oculatus
Saimiri	boliviensis
Condylura	cristata
Trichechus	manatus
Heterocephalus	glaber
Octodon	degus
Paenibacillus	sp.
Klebsiella	pneumoniae
Streptomyces	lividans



26 species with Illumina adapters in the genome

Contaminants

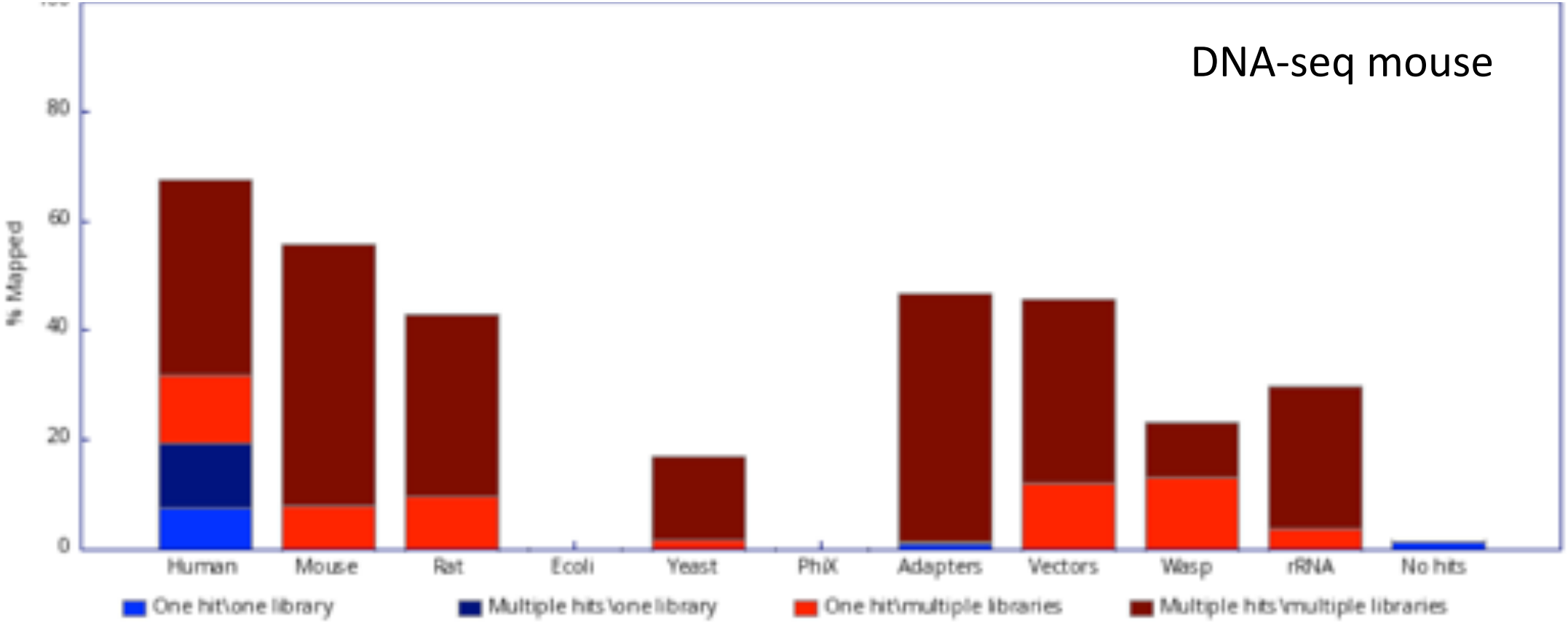


Contamination

FastQ Screen

Contamination screening for NGS data

DNA-seq mouse

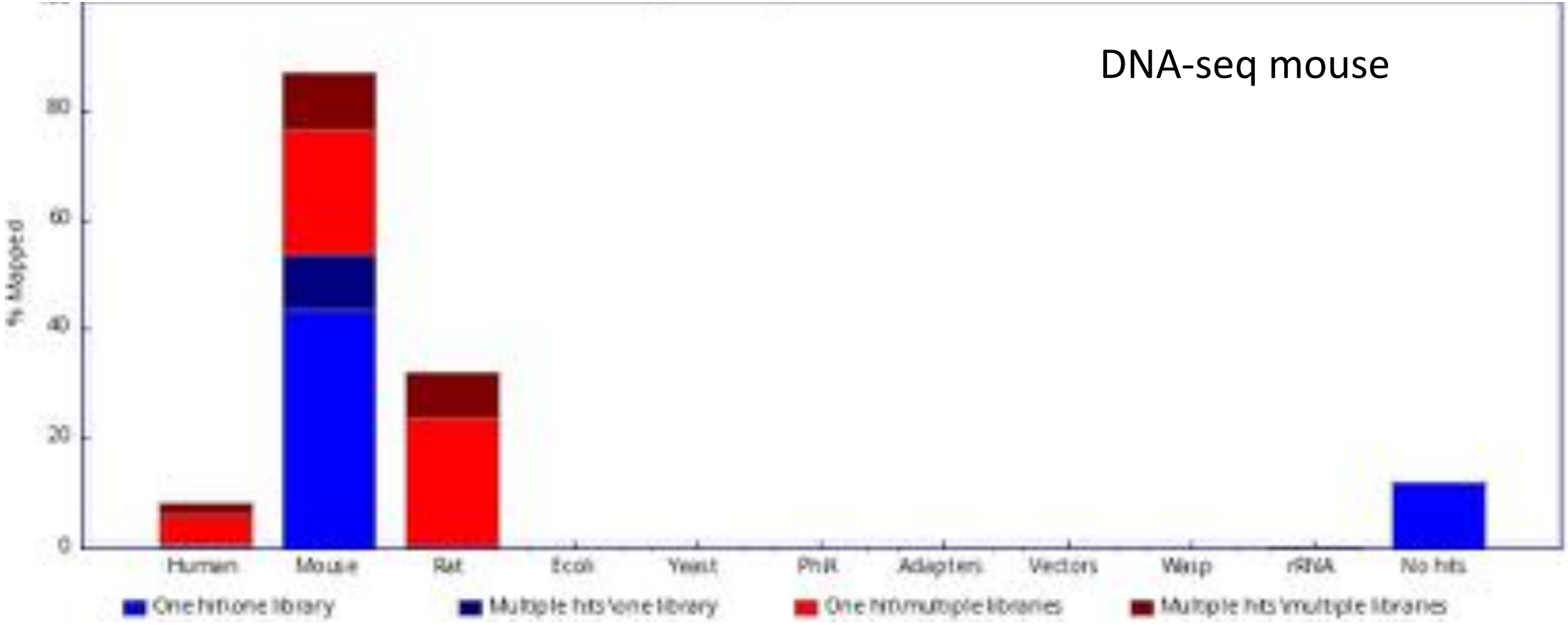


Contamination

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

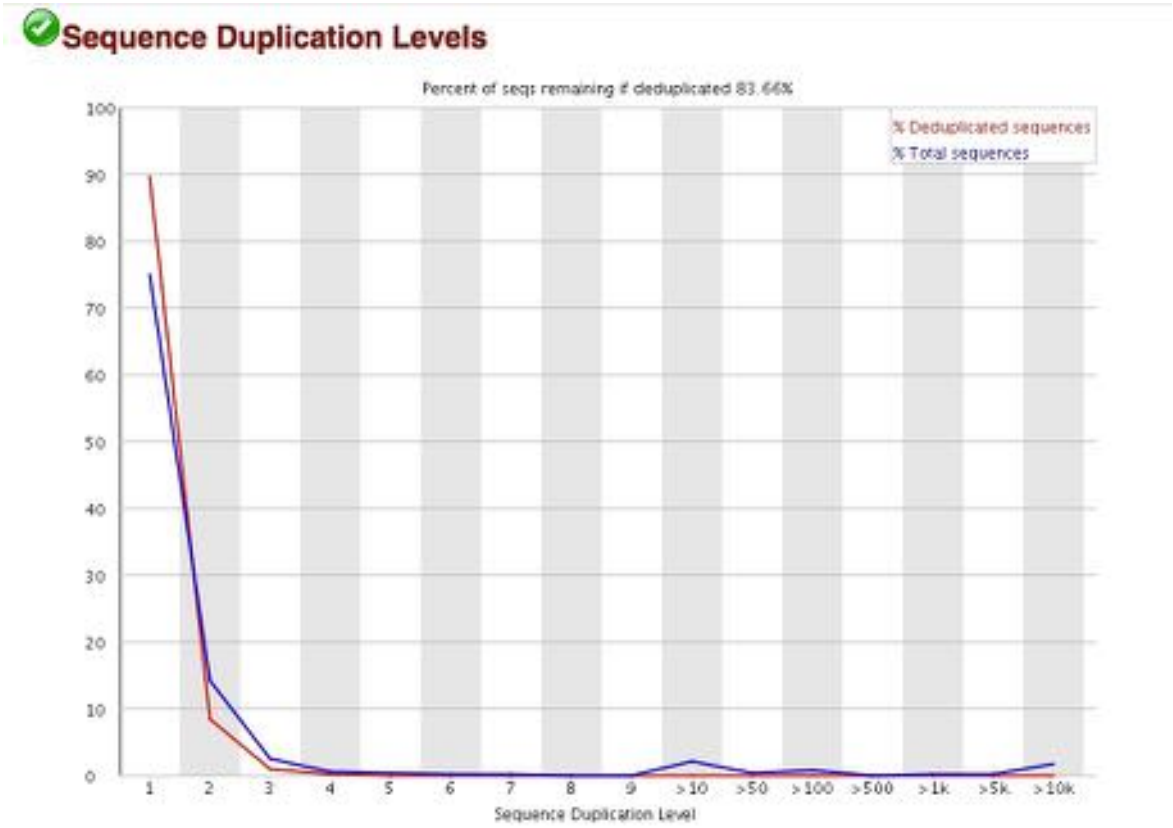
***de novo* assembly in Host-pathogen Systems:**

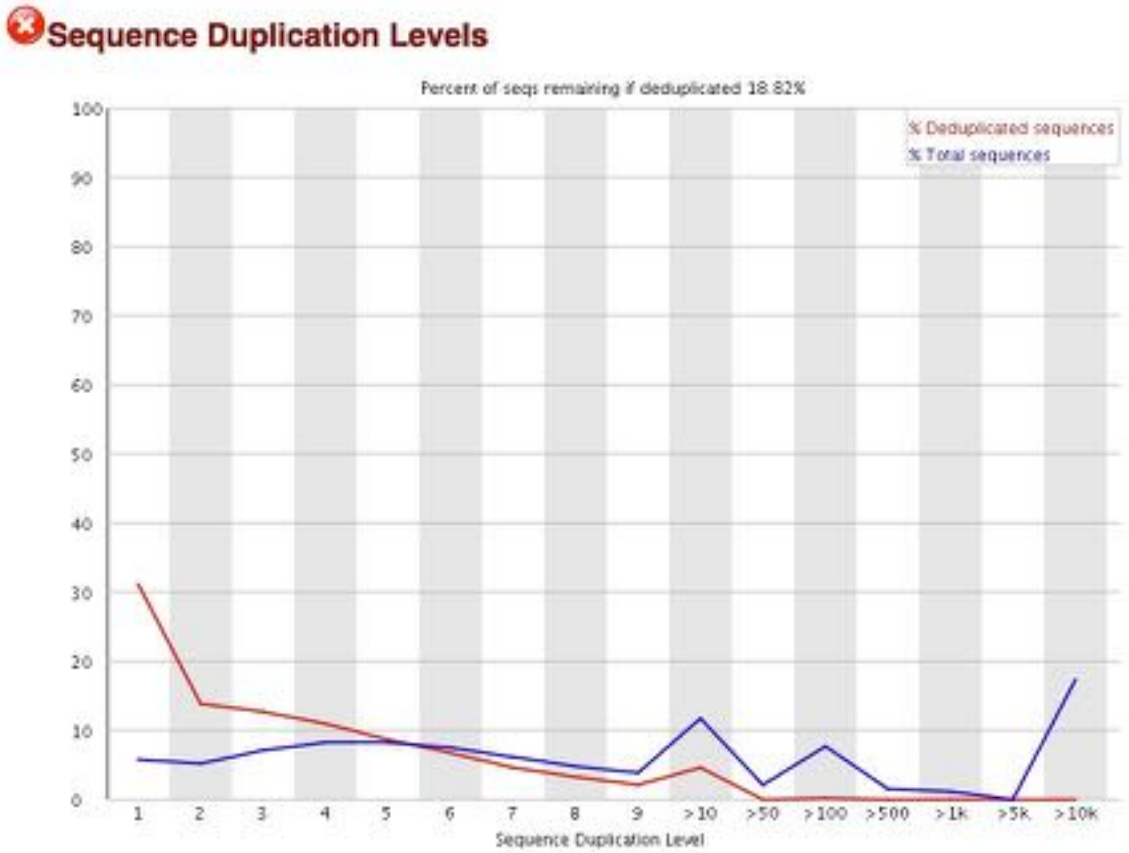
Blast assembled contigs against databases/genome

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

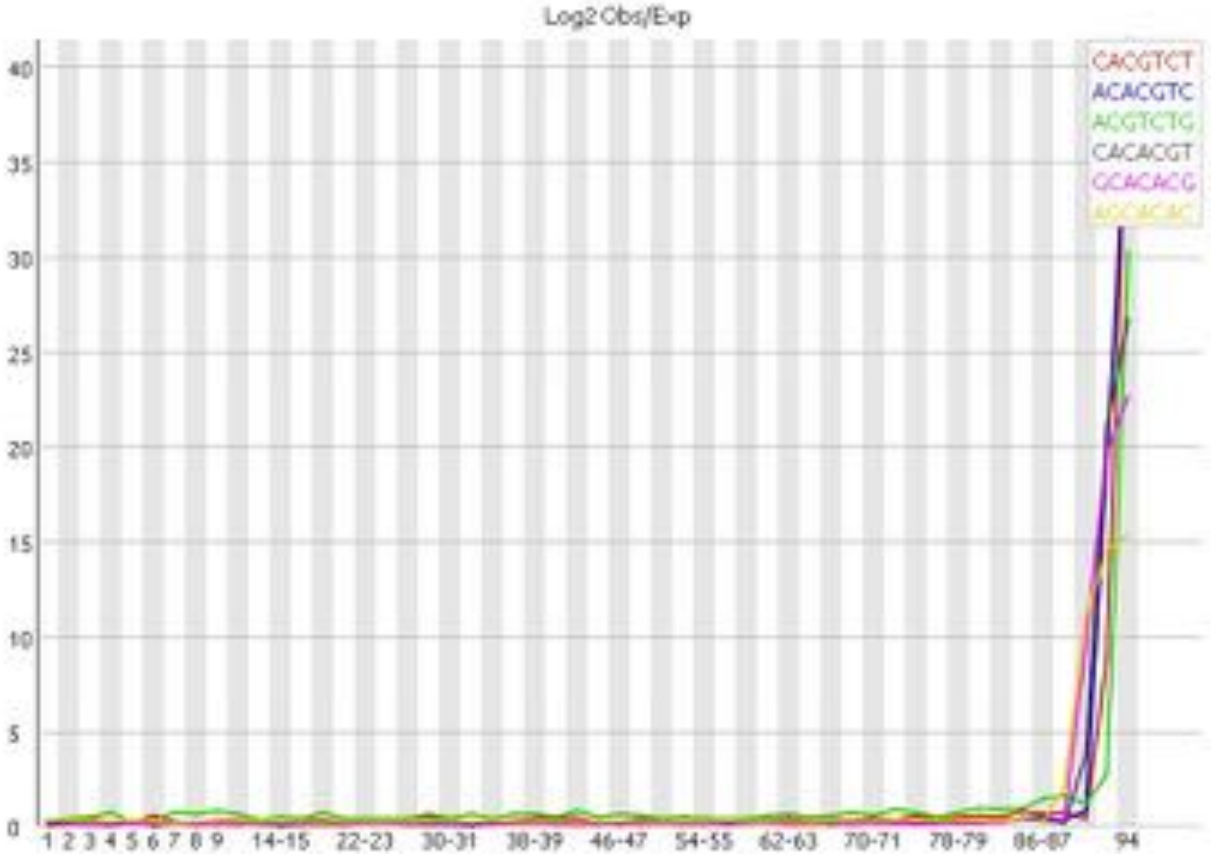
Filter based on GC content

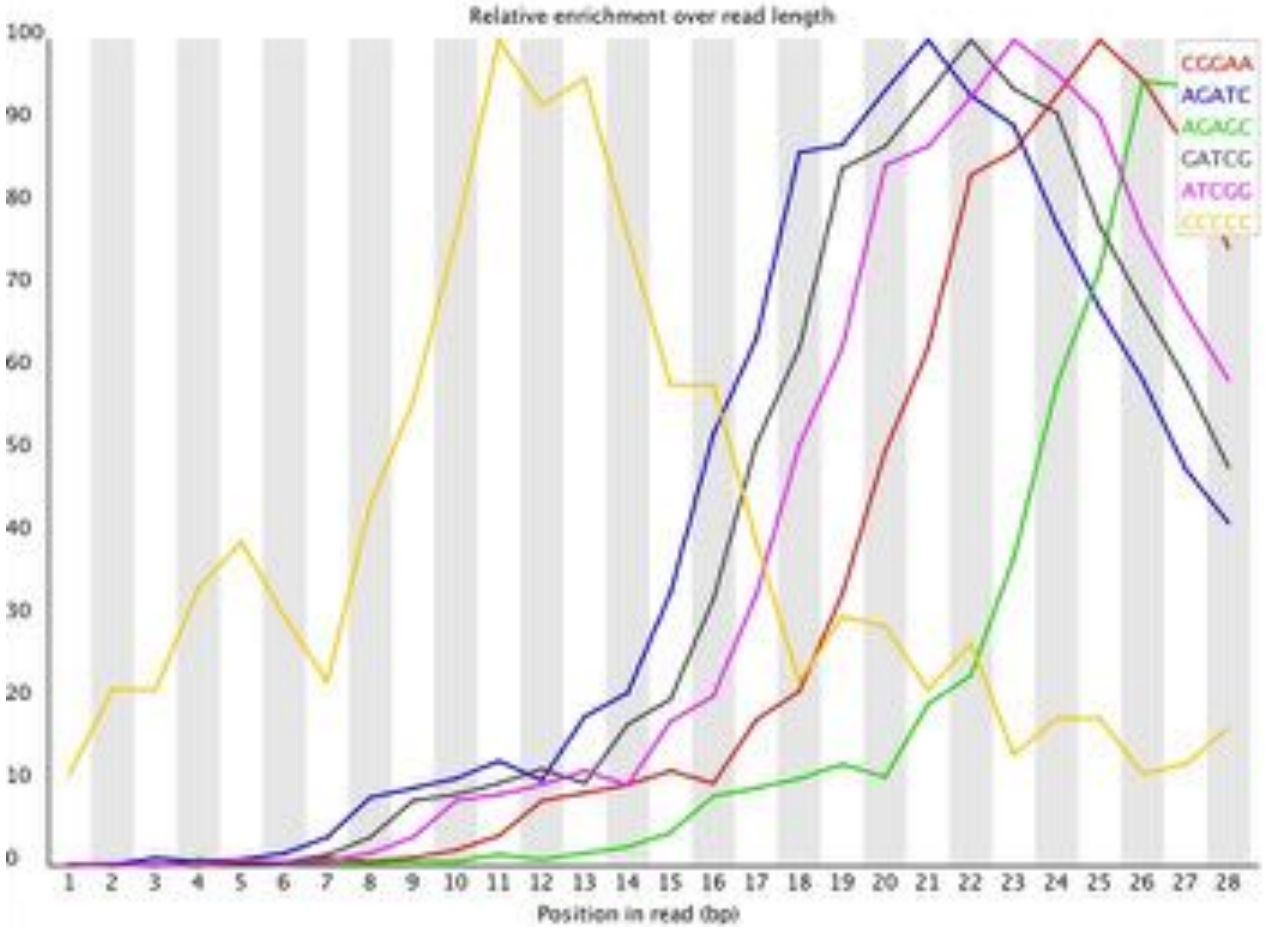
Duplication levels





Kmer content



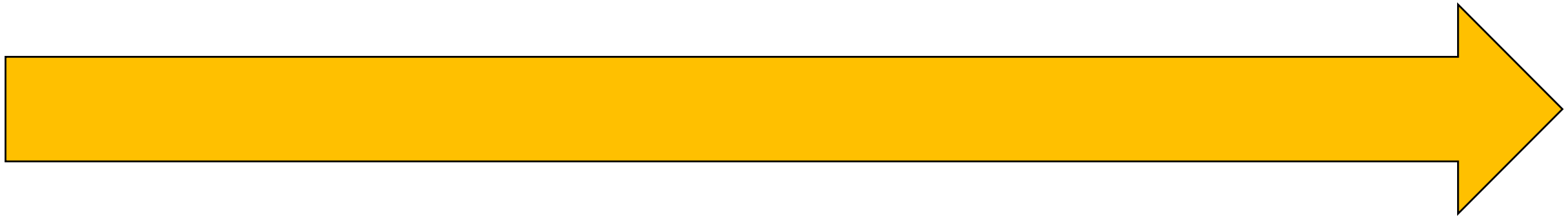


This module will issue a warning if any k-mer is enriched more than 3 fold overall, or more than 5 fold at any individual position.

How stringent do we need to be?

Stringent

No filtering



de novo assembly

Low coverage sequencing

Amplicon sequencing

Re-sequencing

RNA-seq

Take home massage

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