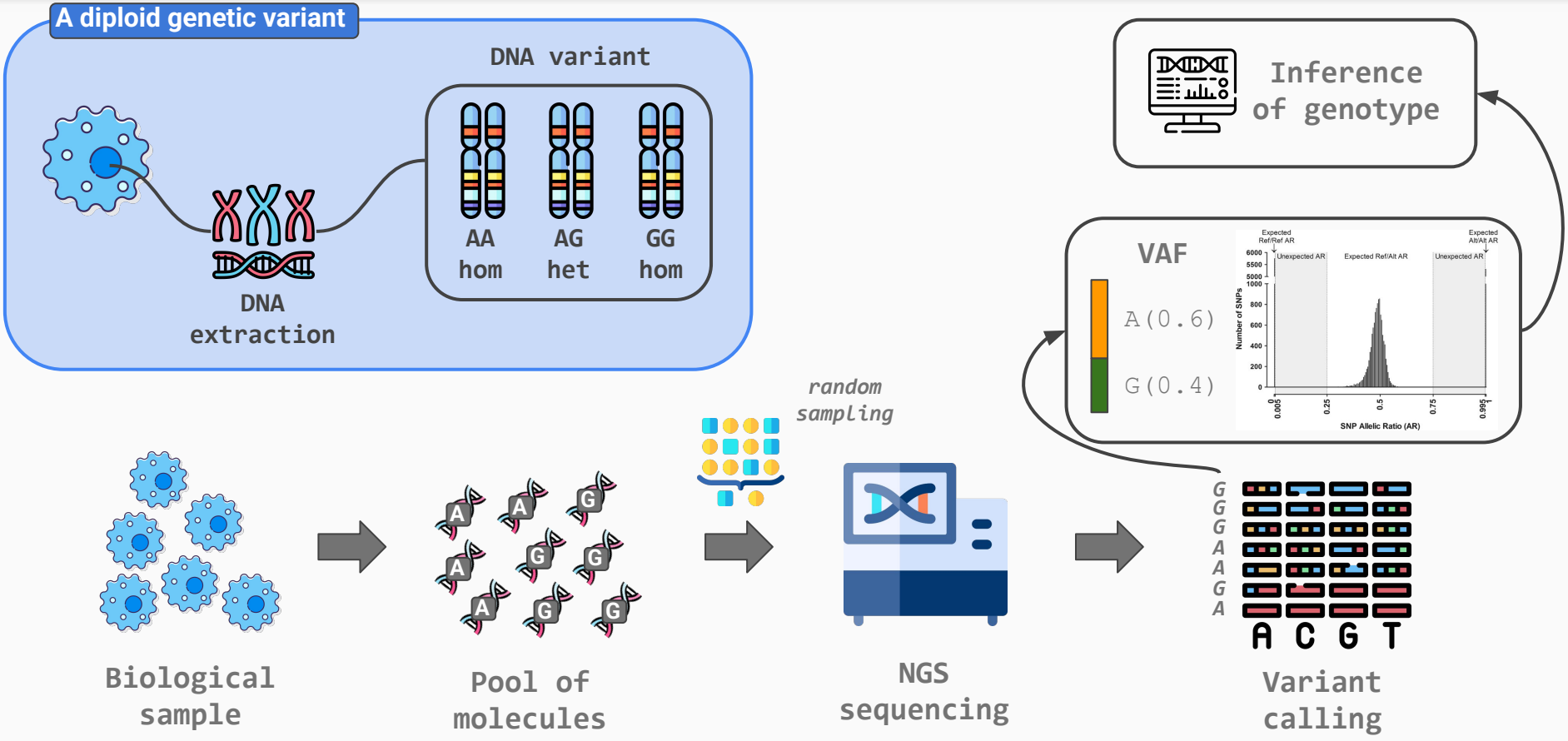
Introduction

NGS basic concepts and short-reads QC

Variant calling from NGS - what is this?



NGS basic concepts

NGS workflow

Essential steps in NGS data generation for short and long-reads

NGS Sequencing technologies

Next-Generation Sequencing

Massive parallel sequencing of DNA fragments

2nd generation

Short fragments (50-300bp)

Needs DNA fragmentation

3rd generation

Long fragments (10-100kb or more)

DNA molecules are directly used for sequencing

NGS short reads - library preparation

Template DNA

Fragmentation

Adapters

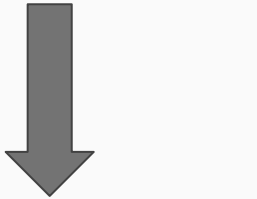
Clonal amplification

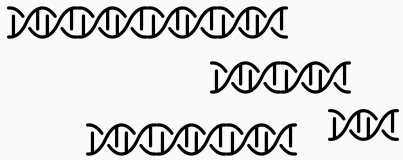
Sequencing

Long reads

UV light shearing

Enzymatic cut





Variable length fragments

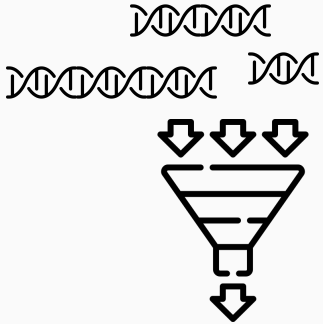
Short reads

Enzymatic cut

Sonication

Nebulization

Targeted amplification



Size selection

Uniform size small fragments

NGS short reads - library preparation

Template DNA

Fragmentation

Adapters

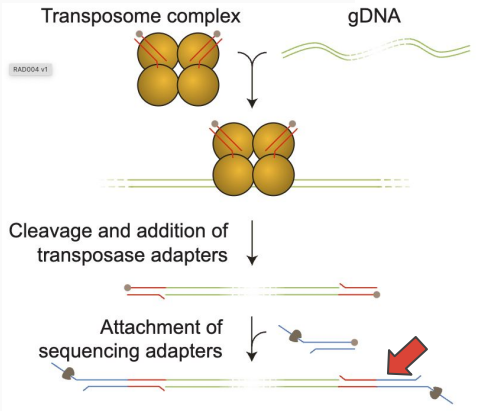
Clonal amplification

Sequencing

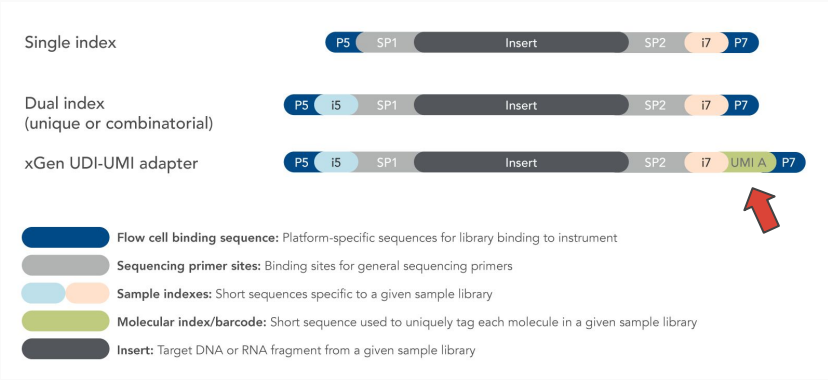
Long reads

PacBio

Oxford NANOPORE Technologies



Short reads



NGS short reads - library preparation

Template DNA

Fragmentation

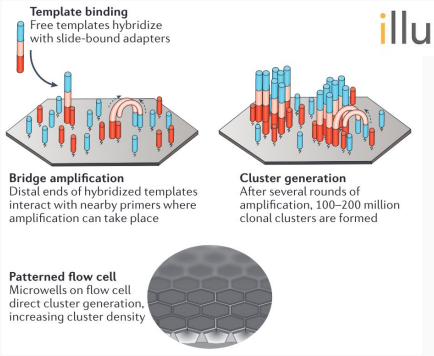
Adapters

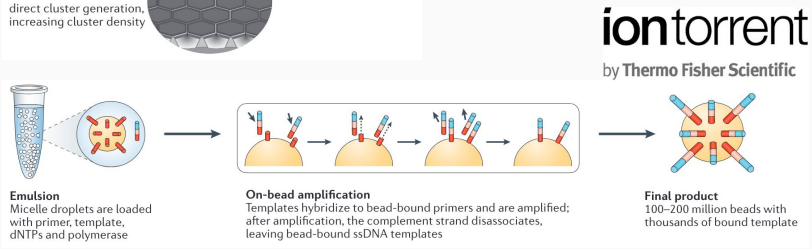
Clonal amplification

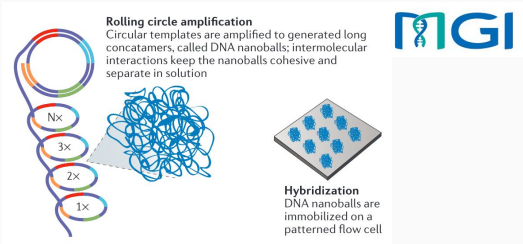
Sequencing

Long reads

Short reads







Goodwin et al., 2016

NGS short reads - library preparation

Template DNA

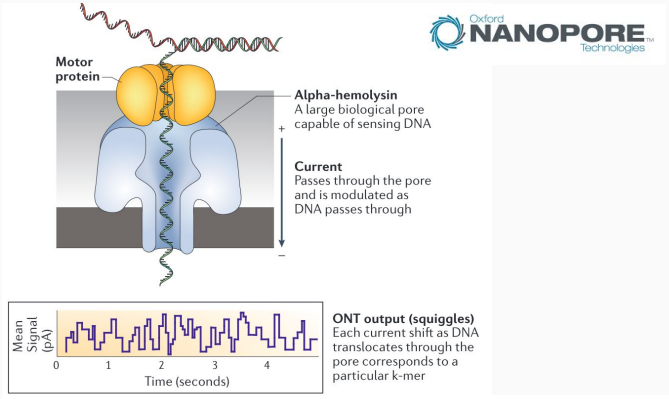
Fragmentation

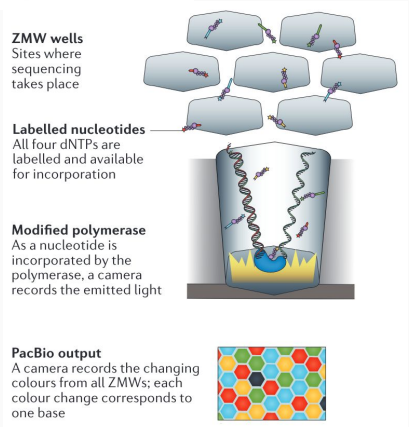
Adapters

Clonal amplification

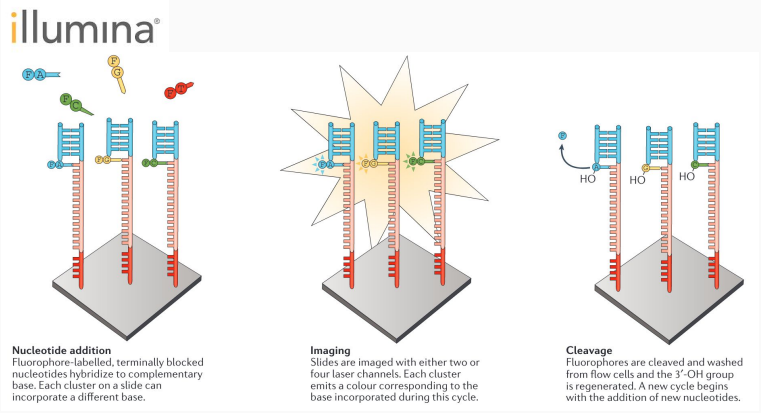
Sequencing

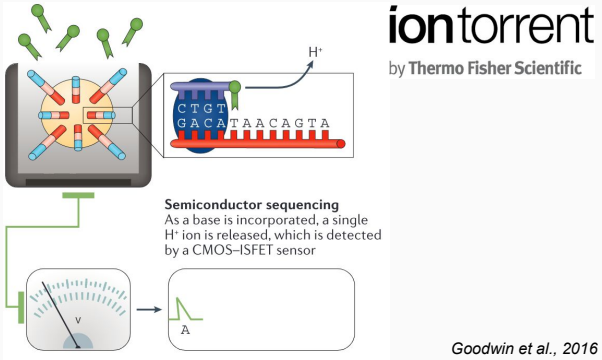
LONG READS





SHORT READS





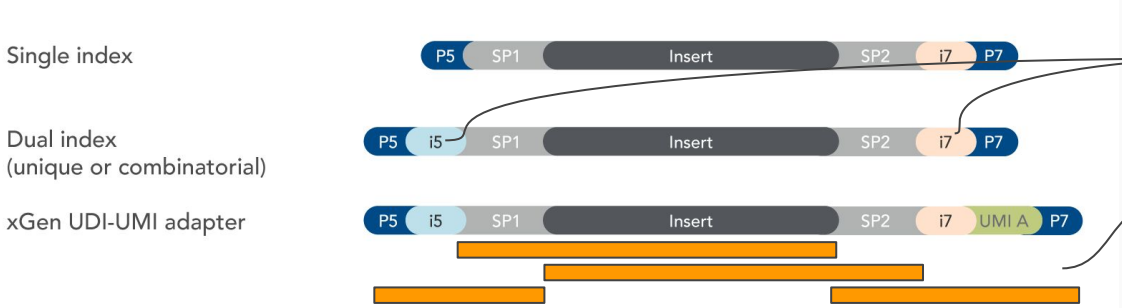
Goodwin et al., 2016

NGS reads structure

Understand the structure of sequencing libraries and the resulting data

- adapters
- read configuration
- UMIs

NGS data - understand library structure

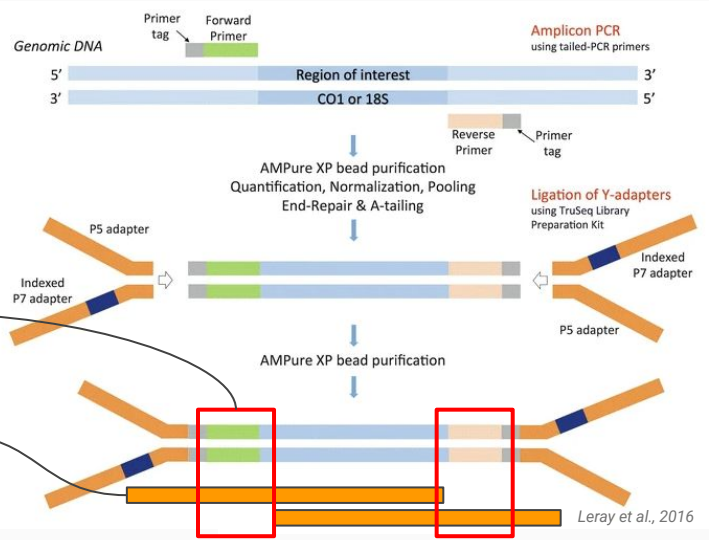


Library barcode(s) for multiplexing

Example of sequencing reads

Standard library preparation

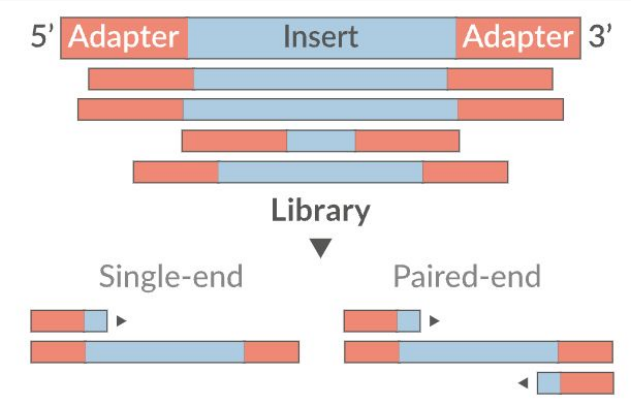
Amplicon targeted library



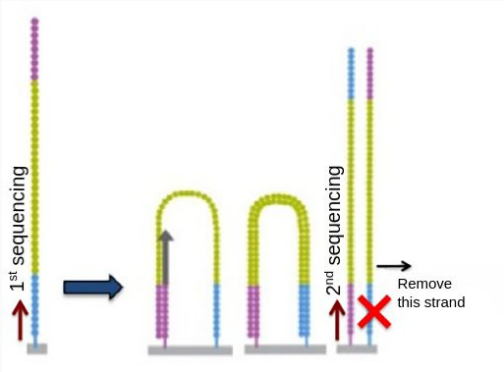
Read contains the PCR primer

Example of sequencing reads

NGS data - single end and paired-end short reads

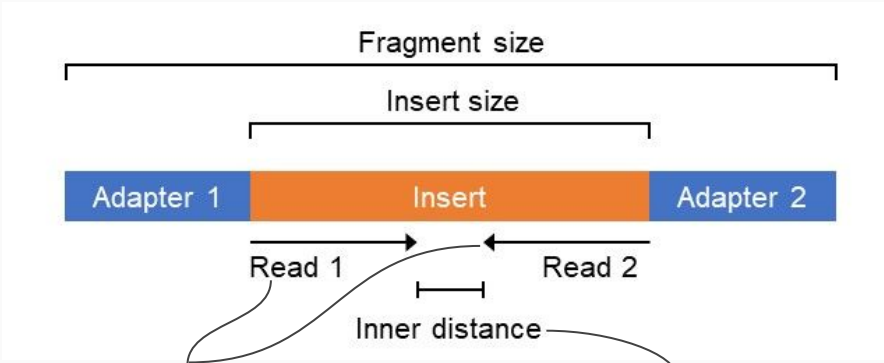


Illumina paired-end sequencing

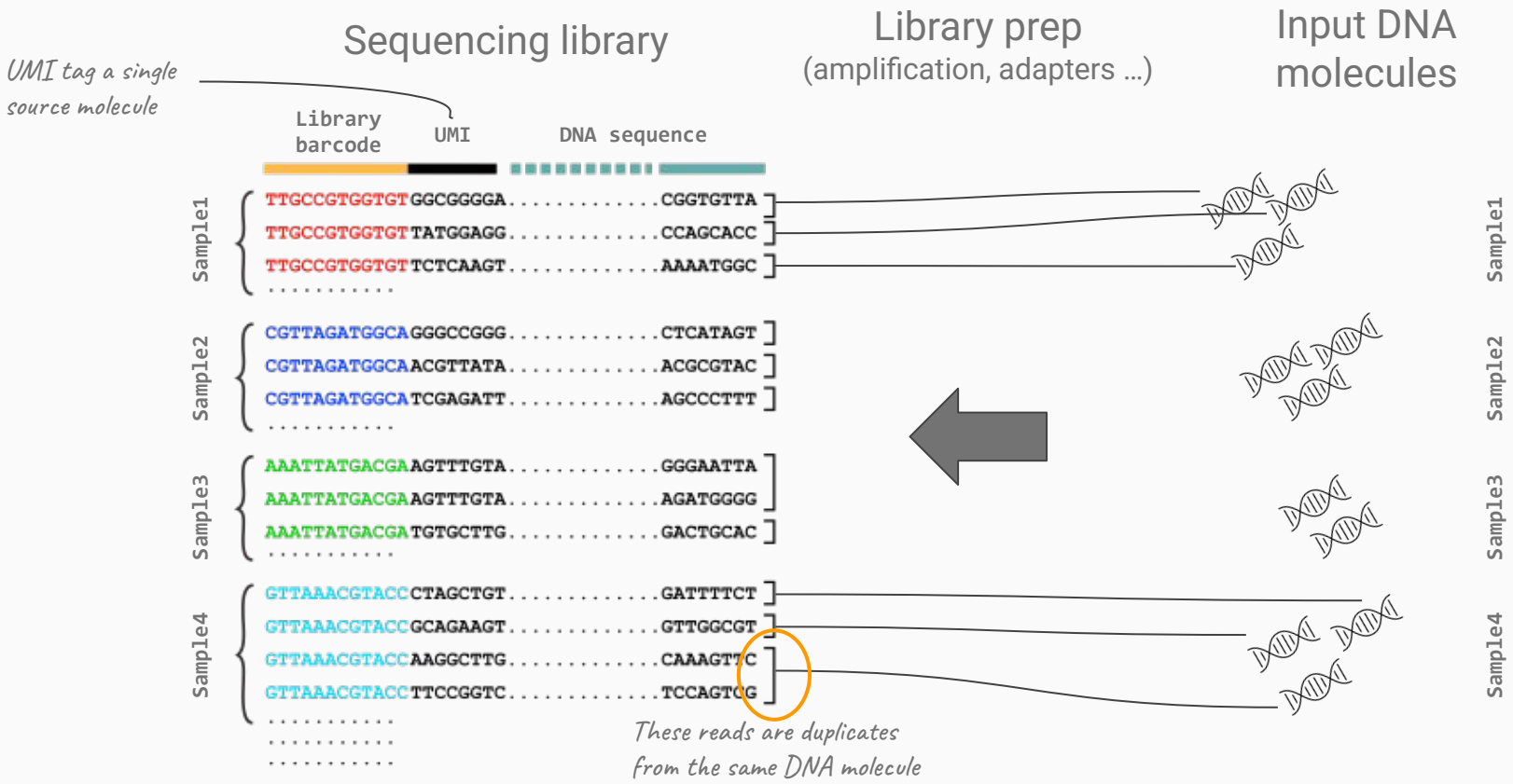


In **single end protocol**, each DNA molecule is sequenced once starting from a specific end.

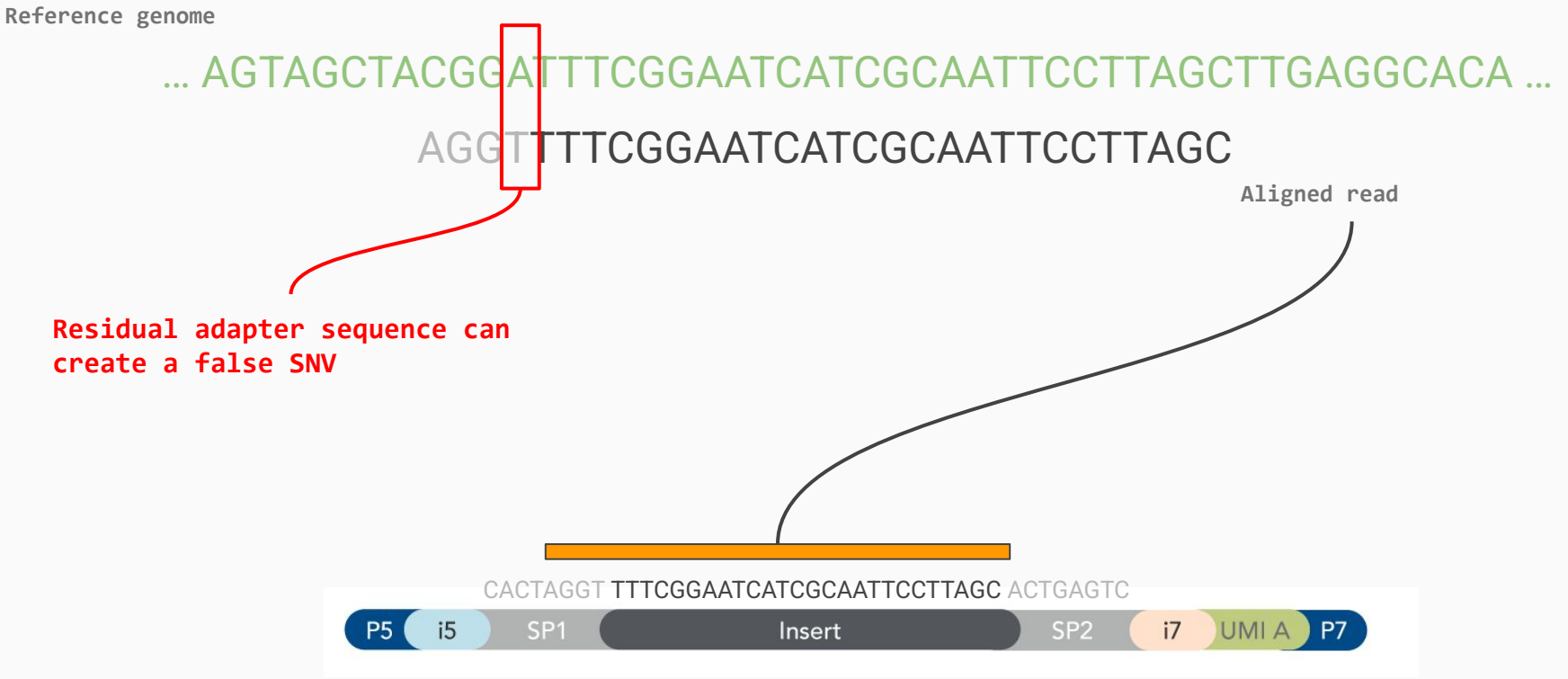
In **paired-end protocol**, it is sequenced twice starting from opposing ends



NGS data - understand UMIs



NGS data - library structure impacts downstream analysis



NGS data - library structure impacts downstream analysis

- A sequencing read derives from multiple molecular manipulations and contains specific elements besides the template DNA
- Knowing the elements and structure of sequencing reads is crucial for proper downstream processing
- Fail to remove adapters parts can results in increased error rate in variant calling
- When a library is generated by targeted amplification, no variant can be detected in the target primers sequences.

Data formats

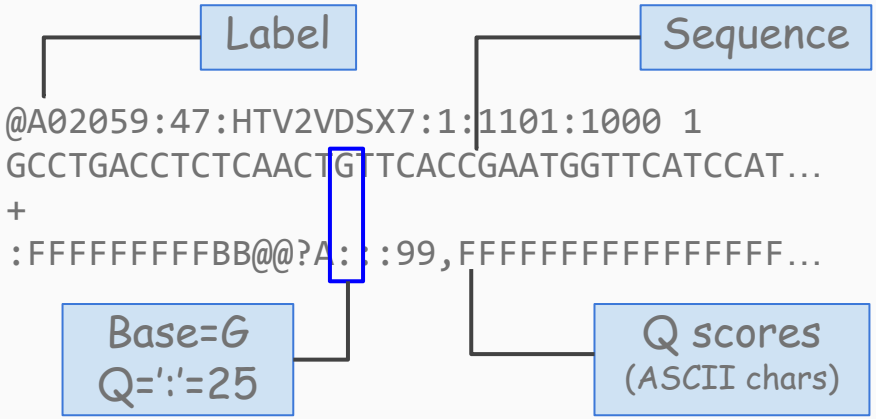
Main file formats

- FASTQ: sequences

Sequences - FASTQ files

Sequence and base quality of each read

```
@A02059:47:HTV2VDSX7:1:1101:2284:1000 1:N:0:TGTAATCGAC+NGCGGTGATC
GCCTGACCTCTCCATCAACTGTTCACCGAATGGTTCATCCATGTTGGGTTTTGCCTAAATCACTTTACATCATTAGAGTTTGAA
+
:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A02059:47:HTV2VDSX7:1:1101:4743:1000 1:N:0:TGTAATCGAC+NGCGGTGATC
GGCCTCGGCCAGGCACGGTGGCTCATGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCAGGCAGATCACCTGAGGTCAGGAGT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A02059:47:HTV2VDSX7:1:1101:5195:1000 1:N:0:TGTAATCGAC+NGCGGTGATC
GTTGAAGTCTGGAAAACCTTTTAGGATCCTTTAAAATGACTAAAATGTTAACATGGTTTGGATAATTATAATGCTTAACATCTA
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A02059:47:HTV2VDSX7:1:1101:8323:1000 1:N:0:TGTAATCGAC+NGCGGTGATC
ACCCTGACCCTTGCATGCCTTATTCCTCTTTCGCCTCATCTCCTCAGTGGGGTTATTAGTCTGATTCACTCTATCATTTCCTAA
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```



- **Label**
  unique identifier for the single read (instrument ID, run number, chip ID, tile, tile XY …)

- **Q scores**
  phred like quality score for each base sequenced encoded as ASCII character

Sequences - Base quality in FASTQ files

PHRED quality

$$Q = -10 \log_{10} P$$

$$P = 10^{\frac{-Q}{10}}$$

Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%

Quality encoding as ASCII characters

```
@SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
CAACGAGTTCACACCTTGGCCGACAGGCCCGGGTAA
+SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
BA@7>B=>:>>7@7@>>9=BAA?;>52;>:9=8.=A
```

Quality converted to single ASCII character

PHRED+33 ⇒ ASCII code ⇒ Char

Char	ASCII code	Phred quality
B	66	33
A	65	32
@	64	31
7	55	22

ASCII Table

Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
64	40	100	@	@	96	60	140	`	`
65	41	101	A	A	97	61	141	a	a
66	42	102	B	B	98	62	142	b	b
67	43	103	C	C	99	63	143	c	c
68	44	104	D	D	100	64	144	d	d
69	45	105	E	E	101	65	145	e	e
70	46	106	F	F	102	66	146	f	f
71	47	107	G	G	103	67	147	g	g
72	48	110	H	H	104	68	150	h	h
73	49	111	I	I	105	69	151	i	i
74	4A	112	J	J	106	6A	152	j	j
75	4B	113	K	K	107	6B	153	k	k
76	4C	114	L	L	108	6C	154	l	l
77	4D	115	M	M	109	6D	155	m	m
78	4E	116	N	N	110	6E	156	n	n
79	4F	117	O	O	111	6F	157	o	o
80	50	120	P	P	112	70	160	p	p
81	51	121	Q	Q	113	71	161	q	q
82	52	122	R	R	114	72	162	r	r

Short-reads QC

Reads cleaning

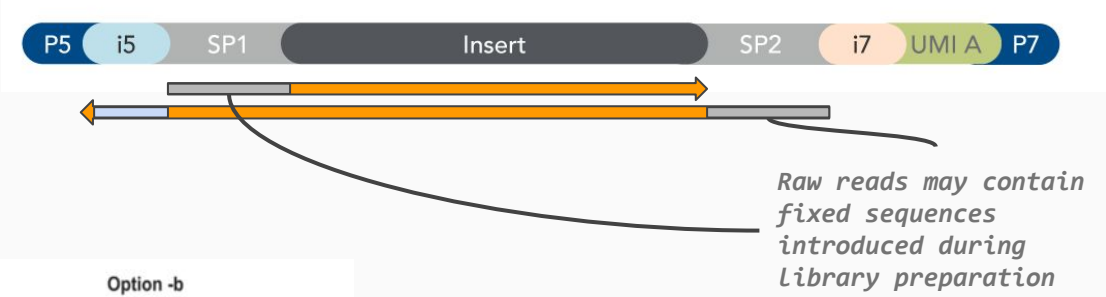
Remove unwanted sequences and poor quality bases

- adapter trimming
- quality trimming
- fixed length trimming

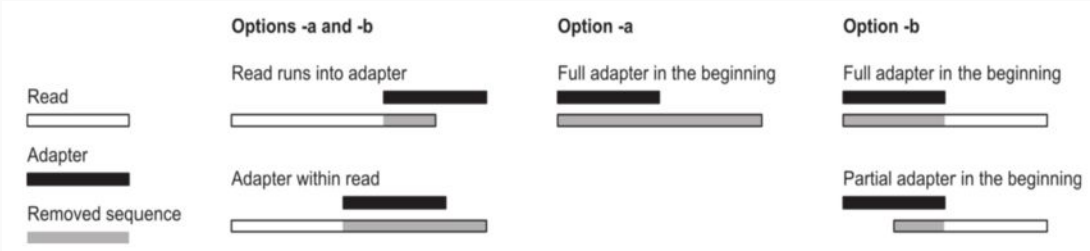
Useful tools

fastp, cutadapt

Adapter trimming - remove unwanted fixed sequences

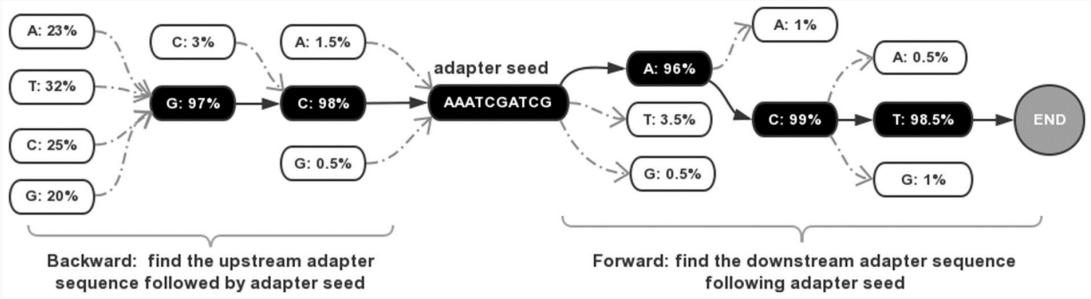


Example from [cutadapt tool]

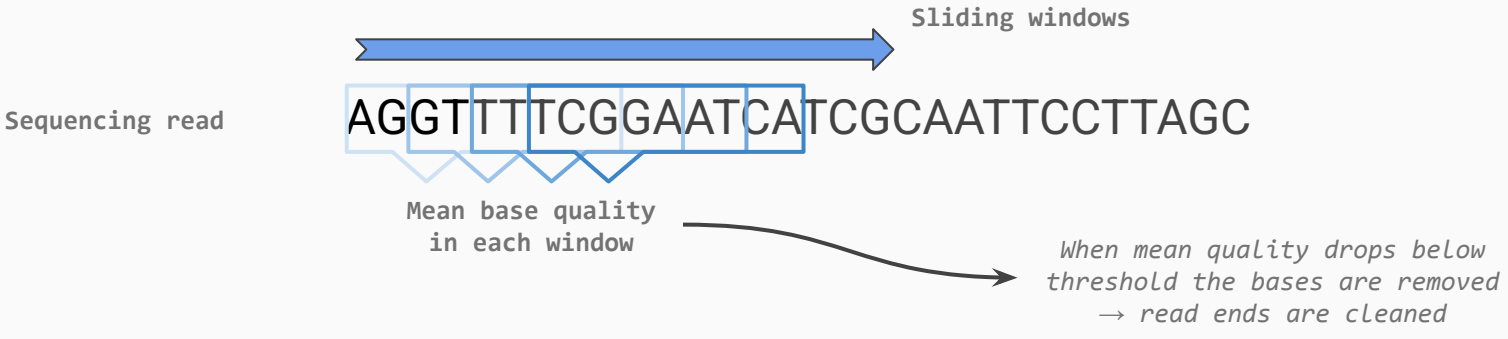


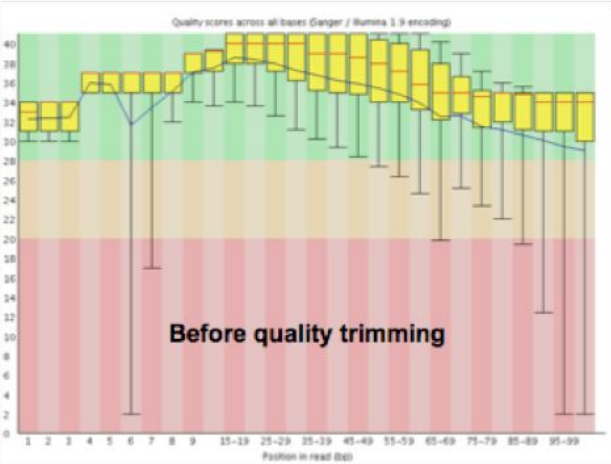
Trimming tools scan the read ends removing fixed sequences

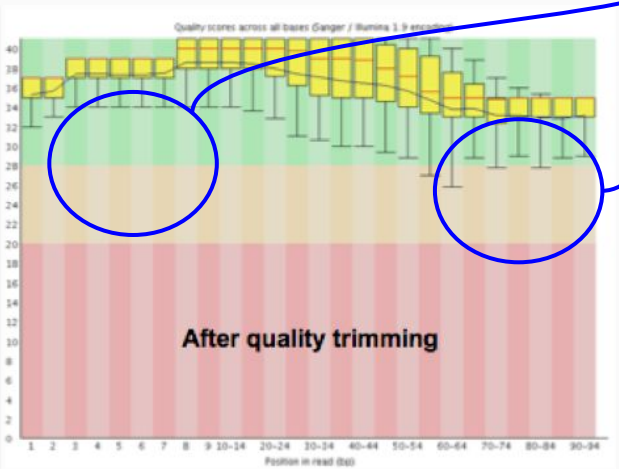
Based on adapter known sequences the algorithm search for partial matches and then extend the matching region to completely remove adapter sequences



Quality trimming - remove low quality bases from read ends

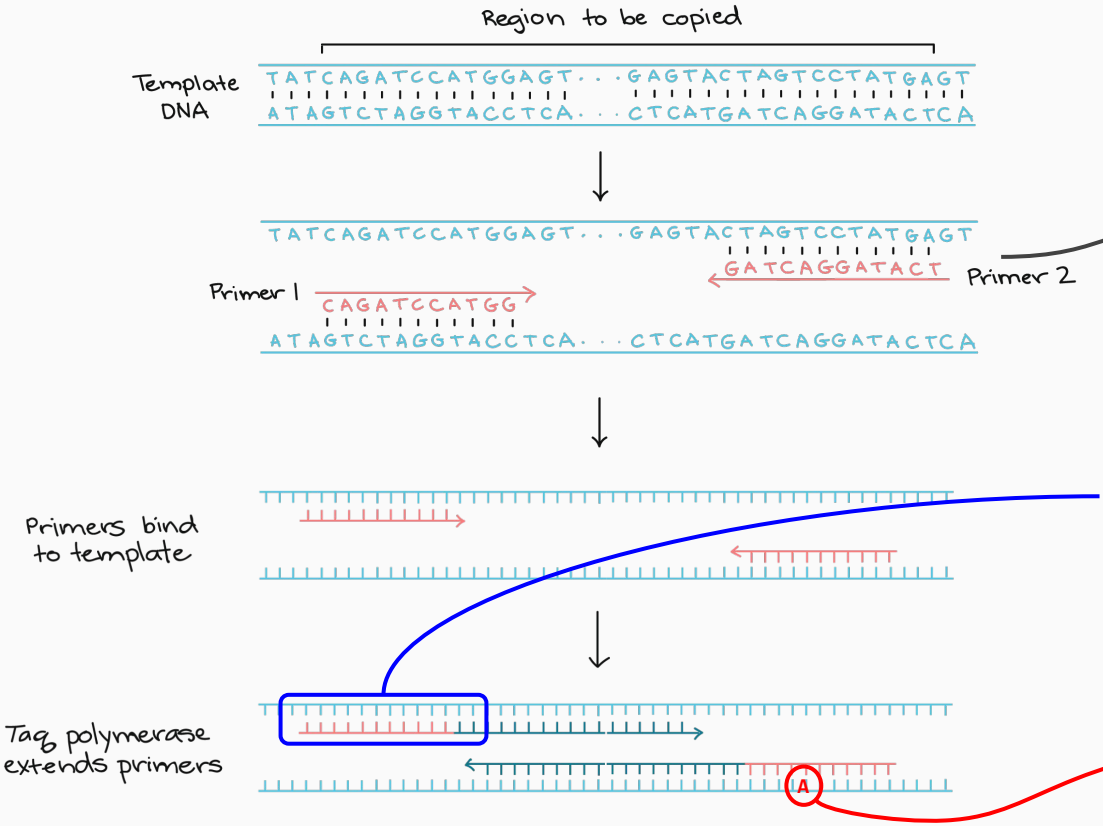






https://h3abionet.github.io/H3ABionet-SOPs/RNA-Seq-2-3.html

Fixed length trimming - remove a fixed amount of bases from read ends



A library is generated started from targeted amplification

Primers are used to amplify region(s) of interest

The primer sequence and NOT the real DNA sequence is included in the amplified fragment
→ primers should be trimmed!

A variant present in this region will not be detected (masked by the primer)

https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/biotechnology/a/polymerase-chain-reaction-pcr

Short-reads QC

- raw reads length and quality
- GC content
- adapter content
- mapping statistics

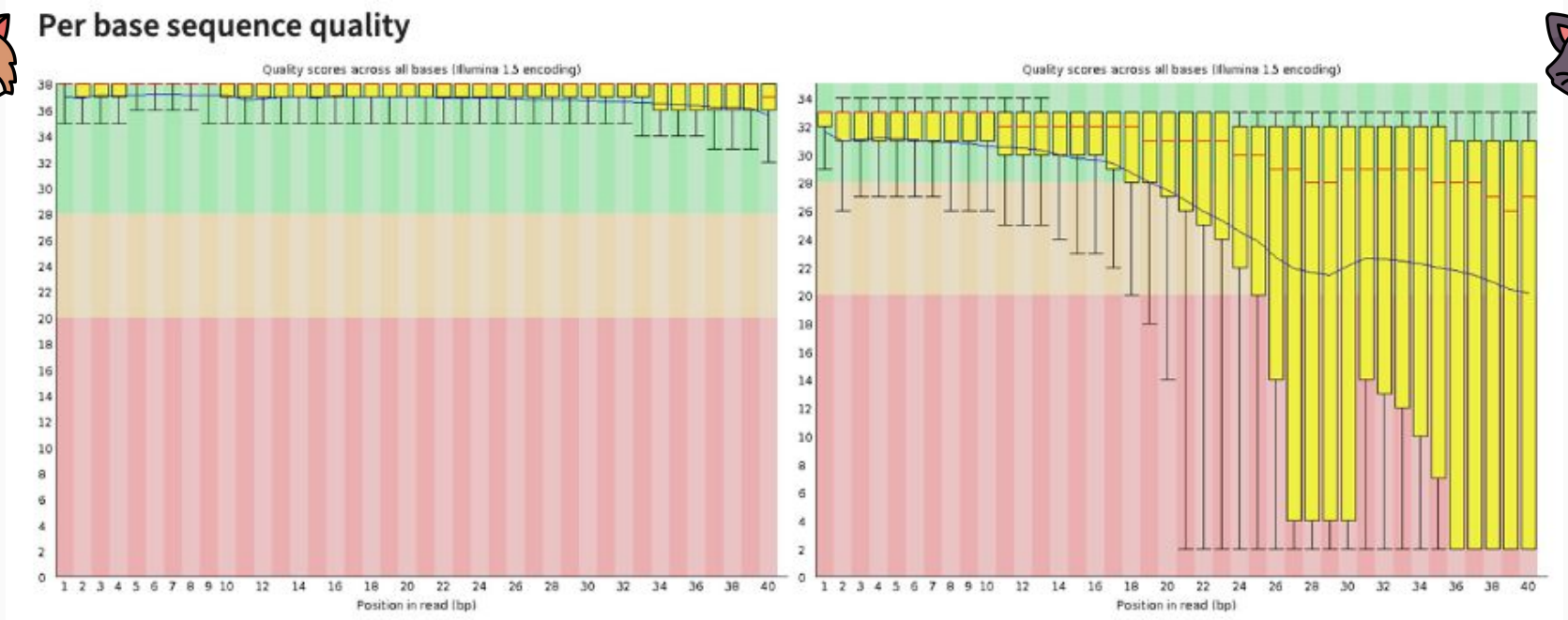
Useful tools

fastp, fastQC, samtools

Reads QC - Per base sequence quality across the reads

fastQC / FASTP

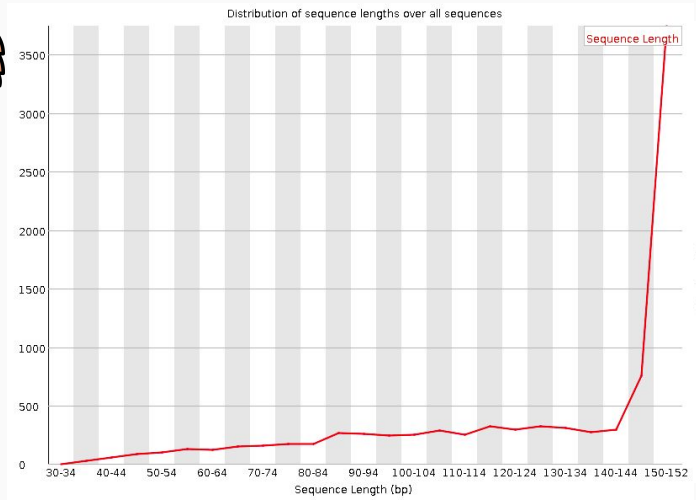
- Ideally, base quality should be >= 30 across all the read
- A little decrease in quality is expected toward the end of the read
- Base quality is considered during variant calling and can affect variant caller performances

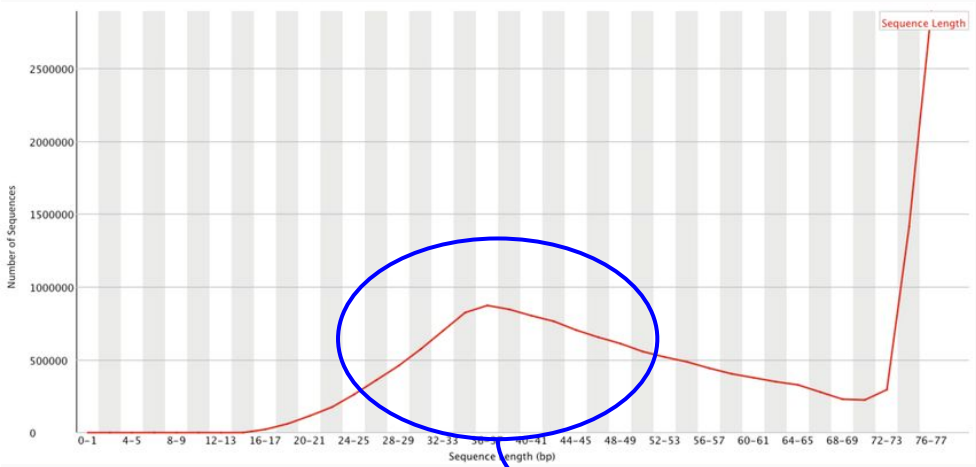


Read QC - sequence length distribution

fastQC / FASTP

- Most read should have the expected read length
- A small tail on the left is acceptable, usually indicate trimming was performed





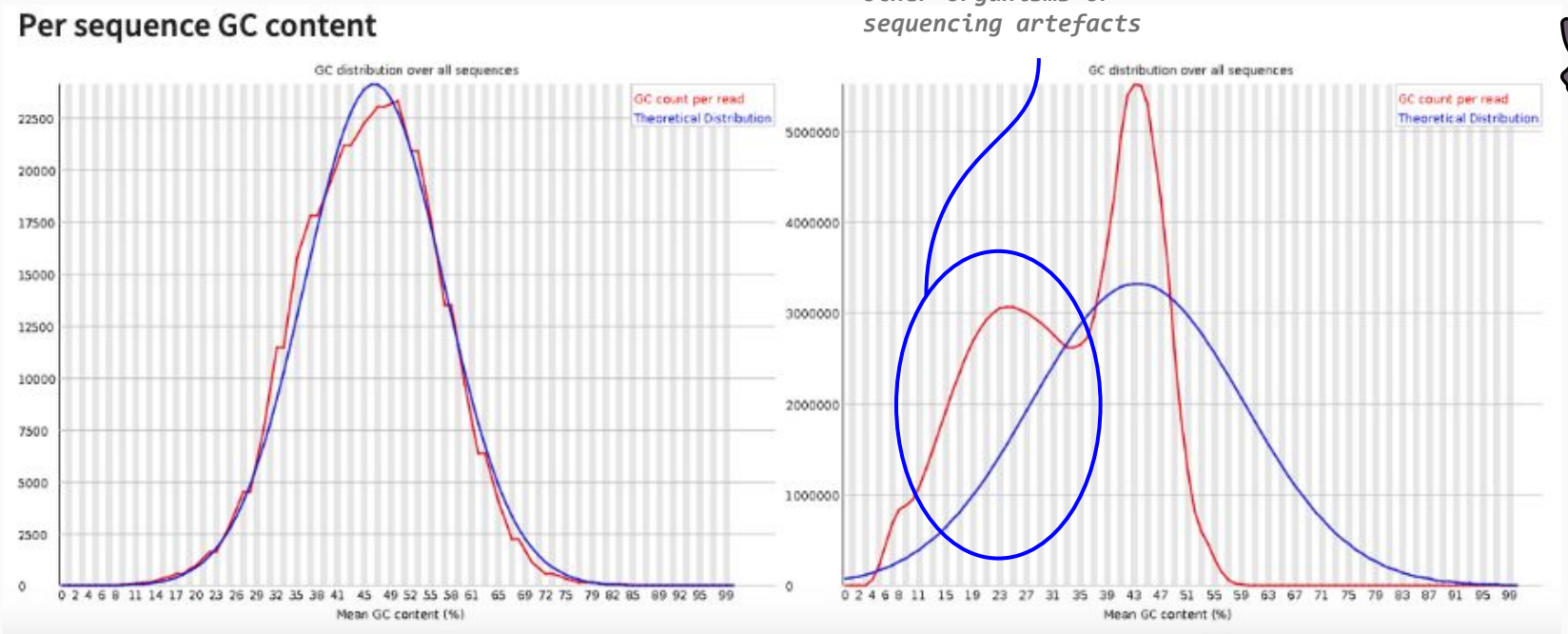
Sequencing problems, low quality reads that got trimmed, adapter dimers

Reads QC - reads GC content

- Distribution should follow the expected for the sequenced organism
- Peak around 45% for human genome samples

fastQC / FASTP

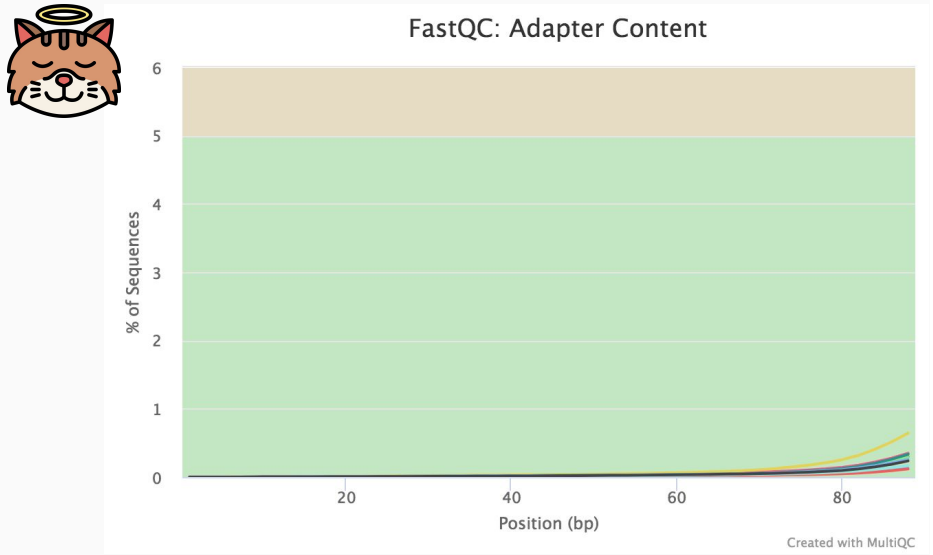
Contamination from DNA of other organisms or sequencing artefacts

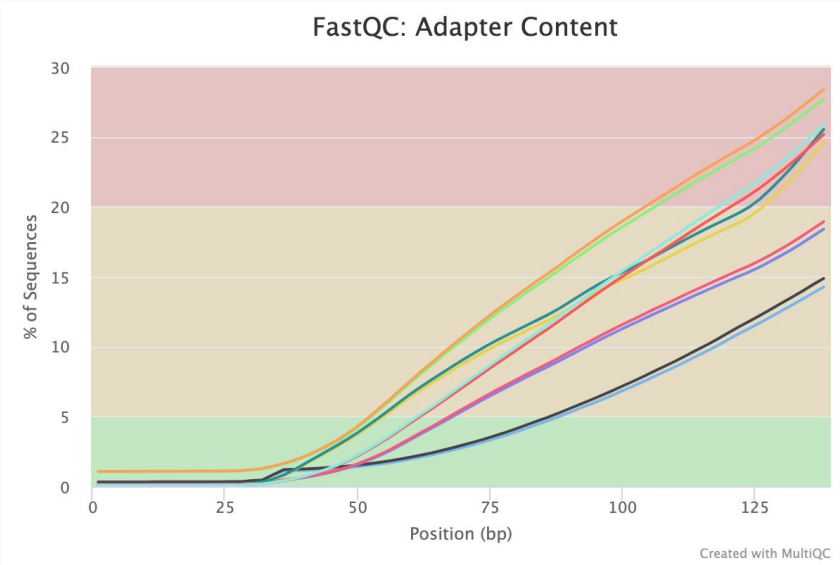


Read QC - residual adapter content

fastQC / FASTP

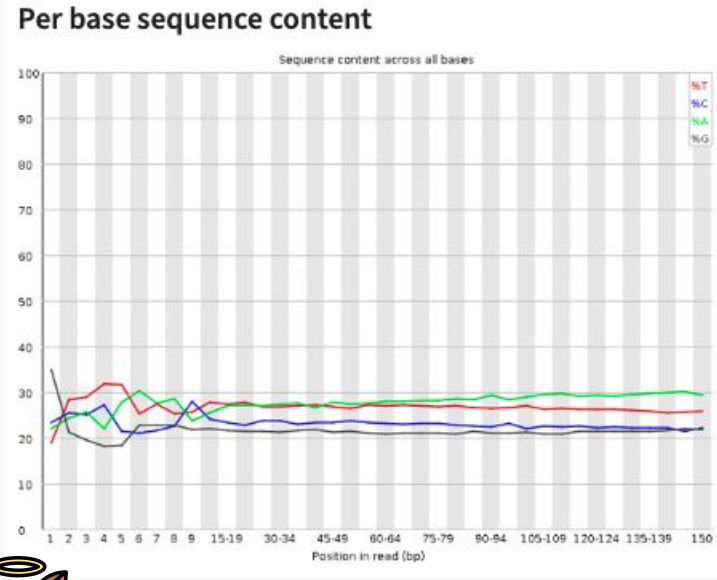
- Low presence of adapter sequences at the read ends
- Residual adapters can be cleaned by trimming



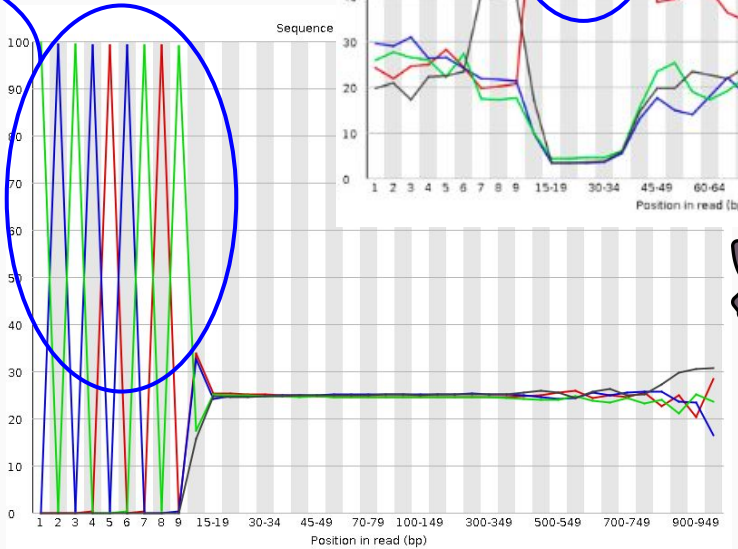


Reads QC - Per base sequence content across the reads

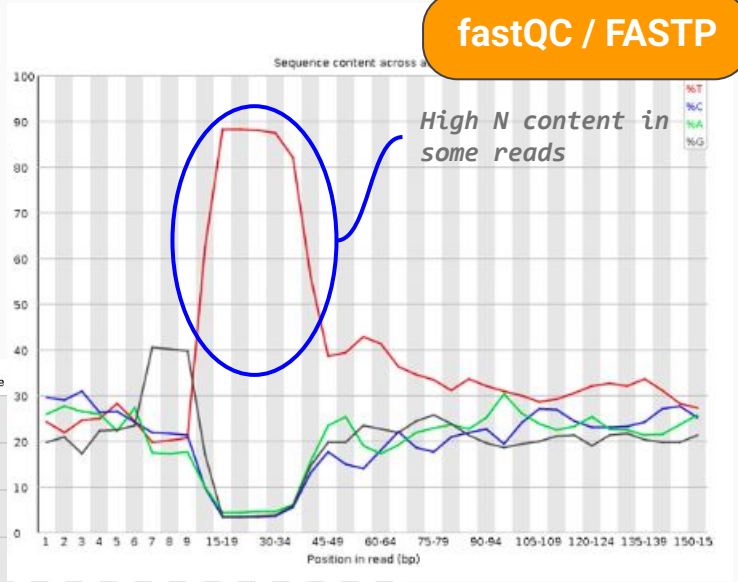
In most cases random sequences are expected
→ balanced base composition across the reads



Residual fixed elements or sequencing artefacts

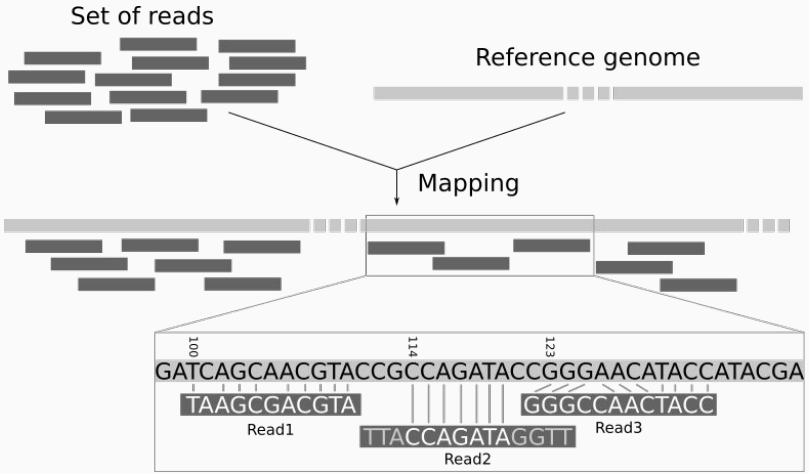


fastQC / FASTP



High N content in some reads

Aligned reads QC - mapping statistics



samtools flagstat

All libraries

- Mapping quality distribution
  (>= 30 for good mapping)
- Fraction of mapped reads
  (>= 90% in good samples)

Paired-end libraries

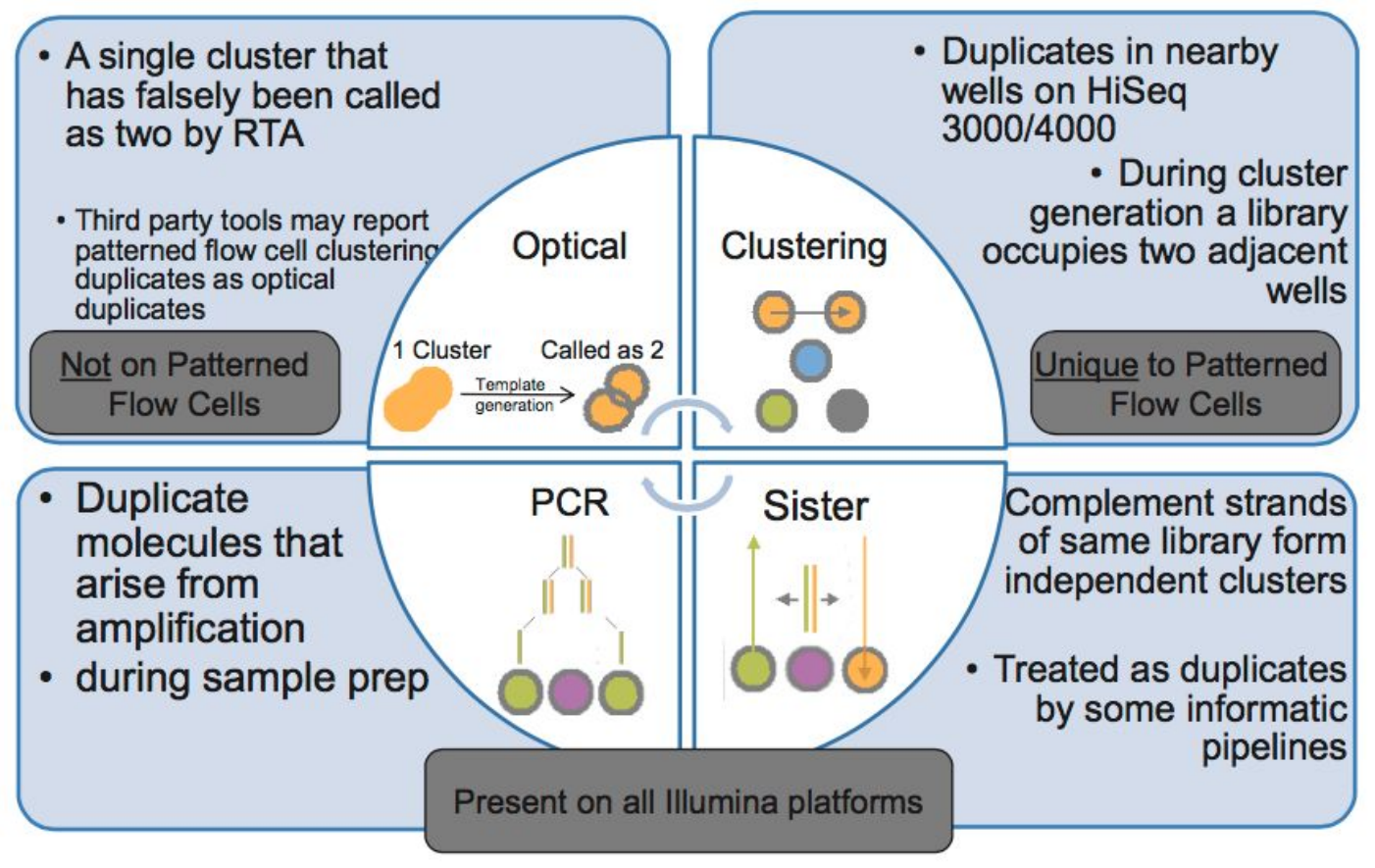
- Insert size distribution
  (distance between F/R read)
- Fraction of reads with a proper pair
  (>= 90% usually)

Reads processing and sample QC

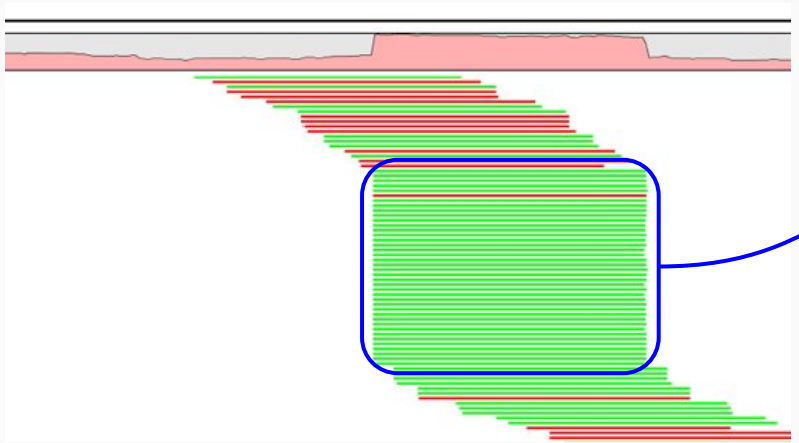
Further clean of sequencing artifacts and check for sample contamination

- duplicated reads
- UMI decomposition
- sex check
- contamination estimates

PCR and optical duplicates



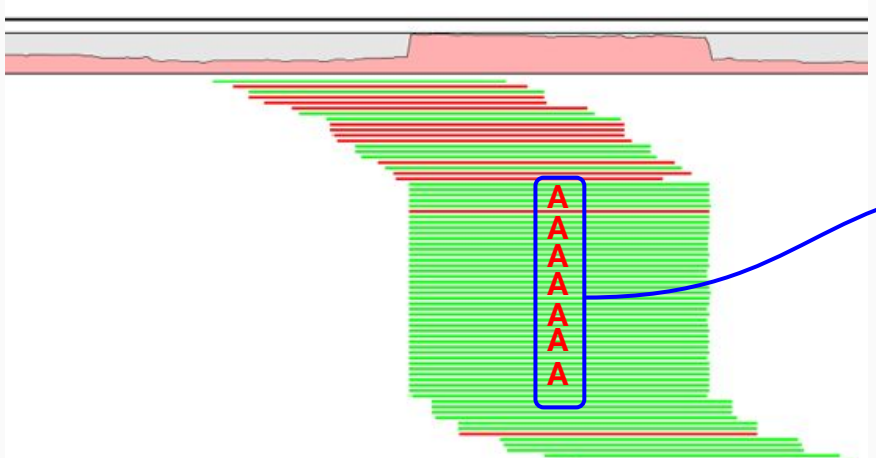
PCR or optical duplicates



samtools flagstat

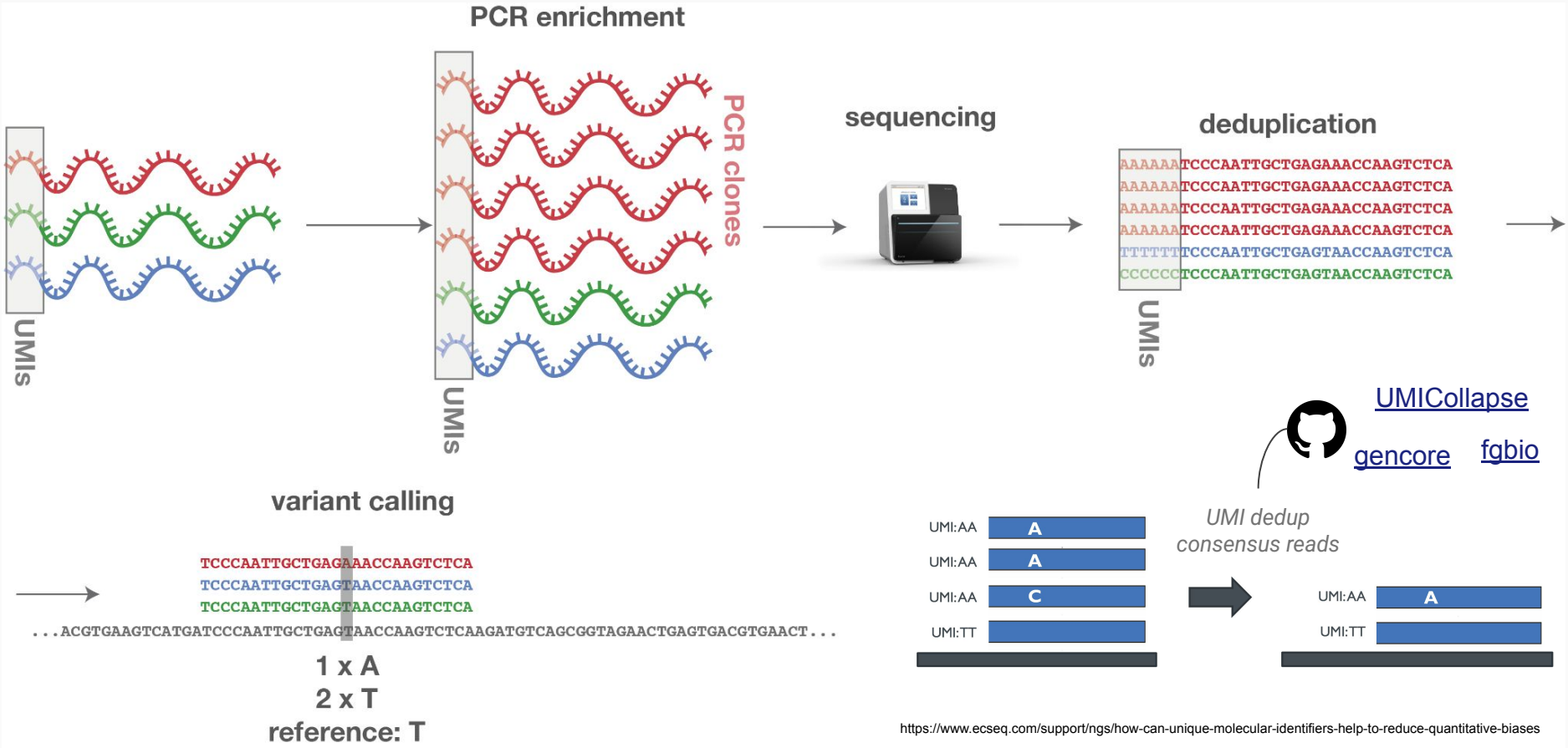
Duplicated reads
A pile or reads that actually represent multiple identical copies originated from the same DNA molecule!

They can be identified as group of reads that have identical start/end points. They are marked and ignored during variant calling. Available tools: picard, samblaster



Duplicated reads may affect genotyping when the duplicated molecule contains an error / SNVs

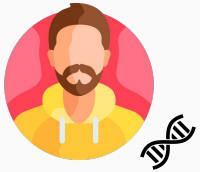
UMI based read deduplication to increase accuracy

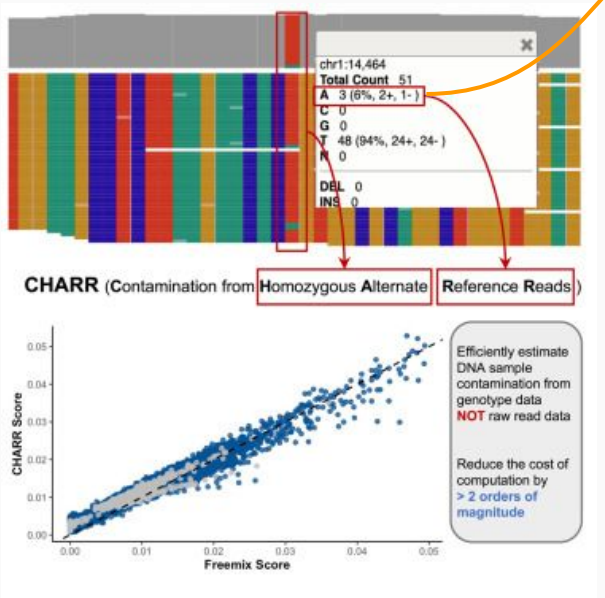


https://www.ecseq.com/support/ngs/how-can-unique-molecular-identifiers-help-to-reduce-quantitative-biases

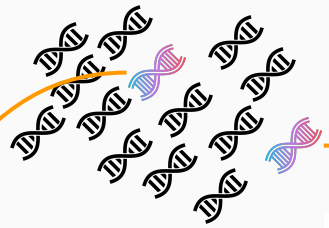
Additional sample-level QC - detect contamination

Target sample



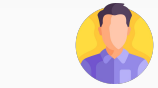


Increased fraction of “unexpected” alleles in homozygous genotypes

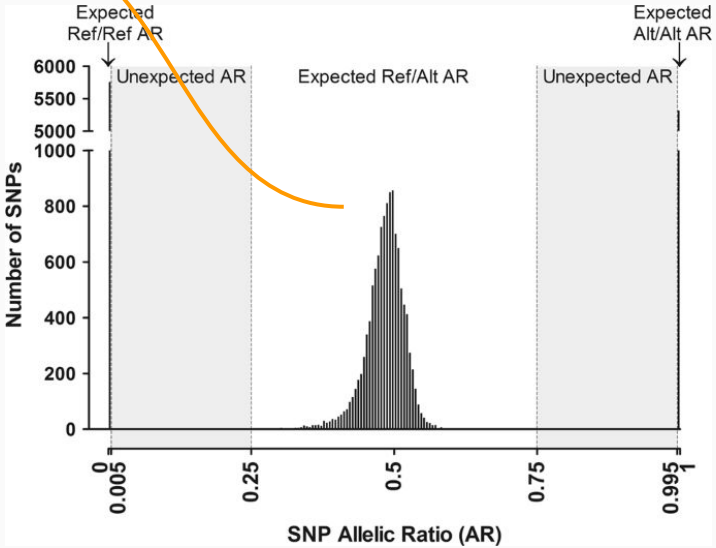


Sequenced material

Contaminating sample

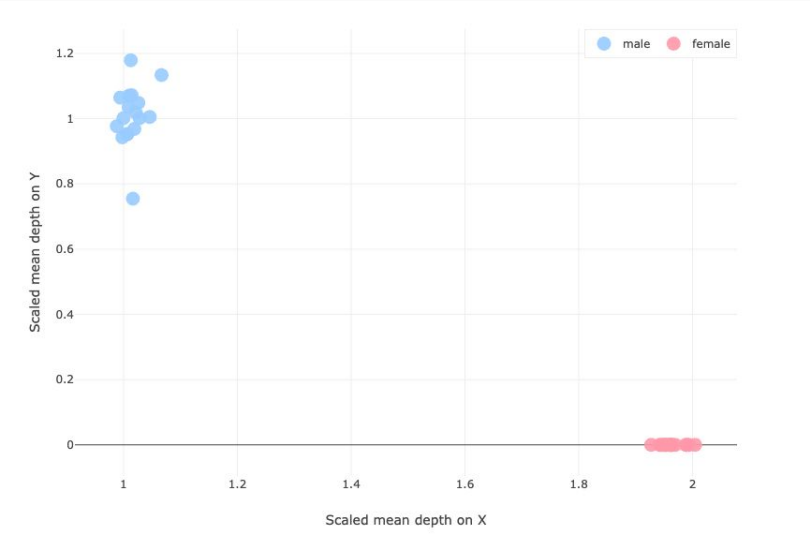


bcftools stat



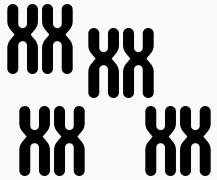
Increased N of heterozygous genotypes
Increased het GT with outlier AR

Additional sample-level QC - sex inference



Scaled mean depth

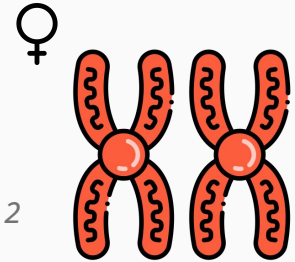
$$\frac{\textit{mean depth chrX}}{\textit{mean depth autosomes}} x\ 2$$



autosomes

mean depth
expected depth for a diploid chromosome

Female
Diploid X
scaled depth = 2



Male
Haploid X
scaled depth = 1
Haploid Y
scaled depth = 1