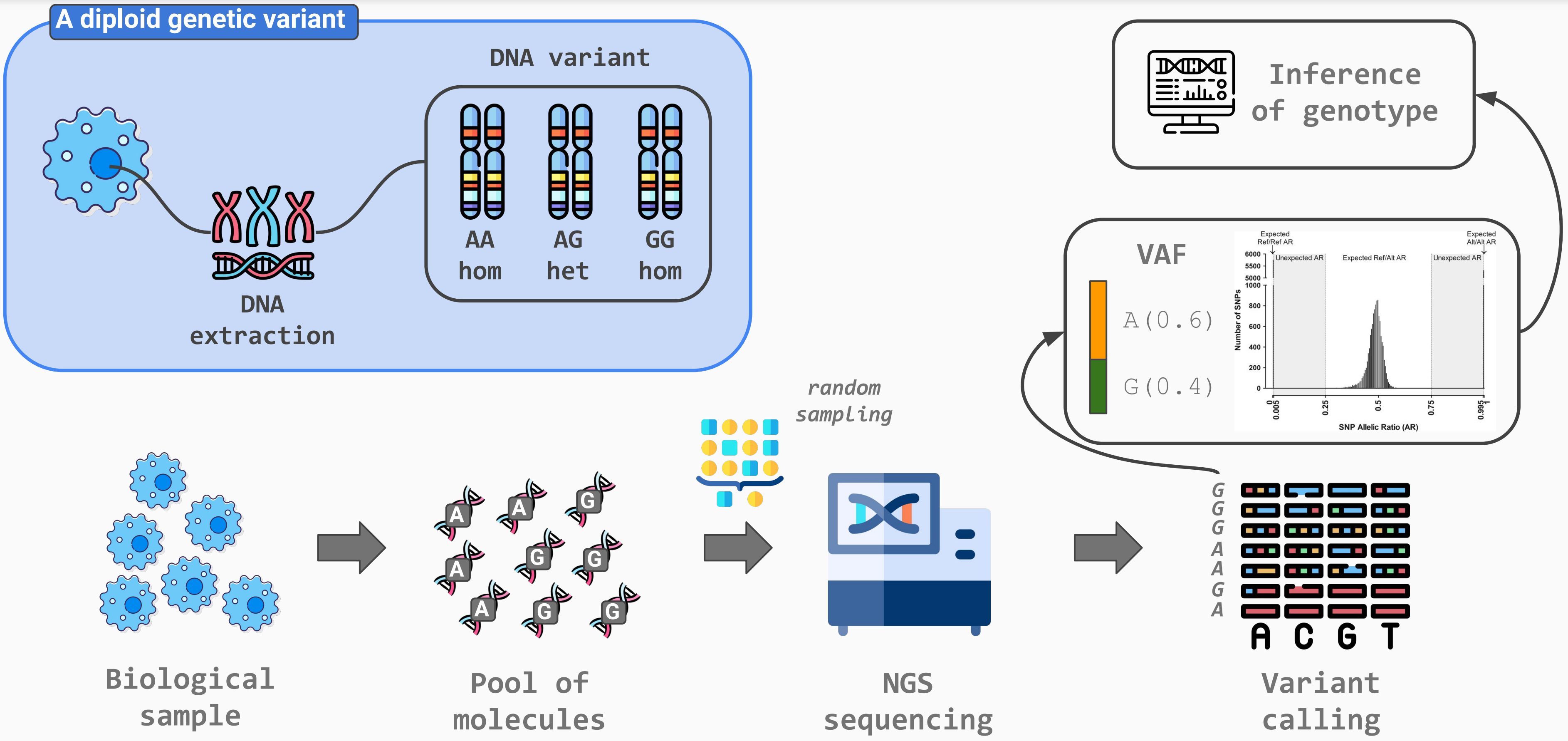
Introduction

NGS basic concepts and short-reads QC

Variant calling from NGS - what is this?



A diploid genetic variant

DNA variant

AA hom

AG het

GG hom

DNA extraction

Inference of genotype

VAF

A(0.6)

G(0.4)

random sampling

Biological sample

Pool of molecules

NGS sequencing

Variant calling

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

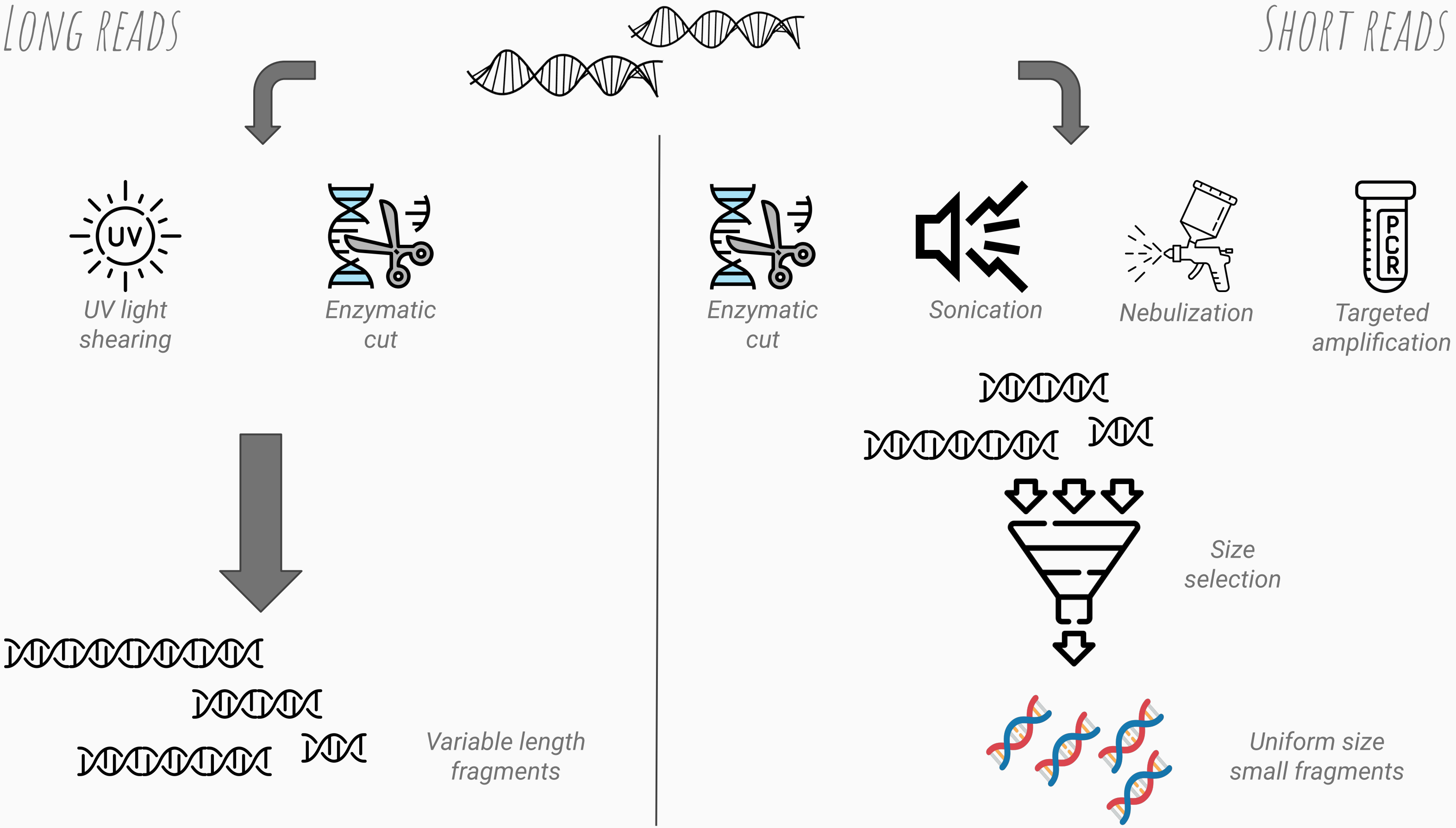
Template DNA

Fragmentation

Adapters

Clonal amplification

Sequencing



NGS short reads - library preparation

Template DNA

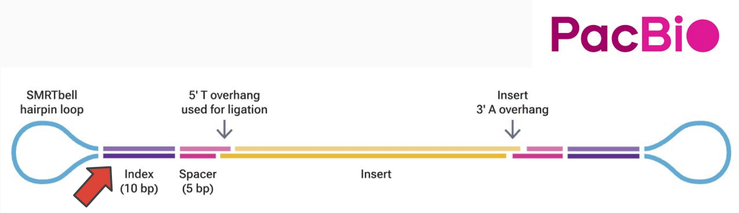
Fragmentation

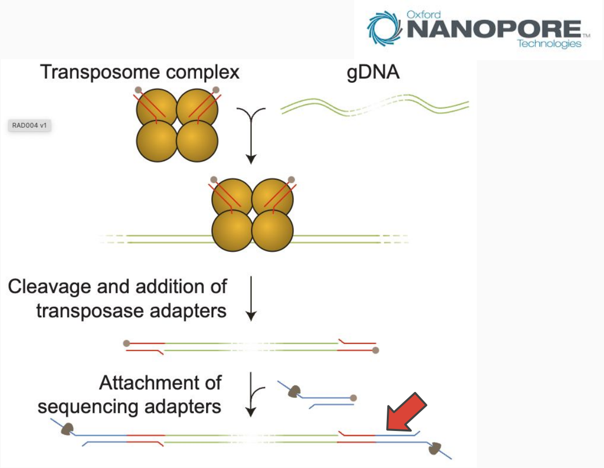
Adapters

Clonal amplification

Sequencing

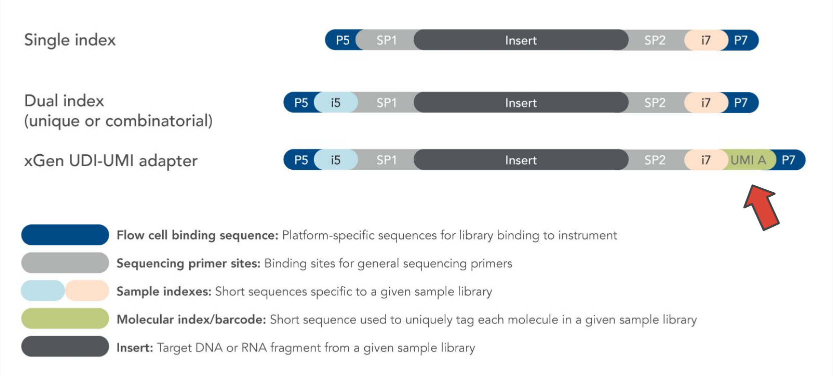
LONG READS





Oxford NANOPORE Technologies
Transposome complex
gDNA
RAD004 v1
Cleavage and addition of transposase adapters
Attachment of sequencing adapters

SHORT READS



- **Flow cell binding sequence:** Platform-specific sequences for library binding to instrument
- **Sequencing primer sites:** Binding sites for general sequencing primers
- **Sample indexes:** Short sequences specific to a given sample library
- **Molecular index/barcode:** Short sequence used to uniquely tag each molecule in a given sample library
- **Insert:** Target DNA or RNA fragment from a given sample library

NGS short reads - library preparation

Template DNA

Fragmentation

Adapters

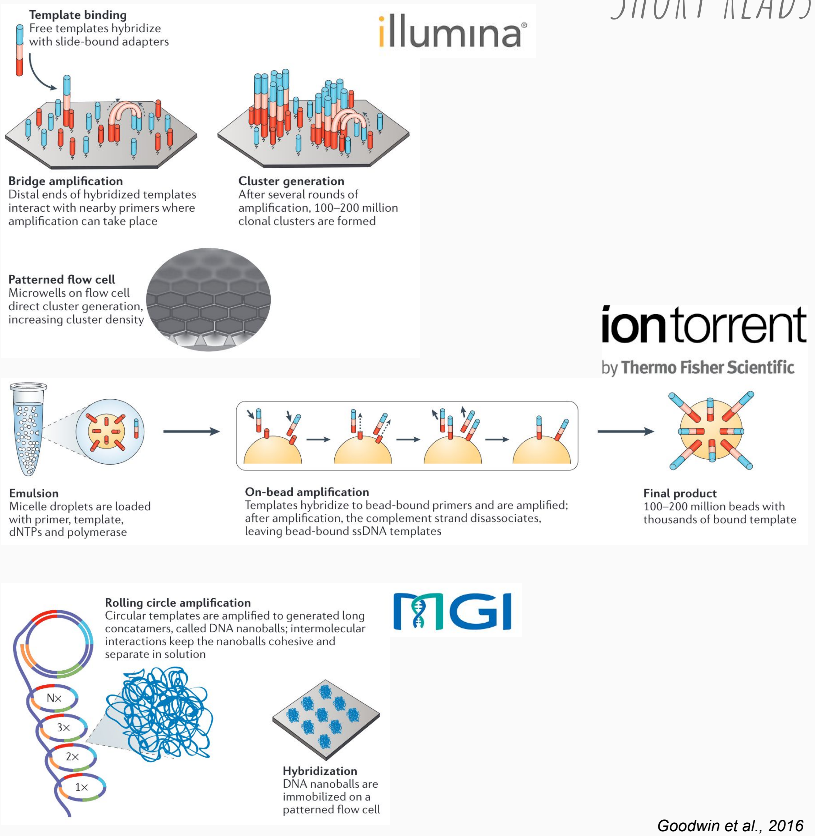
Clonal amplification

Sequencing

LONG READS

NO ACTION REQUIRED

SHORT READS



Template binding
Free templates hybridize with slide-bound adapters

Bridge amplification
Distal ends of hybridized templates interact with nearby primers where amplification can take place

Cluster generation
After several rounds of amplification, 100–200 million clonal clusters are formed

Patterned flow cell
Microwells on flow cell direct cluster generation, increasing cluster density

Emulsion
Micelle droplets are loaded with primer, template, dNTPs and polymerase

On-bead amplification
Templates hybridize to bead-bound primers and are amplified; after amplification, the complement strand disassociates, leaving bead-bound ssDNA templates

Final product
100–200 million beads with thousands of bound template

Rolling circle amplification
Circular templates are amplified to generated long concatamers, called DNA nanoballs; intermolecular interactions keep the nanoballs cohesive and separate in solution

Hybridization
DNA nanoballs are immobilized on a patterned flow cell

Goodwin et al., 2016

NGS short reads - library preparation

Template DNA

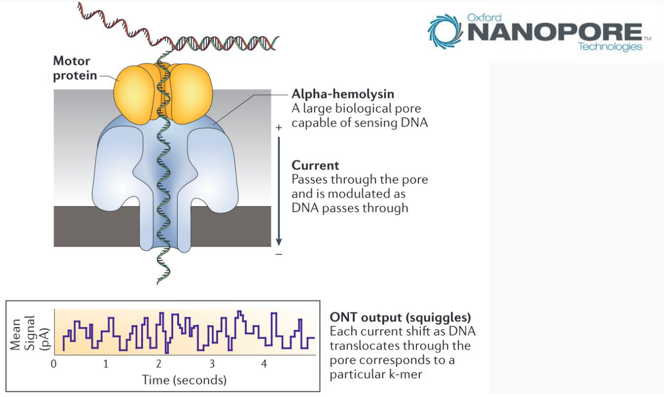
Fragmentation

Adapters

Clonal amplification

Sequencing

Long reads

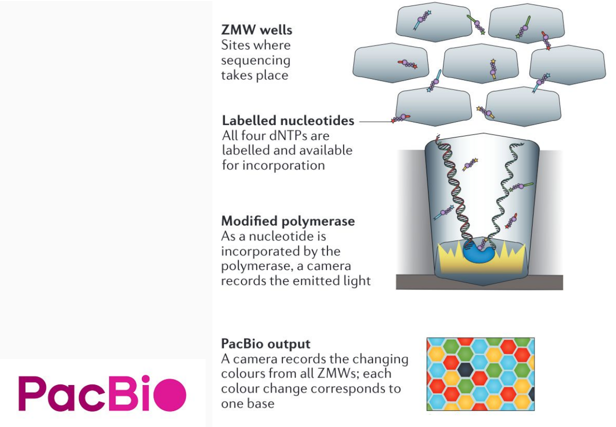


Motor protein

Alpha-hemolysin
A large biological pore capable of sensing DNA

Current
Passes through the pore and is modulated as DNA passes through

ONT output (squiggles)
Each current shift as DNA translocates through the pore corresponds to a particular k-mer



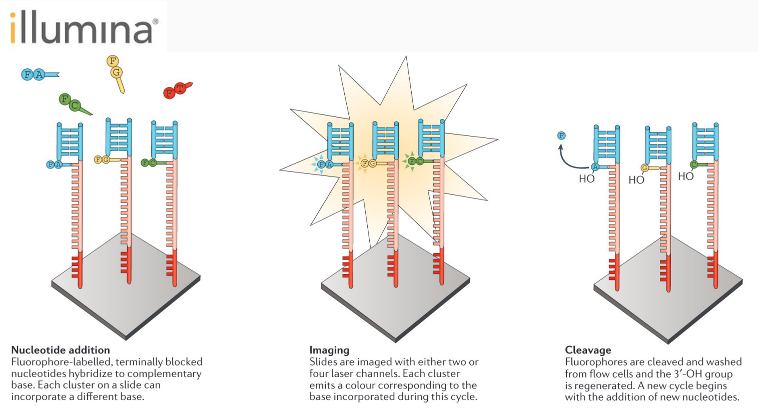
ZMW wells
Sites where sequencing takes place

Labelled nucleotides
All four dNTPs are labelled and available for incorporation

Modified polymerase
As a nucleotide is incorporated by the polymerase, a camera records the emitted light

PacBio output
A camera records the changing colours from all ZMWs; each colour change corresponds to one base

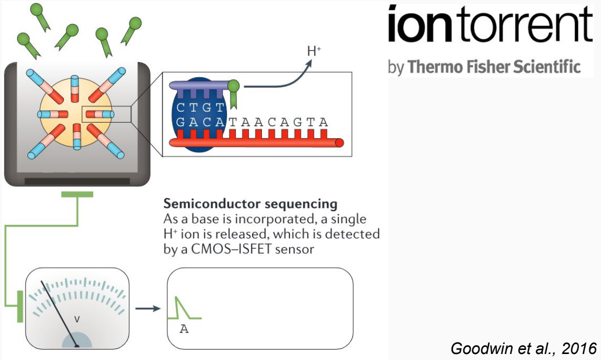
Short reads



Nucleotide addition
Fluorophore-labelled, terminally blocked nucleotides hybridize to complementary base. Each cluster on a slide can incorporate a different base.

Imaging
Slides are imaged with either two or four laser channels. Each cluster emits a colour corresponding to the base incorporated during this cycle.

Cleavage
Fluorophores are cleaved and washed from flow cells and the 3′-OH group is regenerated. A new cycle begins with the addition of new nucleotides.



Semiconductor sequencing
As a base is incorporated, a single H^+ ion is released, which is detected by a CMOS–ISFET sensor

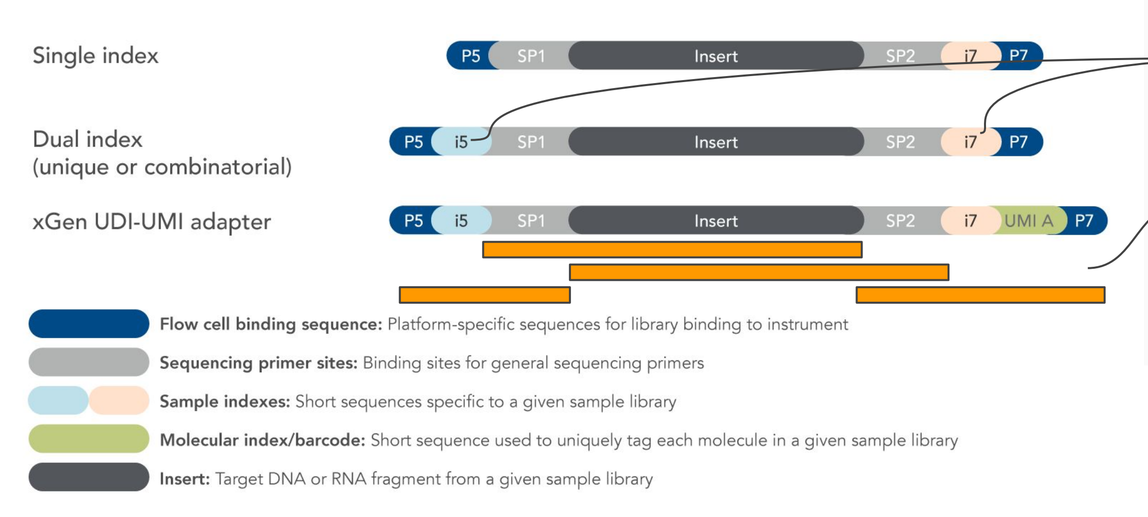
Goodwin et al., 2016

NGS reads structure

Understand the structure of sequencing libraries and the resulting data

- adapters
- read configuration
- UMIs

NGS data - understand library structure



Flow cell binding sequence: Platform-specific sequences for library binding to instrument

Sequencing primer sites: Binding sites for general sequencing primers

Sample indexes: Short sequences specific to a given sample library

Molecular index/barcode: Short sequence used to uniquely tag each molecule in a given sample library

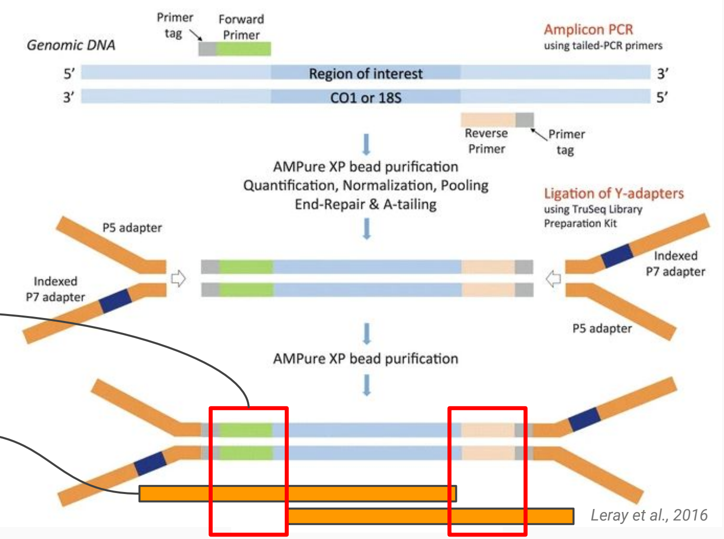
Insert: Target DNA or RNA fragment from a given sample library

Library barcode(s) for multiplexing

Example of sequencing reads

Standard library preparation

Amplicon targeted library

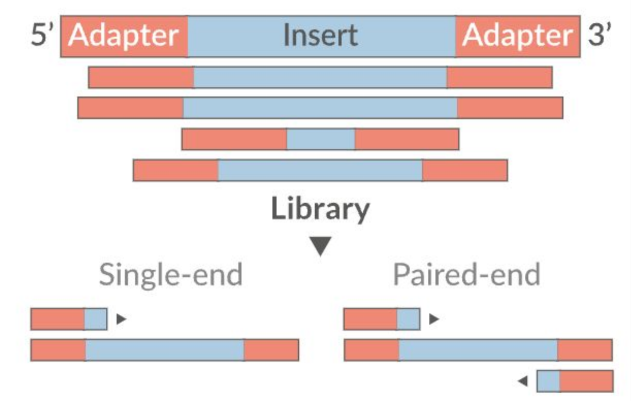


Leray et al., 2016

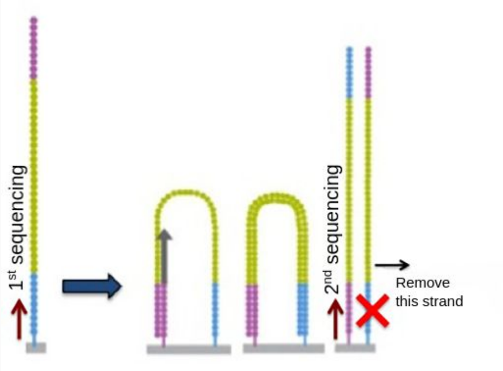
Read contains the PCR primer

Example of sequencing reads

NGS data - single end and paired-end short reads

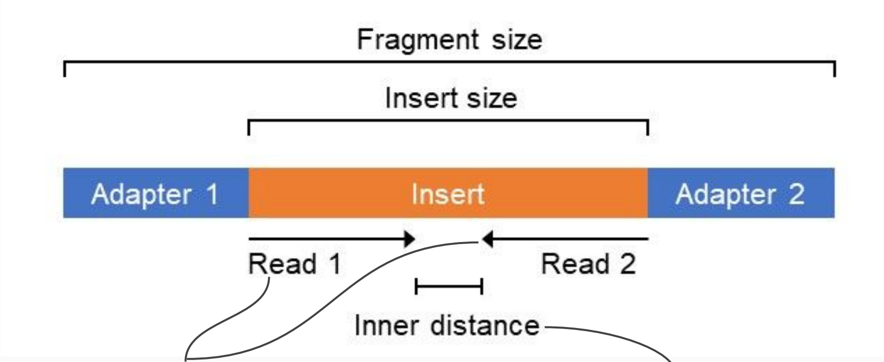


Illumina paired-end sequencing



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

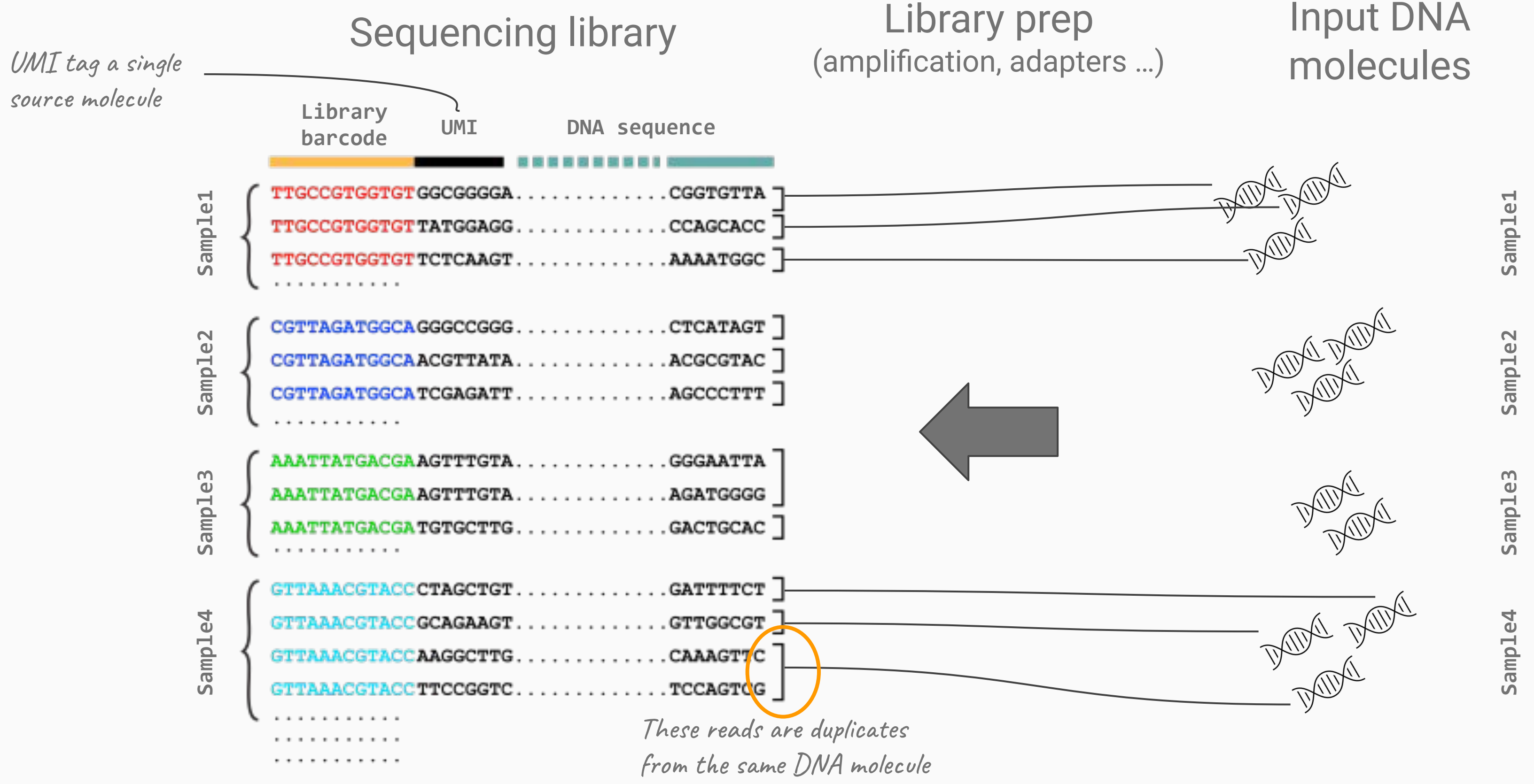
In **paired-end protocol**, it is sequenced twice starting from opposing ends



R1 and R2 will have known reciprocal position and orientation

This distance is known and can be leveraged in variant calling

NGS data - understand UMIs



Sequencing library

Library prep
(amplification, adapters ...)

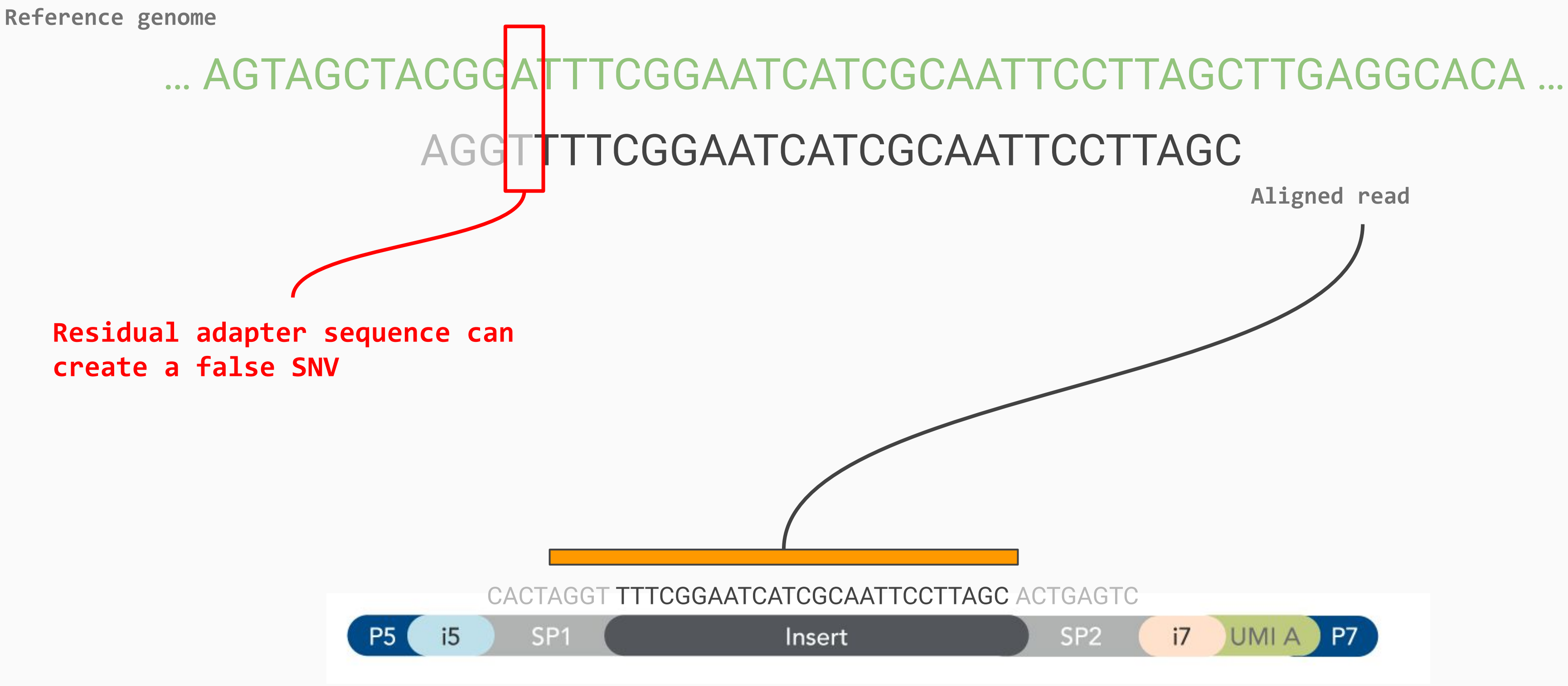
Input DNA molecules

UMI tag a single source molecule

```
TTGCCGTGGTGTGGCGGGGA.............CGGTGTTA
TTGCCGTGGTGTTATGGAGG.............CCAGCACC
TTGCCGTGGTGTTCTCAAGT.............AAAATGGC
...........

CGTTAGATGGCAGGGCCGGG.............CTCATAGT
CGTTAGATGGCAACGTTATA.............ACGCGTAC
CGTTAGATGGCATCGAGATT.............AGCCCTTT
...........

AAATTATGACGAAGTTTGTA.............GGGAATTA
AAATTATGACGAAGTTTGTA.............AGATGGGG
AAATTATGACGATGTGCTTG.............GACTGCAC
...........

GTTAAACGTACCCTAGCTGT.............GATTTTCT
GTTAAACGTACCGCAGAAGT.............GTTGGCGT
GTTAAACGTACCAAGGCTTG.............CAAAGTTC
GTTAAACGTACCTTCCGGTC.............TCCAGTCG
...........
...........
...........
```

These reads are duplicates from the same DNA molecule

NGS data - library structure impacts downstream analysis



Reference genome

… AGTAGCTACGGATTTCGGAATCATCGCAATTCCTTAGCTTGAGGCACA …

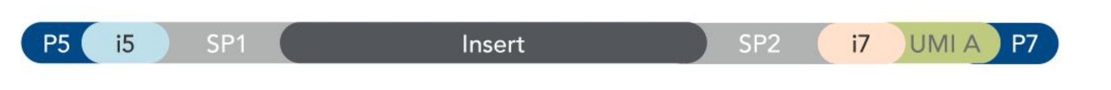
AGGTTTTCGGAATCATCGCAATTCCTTAGC

Aligned read

Residual adapter sequence can create a false SNV

CACTAGGT TTTCGGAATCATCGCAATTCCTTAGC ACTGAGTC

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.

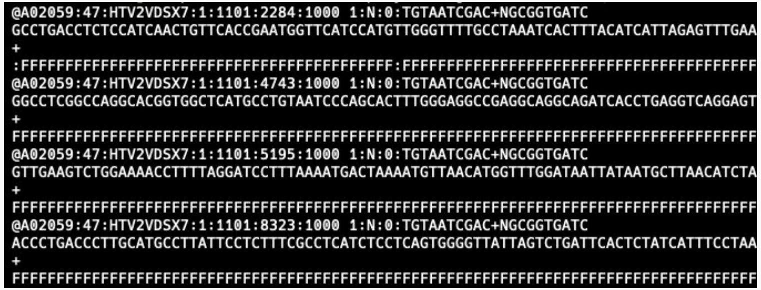
Data formats

Main file formats

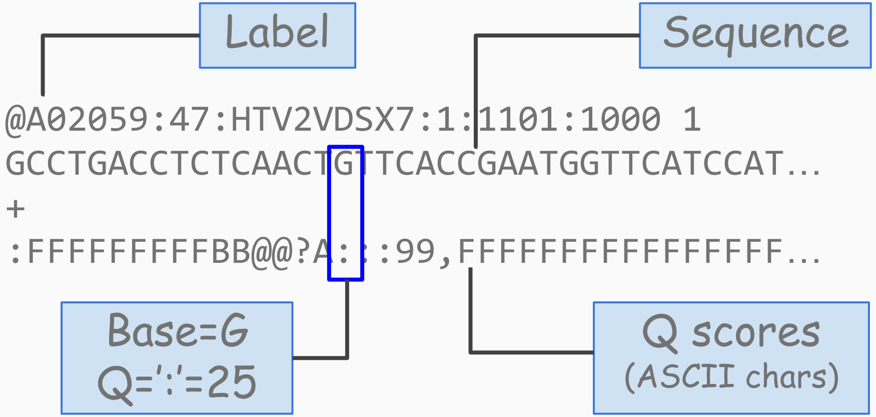
- FASTQ: sequences

Sequences - FASTQ files

Sequence and base quality of each read



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



Label
Sequence
@A02059:47:HTV2VDSX7:1:1101:1000 1
GCCTGACCTCTCAACTGTTCACCGAATGGTTCATCCAT...
+
:FFFFFFFFFBB@@?A:::99,FFFFFFFFFFFFFFFF...
Base=G
Q=':'=25
Q scores
(ASCII chars)

- **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

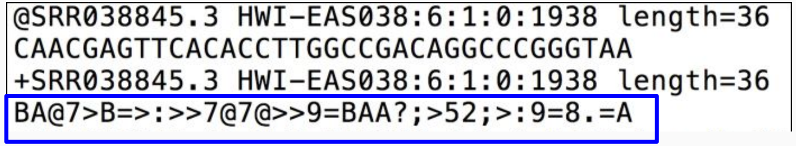
PHRED quality

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

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

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

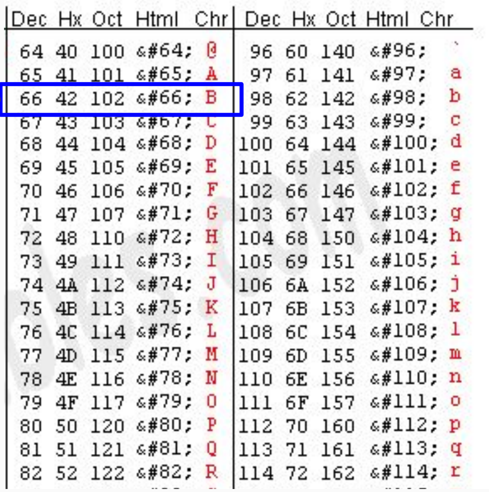


```
@SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
CAACGAGTTCACACCTTGGCCGACAGGCCCGGGTAA
+SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
BA@7>B=>:>>7@7@>>9=BAA?;>52;>:9=8.=A
```

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

Short-reads QC

Reads cleaning

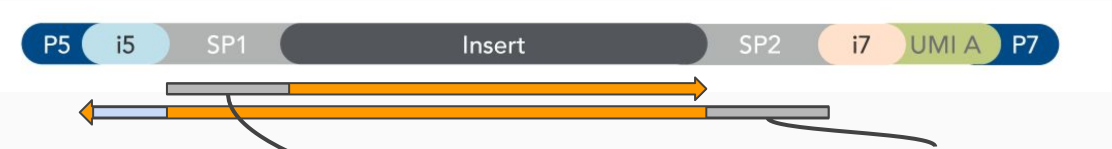
Remove unwanted sequences and poor quality bases

- adapter trimming
- quality trimming
- fixed length trimming

Useful tools

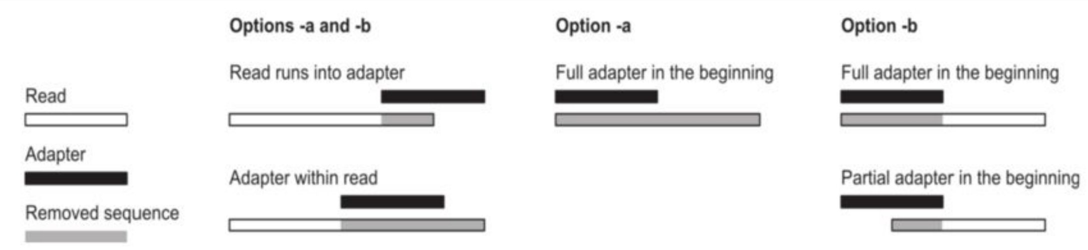
fastp, cutadapt

Adapter trimming - remove unwanted fixed sequences



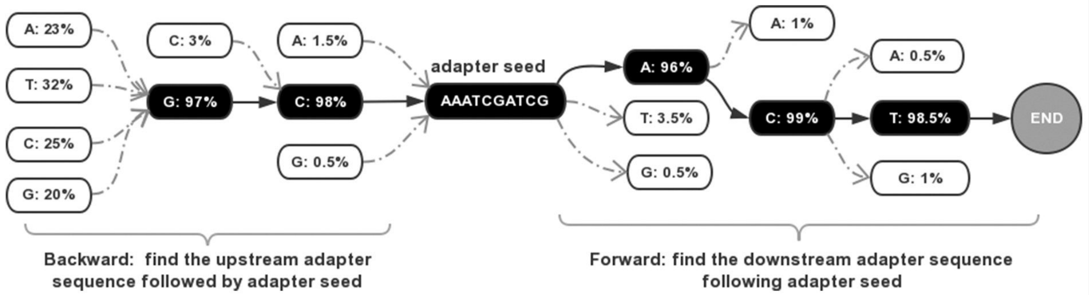
Raw reads may contain fixed sequences introduced during library preparation

Example from [cutadapt tool]()

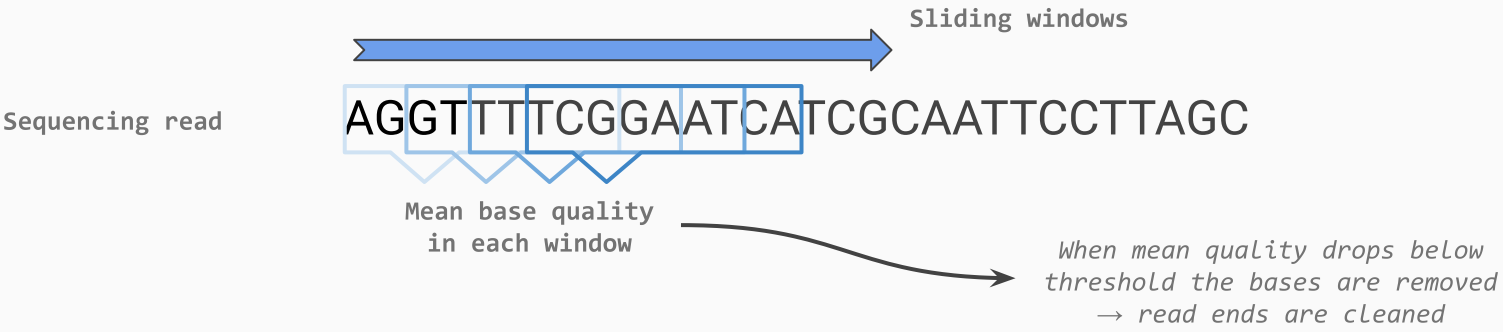


Trimming tools scan the read ends removing fixed sequences

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


A: 23%
T: 32%
C: 25%
G: 20%
C: 3%
G: 97%
A: 1.5%
C: 98%
G: 0.5%
adapter seed
AAATCGATCG
A: 96%
T: 3.5%
G: 0.5%
A: 1%
C: 99%
A: 0.5%
T: 98.5%
G: 1%
END
Backward: find the upstream adapter sequence followed by adapter seed
Forward: find the downstream adapter sequence following adapter seed

Quality trimming - remove low quality bases from read ends



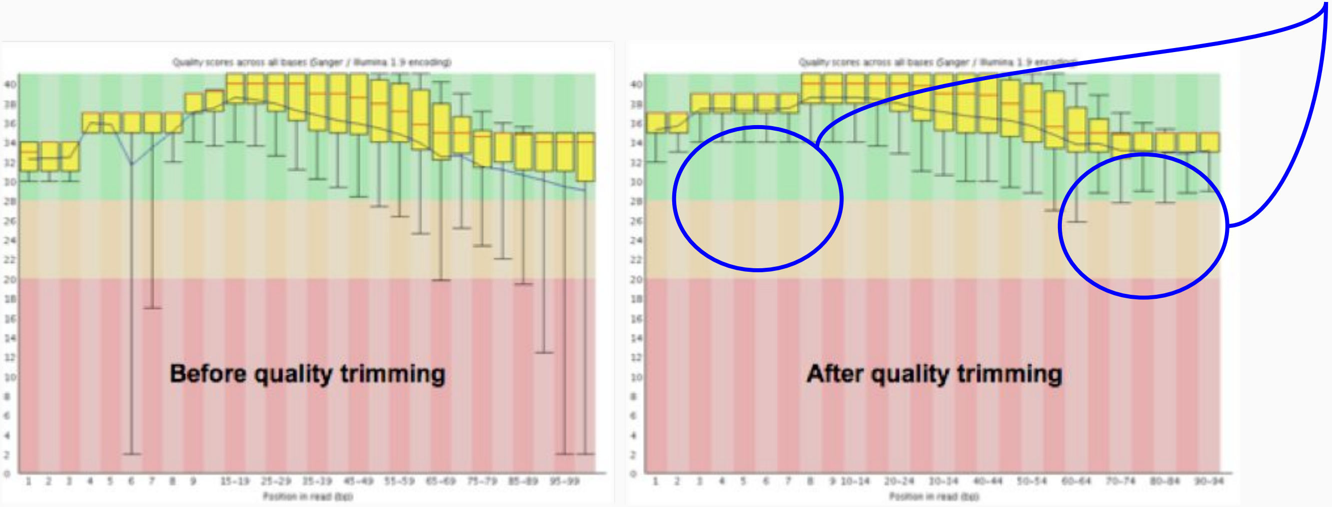
Sliding windows

Sequencing read

AGGTTTTCGGAATCATCGCAATTCCTTAGC

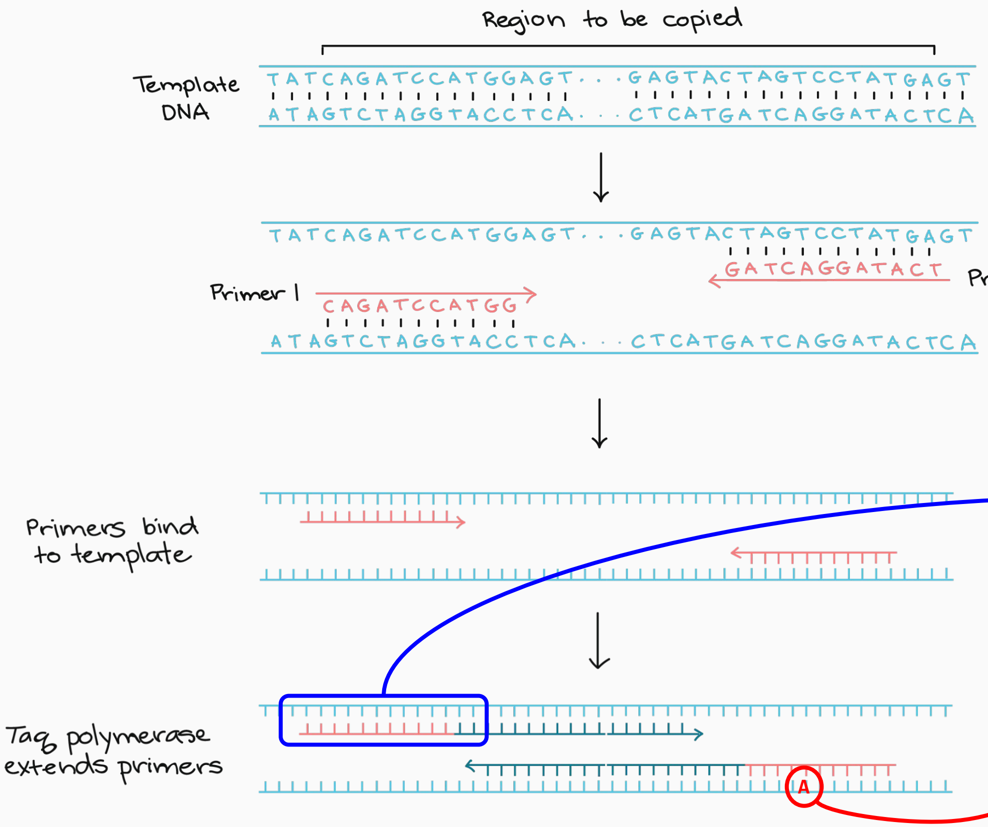
Mean base quality in each window

When mean quality drops below threshold the bases are removed → read ends are cleaned



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

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



Region to be copied
Template DNA
TATCAGATCCATGGAGT...GAGTACTAGTCCTATGAGT
ATAGTCTAGGTACCTCA...CTCATGATCAGGATACTCA
TATCAGATCCATGGAGT...GAGTACTAGTCCTATGAGT
GATCAGGATACT
Primer 2
Primer 1
CAGATCCATGG
ATAGTCTAGGTACCTCA...CTCATGATCAGGATACTCA
Primers bind to template
Taq polymerase extends primers
A

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

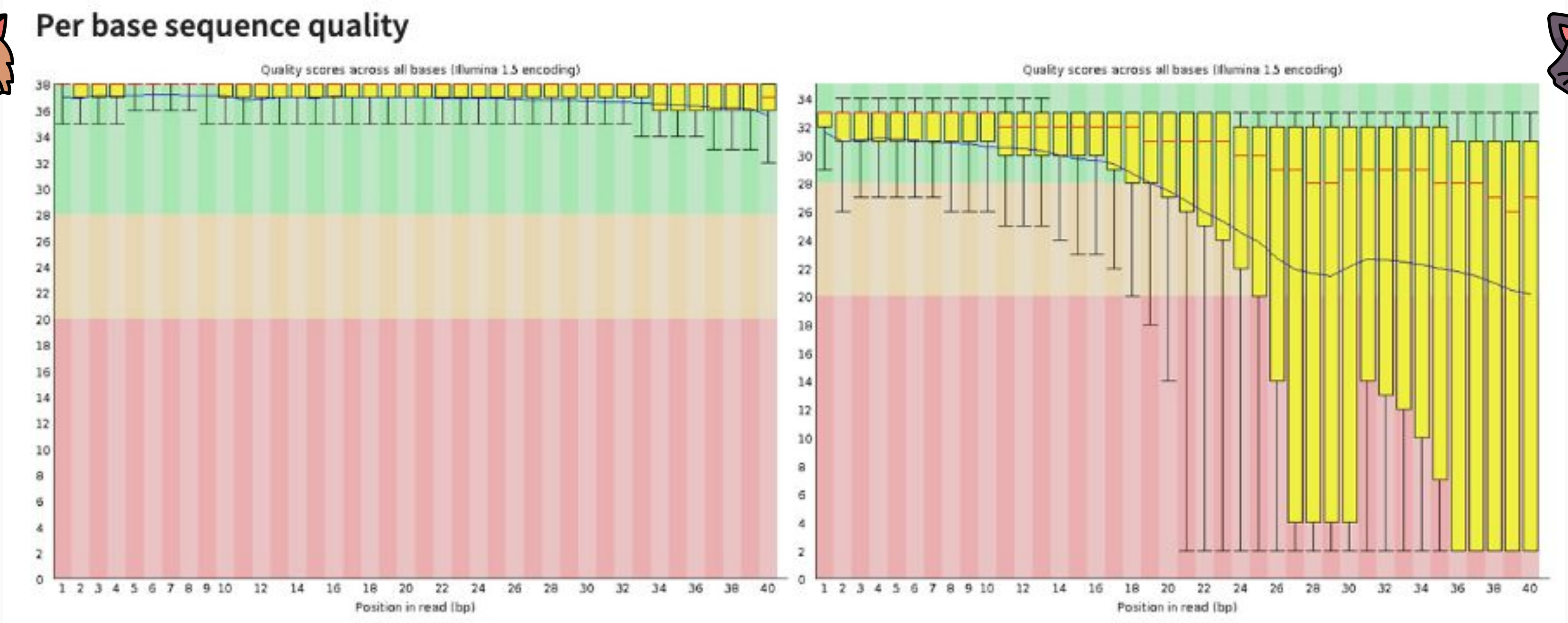
- raw reads length and quality
- GC content
- adapter content
- mapping statistics

Useful tools

fastp, fastQC, samtools

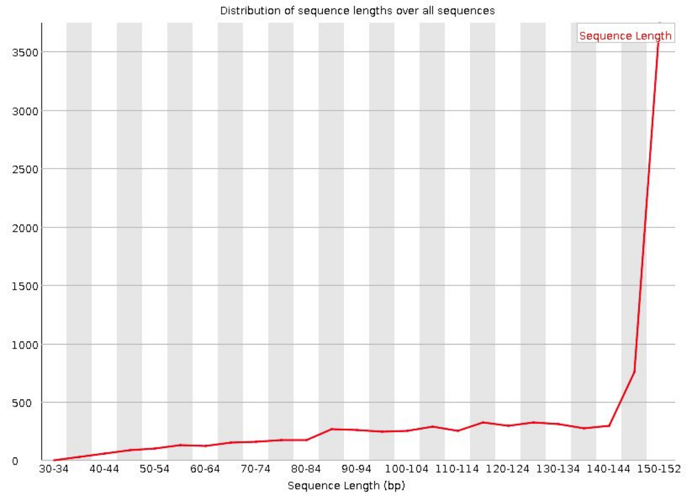
Reads QC - Per base sequence quality across the reads

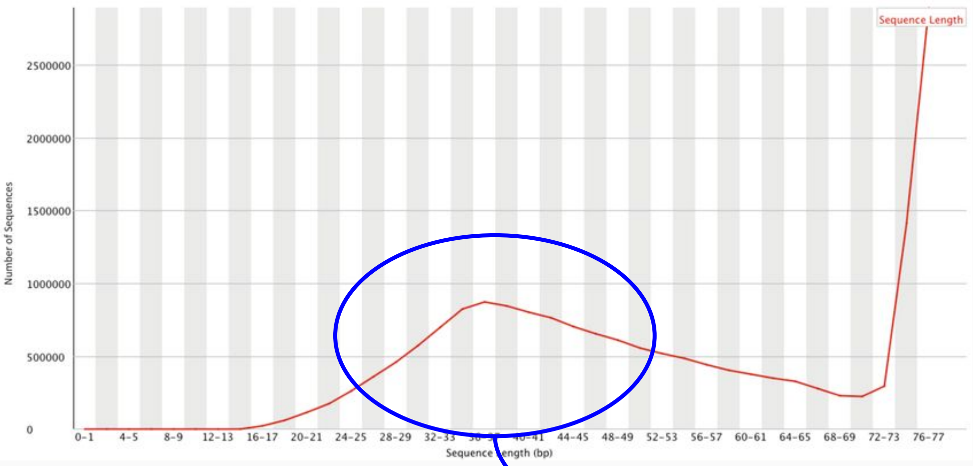
- Ideally, base quality should be >= 30 across all the read
- A little decrease in quality is expected toward the end of the read
- Base quality is considered during variant calling and can affect variant caller performances


Per base sequence quality

Read QC - sequence length distribution

- Most read should have the expected read length
- A small tail on the left is acceptable, especially after trimming

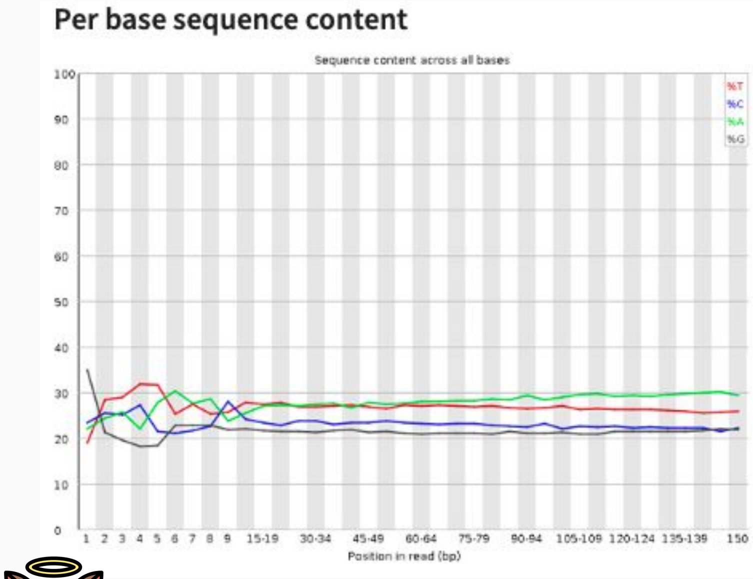




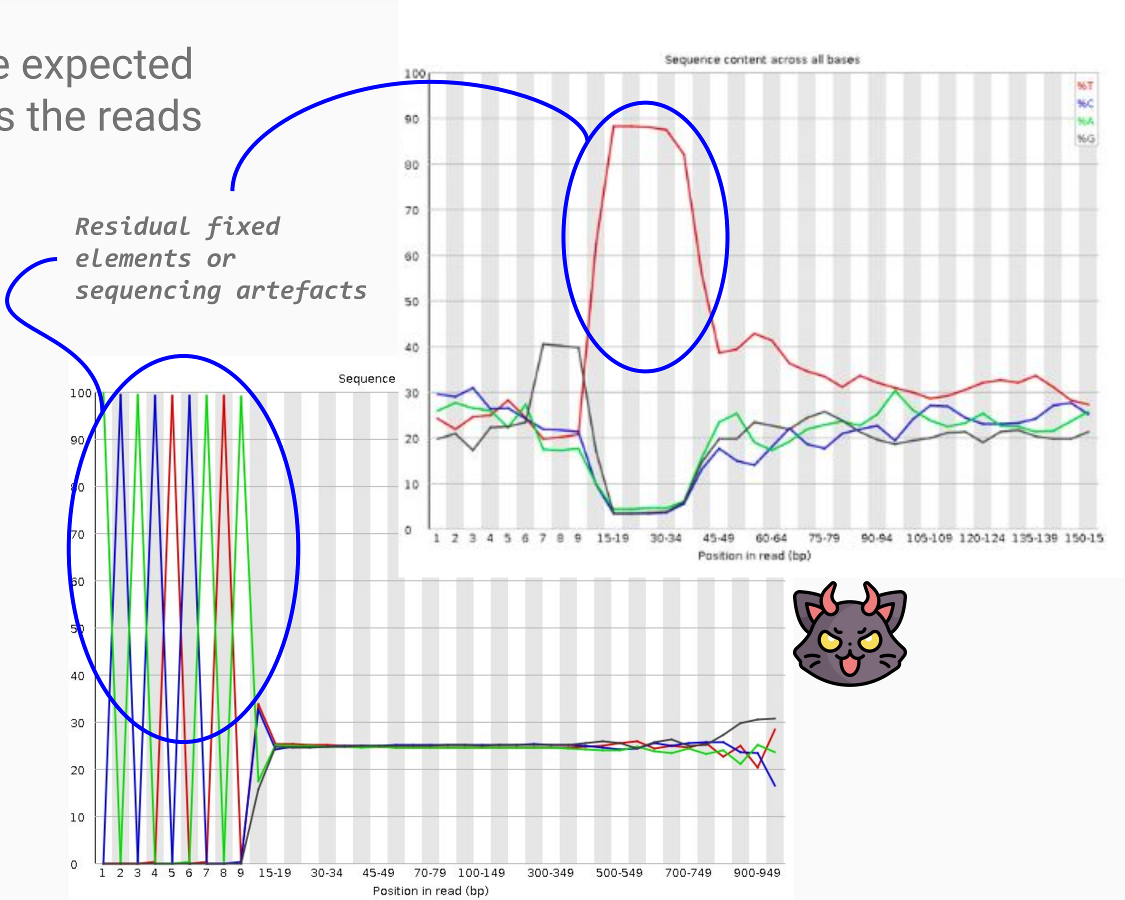
Sequencing problems, low quality reads that got trimmed, adapter dimers

Reads QC - Per base sequence content across the reads

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



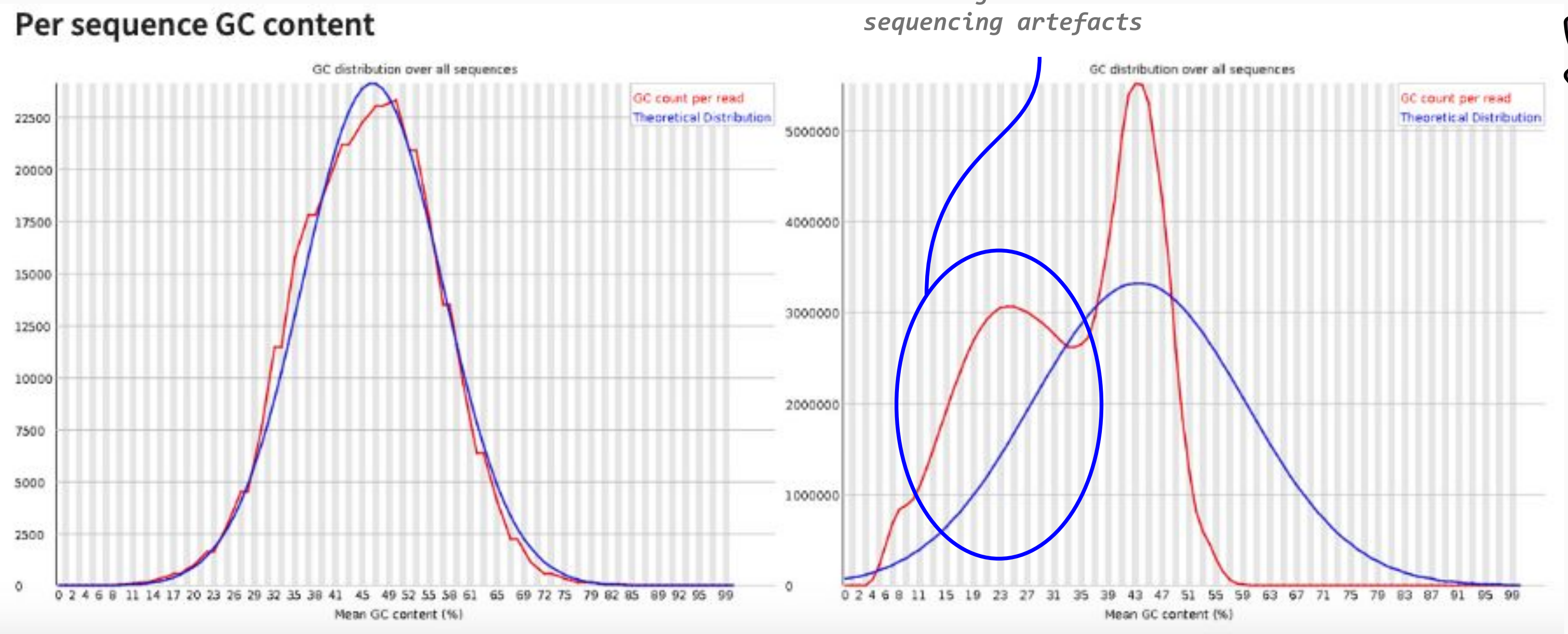
Per base sequence content



Residual fixed elements or sequencing artefacts

Reads QC - reads GC content

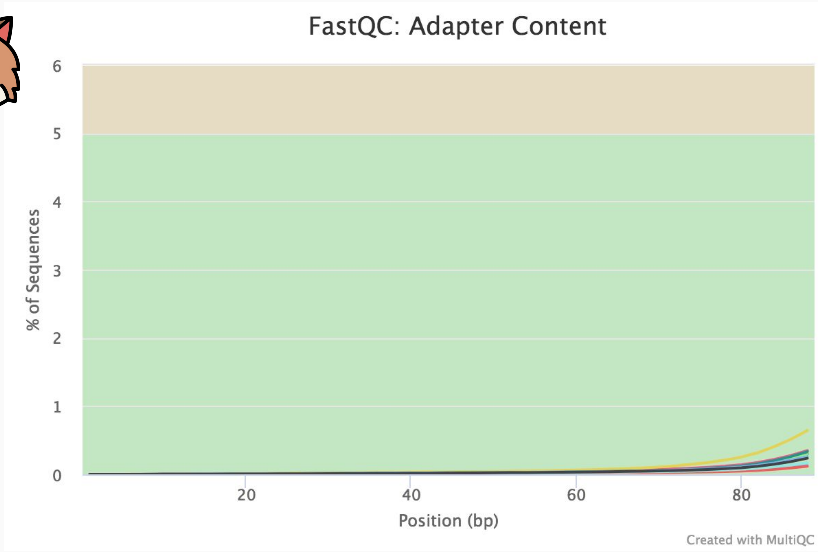
- Distribution should follow the expected for the sequenced organism
- Peak around 45% for human genome samples

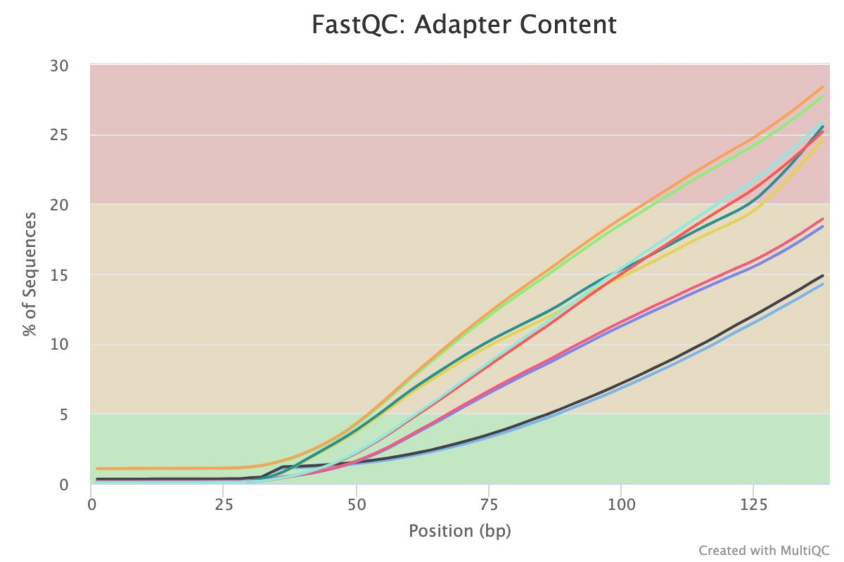


Contamination from DNA of other organisms or sequencing artefacts

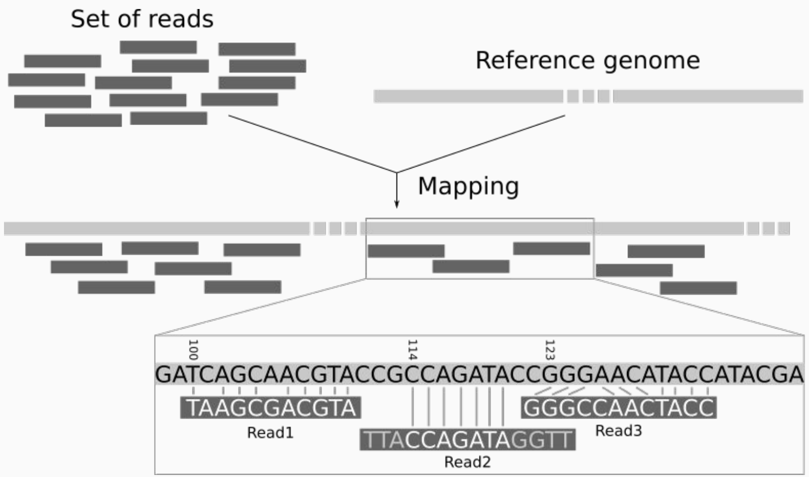
Read QC - residual adapter content

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





Aligned reads QC - mapping statistics



All libraries

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

Paired-end libraries

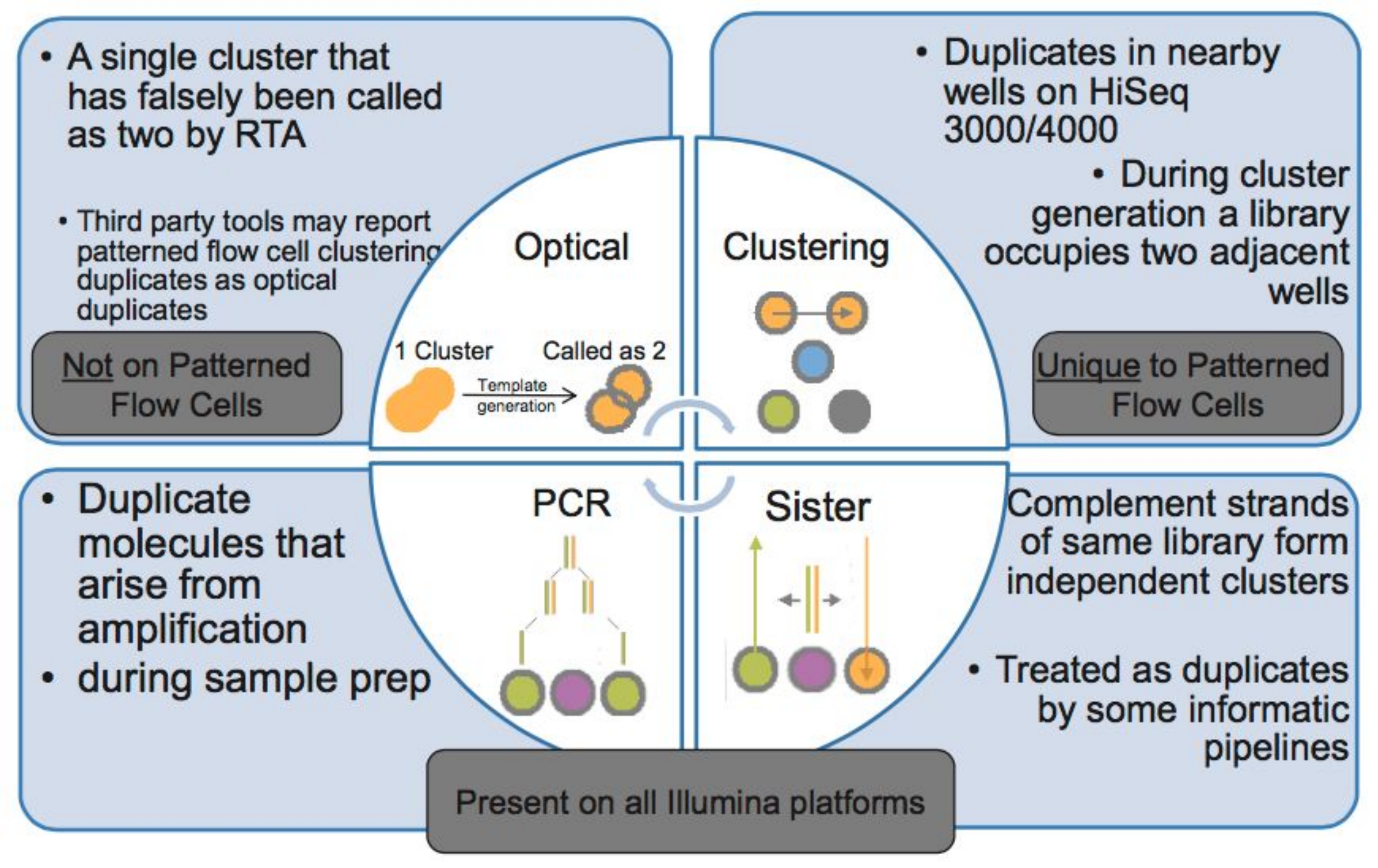
- Insert size distribution
  (distance between F/R read)
- Fraction of reads with a proper pair
  (>= 90% usually)

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



- A single cluster that has falsely been called as two by RTA
  - Third party tools may report patterned flow cell clustering duplicates as optical duplicates

Not on Patterned Flow Cells

- Duplicates in nearby wells on HiSeq 3000/4000
  - During cluster generation a library occupies two adjacent wells

Unique to Patterned Flow Cells

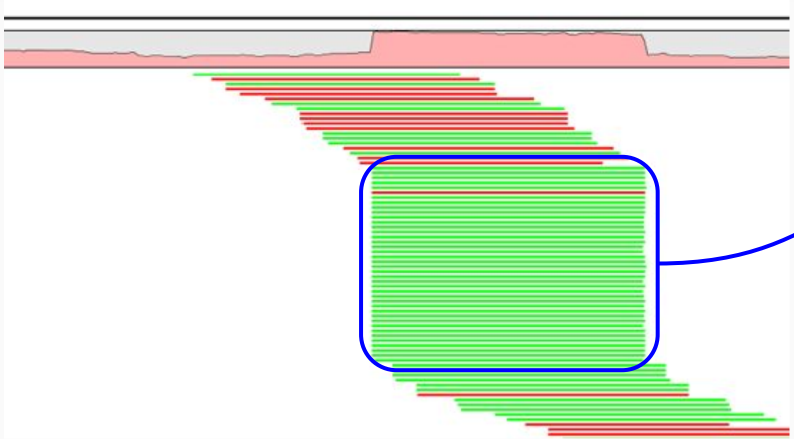
- Duplicate molecules that arise from amplification
- during sample prep

Complement strands of same library form independent clusters

- Treated as duplicates by some informatic pipelines

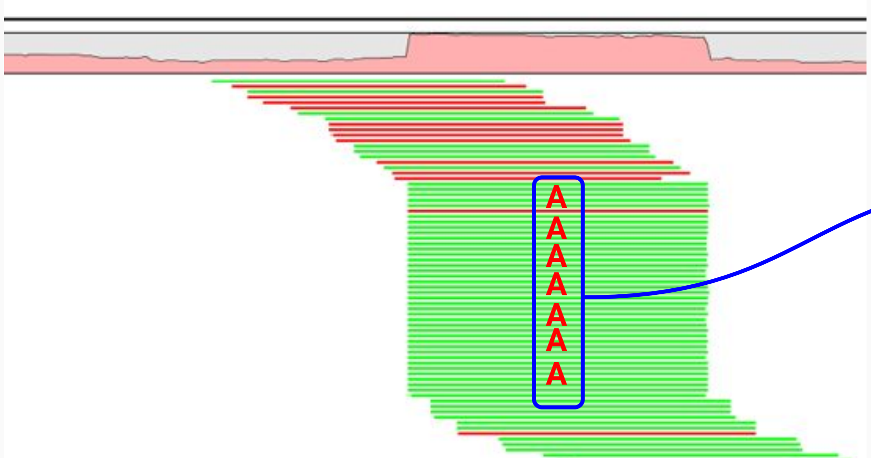
Present on all Illumina platforms

PCR or optical duplicates



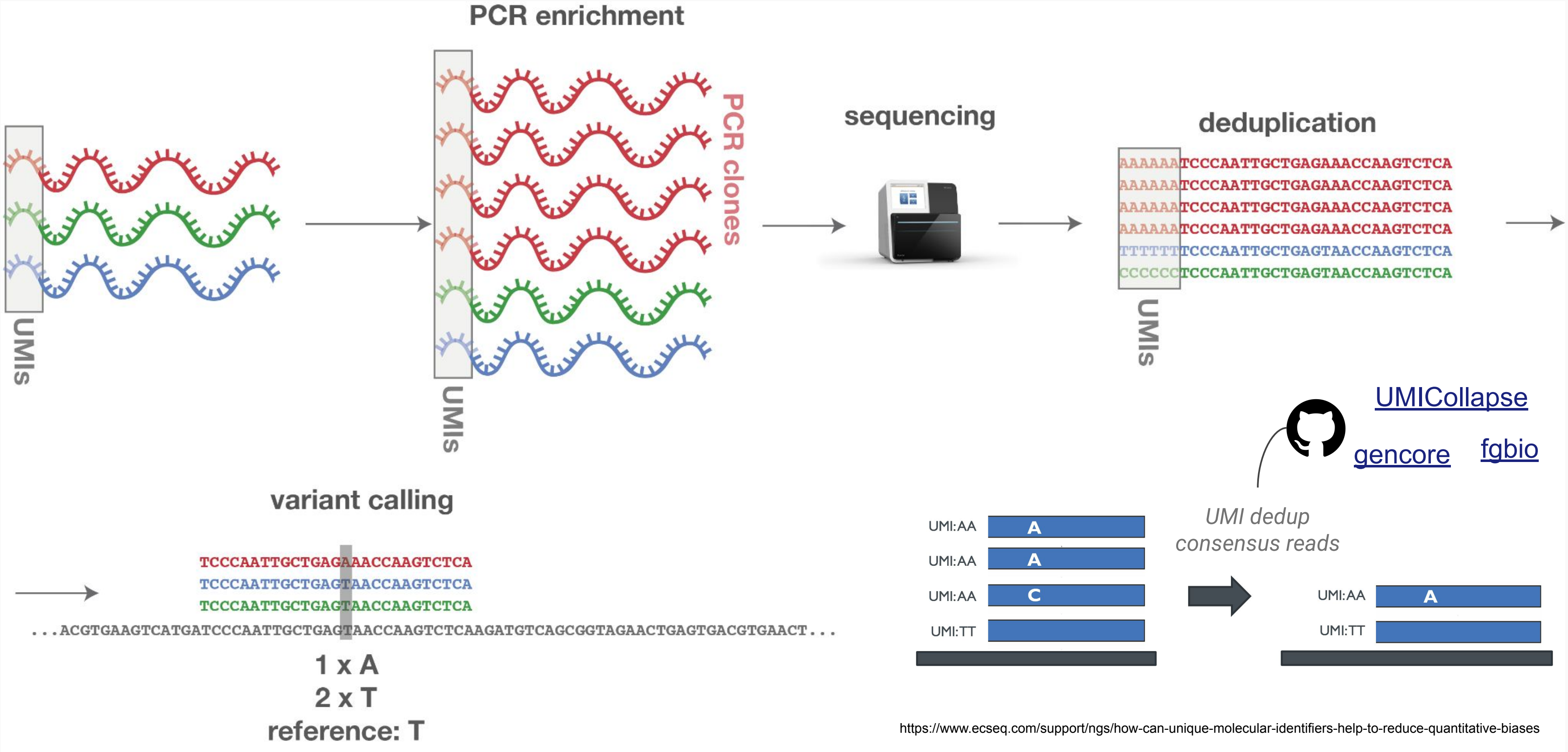
Duplicated reads
A pile or reads that actually represent multiple identical copies originated from the same DNA molecule!

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



PCR enrichment

sequencing

deduplication

variant calling

UMIs

PCR clones

AAAAAATCCCAATTGCTGAGAAACCAAGTCTCA
AAAAAATCCCAATTGCTGAGAAACCAAGTCTCA
AAAAAATCCCAATTGCTGAGAAACCAAGTCTCA
AAAAAATCCCAATTGCTGAGAAACCAAGTCTCA
TTTTTTTCCCAATTGCTGAGTAACCAAGTCTCA
CCCCCCTCCCAATTGCTGAGTAACCAAGTCTCA

TCCCAATTGCTGAGAAACCAAGTCTCA
TCCCAATTGCTGAGTAACCAAGTCTCA
TCCCAATTGCTGAGTAACCAAGTCTCA
...ACGTGAAGTCATGATCCCAATTGCTGAGTAACCAAGTCTCAAGATGTCAGCGGTAGAACTGAGTGACGTGAACT...

1 x A
2 x T
reference: T

UMICollapse

gencore

fgbio

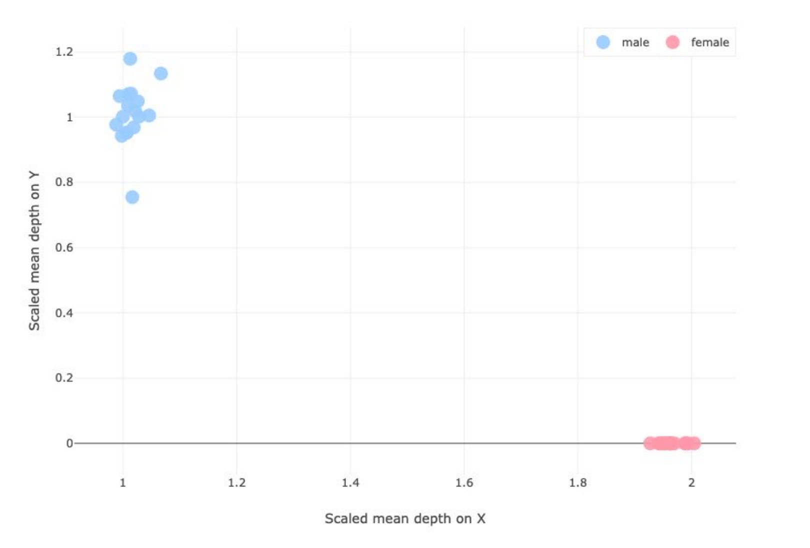
UMI dedup consensus reads

UMI:AA	A
UMI:AA	A
UMI:AA	C
UMI:TT	

UMI:AA	A
UMI:TT	

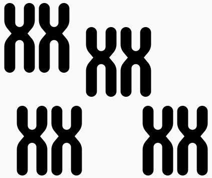
https://www.ecseq.com/support/ngs/how-can-unique-molecular-identifiers-help-to-reduce-quantitative-biases

Additional sample-level QC - sex inference



Scaled mean depth

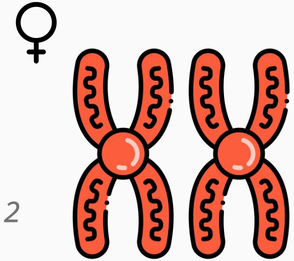
$$\frac{\textit{mean depth chrX}}{\textit{mean depth autosomes}} x\ 2$$



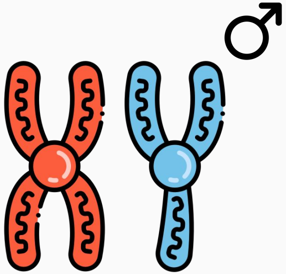
autosomes

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

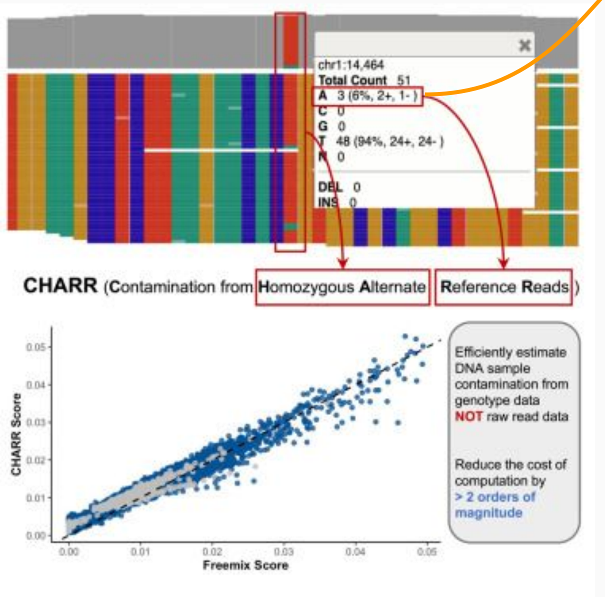


Additional sample-level QC - detect contamination

Target sample

Sequenced material

Contaminating sample



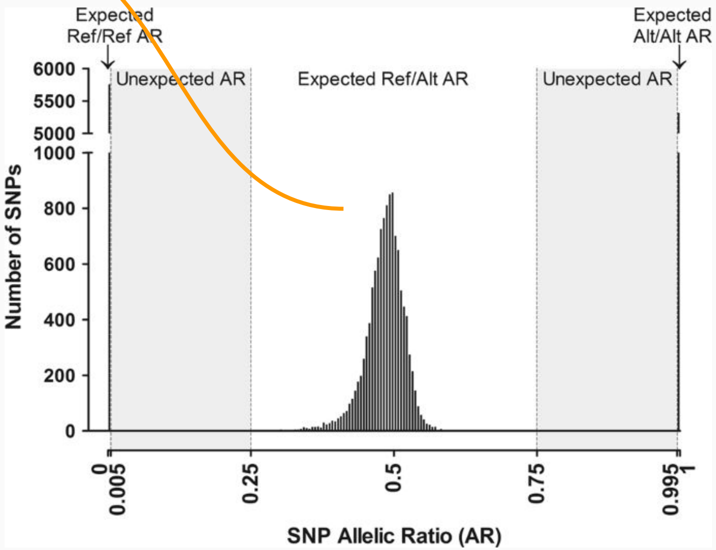
chr1:14,464
Total Count 51
A 3 (6%, 2+, 1-)
C 0
G 0
T 48 (94%, 24+, 24-)
N 0
DEL 0
INS 0

CHARR (Contamination from Homozygous Alternate Reference Reads)

Efficiently estimate DNA sample contamination from genotype data **NOT** raw read data

Reduce the cost of computation by > 2 orders of magnitude

Increased fraction of “unexpected” alleles in homozygous genotypes



Increased N of heterozygous genotypes
Increased het GT with outlier AR