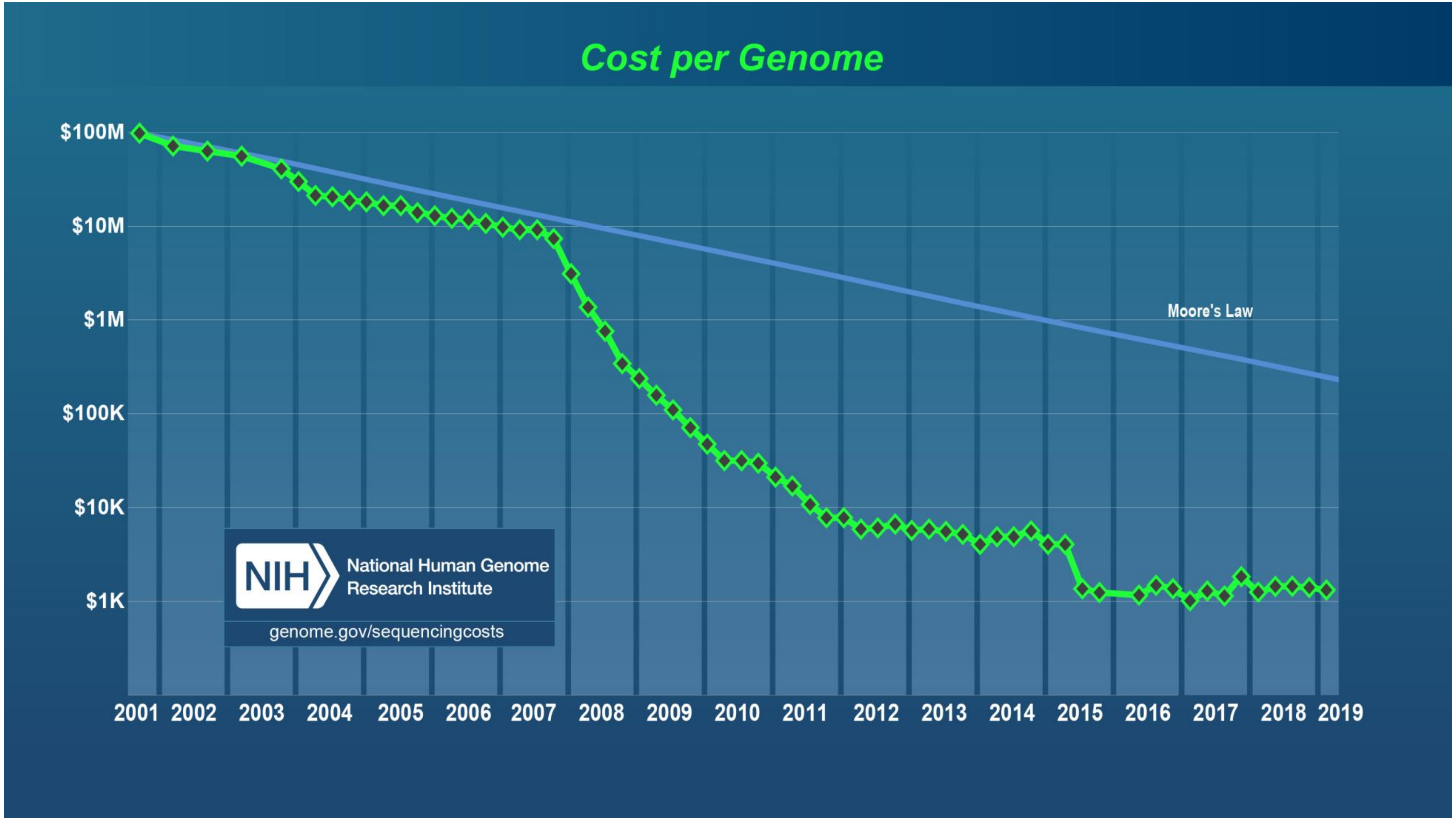
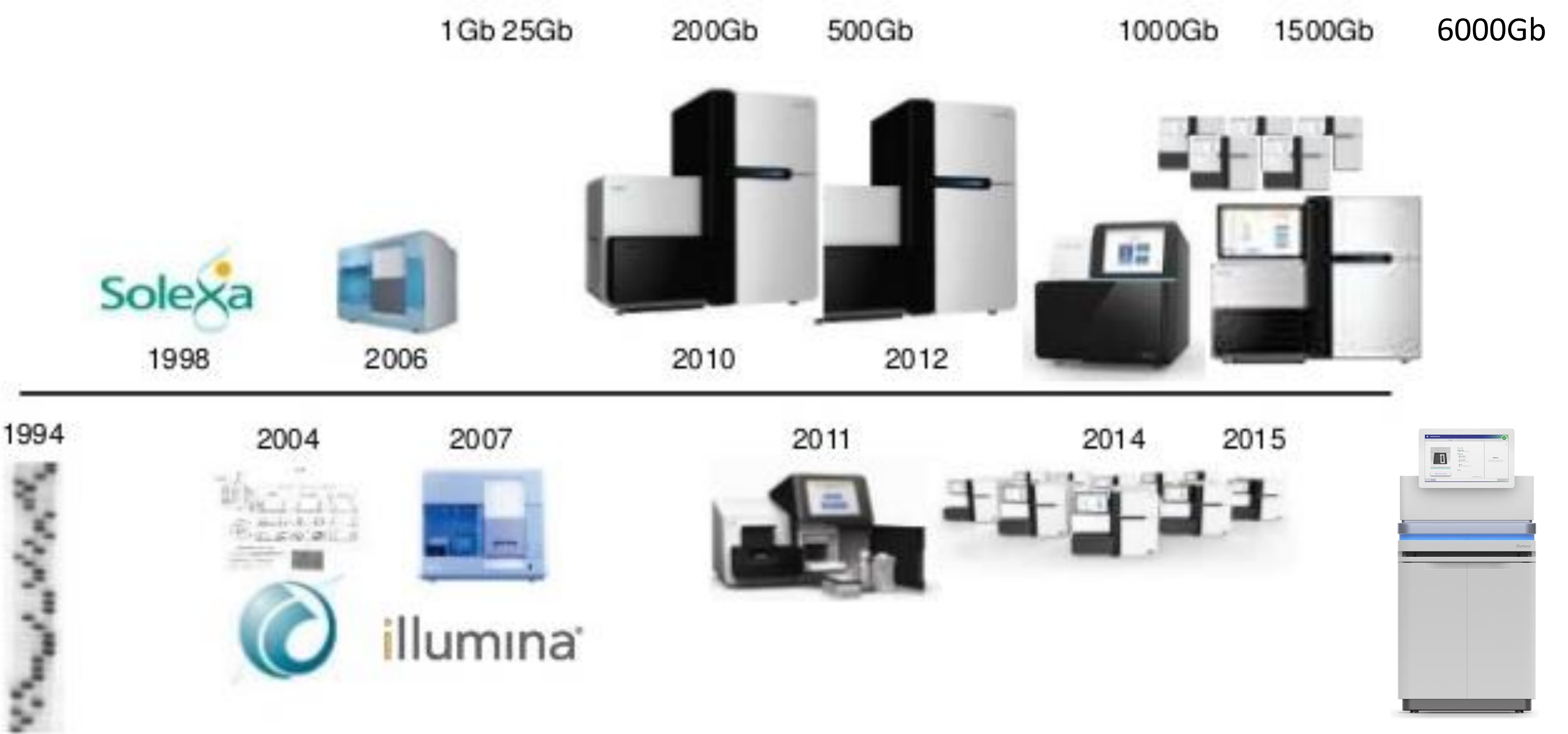
NGS introduction

Joana Meier

Sequencing costs have decreased massively



History of Illumina sequencing



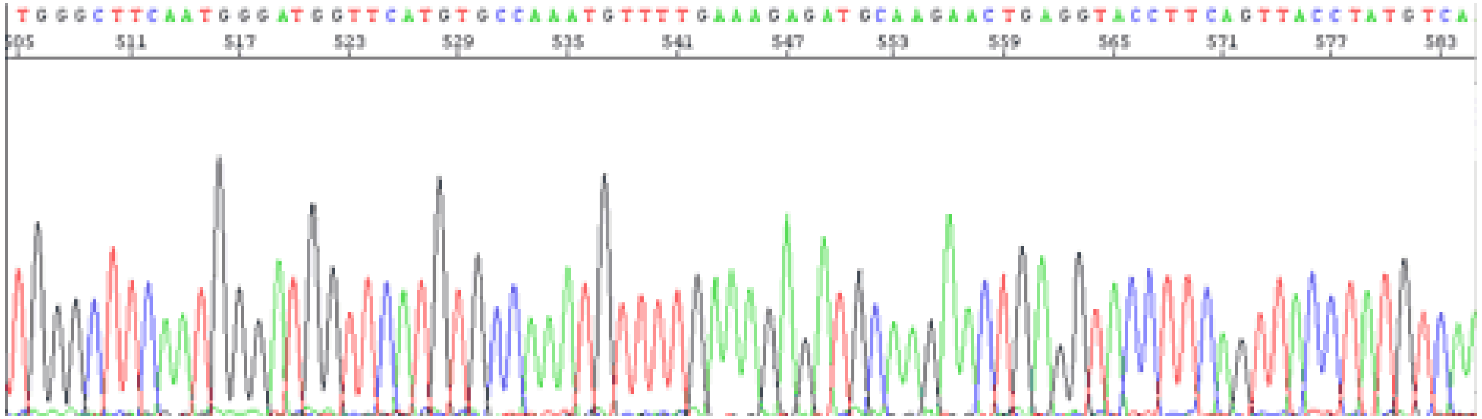
NovaSeq 6000

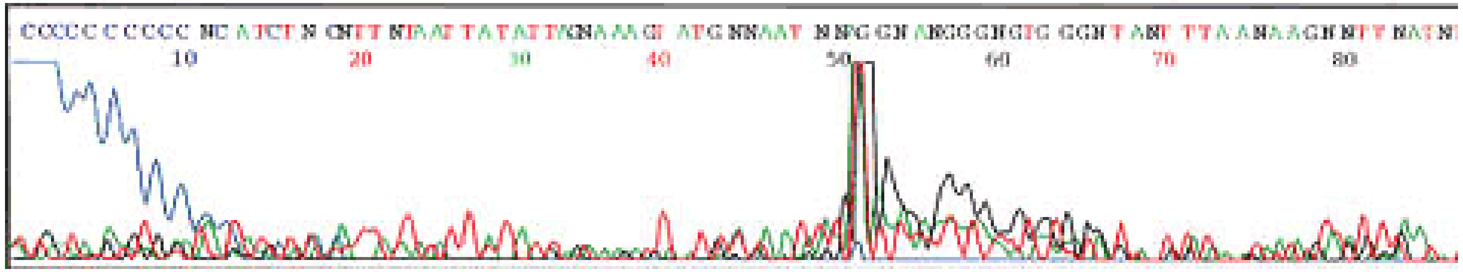
High Throughput Sequencing (=Next Generation Sequencing)

- **Short-read sequencing technologies (2nd generation):**
  - Sequence millions of clonally amplified molecules
  - E.g. Illumina, Ion Torrent, SOLiD

- **Long-read technologies (3rd generation):**
  - Single molecules are sequenced in real-time, fast but expensive and high error rates
  - E.g. PacBio (bought by Illumina): ~12kb reads, single molecules are read multiple times to reduce error rate
  - E.g. Oxford Nanopore: up to 900 kb reads, high sequencing error rate (5-15%) and non-random errors, each DNA fragment can only be read 2x

Sanger Sequencing

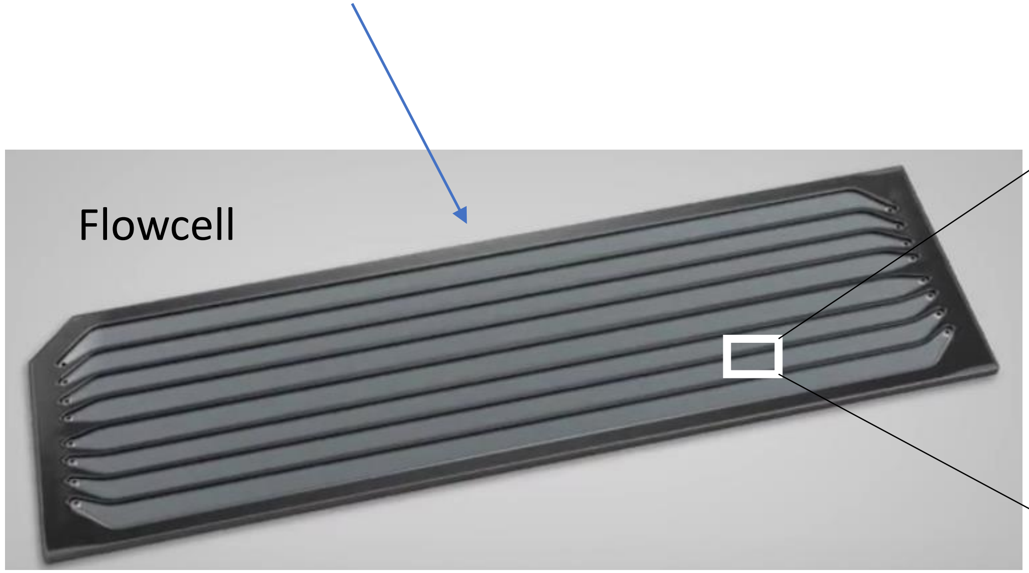




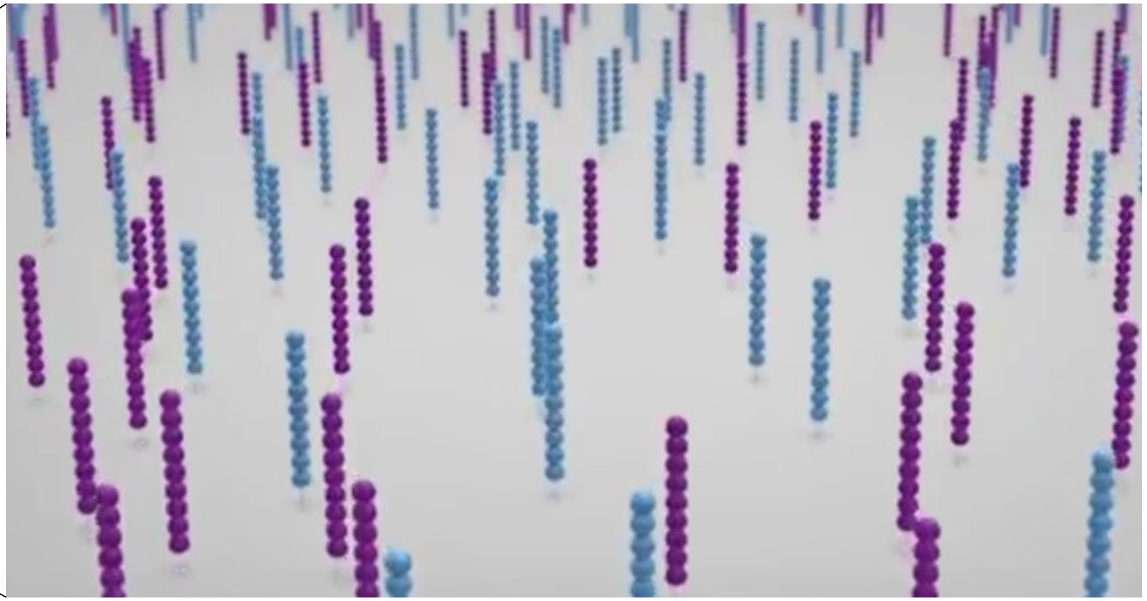
- Manually check each sequence
- Resequence failed sequences

Illumina flowcell: millions of DNA sequences

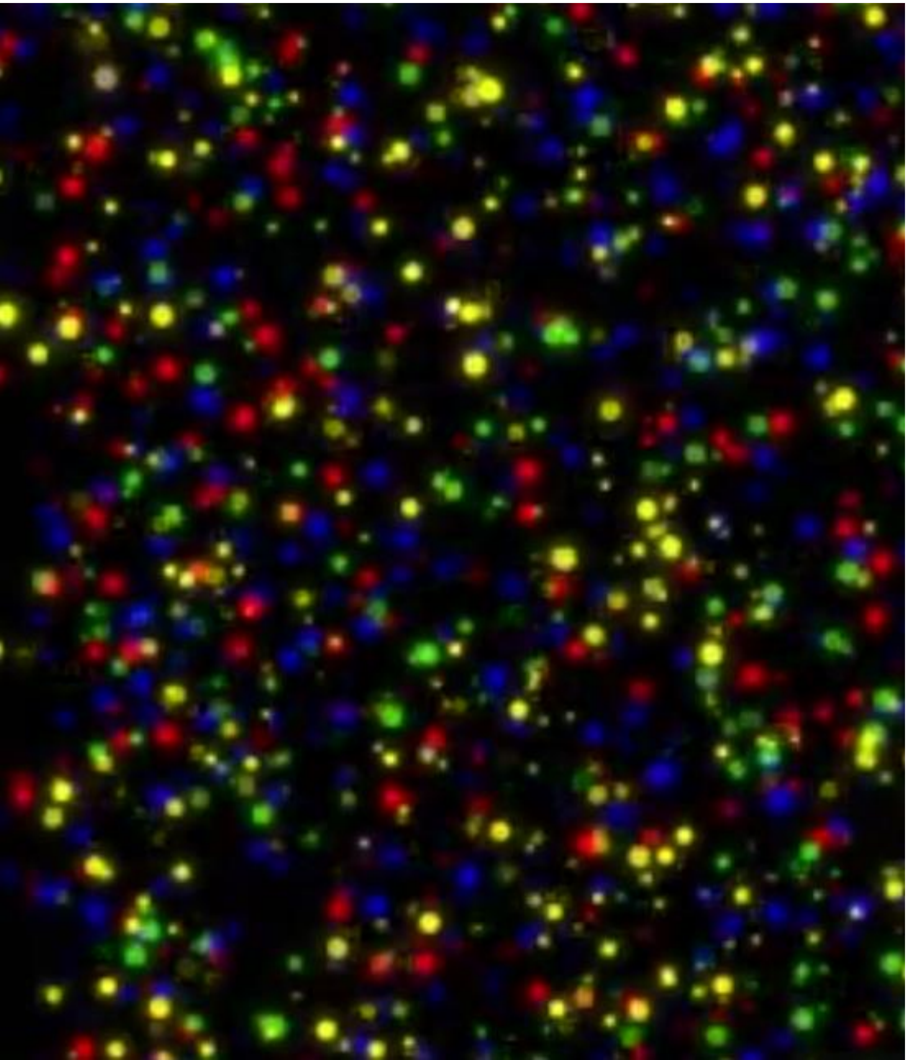
DNA fragments with Illumina adapters

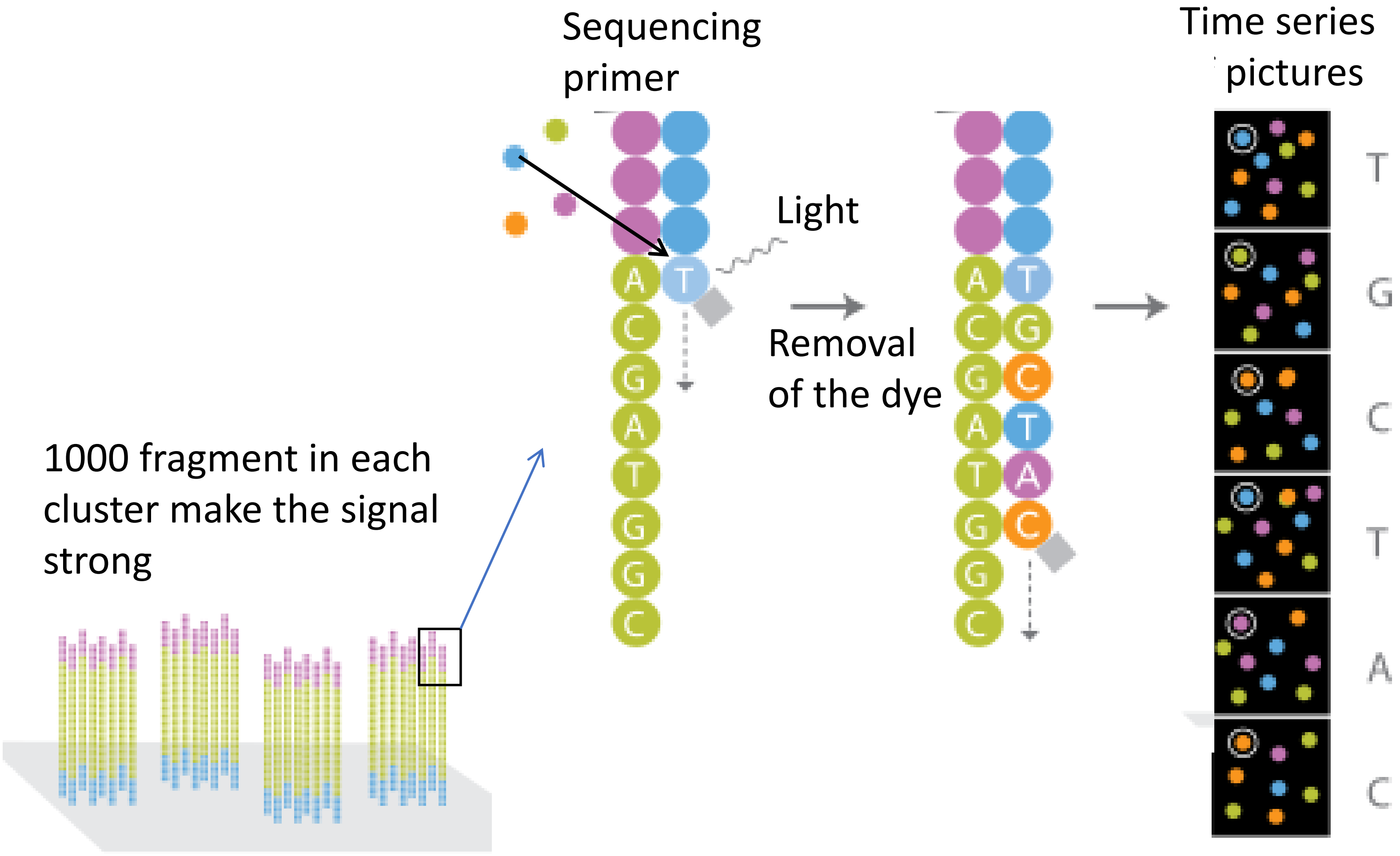


Each lane contains a dense lawn of Illumina primers



Sequencing by synthesis by Illumina





Sequencing primer

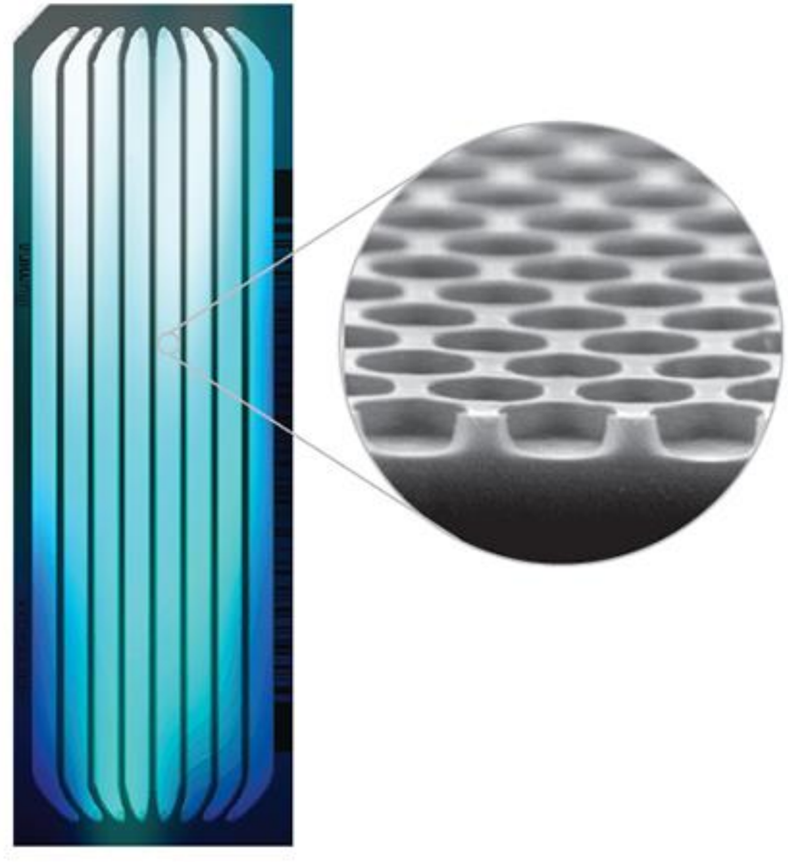
Light

Removal of the dye

Time series pictures

1000 fragment in each cluster make the signal strong

Illumina HiSeq4000 and X Ten

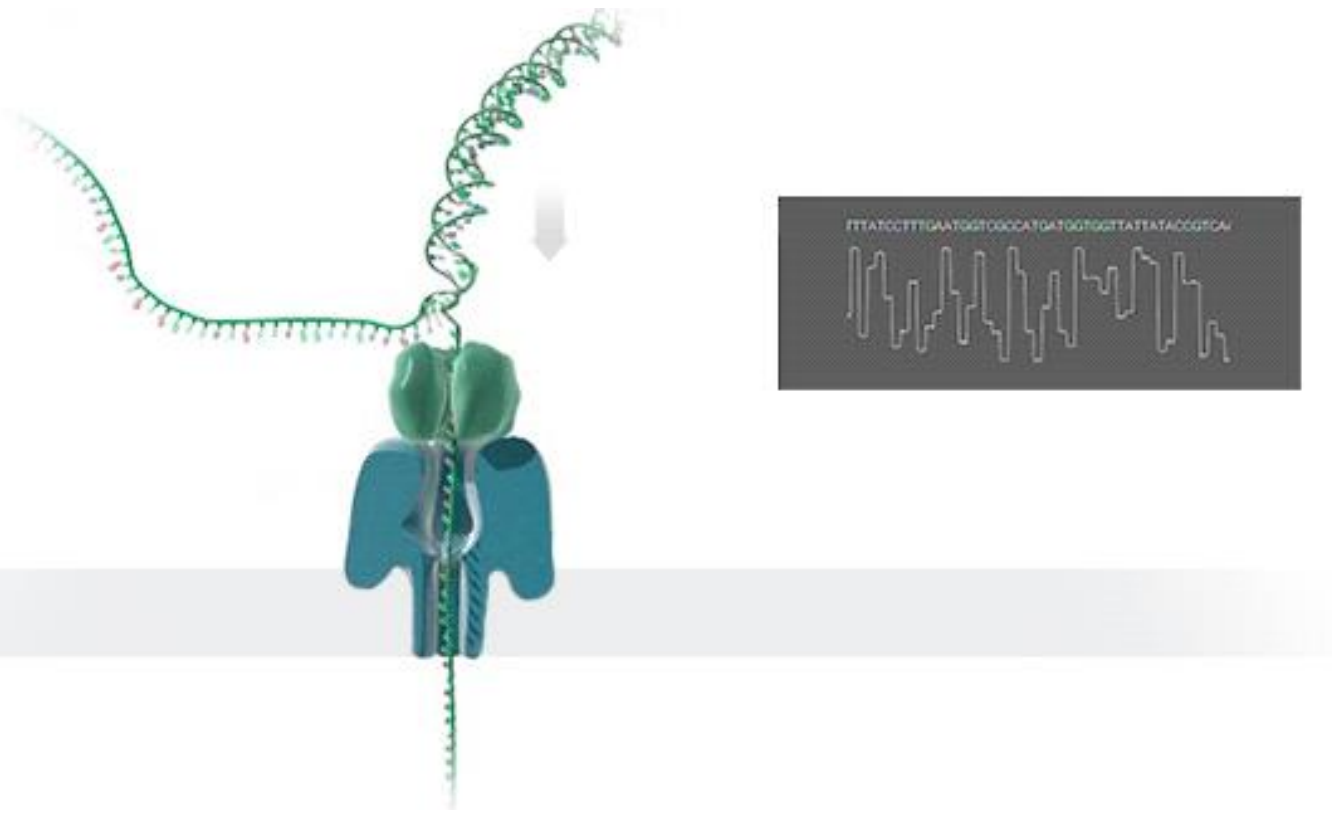


Problem:
Illumina barcode switching
(Index hopping)

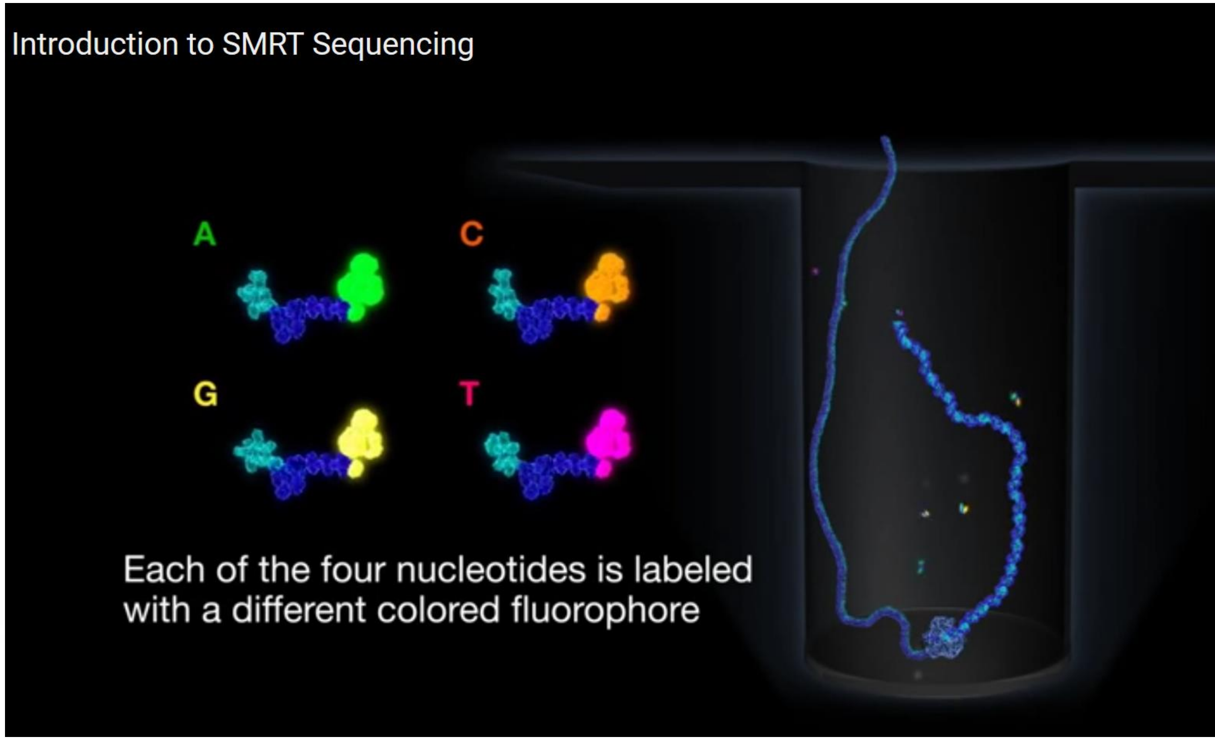
-> use double-indexing
Different P1 and P2 indices

Long read sequencing technologies

Nanopore

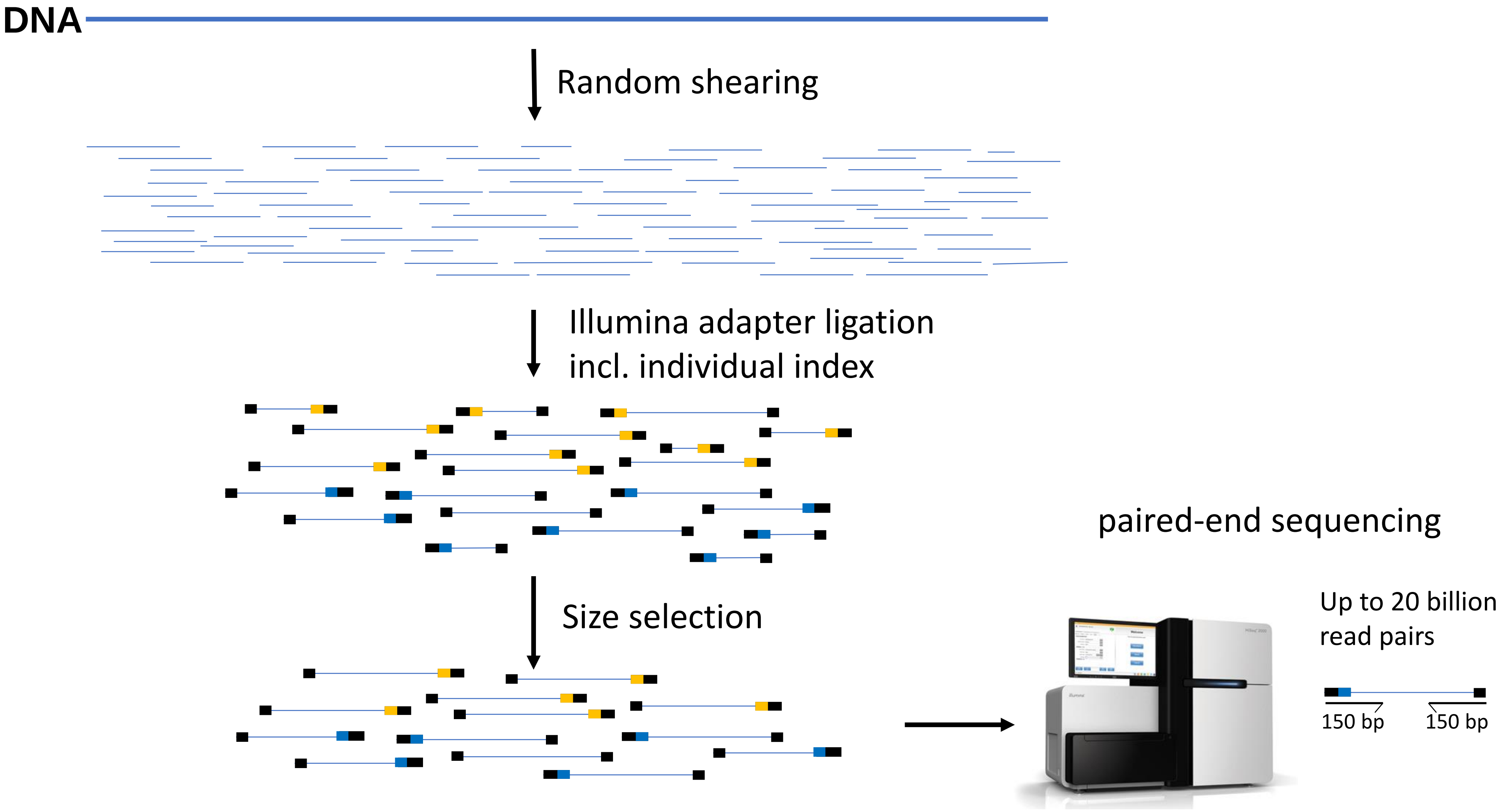


PacBio


Introduction to SMRT Sequencing

Each of the four nucleotides is labeled with a different colored fluorophore

Whole-genome sequencing (shotgun sequencing)



DNA

Random shearing

Illumina adapter ligation
incl. individual index

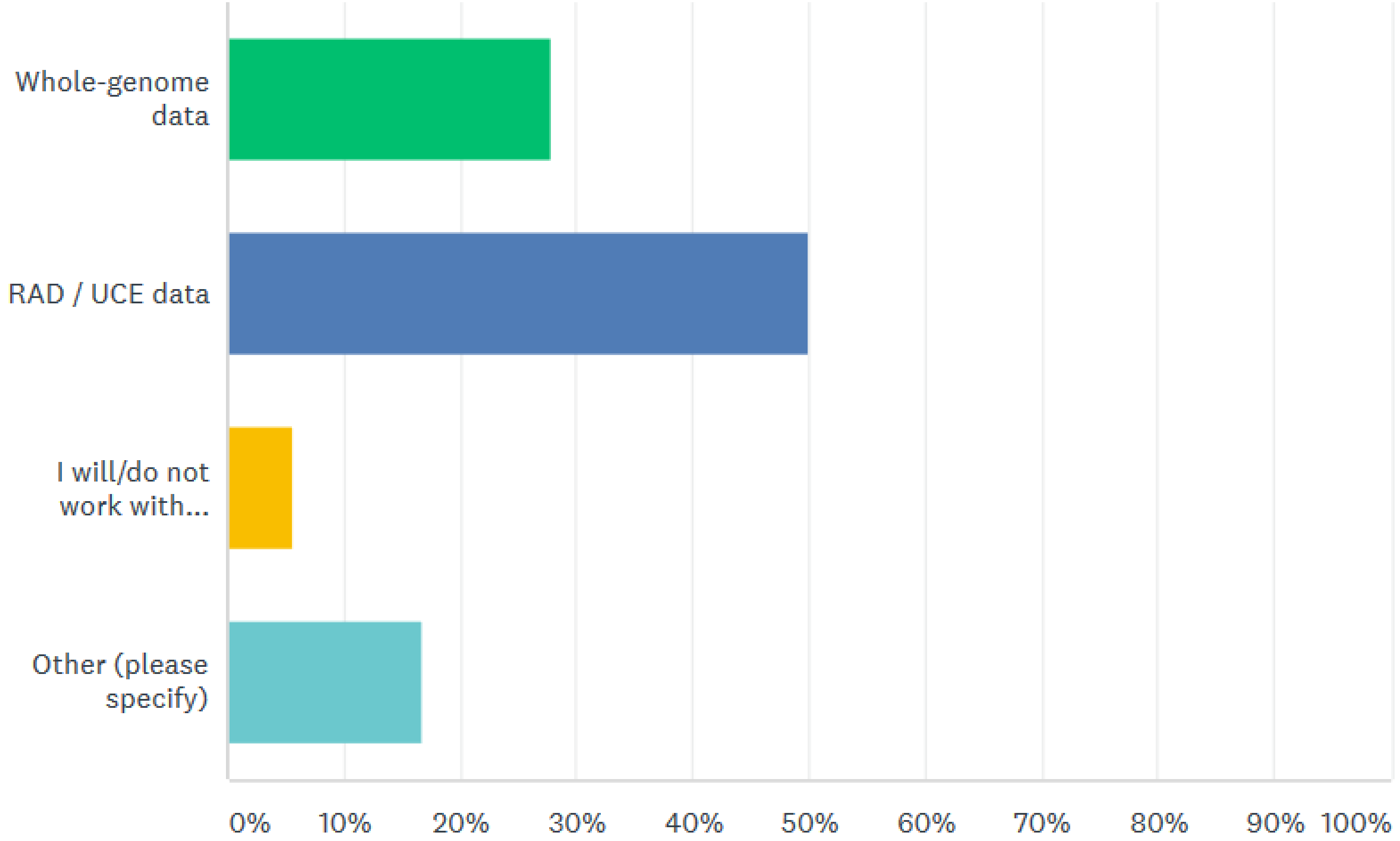
Size selection

paired-end sequencing

Up to 20 billion
read pairs

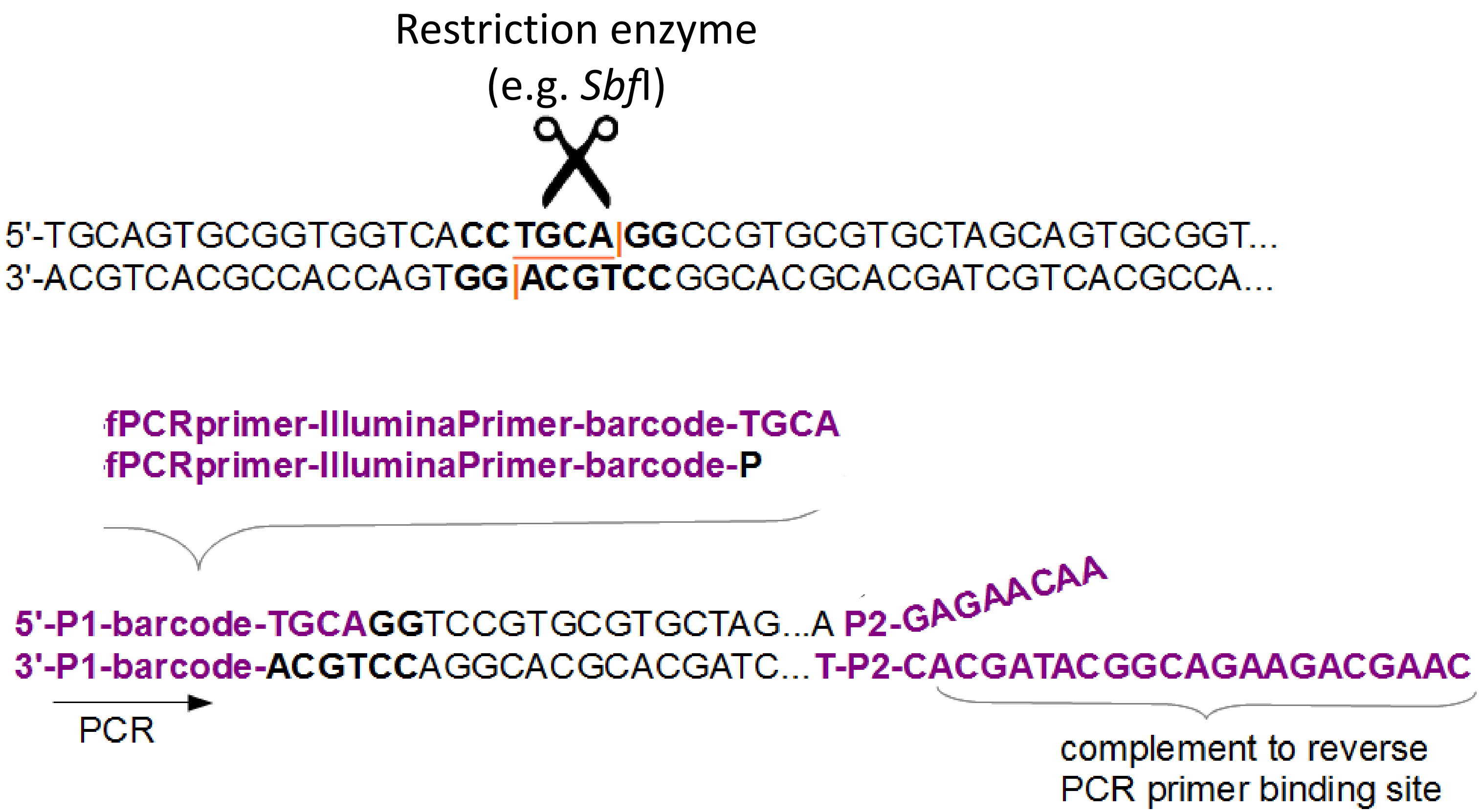
What kind of genomic data will/are you working on?

Answered: 18 Skipped: 1



RAD sequencing

Restriction **A**ssociated **D**NA **sequencing**



Restriction enzyme
(e.g. *Sbf*I)

5'-TGCAGTGCGGTGGTCA**CCTGCA**|**GG**CCGTGCGTGCTAGCAGTGCGGT...
3'-ACGTCACGCCACCAGT**GG**|**ACGTCC**GGCACGCACGATCGTCACGCCA...

fPCRprimer-IlluminaPrimer-barcode-TGCA
fPCRprimer-IlluminaPrimer-barcode-P

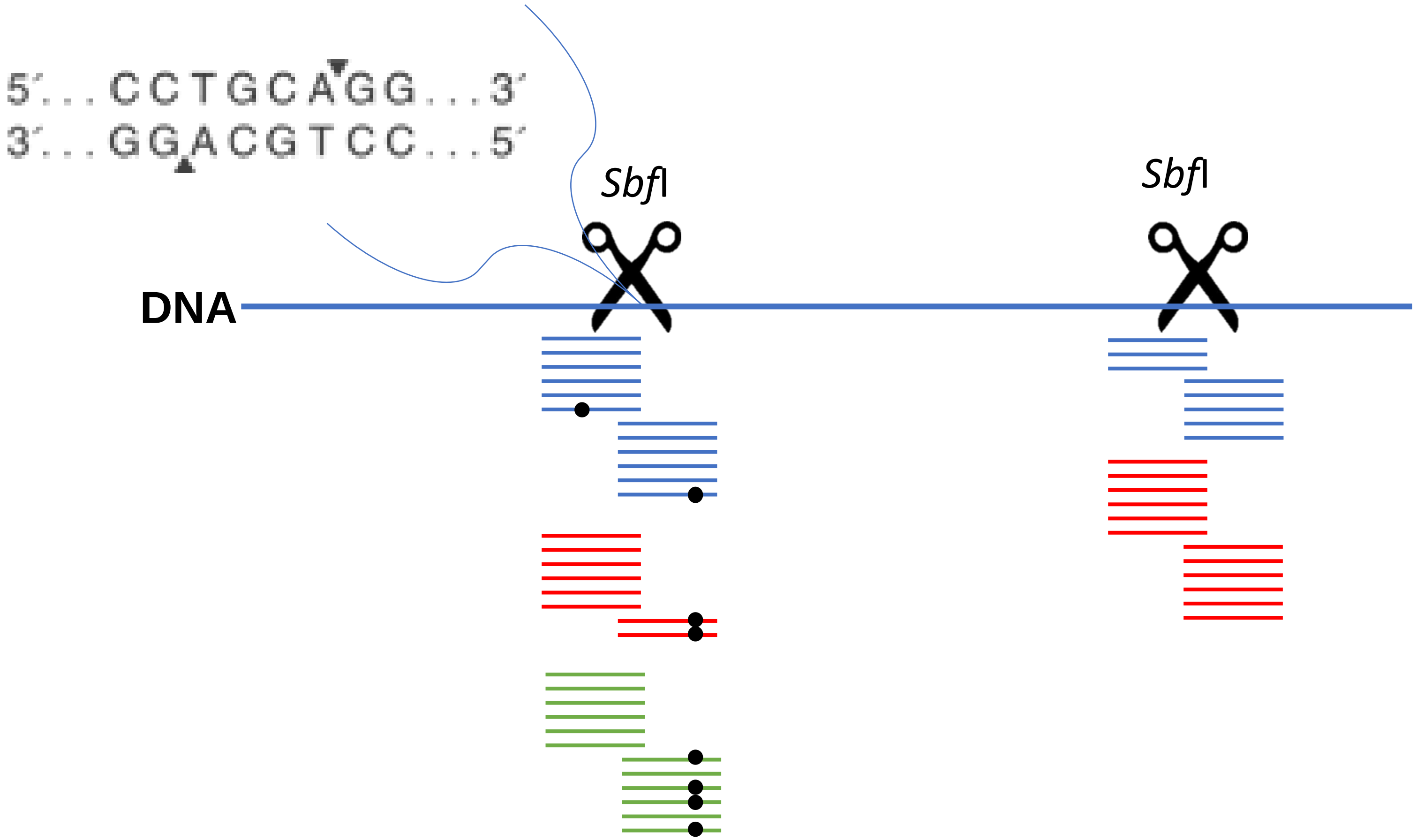
5'-P1-barcode-TGCAGGTCCGTGCGTGCTAG...A **P2-GAGAACAA**
3'-P1-barcode-ACGTCCAGGCACGCACGATC... **T-P2-CACGATACGGCAGAAGACGAAC**

PCR

complement to reverse
PCR primer binding site

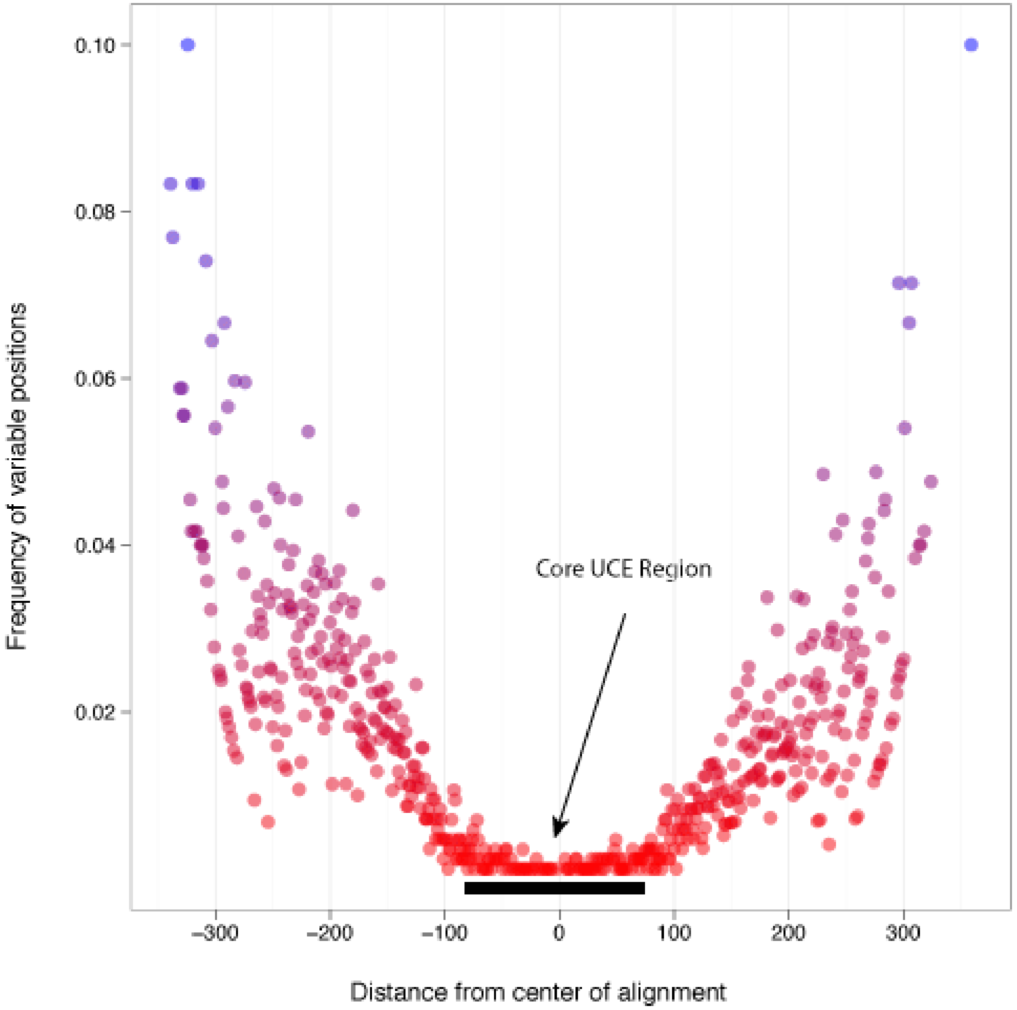
RAD sequencing

Restriction **A**ssociated **DNA sequencing**



Other «reduced-representation» techniques

- **CRoPS/ddRAD sequencing** (double-digest restriction enzyme and size selection inste
- **GBS** (genotyping by sequencing): no sheari and sequencing select short fragments
- **UCE**: Selection of DNA fragments through s on ultraconserved elements (conserved ac
- **2b-RAD**: type IIB restriction enzymes that c downstream of the restriction site
- **Transcriptome sequencing**: RNAseq, only c to annotate, depth informative on expressi
- **Targeted resequencing:** Sequence capture



Trade-offs: Splitting reads (i.e. costs) among:

- Number of sites to sequence
- Number of samples
- Depth of coverage

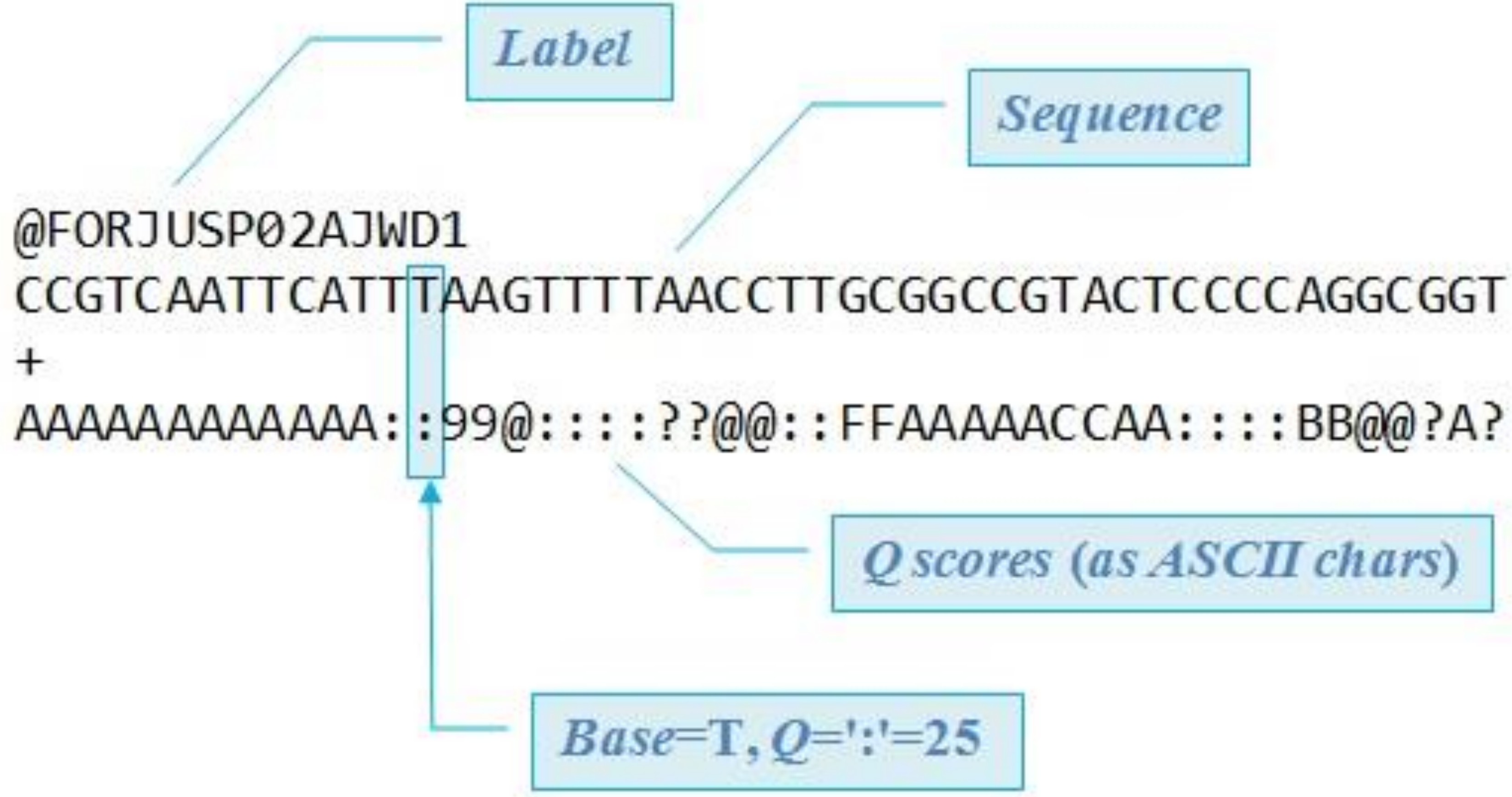
- Example: 1 Hiseq2500 flow cell (about 1000 Euro) ~250 mio read pairs of 125 bp each -> 75 Gb data
  - 5 whole-genomes of a species with 1 Gb genome size at 15x coverage
  - 50 whole-genomes of a species with 500 Mb genome size at 3x coverage
  - 30 Mbp sequenced for 100 samples with a reduced-representation technique at a sequencing depth of 25

Considerations in choosing the library preparation and sequencing techniques

- Research question and planned analyses
- Genome size
- Availability & quality of reference genome (no ref genome -> not wgs)
- Available budget
- Number of samples to sequence (tradeoff with sequencing depth)
- Amounts of DNA available
- Sequencing depth aimed at
- SNP density required
- Divergence between samples
- Heterozygosity of samples
- Phase required
- Accuracy of each single position (if high needed, avoid PCR-based methods)
- Importance of annotations
- Neutral dataset or specific regions wanted

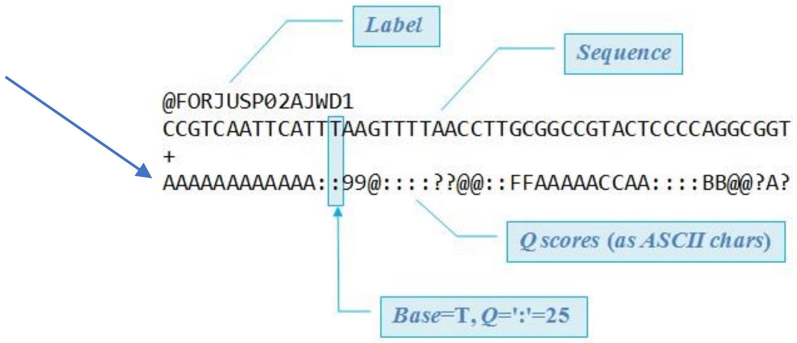
Fastq format

Header



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

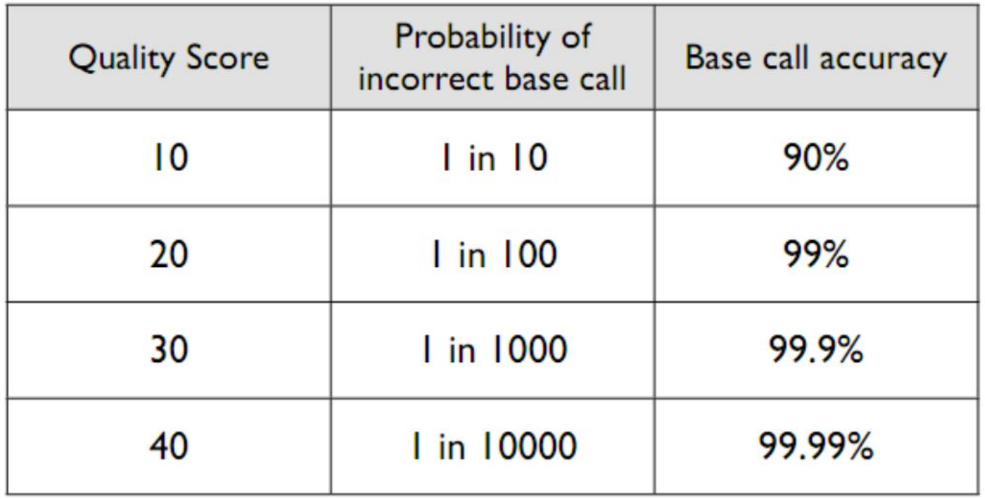
Quality scores



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

$$\text{Phred} = -10 \log_{10} p$$

p = Probability call is incorrect



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 10000	99.99%

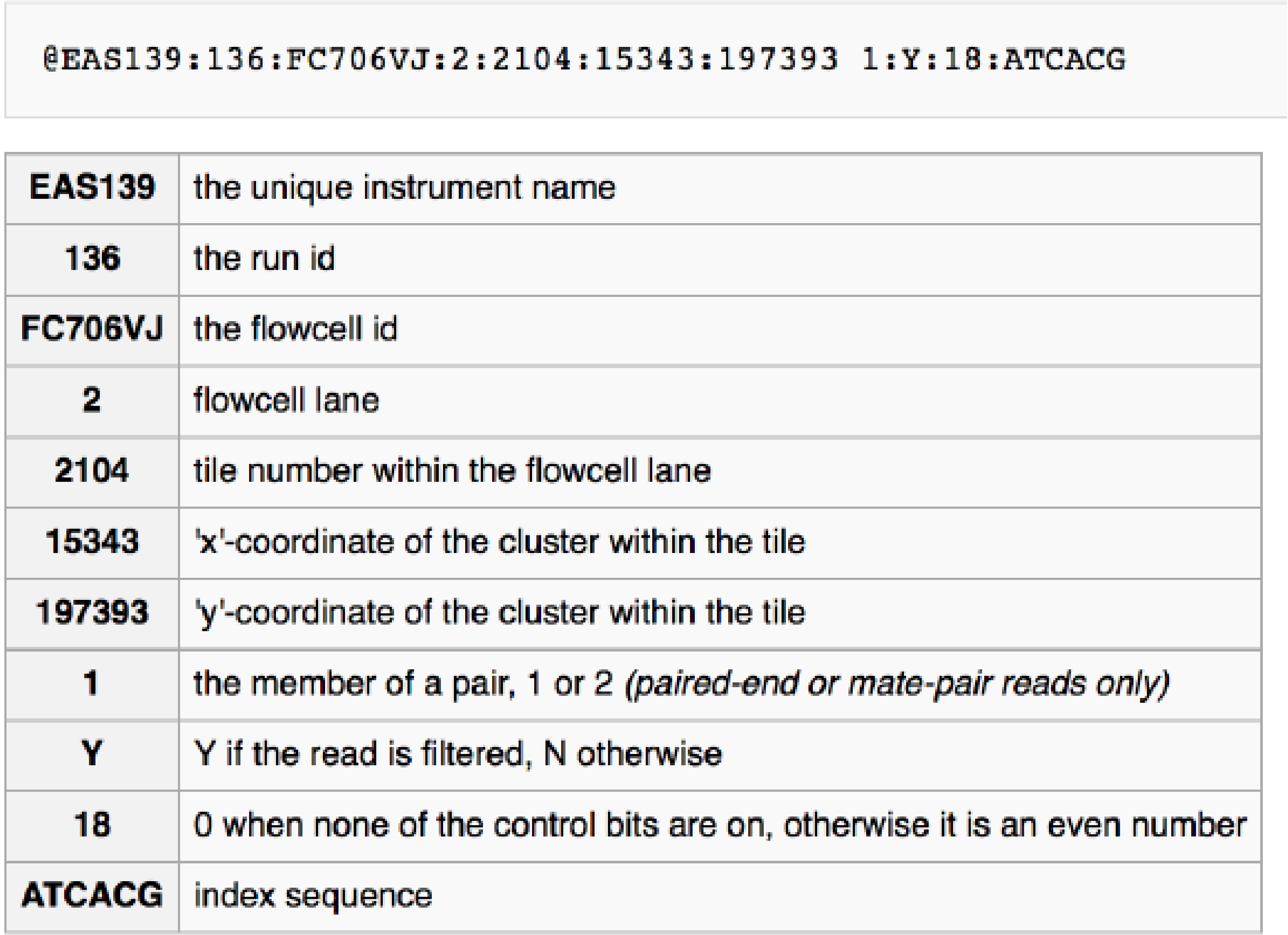
ASCII encoding

40:@
41:A
42:B
43:C
44:D
45:E
… :…

90:Z
91:[
92:\
93:]
94:^
95:_
… :…

141:a
142:b
143:c
144:d
145:e
146:f
… :…

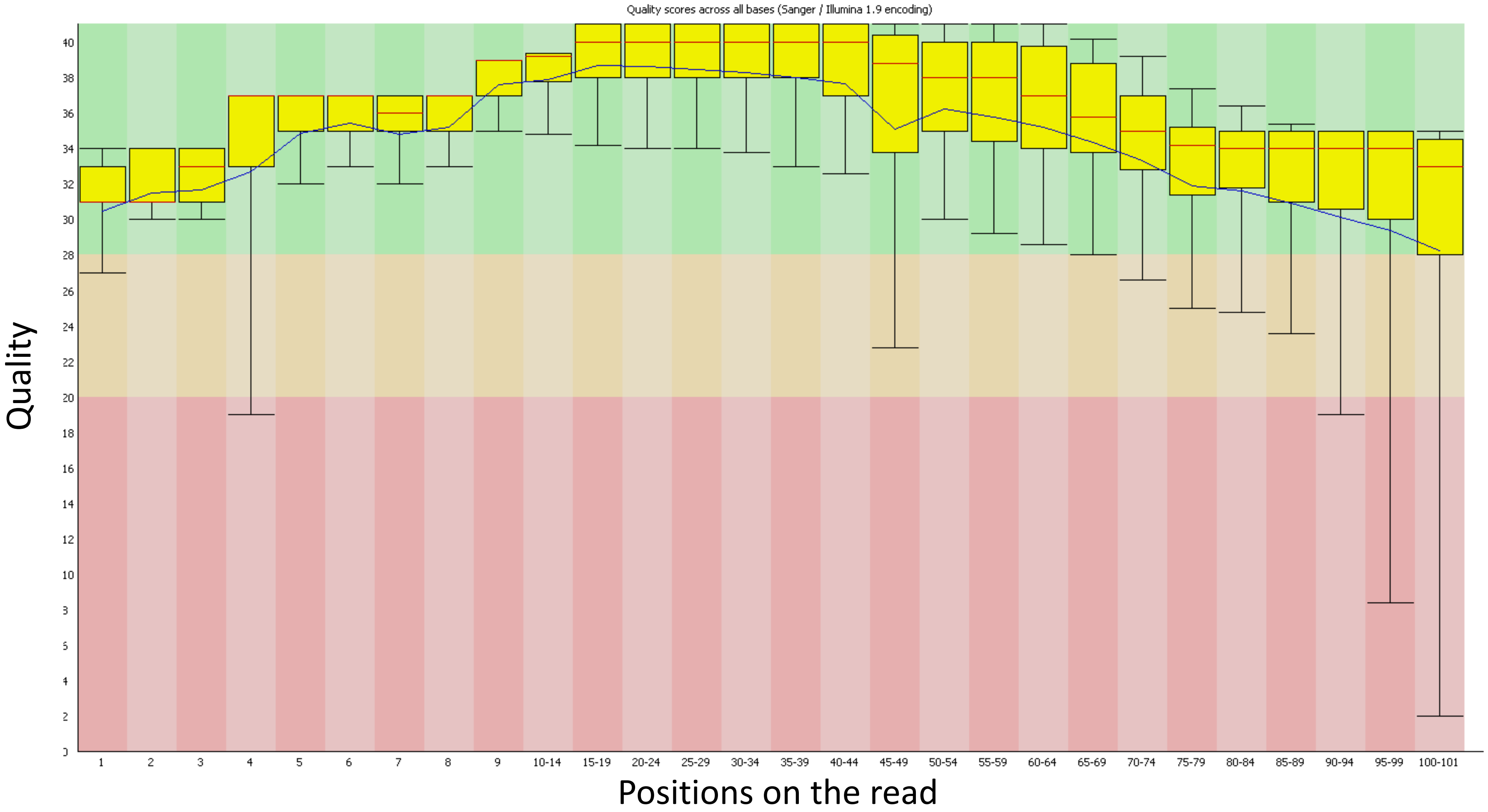
Read header



```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

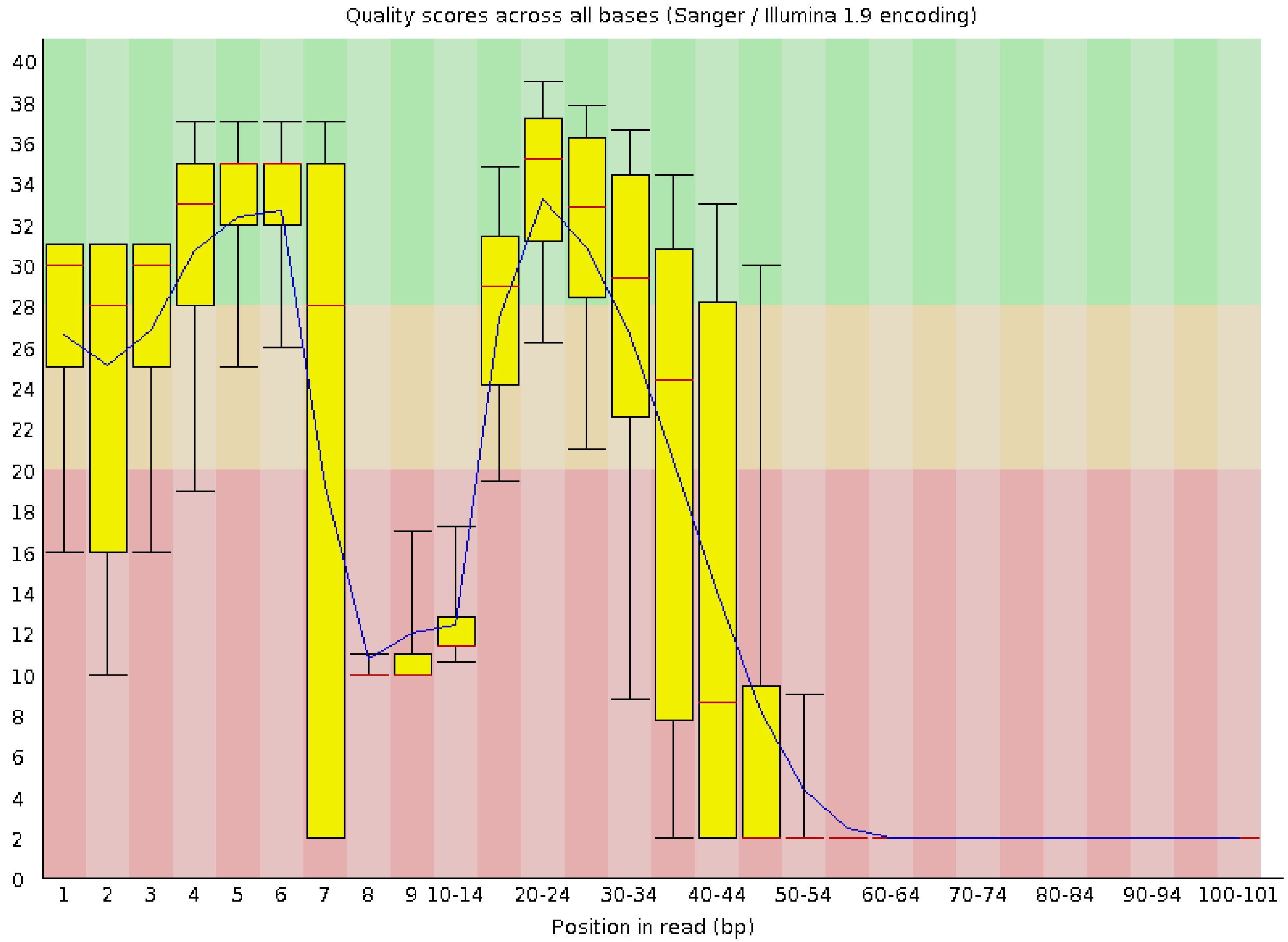
FastQC: Quality across bases (good example)



>Q30
Error prob. < 0.001

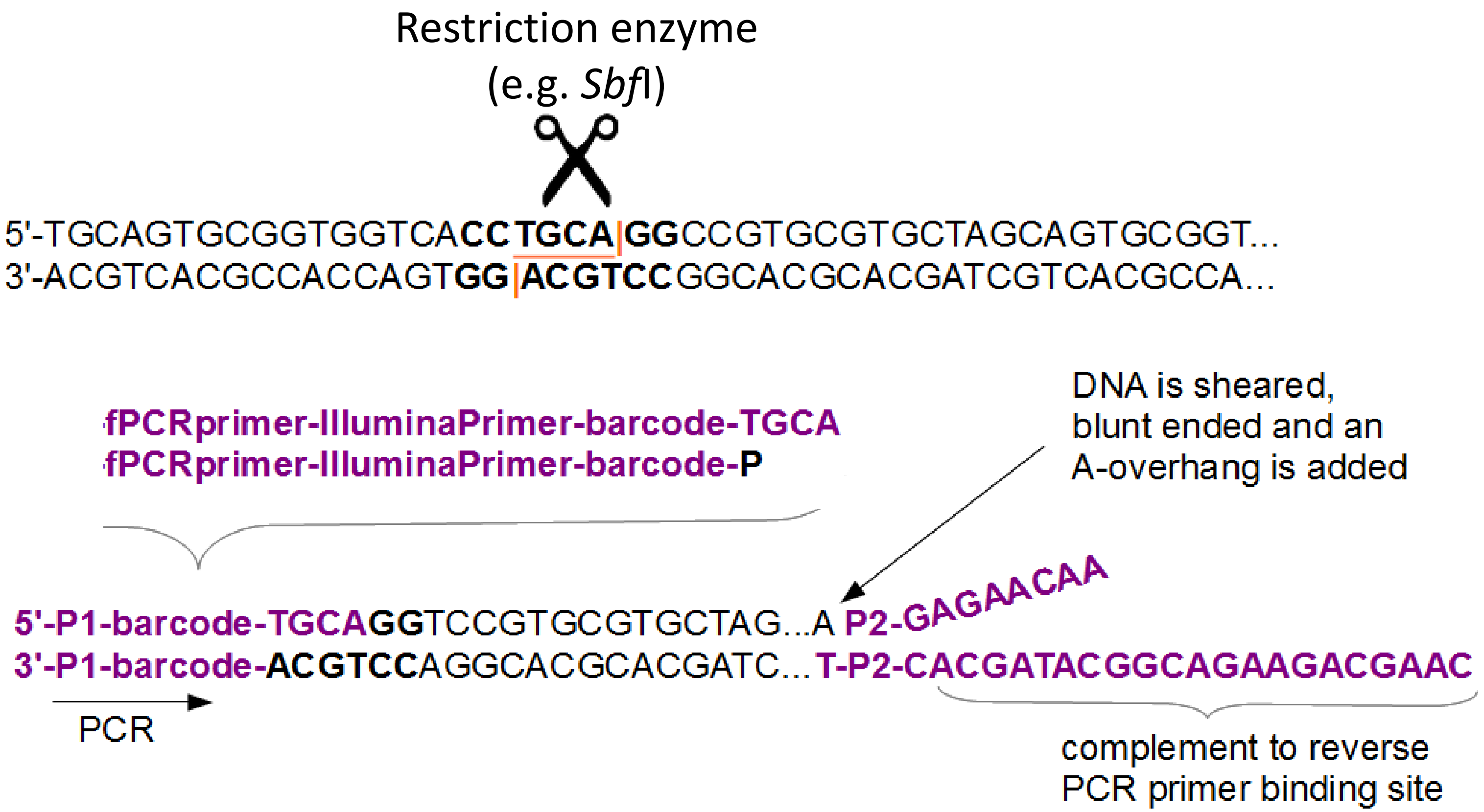
<Q20
Error prob. < 0.01

FastQC: Quality across bases (bad example)



Let’s have a look at the first few sequences and check the sequencing quality with fastqc

RAD/GBS



Restriction enzyme
(e.g. *Sbf*I)

5'-TGCAGTGCGGTGGTCA**CCTGCAGG**CCGTGCGTGCTAGCAGTGCGGT...
3'-ACGTCACGCCACCAGT**GGACGTCC**GGCACGCACGATCGTCACGCCA...

fPCRprimer-IlluminaPrimer-barcode-TGCA
fPCRprimer-IlluminaPrimer-barcode-P

DNA is sheared,
blunt ended and an
A-overhang is added

5'-P1-barcode-TGCAGGTCCGTGCGTGCTAG...A **P2-GAGAACAA**
3'-P1-barcode-ACGTCCAGGCACGCACGATC... **T-P2-CACGATACGGCAGAAGACGAAC**

PCR

complement to reverse
PCR primer binding site

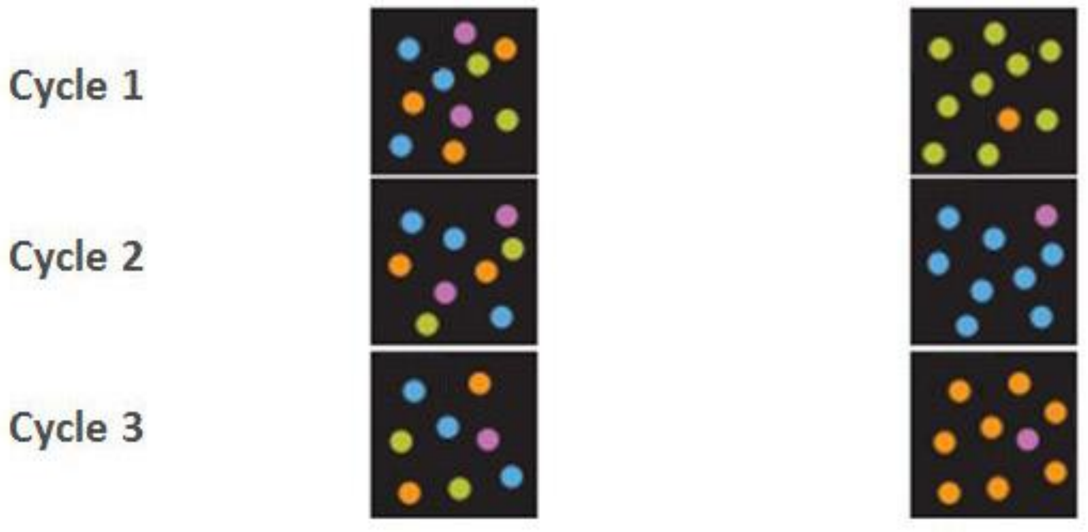
Each read: will start with the barcode, then the restriction site, then a variable sequence

Issues with cluster identification

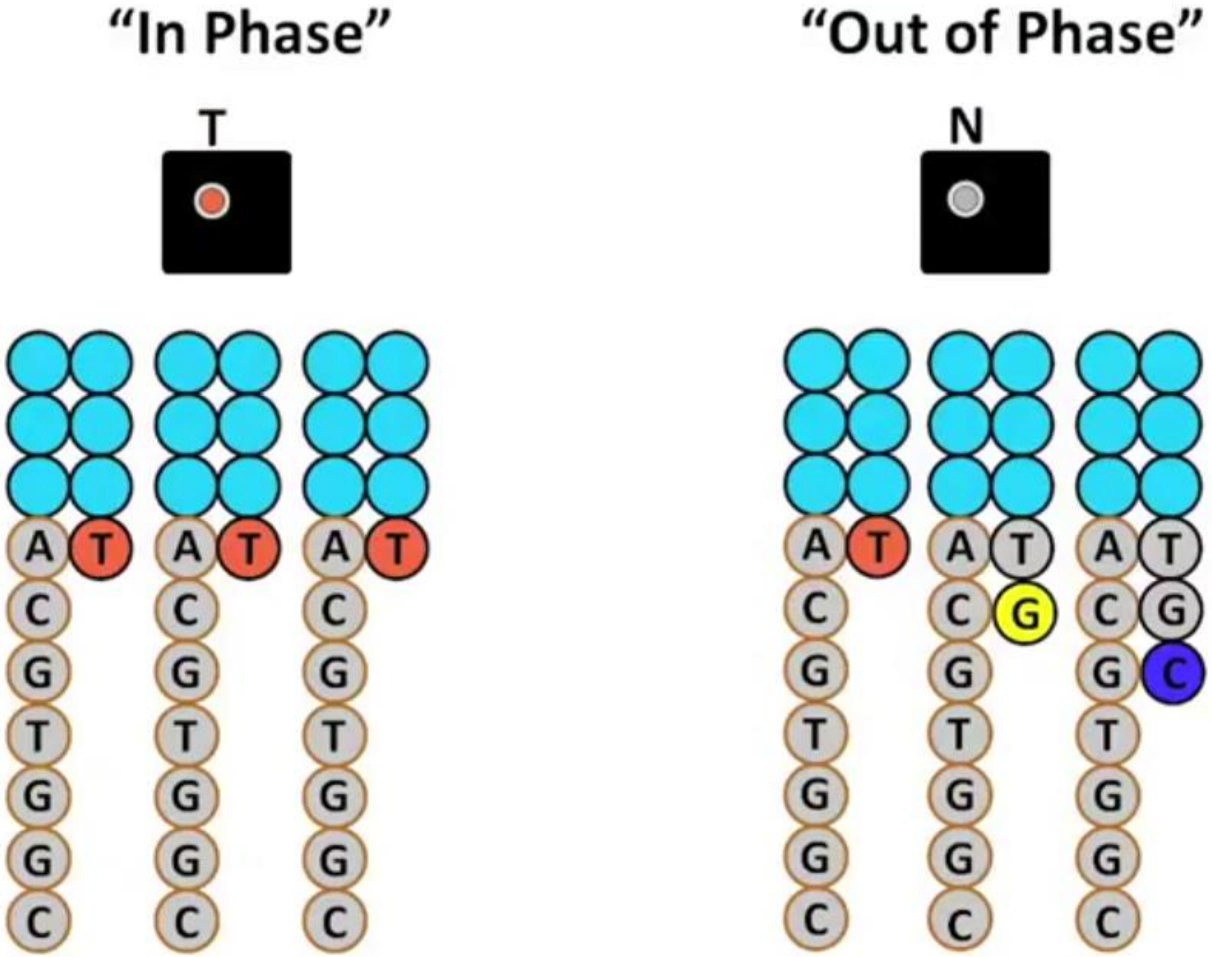
Due to low complexity at the beginning of the sequence,
Illumina cannot distinguish if a signal comes from one or two clusters

balanced

Low complexity



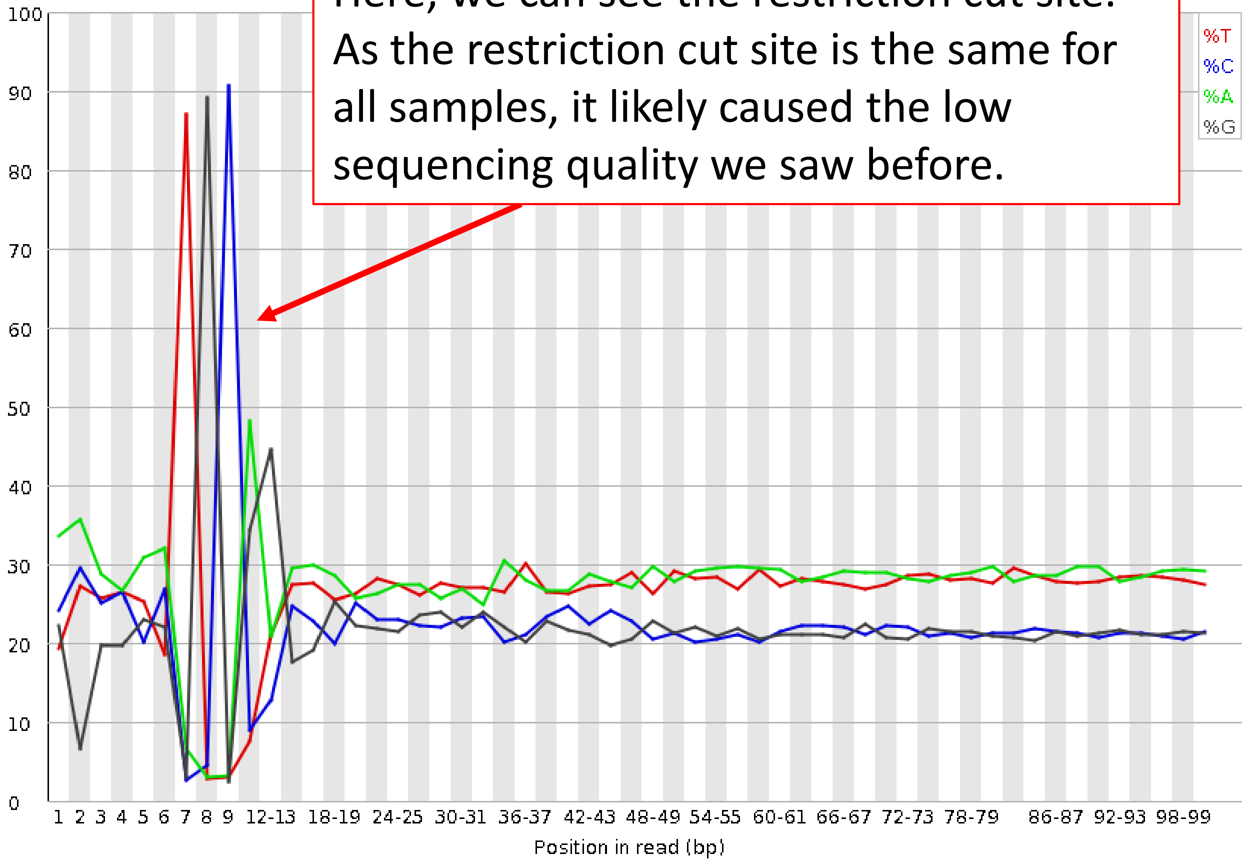
Phasing issues



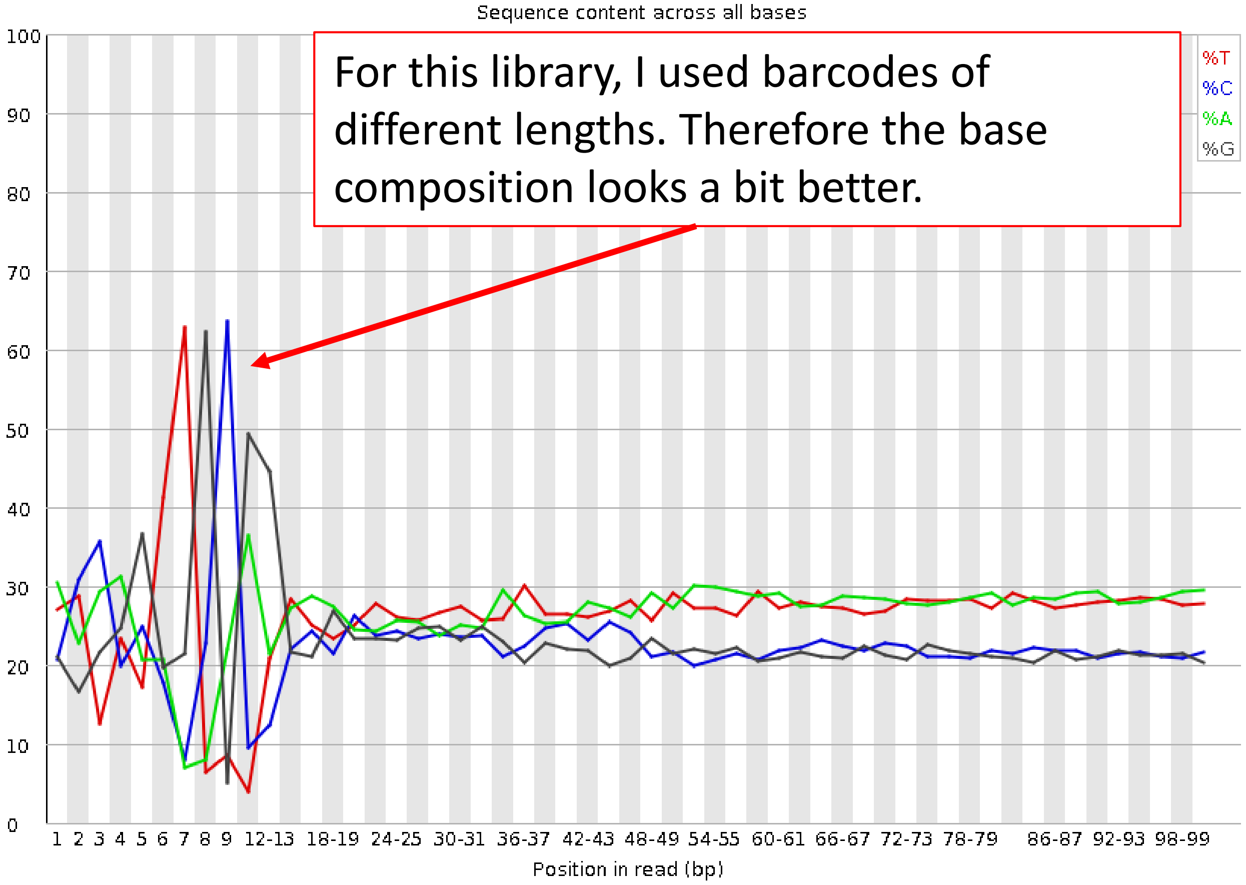
The first 12 nucleotides are also used for «phasing», i.e. correcting for reads that are out of phase. The algorithm expects random nucleotide distribution!

-> Barcodes of the same length may lead to low quality overall

Per base sequence content



Here, we can see the restriction cut site. As the restriction cut site is the same for all samples, it likely caused the low sequencing quality we saw before.



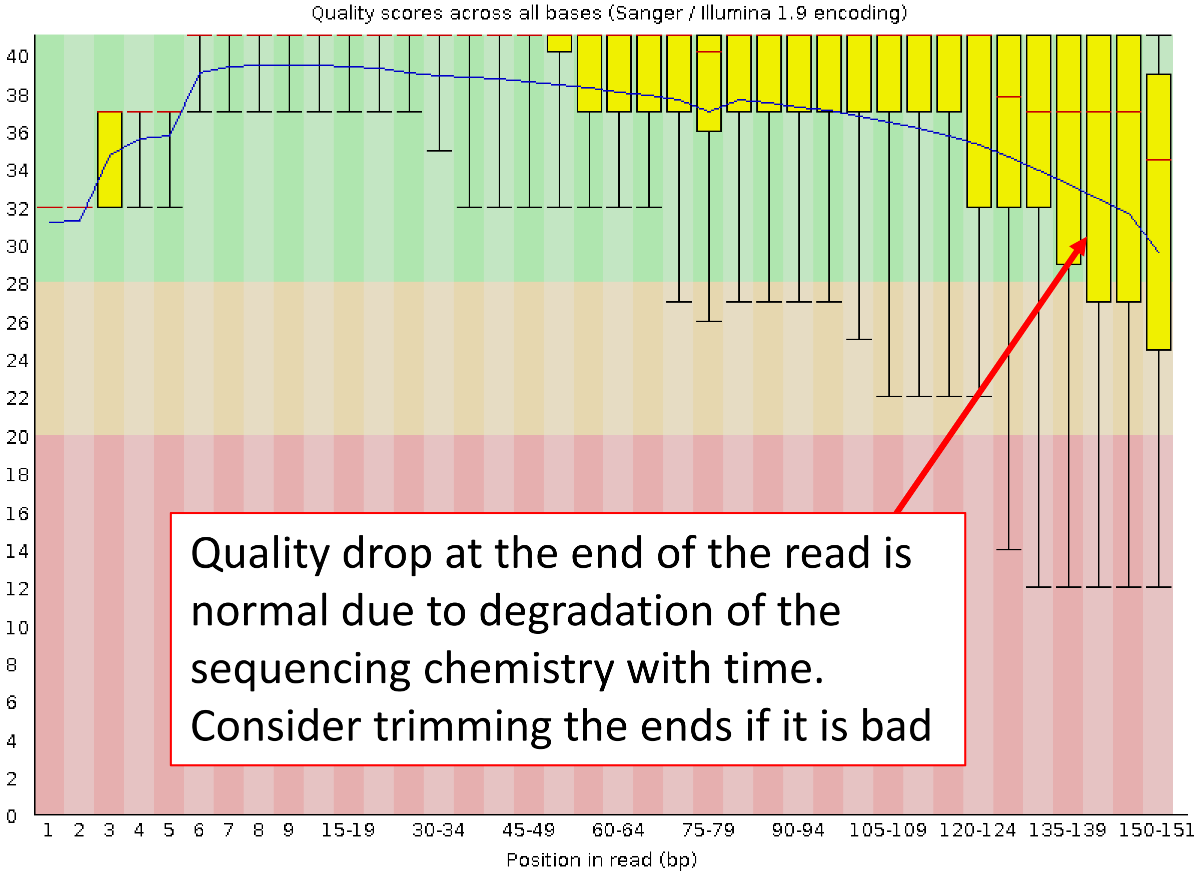
For this library, I used barcodes of different lengths. Therefore the base composition looks a bit better.

How to minimize the problem

- Use barcodes of different lengths to shift the restriction enzyme cut site
- Add PhiX virus DNA to the RAD libraries to increase the complexity of reads (‘spiking’)
- Reduce loading concentrations of Illumina plates
- Potentially: filter out bad reads

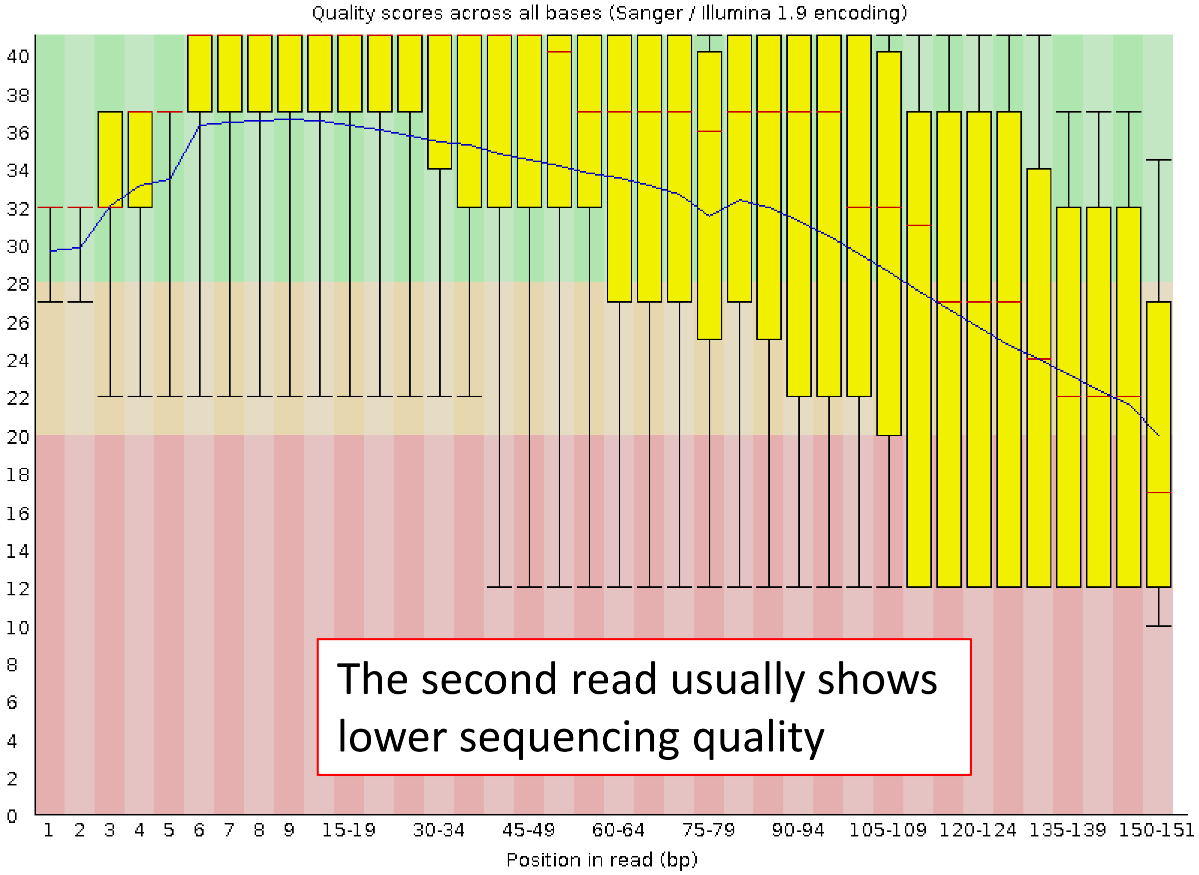
Quality scores across bases: Whole genome sequencing (PCR free library prep)

forward



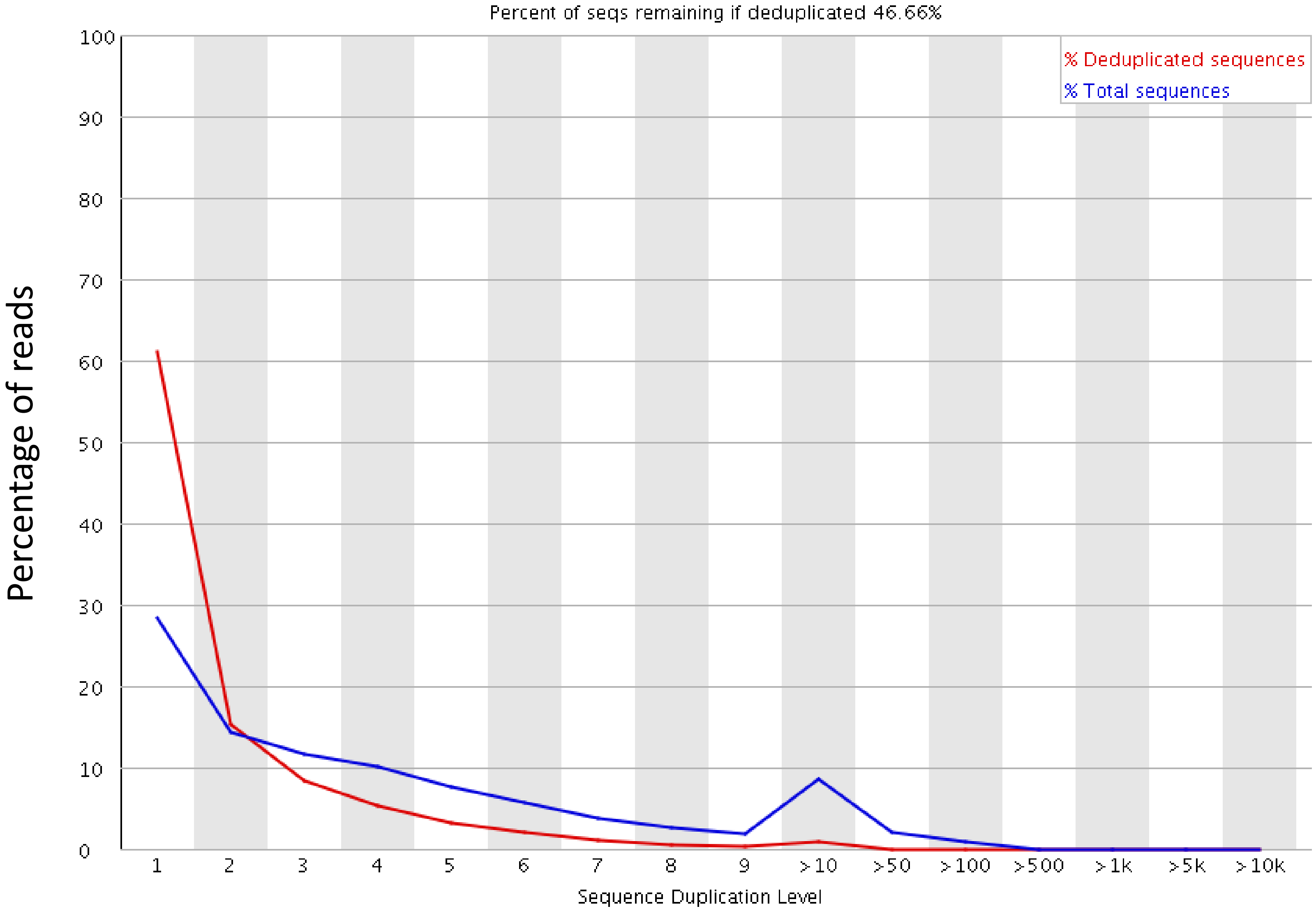
Quality drop at the end of the read is normal due to degradation of the sequencing chemistry with time. Consider trimming the ends if it is bad

reverse



The second read usually shows lower sequencing quality

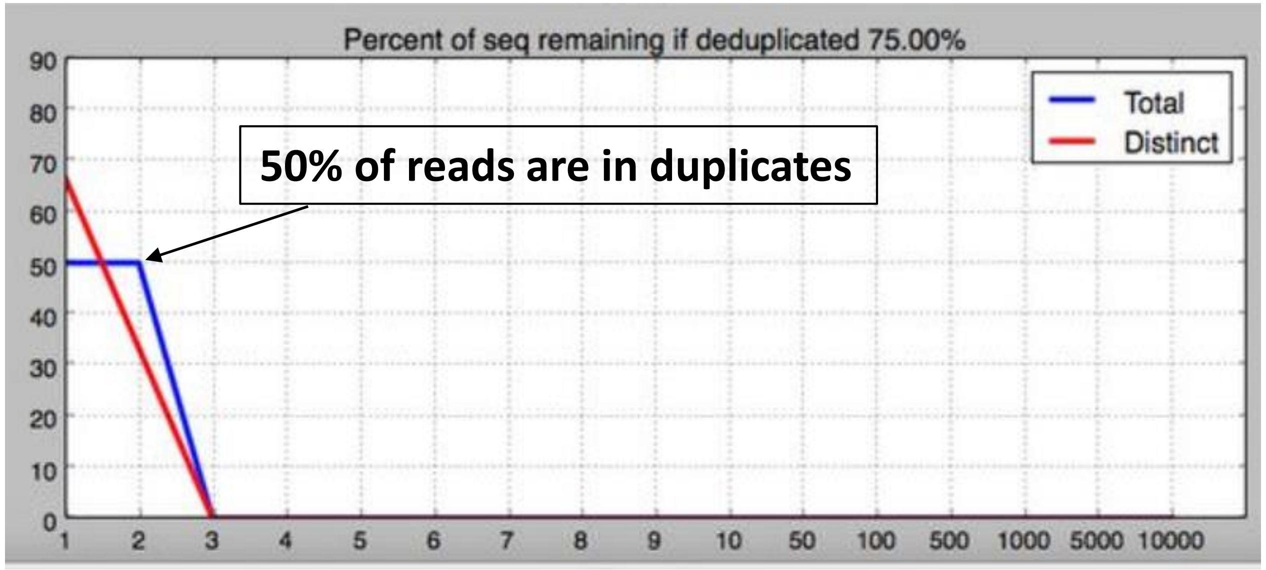
Sequence duplication level



**Duplication level:
= Percentage of reads that have x copies**

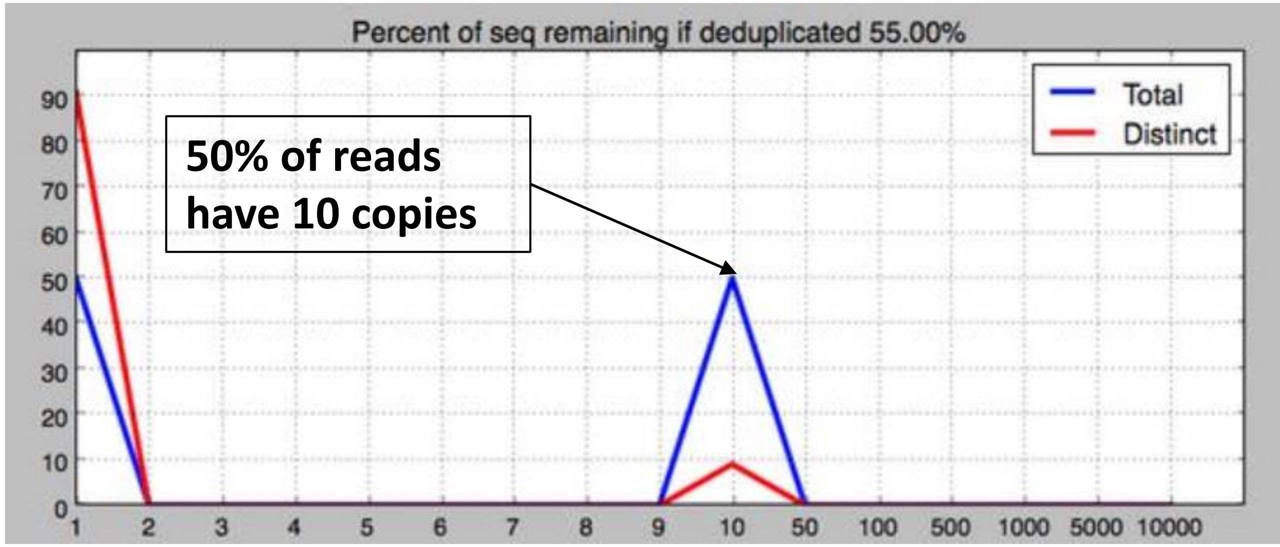
Sequence duplication level

Example 1: 20 reads total
10 unique sequences + 5 sequences each present twice



50% of reads are in duplicates

Example 2: 20 reads total
10 unique sequences + 1 sequence present 10x



50% of reads have 10 copies

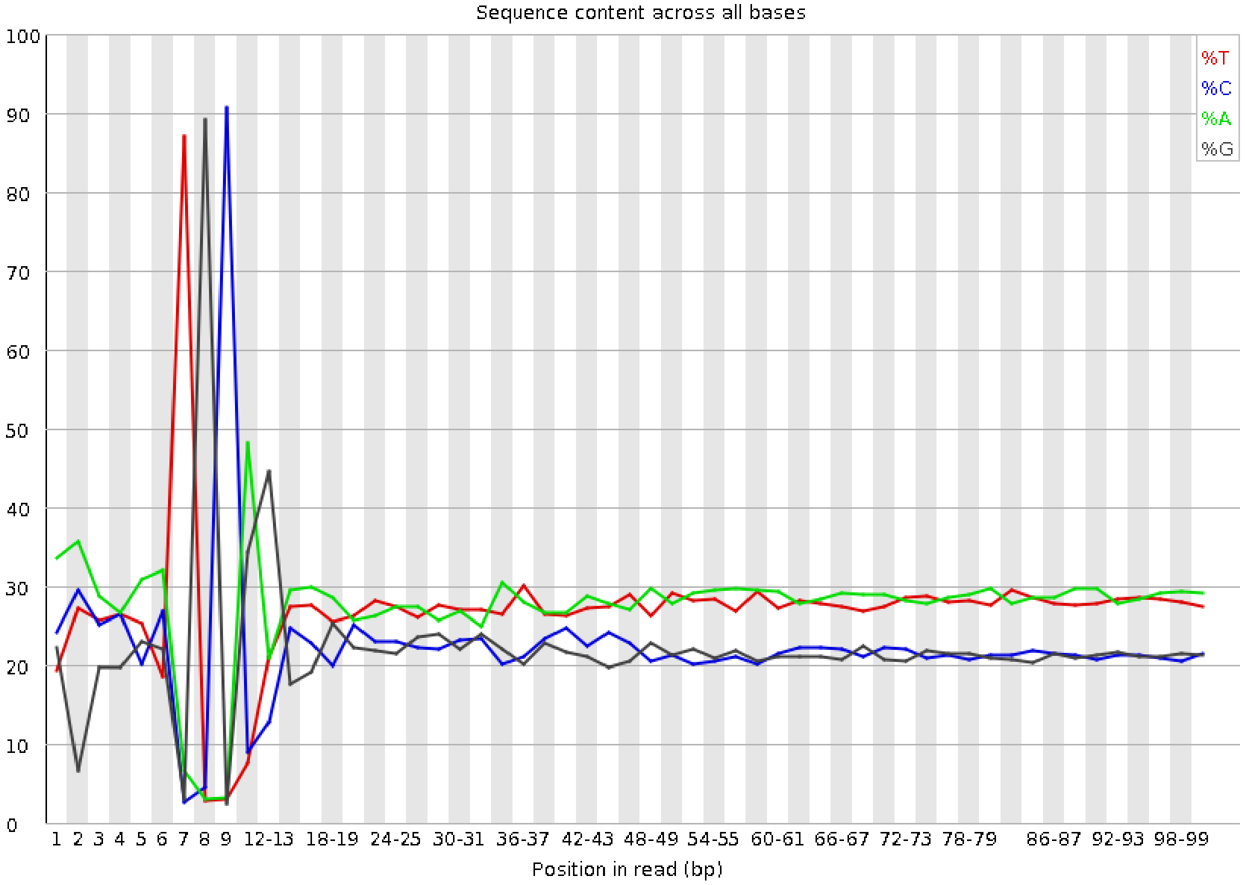
Deduplicated sequences (=number of distinct copies)
15 distinct sequences are distributed as 10 singletons and 5 duplicates, 10/15=66% and 5/15=33% is the slope of the red line. Thus 15/20=75% remaining after deduplication (distinct reads).

Deduplicated sequences (=number of distinct copies)
11 total groups where 10/11=91% are singletons and 1/11=9% of the groups form at duplication rate of 10x. Therefore, 11/20 = 55% distinct reads.

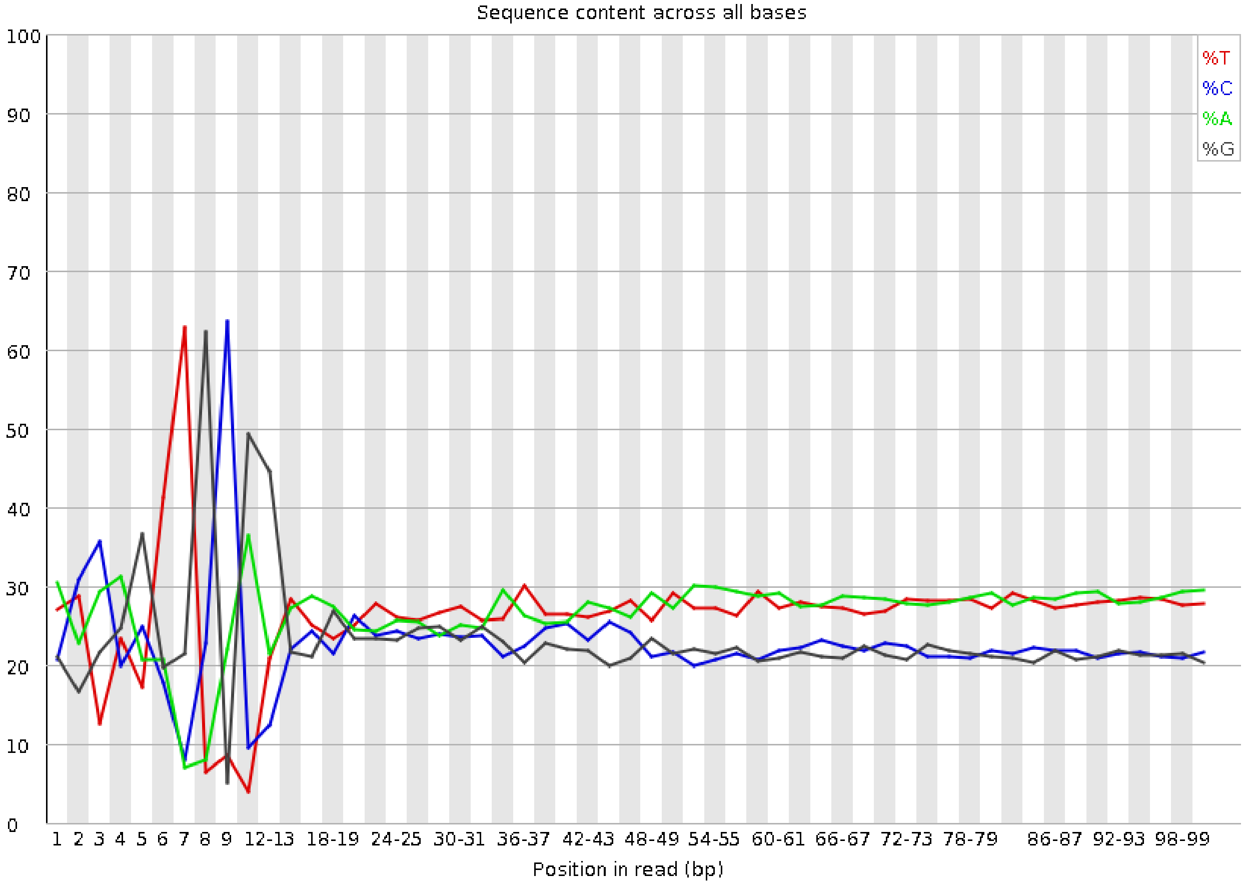
Examples and plots from https://www.biostars.org/p/107402/

Per base sequence content

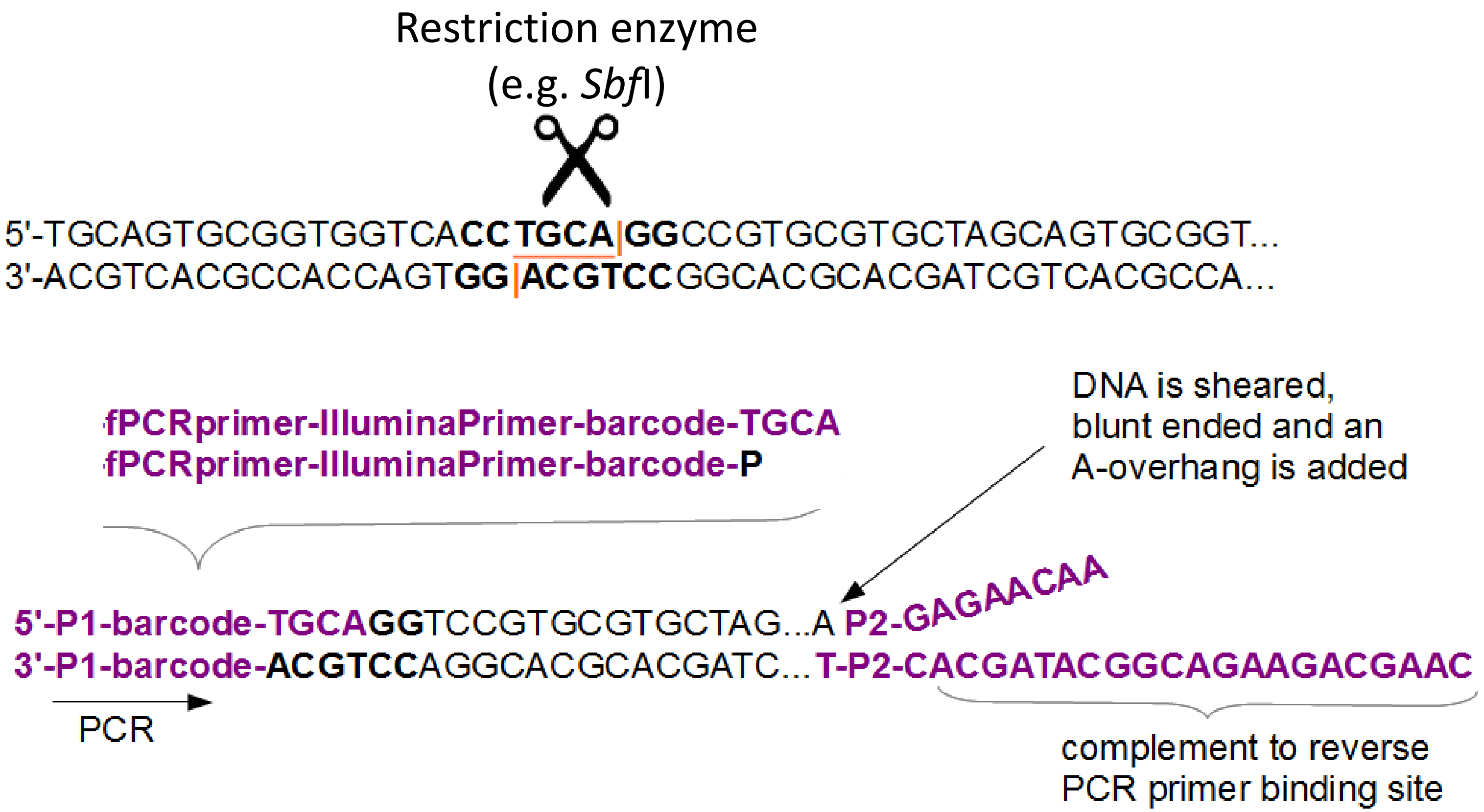
RAD1



RAD2



RAD/GBS



Restriction enzyme
(e.g. *Sbf*I)

5'-TGCAGTGCGGTGGTCA**CCTGCA**|**GG**CCGTGCGTGCTAGCAGTGCGGT...
3'-ACGTCACGCCACCAGT**GG**|**ACGTCC**GGCACGCACGATCGTCACGCCA...

fPCRprimer-IlluminaPrimer-barcode-TGCA
fPCRprimer-IlluminaPrimer-barcode-P

DNA is sheared,
blunt ended and an
A-overhang is added

5'-P1-barcode-TGCAGGTCCGTGCGTGCTAG...A **P2-GAGAACAA**
3'-P1-barcode-ACGTCCAGGCACGCACGATC... **T-P2-CACGATACGGCAGAAGACGAAC**

PCR

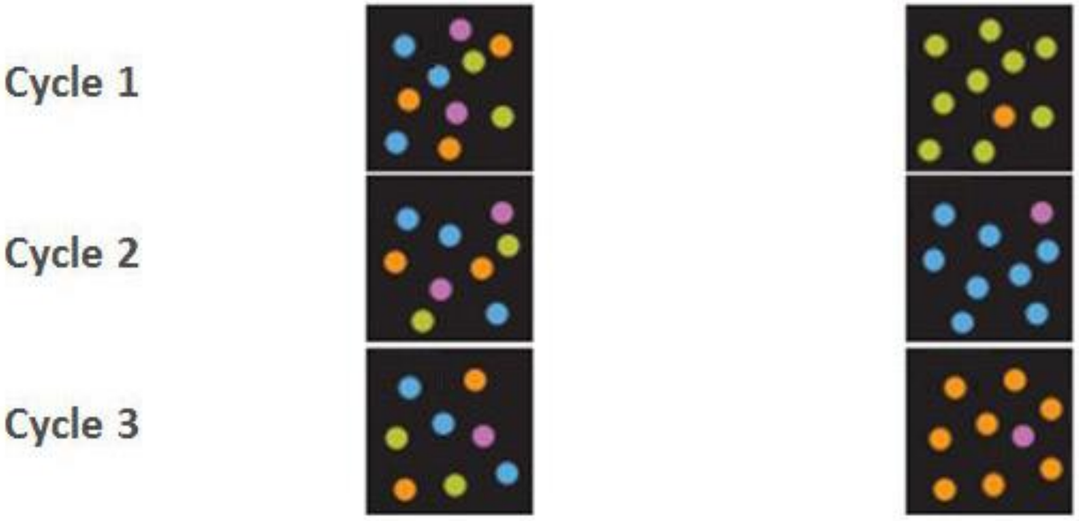
complement to reverse
PCR primer binding site

Each read: will start with the barcode, then the restriction site, then a variable sequence

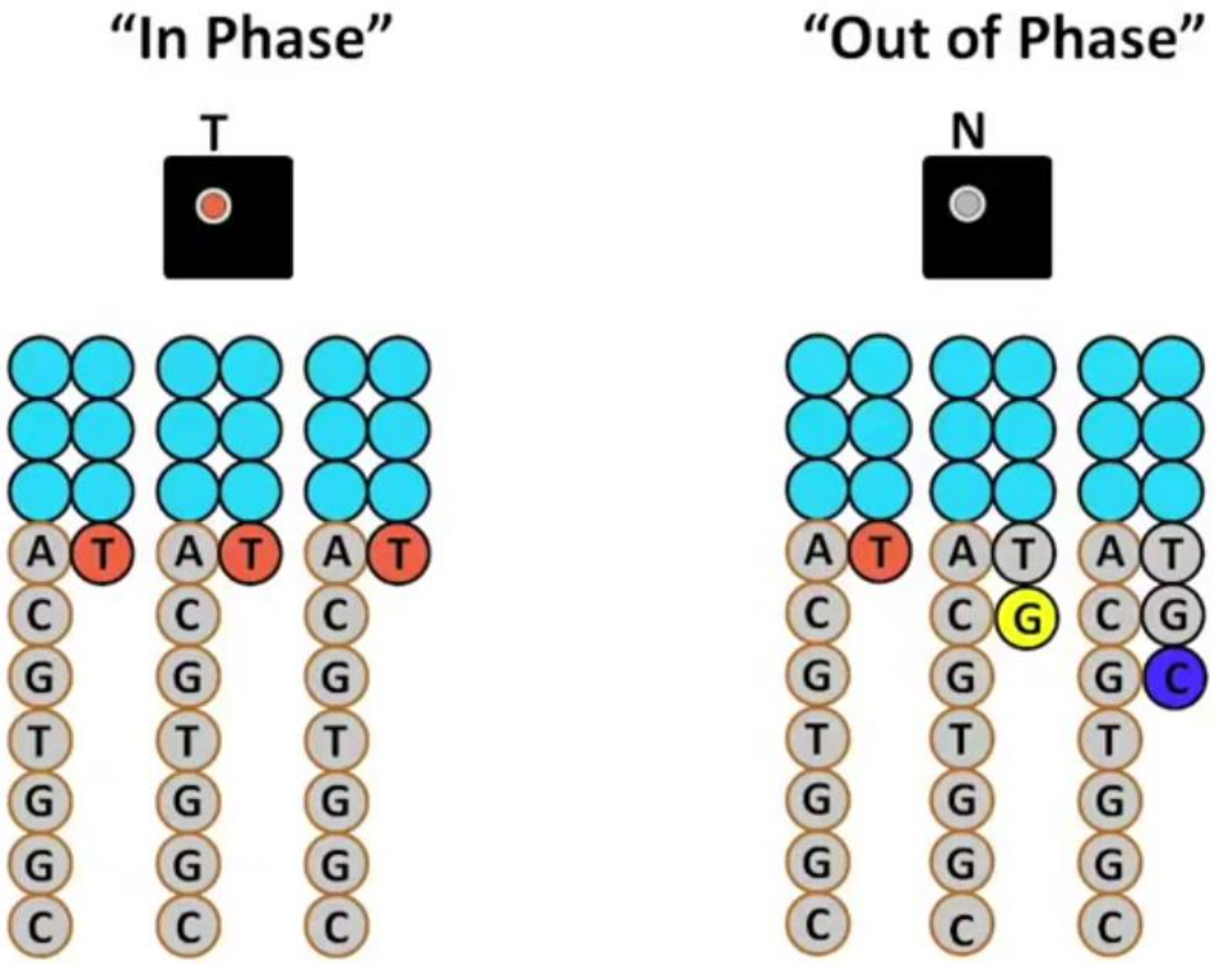
Issues with cluster identification

Due to low complexity at the beginning of the sequence,
Illumina cannot distinguish if a signal comes from one or two clusters

balanced Low complexity



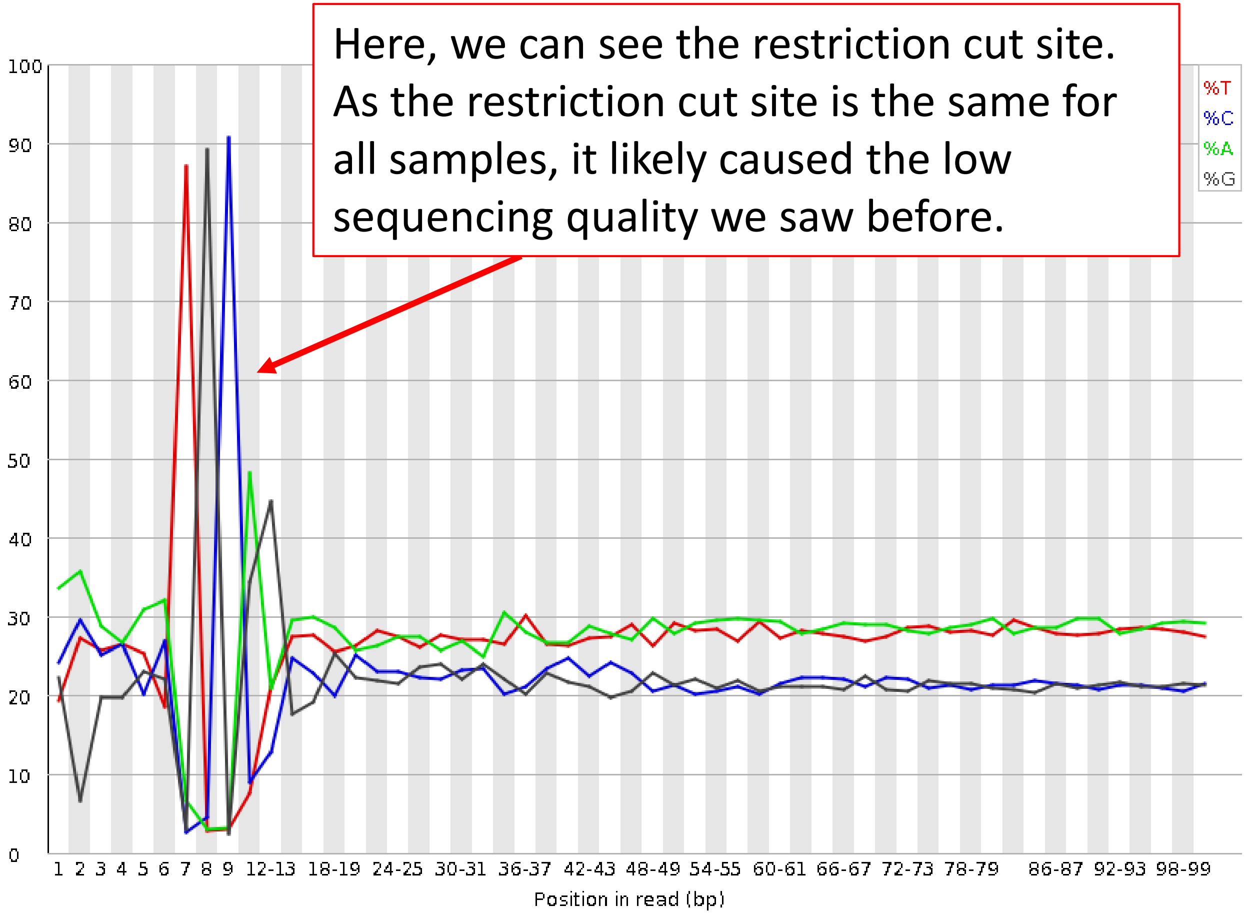
Phasing issues

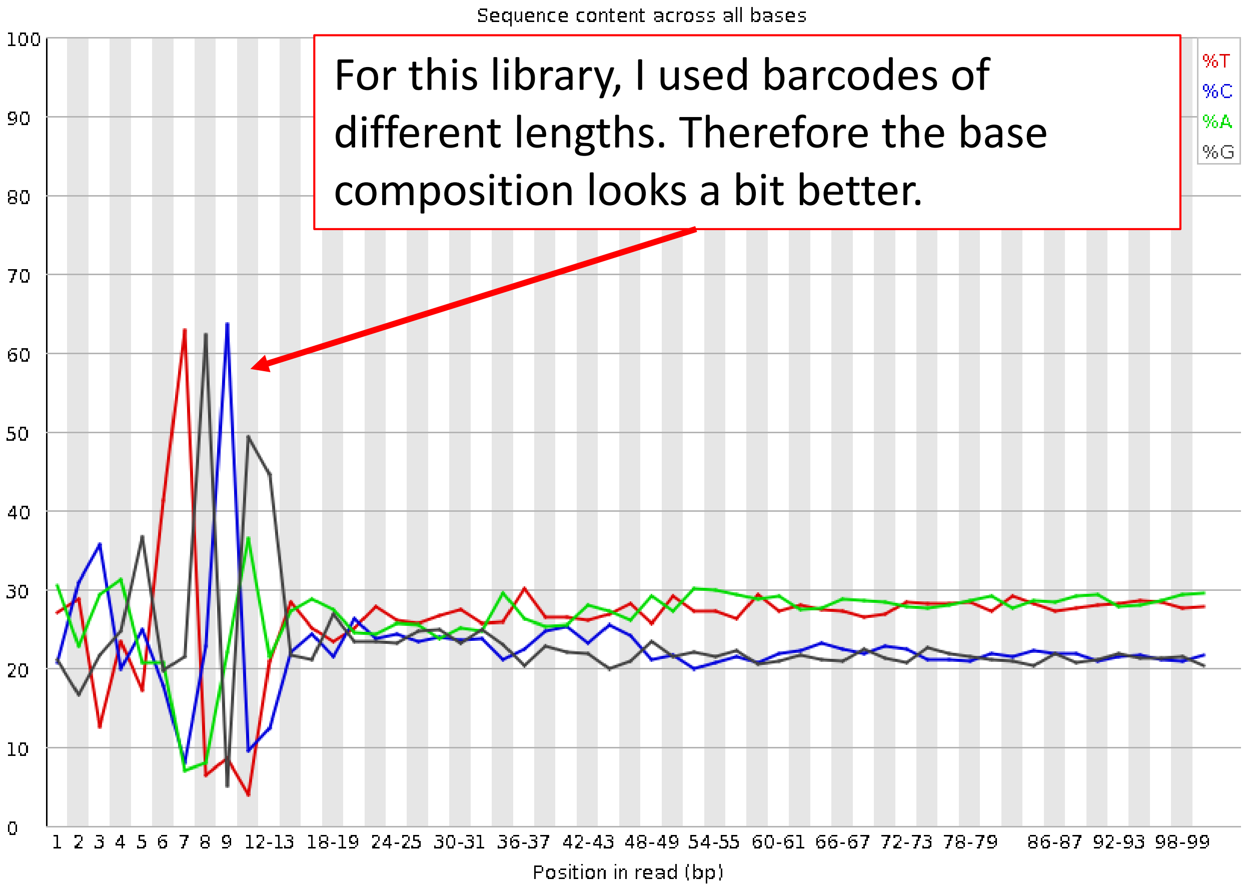


The first 12 nucleotides are also used for «phasing», i.e. correcting for reads that are out of phase. The algorithm expects random nucleotide distribution!

-> Barcodes of the same length may lead to low quality overall

Per base sequence content


Here, we can see the restriction cut site. As the restriction cut site is the same for all samples, it likely caused the low sequencing quality we saw before.

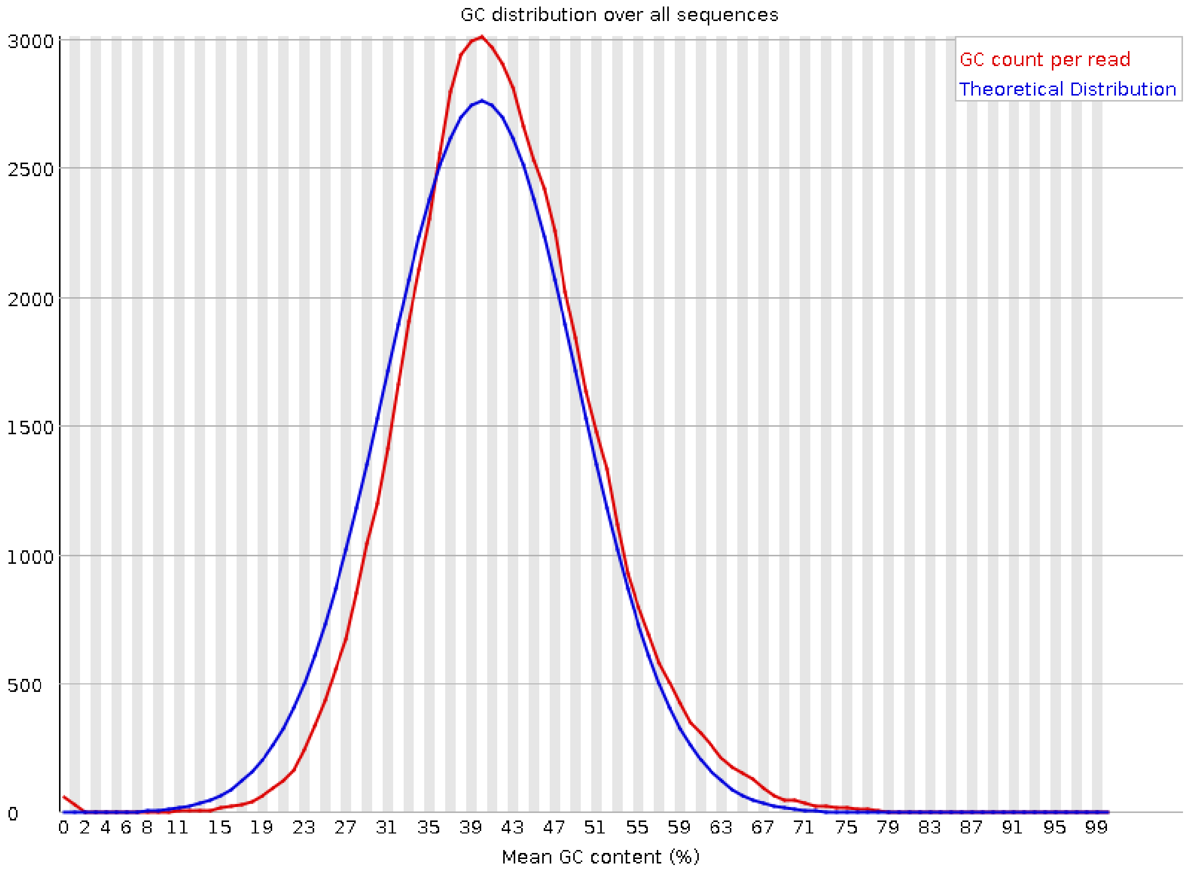

For this library, I used barcodes of different lengths. Therefore the base composition looks a bit better.

How to minimize the problem

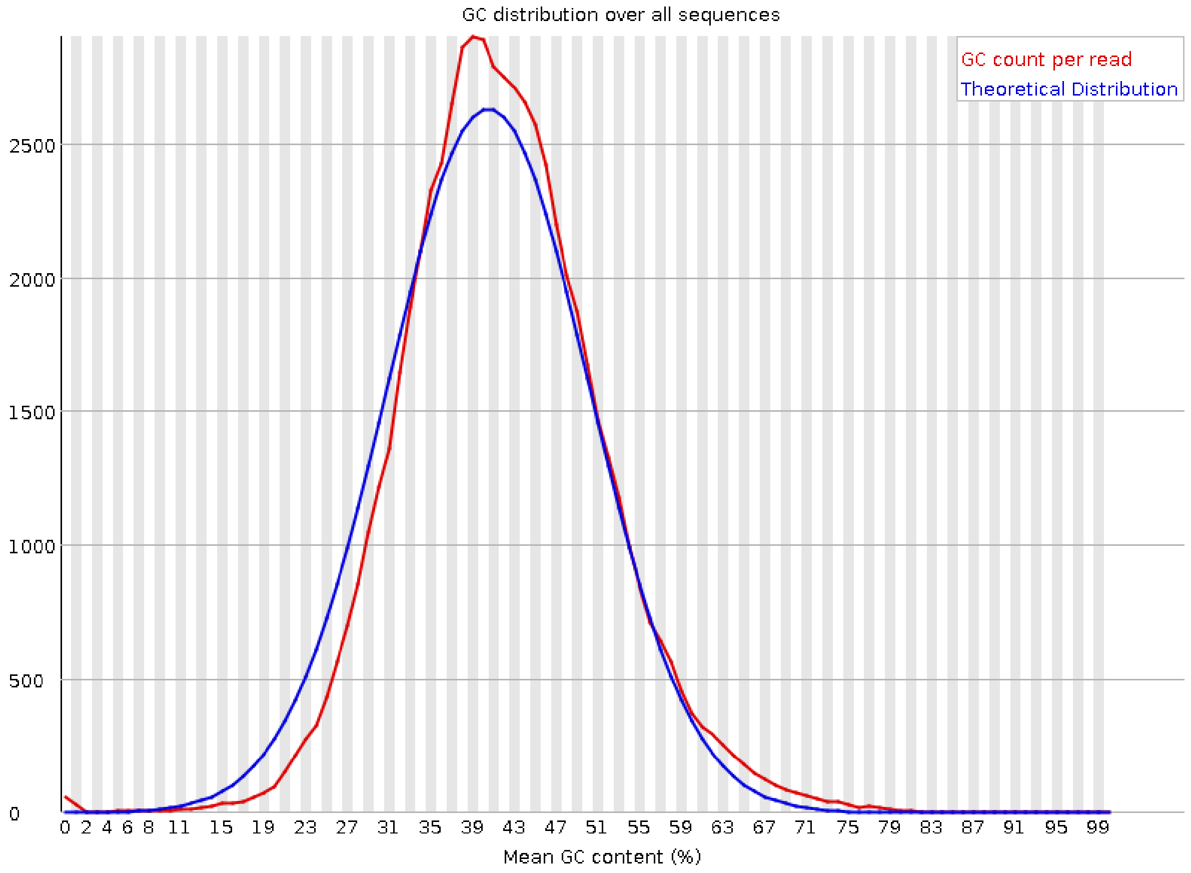
- Use barcodes of different lengths to shift the restriction enzyme cut site
- Add PhiX virus DNA to the RAD libraries to increase the complexity of reads (‘spiking’)
- Reduce loading concentrations of Illumina plates
- Potentially: filter out bad reads

GC distribution over all sequences

forward

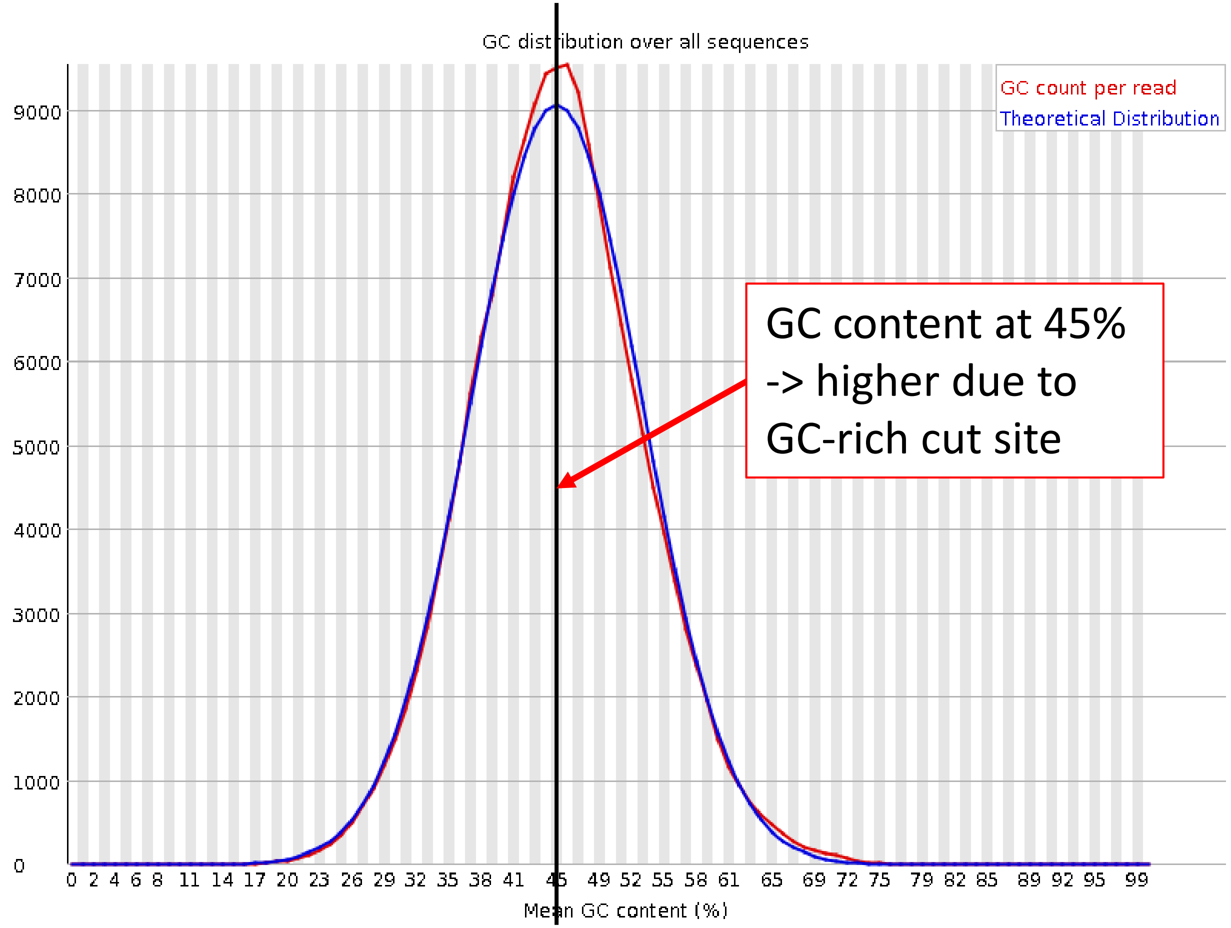


reverse



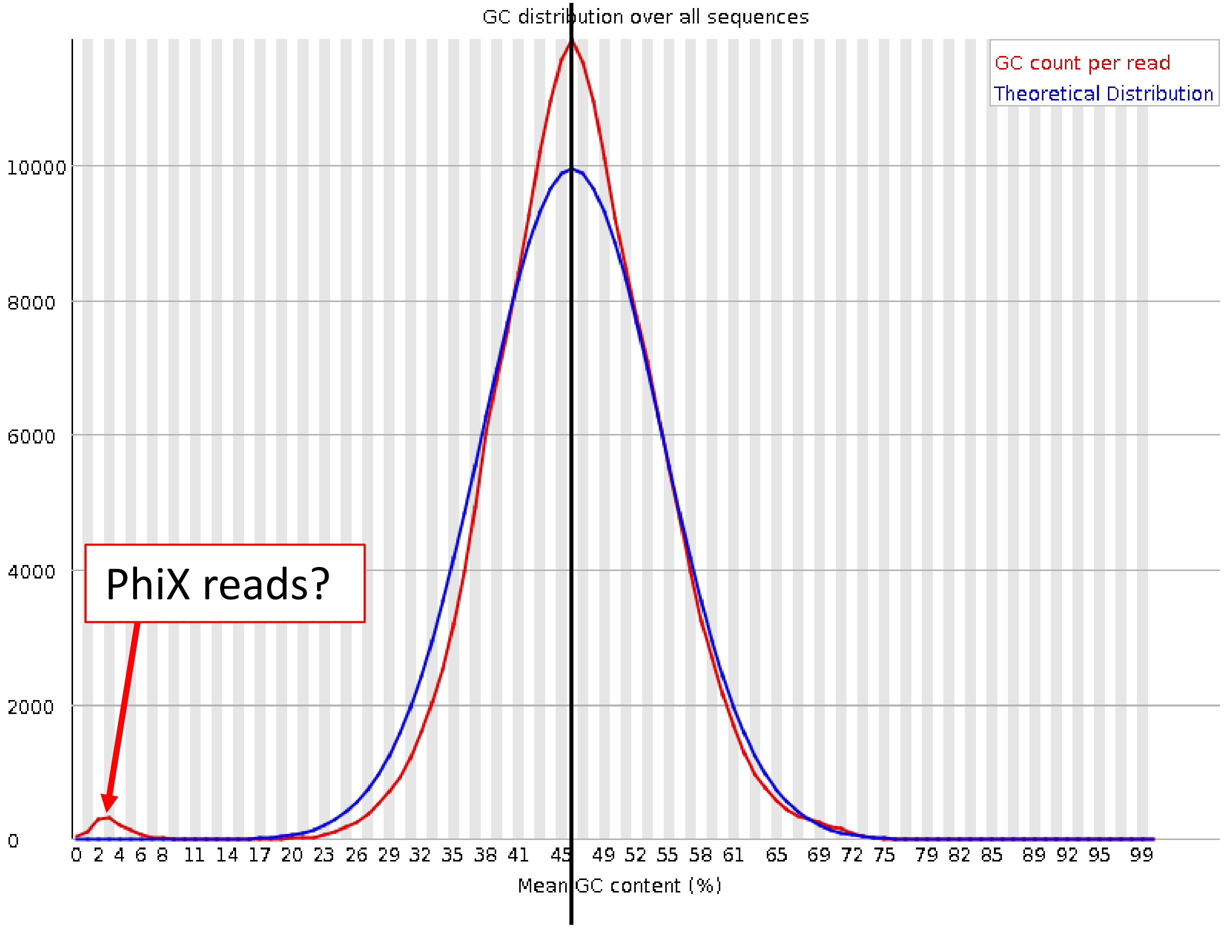
GC distribution over all sequences

RAD1



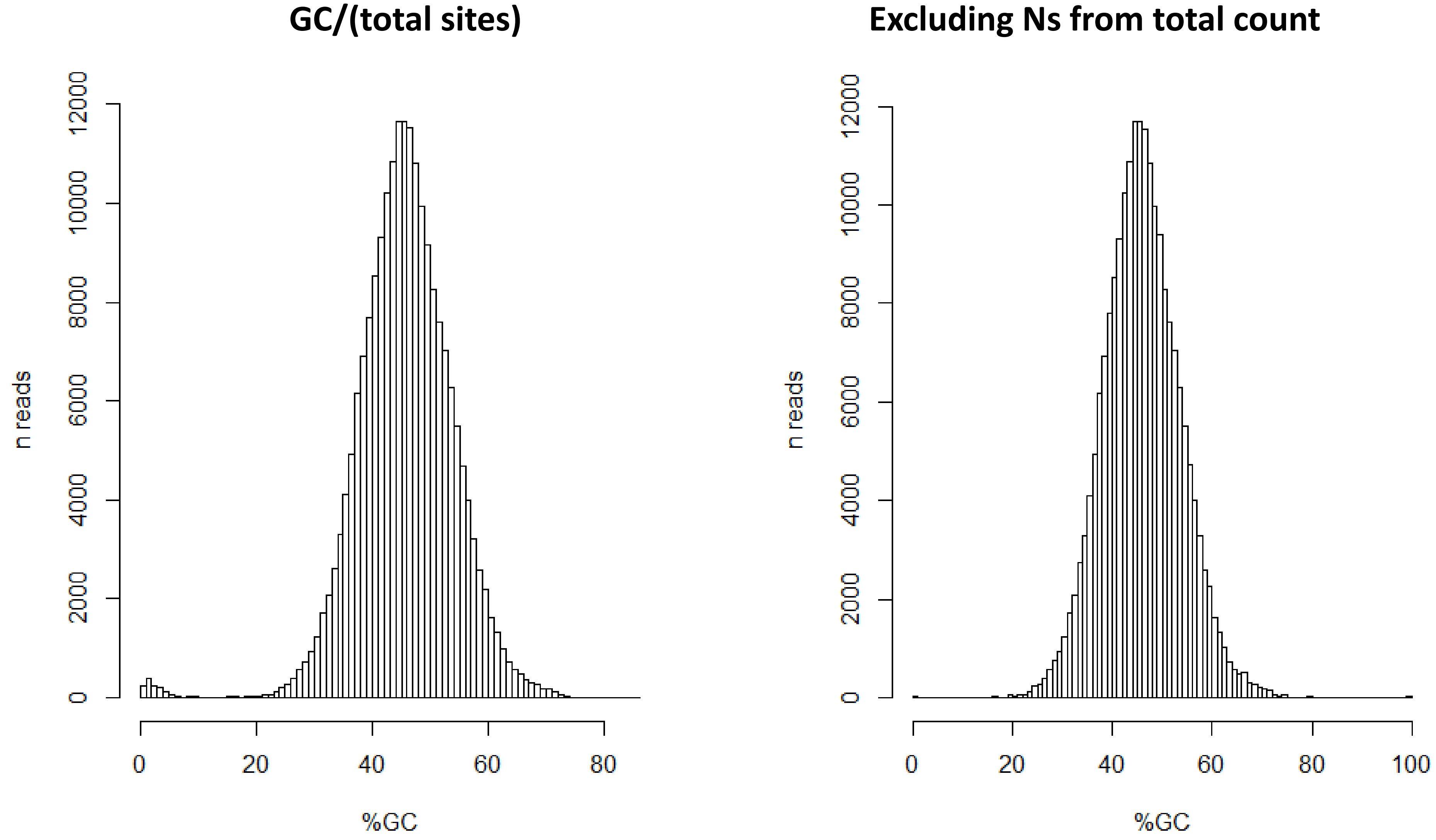
GC content at 45% -> higher due to GC-rich cut site

RAD2



PhiX reads?

GC distribution over all sequences

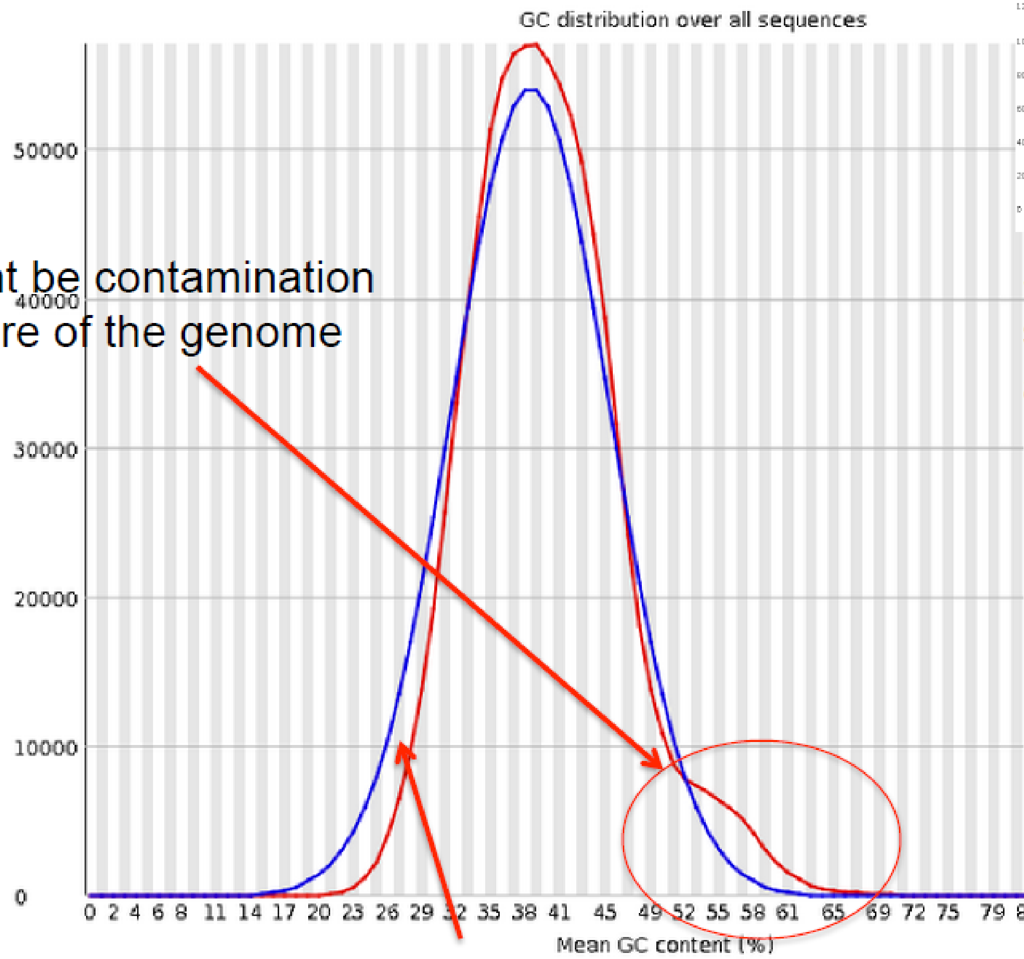


GC/(total sites)

Excluding Ns from total count

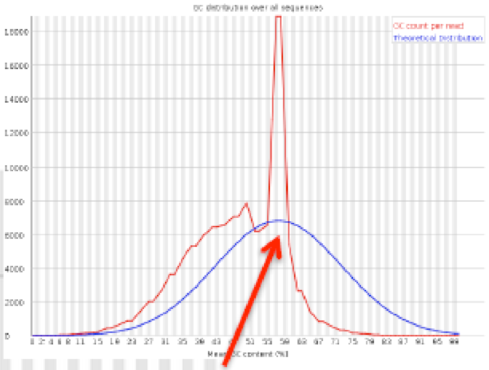
Fastqc: Per sequence GC content

Per sequence GC content

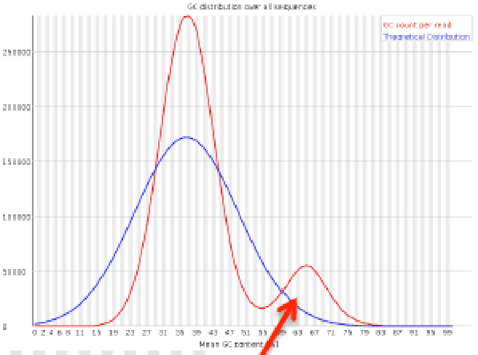


This might be contamination or a feature of the genome

Expected: Normal/Gaussian Distribution

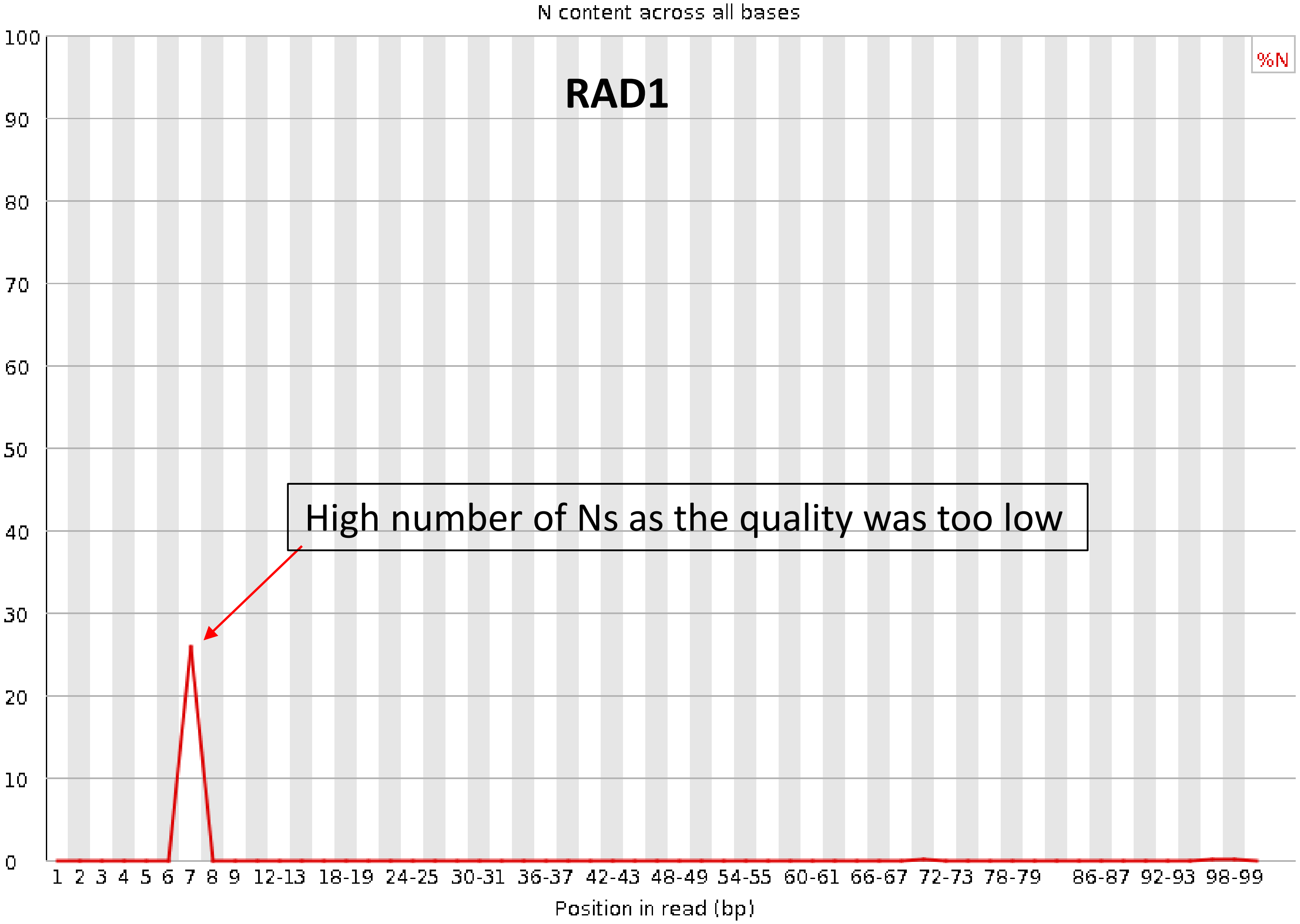


Sharp peak indicates specific motif. Adapters are the usual suspect.



Wider or multiple distributions suggest contamination.

Per base N content

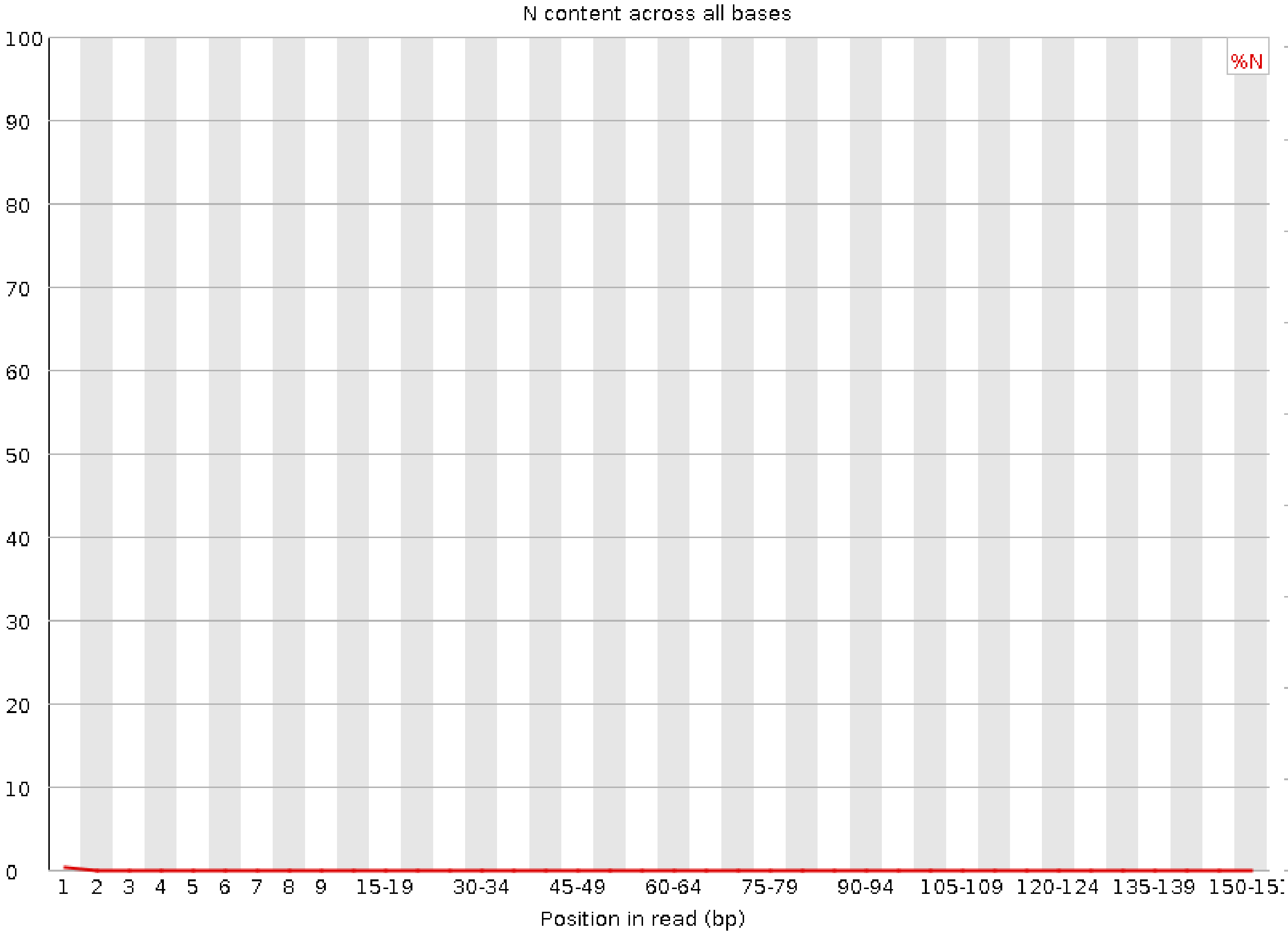


RAD1

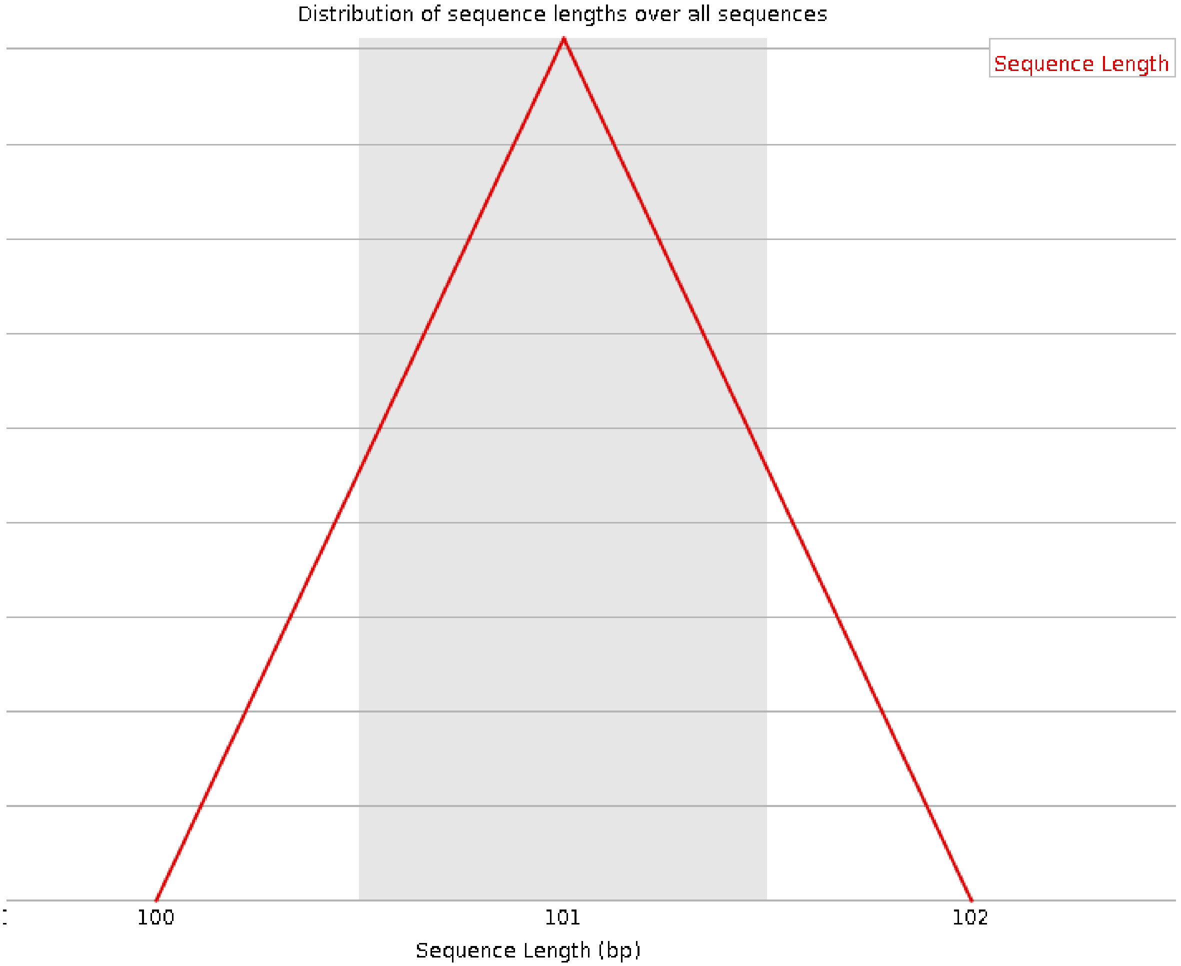
High number of Ns as the quality was too low

Sequence Length Distribution

wgs

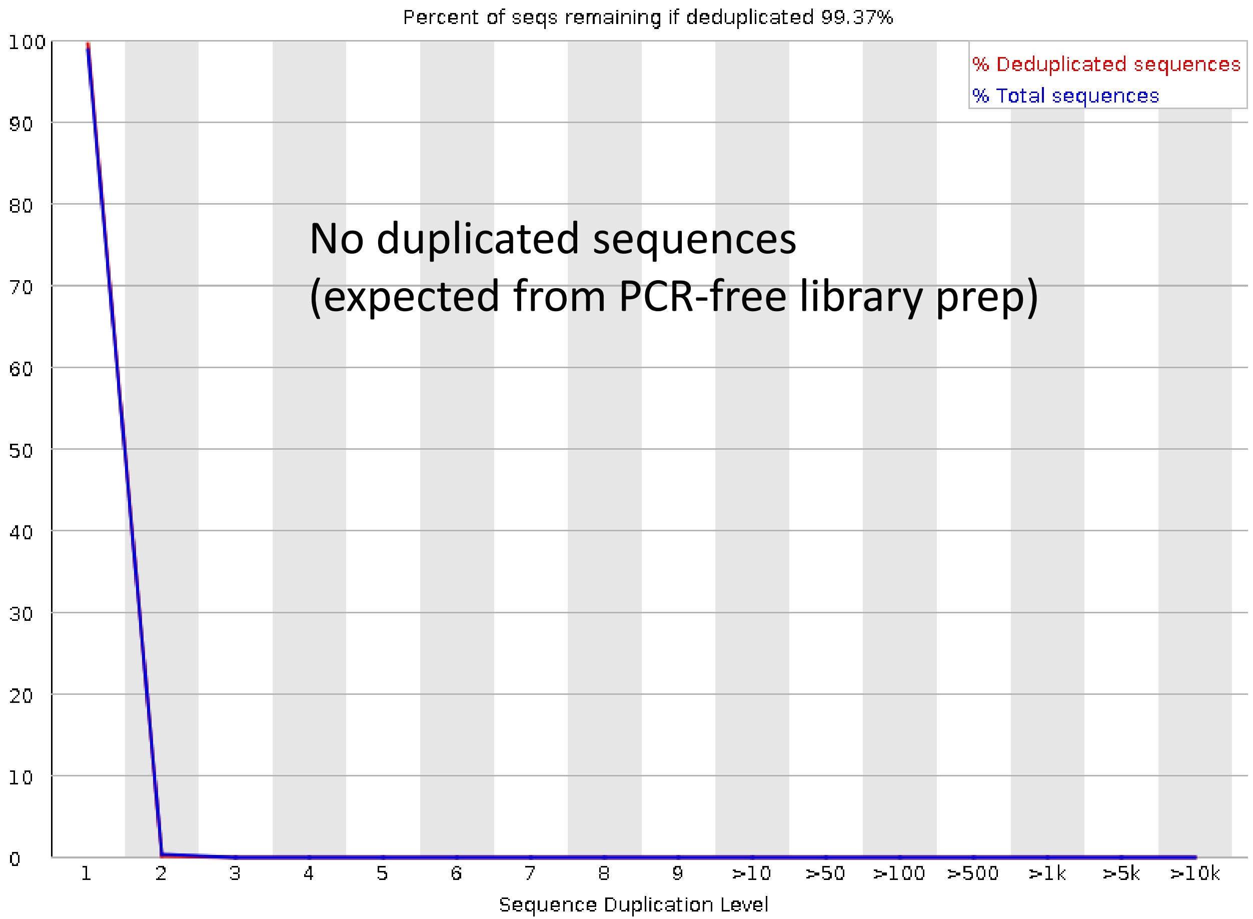


RAD



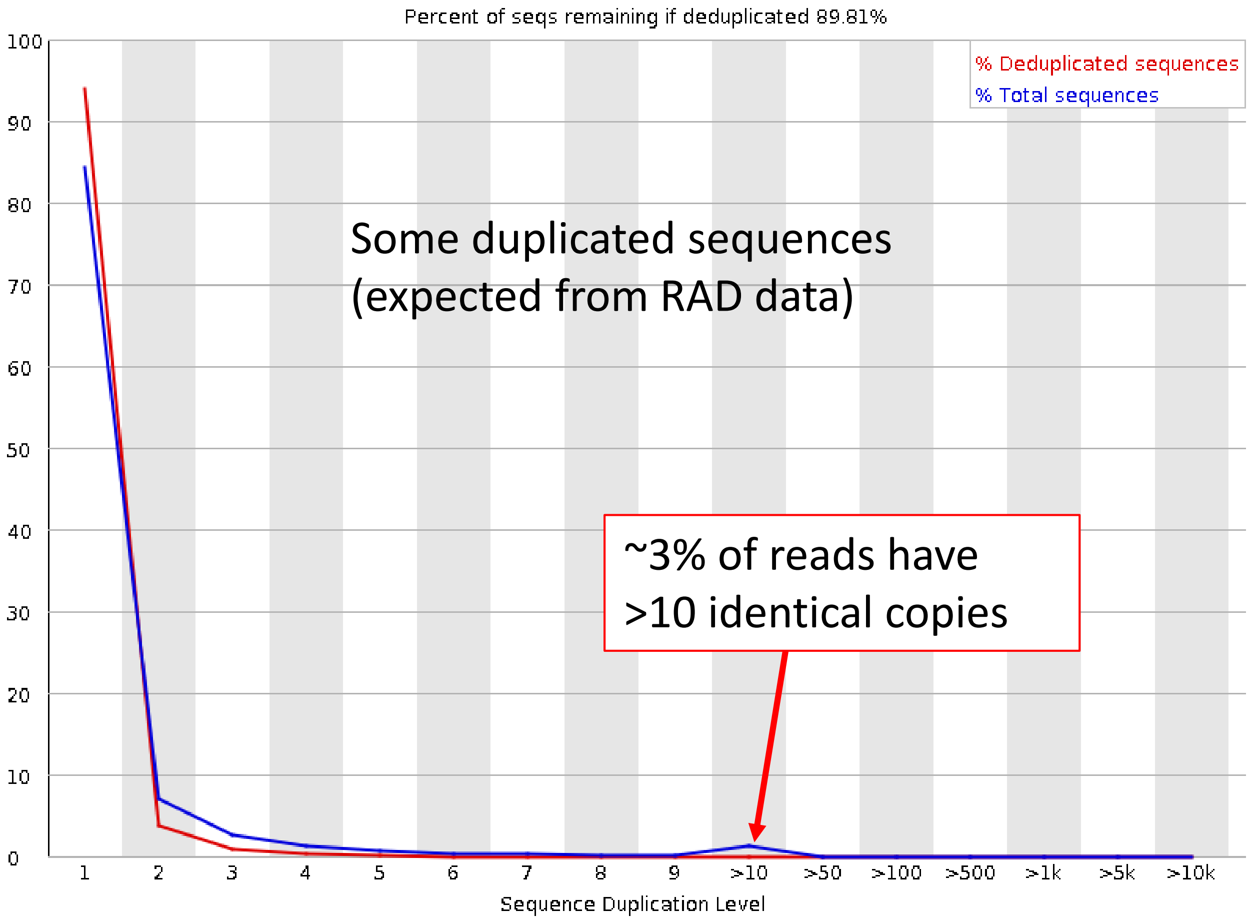
Sequence duplication level

wgs



No duplicated sequences
(expected from PCR-free library prep)

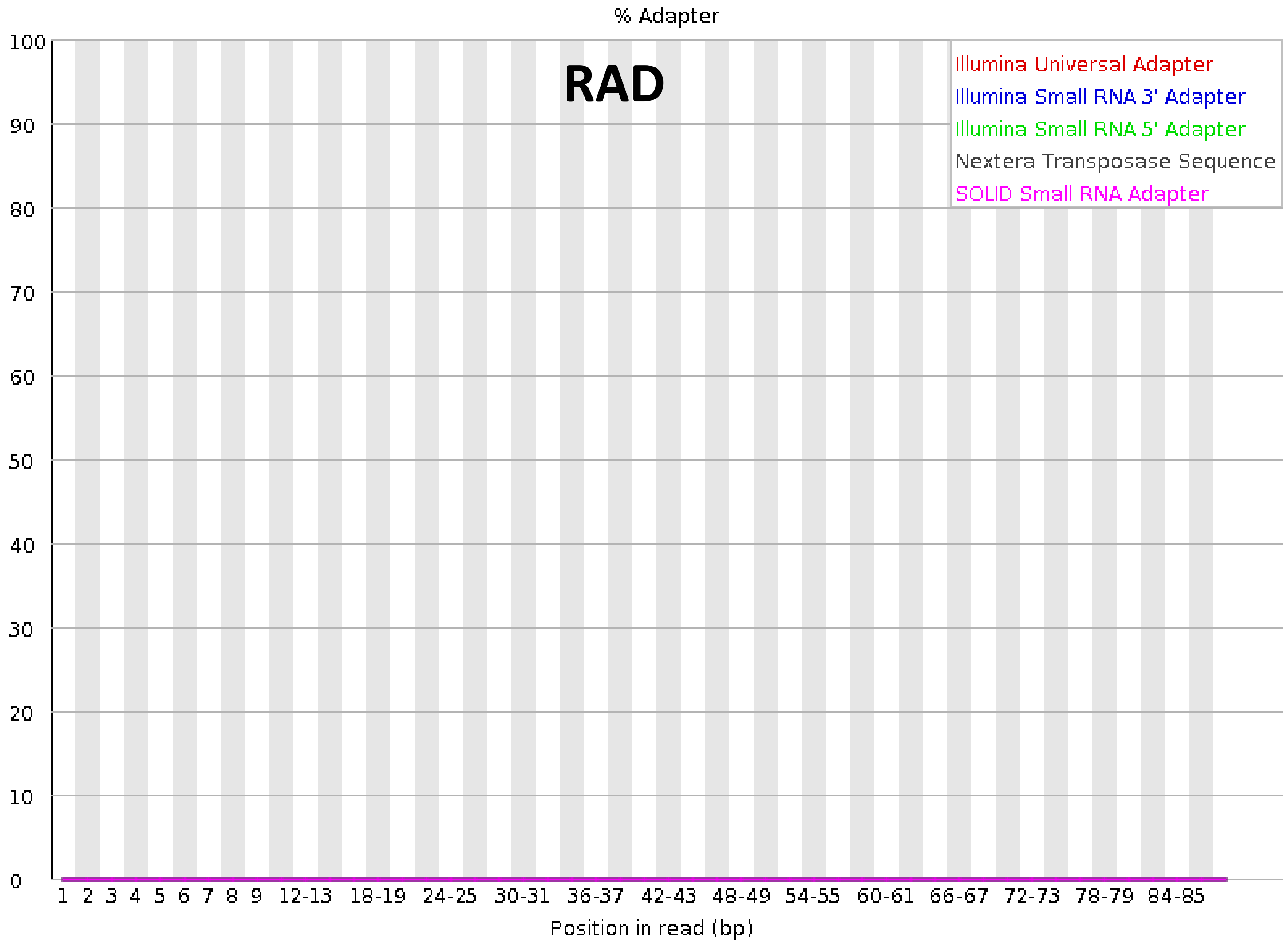
RAD



Some duplicated sequences
(expected from RAD data)

~3% of reads have
>10 identical copies

Adapter content

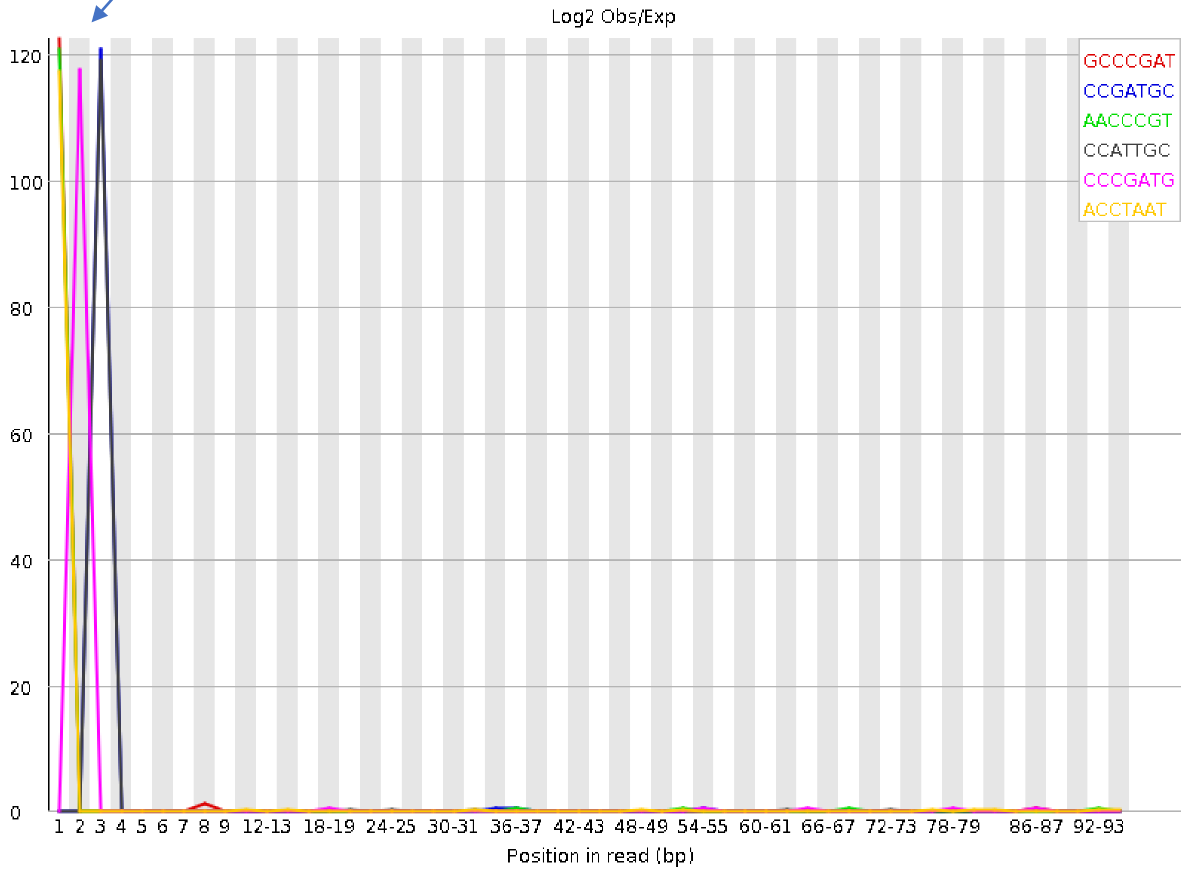


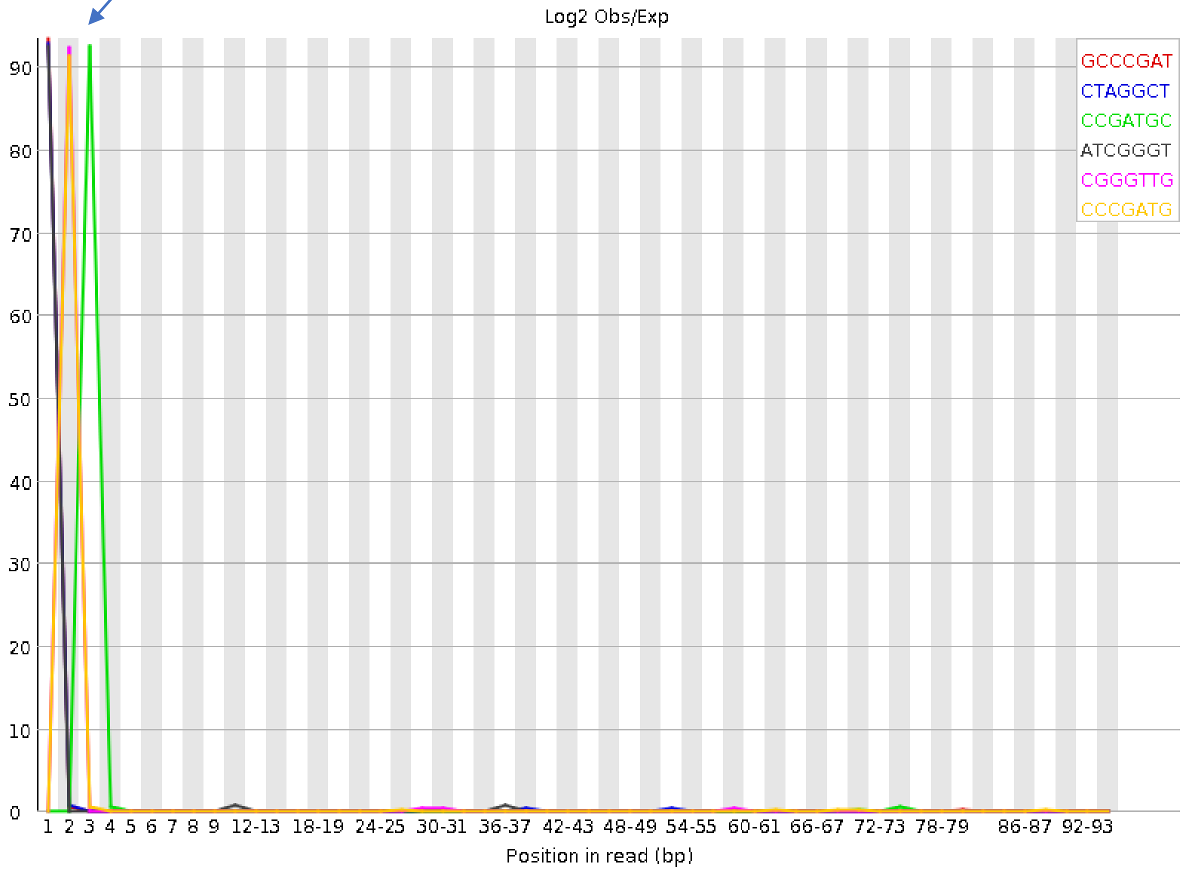
In wgs datasets not even shown

Kmer content

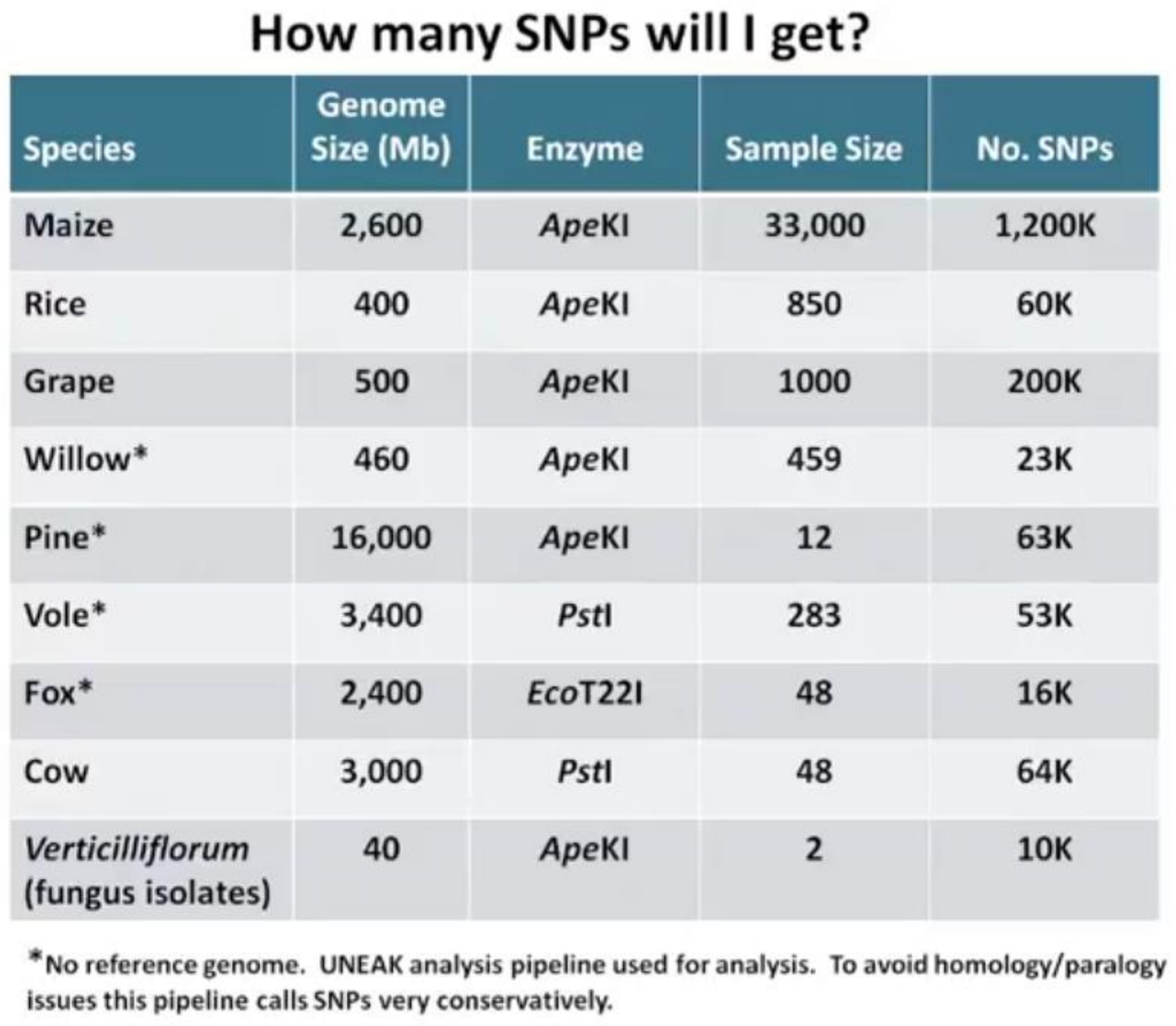
Most common barcodes + cut site

Most common barcodes + cut site





GBS



How many SNPs will I get?

Species	Genome Size (Mb)	Enzyme	Sample Size	No. SNPs
Maize	2,600	*Ape*KI	33,000	1,200K
Rice	400	*Ape*KI	850	60K
Grape	500	*Ape*KI	1000	200K
Willow*	460	*Ape*KI	459	23K
Pine*	16,000	*Ape*KI	12	63K
Vole*	3,400	*Pst*I	283	53K
Fox*	2,400	*Eco*T22I	48	16K
Cow	3,000	*Pst*I	48	64K
Verticilliflorum (fungus isolates)	40	*Ape*KI	2	10K

*No reference genome. UNEAK analysis pipeline used for analysis. To avoid homology/paralogy issues this pipeline calls SNPs very conservatively.