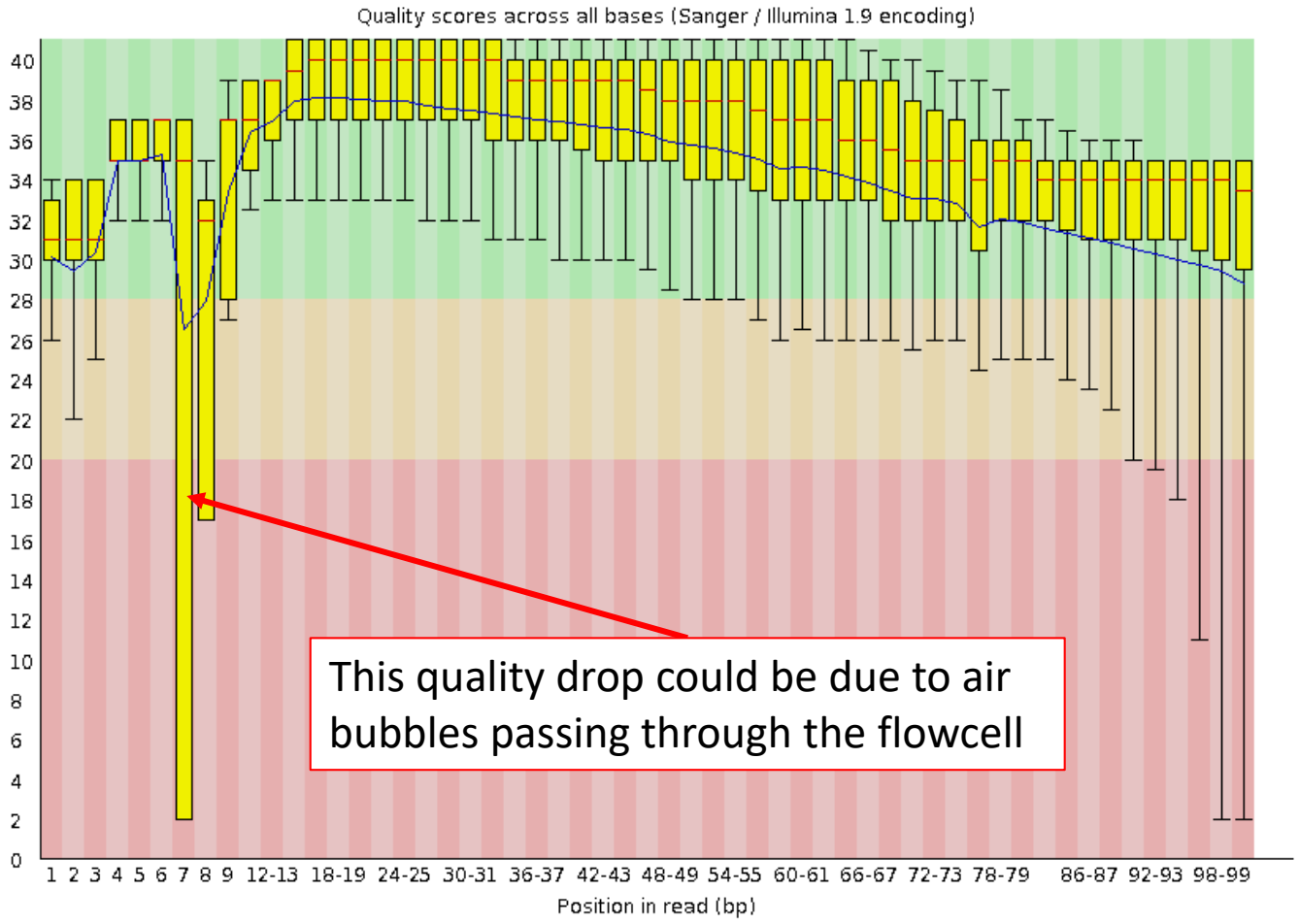
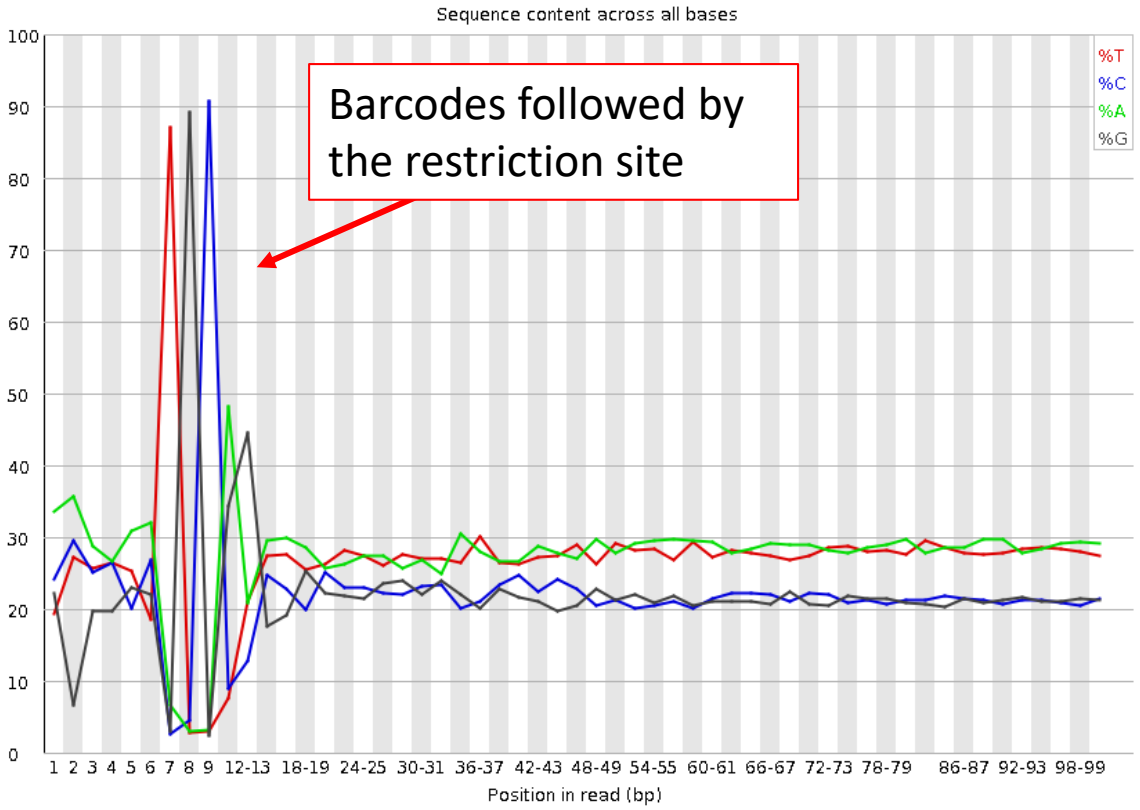
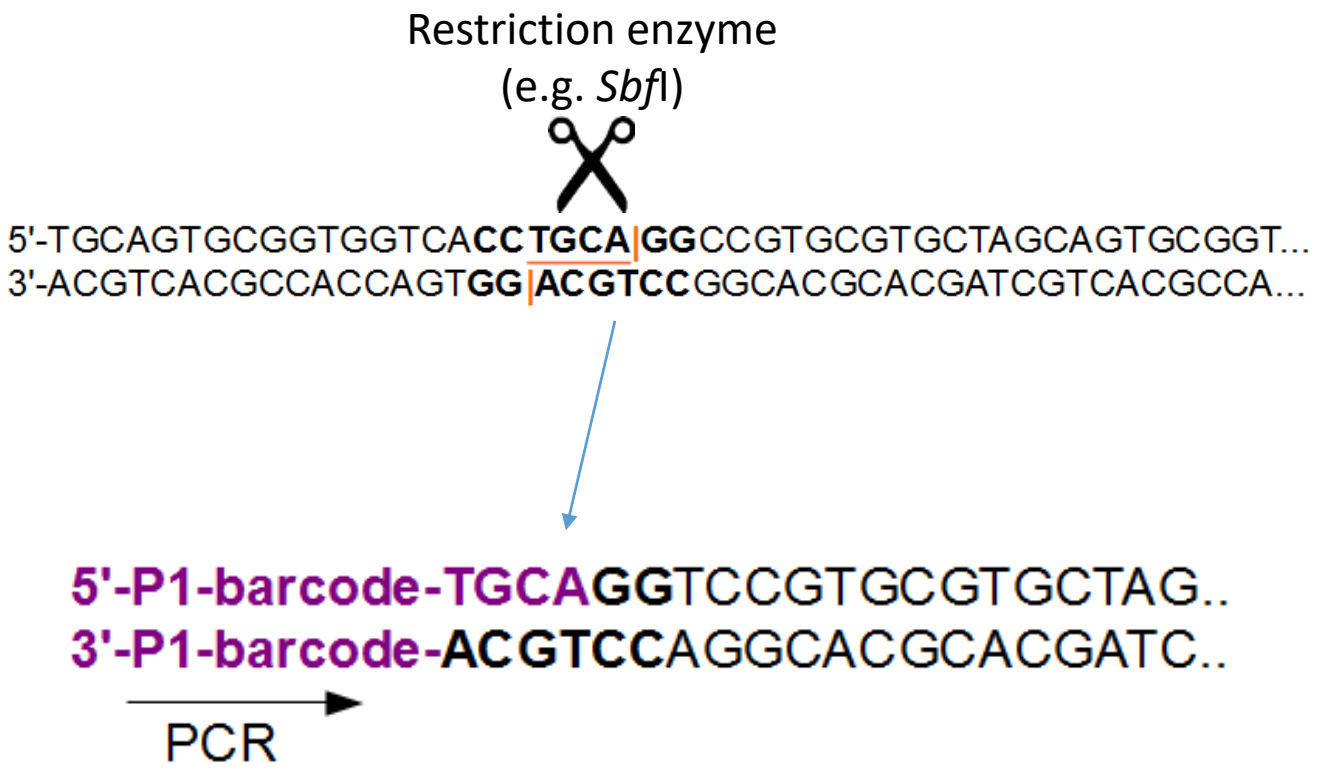
Quality scores across bases: RAD datasets



Nucleotide composition



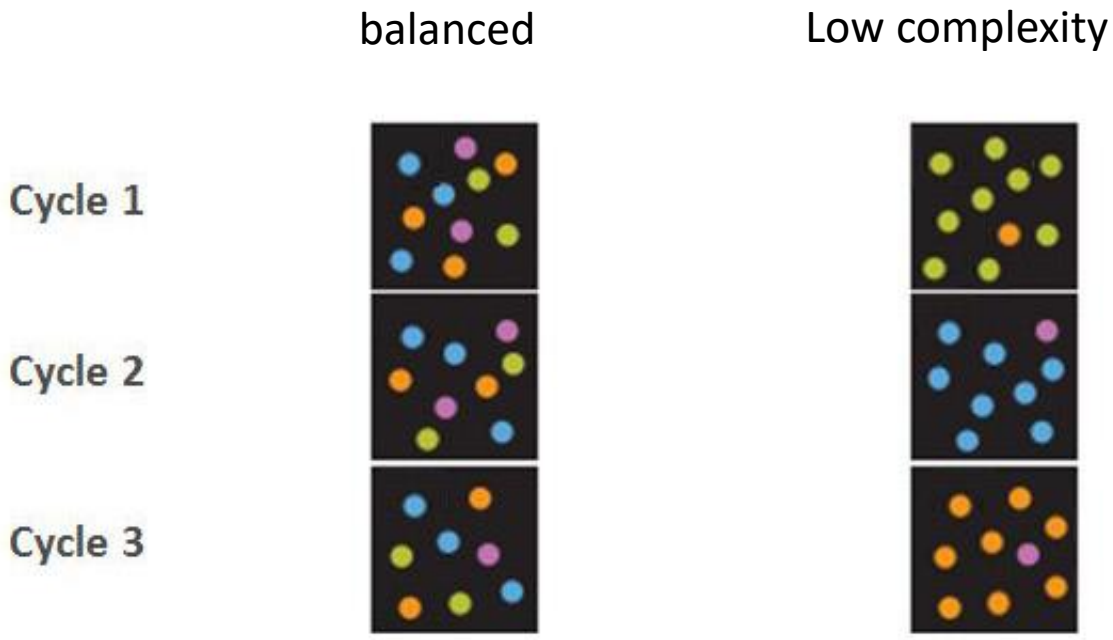
RAD/GBS



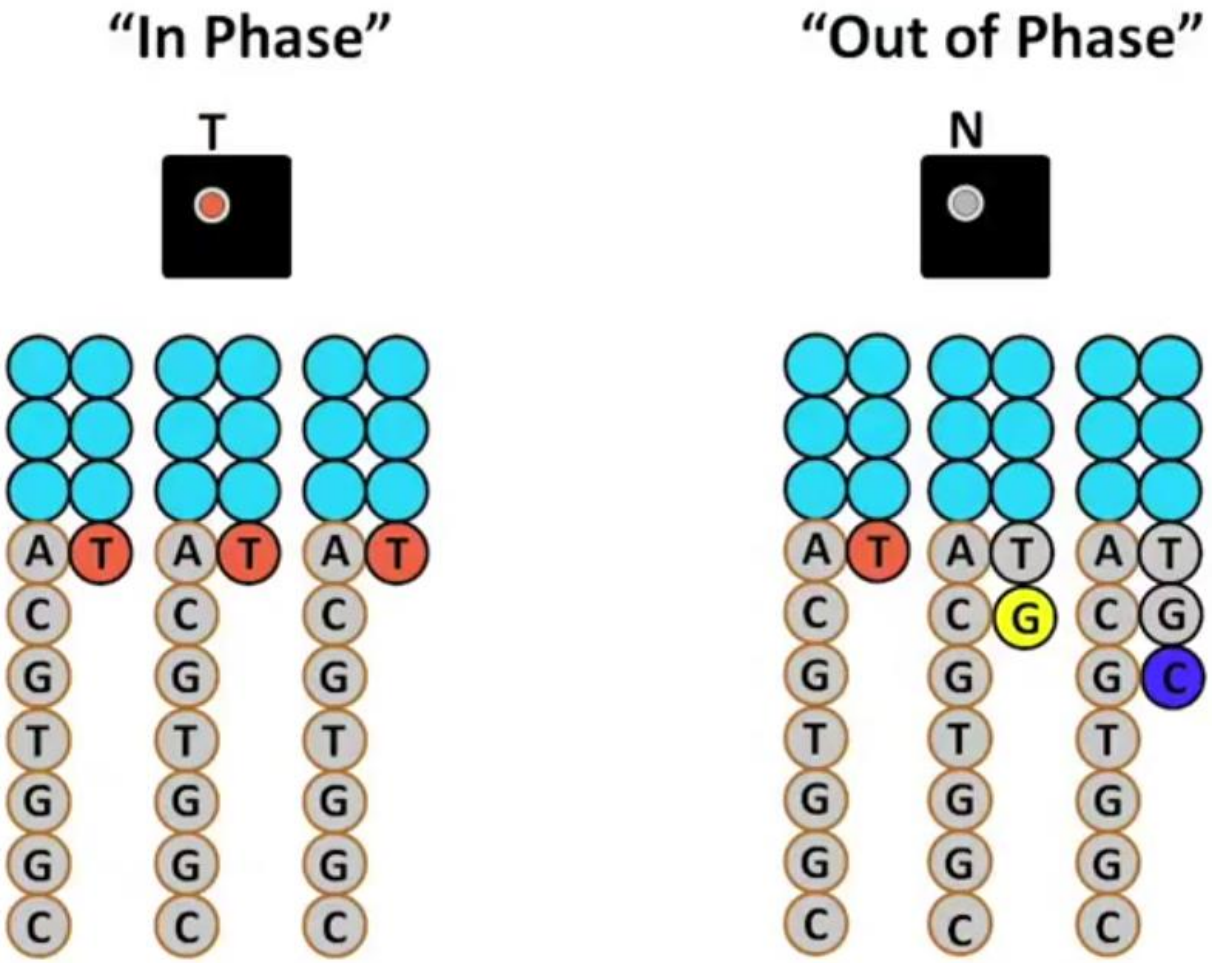
Each read starts with the barcode, then the restriction site, then a variable sequence

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



Phasing issues



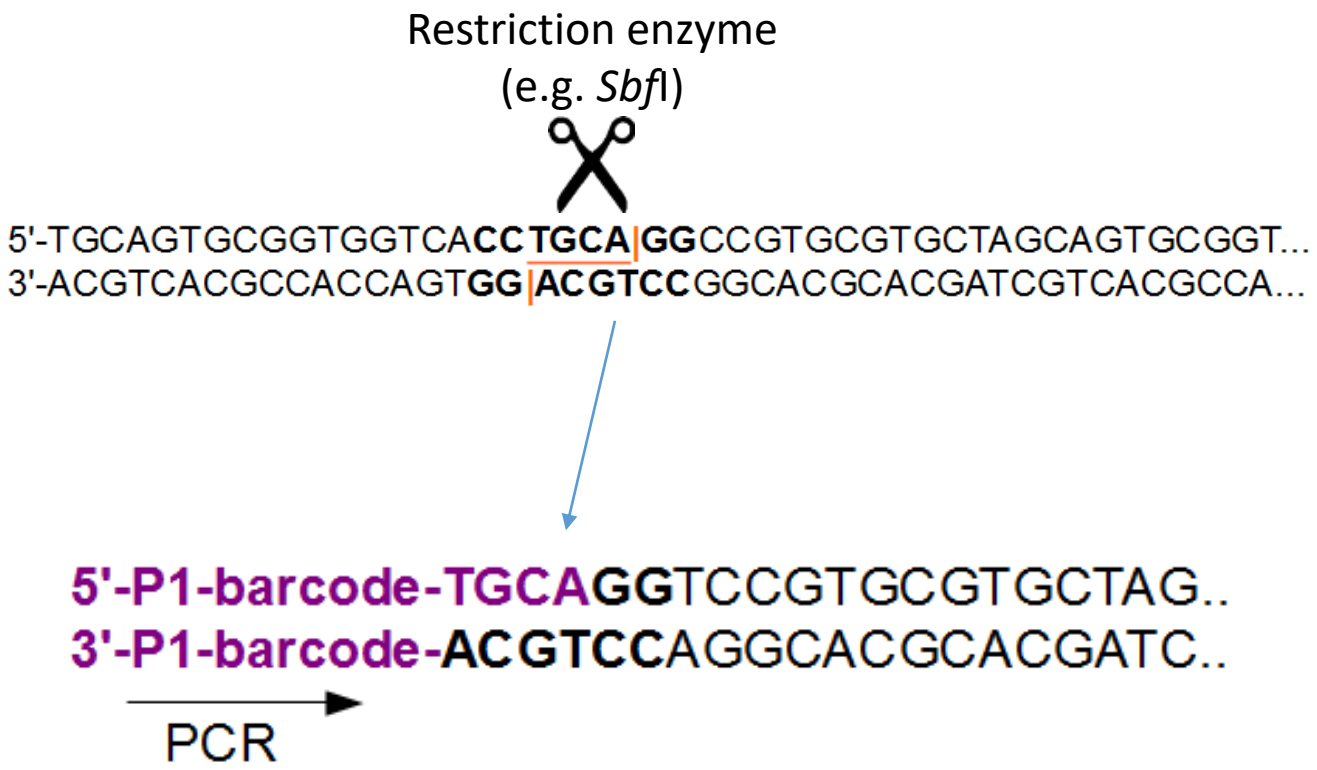
The first 12 nucleotides are also used for «phasing», i.e. correcting for reads that are out of phase. The algorithm expects random nucleotide distribution!

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

How to minimize the problem

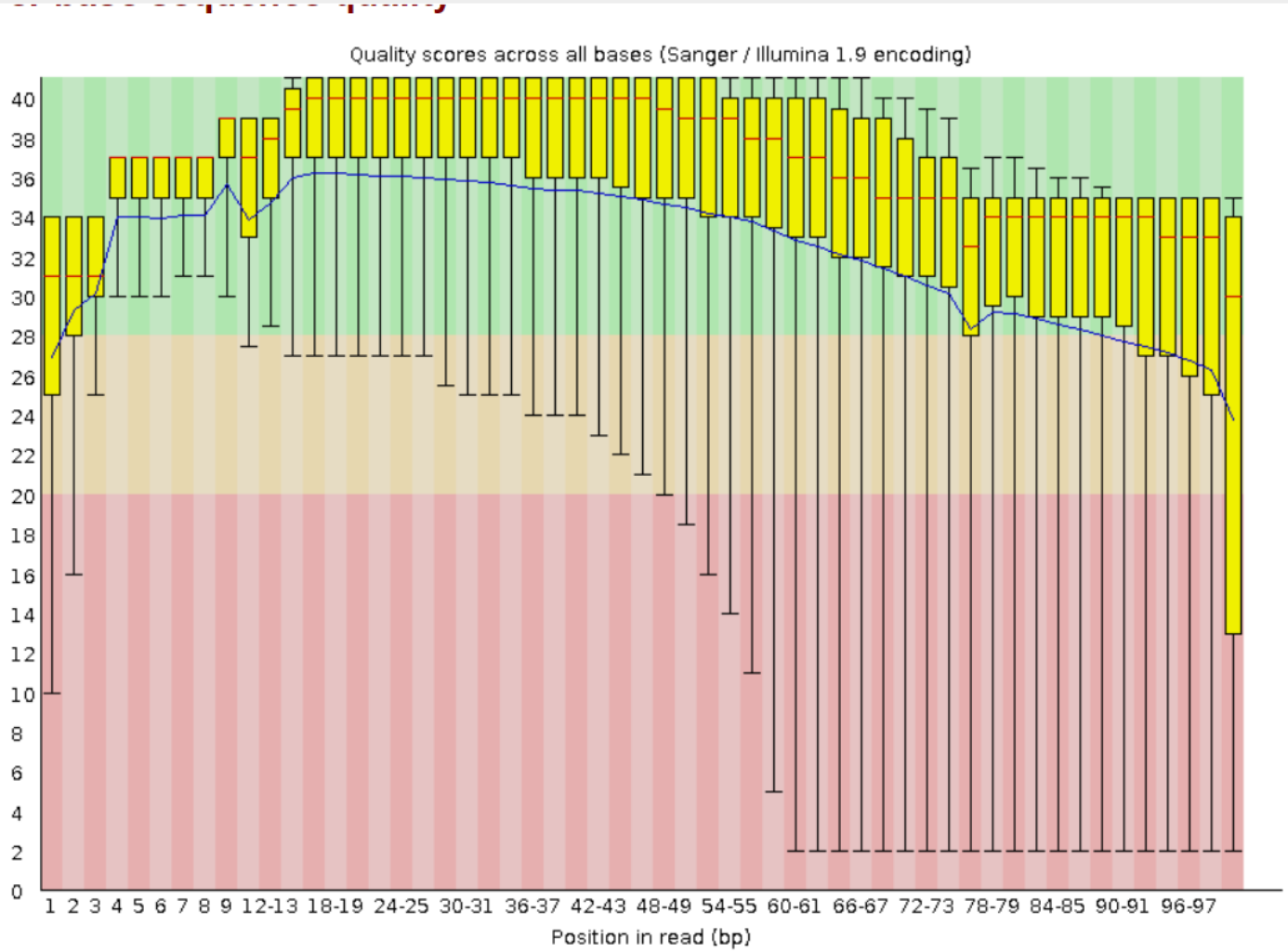
- Use barcodes of different lengths to shift the restriction enzyme cut site
- Add PhiX virus DNA to the RAD libraries to increase the complexity of reads (‘spiking’)
- Reduce loading concentrations of Illumina plates
- Potentially: filter out bad reads

RAD/GBS



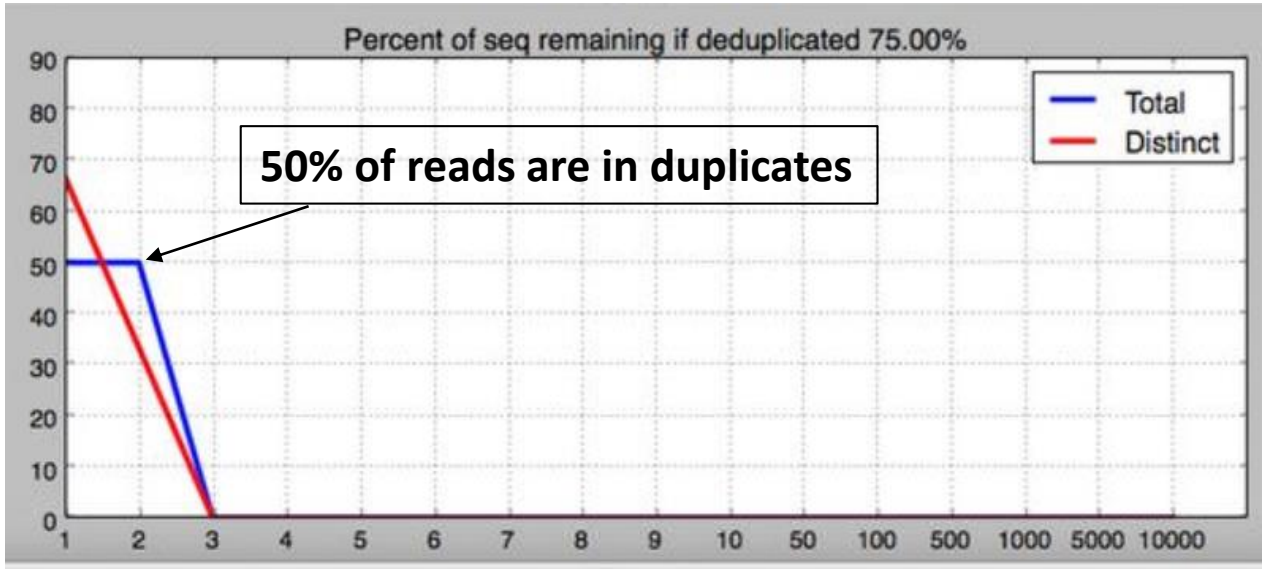
Each read starts with the barcode, then the restriction site, then a variable sequence

RAD 2: with barcodes of different lengths



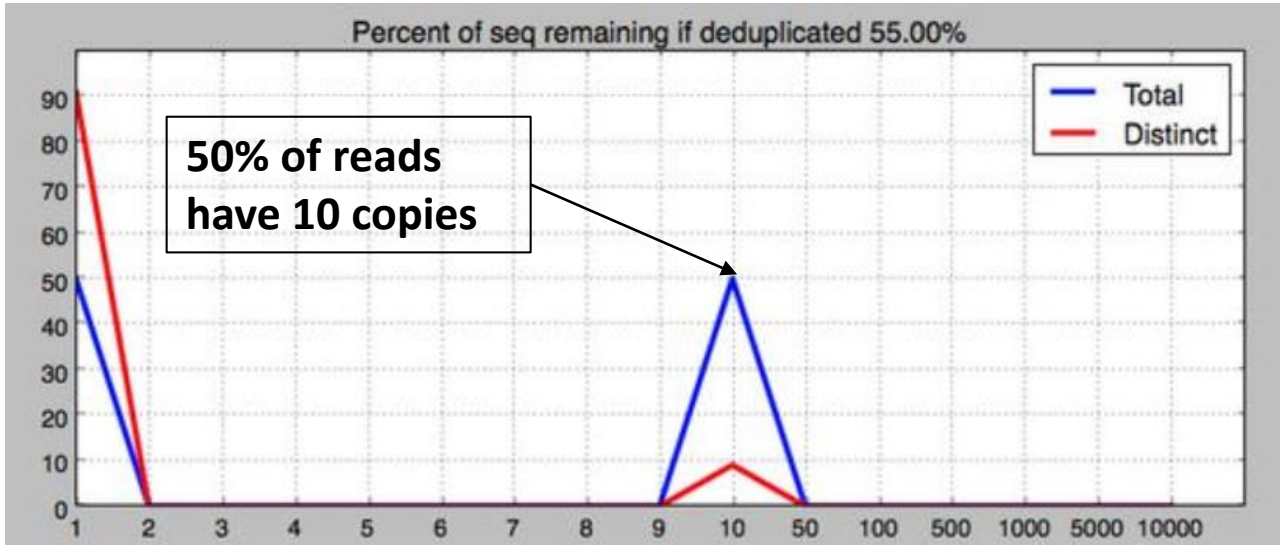
Sequence duplication level

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



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).

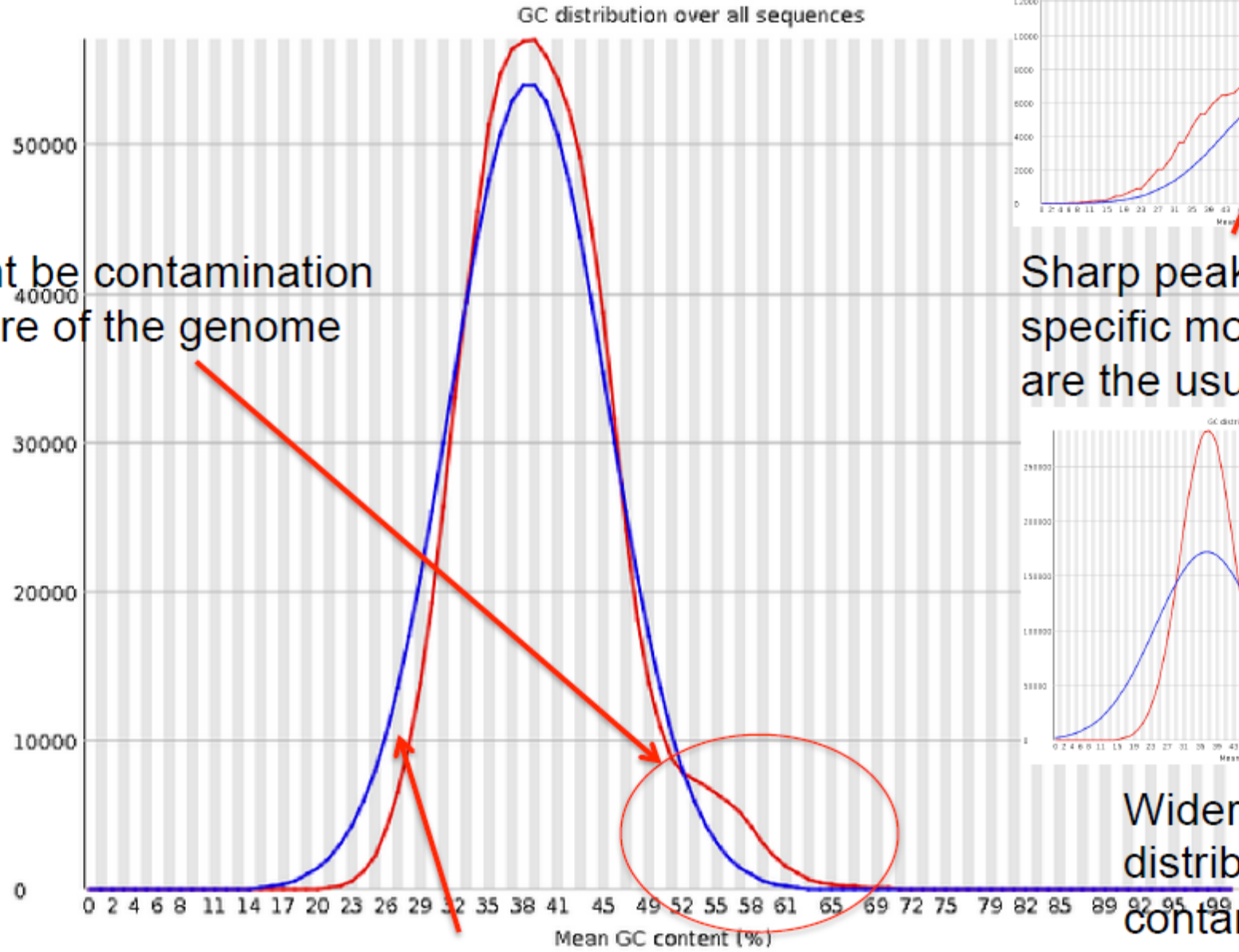
Example 2: 20 reads total
10 unique sequences + 1 sequence present 10x



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.

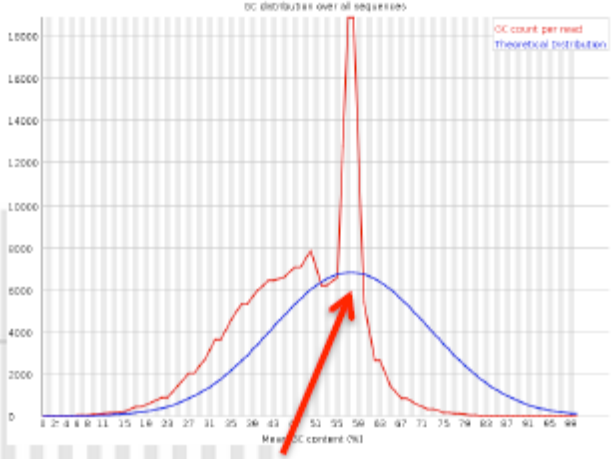
Examples and plots from https://www.biostars.org/p/107402/

Per sequence GC content

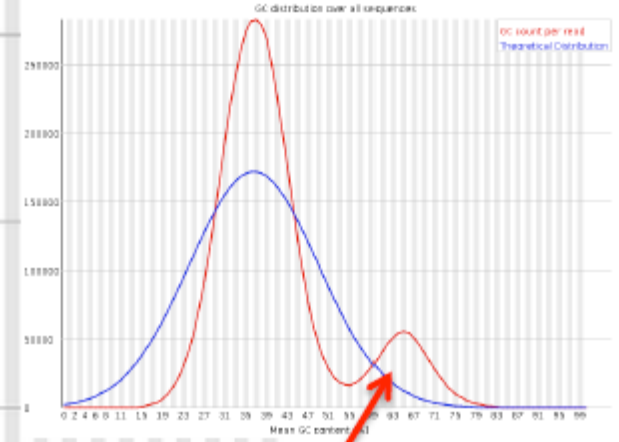


This might be contamination or a feature of the genome

Expected: Normal/Gaussian Distribution

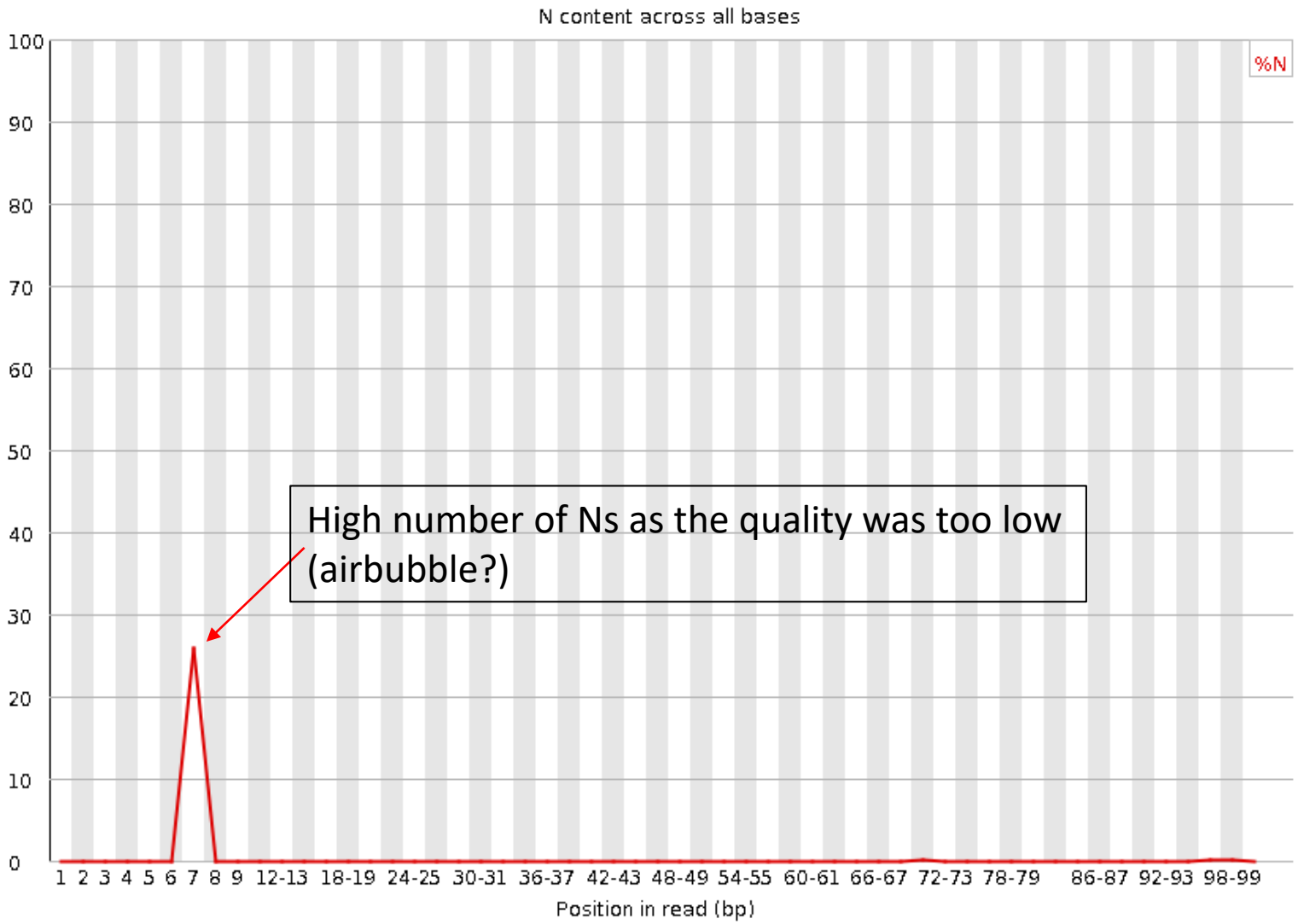


Sharp peak indicates specific motif. Adapters are the usual suspect.

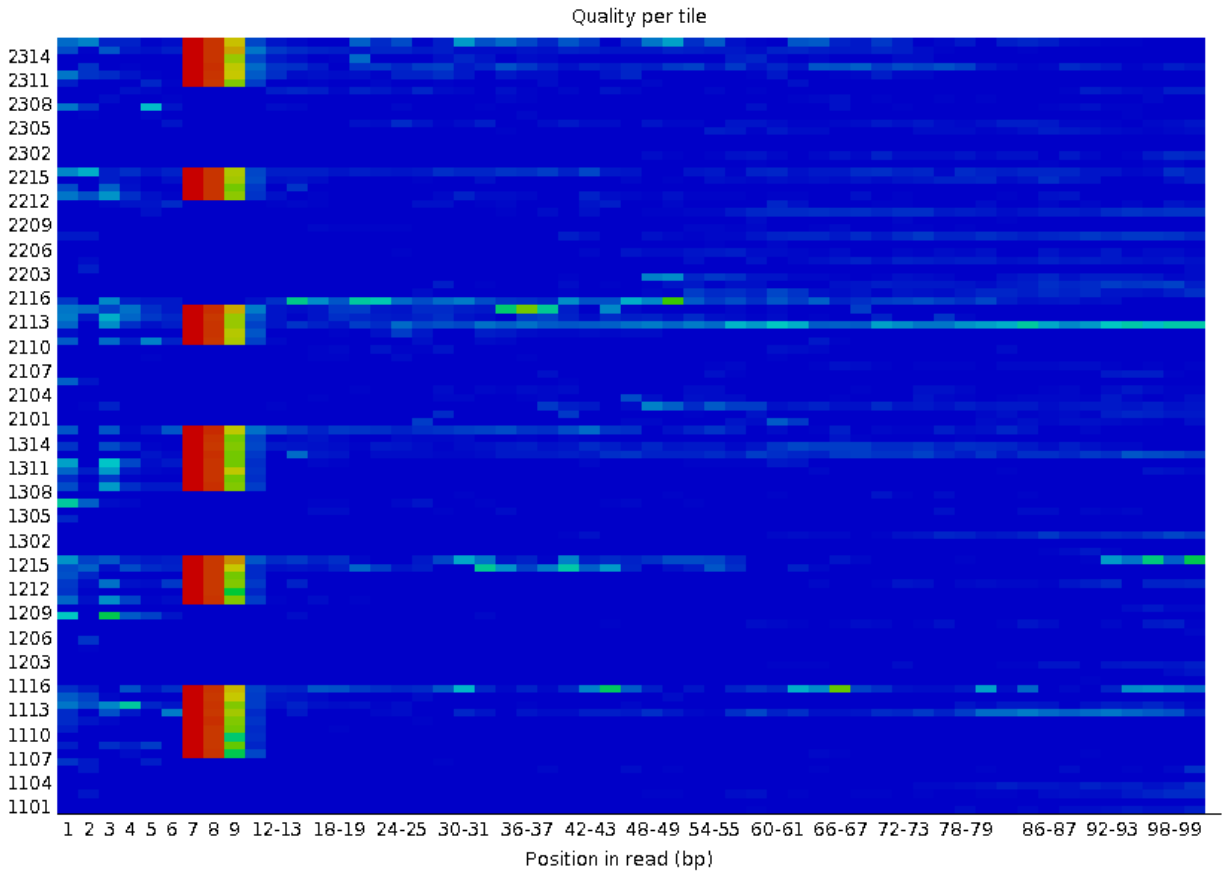


Wider or multiple distributions suggest contamination.

Per base N content

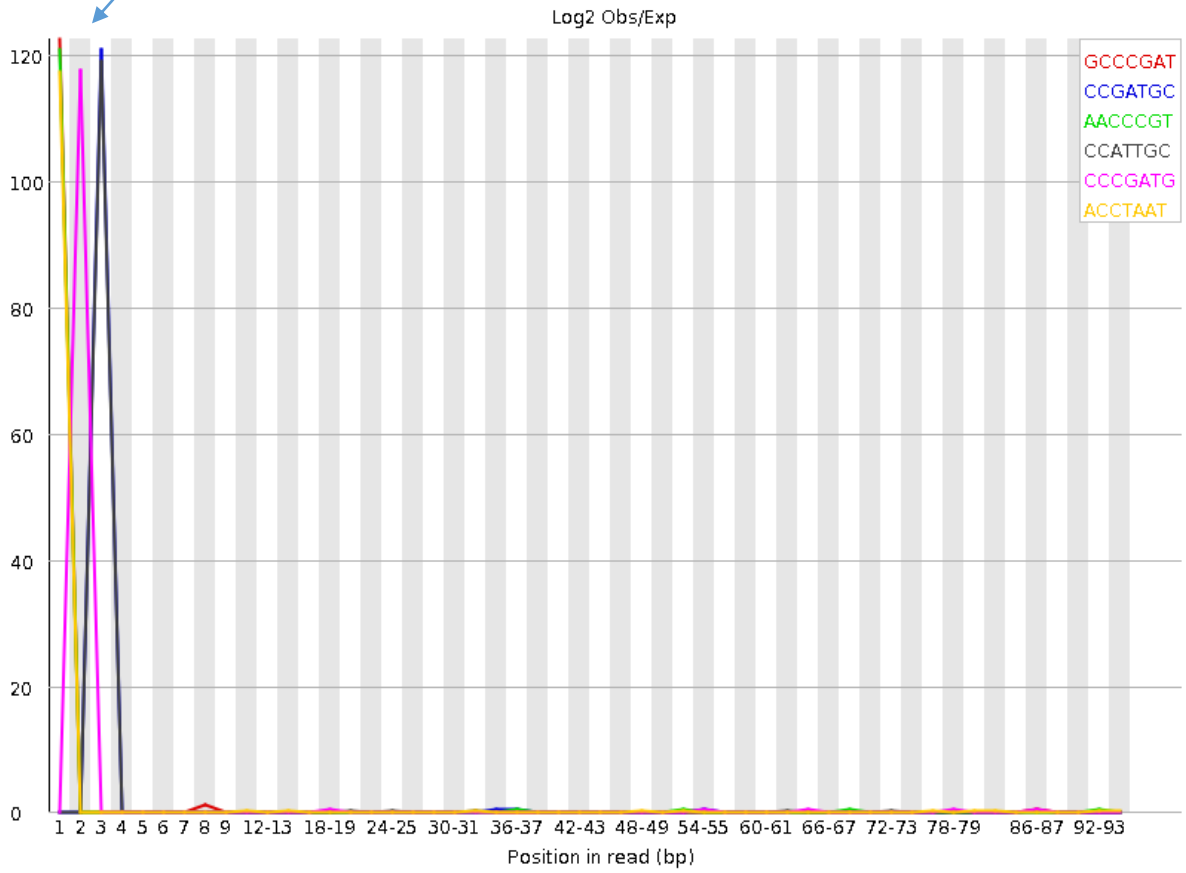


Per tile sequencing quality

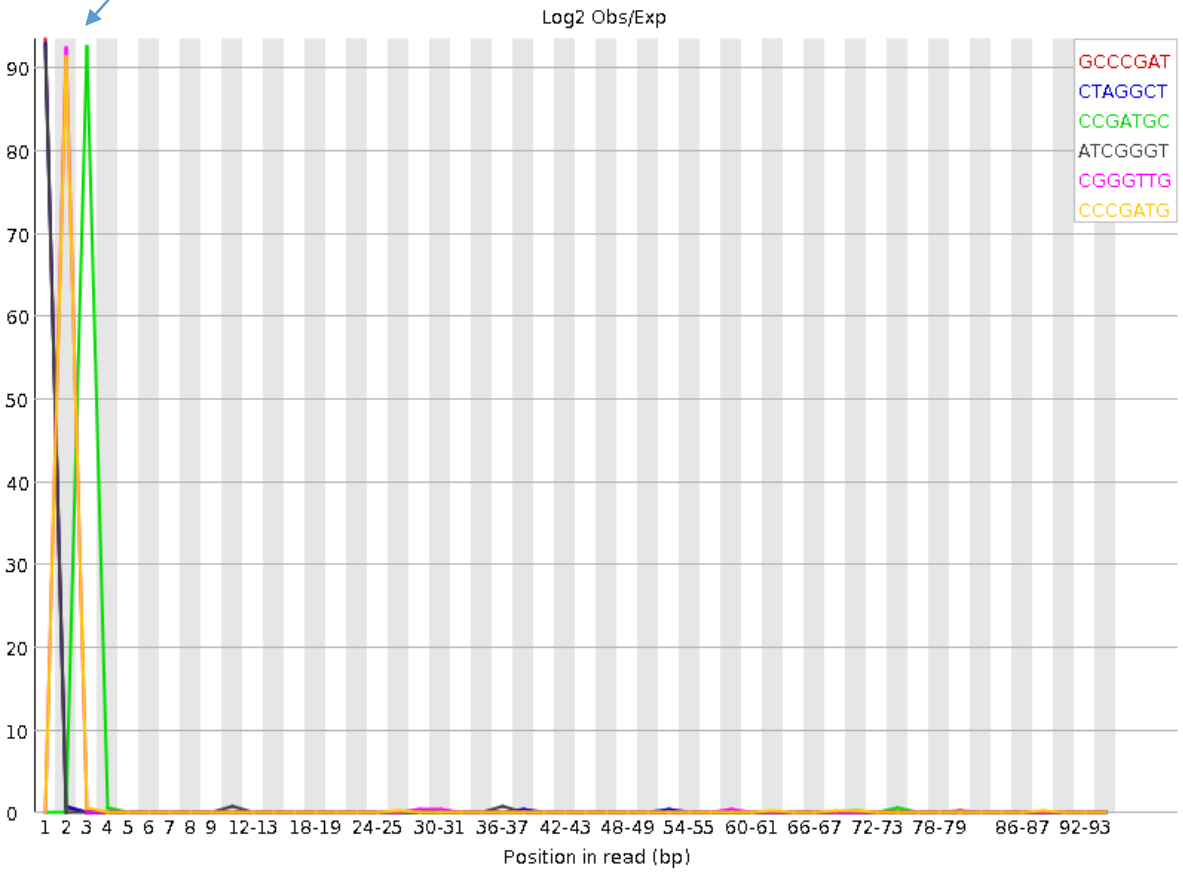


Kmer content in RAD data:

Most common barcodes + cut site



Most common barcodes + cut site

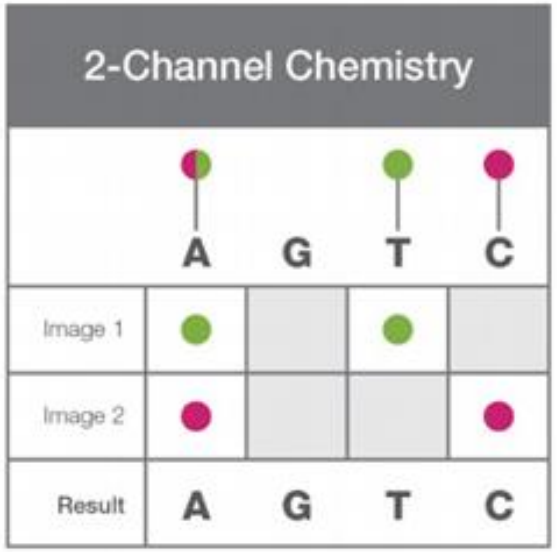


wgs.Novaseq.R2

Overrepresented sequences

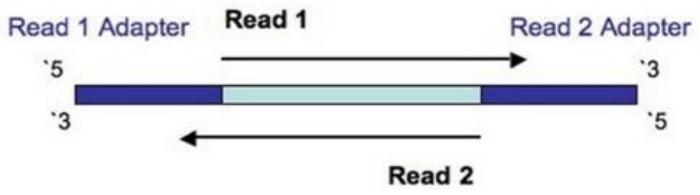
Sequence	Count	Percentage	Possible Source
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	28471	0.17023848994734272	No Hit

polyG tail due to 2-colour SBS technology



Trimming and filtering reads

- Remove adapter sequences



If the DNA fragment is very short, it will sequence into the second adapter. Therefore, there will be adapter sequences in the read which need to be removed or the read will not map.

- Remove polyG tails (if 2-base SBS technology used)
- Trim (cut off) ends of reads with low sequencing quality

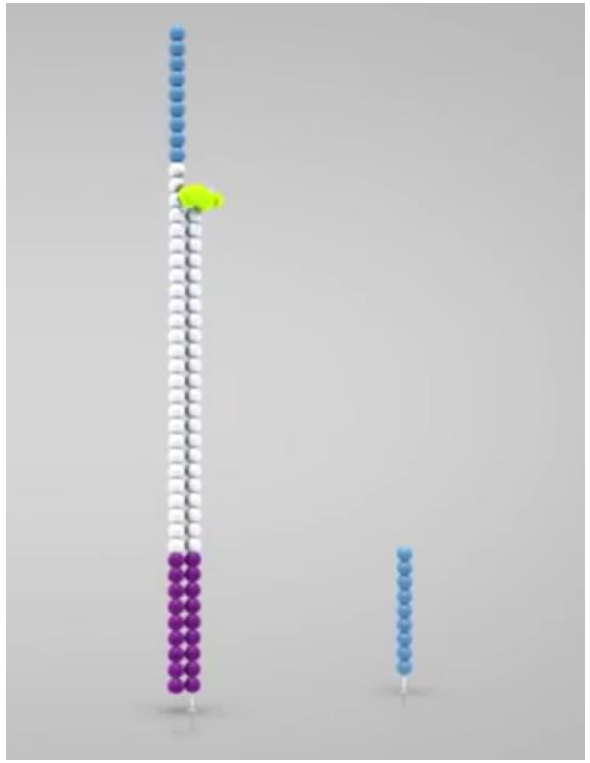
Sequencing by synthesis by Illumina: Read1

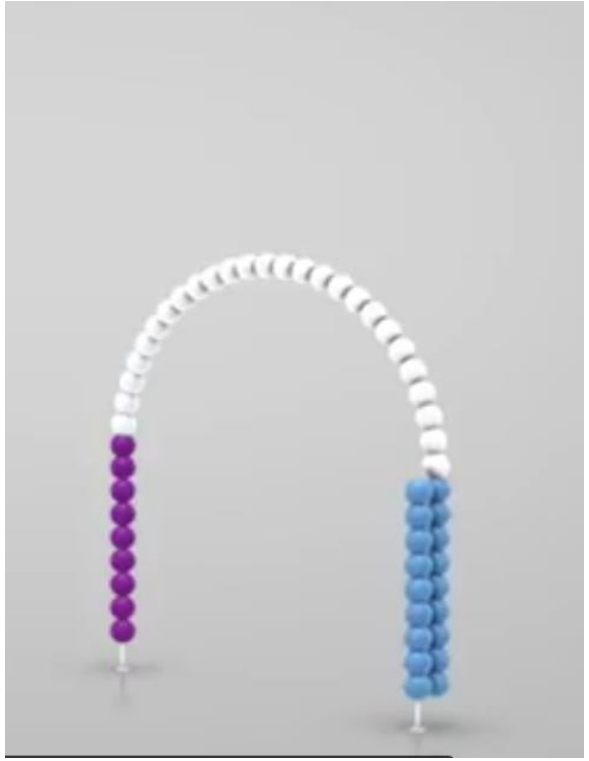
1. DNA fragments bind to the P1 primer
2. polymerase makes it double-stranded
3. Template strand is washed away (denaturated)

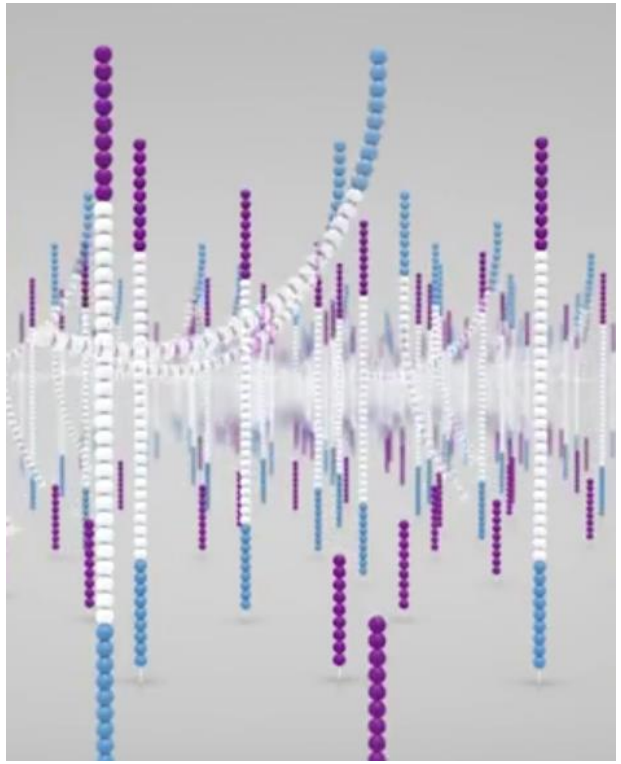
4. DNA strand forms a bridge and binds to the P2 primer
5. Polymerase makes it double-stranded
6. Denaturation -> two single stranded DNA fragments

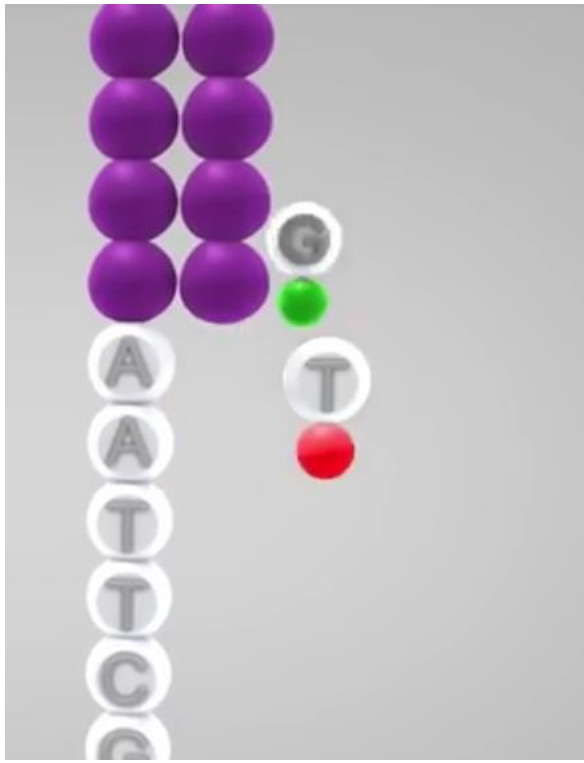
Repeat many times to form clusters of thousands of identical DNA strands
7. The reverse strands are cleaved and washed off

8. Primer annealing
9. Complementary fluorescently tagged nucleotides are incorporated in each cycle
10. Repeat step 9 150x



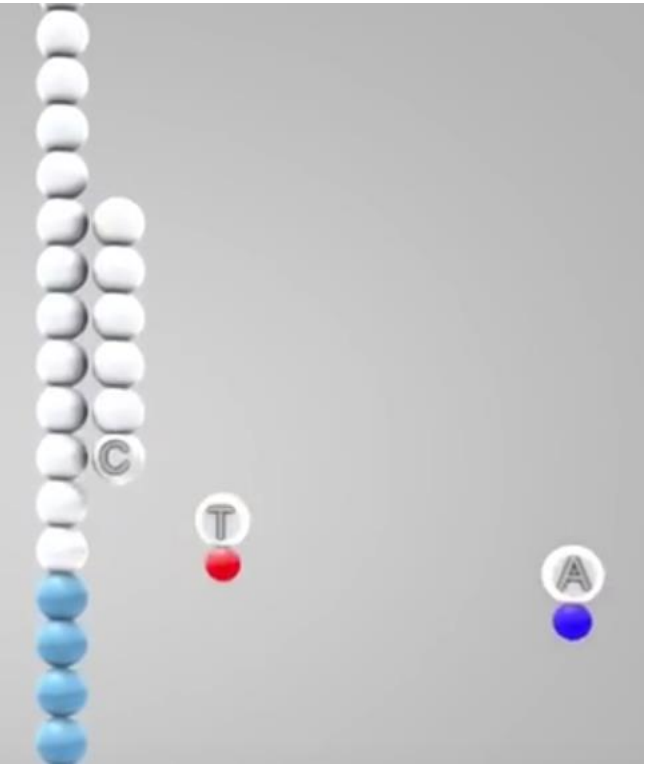






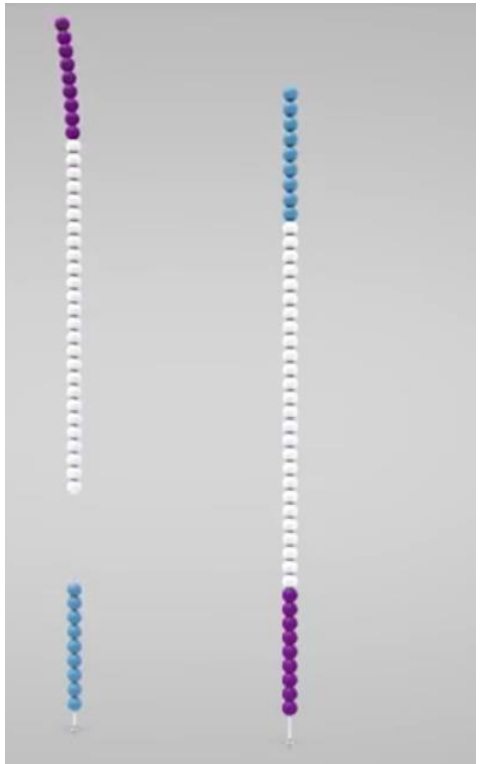
Sequencing by synthesis by Illumina: Read2

11. Denaturation
12. Primer index 1 is added and sequenced
13. The 3’ end is deprotected

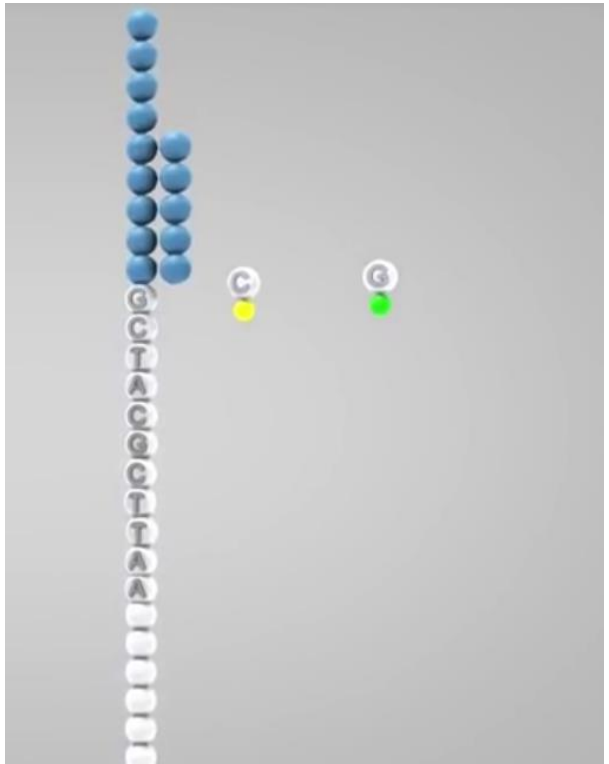


14. The DNA fragment forms a bridge to the reverse primer
15. Sequencing of index 2
16. Polymerase makes bridge double-stranded

14. Denaturation leads to single-stranded fragments bound to the flowcell
15. Forward strands are cleaved and washed off



16. Second read is sequenced as the first read with sequencing-by-synthesis



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.