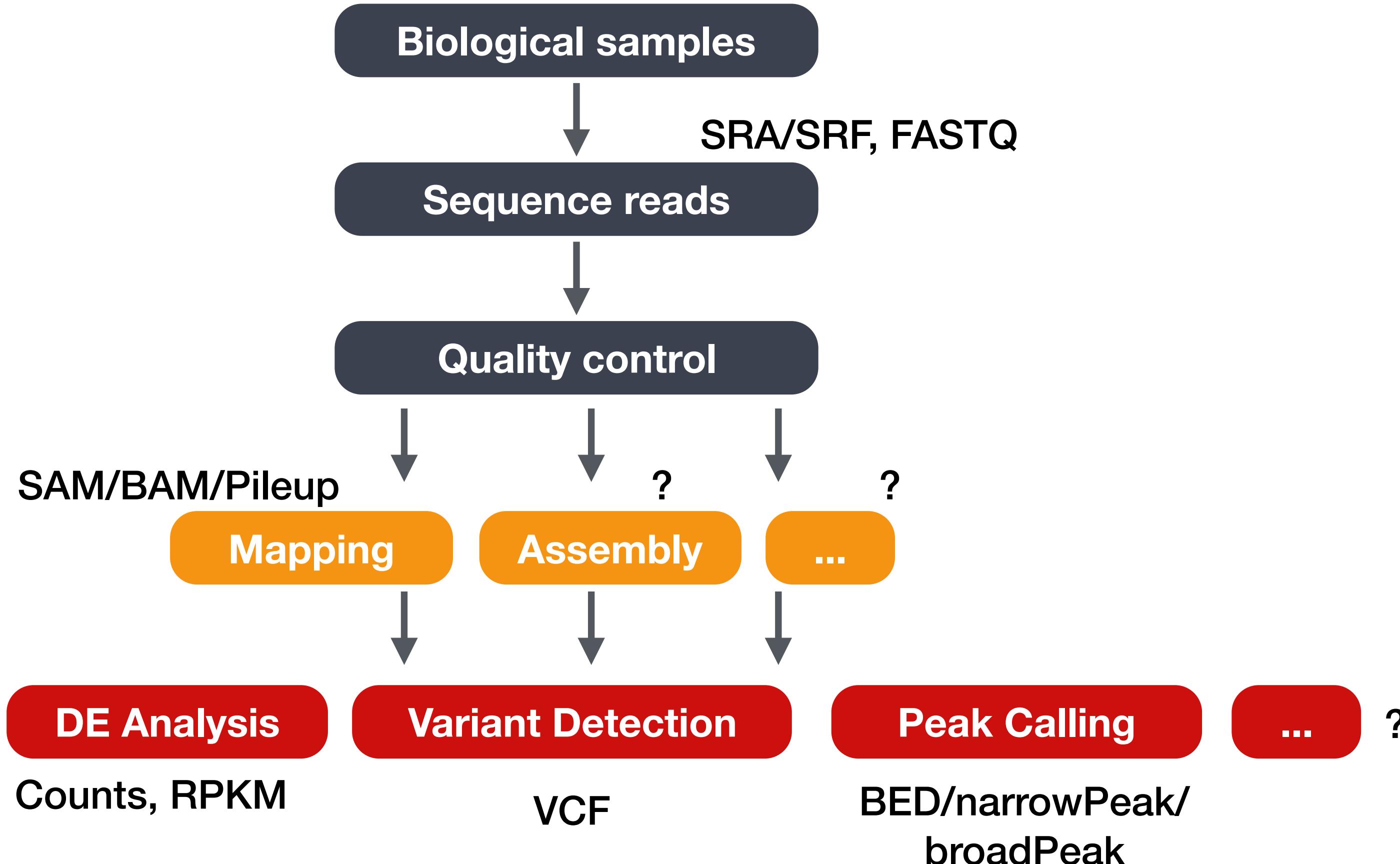


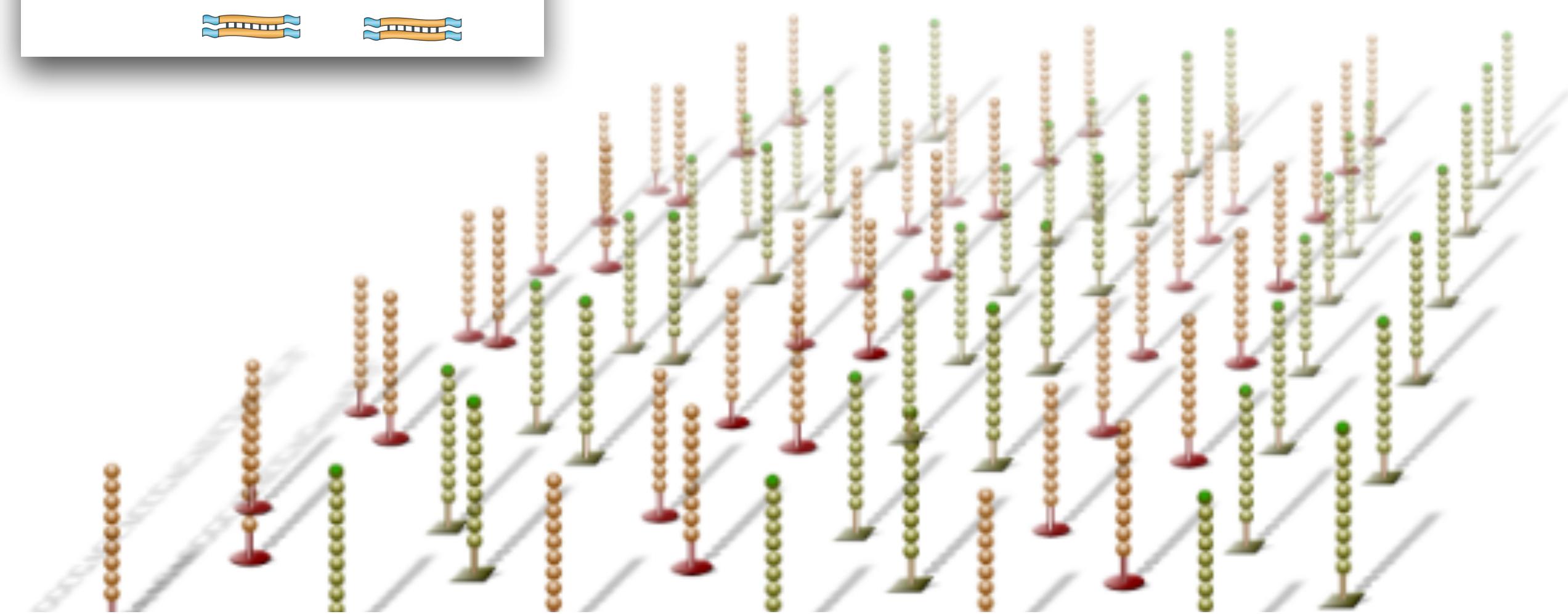
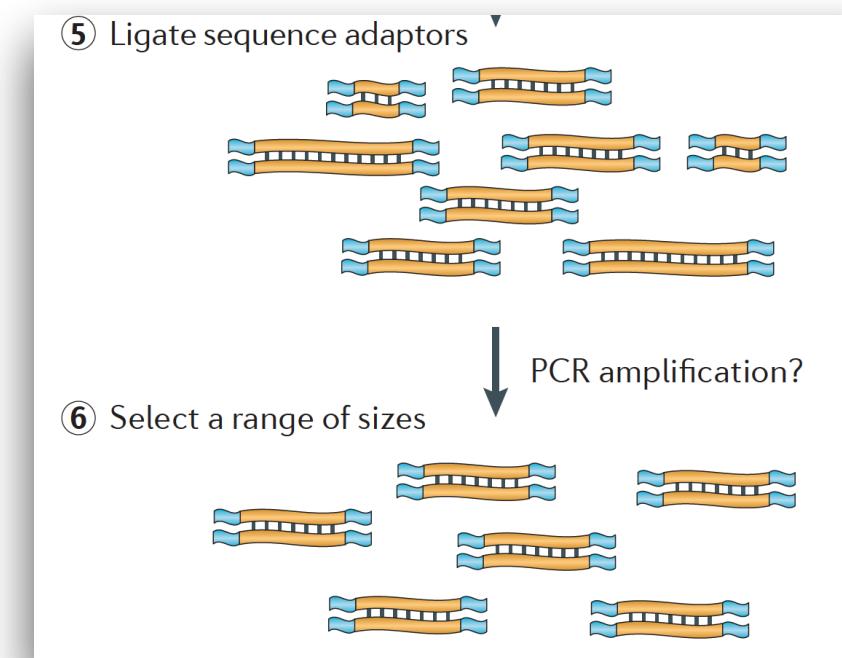
Illumina Sequencing Error Profiles and Quality Control



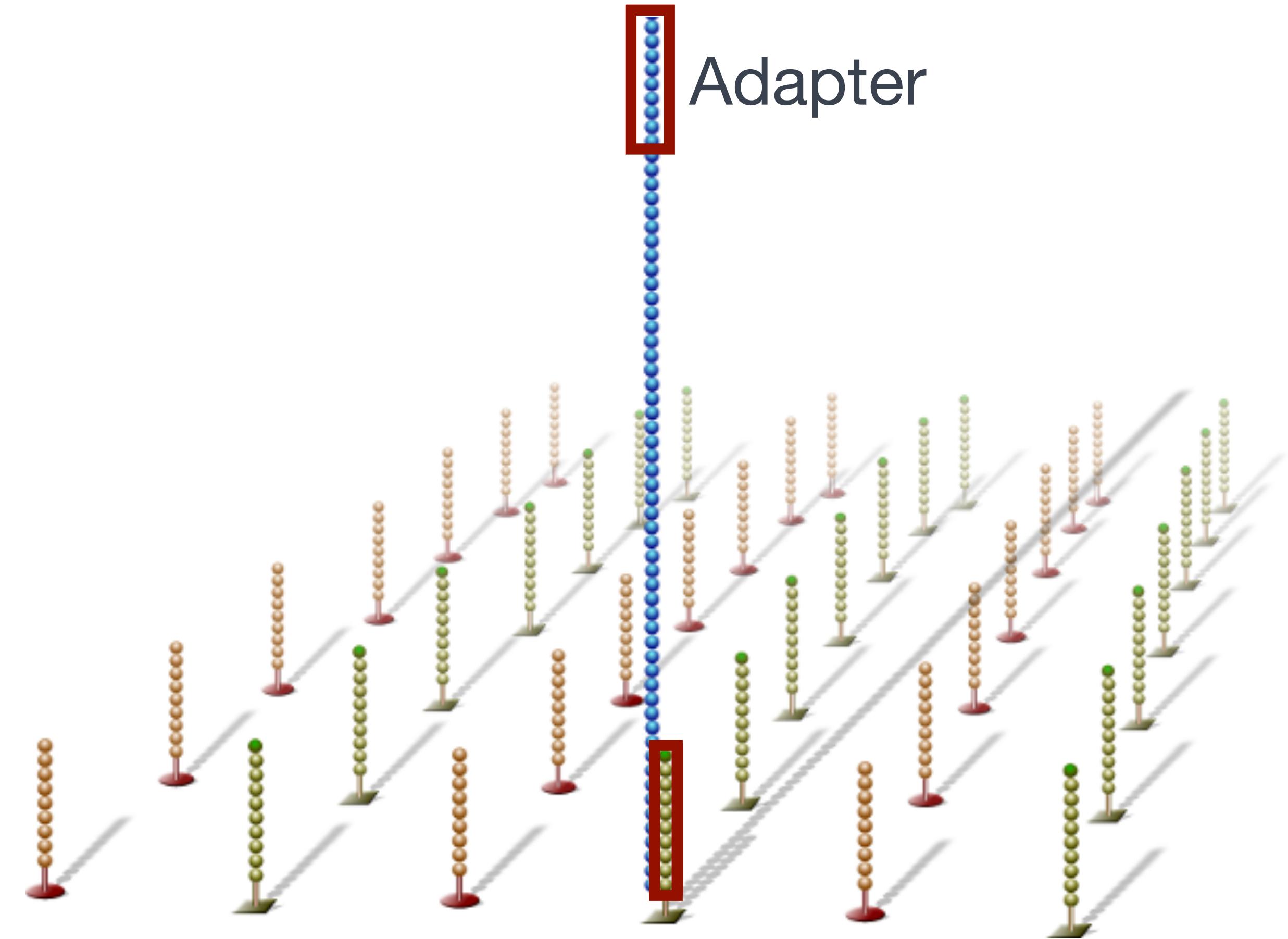
Workflows and Data Standards

Illumina sequencing

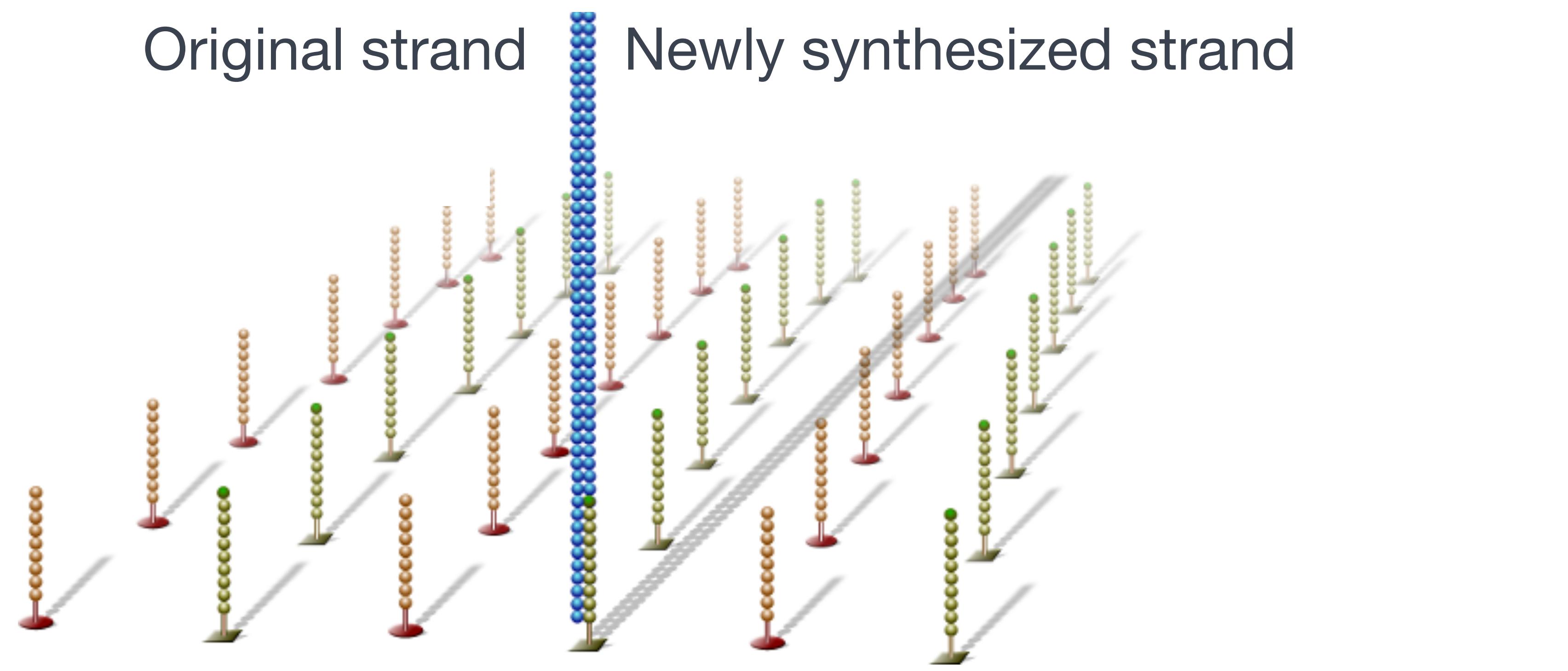
TTAATGATACTGGGACCCCGAGAUCTACAC-3'
TTCAAGGAGAACGGCATACGAG0x0AT-3'



Illumina: flow cell

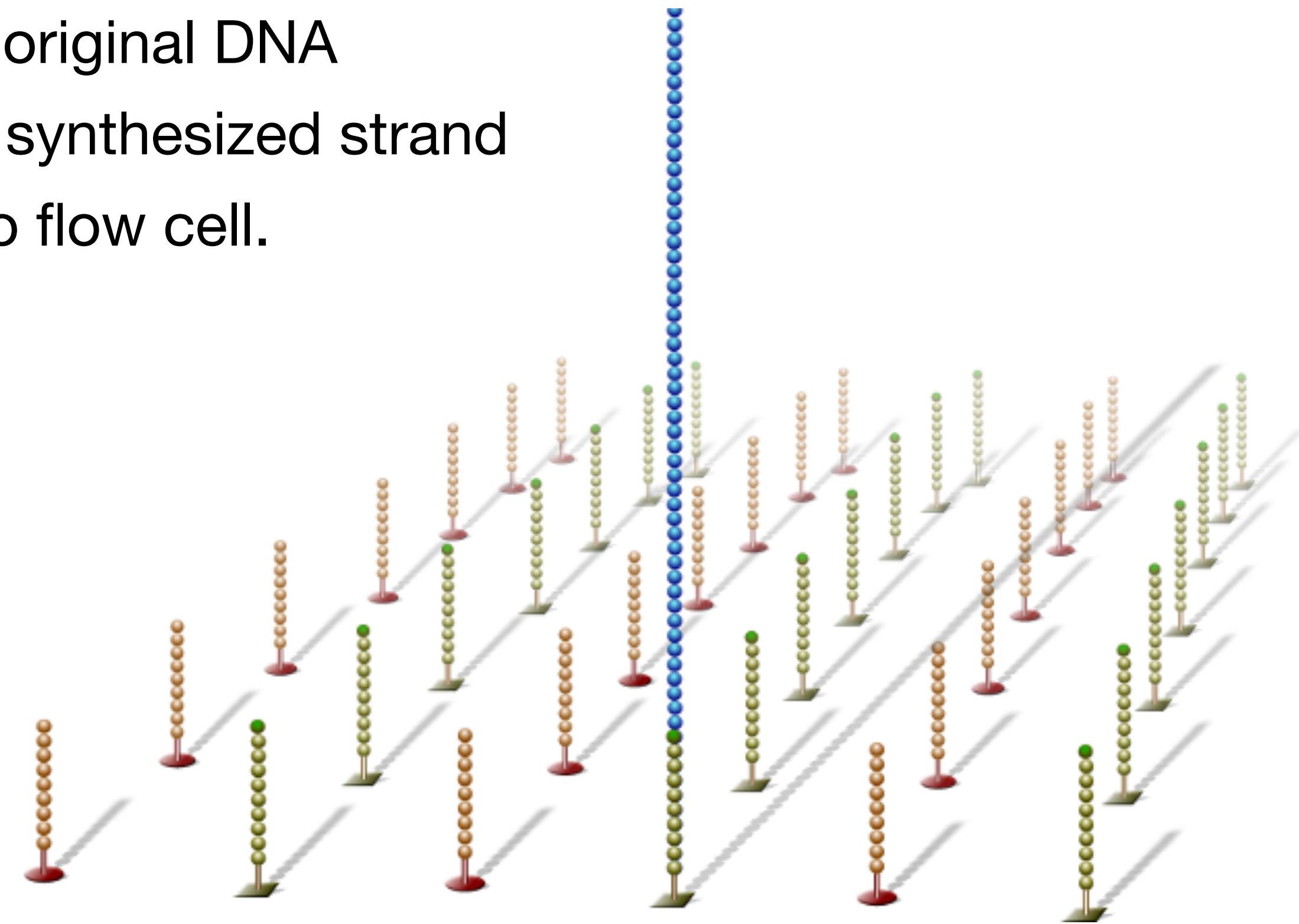


Illumina: cluster generation

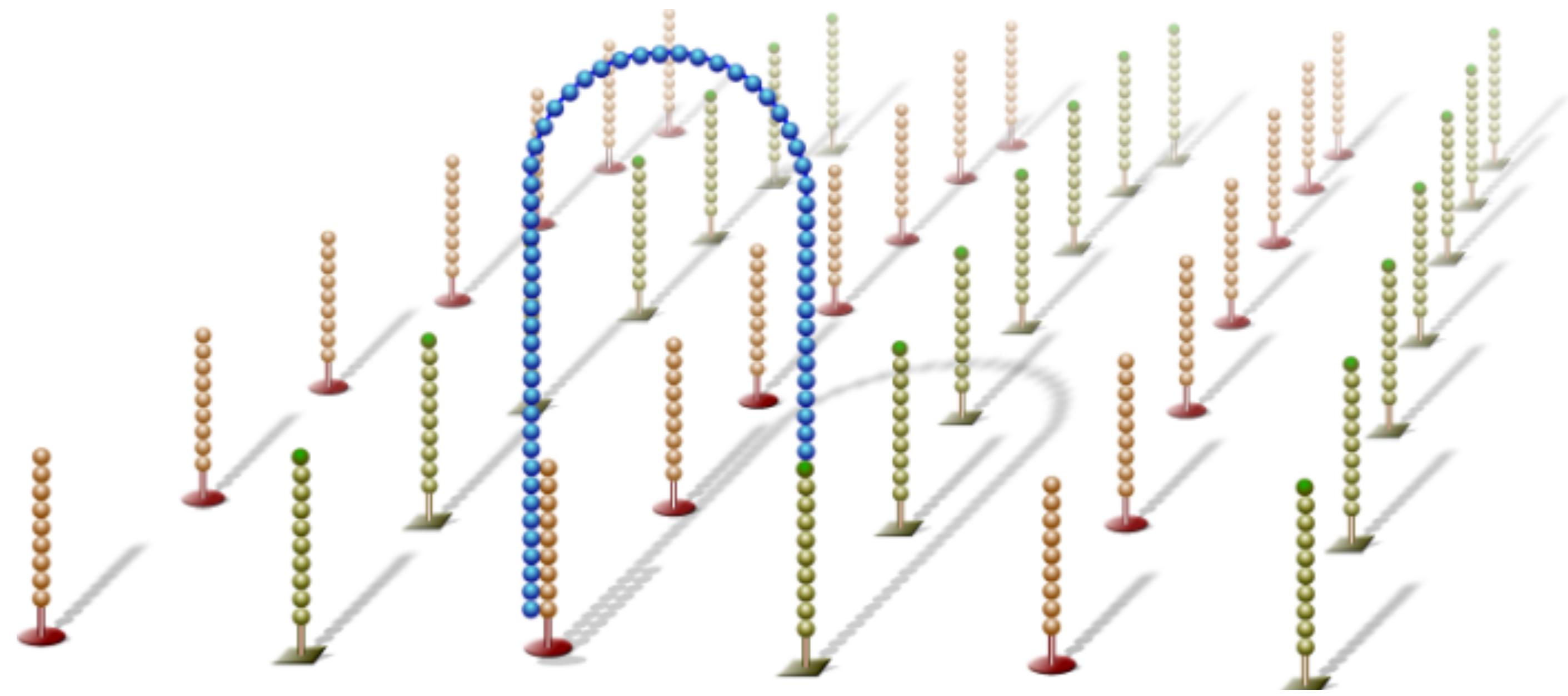


Illumina: cluster generation

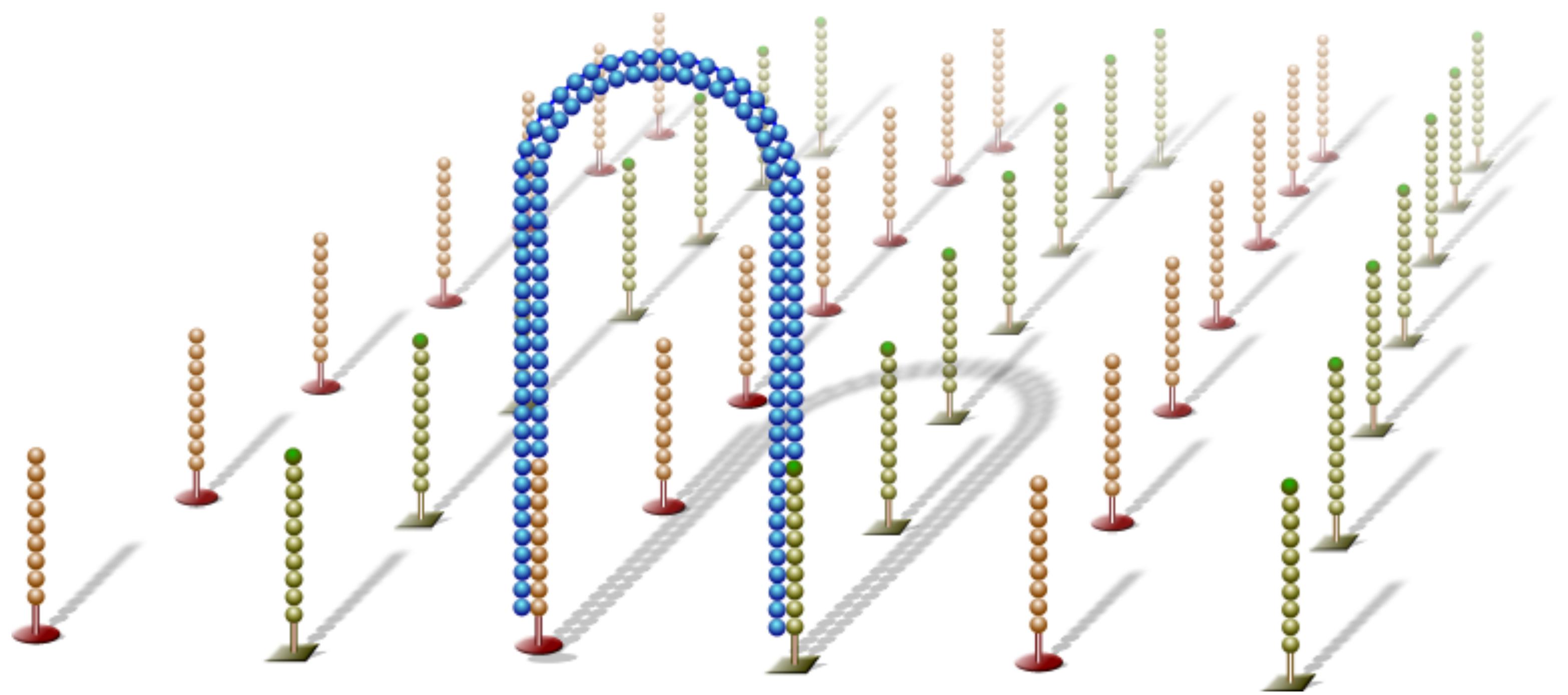
dsDNA is denatured, original DNA washed away. Newly synthesized strand is covalently bound to flow cell.



Illumina: cluster generation



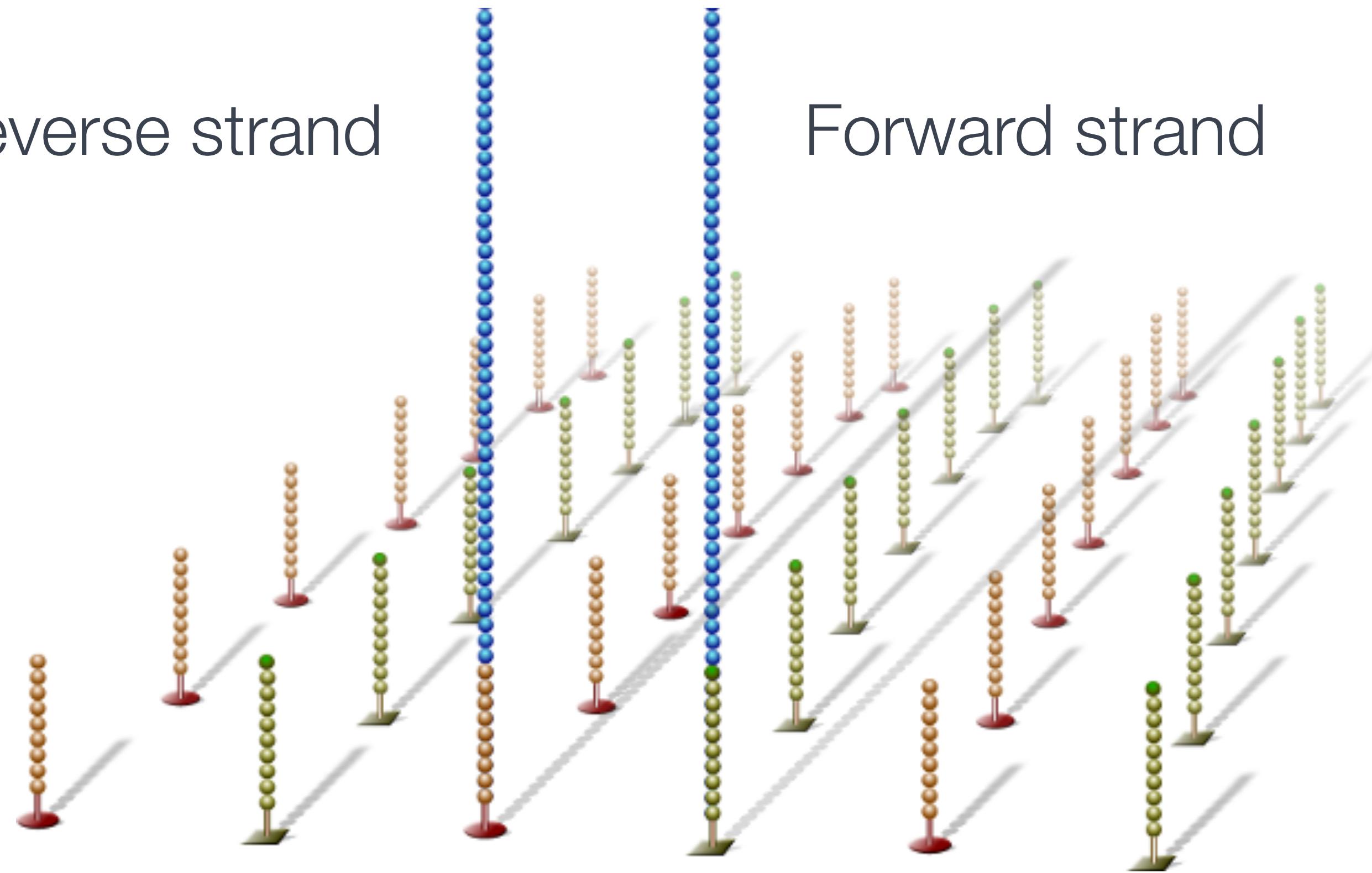
Illumina: bridge amplification



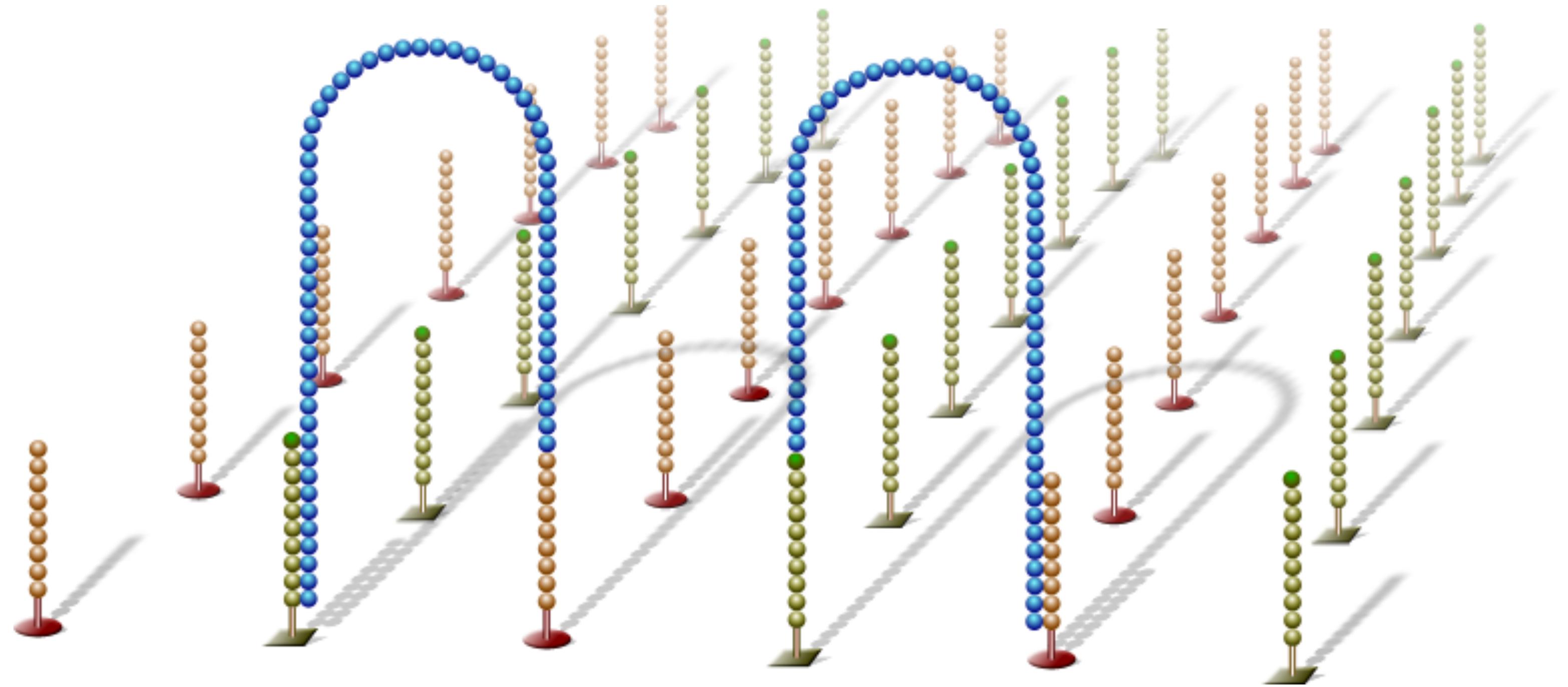
Illumina: bridge amplification

Reverse strand

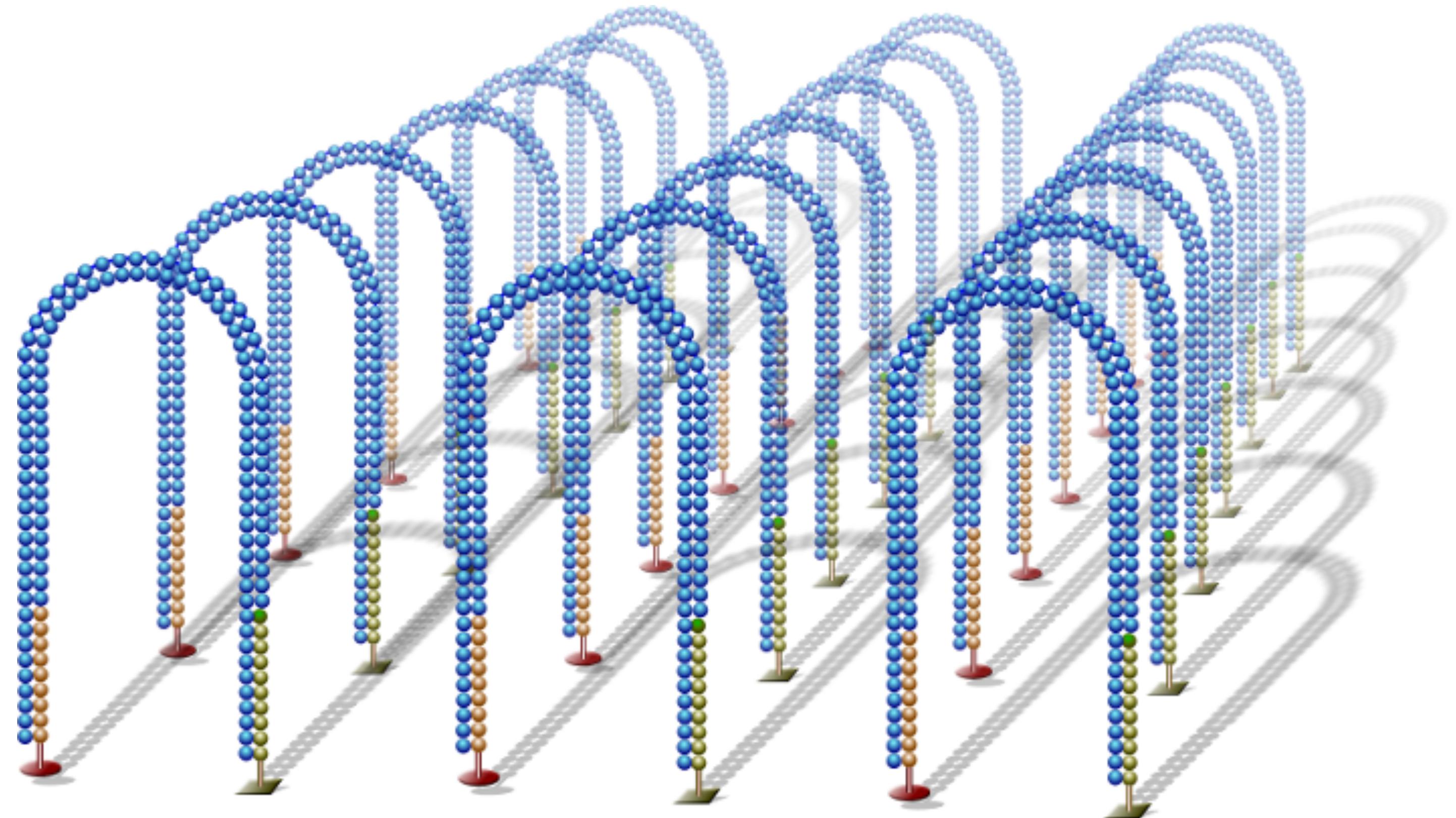
Forward strand



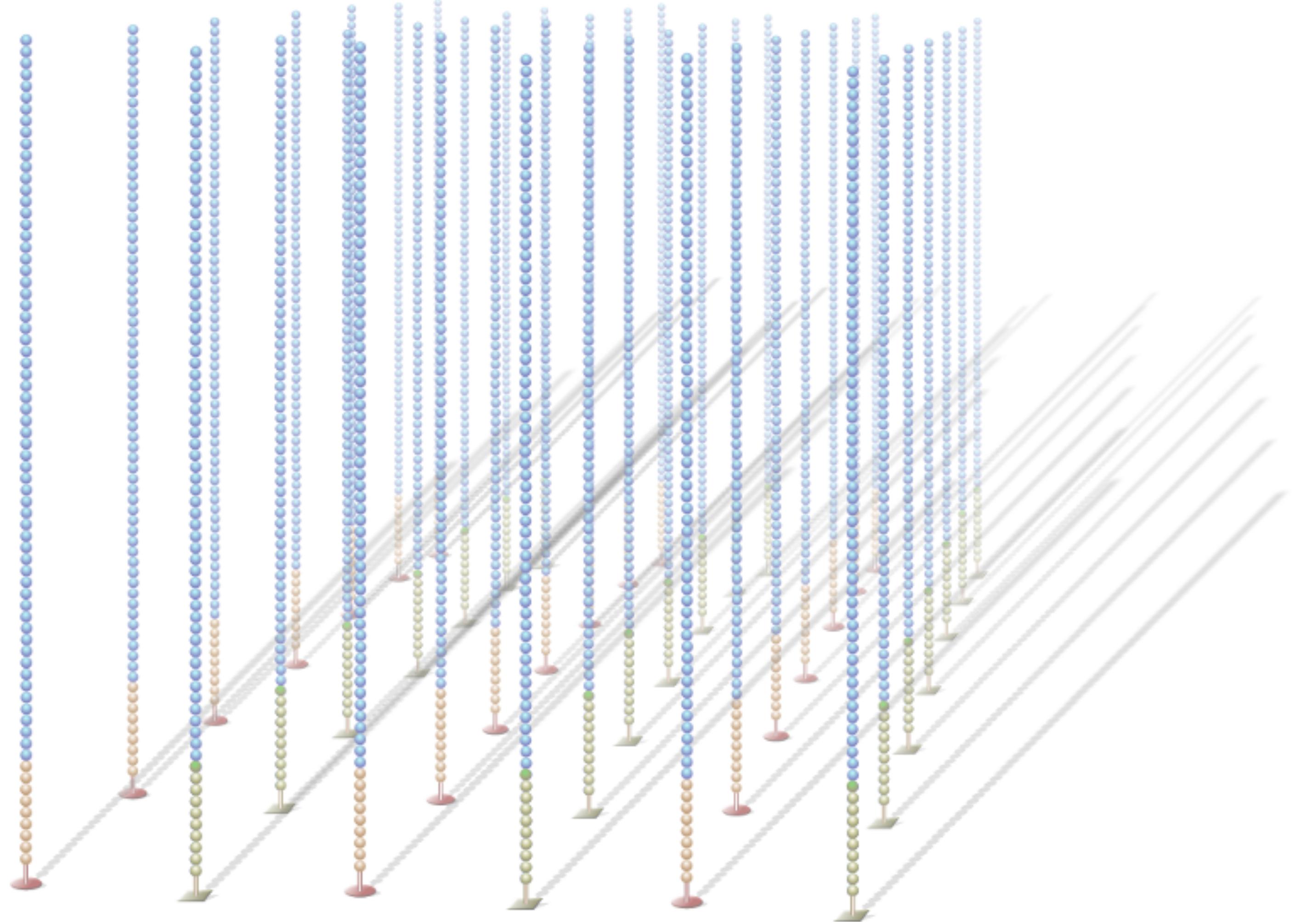
Illumina: bridge amplification



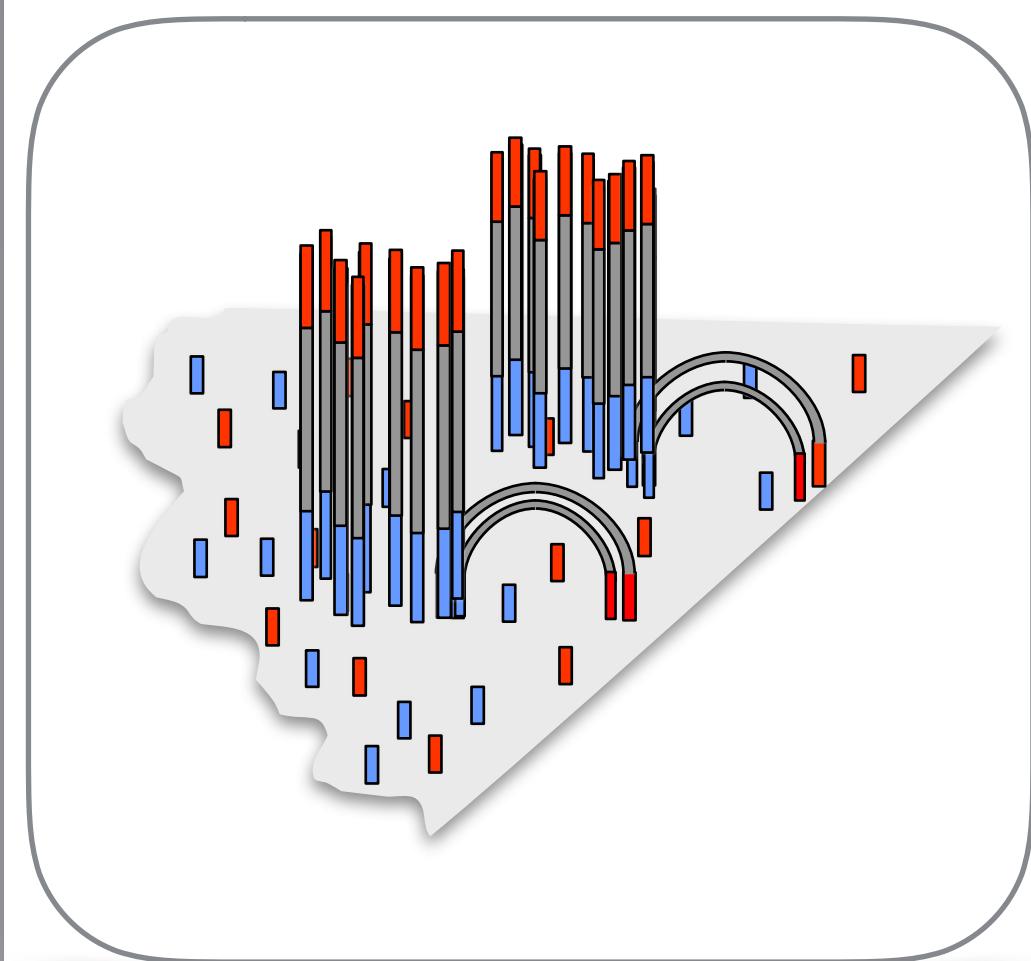
Illumina: bridge amplification

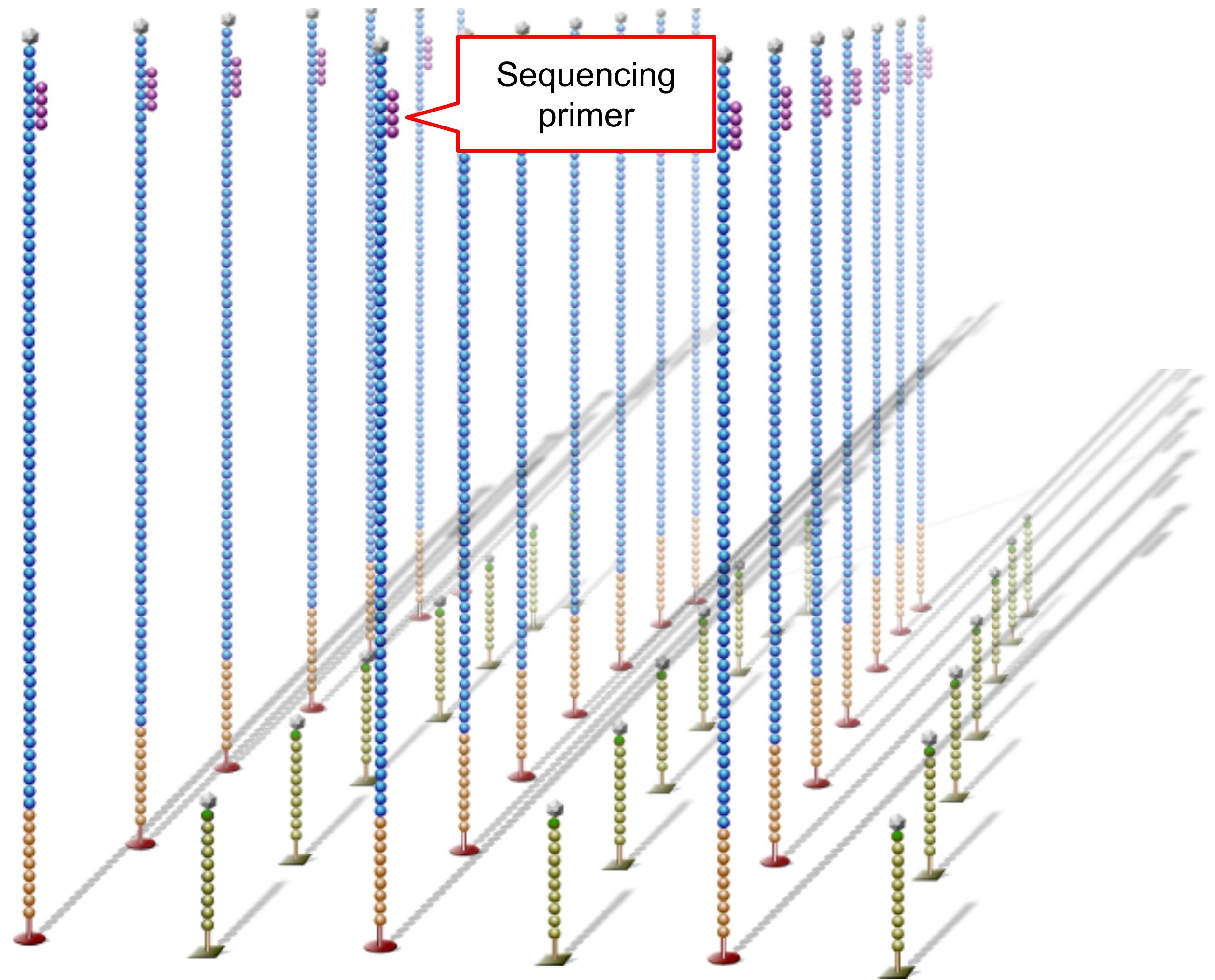


Illumina: bridge amplification

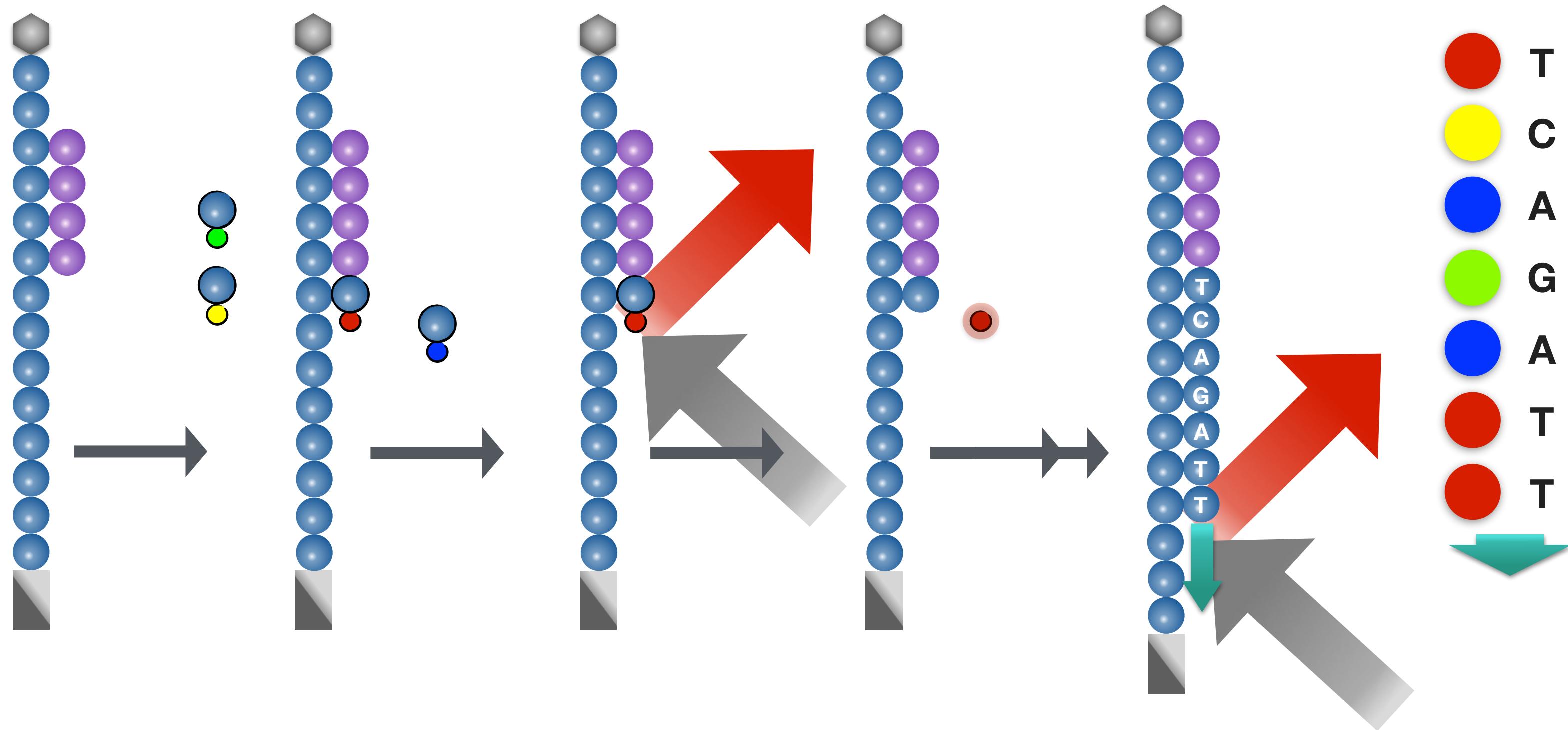


Illumina: cluster generation

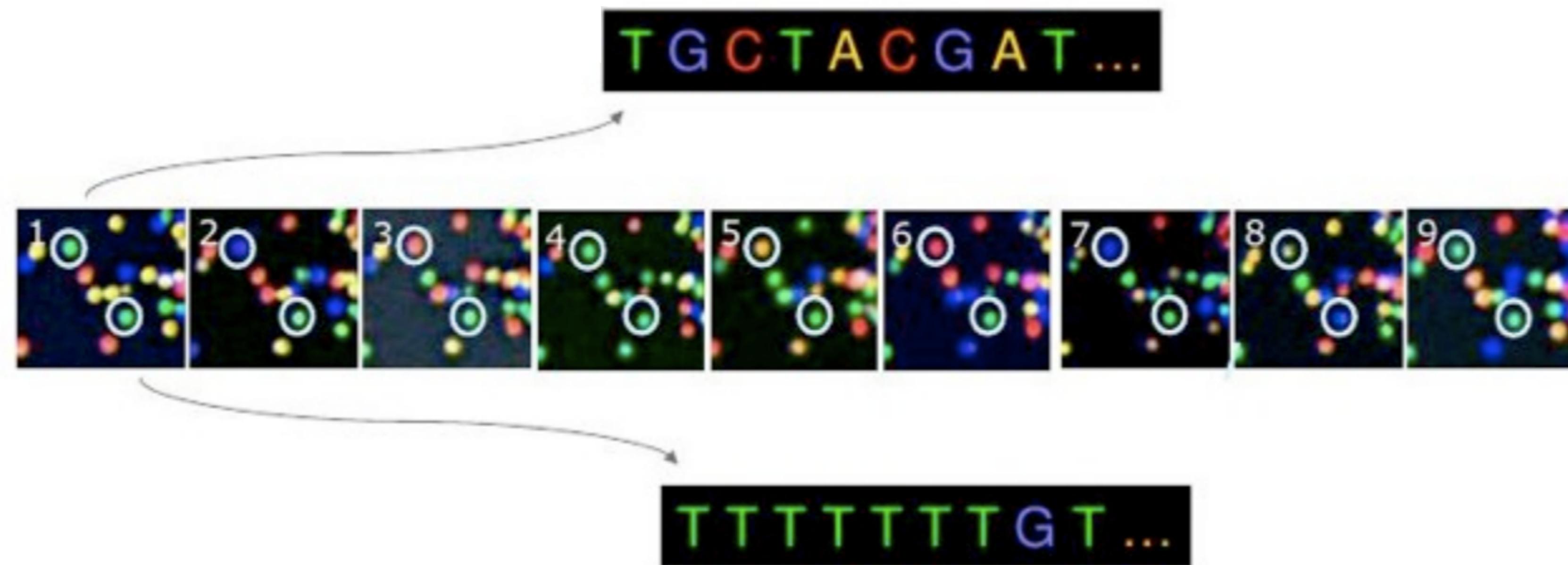




Illumina: Prepare for sequencing



Illumina: sequencing by synthesis



Illumina: base calling

Off the sequencer

FASTA

>SRR014849.1 EIXKN4201CFU84 length=93

GGGGGGGGGGGGGGGGCTTTTGTGAACCGAAAGGGTTTGAAATTCAAACCTTTCGTTCAAACCTCCAAAGCAATGCCAATA

>gi|340780744|ref|NC_015850.1| Acidithiobacillus caldus SM-1 chromosome, complete genome

ATGAGTAGTCATTCAAGCGCCGACAGCGTTGCAAGATGGAGCCGCGCTGTGGTCCGCCCTATGCGTCCAACTGGAGCTCGTCACGAG
TCCGCAGCAGTTCAATACCTGGCTGC GGCCCTGCGTGGCGAATTGCAGGGTCATGAGCTGCGCCTGCTCGCCCCAATCCCTCG
TCCGCGACTGGGTGCGTGAACGCATGCCGAACCTCGTAAGGAACAGCTGCAGCGGATCGCTCCGGTTTGAGCTGGTCTCGCT
CTGGACGAAGAGGCAGCAGCGGCACATCGCACCGACCAGCGAGCATTGCGCCCGAGCGCAGCGCACCCGGTGGTCACCGCCT
CAACCCAGCCTCAACTTCCAGTCCTACGTGCAAGGGAAGTCCAATCAGCTGCCCTGGCGGAGCCGCCAGGTTGCCAGCATC
CAGGCAAATCCTACAACCCACTGTACATTATGGTGGTGGGCCTCGGCAAGACGCACCTCATGCAGGCCGTGGCAACGATATC
CTGCAGCGGCAACCGAGGCCAAGGTGCTCTATCAGCTCCGAAGGCTTCATCATGGATATGGTGCCTCGCTGCAACACAATAC
CATCAACGACTCAAACAGCGTTATCGCAAGCTGGACGCCCTGCTCATCGACGACATCCAGTTGGCAAGGACCGCACCC

>gi|129295|sp|P01013|OVAX_CHICK GENE X PROTEIN (OVALBUMIN-RELATED)

QIKDLLVSSSTDLDLTLVNAIYFKGMWKTA FNAEDTREMPFHVTQESKPVQMMCMNNSFNVATLPAE

FASTQ: FASTA with Quality scores

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTGTGGAAACGAAAGGGTTTGAATTCAAACCCTTCGGTTCCAACCTCCAAAGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#""""""""7F@71,'";C?,B;?6B;:EA1EA1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@/=<?7=9<2A8==
```

Line	Description
1	Always begins with '@' and then information about the read
2	The actual DNA sequence
3	Always begins with a '+' and sometimes the same info in line 1
4	Has a string of characters which represent the quality score

FASTQ Quality Encoding

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTGTGGAACCGAAAGGGTTTGAATTCAAACCCTTCGGTTCCAACCTCCAAAGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#""""""""7F@71,'";C?,B;?6B;:EA1EA1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@/=<?7=9<2A8==
```

Quality encoding:	!"#\$%&'()*+,-./0123456789:@<=>?@ABCDEFGHI
Quality score:	0.....10.....20.....30.....40

The legend above provides the mapping of quality scores (Phred-33) to the quality encoding characters.

Different quality encoding scales exist (differing by offset in the ASCII table), but note the most commonly used one is fastqsanger.

FASTQ Quality Encoding

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTGTGGAACCGAAAGGGTTTGAATTCAAACCCTTCGGTTCCAACCTCCAAAGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#""""""""7F@71,'";C?,B;?6B;:EA1EA1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@/=<?7=9<2A8==
```

Quality encoding:	!"#\$%&'()*+,-./0123456789:@>?=ABCDEF
Quality score:	0.....10.....20.....30.....40

$Q = -10 \times \log_{10}(P)$, where P is the probability that a base call is erroneous

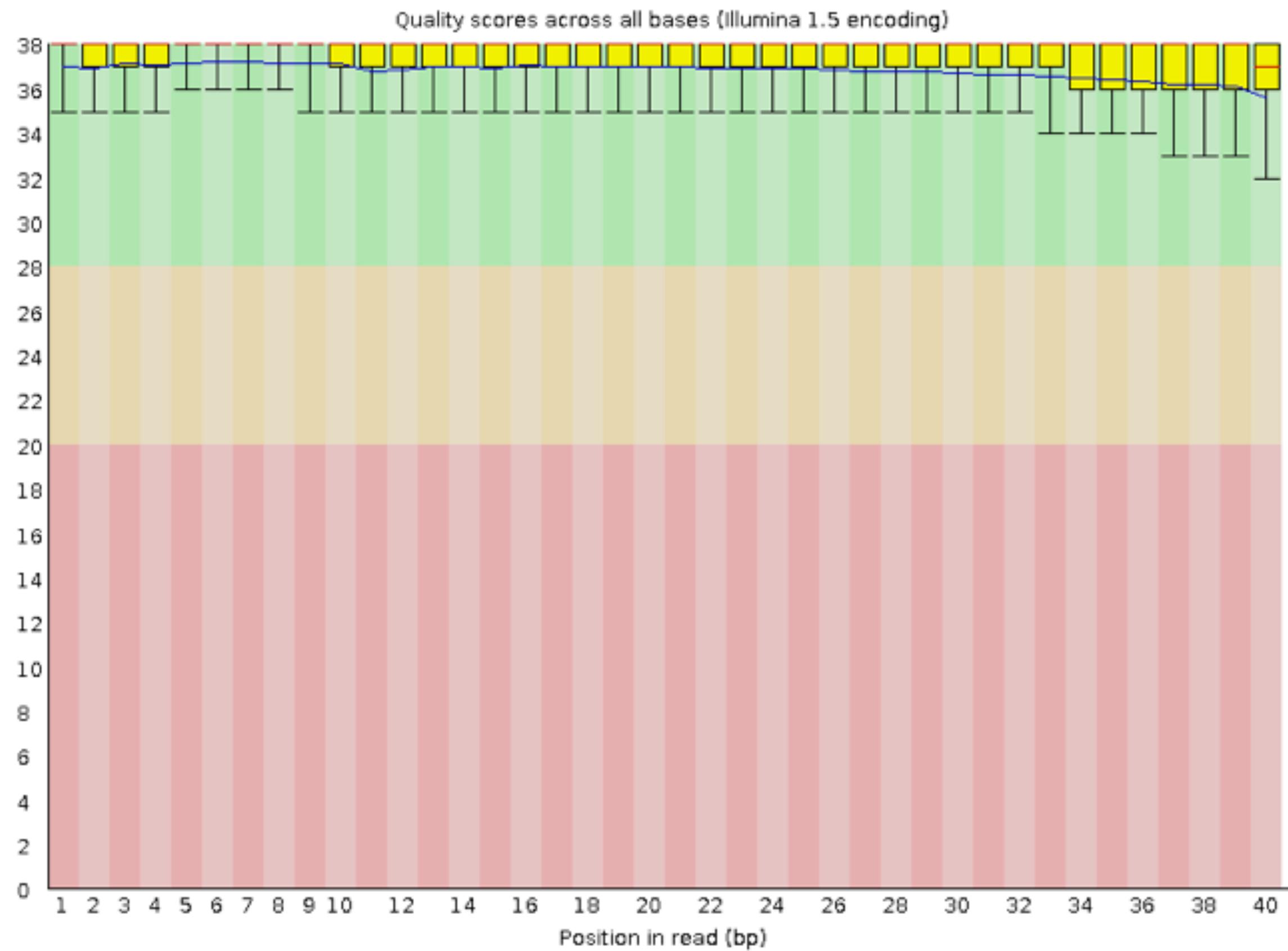
The legend above provides the mapping of quality scores (Phred-33) to the quality encoding characters.

Different quality encoding scales exist (differing by offset in the ASCII table), but note the most commonly used one is fastqsanger.

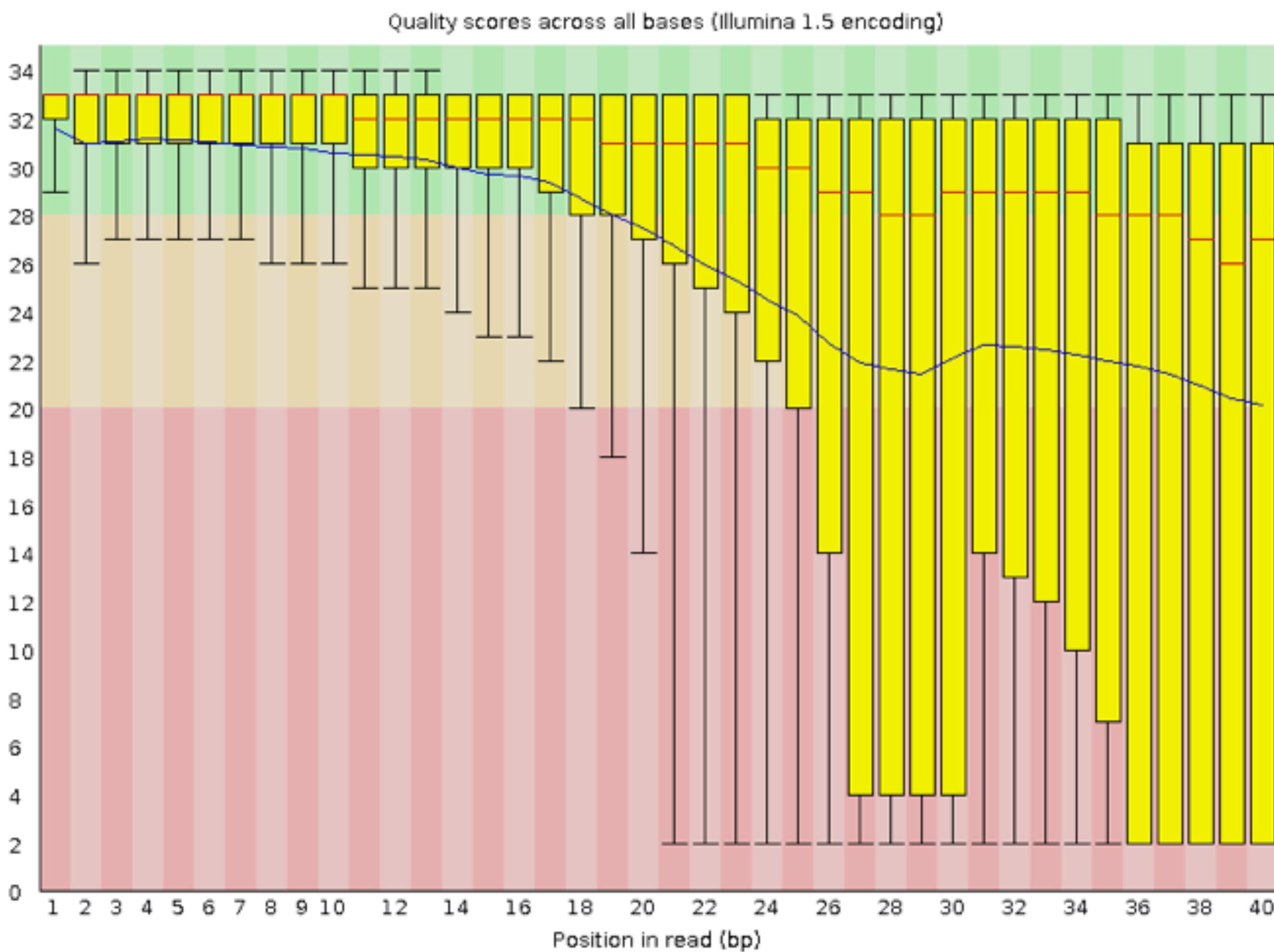
FASTQ Quality Scores

These probability values are the results from the base calling algorithm and dependent on how much signal was captured for the base incorporation. The score values can be interpreted as follows:

Phred 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%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%



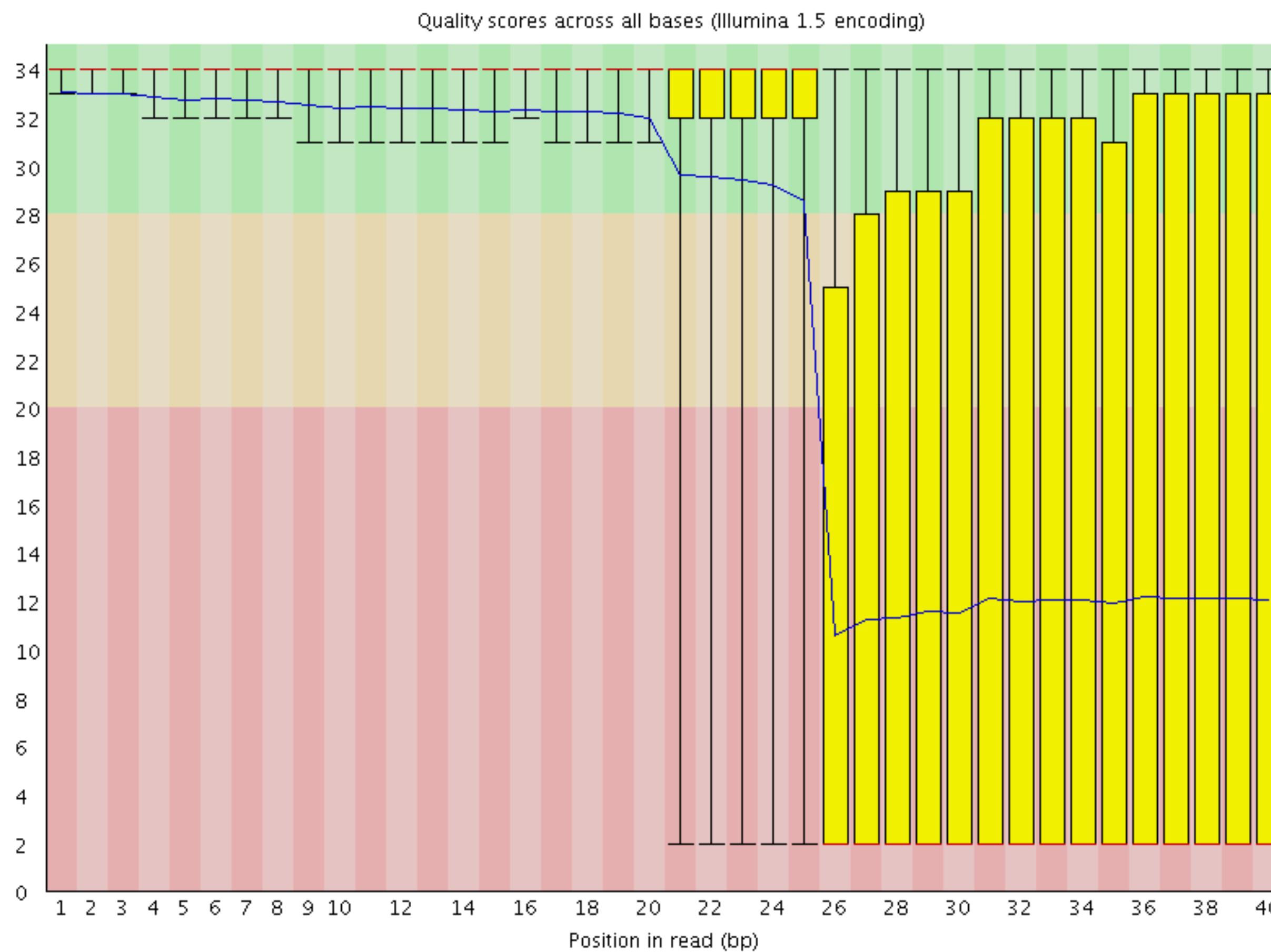
A good quality sample



A not-so-good quality sample

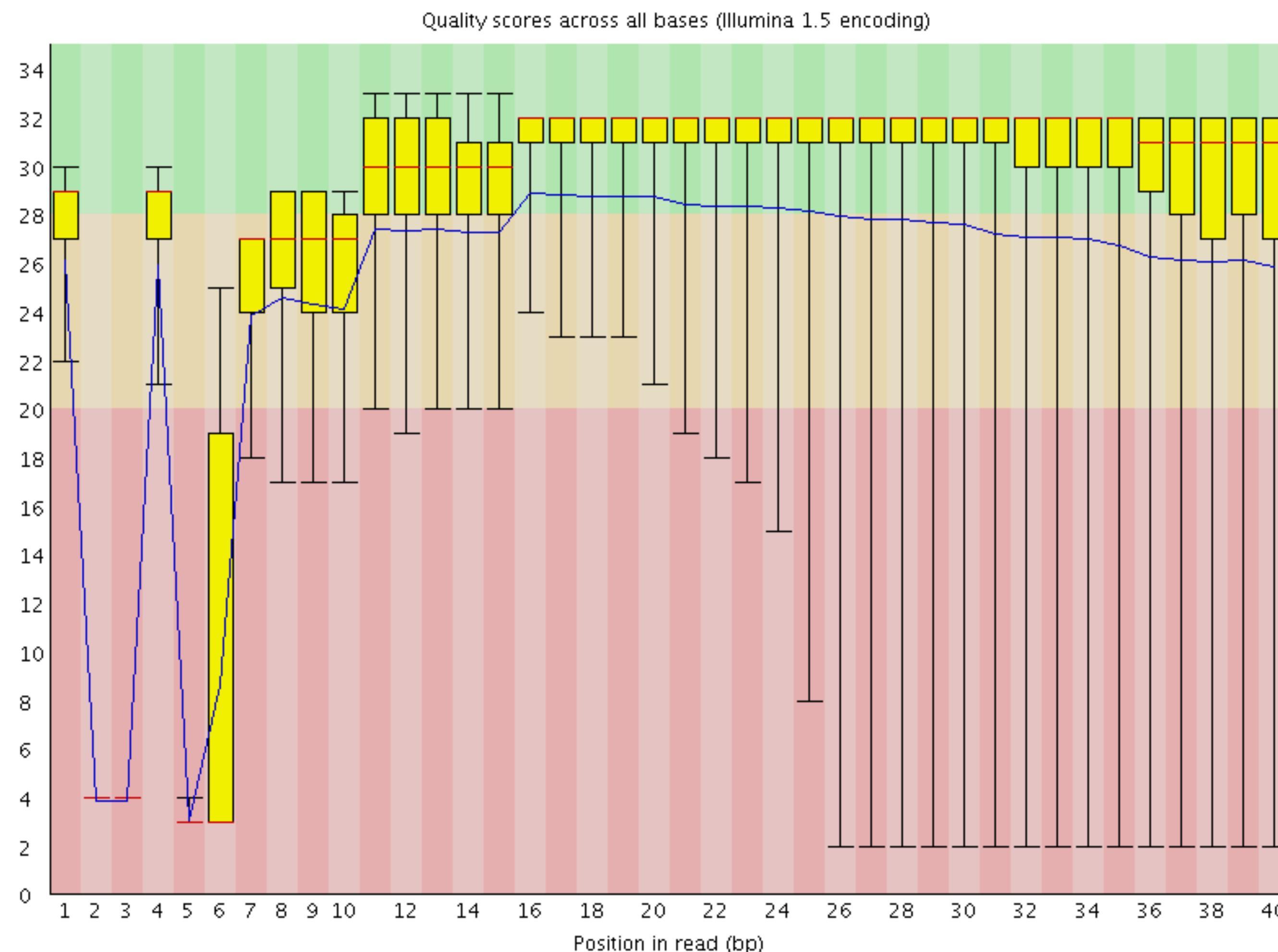
Error profiles: Technical Sequencer Problems

Manifold burst in cycle 26

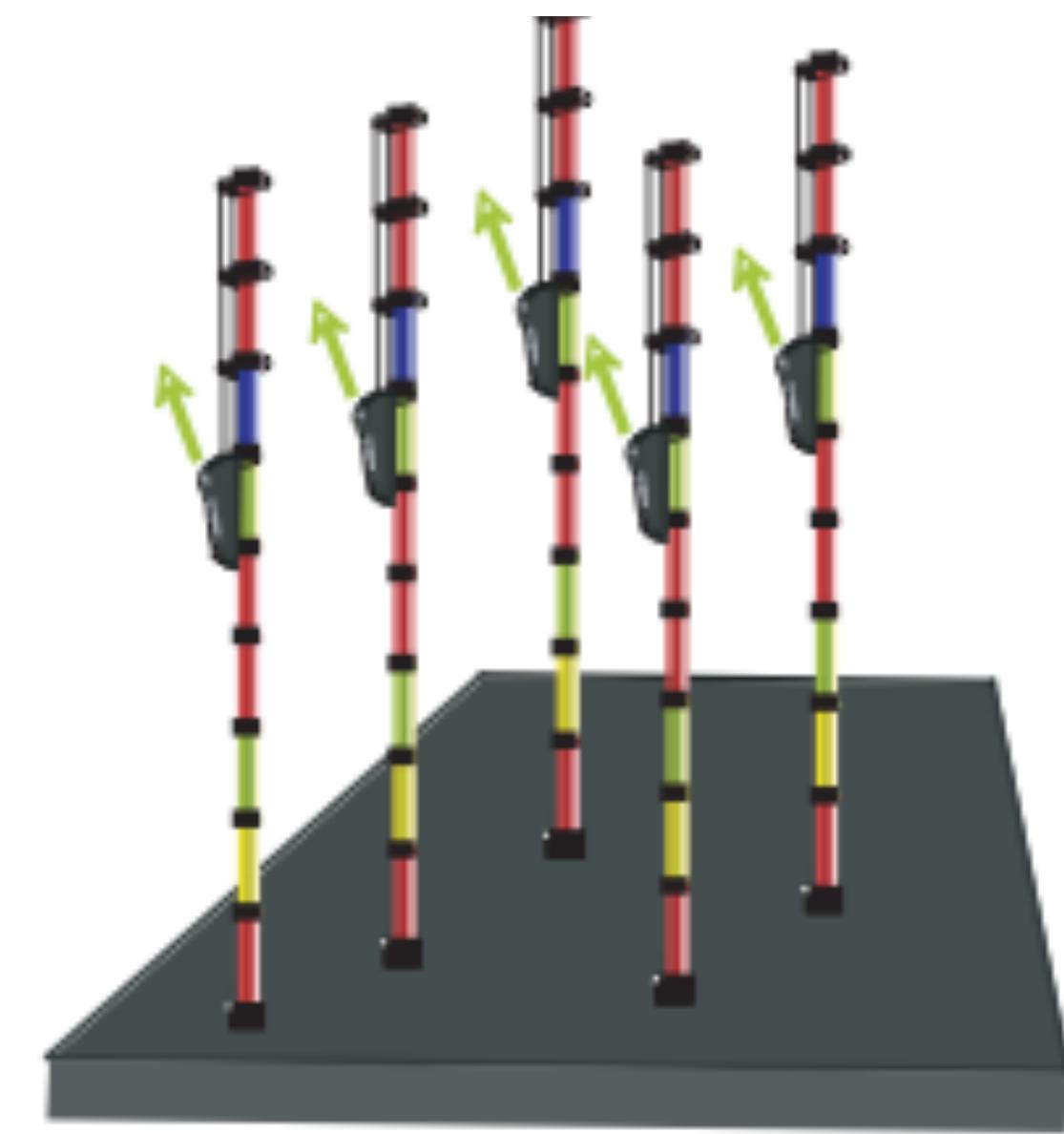


See http://bioinfo-core.org/index.php/9th_Discussion-28_October_2010 for more example

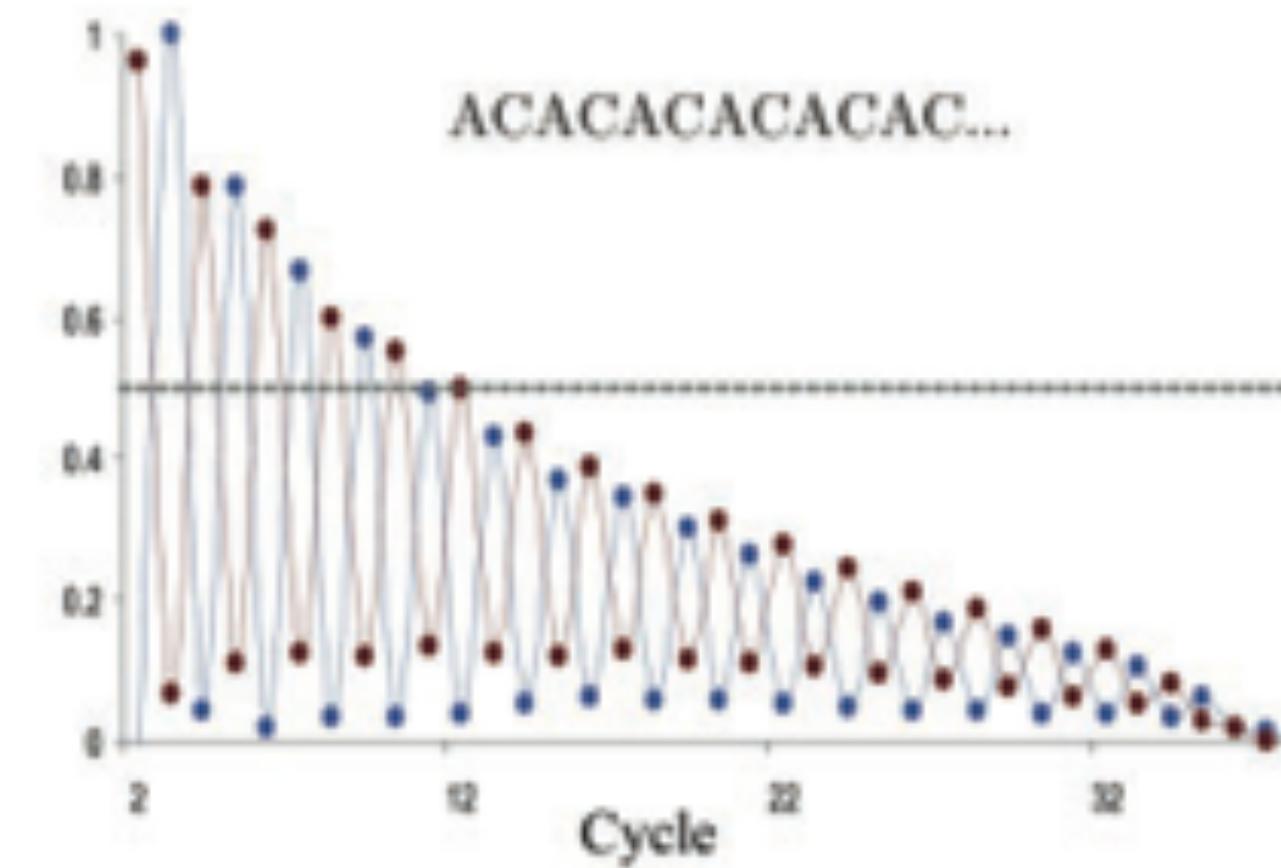
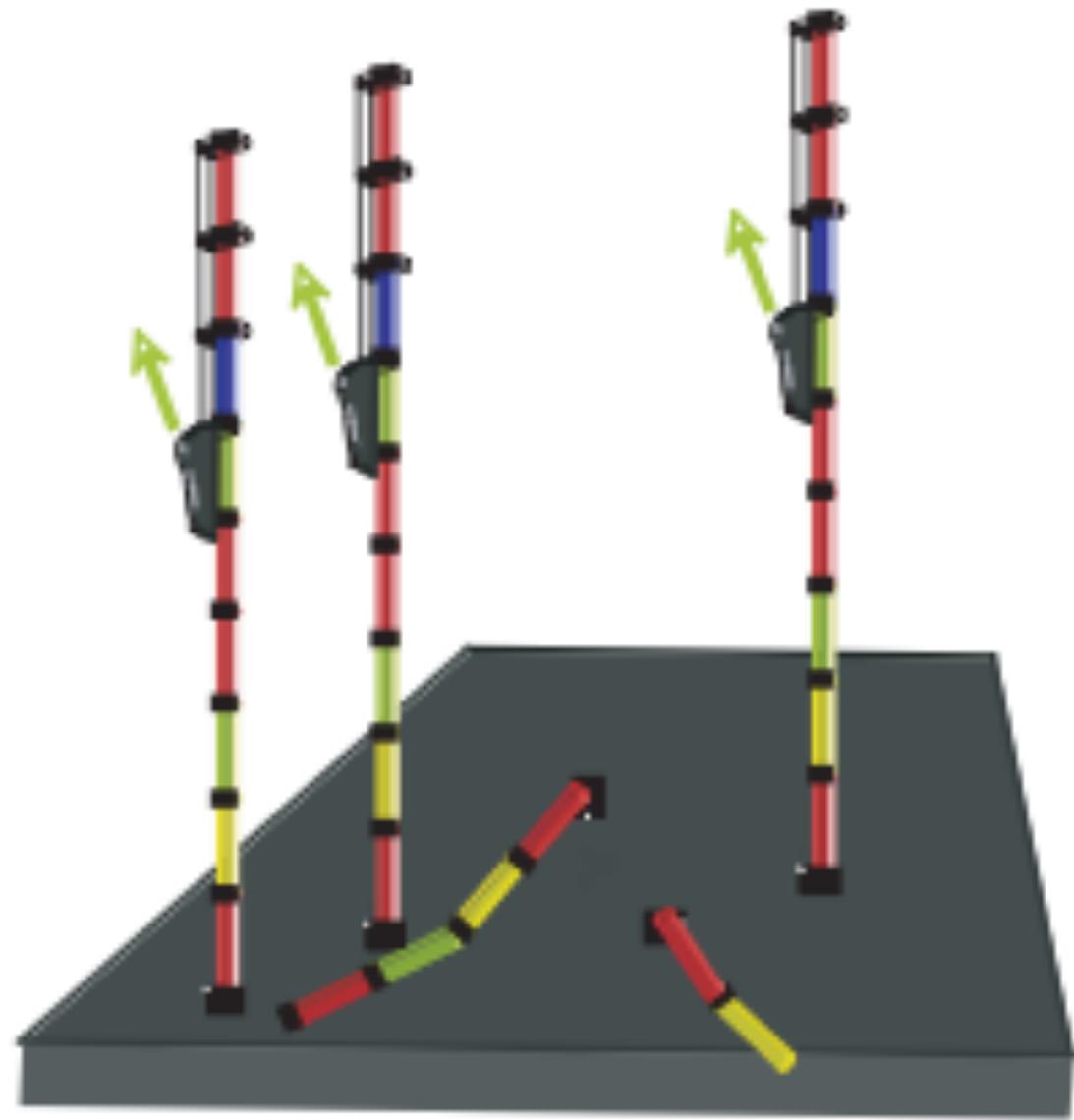
Specific cycles lost



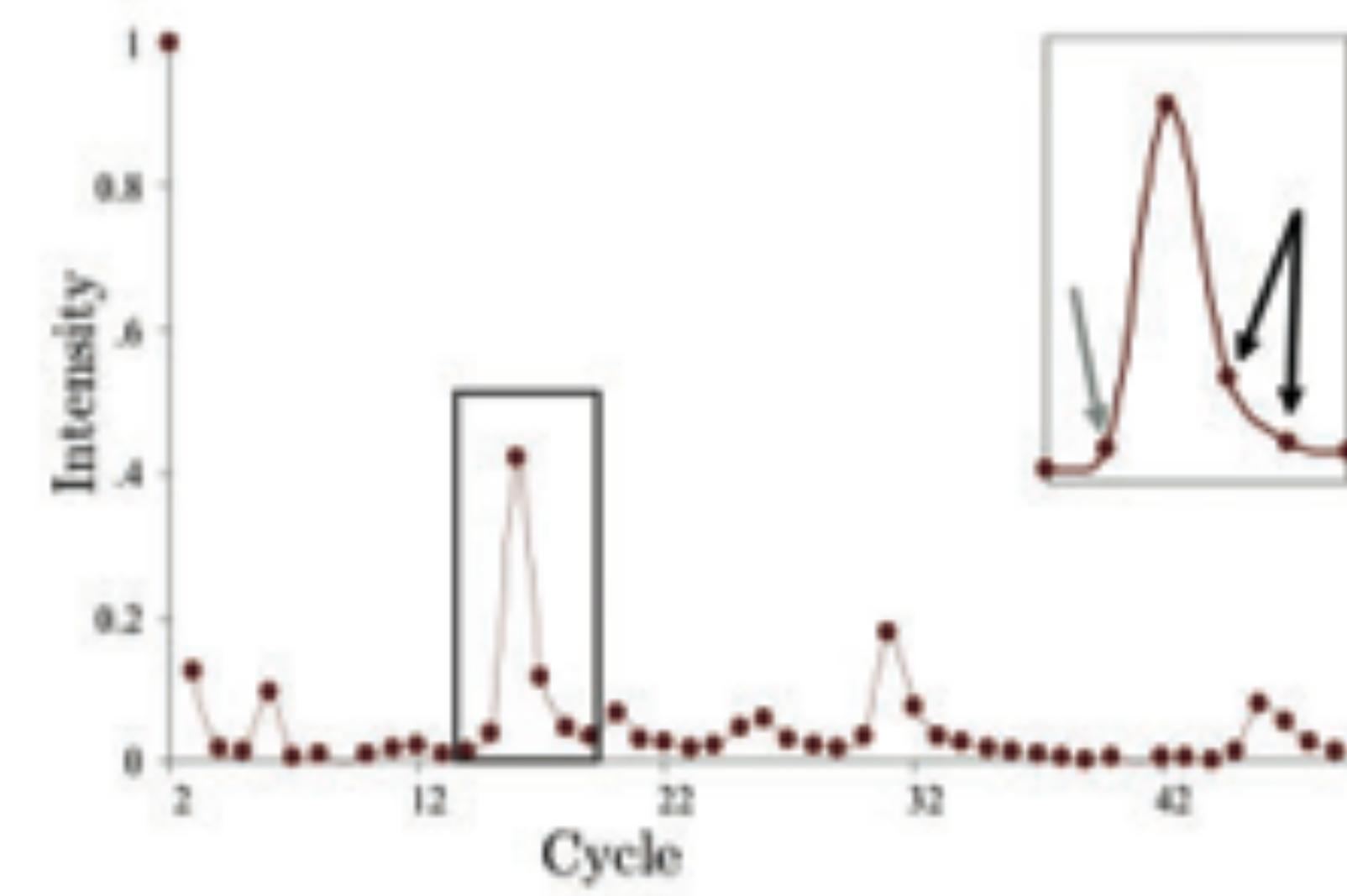
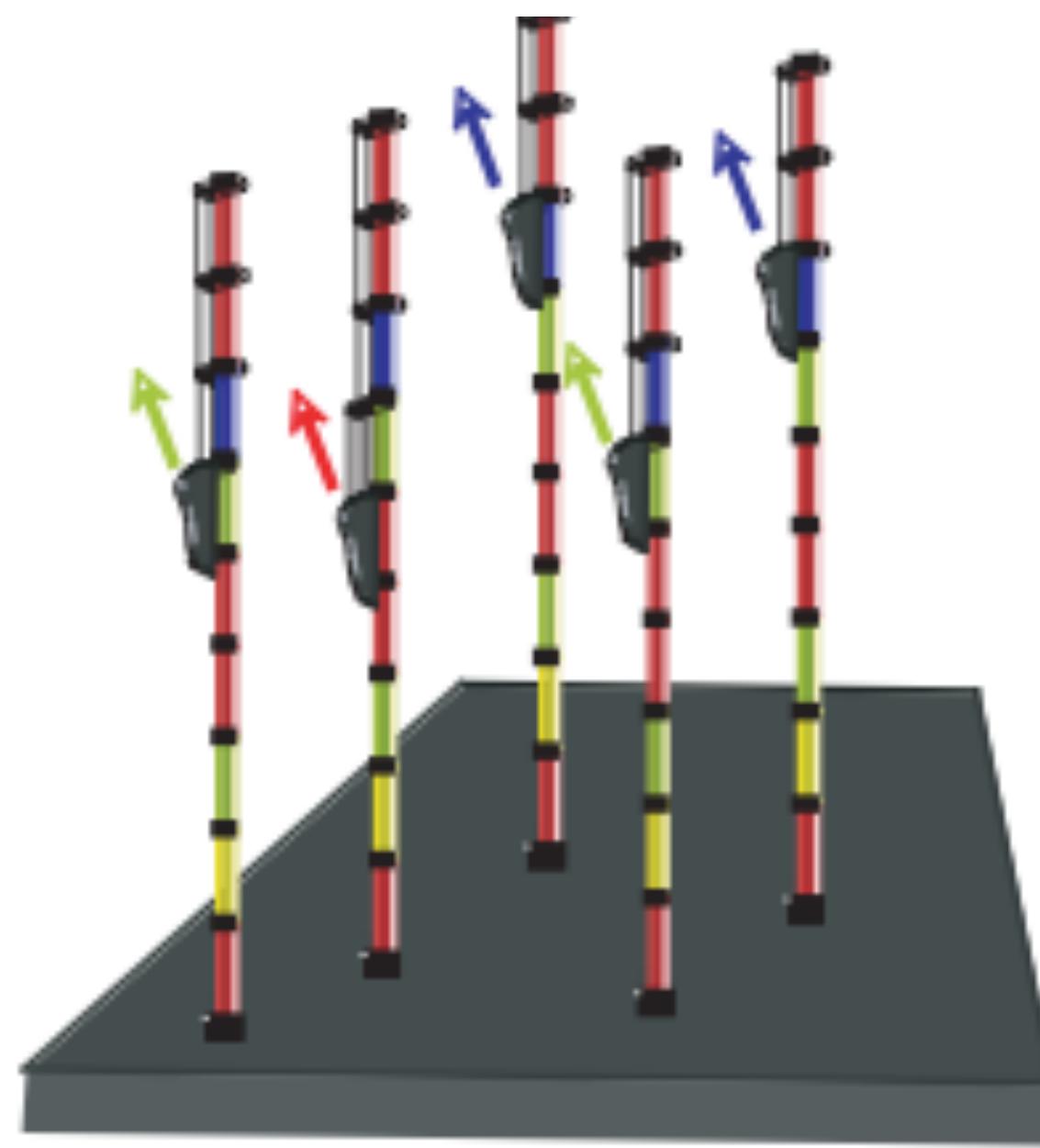
Error dependency on technology



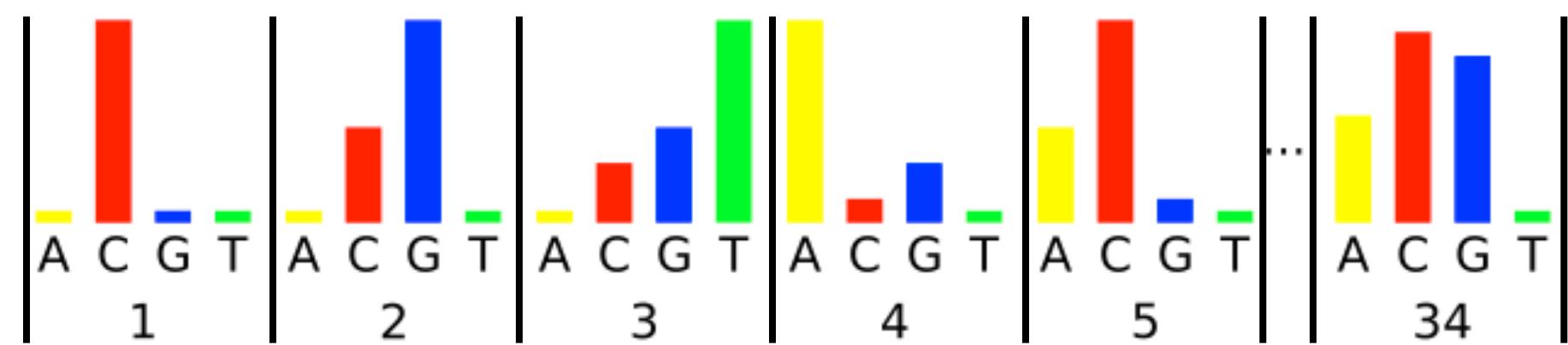
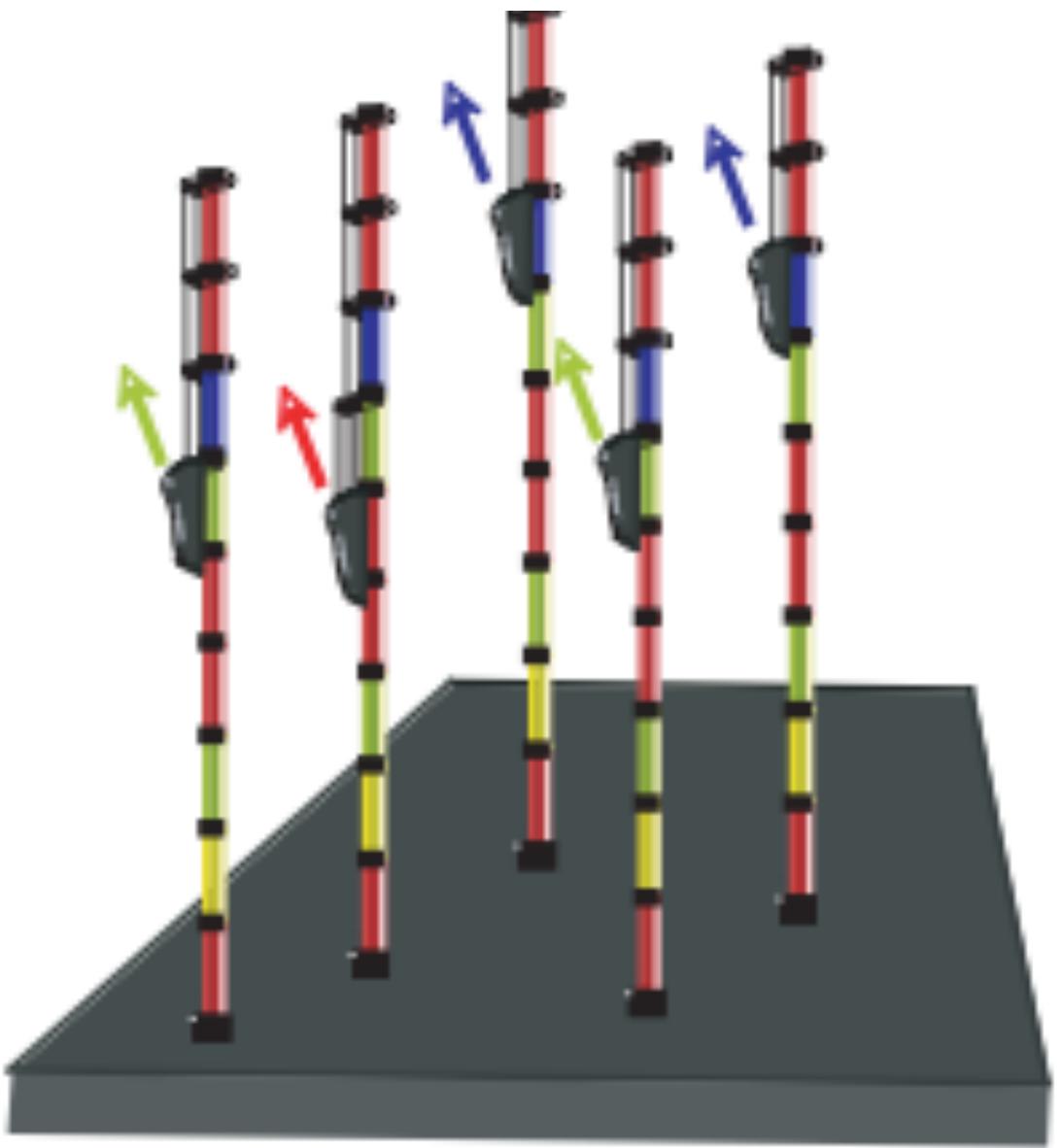
Illumina



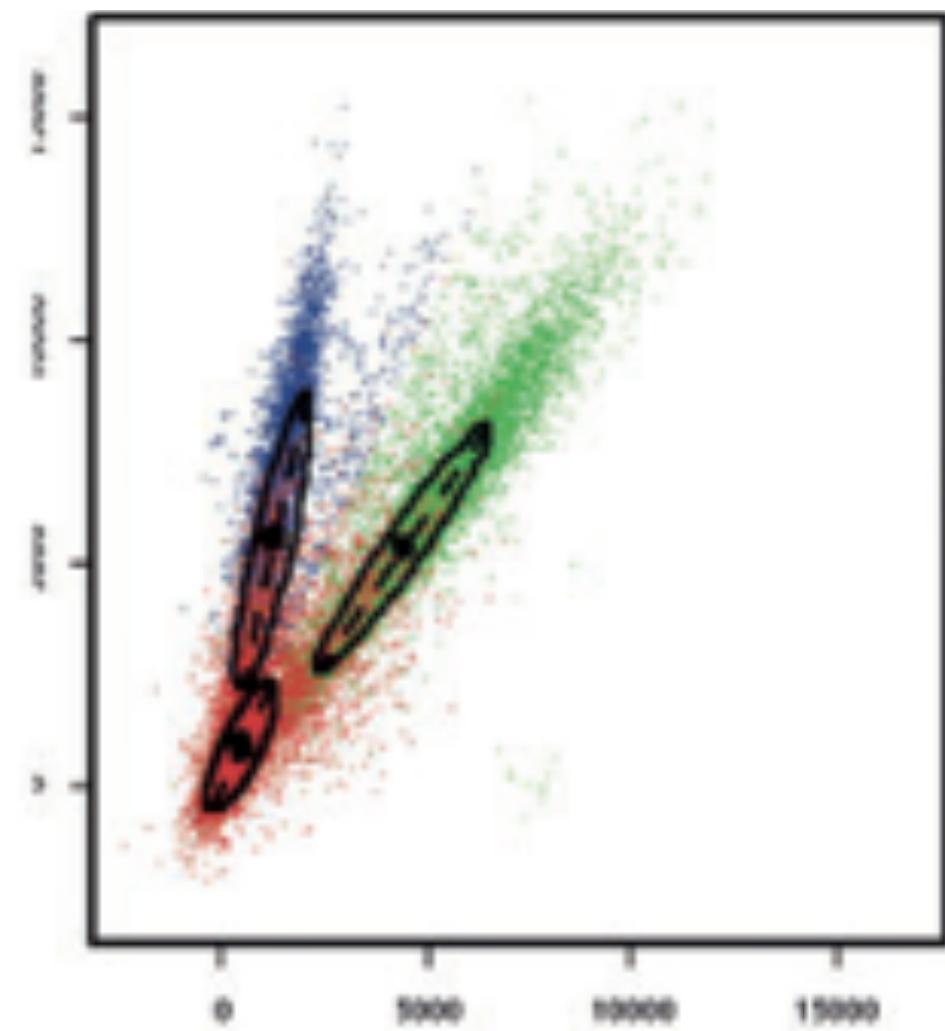
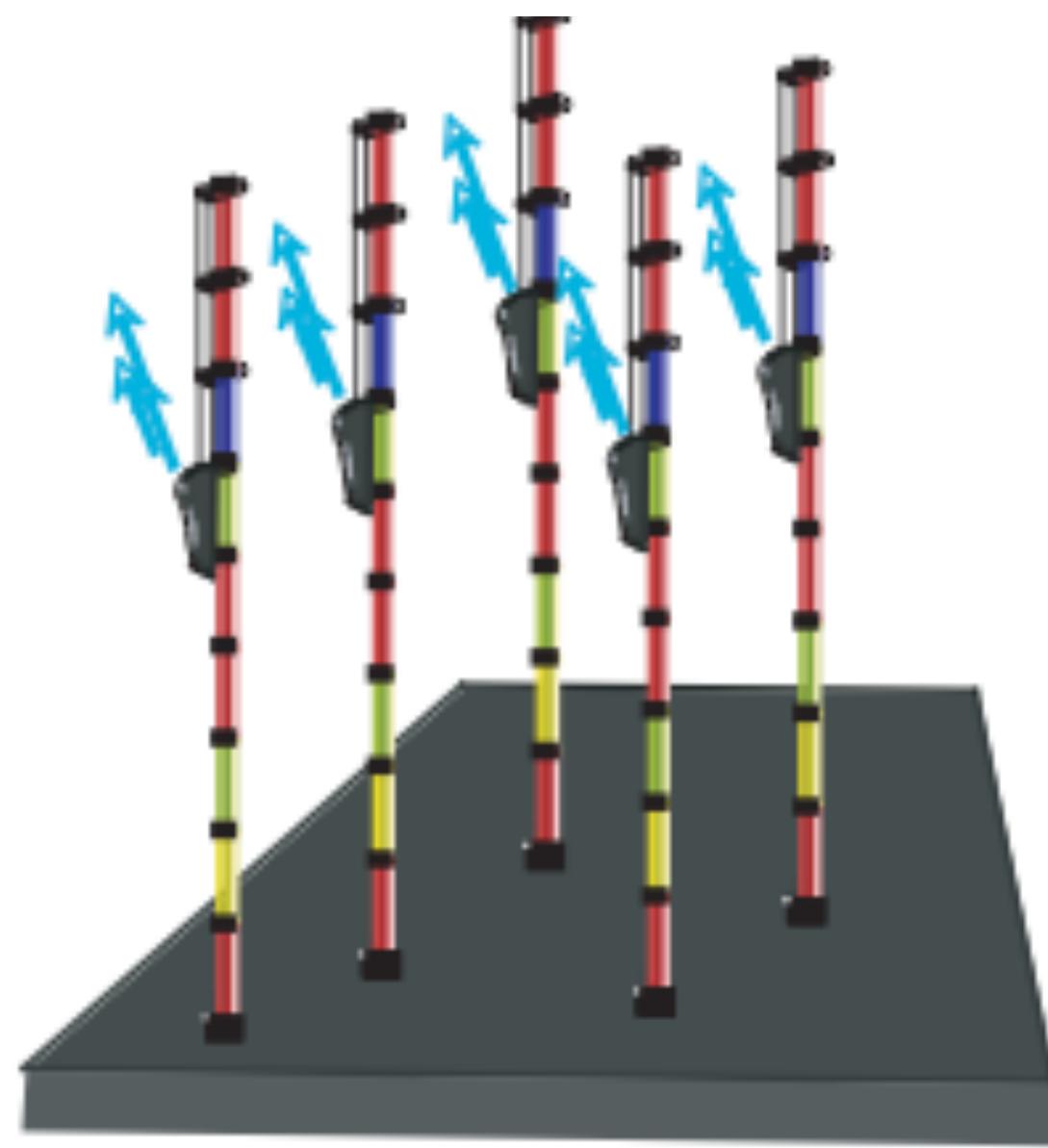
Illumina: signal decay



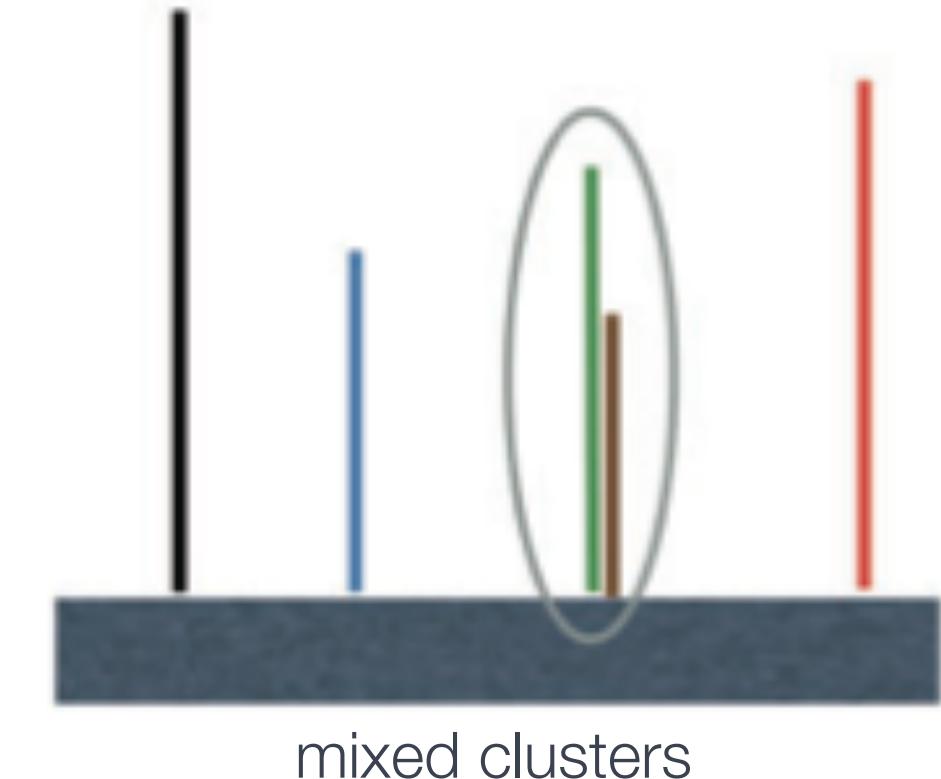
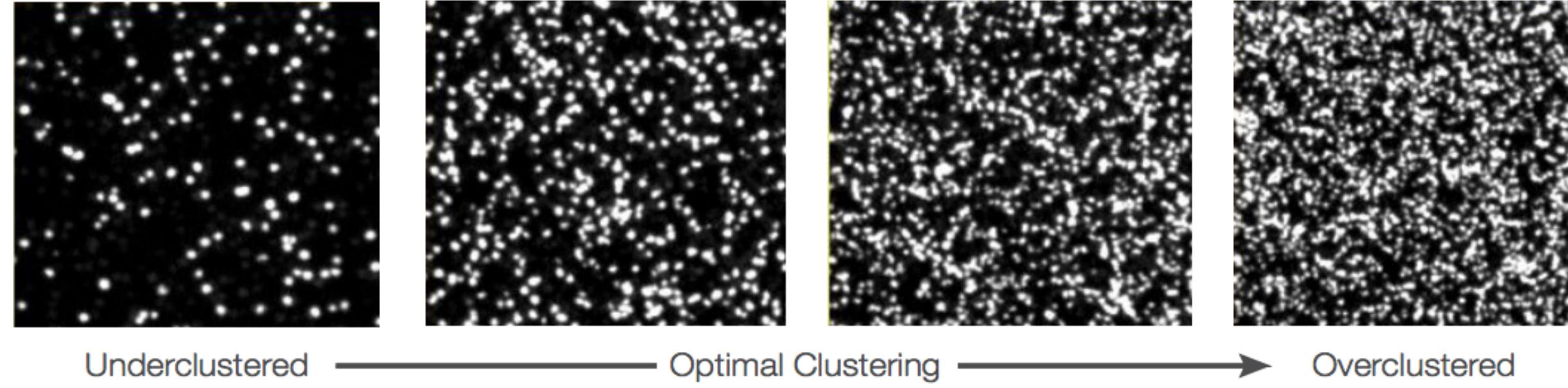
Illumina: phasing



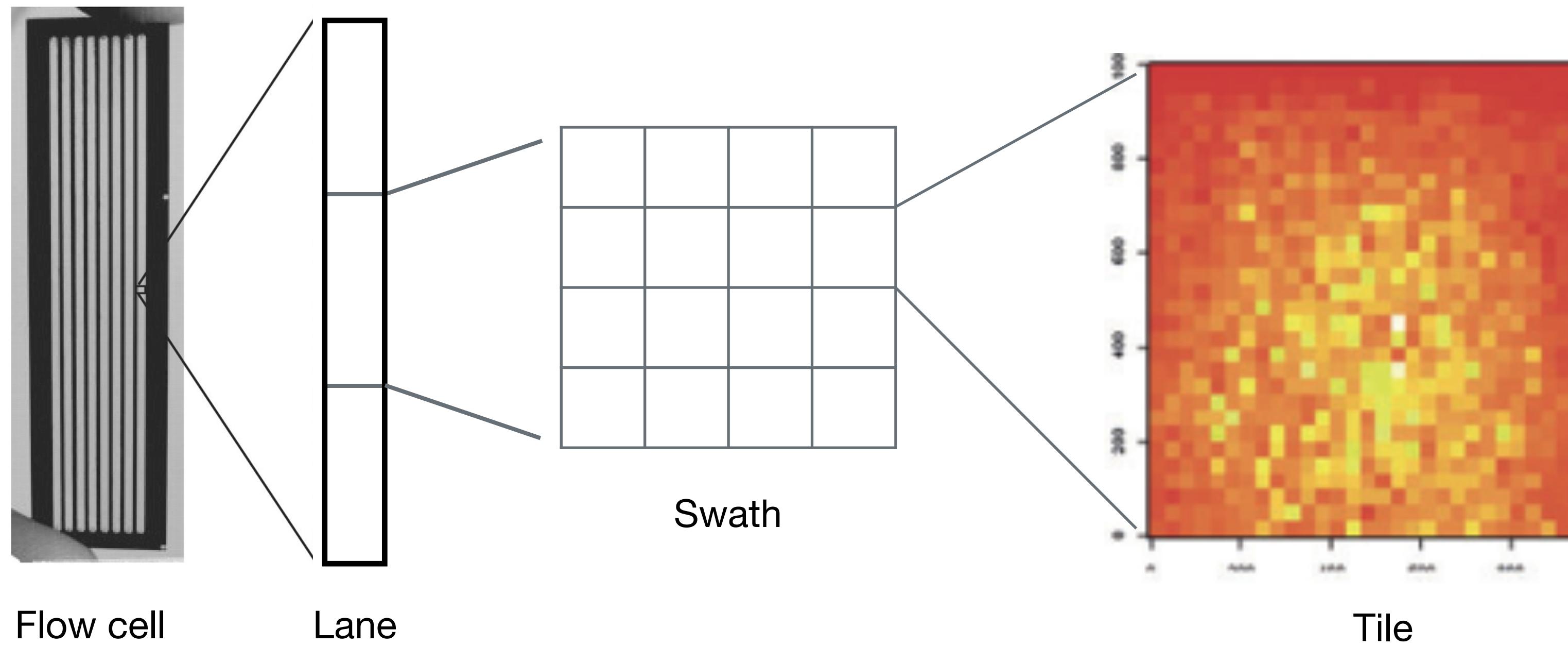
Illumina: phasing



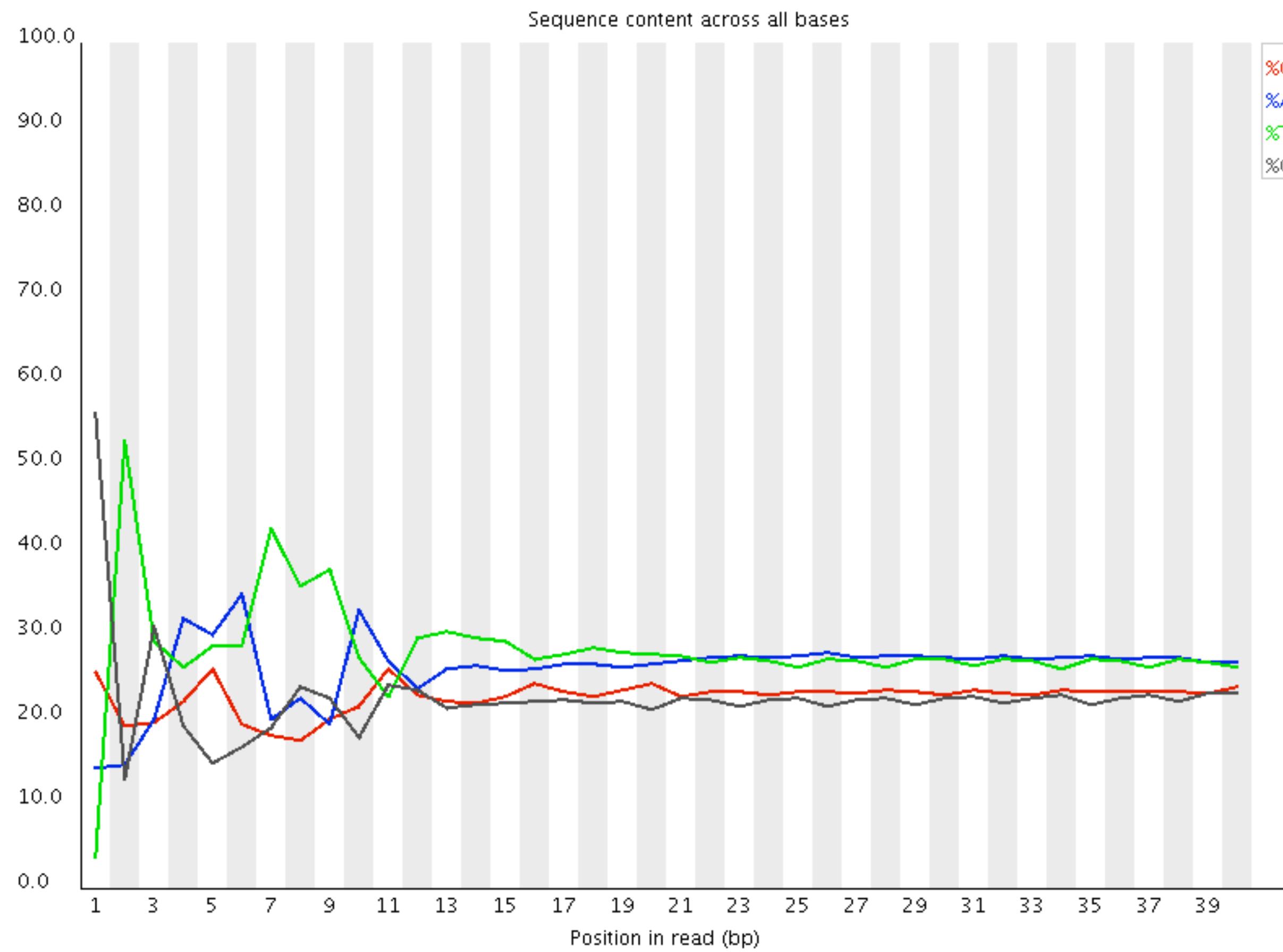
Illumina: cross-talk



Illumina: flow cell clusters

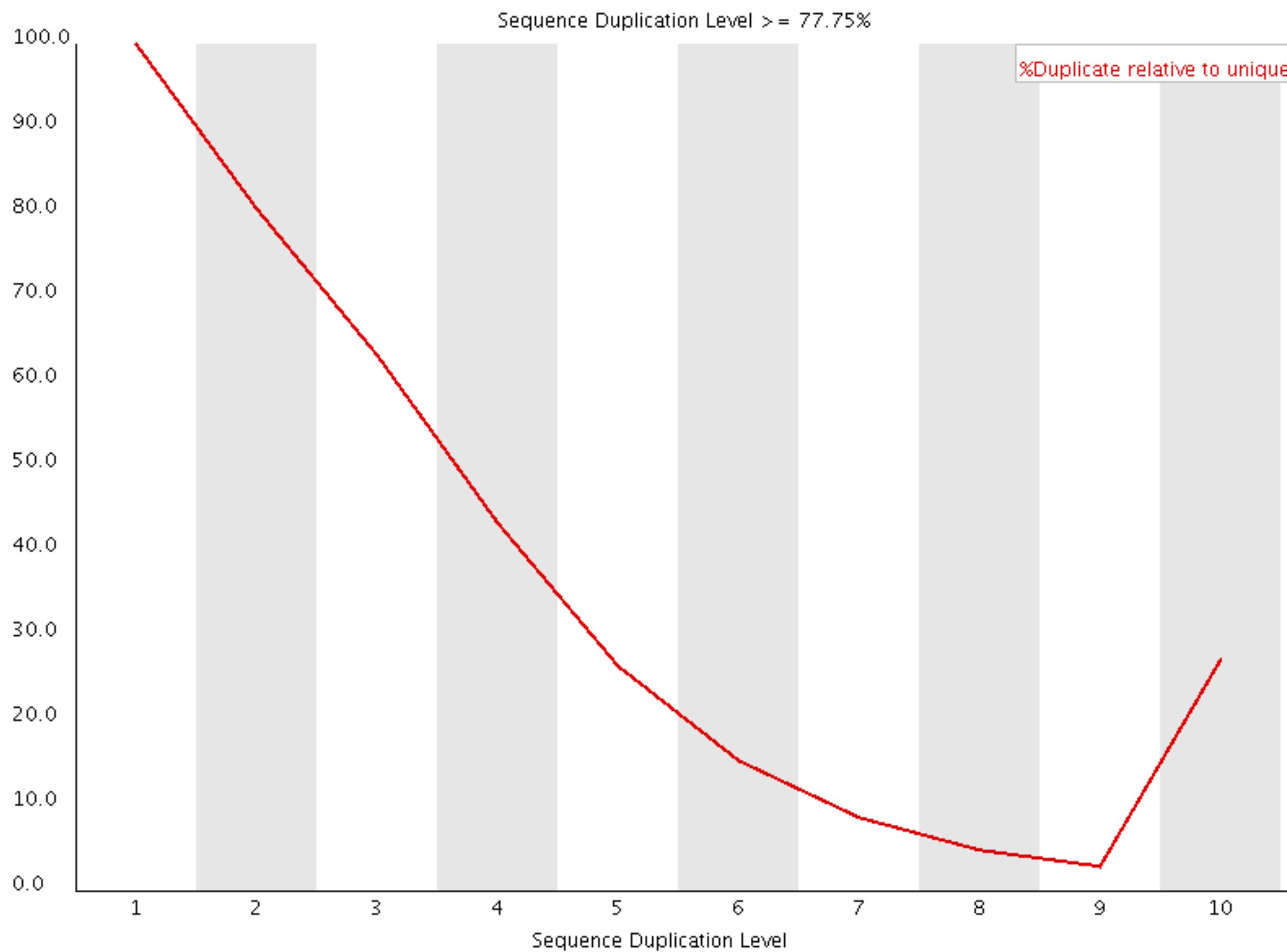


Illumina: optical effects



Positional sequence bias

PCR Artifacts



Duplicated sequences

Over-represented sequences

		sequence	count	lane
1051	A	AAAAAAAAAAAAAAAAAAAAA	70947	s_5_1_export.txt
451	A	AAAAAAAAAAAAAAAAAAAAA	69116	s_4_1_export.txt
601	A	AAAAAAAAAAAAAAAAAAAAA	66776	s_6_1_export.txt
301	A	AAAAAAAAAAAAAAAAAAAAA	63998	s_3_1_export.txt
751	A	AAAAAAAAAAAAAAAAAAAAA	55729	s_7_1_export.txt
151	A	AAAAAAAAAAAAAAAAAAAAA	54828	s_2_1_export.txt
901	A	AAAAAAAAAAAAAAAAAAAAA	40359	s_8_1_export.txt
1	A	NNNNNNNNNNNNNNNNNNN	30880	s_1_1_export.txt
152	A	NNNNNNNNNNNNNNNNNNN	30485	s_2_1_export.txt
153	C	NNNNNNNNNNNNNNNNNNN	26476	s_2_1_export.txt
2	T	NNNNNNNNNNNNNNNNNNN	25600	s_1_1_export.txt
154	G	NNNNNNNNNNNNNNNNNNN	25594	s_2_1_export.txt
3	C	NNNNNNNNNNNNNNNNNNN	25063	s_1_1_export.txt
155	T	NNNNNNNNNNNNNNNNNNN	24965	s_2_1_export.txt
4	G	NNNNNNNNNNNNNNNNNNN	24164	s_1_1_export.txt
302	A	NNNNNNNNNNNNNNNNNNN	22501	s_3_1_export.txt
5	A	AAAAAAAAAAAAAAAAAAAAA	20996	s_1_1_export.txt
452	T	NNNNNNNNNNNNNNNNNNN	20842	s_4_1_export.txt

Filtering

	sequence	count
1	ATTAACCCTCACTAAAGGGACTAGTCCTGCAGGTTAACGAATTGCC	482185
151	ATTAACCCTCACTAAAGGGACTAGTCCTGCAGGTTAACGAATTGCC	271724
2	TAATACGACTCACTATAAGGCGAATTGAATTAGC GGCCCGAATTGCC	159936
152	TAATACGACTCACTATAAGGCGAATTGAATTAGC GGCCCGAATTGCC	105273
153	CTTAACCCTCACTAAAGGGACTAGTCCTGCAGGTTAACGAATTGCC	46872
3	CTTAACCCTCACTAAAGGGACTAGTCCTGCAGGTTAACGAATTGCC	43212
4	NN	13142

Read Frequency Distribution

Contamination

> gnl|uv|NGB00105.1:1-219 pCR4-TOPO multiple cloning site
Length=219

Score = 100 bits (50), Expect = 9e-19
Identities = 50/50 (100%), Gaps = 0/50 (0%)
Strand=Plus/Plus

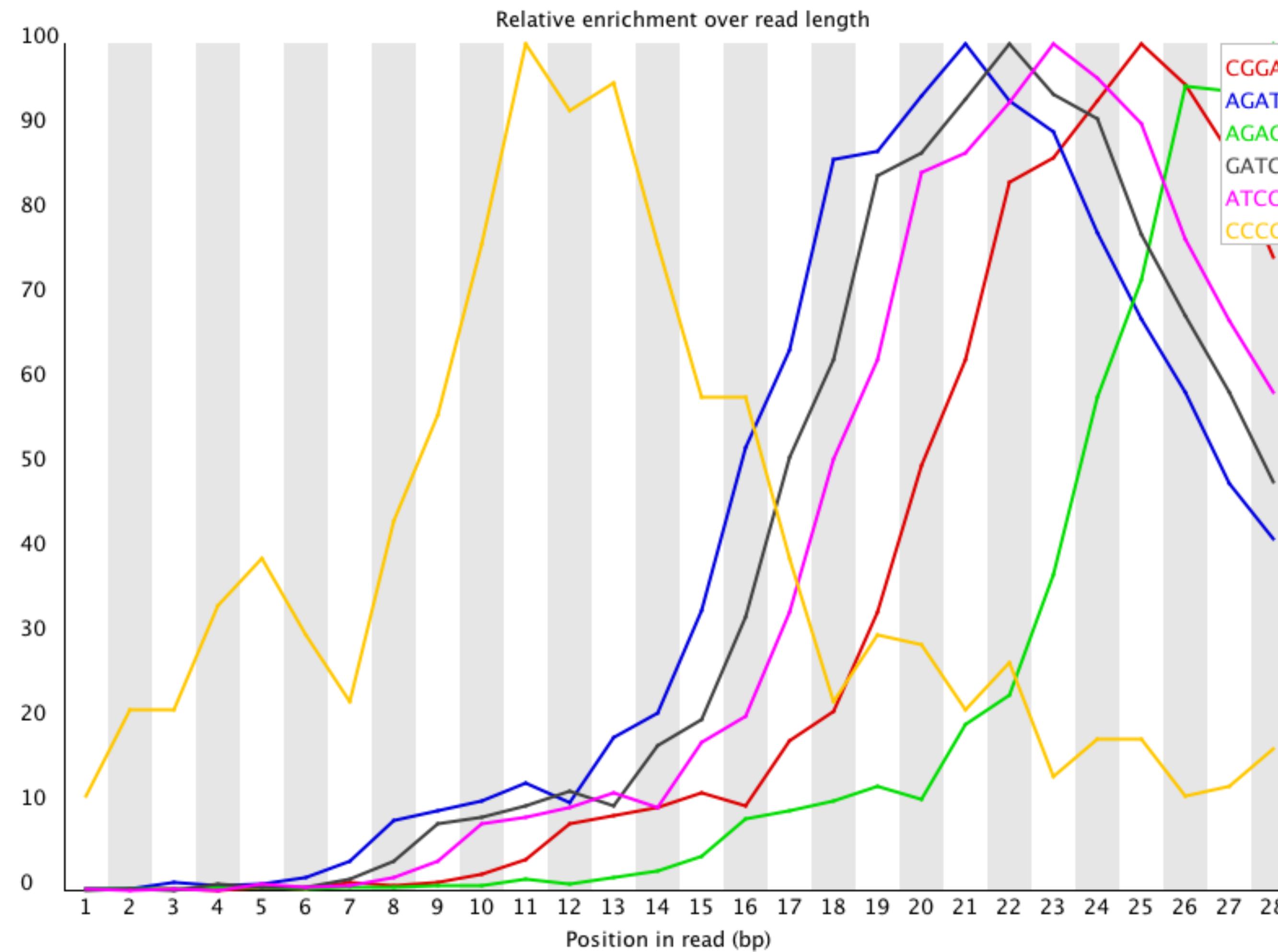
Query 1

ATTAACCCTCACTAAAGGGACTAGTCCTGCAGGTTAACGAATTGCC 50

|||||||||||||||||||||||||||||||||||||||||||||||

Sbjct 43

ATTAACCCTCACTAAAGGGACTAGTCCTGCAGGTTAACGAATTGCC 92



Adaptor contamination

These materials have been developed by members of the teaching team at the Harvard Chan Bioinformatics Core (HBC). These are open access materials distributed under the terms of the Creative Commons Attribution license (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

