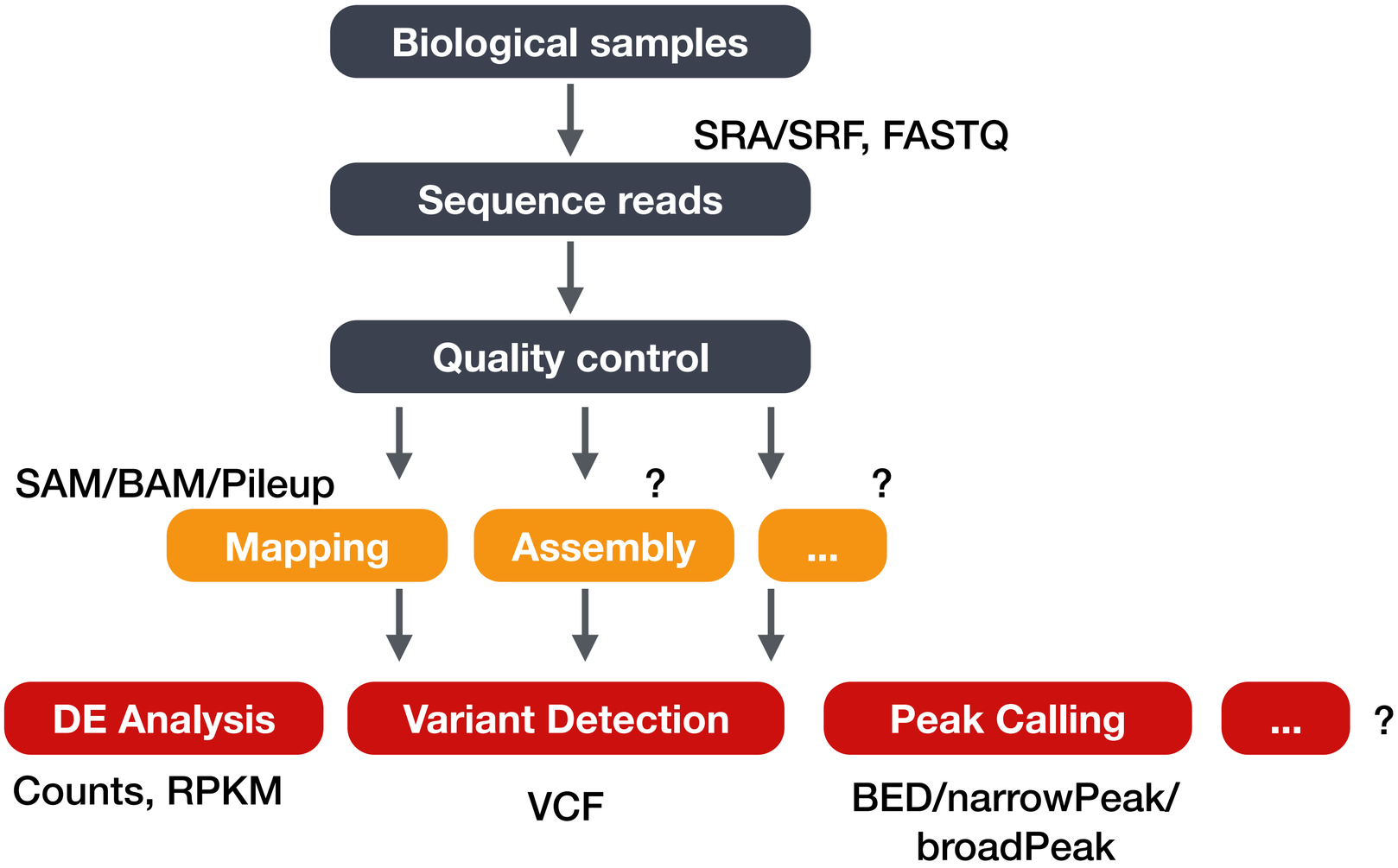
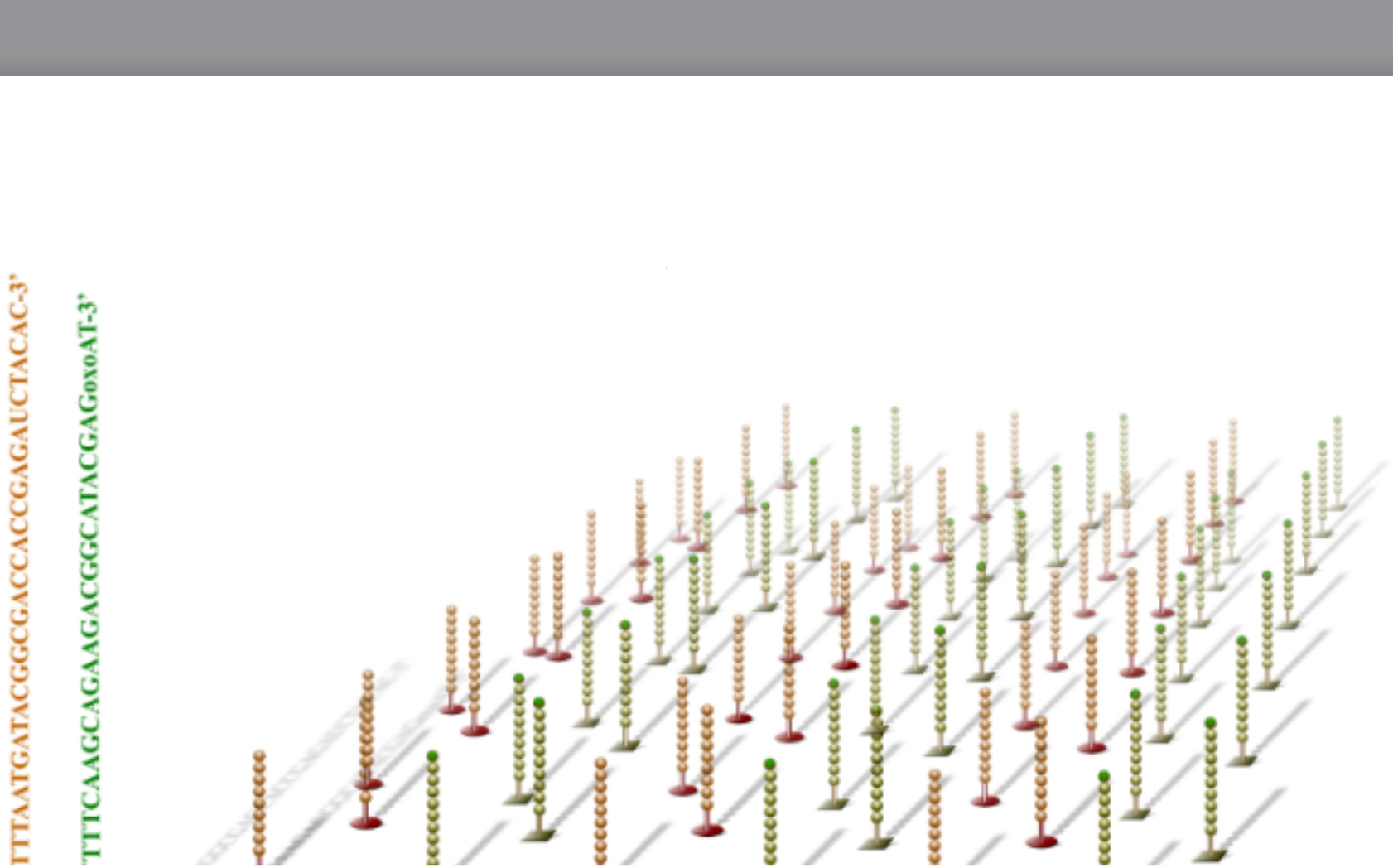
Workflows and Data Standards for Next-Generation Sequencing Data



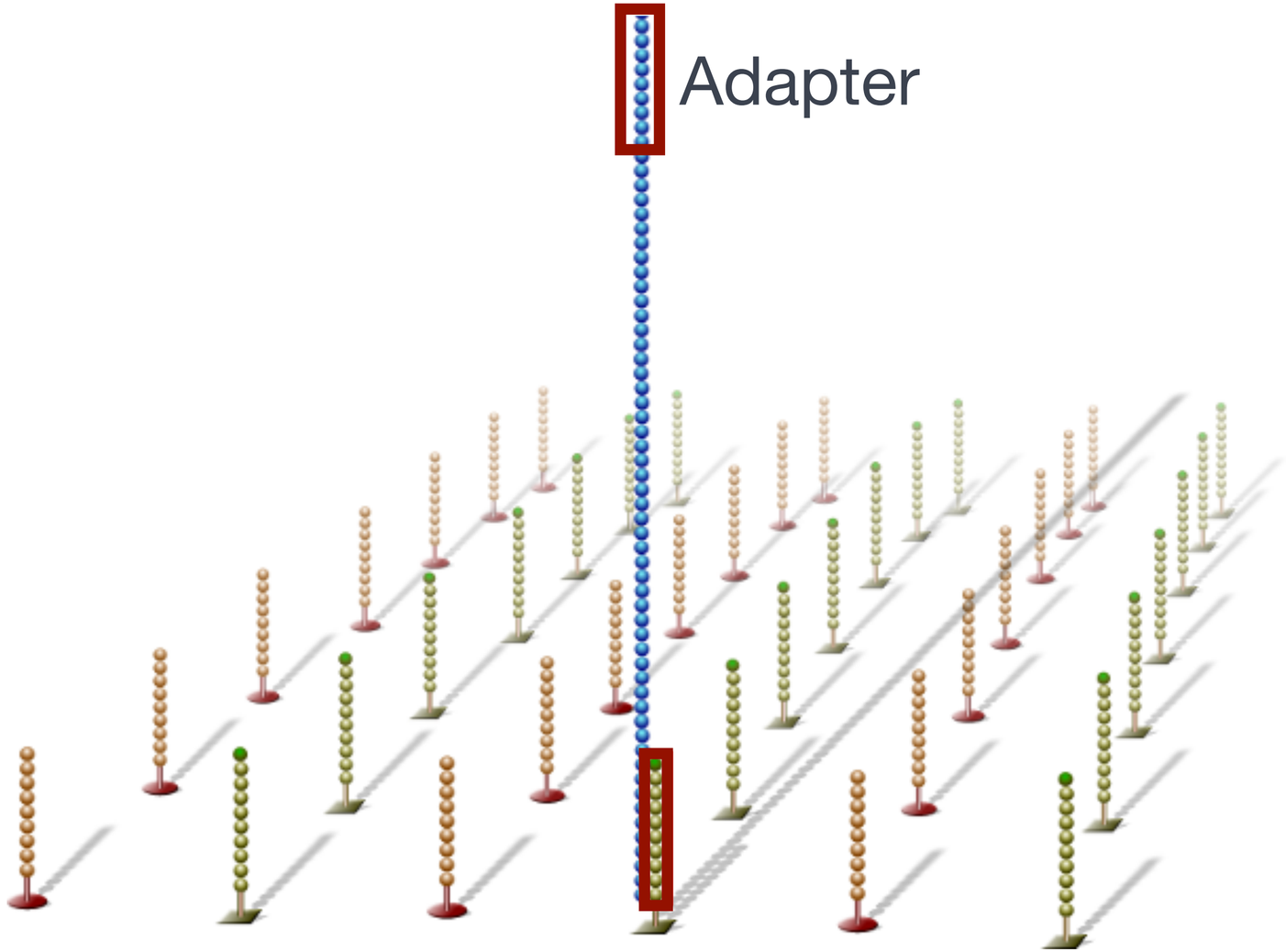
Workflows and Data Standards

Illumina sequencing



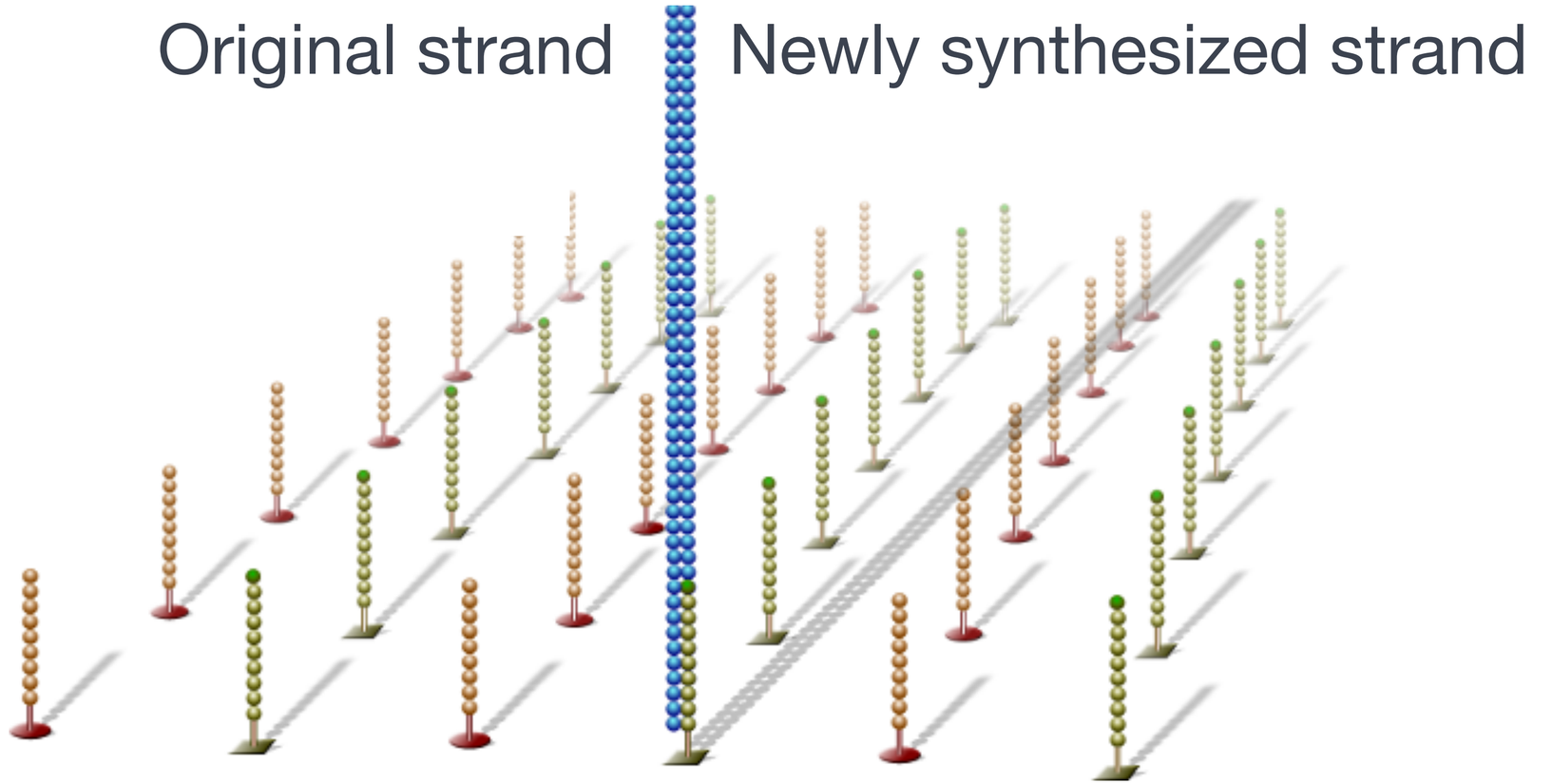
Illumina: flow cell

http://training.bioinformatics.ucdavis.edu/docs/2014/09/september-2014-workshop/Monday_JF_HTS_lecture.html



Illumina: cluster generation

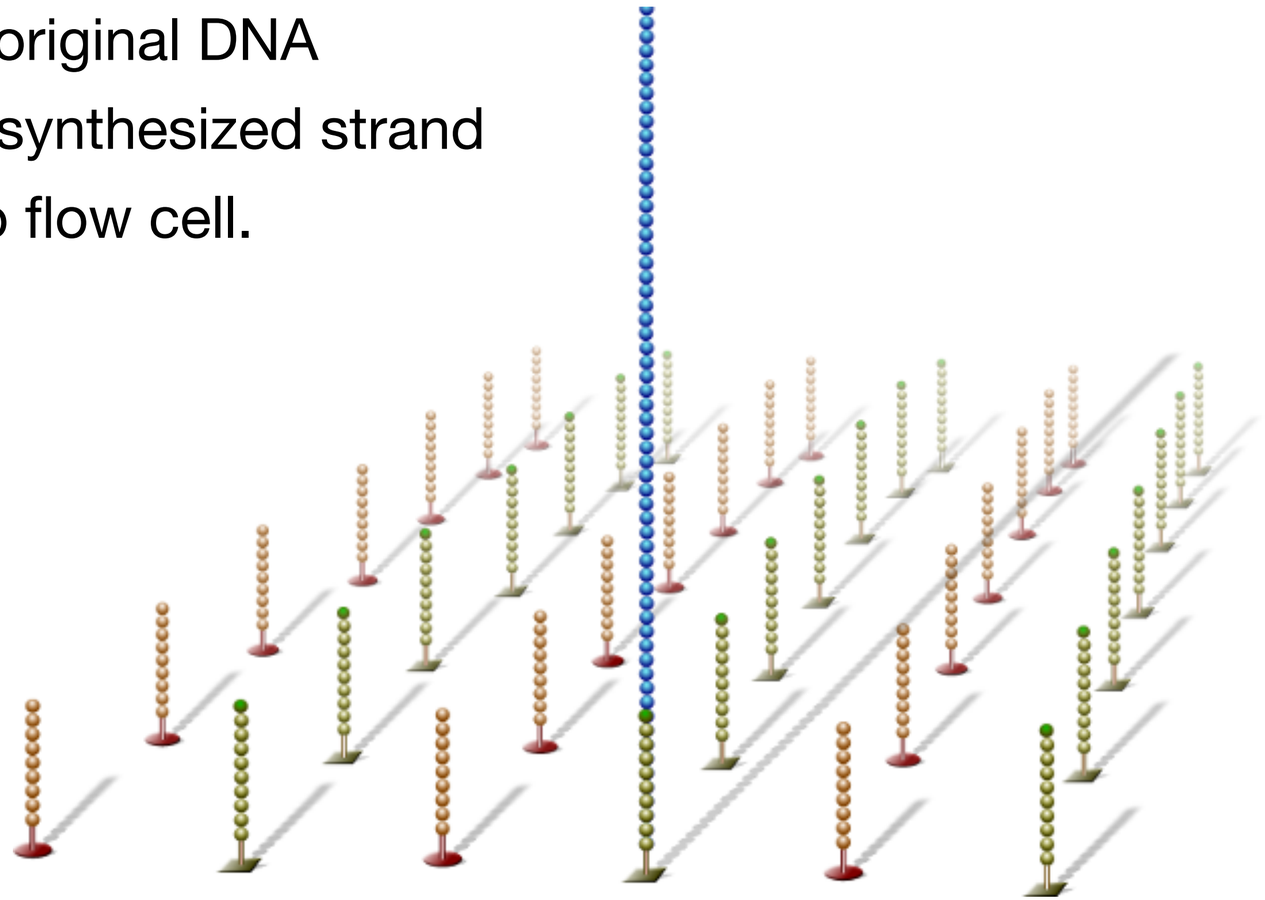
http://training.bioinformatics.ucdavis.edu/docs/2014/09/september-2014-workshop/Monday_JF_HTS_lecture.html



Illumina: cluster generation

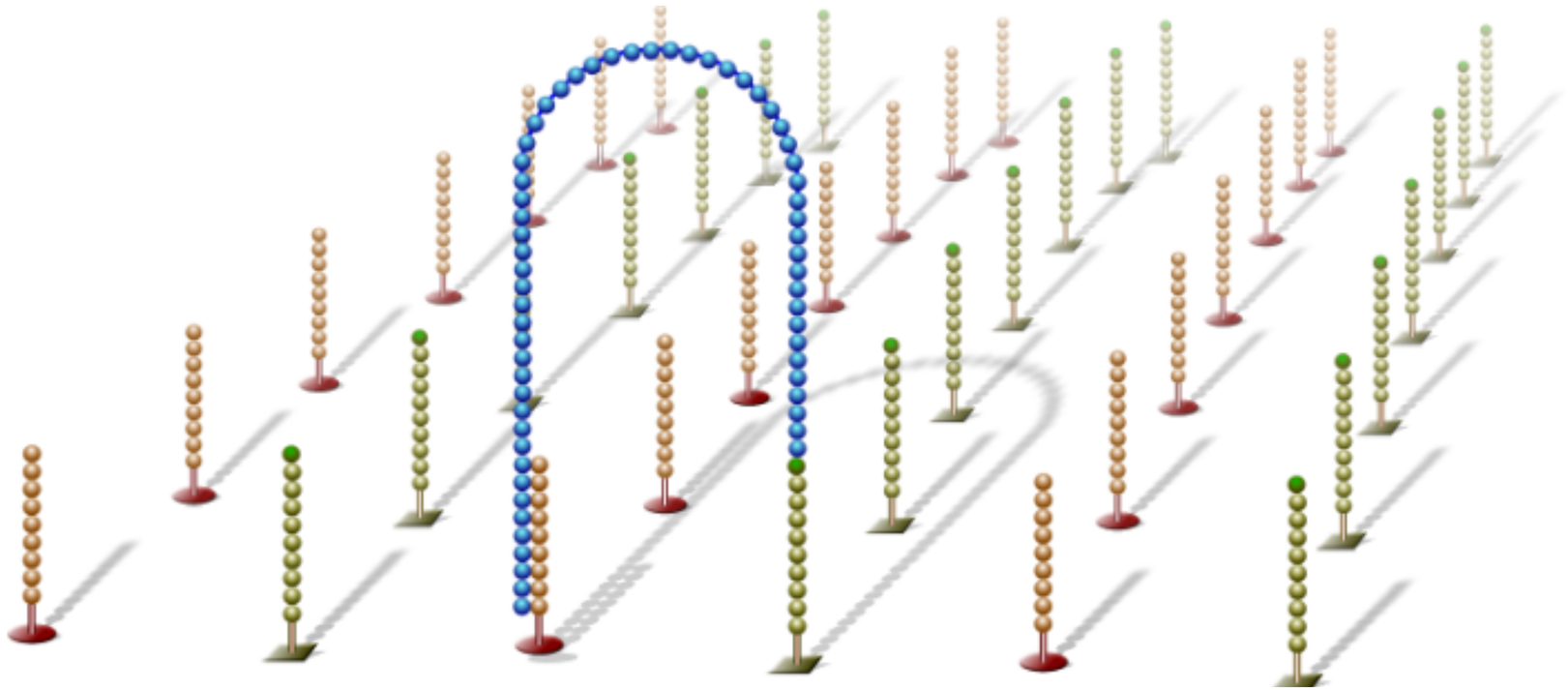
http://training.bioinformatics.ucdavis.edu/docs/2014/09/september-2014-workshop/Monday_JF_HTS_lecture.html

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



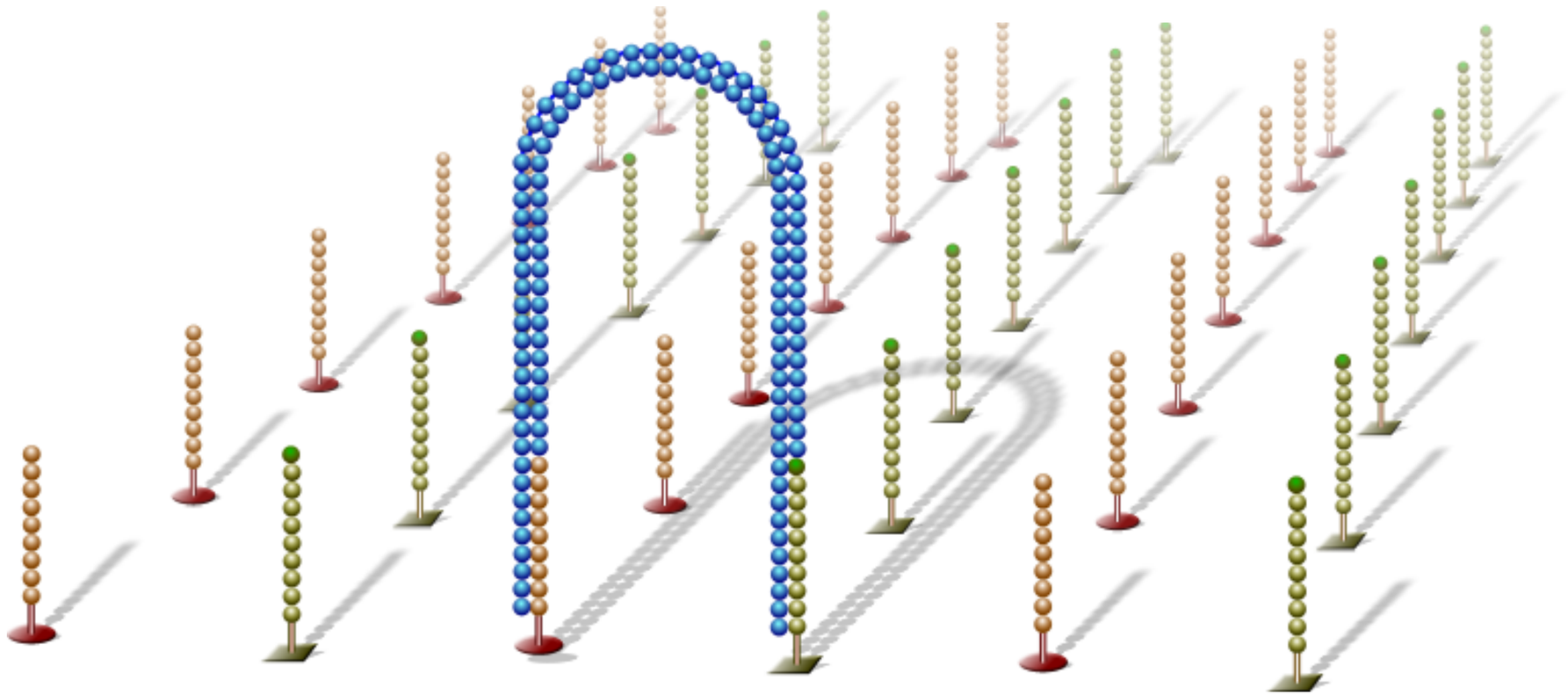
Illumina: cluster generation

http://training.bioinformatics.ucdavis.edu/docs/2014/09/september-2014-workshop/Monday_JF_HTS_lecture.html



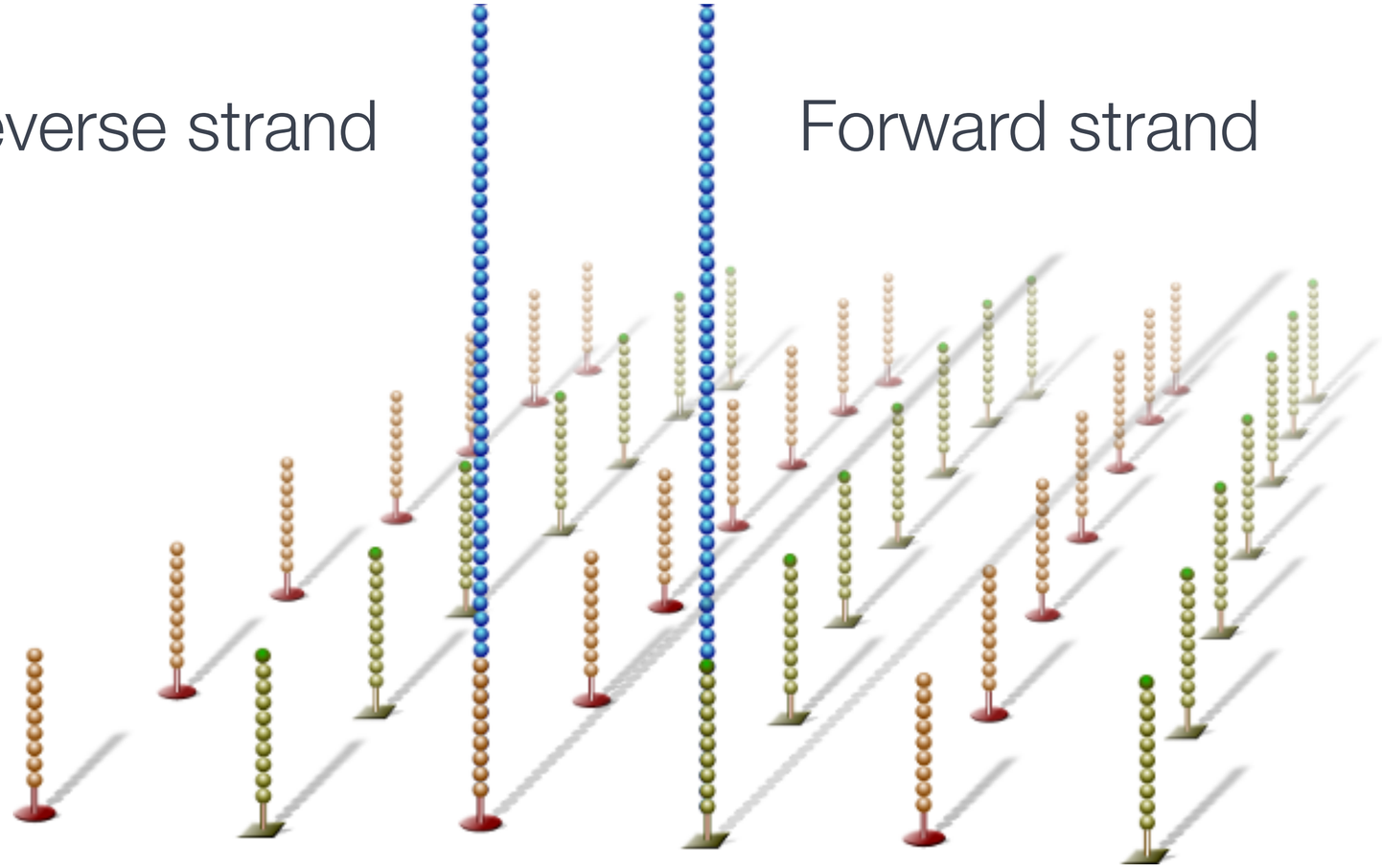
Illumina: bridge amplification

http://training.bioinformatics.ucdavis.edu/docs/2014/09/september-2014-workshop/Monday_JF_HTS_lecture.html



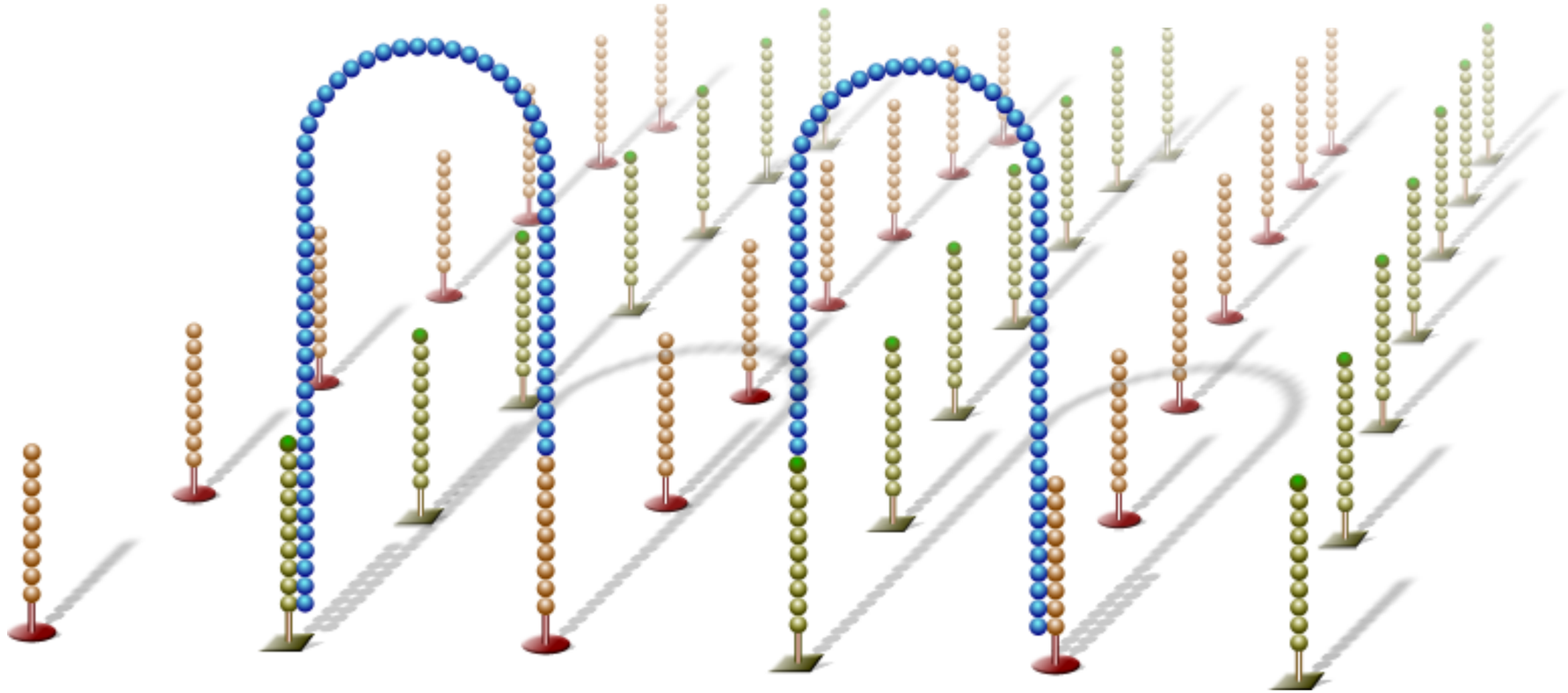
Illumina: bridge amplification

http://training.bioinformatics.ucdavis.edu/docs/2014/09/september-2014-workshop/Monday_JF_HTS_lecture.html



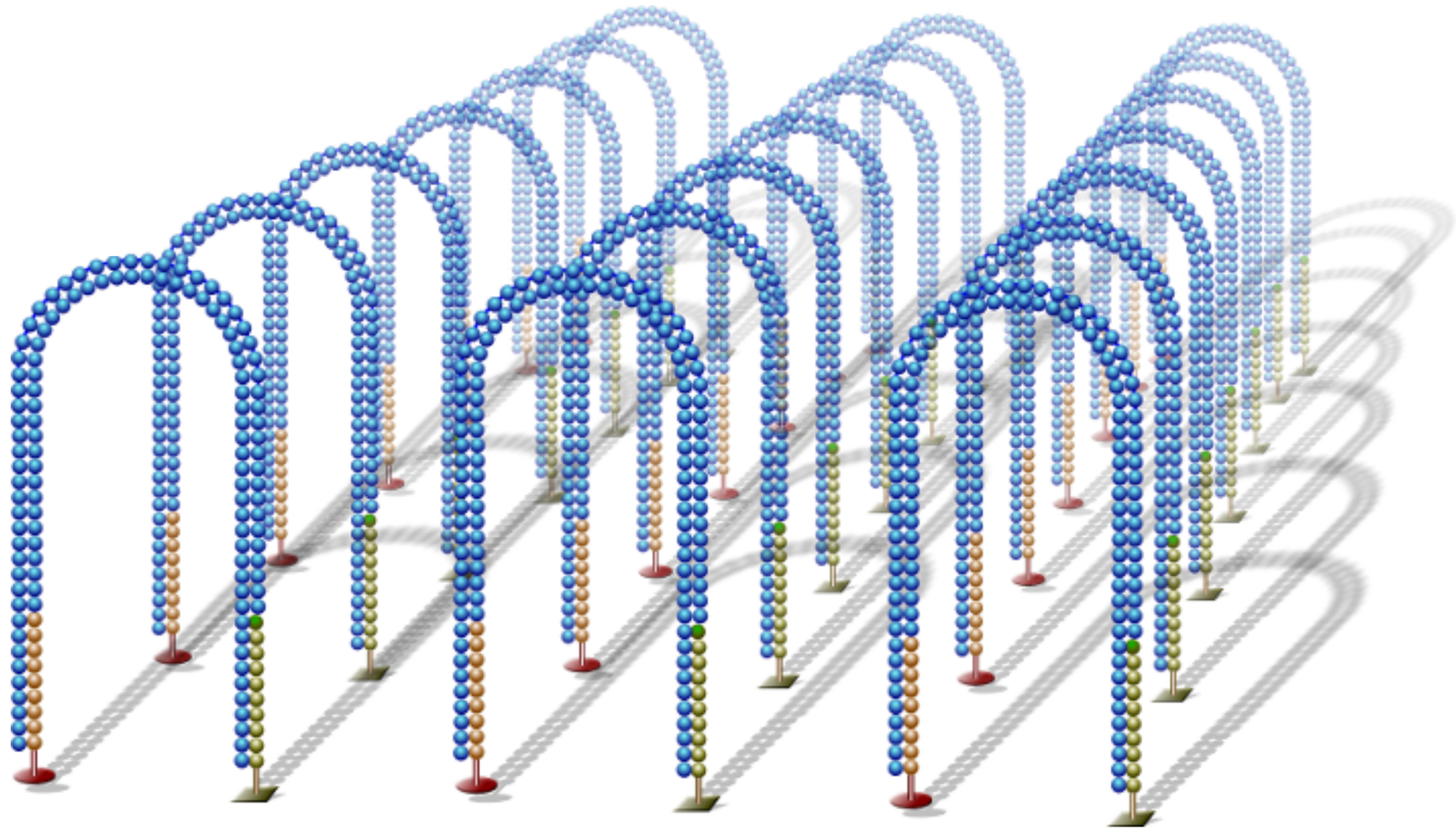
Illumina: bridge amplification

http://training.bioinformatics.ucdavis.edu/docs/2014/09/september-2014-workshop/Monday_JF_HTS_lecture.html



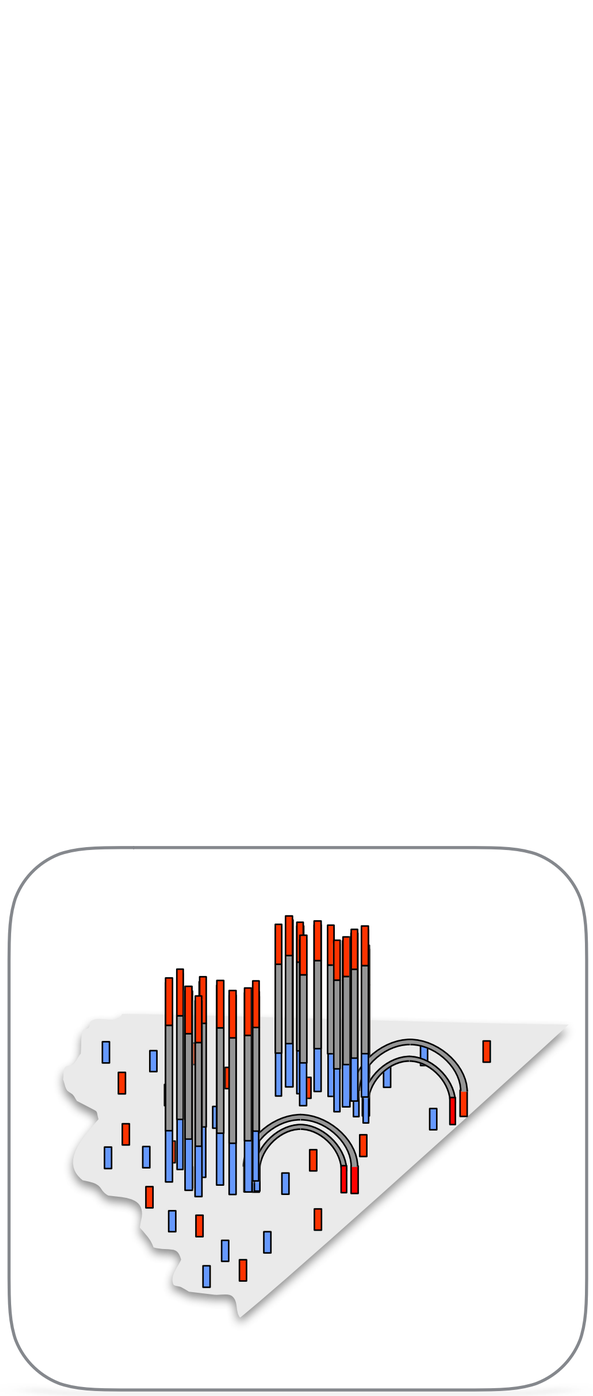
Illumina: bridge amplification

http://training.bioinformatics.ucdavis.edu/docs/2014/09/september-2014-workshop/Monday_JF_HTS_lecture.html

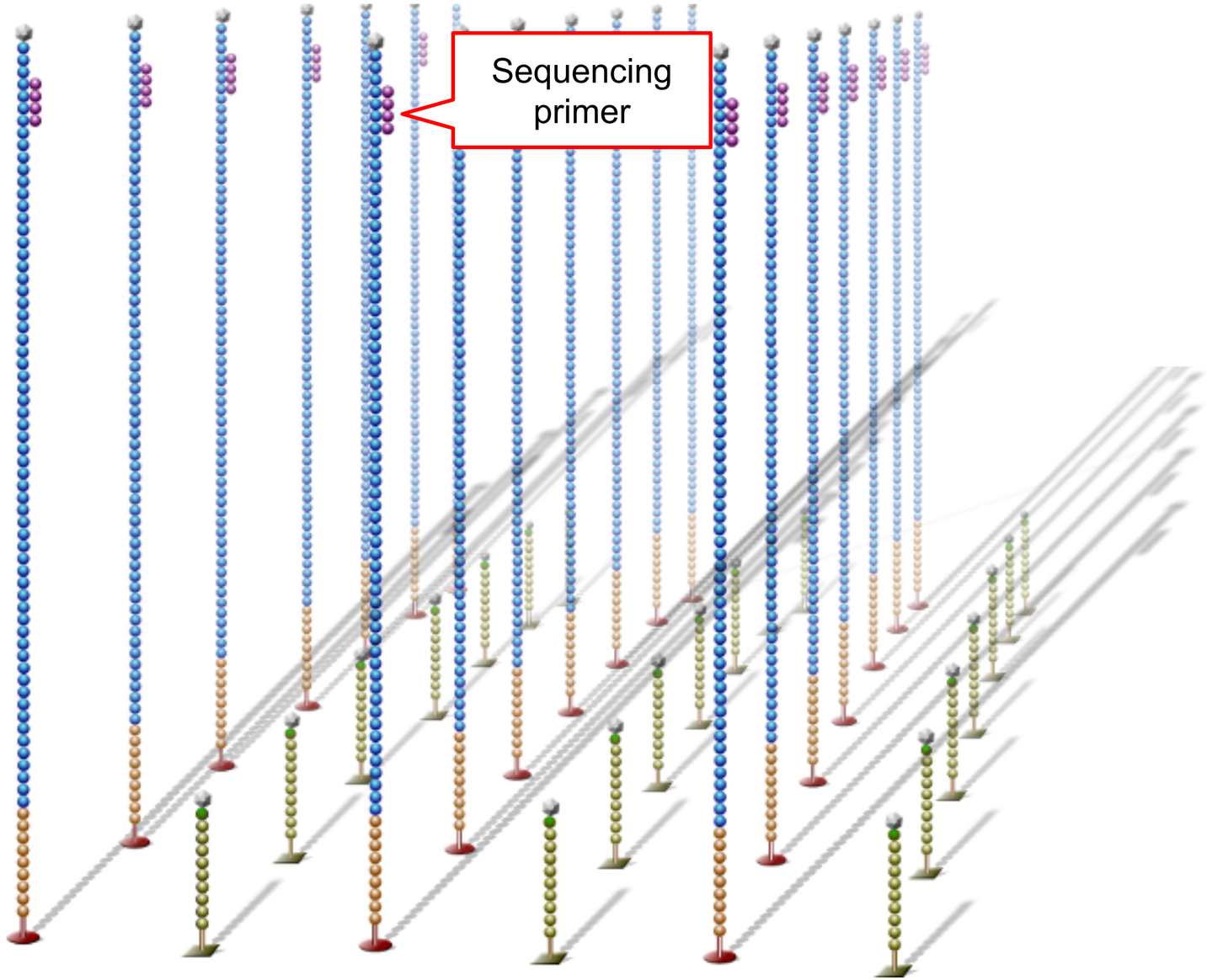


Illumina: bridge amplification

http://training.bioinformatics.ucdavis.edu/docs/2014/09/september-2014-workshop/Monday_JF_HTS_lecture.html

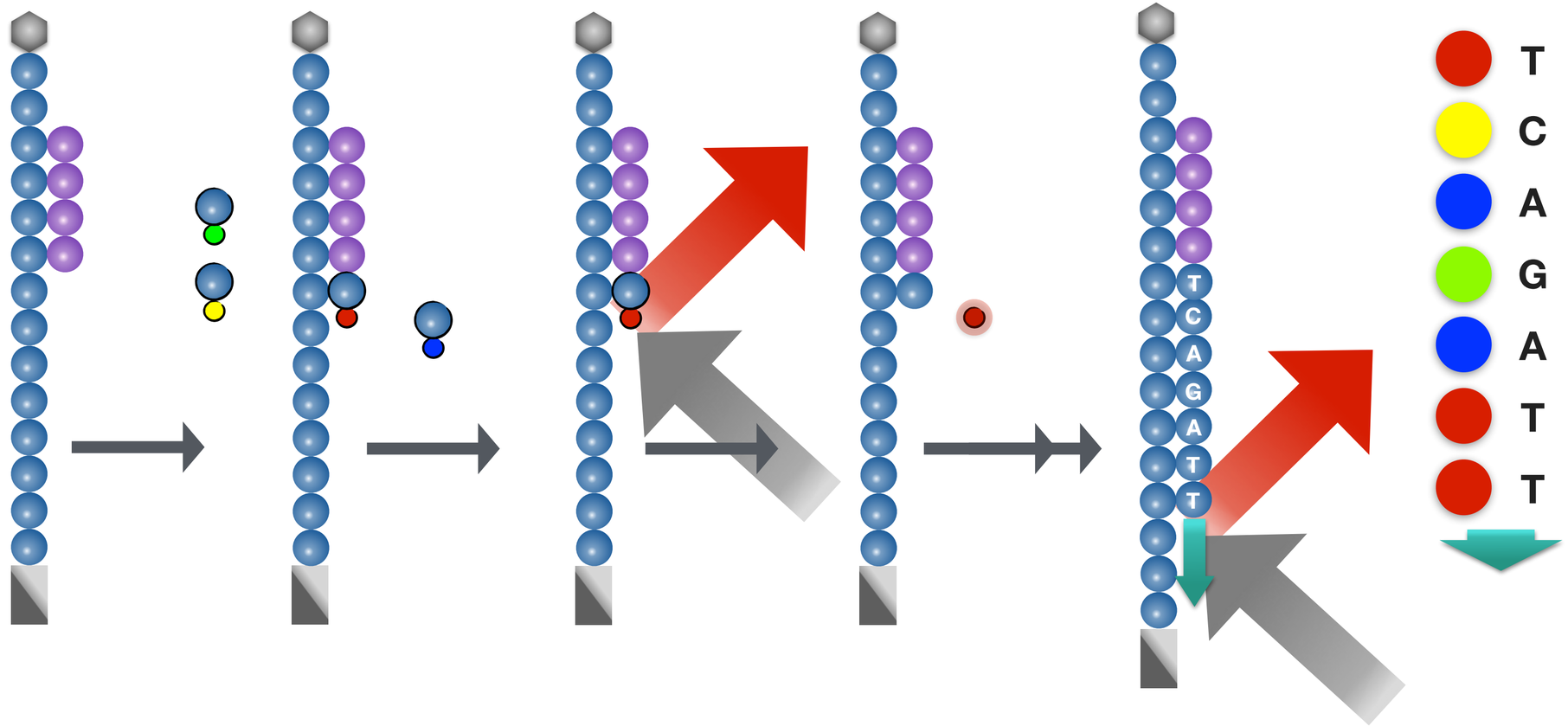


Illumina: cluster generation

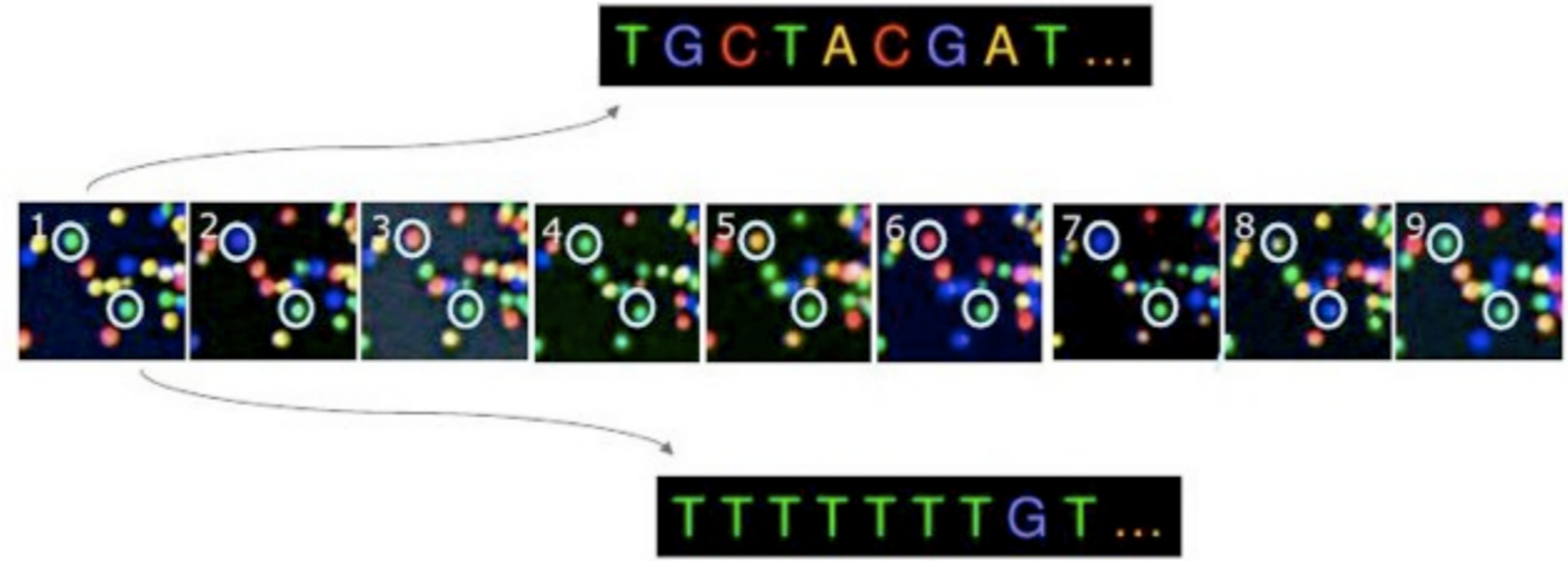


Illumina: Prepare for sequencing

http://training.bioinformatics.ucdavis.edu/docs/2014/09/september-2014-workshop/Monday_JF_HTS_lecture.html



Illumina: sequencing by synthesis



Illumina: base calling

http://www.slideshare.net/CRS4/chris-jones-crs4-staff-meeting-24032010

Off the sequencer

FASTA

```
>SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGGGTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAAAGCAATGCCAATA
```

```
>gi|340780744|ref|NC_015850.1| Acidithiobacillus caldus SM-1 chromosome, complete genome
ATGAGTAGTCATTCAGCGCCGACAGCGTTGCAAGATGGAGCCGCGCTGTGGTCCGCCCTATGCGTCCAACTGGAGCTCGTCACGAG
TCCGCAGCAGTTCAATACCTGGCTGCGGCCCCTGCGTGGCGAATTGCAGGGTCATGAGCTGCGCCTGCTCGCCCCCAATCCCTTCG
TCCGCGACTGGGTGCGTGAACGCATGGCCGAACTCGTCAAGGAACAGCTGCAGCGGATCGCTCCGGGTTTTGAGCTGGTCTTCGCT
CTGGACGAAGAGGCAGCAGCGGCGACATCGGCACCGACCGCGAGCATTGCGCCCGAGCGCAGCAGCGCACCCGGTGGTCACCGCCT
CAACCCAGCCTTCAACTTCCAGTCCTACGTCGAAGGGAAGTCCAATCAGCTCGCCCTGGCGGCAGCCCGCCAGGTTGCCCAGCATC
CAGGCAAATCCTACAACCCACTGTACATTTATGGTGGTGTGGGCCTCGGCAAGACGCACCTCATGCAGGCCGTGGGCAACGATATC
CTGCAGCGGCAACCCGAGGCCAAGGTGCTCTATATCAGCTCCGAAGGCTTCATCATGGATATGGTGCGCTCGCTGCAACACAATAC
CATCAACGACTTCAAACAGCGTTATCGCAAGCTGGACGCCCTGCTCATCGACGACATCCAGTTCTTTGCGGGCAAGGACCGCACCC
```

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

FASTQ: FASTA with Quality scores

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGGGTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAAAGCAATGCCAATA
+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 Quailty Encoding

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGGGTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAAAGCAATGCCAATA
+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 Quailty Encoding

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGGGTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAAAGCAATGCCAATA
+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
```

```
Q = -10 x log10(P), where P is the probability that a base call is erroneous
```

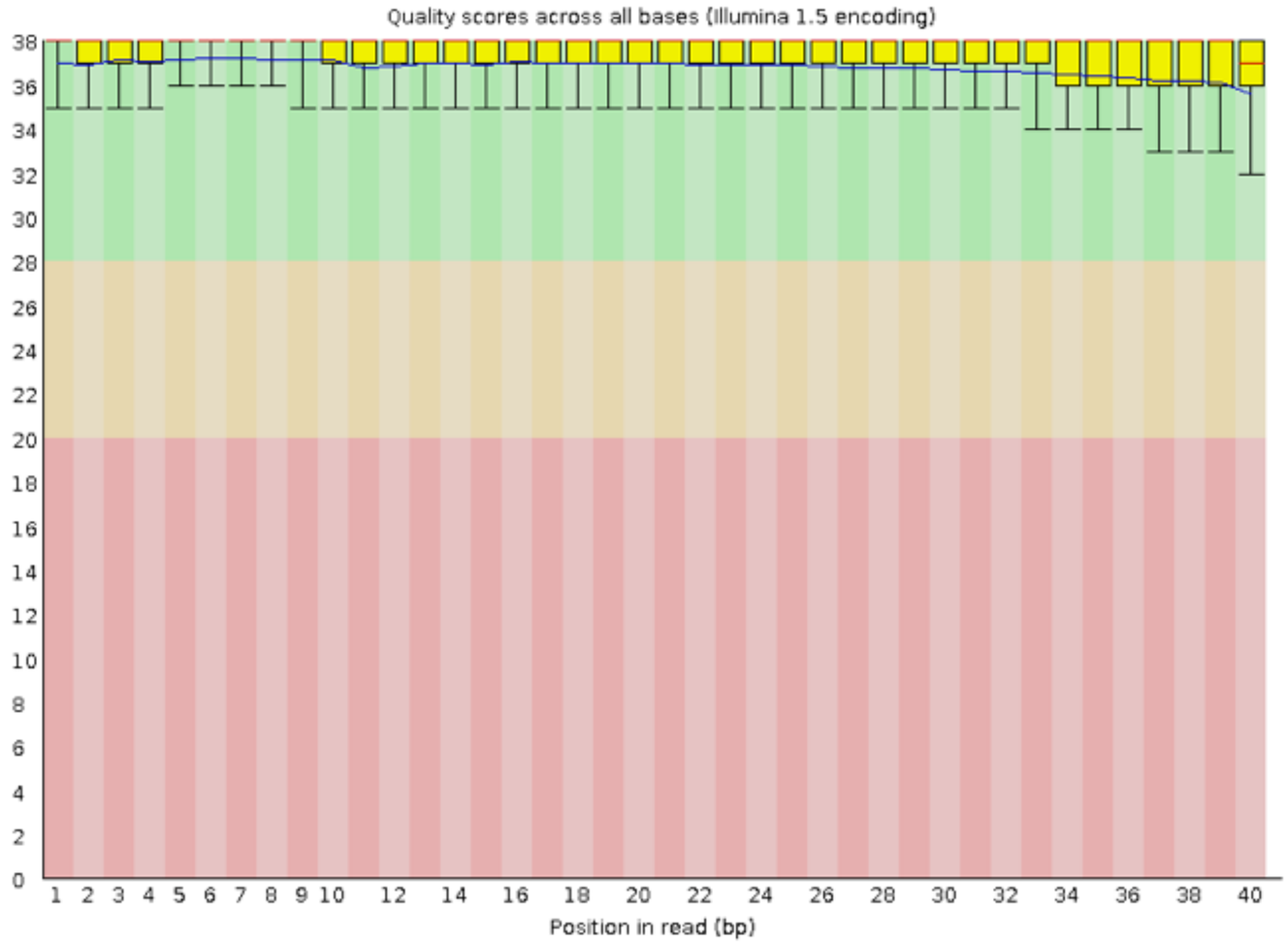
The legend above provides the mapping of quality scores (Phred-33) to the quality encoding characters.

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

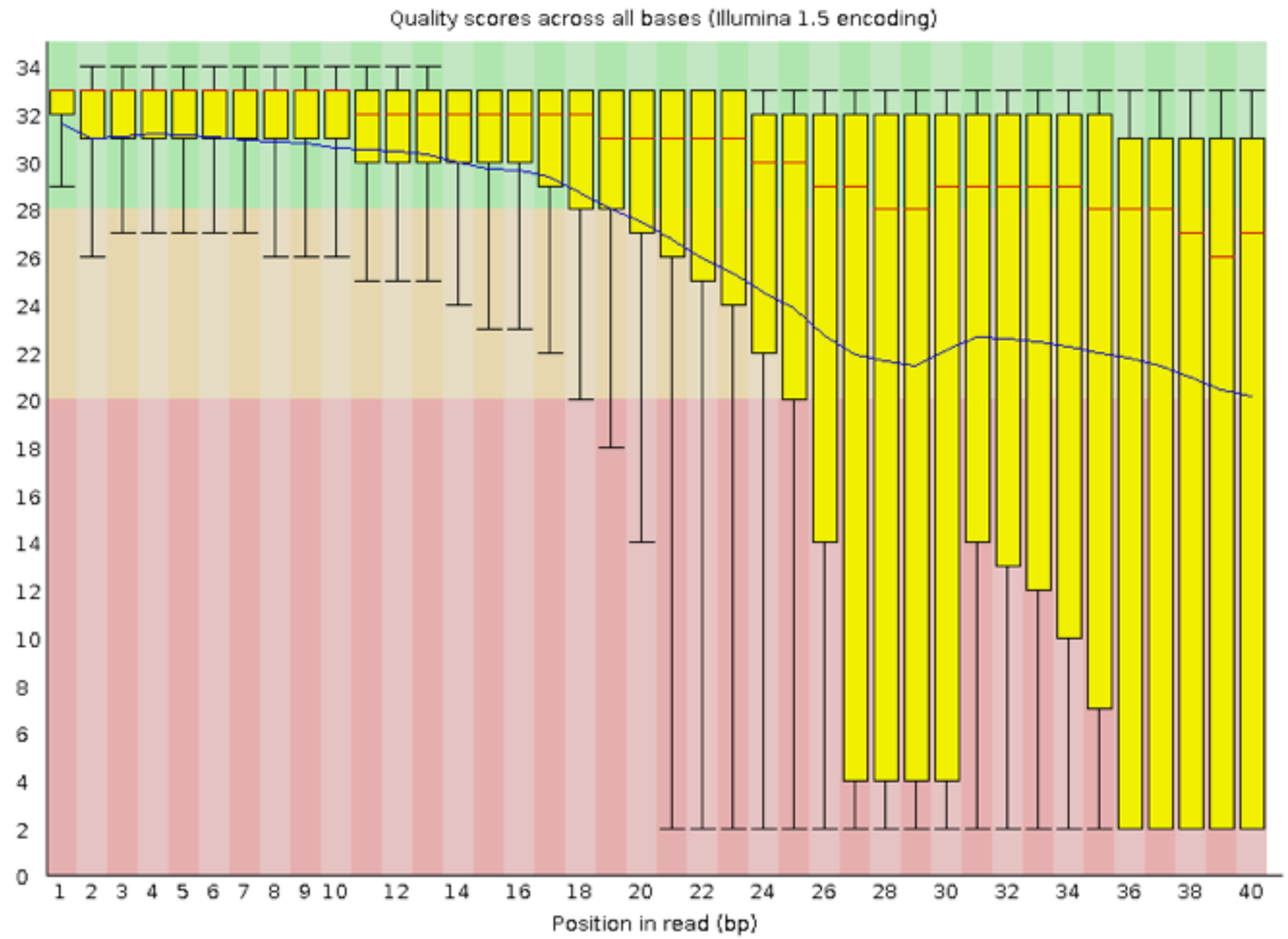
FASTQ Quality Scores

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%



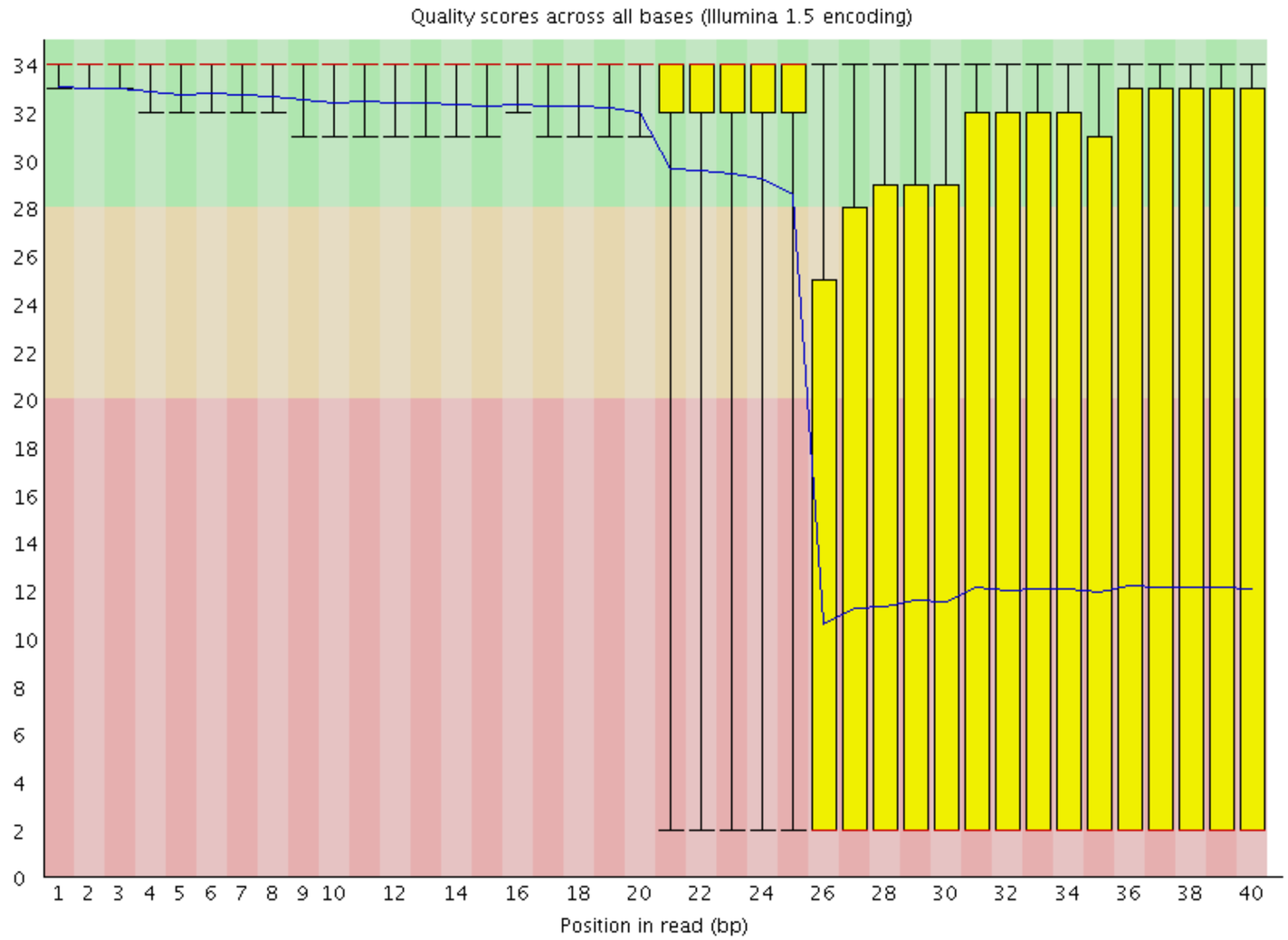
A good quality sample



A not-so-good quality sample

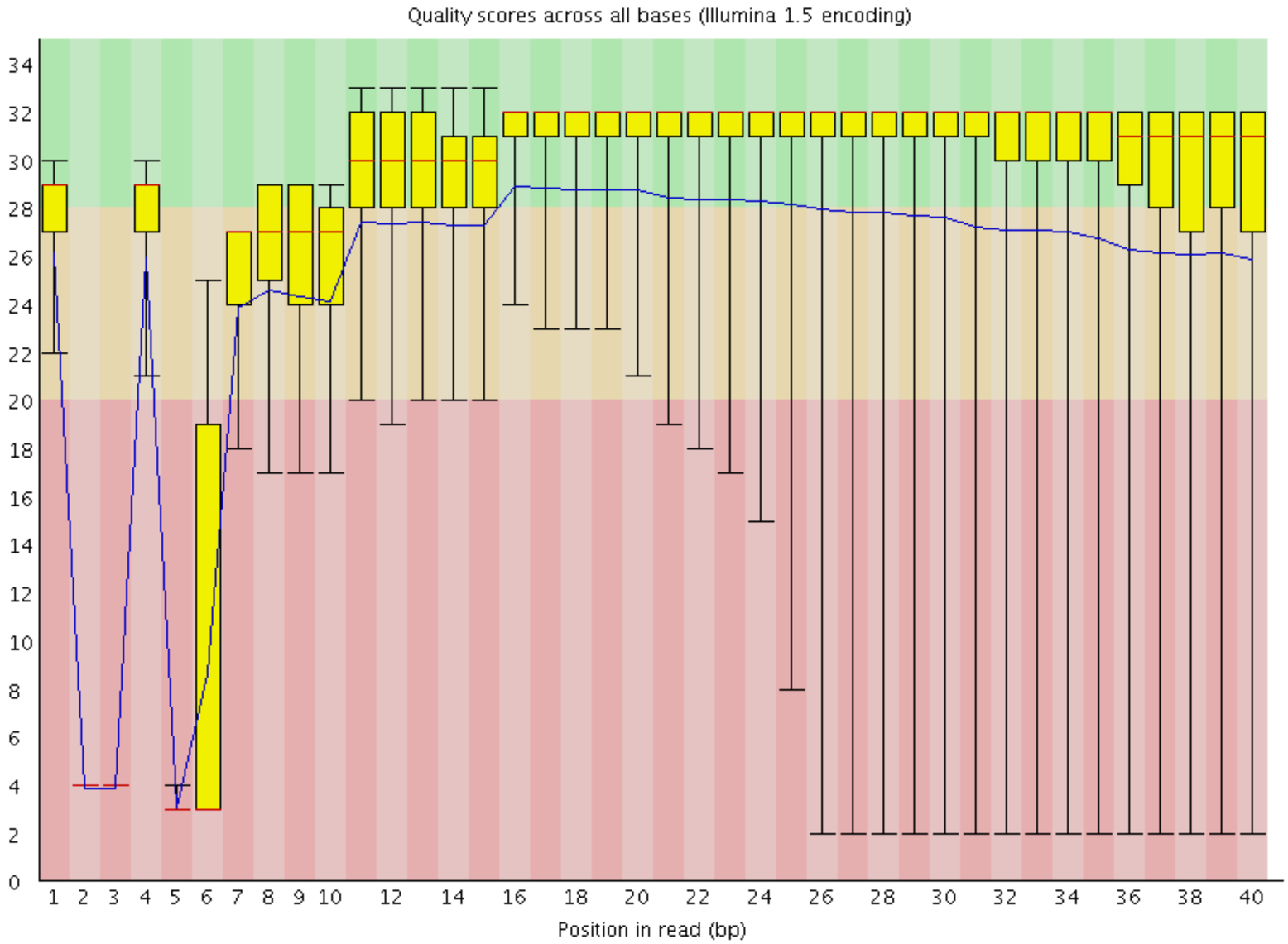
Error profiles:
Technical Sequencer Problems

Manifold burst in cycle 26

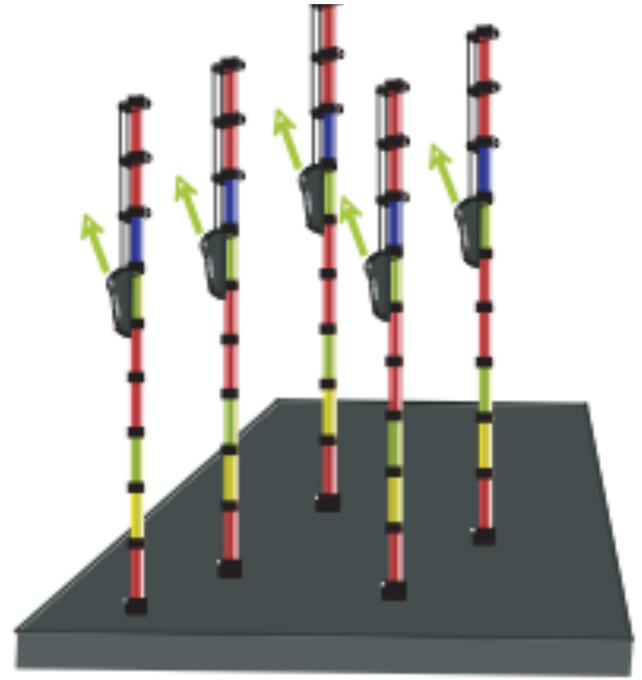


See http://bioinfo-core.org/index.php/9th_Discussion-28_October_2010 for more example

Specific cycles lost

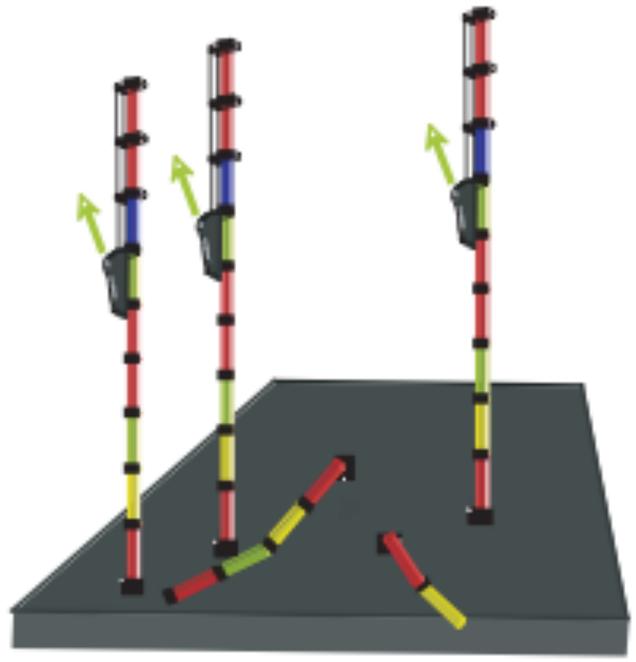


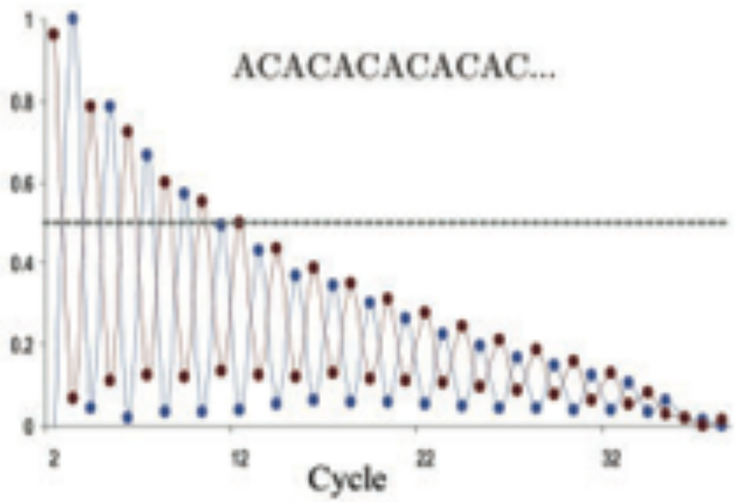
Error dependency on technology



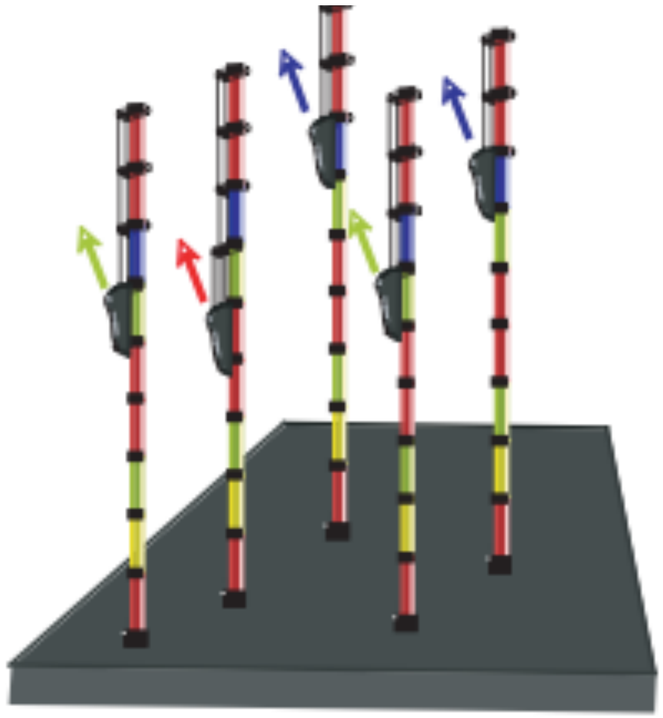
Illumina

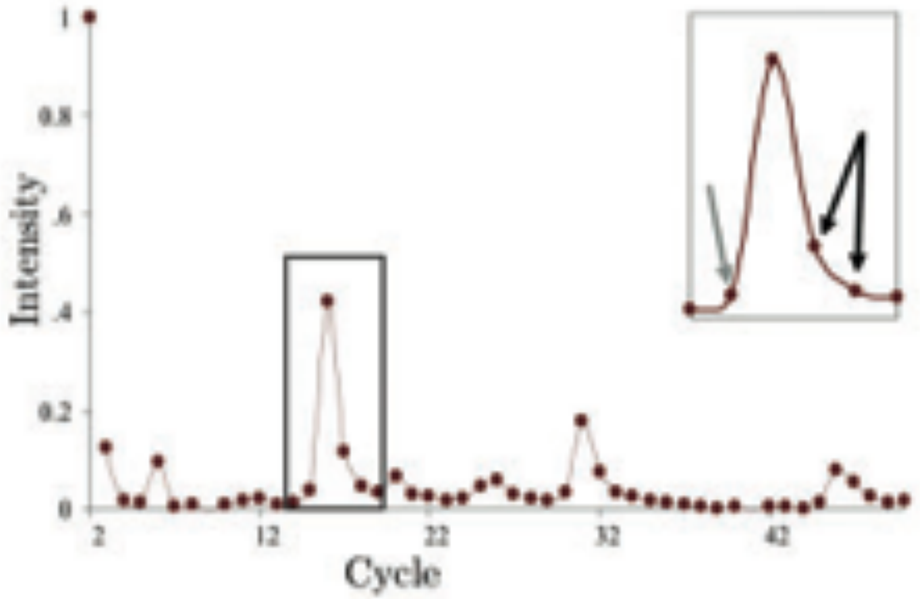
Base-calling for next-generation sequencing platforms. Brief Bioinform 2011, 12(5):489-497



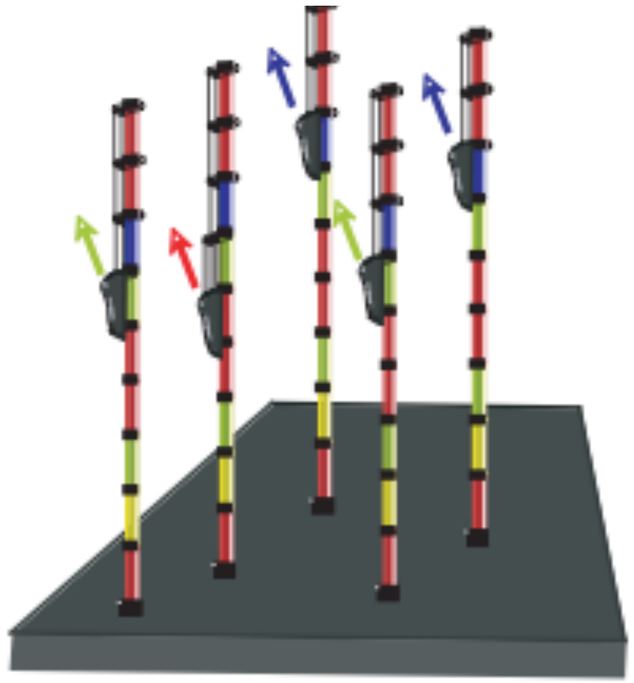


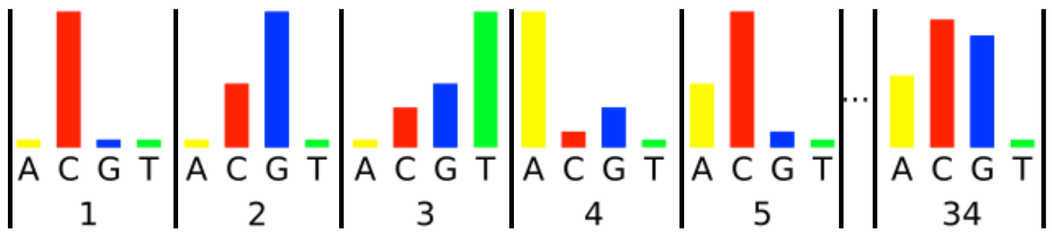
Illumina: signal decay



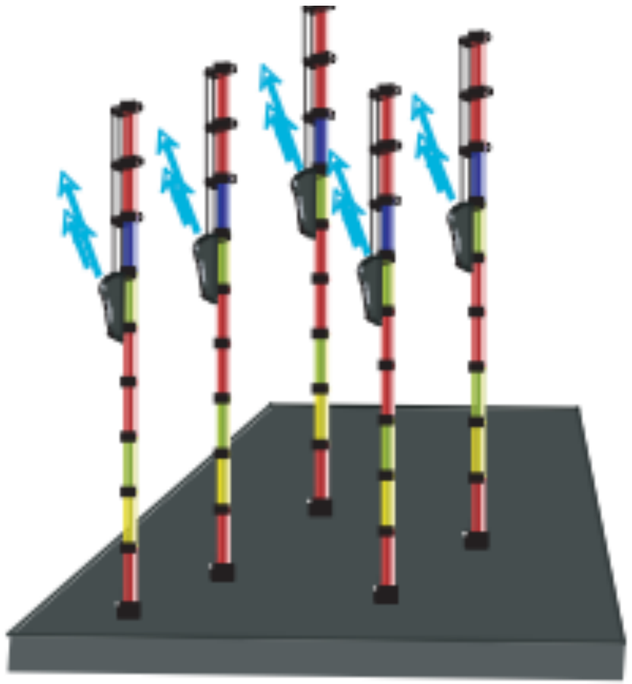


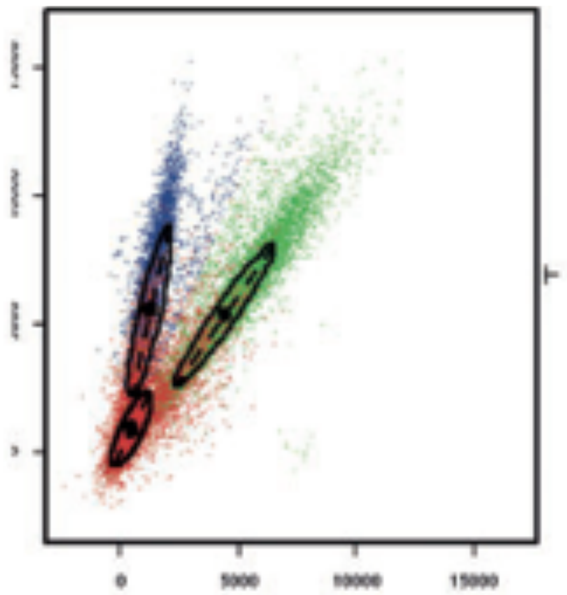
Illumina: phasing



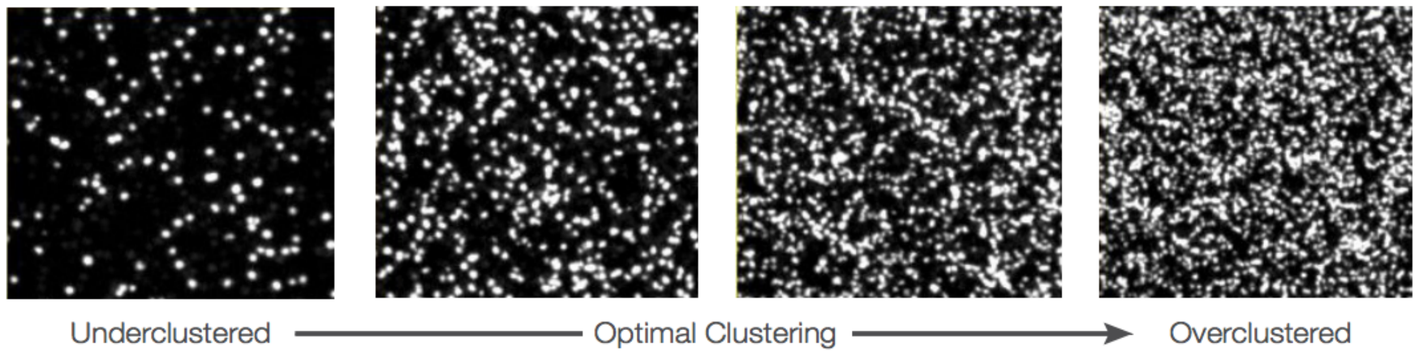


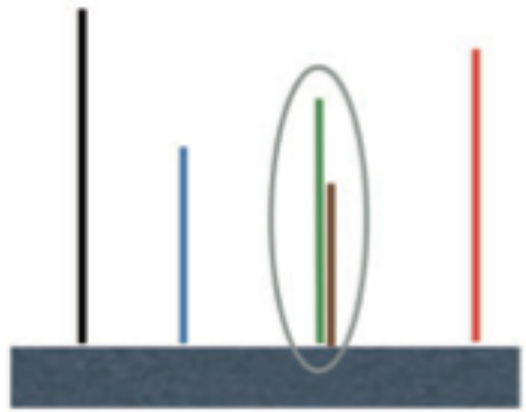
Illumina: phasing



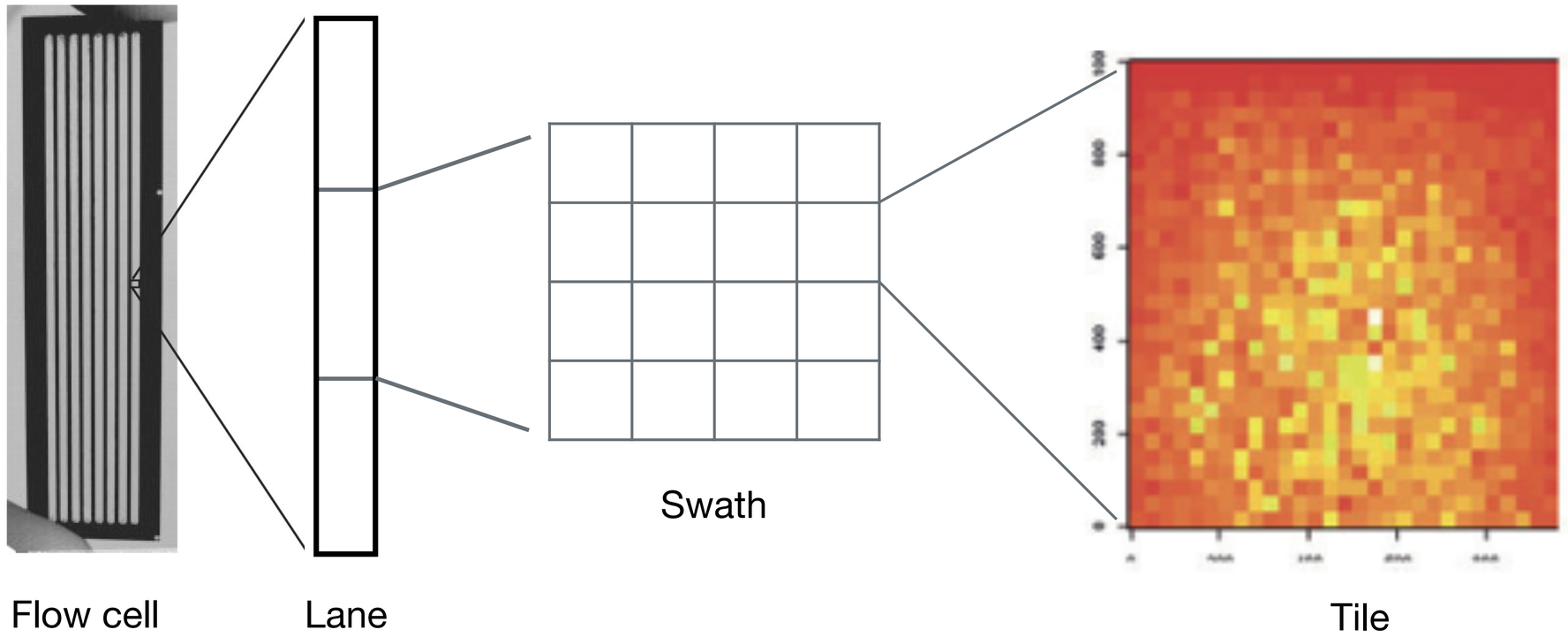


Illumina: cross-talk

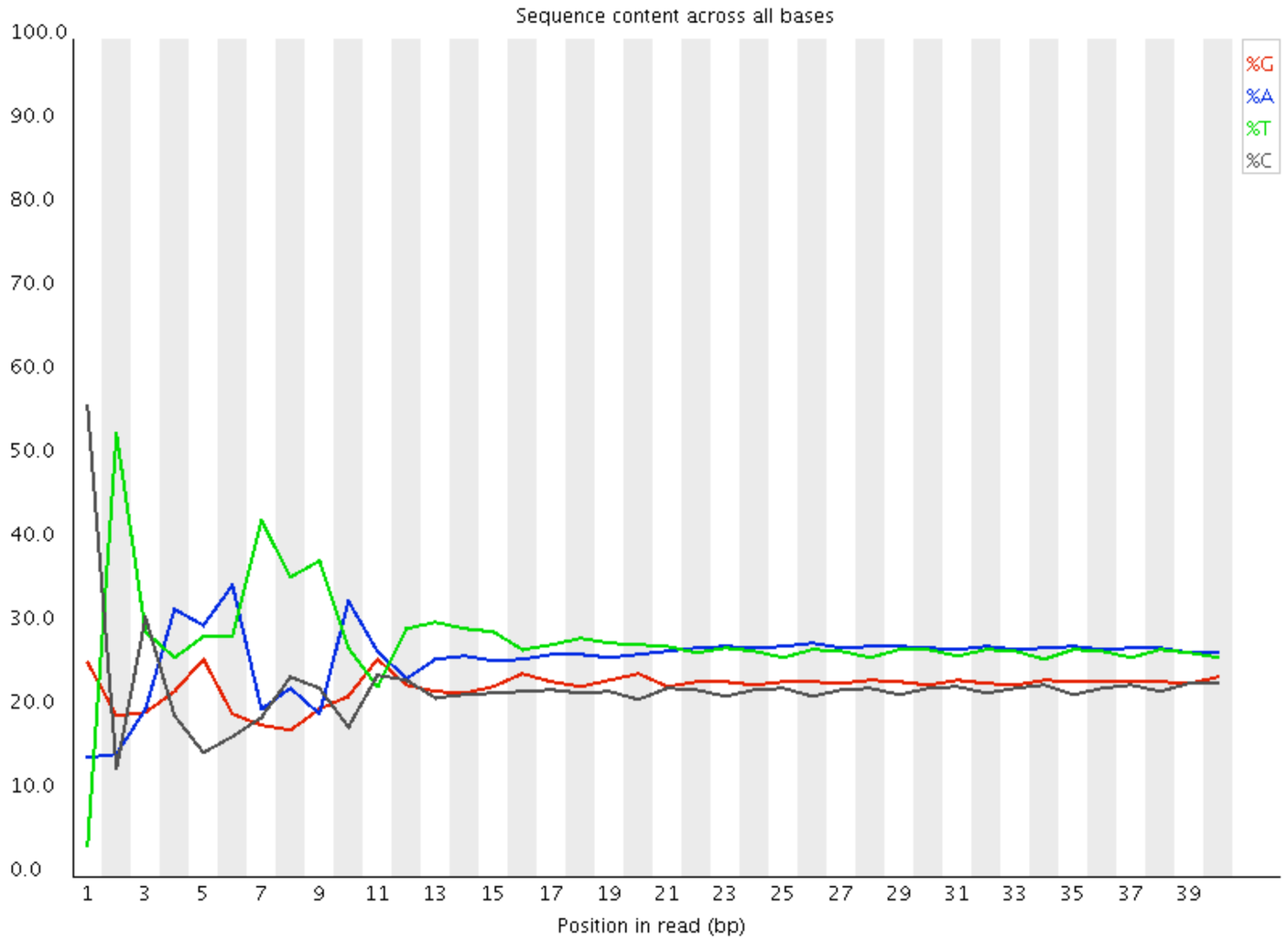




Illumina: flow cell clusters

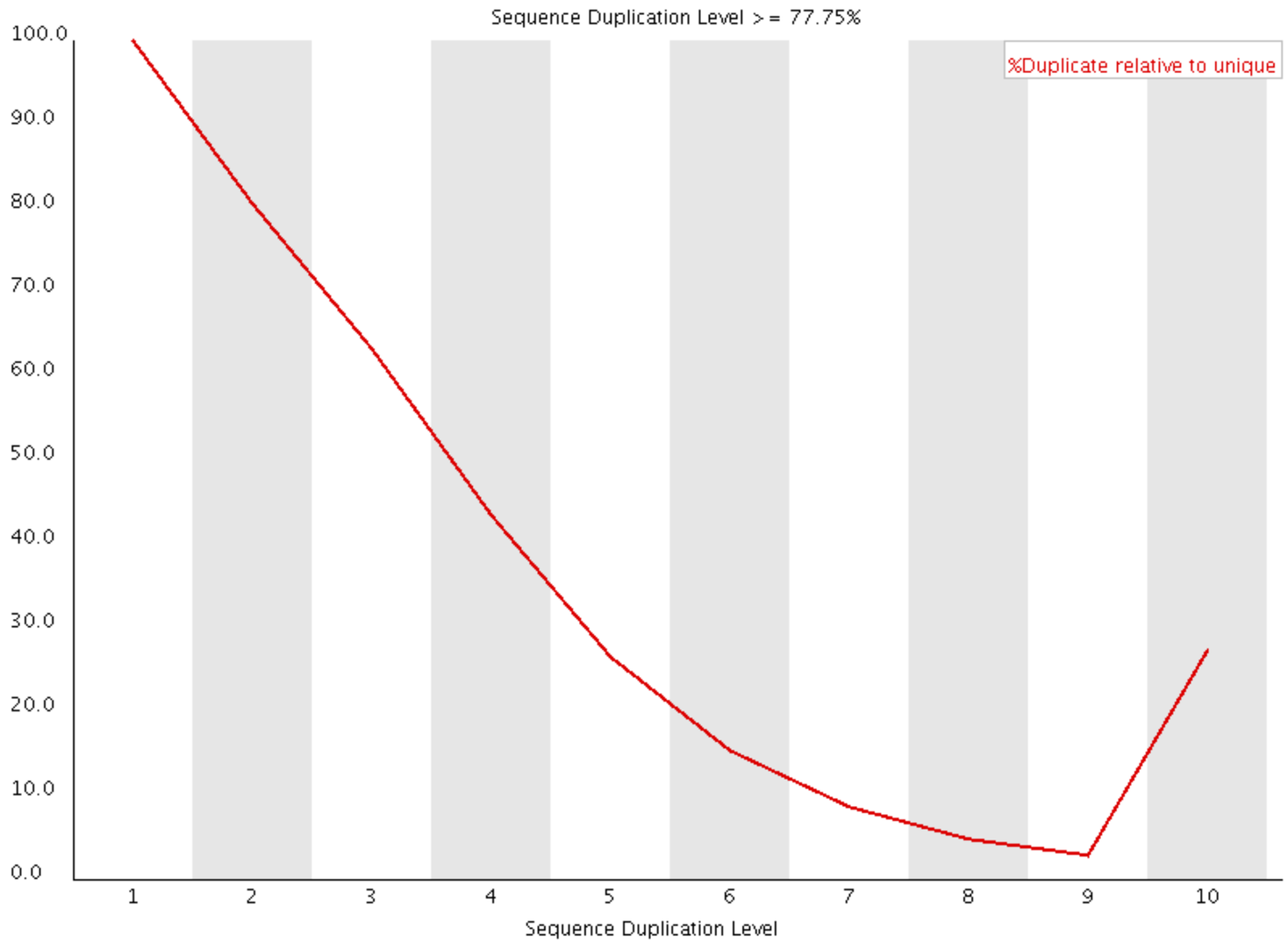


Illumina: optical effects



Positional sequence bias

PCR Artifacts



Duplicated sequences

Over-represented sequences

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

Filtering

	sequence	count
1	ATTAACCCTCACTAAAGGGACTAGTCCTGCAGGTTTAAACGAATTCGCCC	482185
151	ATTAACCCTCACTAAAGGGACTAGTCCTGCAGGTTTAAACGAATTCGCCC	271724
2	TAATACGACTCACTATAGGGCGAATTGAATTTAGCGGCCGCGAATTCGCC	159936
152	TAATACGACTCACTATAGGGCGAATTGAATTTAGCGGCCGCGAATTCGCC	105273
153	CTTAACCCTCACTAAAGGGACTAGTCCTGCAGGTTTAAACGAATTCGCCC	46872
3	CTTAACCCTCACTAAAGGGACTAGTCCTGCAGGTTTAAACGAATTCGCCC	43212
4	NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	13142

Read Frequency Distribution

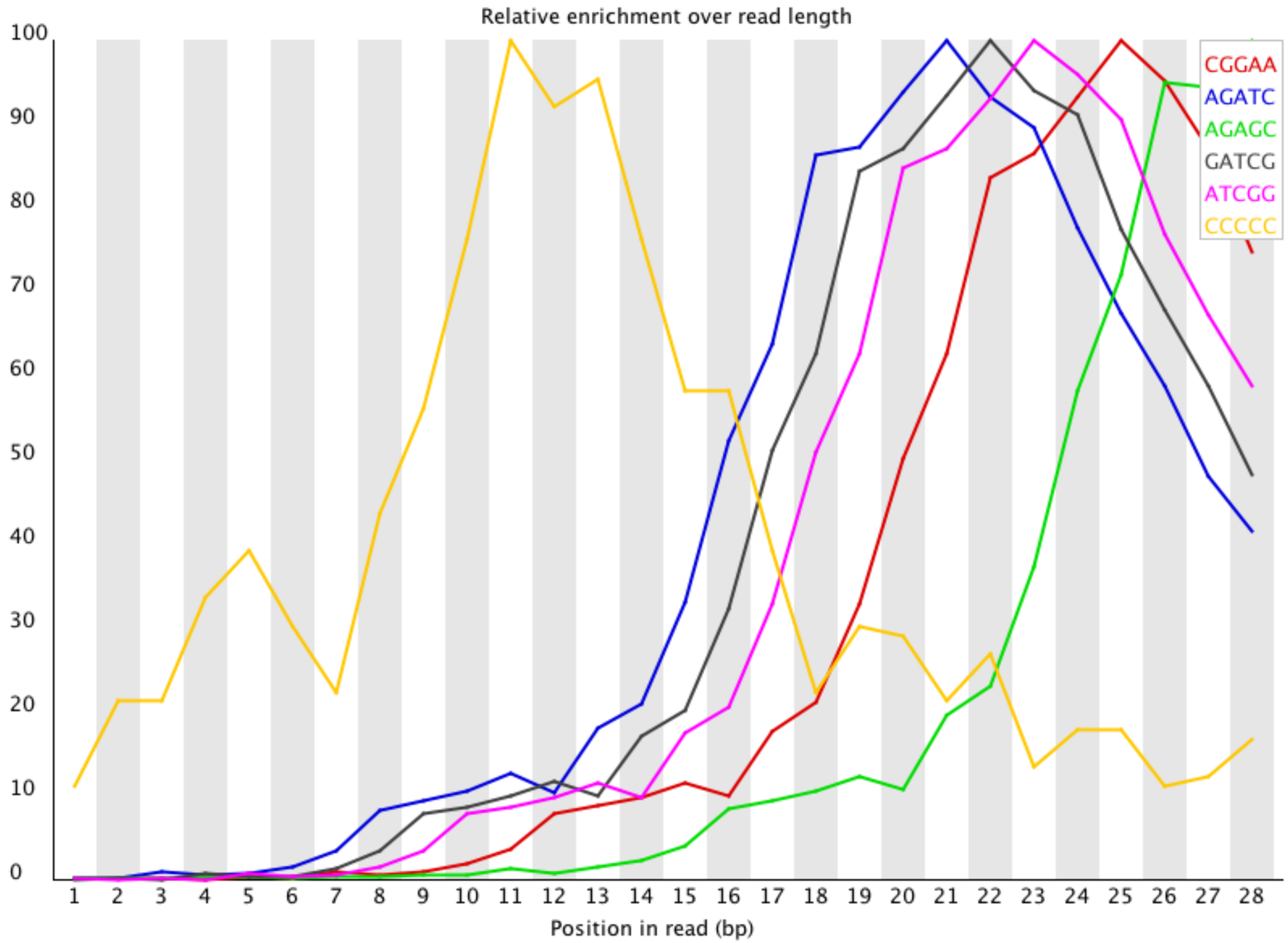
Contamination

> 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
ATTAACCCTCACTAAAGGGACTAGTCCTGCAGGTTTAAACGAATTCGCCC  50

||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  43
ATTAACCCTCACTAAAGGGACTAGTCCTGCAGGTTTAAACGAATTCGCCC  92
```



Adaptor contamination