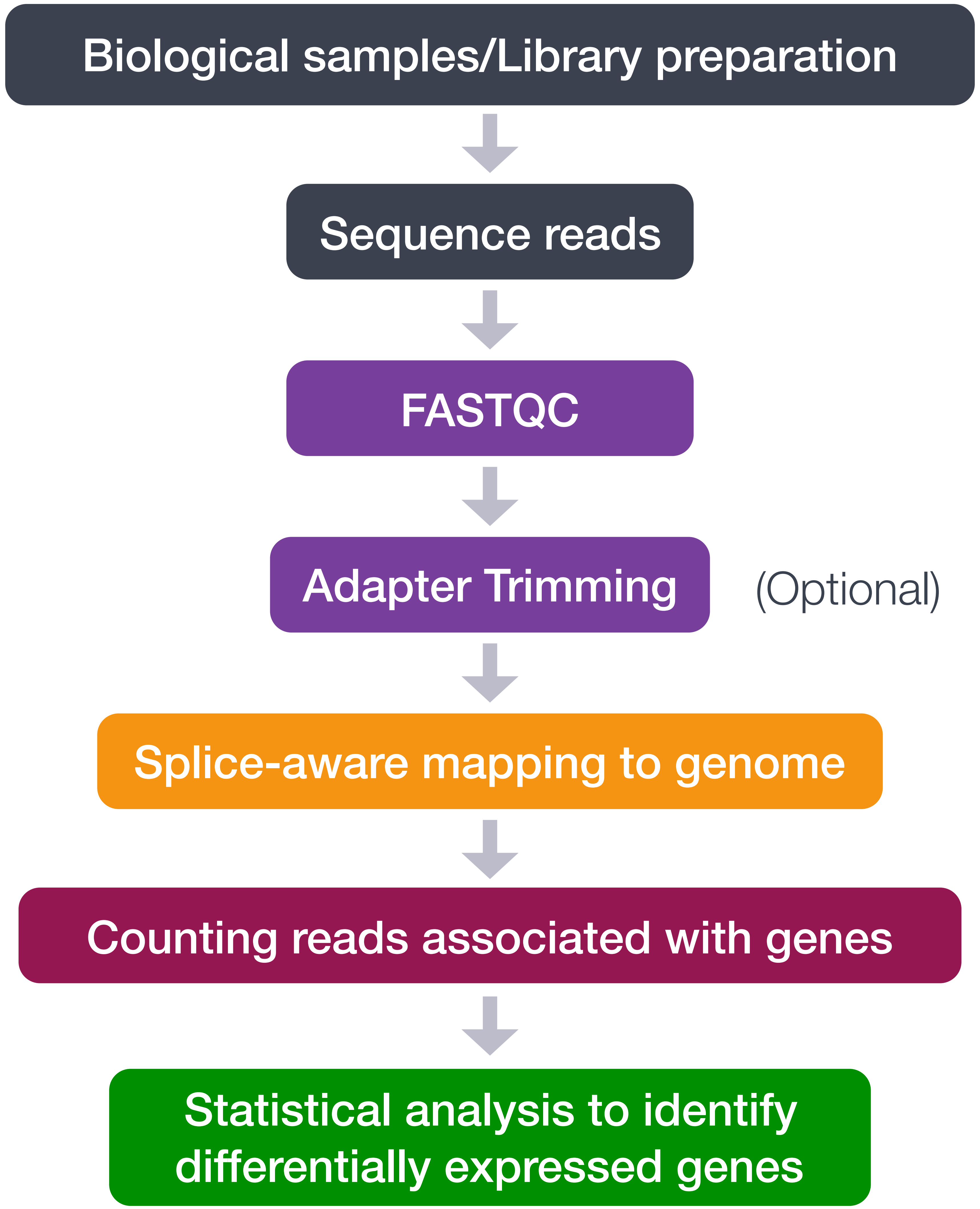
Illumina Sequencing Error Profiles

and

Quality Control

RNA-seq Workflow



Biological samples/Library preparation

Sequence reads

FASTQC

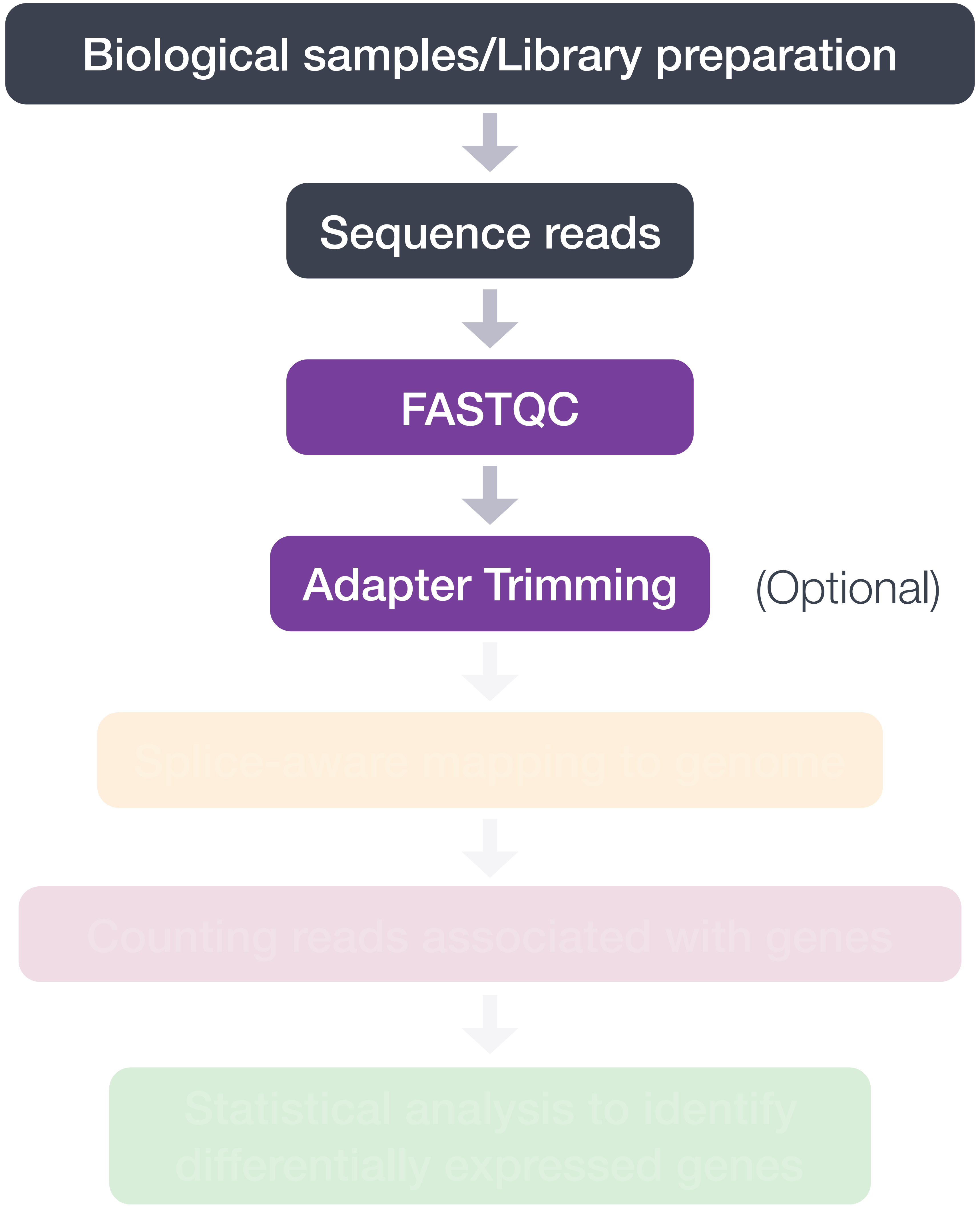
Adapter Trimming (Optional)

Splice-aware mapping to genome

Counting reads associated with genes

Statistical analysis to identify differentially expressed genes

Quality Checks: Raw Data



Biological samples/Library preparation

Sequence reads

FASTQC

Adapter Trimming

(Optional)

Splice-aware mapping to genome

Counting reads associated with genes

Statistical analysis to identify differentially expressed genes

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
```

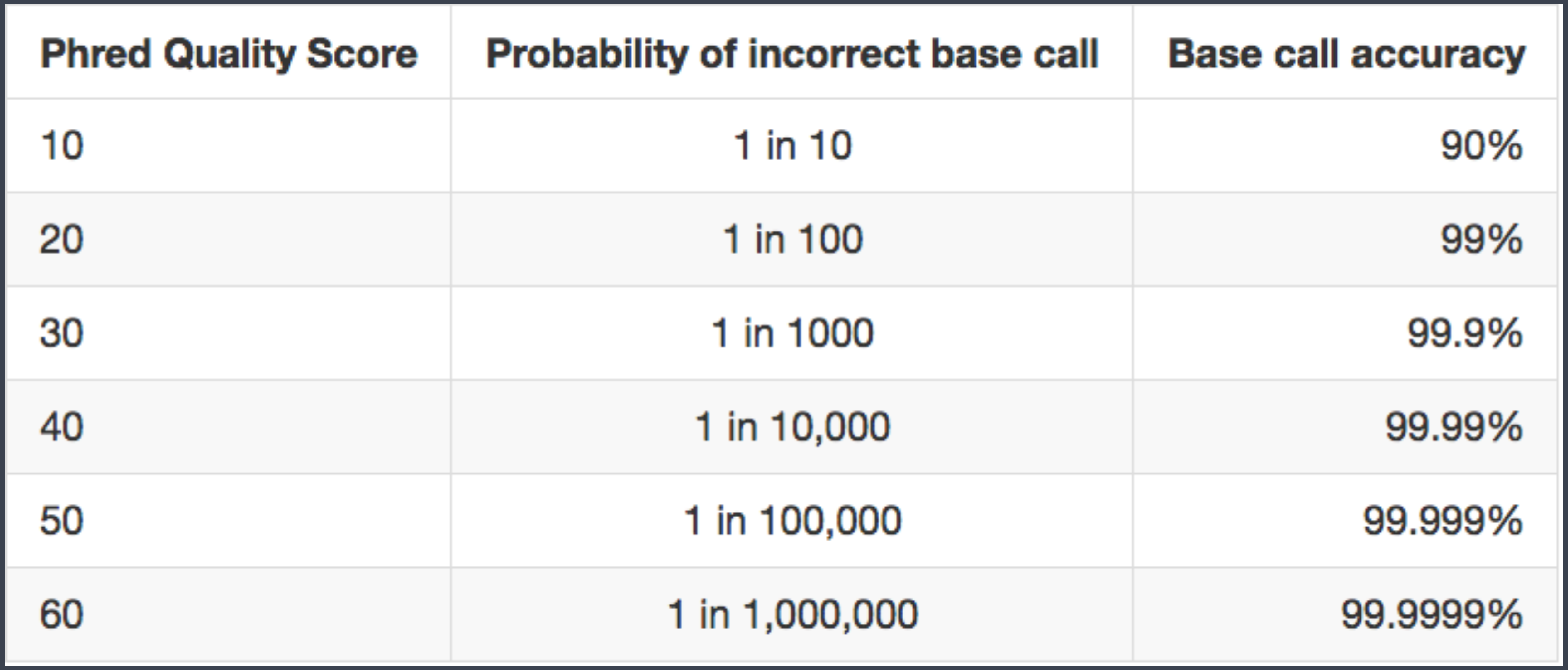
```
Q = -10 x log10(P), where P is the probability that a base call is erroneous
```

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

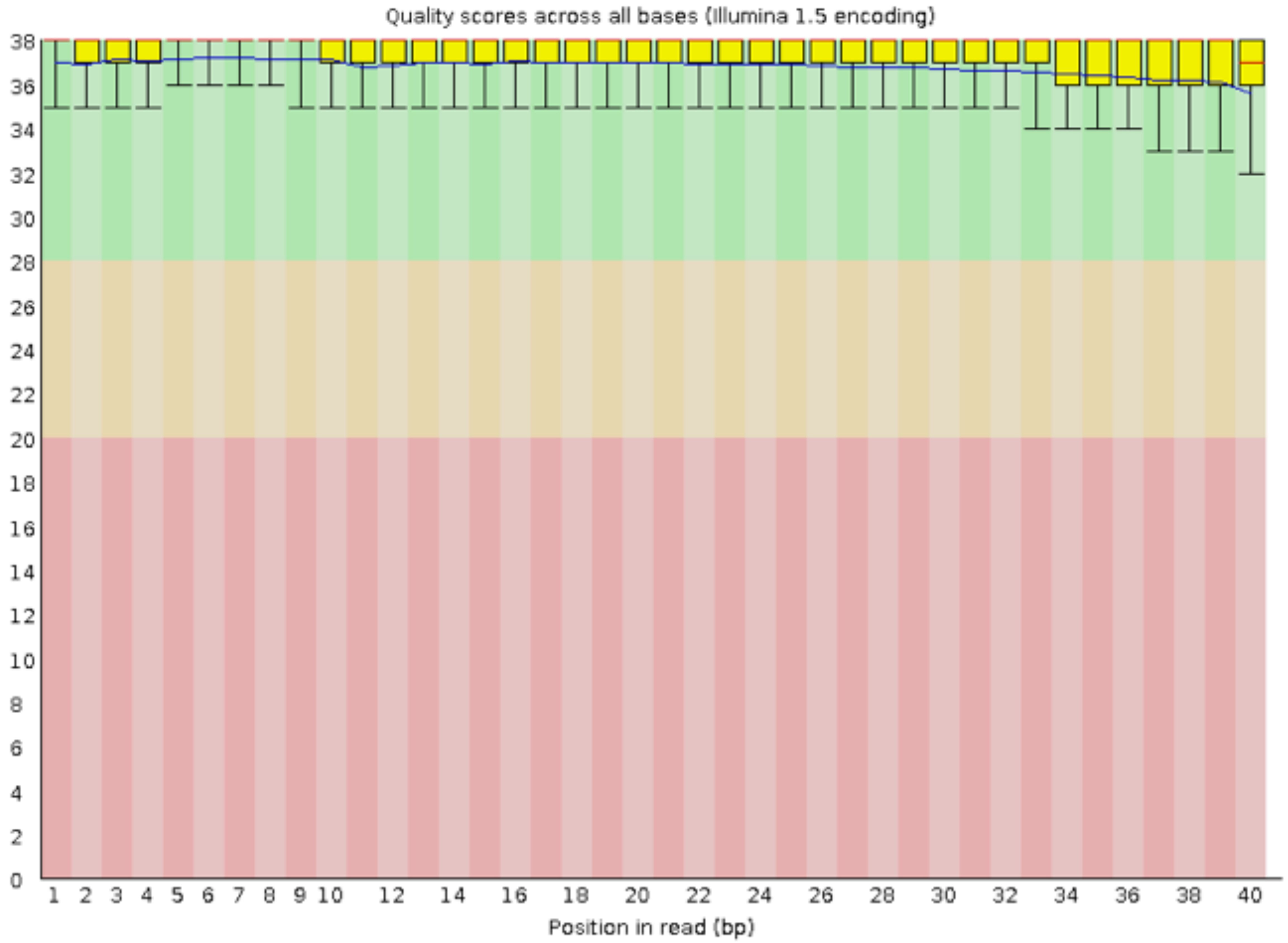
Different quality encoding scales exist (differing by offset in the ASCII table), but note the most commonly used one is fastqsanger.

FASTQ Quality Scores

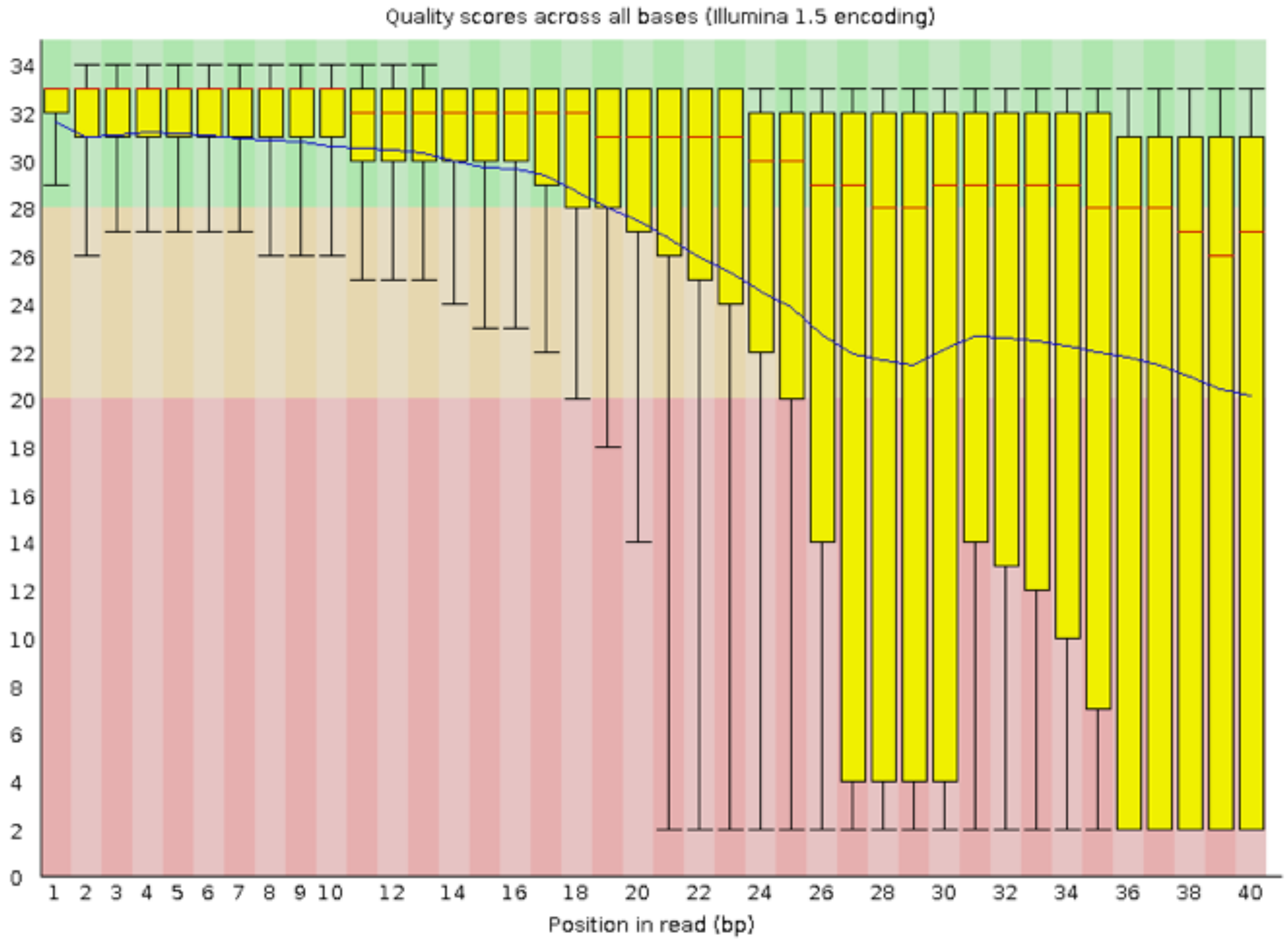
These probability values are the results from the base calling algorithm and dependent on how much signal was captured for the base incorporation. The score values can be interpreted as follows:



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%



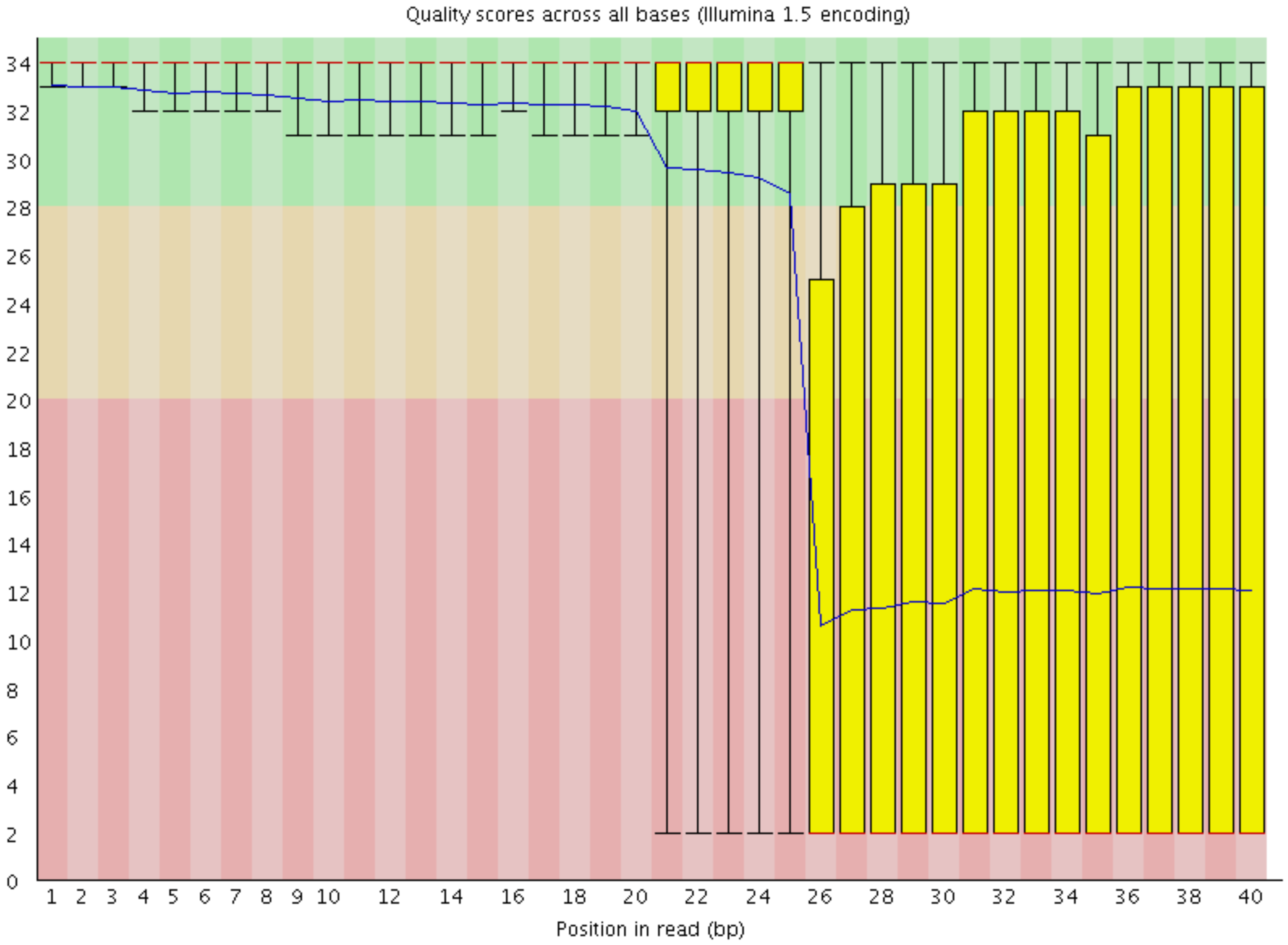
A good quality sample



A not-so-good quality sample

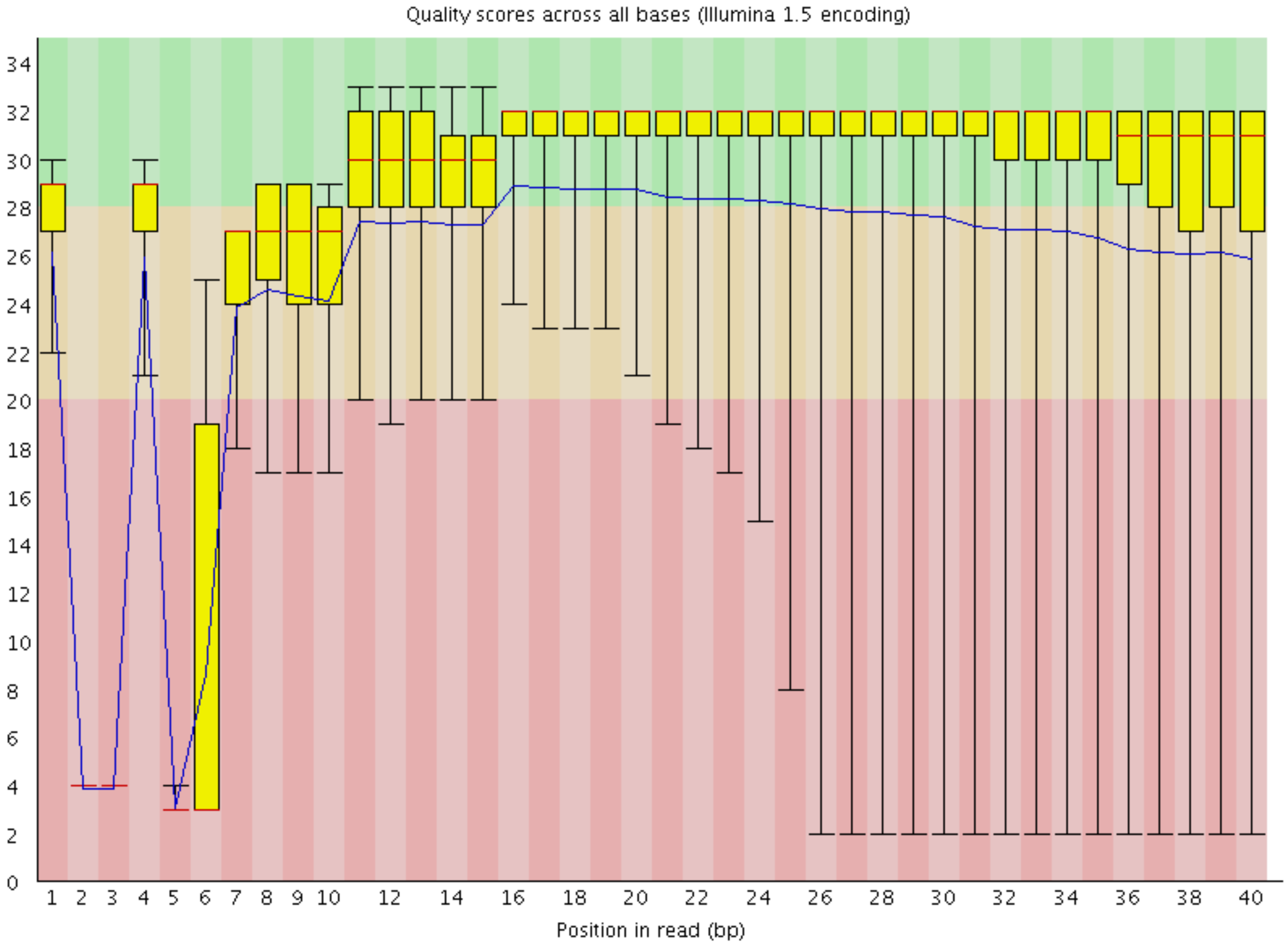
Error profiles:
Technical Sequencer Problems

Manifold burst in cycle 26

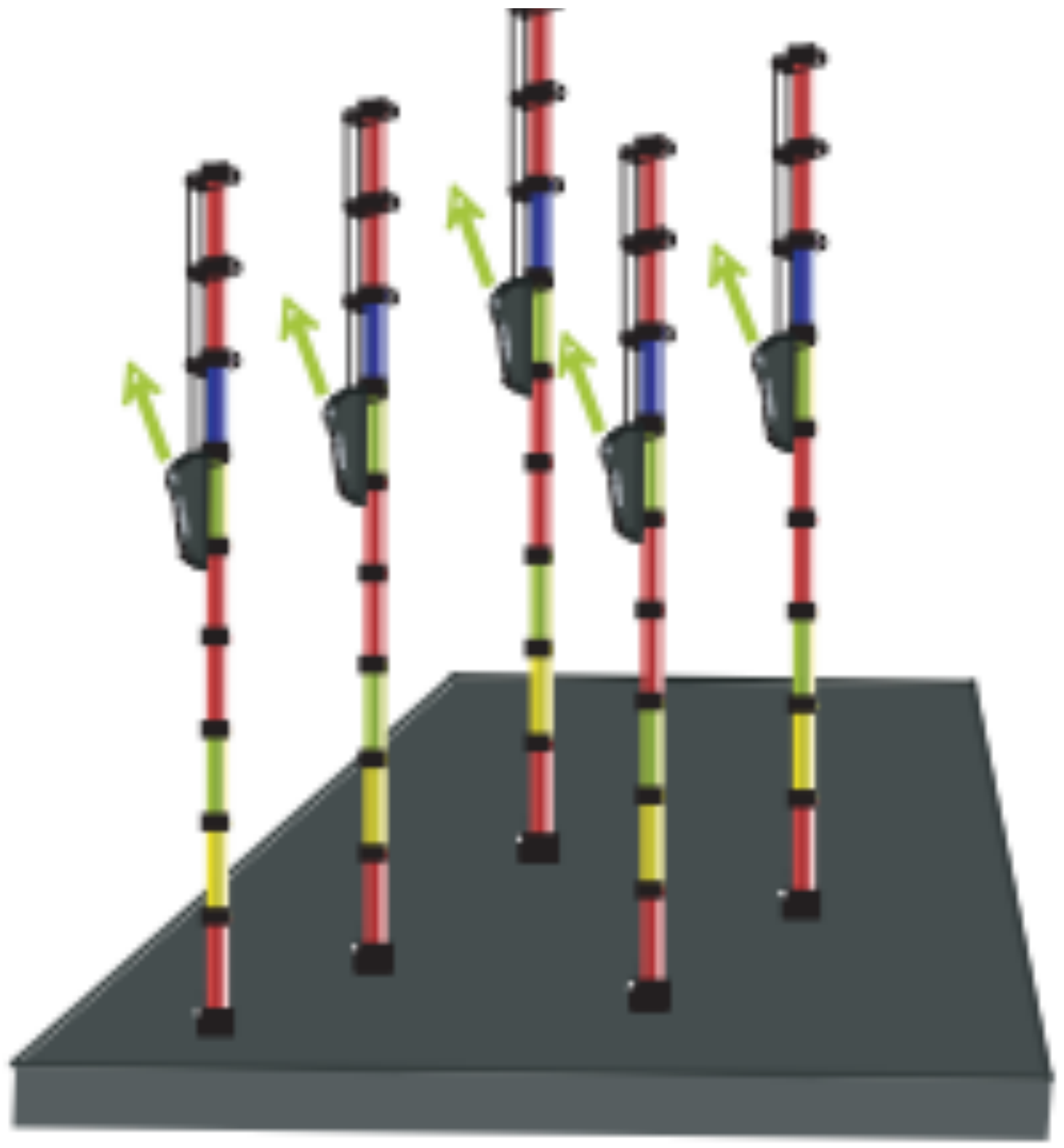


See http://bioinfo-core.org/index.php/9th_Discussion-28_October_2010 for more example

Specific cycles lost

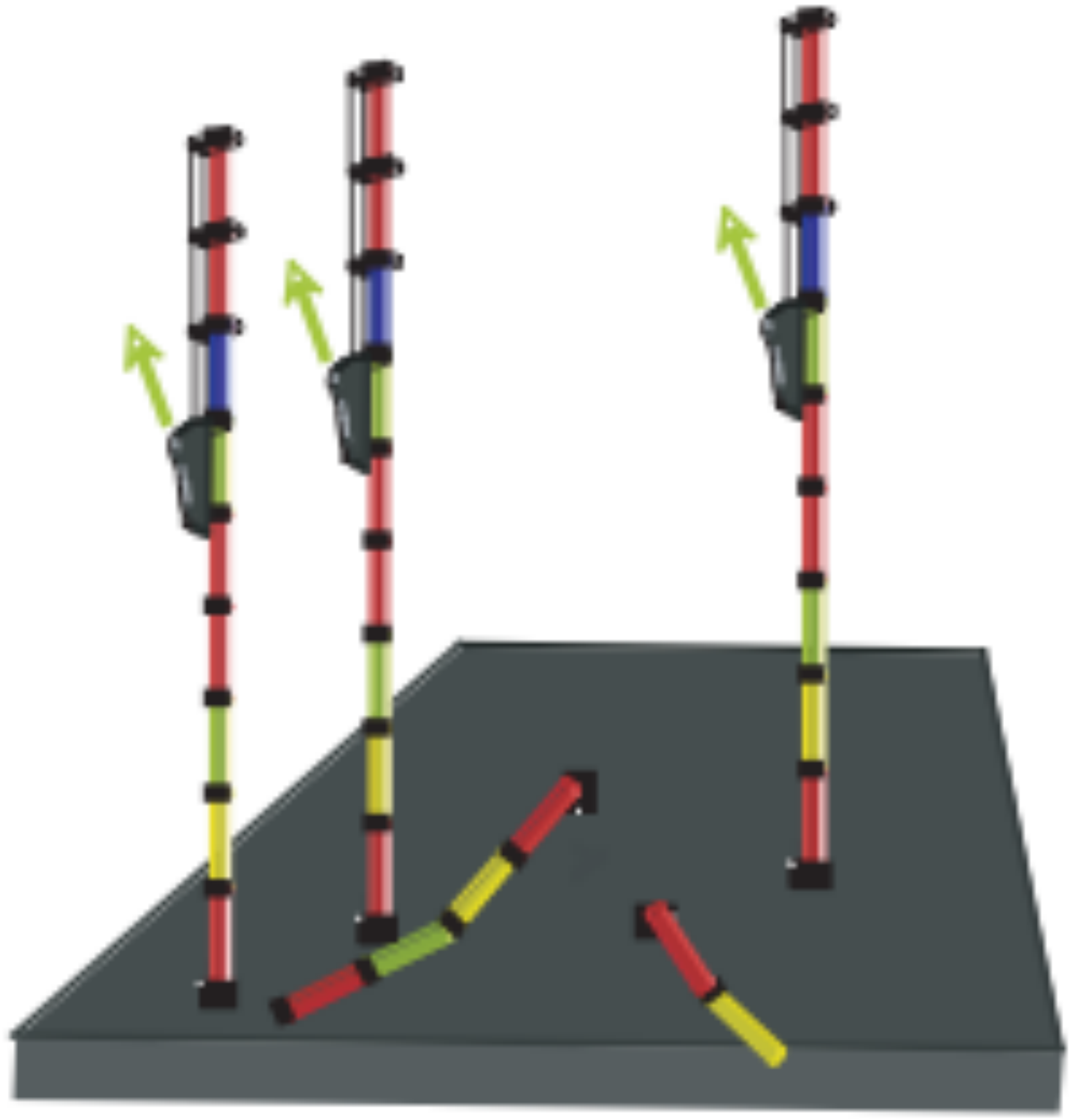


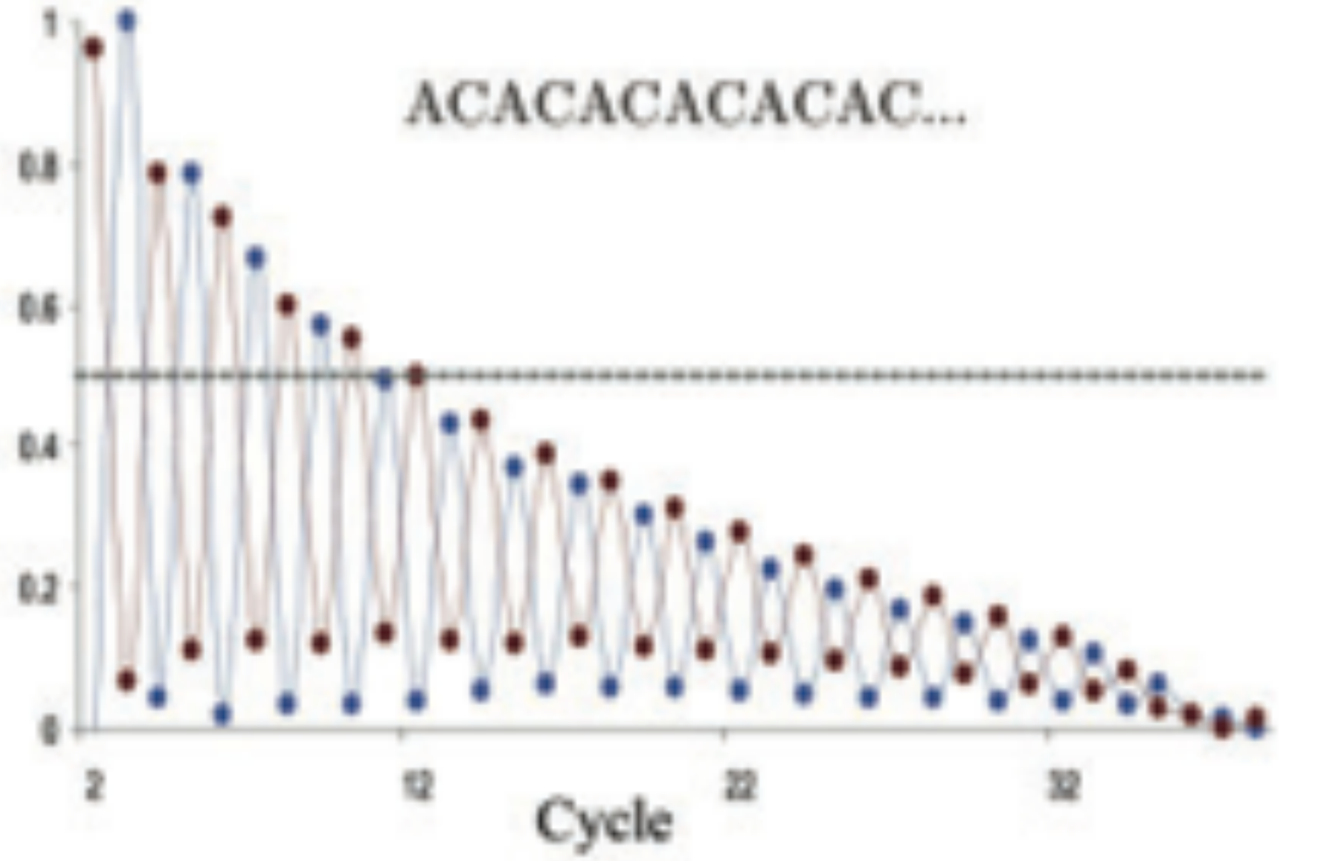
Error dependency on technology



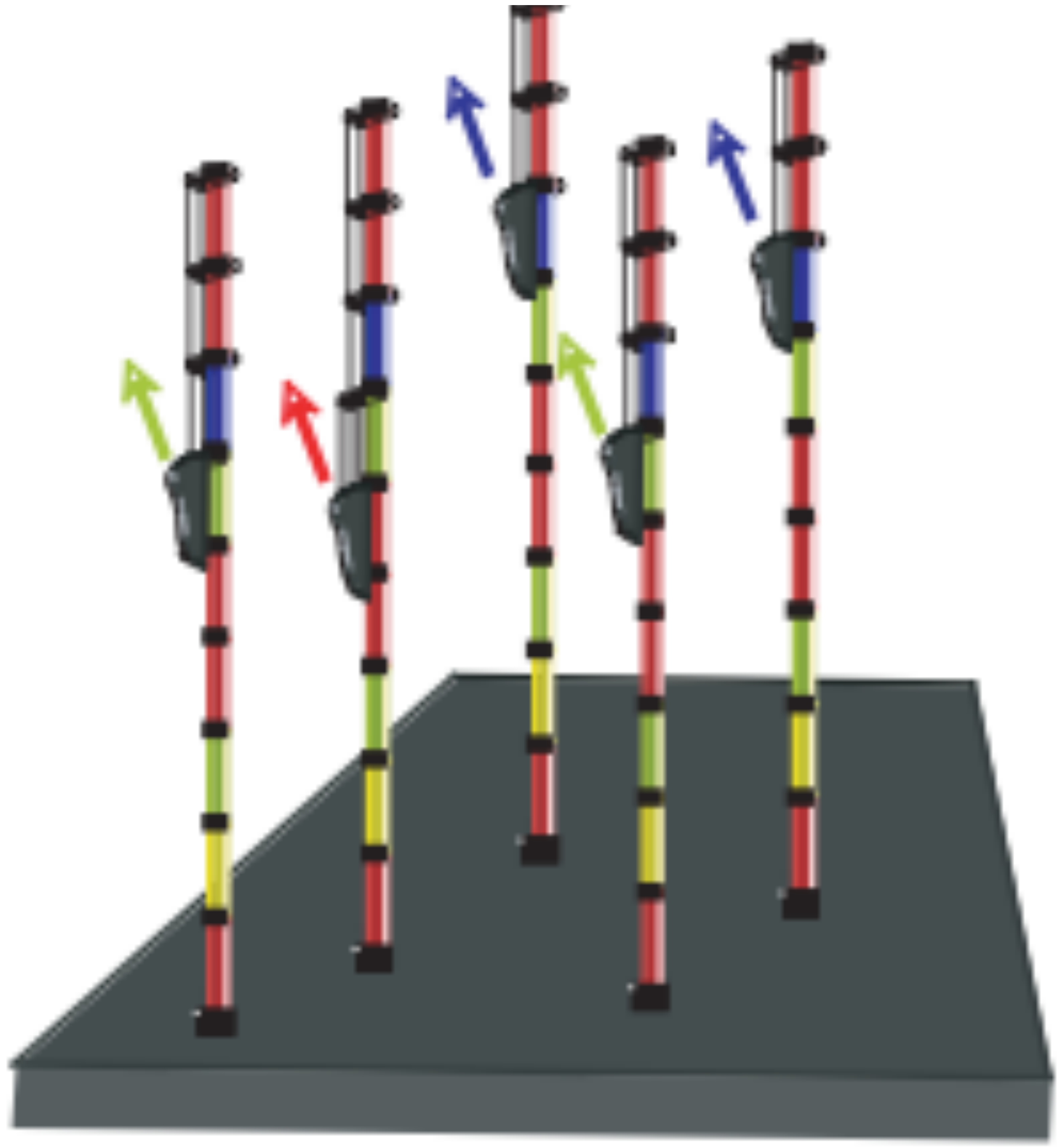
Illumina

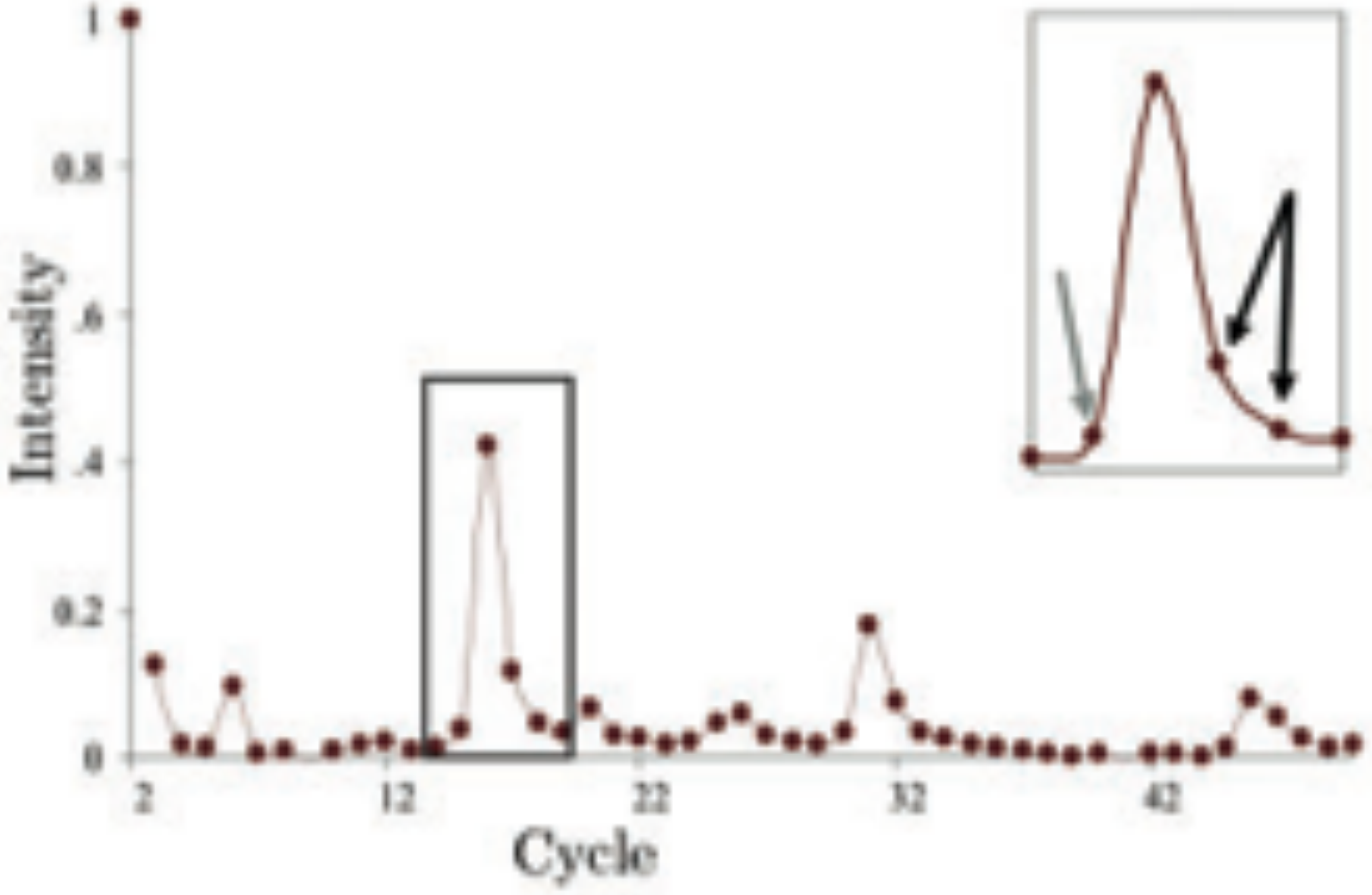
Base-calling for next-generation sequencing platforms. Brief Bioinform 2011, 12(5):489-497



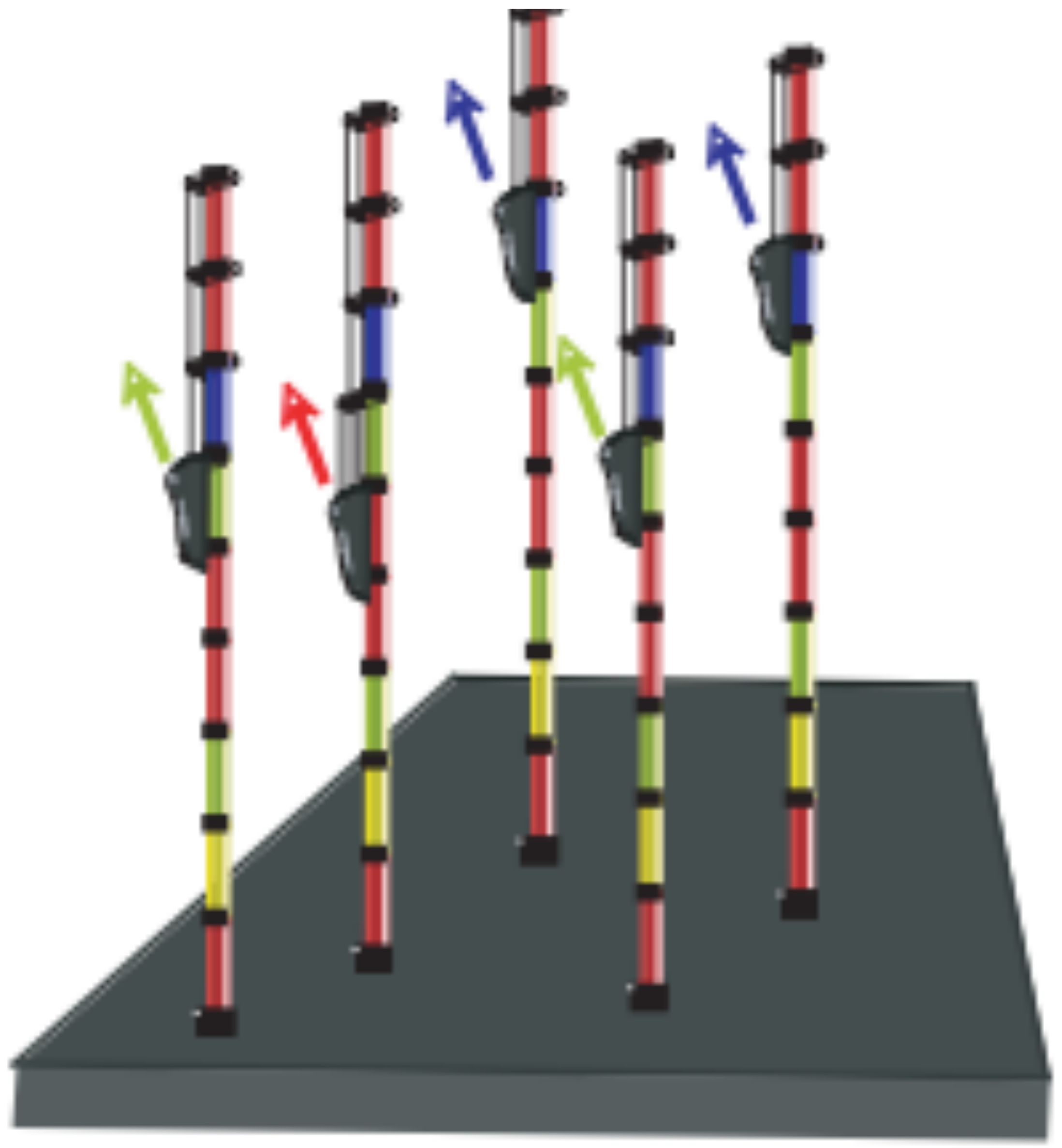


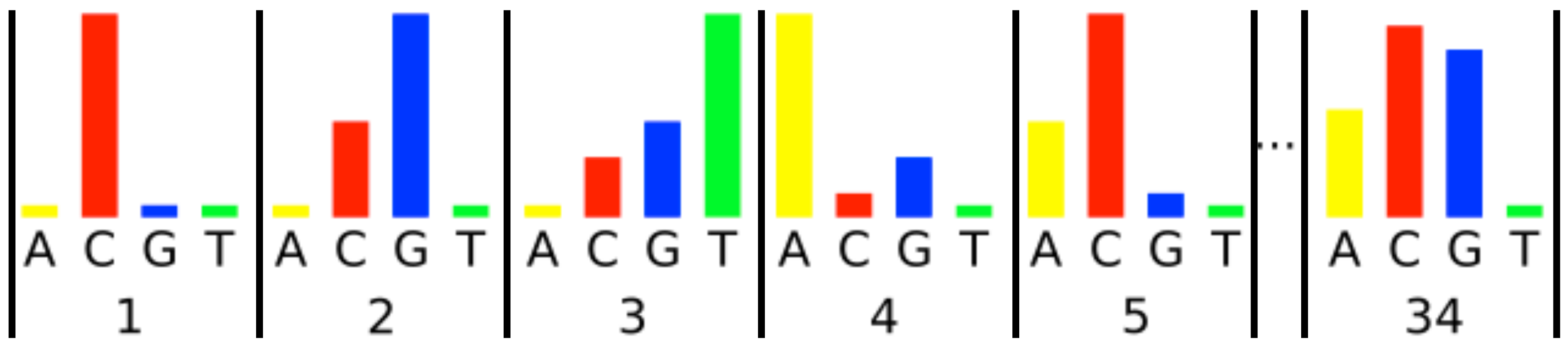
Illumina: signal decay



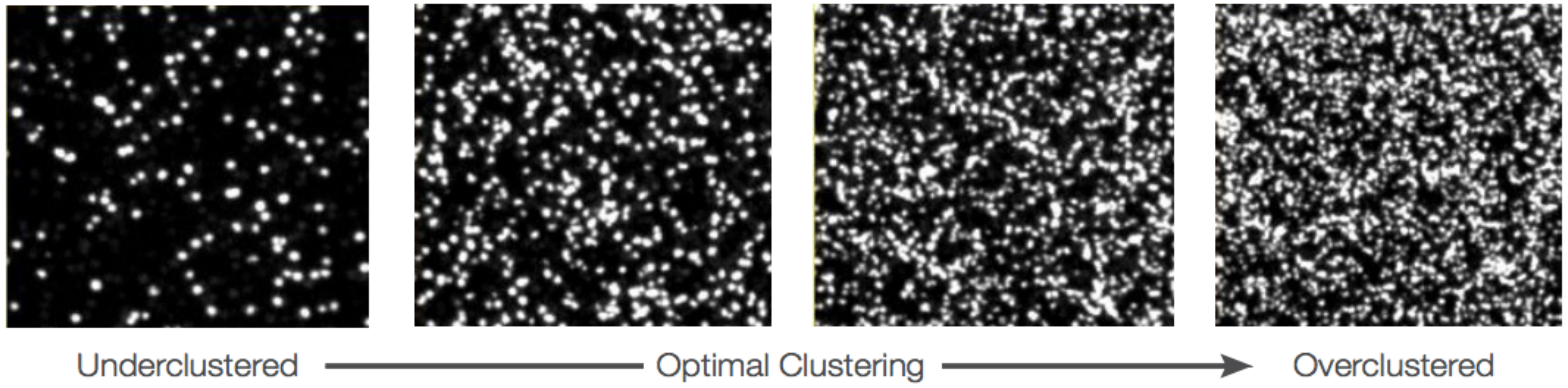


Illumina: phasing

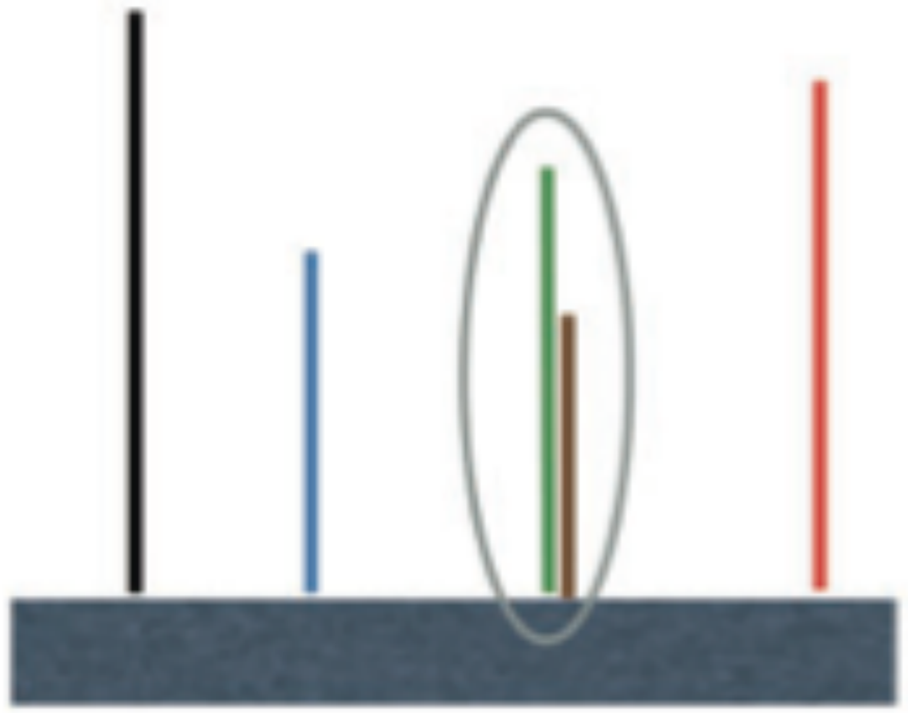




Illumina: phasing

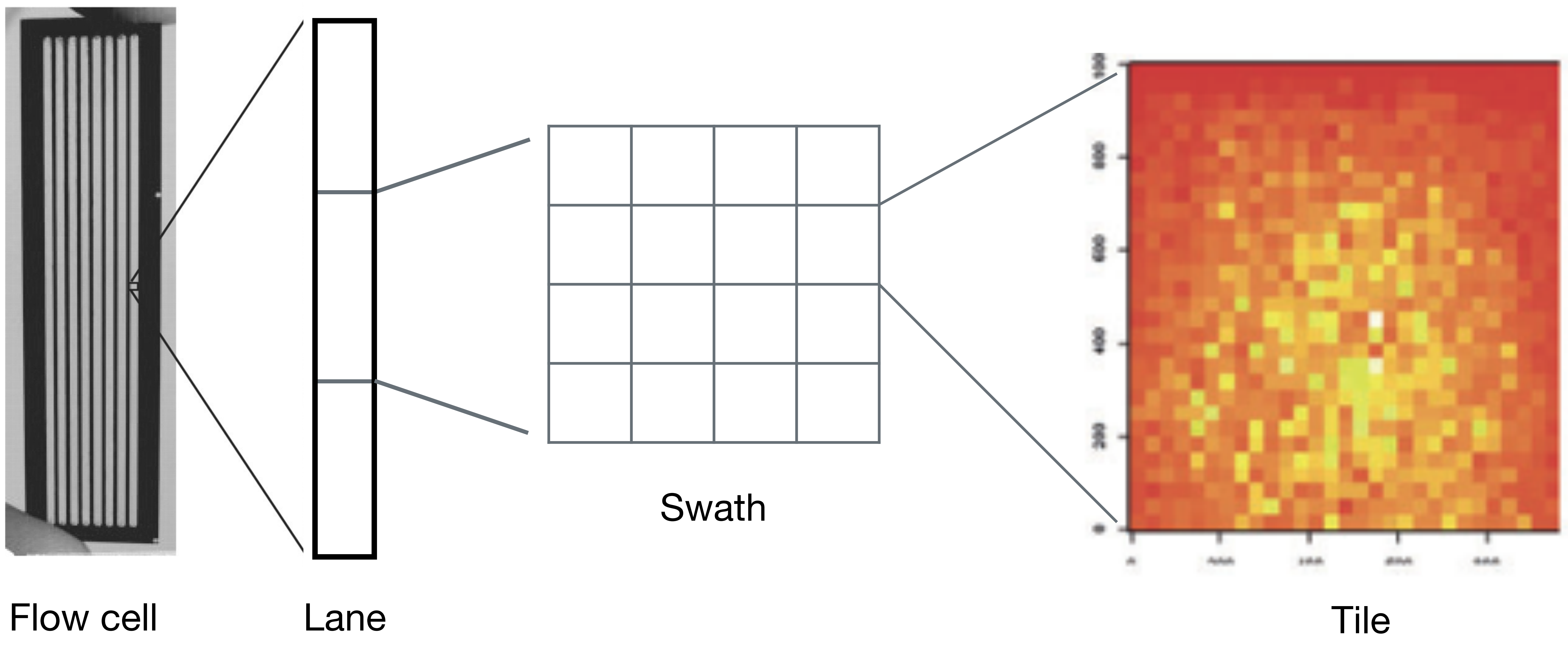


Underclustered → Optimal Clustering → Overclustered

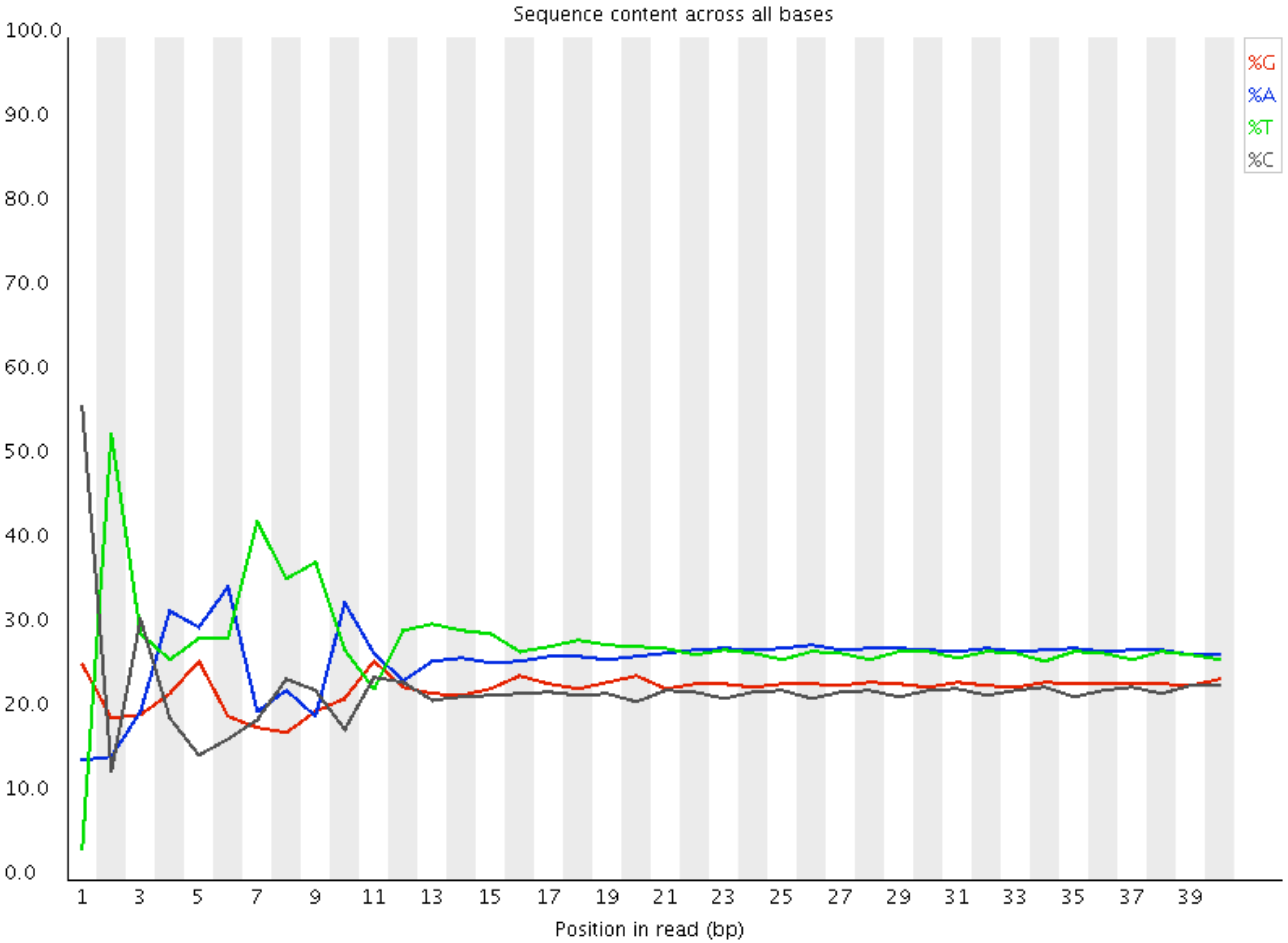


mixed clusters

Illumina: flow cell clusters

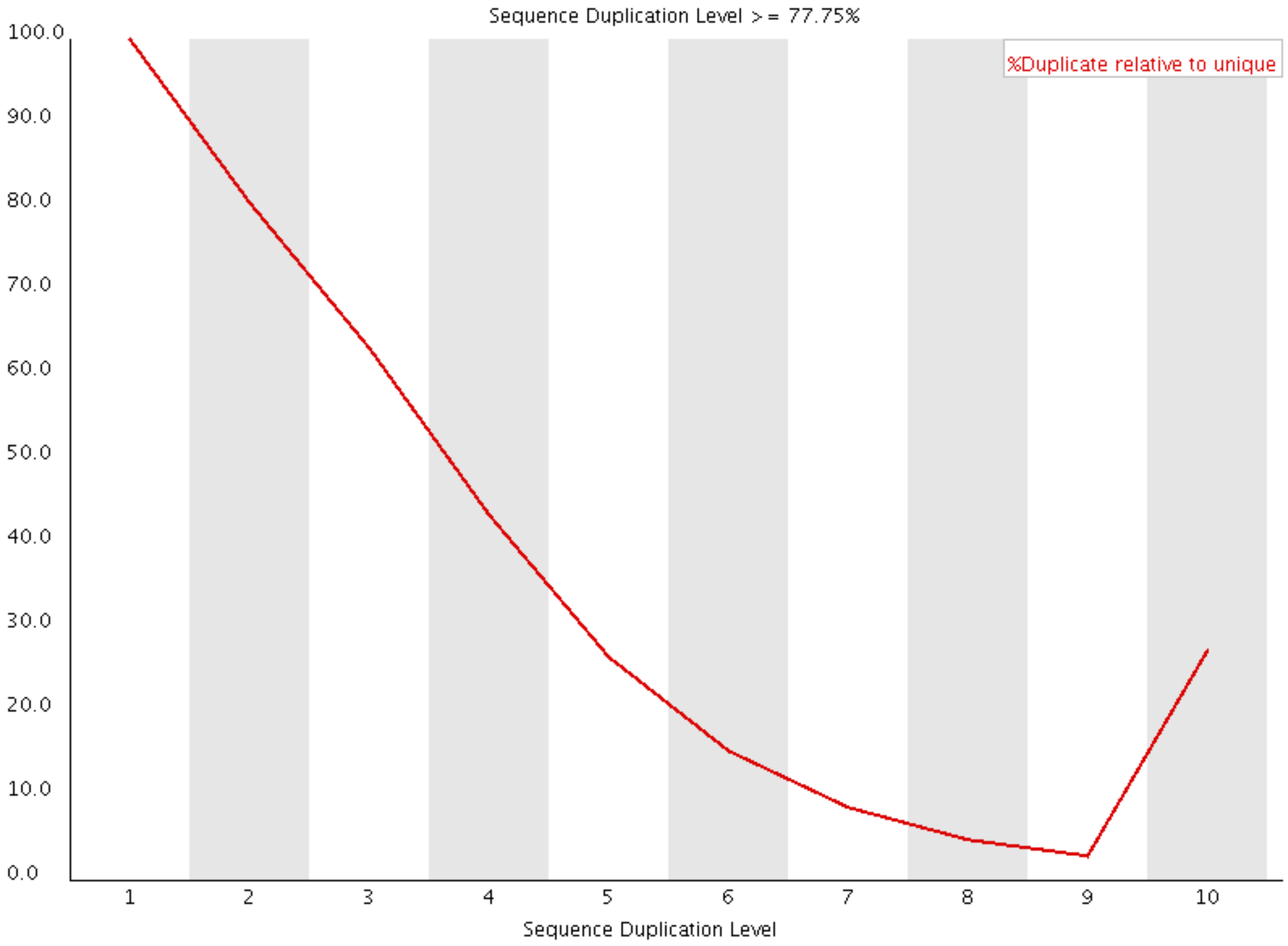


Illumina: optical effects



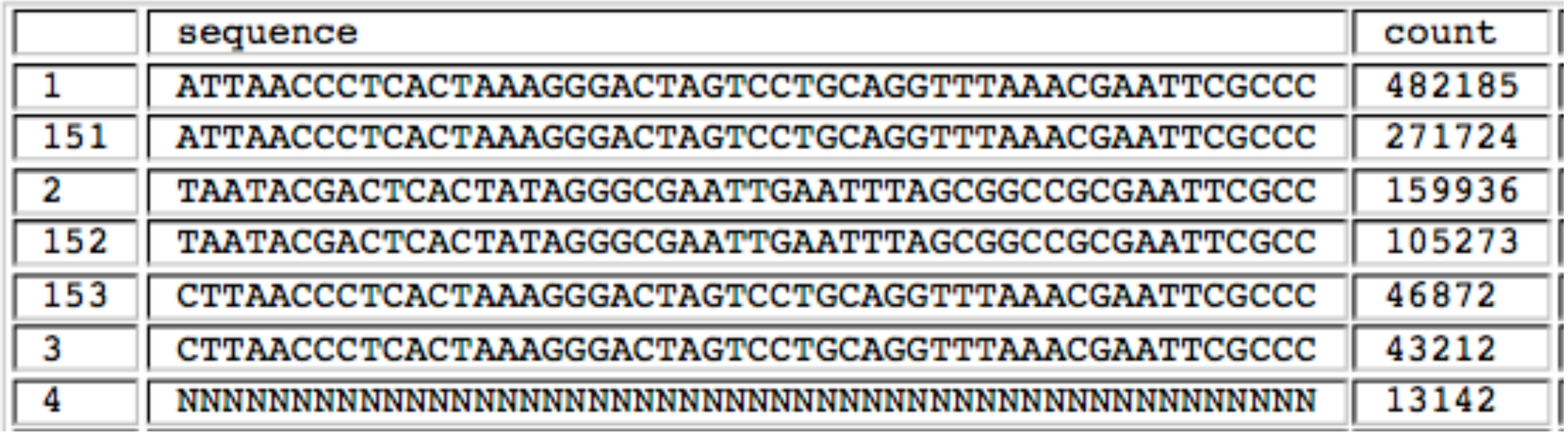
Positional sequence bias

PCR Artifacts



Duplicated sequences

Over-represented sequences



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

Quality Checks for Raw Data

Quality Checks: Raw Data

All NGS analyses require that the **quality of the raw data** is assessed prior to any downstream analysis.

The quality checks at this stage in the workflow include:

1. Checking the **quality of the base calls** to ensure that there were no issues during sequencing

2. Examining the reads to ensure their **quality metrics adhere to our expectations** for our experiment

3. Exploring reads for **contamination**

The tool **FASTQC** is often used to assess these metrics, and it generates a QC report for each sample.

Quality Checks: Raw Data

Raw Data QC Goals:

- **Identify sequencing problems and determine whether there is a need to contact the sequencing facility**
- Identify over-represented contaminating sequences
- Gain insight into library complexity (rRNA contamination, duplications)
- Ensure organism is properly represented by %GC content