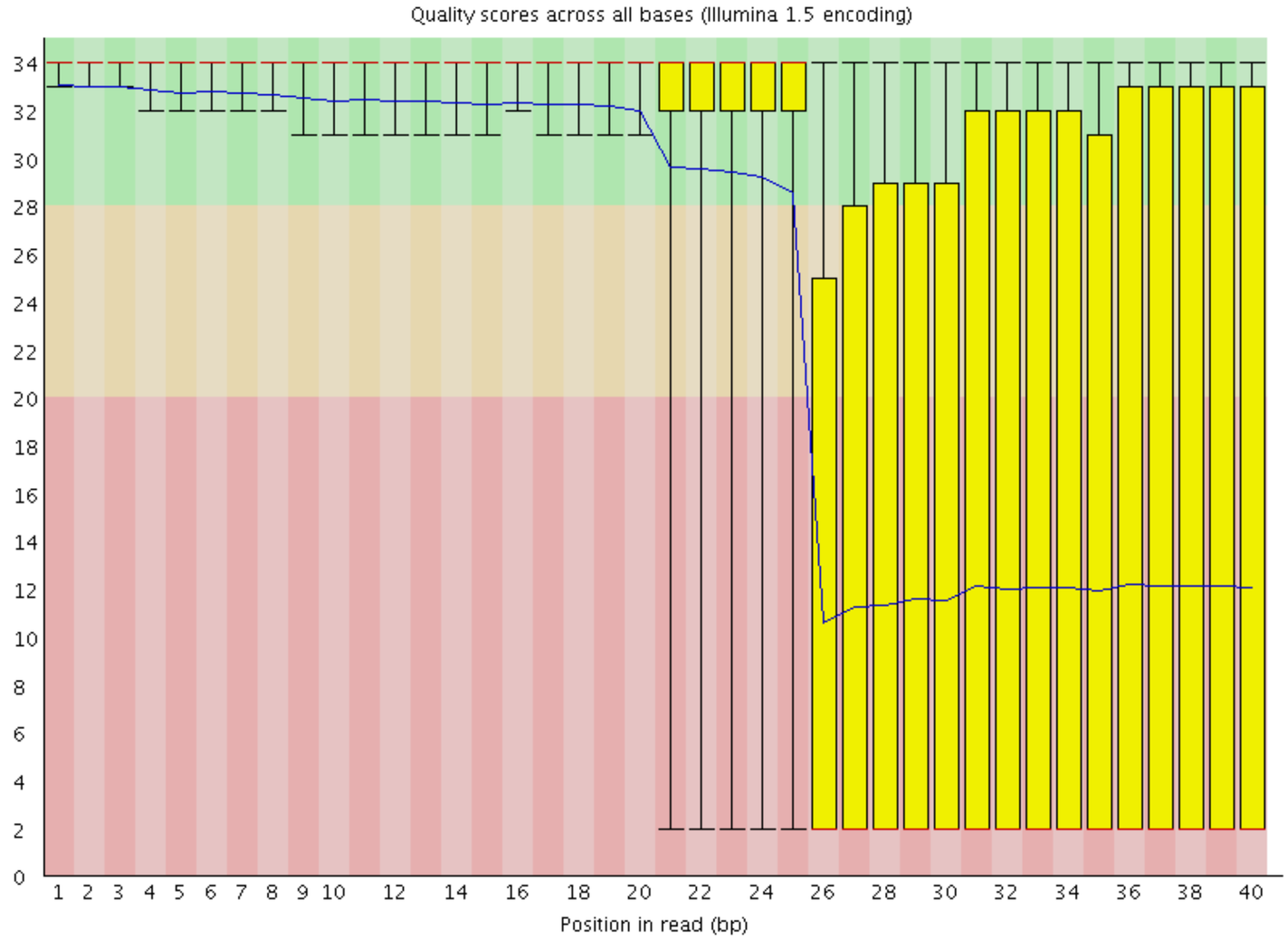
NGS Error Profiles

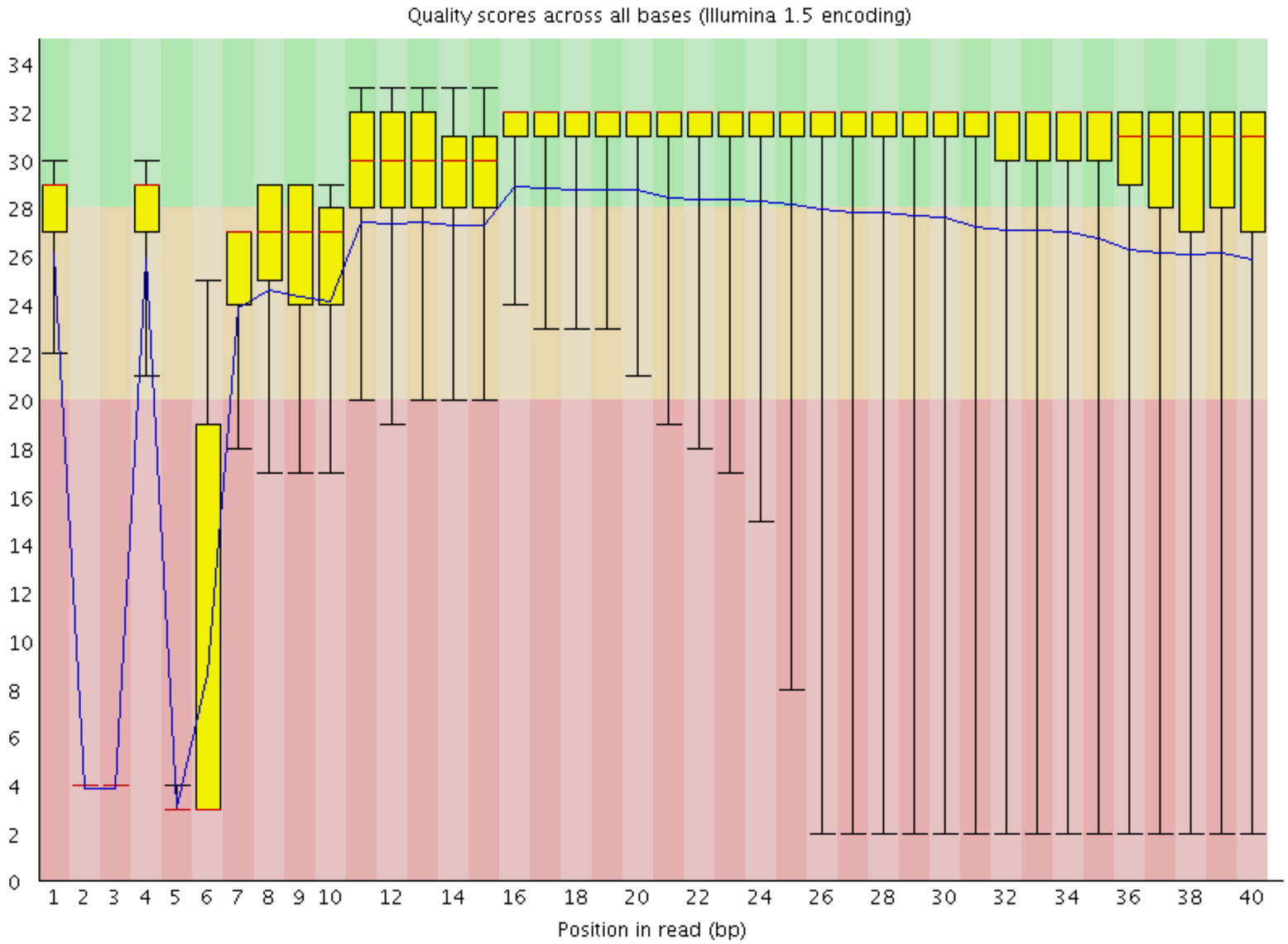
Technical Sequencer Problems

Manifold burst in cycle 26

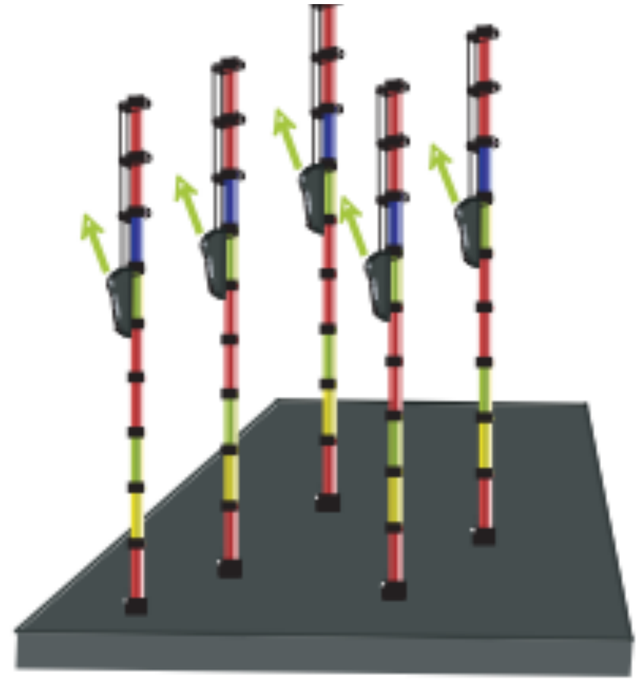


See http://bioinfo-core.org/index.php/9th_Discussion-28_October_2010 for more example

Specific cycles lost

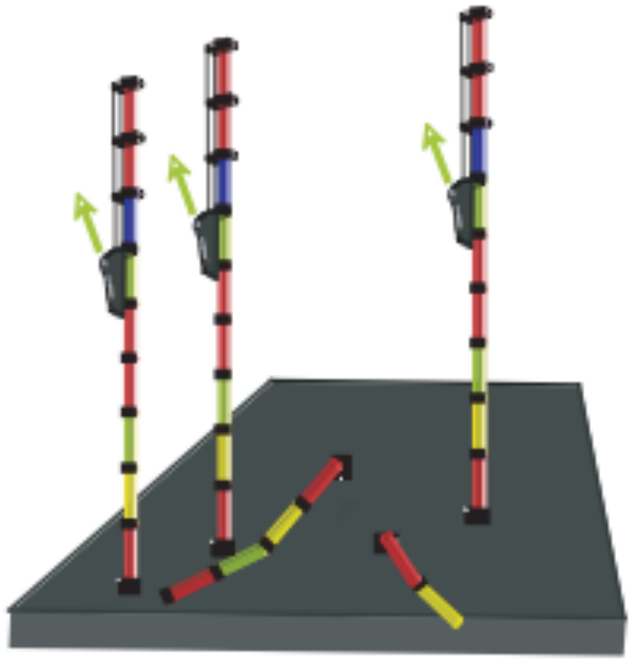


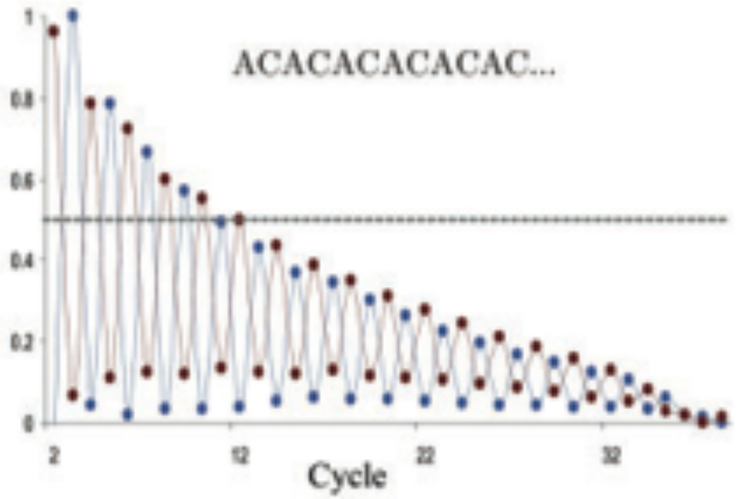
Error dependency on technology



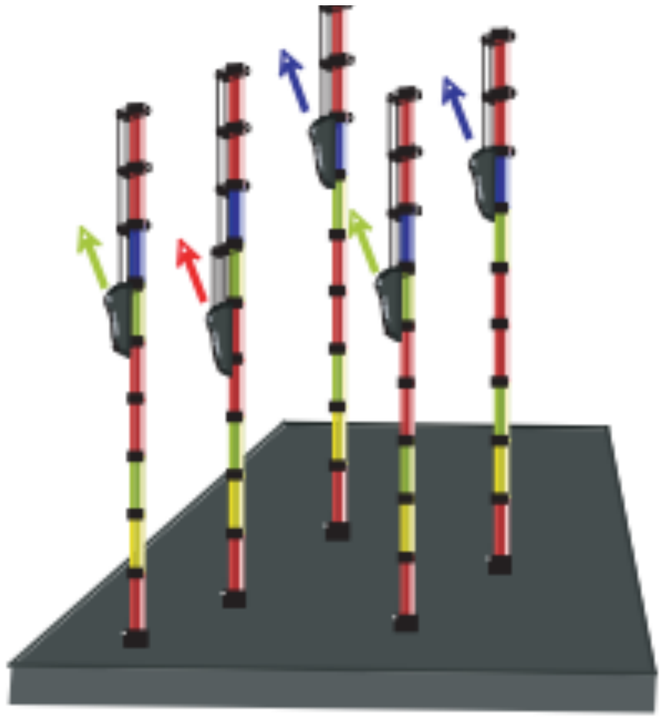
Illumina

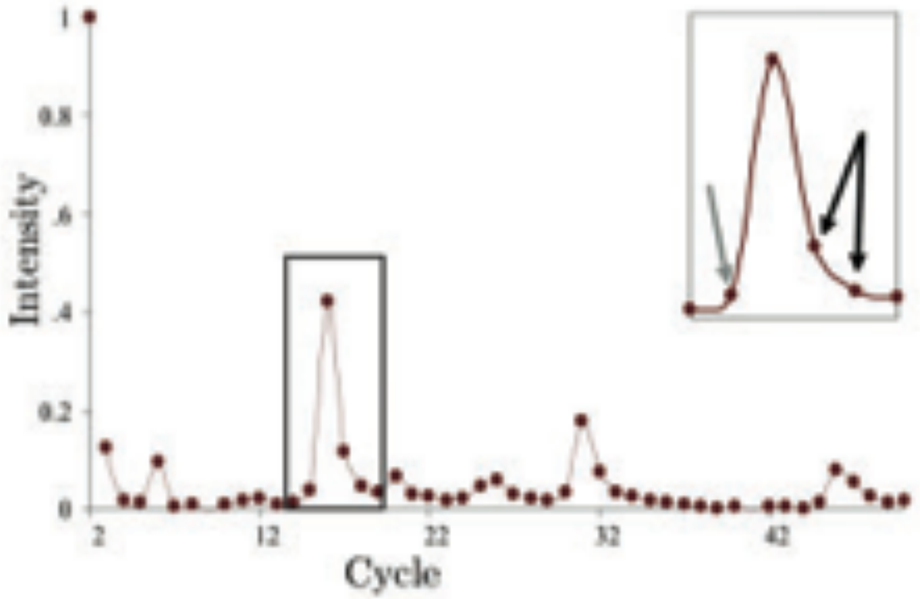
Base-calling for next-generation sequencing platforms. Brief Bioinform 2011, 12(5):489-497



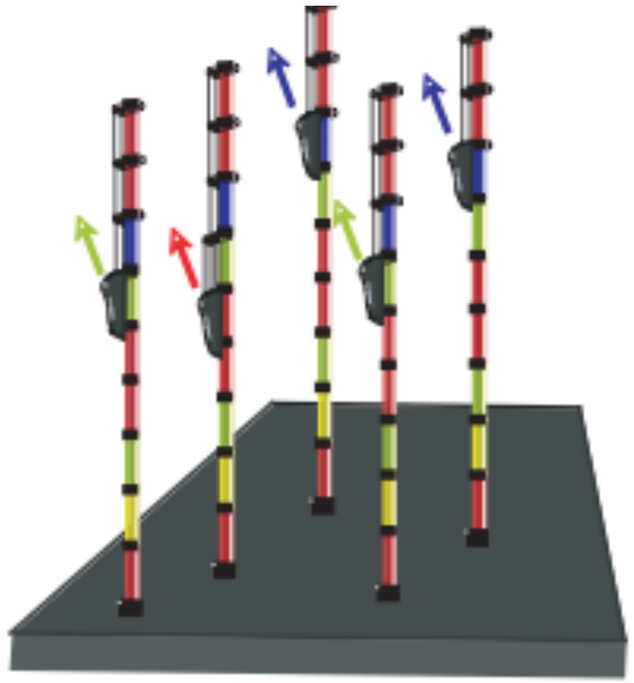


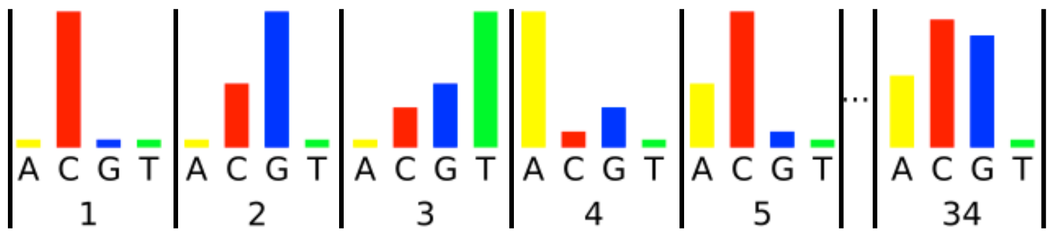
Illumina: signal decay



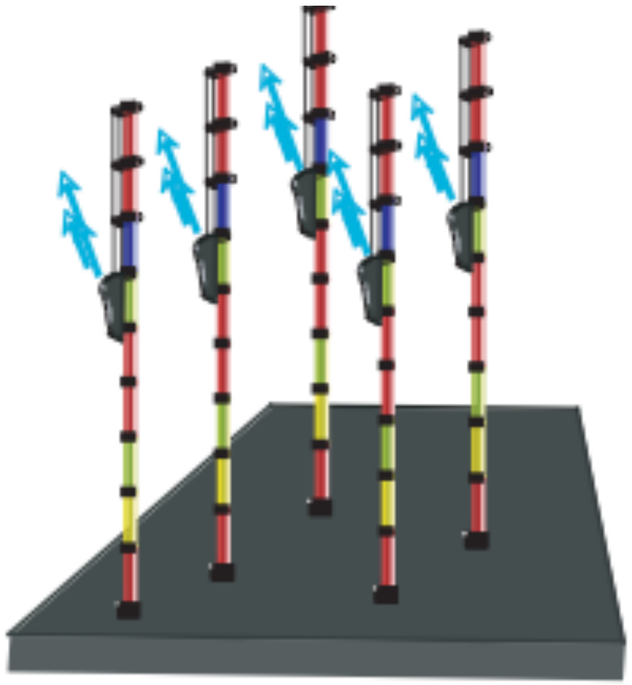


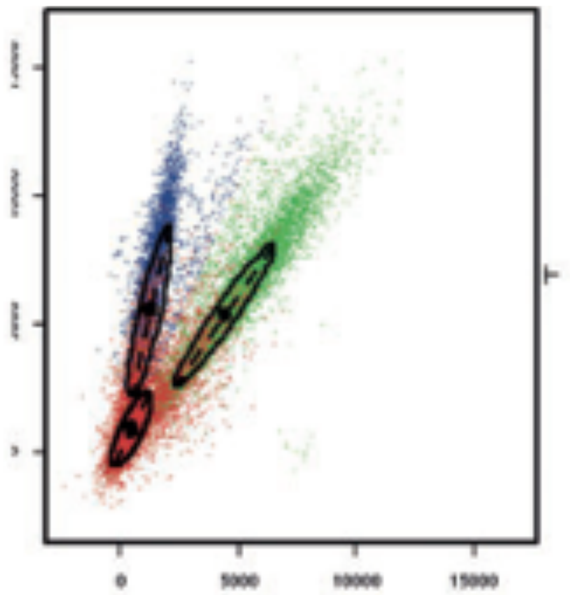
Illumina: phasing



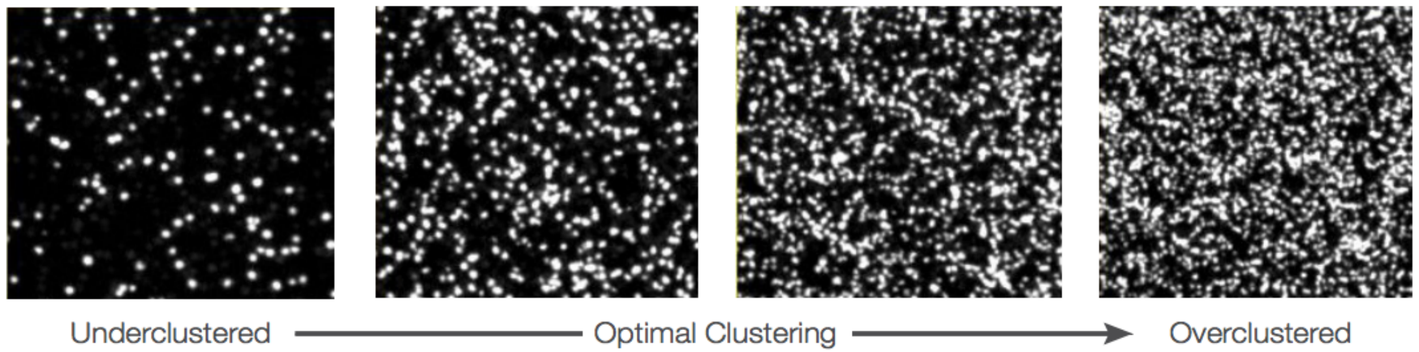


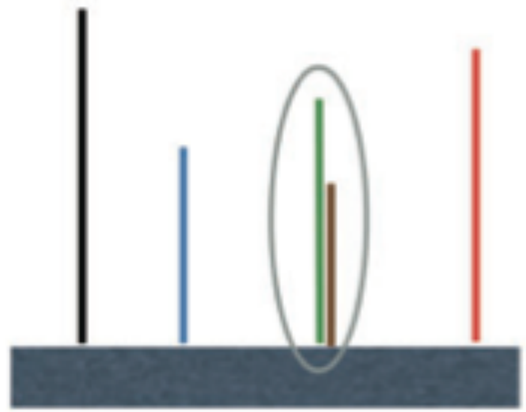
Illumina: phasing



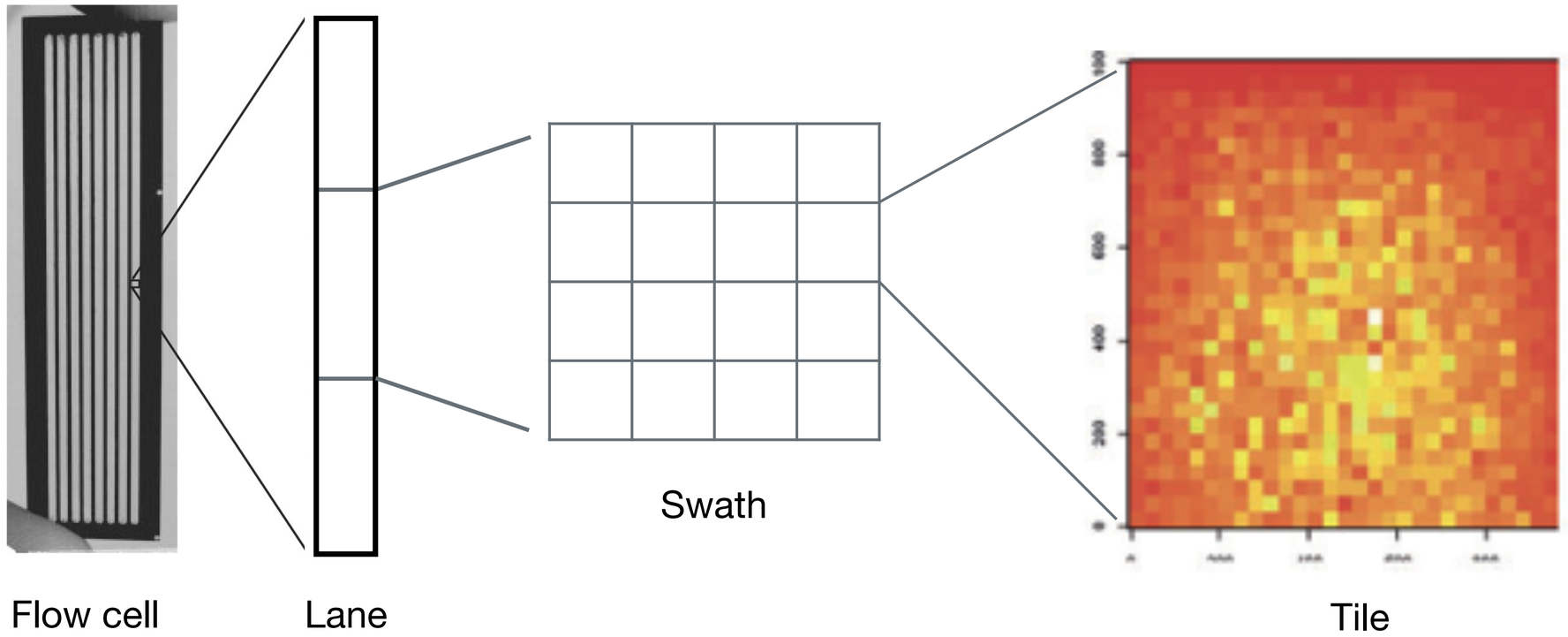


Illumina: cross-talk

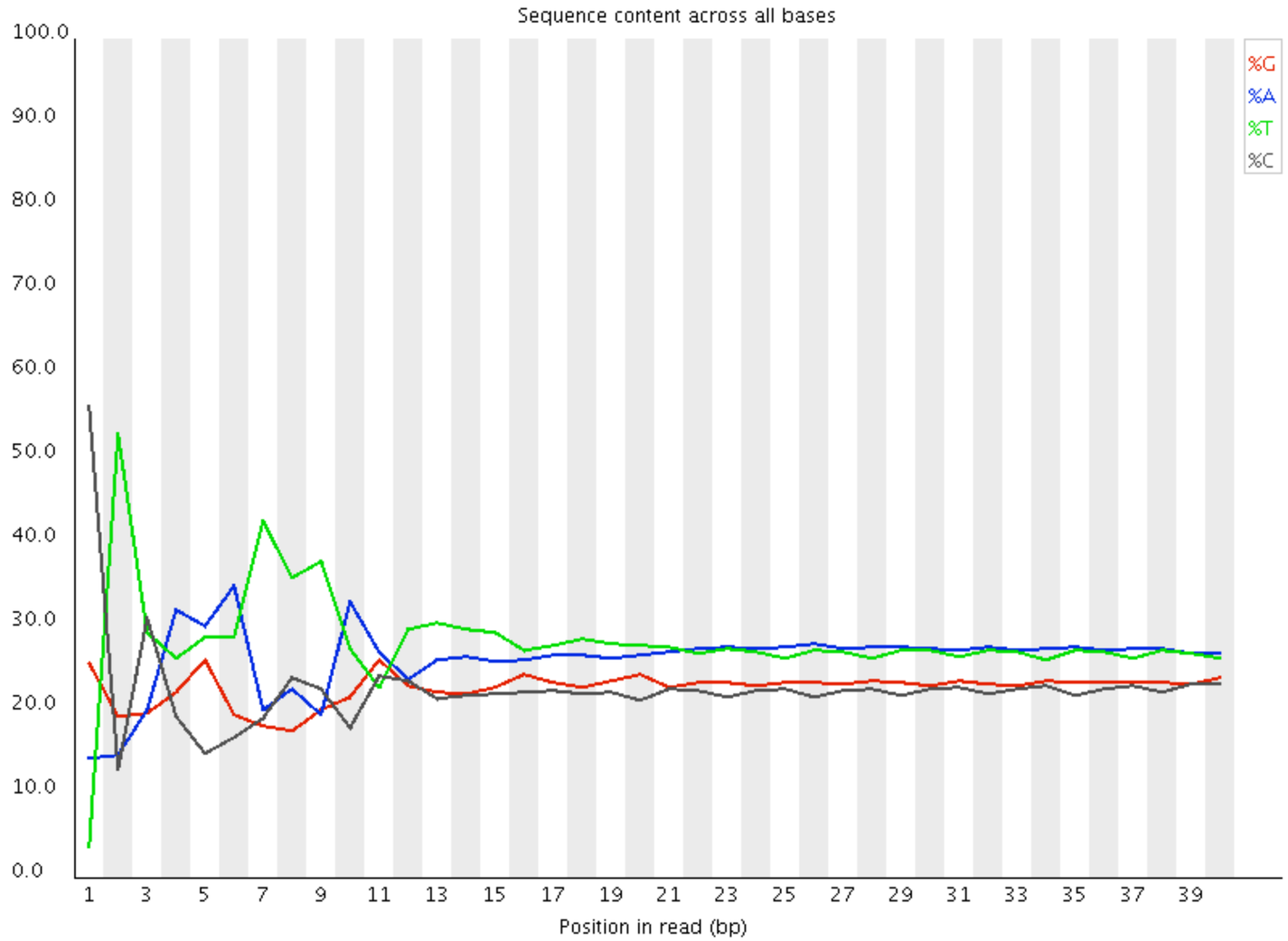




Illumina: flow cell clusters



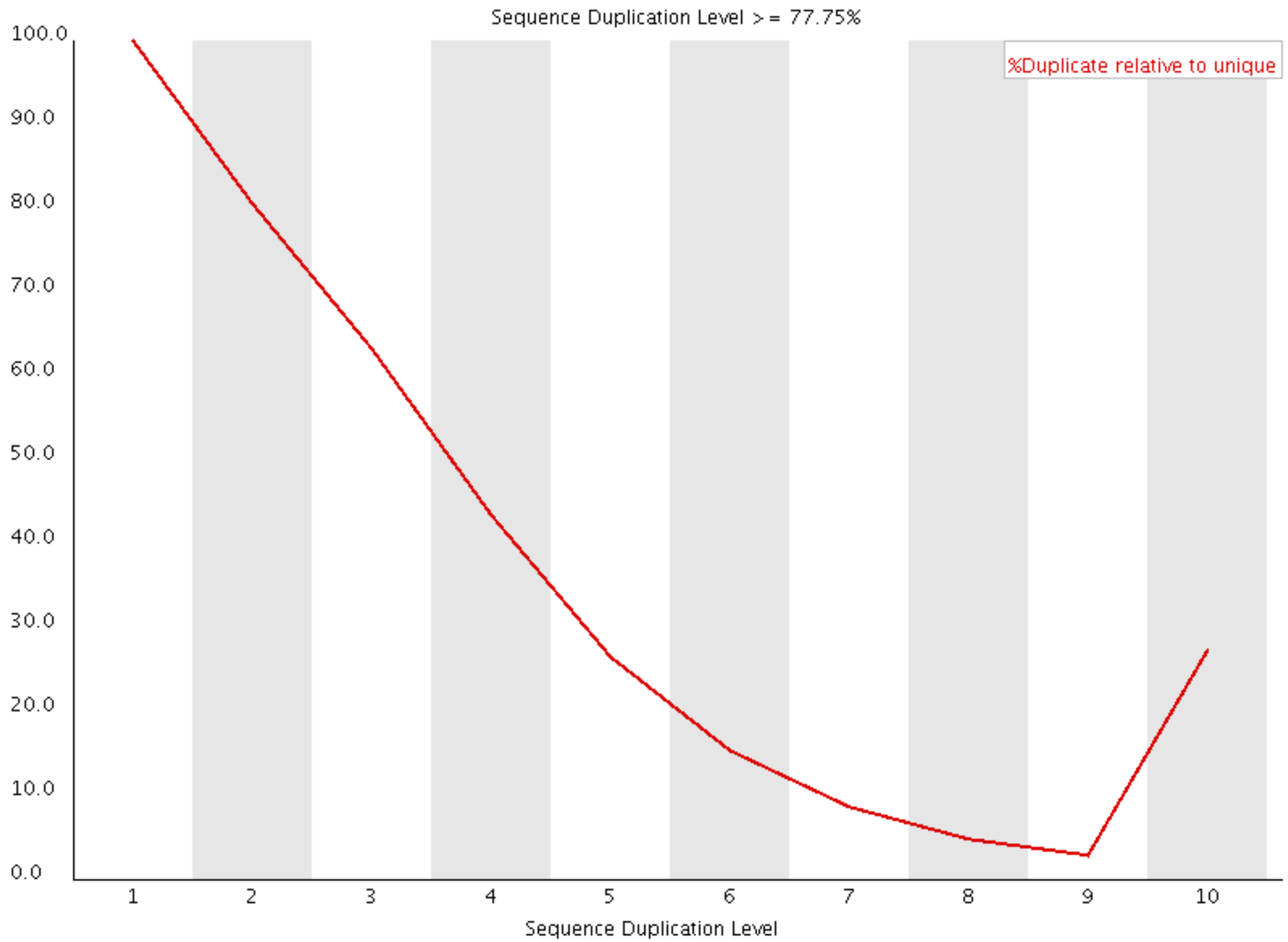
Illumina: optical effects



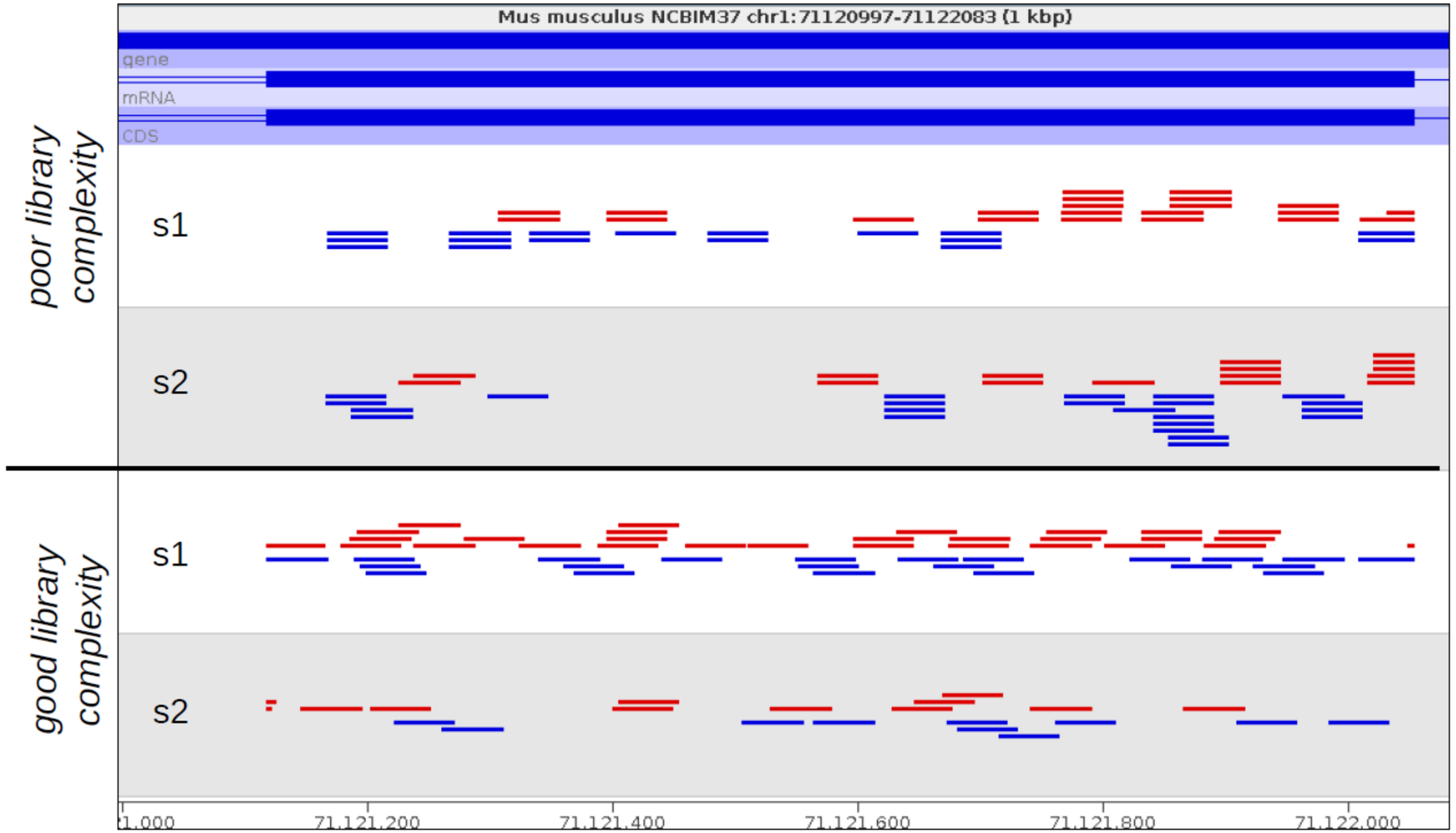
Positional sequence bias

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

PCR Artifacts



Duplicated sequences


Mus musculus NCBIM37 chr1:71120997-71122083 (1 kbp)
gene
mRNA
CDS
poor library complexity
s1
s2
good library complexity
s1
s2
1,000
71,121,200
71,121,400
71,121,600
71,121,800
71,122,000

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

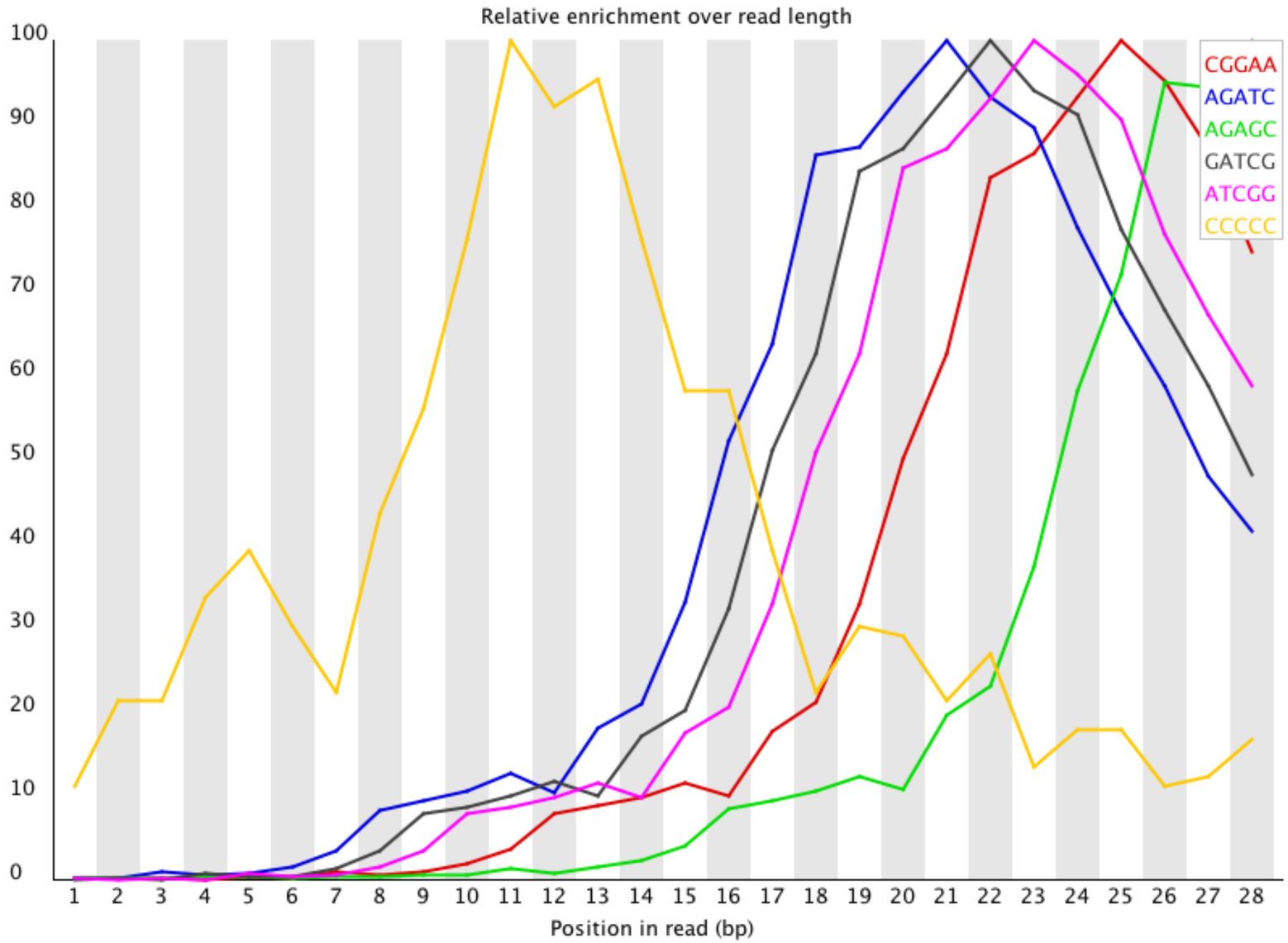
Contamination

```
> 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