

RNA-SEQ DATA ANALYSIS PIPELINE: QUALITY CONTROL

TRANSCRIPTOME RECONSTRUCTION

Guided reconstruction of the transcriptome

1. Quality control
2. Read mapping
3. Expression estimation
4. Differential expression estimation
5. Functional analysis

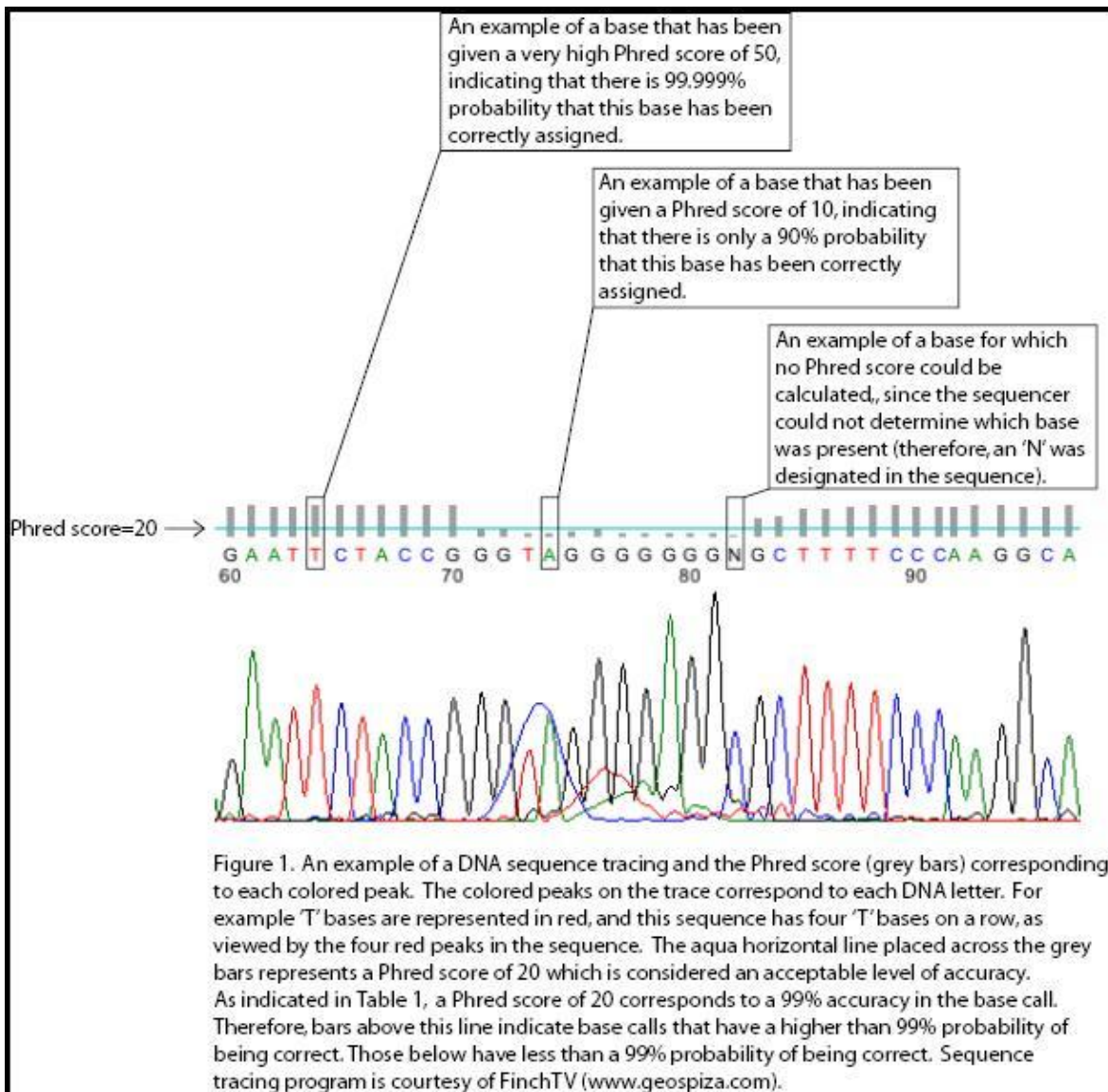
QUALITY CONTROL

```
@SEQUENCE1
GCCCCGGCGGGTTCATGCTGAAGAAAGGCGAAGTGTTTCGGTTGGGCGGC
+
ffffffffffe^eeceedffdc^dXecffbeed`Reebe`db\]XWSS
```

A FASTQ file employs 4 rows for each read:

- The first row starts with @ followed by the unique read identifier
- The second row contains the sequence.
- The third row starts with a + symbol and it might contain the read unique identifier
- The fourth row contains the quality scores for each nucleotide in the read sequence, encoded as decimal conversion of the ASCII code (**A**merican **S**tandard **C**ode for **I**nformation **I**nterchange) of the corresponding character (es.]=93,f=102).

QUALITY CONTROL



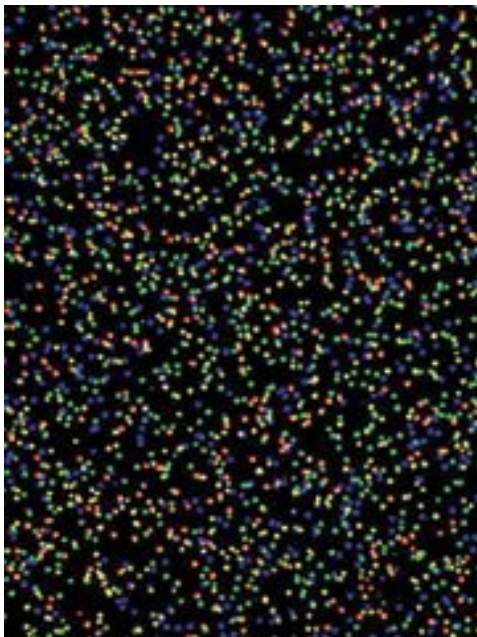
PHRED – **PHil's** **R**ead
EDitor (Phil Green)

PHRED quality =
 $-10 \times \log_{10} \text{Prob}(\text{Error})$



QUALITY CONTROL

- Software analyze each cycle images, interpret the fluorescence signals, make the base call, and estimate the probability that the base call is wrong



```
@NA12878:1463:NA12892:NA12891:F_IL20_290:1:80:114:644
TTTGCATTTAACAAATAATATGAGAACCGTTGACTG
+
6@<?3@@5@7@AAABB1A;;;BBABABB<@==<9/.
@NA12878:1463:NA12892:NA12891:F_IL20_290:3:97:342:584
GCATTTAACAAATAATATGAGAACCGTTGACTGAAA
+
@@AA@AAABAAABBABBABB>>BABAACA=@@A@<<
@NA12891:1463:::M_IL6_344:6:73:359:297.2
TTTCAGTCAACGGTTCTCATATTATTGTTAAATGC
+
????>>??@?@@@AAA;A@AAA@:@@AA@@;4-4;:
```

QUALITY CONTROL

- Each sequencing platform has its own way to evaluate sequencing errors, estimating the probability that the called nucleotide is wrong
- These probabilities are converted into a score analogous to the PHRED, then reported in compact form onto the sequencing platforms output
- Not all PHRED scores have the same meaning: different platforms use different scales, and a PHRED score 20 might mean good quality one platform and bad for another

QUALITY CONTROL

- Hence, each nucleotide has associated a quality score, called Phred score also for high-throughput sequencing platforms
- The Phred score is determined by platform-specific algorithms, and its range can also vary

Base Quality	$P_{\text{error}}(\text{obs. base})$
3	50 %
5	32 %
10	10 %
20	1 %
30	0.1 %
40	0.01 %

QUALITY CONTROL

Sequencing flavors:

- **Single end sequencing**: only one end of each DNA fragment is sequenced
- **Paired end sequencing**: both ends of a short (few hundreds base pairs) DNA fragment are sequenced
- **Mate pair sequencing** (jump libraries): both ends of a long (up to several Kbp) are sequenced, by circularization of a long DNA fragment, biotinylation of its junction, fragmentation and selection of the biotinylated fragments, which are then sequenced

QUALITY CONTROL

my_sequence.fastq

```
@HWI-BRUNOP16X_0001:1:1:1466:1018#0/1
AAGGAAGTGCTTGTCTGGCTAACACAGCNAGNCACGTGAC
+
aVfbe`^^^_TTSSdffffdfffabbZbbfebafbbbbbb
```

Single-end

my_sequence_1.fastq

```
@HWI-BRUNOP16X_0001:1:1:1278:989#0/1
NAAATTTCTGAATTTCTGTGAAGTAAGCATCTTCTTTGTCAT
+
BJJGGKIINN^^^^QNTUQ00TTTTRTOTY^^Y^\\^^^\\
```

Paired-end

my_sequence_2.fastq

```
@HWI-BRUNOP16X_0001:1:1:1278:989#0/2
AACCCACACAGGAGAGCAGCCTTACAGATGCAAATACTGTG
+
]K___fffffgggghgeggggggdgggggfggggggegggghh
```

QUALITY CONTROL

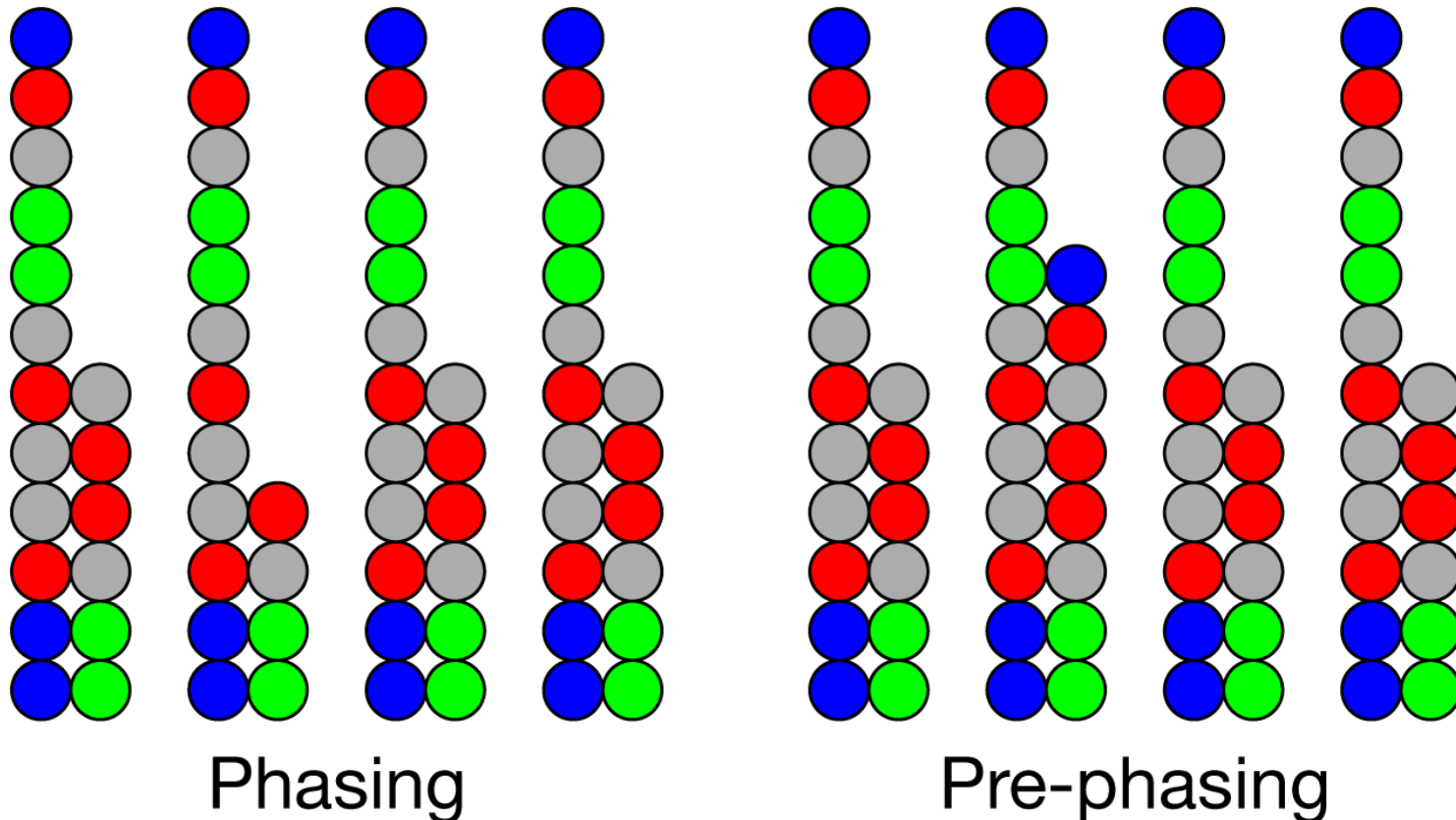
TYPE OF ERROR	EXAMPLE
Miscalled base	aaac ^c ttt c wrong
Insertion	tatt ^t caa extra t
Deletion	tgt-aag missing nt
Short homopolymer	aaa-taa missing a
Long homopolymer	gat ^{aaa} aat extra a
Homopolymer indel	gg ^g aag g should be a
Homopolymer incomplete extension	aaaaact ^a gg extra a
HMP incompl exten with indel	aact ^a gg a should be g
N's	ⁿ
Other – not easily recognized	tct-aaa mixed case

QUALITY CONTROL

- In the Illumina sequencing, the most common error type is **base substitution**. The overall accuracy is high, having an error rate less than **0.1%**. Indels are rare.
- The substitution frequency is not uniform for all possible substitutions, and the error frequency tends to increase towards read ends
- Substitution patterns might be dependent on the genomic fragment base composition: base substitutions are more frequent at high GC content, while insertion and deletion rates increase in homopolymeric regions and in the presence of inverted repeats

QUALITY CONTROL

- The increase of base substitution rates when progressing through the reaction cycles is due to the loss of synchronization of the amplified fragments in a cluster (**phasing**)

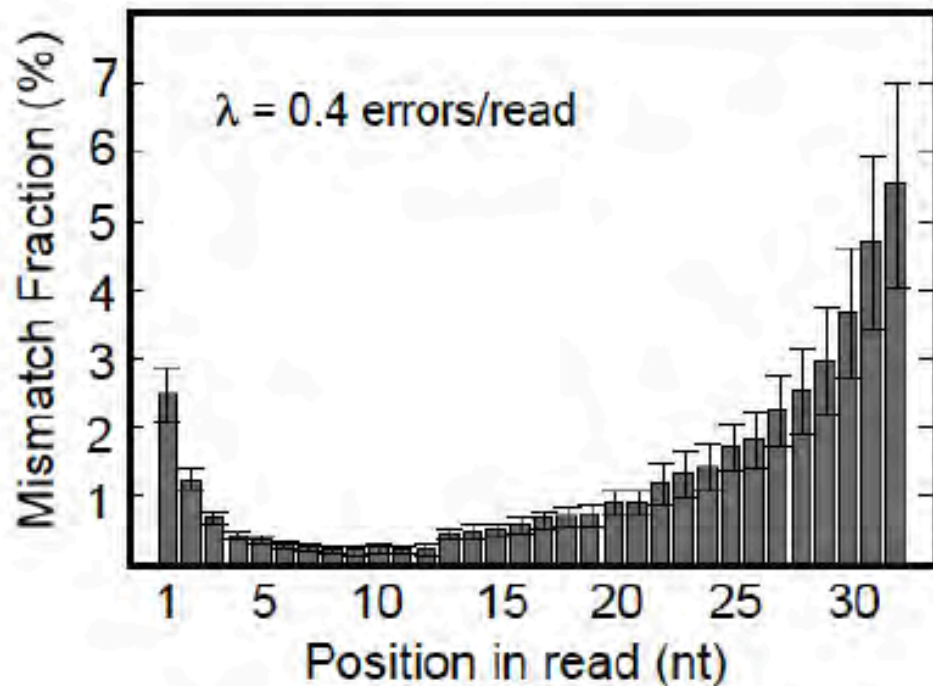


QUALITY CONTROL

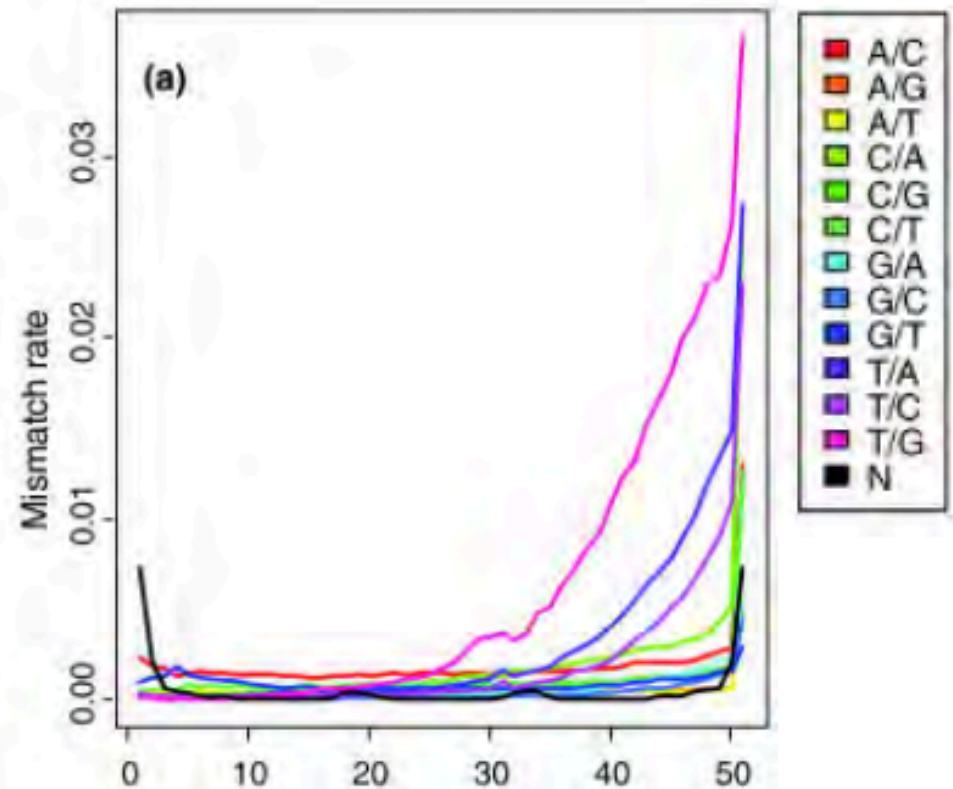
Post-phasing and pre-phasing are caused by:

- incomplete removal of the 3' terminators and fluorophores
- sequences in the cluster missing an incorporation cycle
- incorporation of nucleotides without effective 3' terminators or fluorophores

QUALITY CONTROL



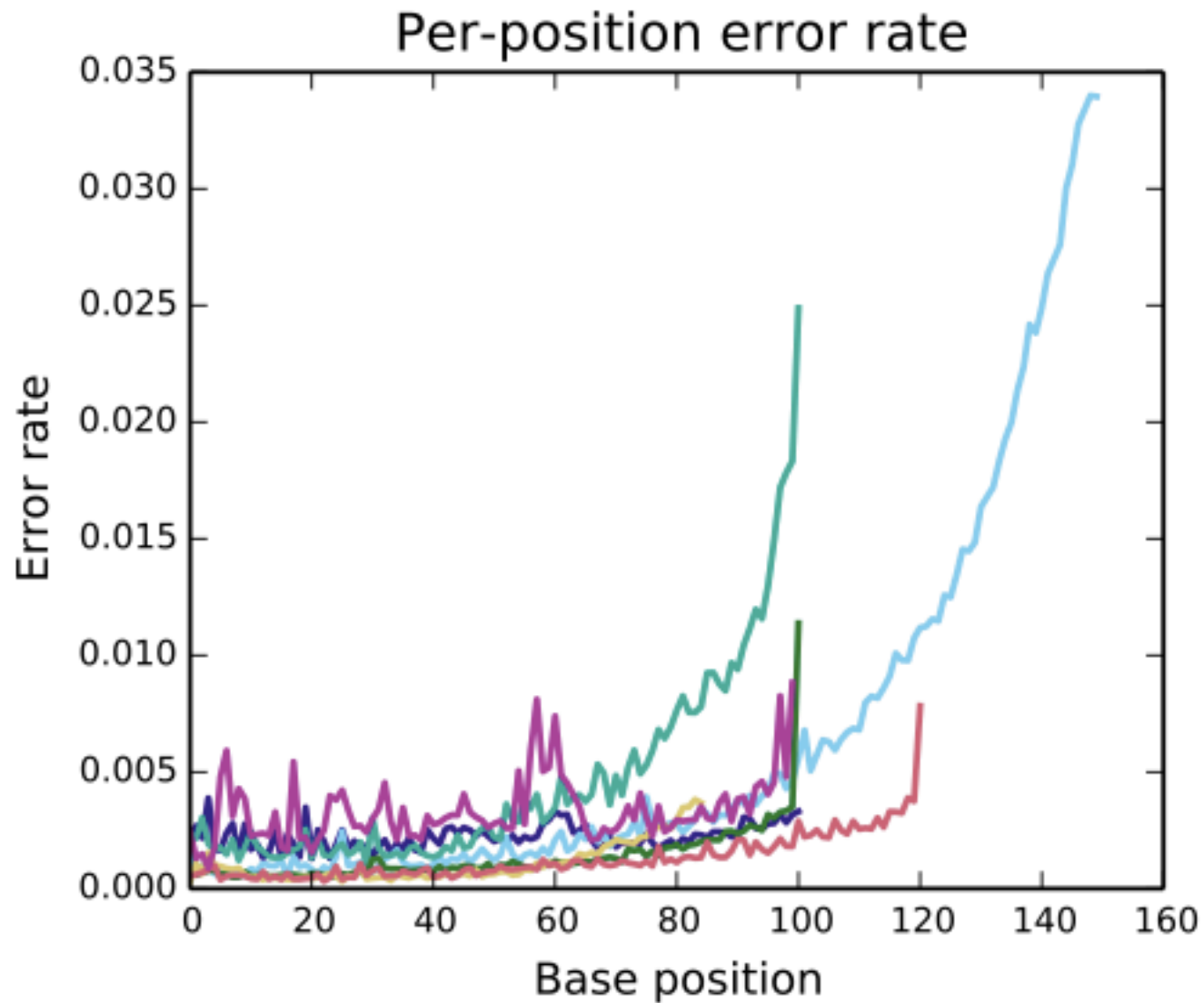
Wang et al, *Nature* 456: 470, 2008



Kircher et al. *Genome Biology* 2009

Error rate is high in the first 1-2 bases, and usually increases dramatically towards the read end

QUALITY CONTROL



QUALITY CONTROL

What should you do when you receive the sequencing output files (i.e. the *fastq* files):

1. Check **read quality**
2. Identify the presence of general issues (problems during the run, contaminants)
3. Discard whole reads, or part of reads, having low quality (**trimming**); discard reads that are too short; remove adapter sequences

QUALITY CONTROL

FastQC

<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>



Babraham Bioinformatics



[About](#) | [People](#) | [Services](#) | [Projects](#) | [Training](#) | [Publications](#)

FastQC

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later .
Initial Contact	Simon Andrews
Download Now	

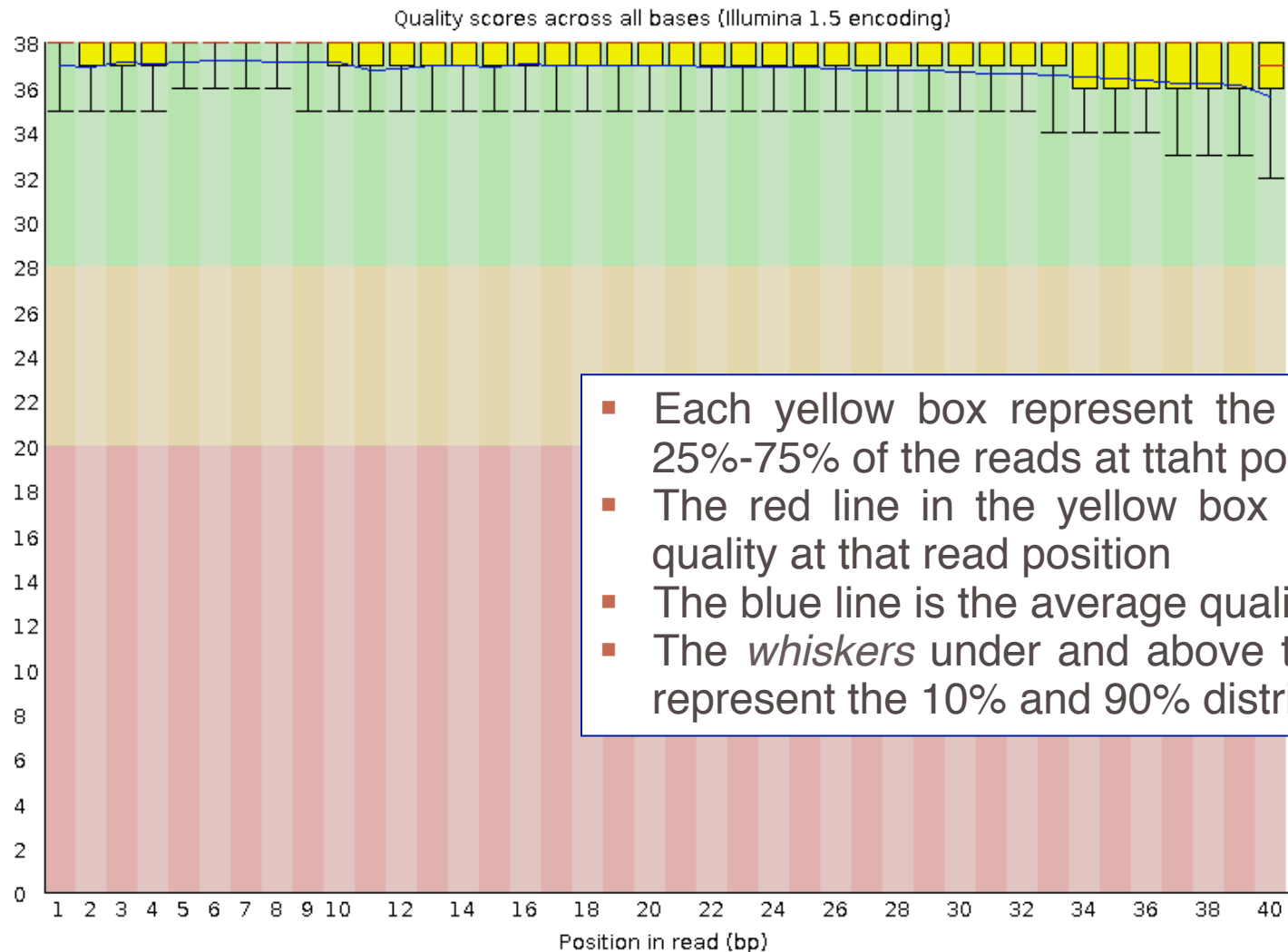
QUALITY CONTROL

FastQC output:

- Average read quality distribution
- Average read quality per position
- Nucleotide content per position
- GC content distribution per read
- GC content per position
- Undetermined bases (N) per position
- Read length distribution
- Over-represented/duplicate sequences
- K-mers content

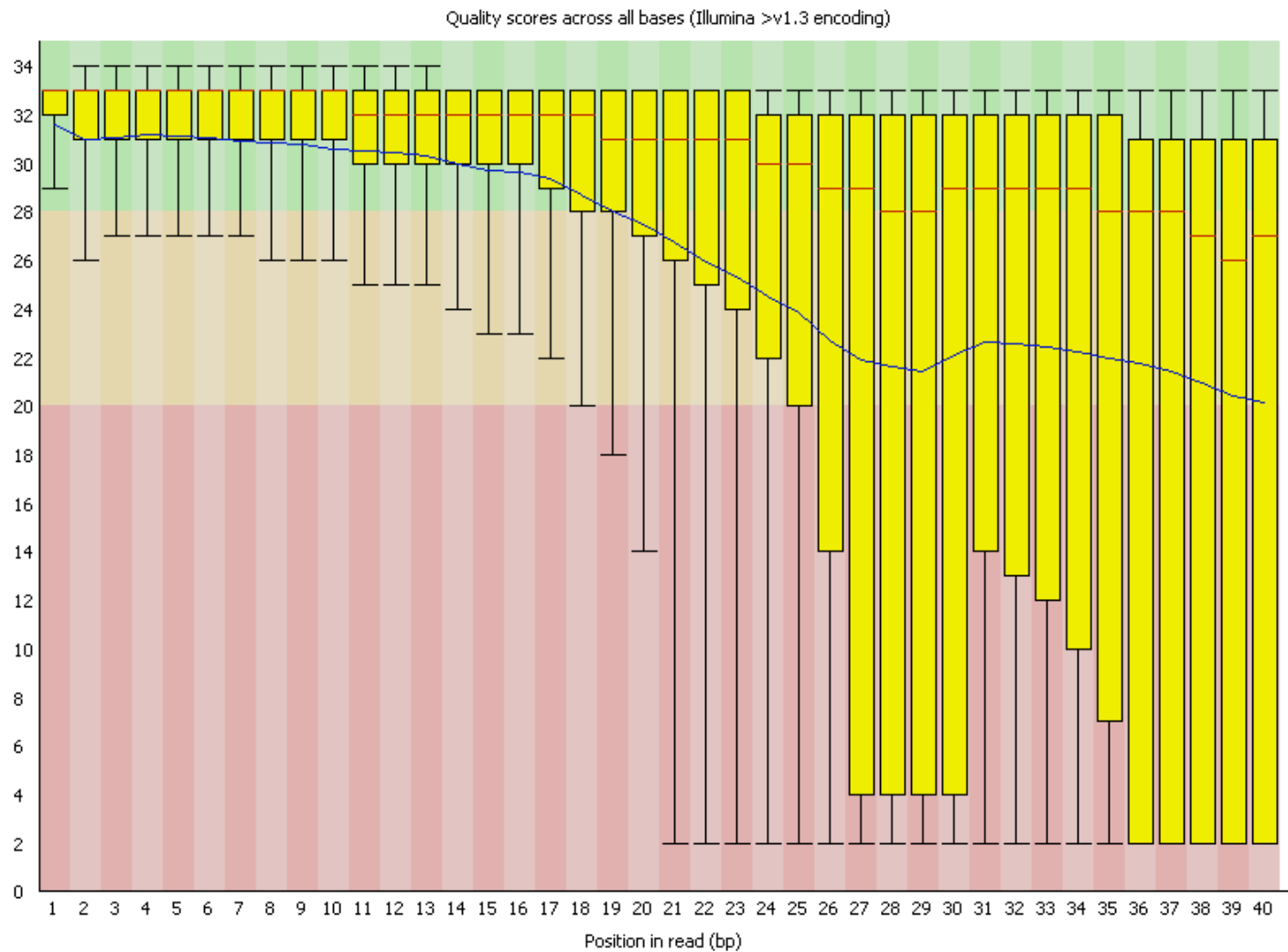
QUALITY CONTROL

Average read quality per position



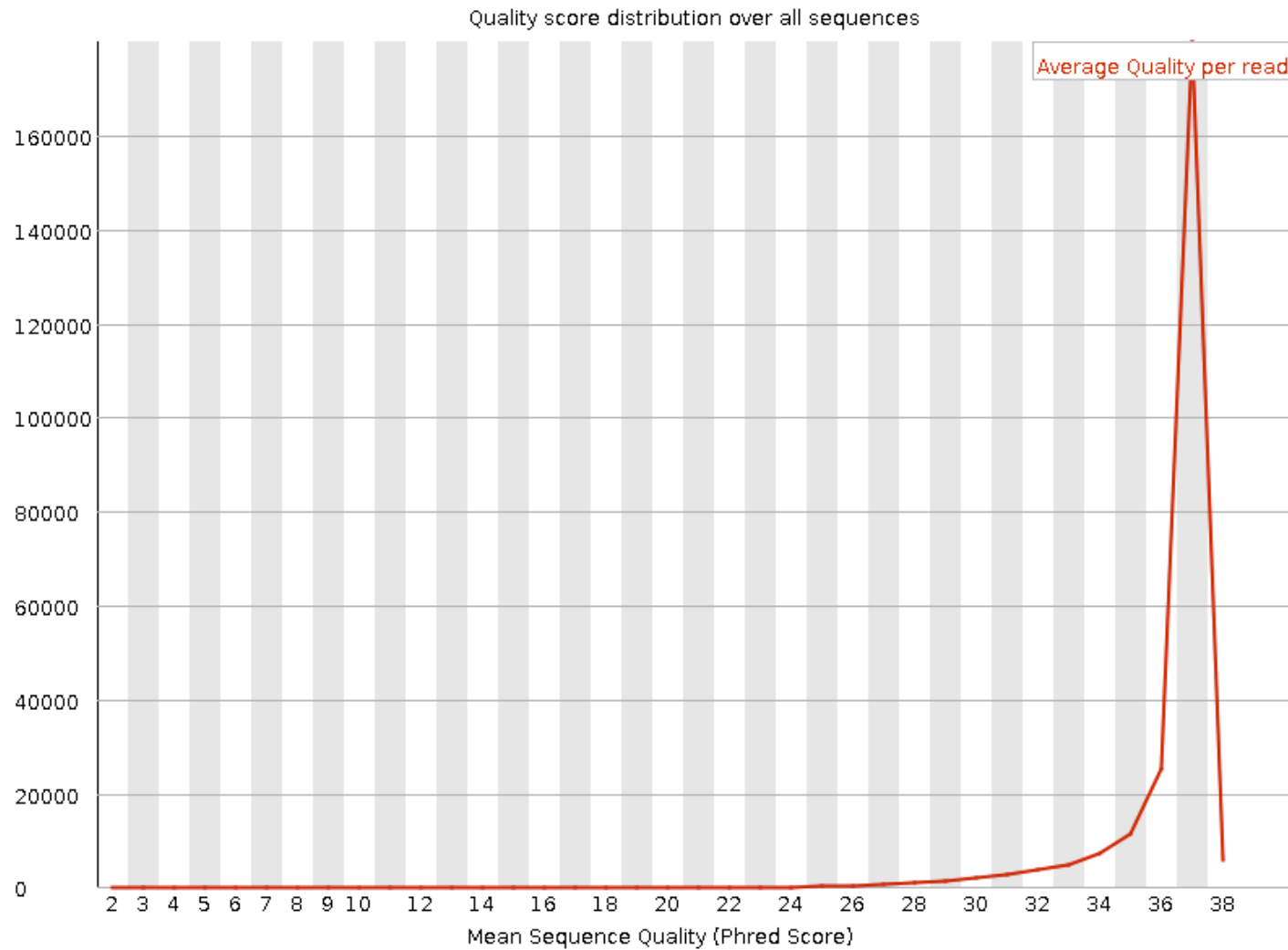
QUALITY CONTROL

Average read quality per position



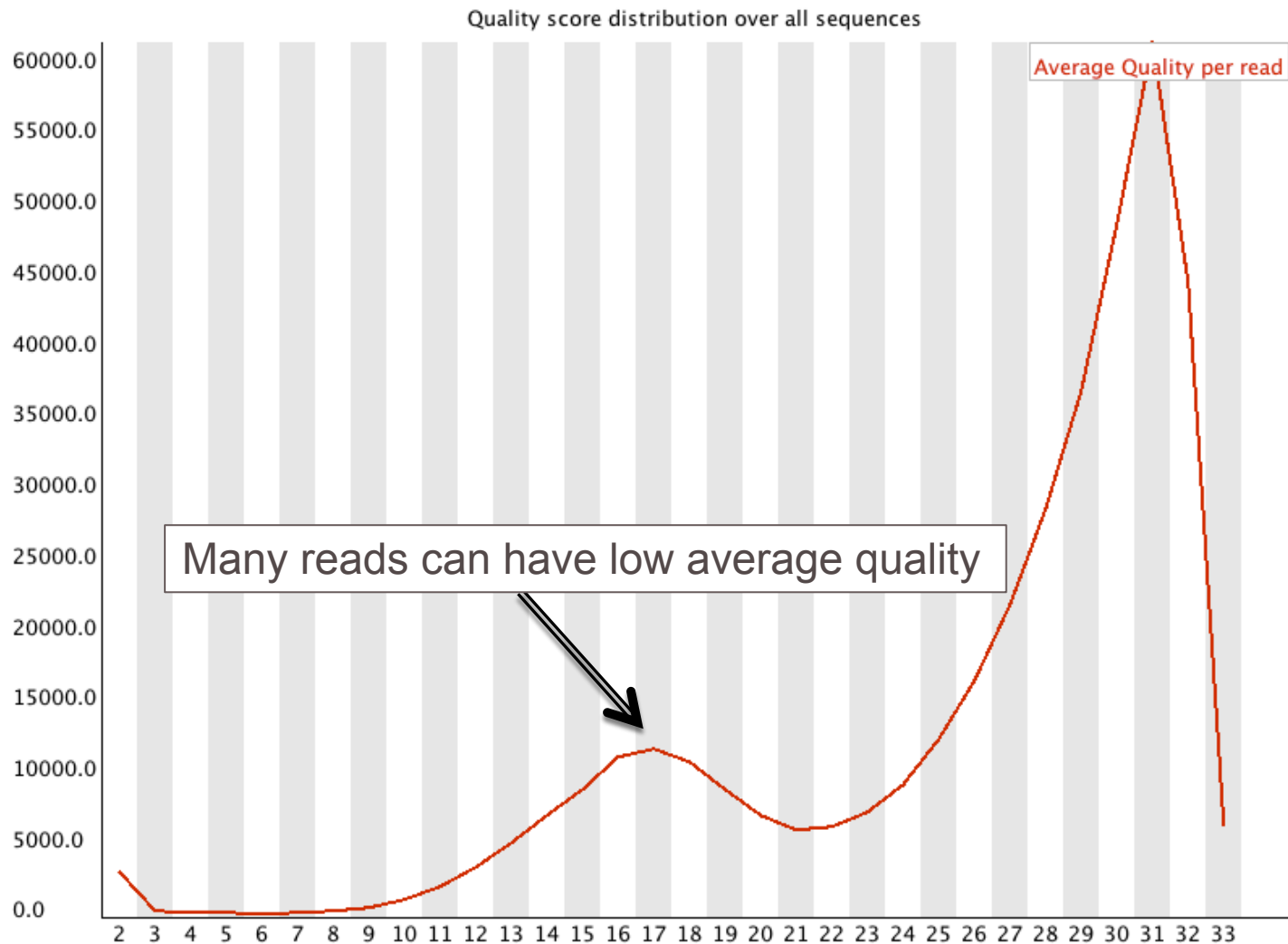
QUALITY CONTROL

Read average quality distribution



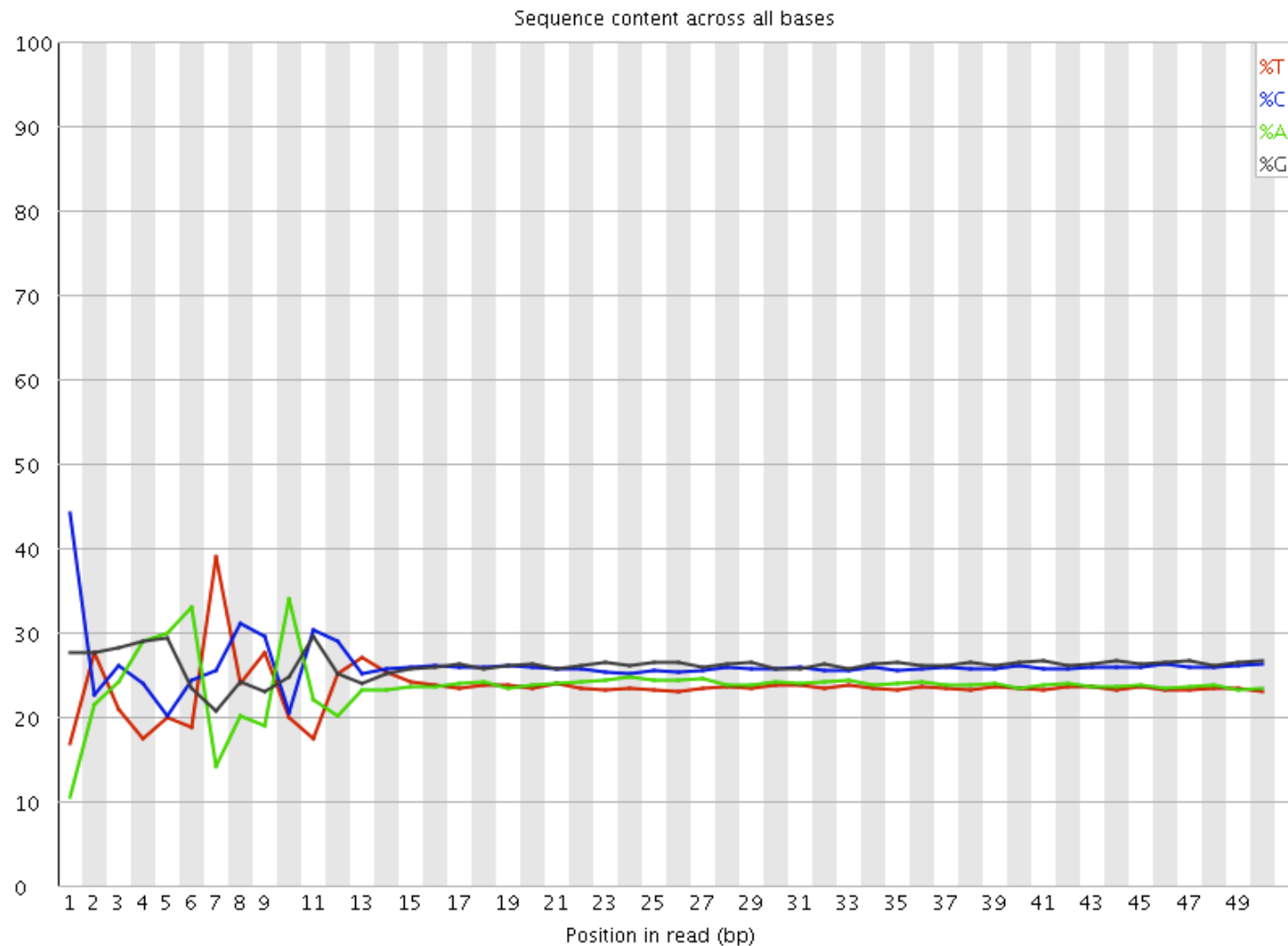
QUALITY CONTROL

Read average quality distribution



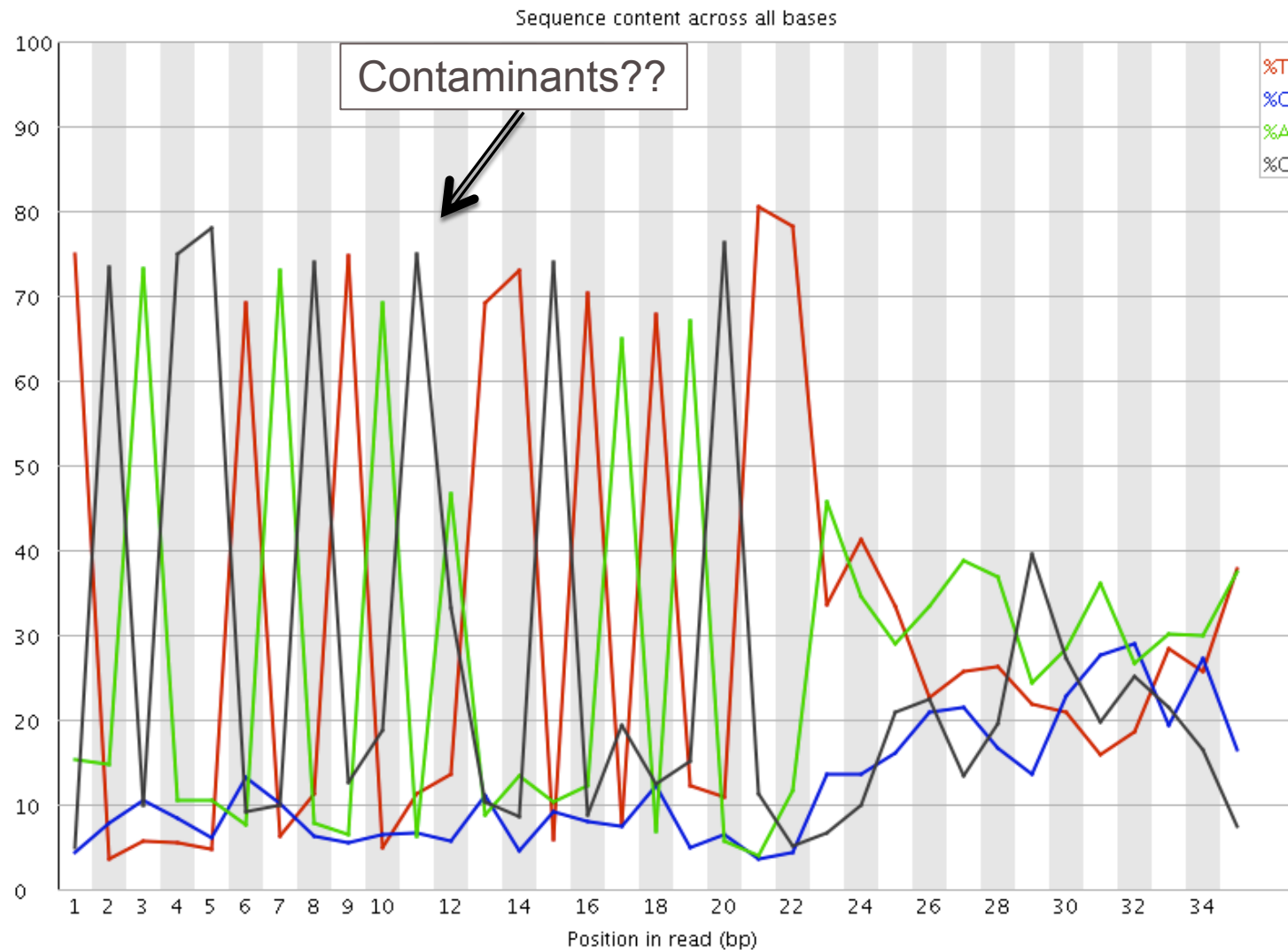
QUALITY CONTROL

Nucleotide content per position



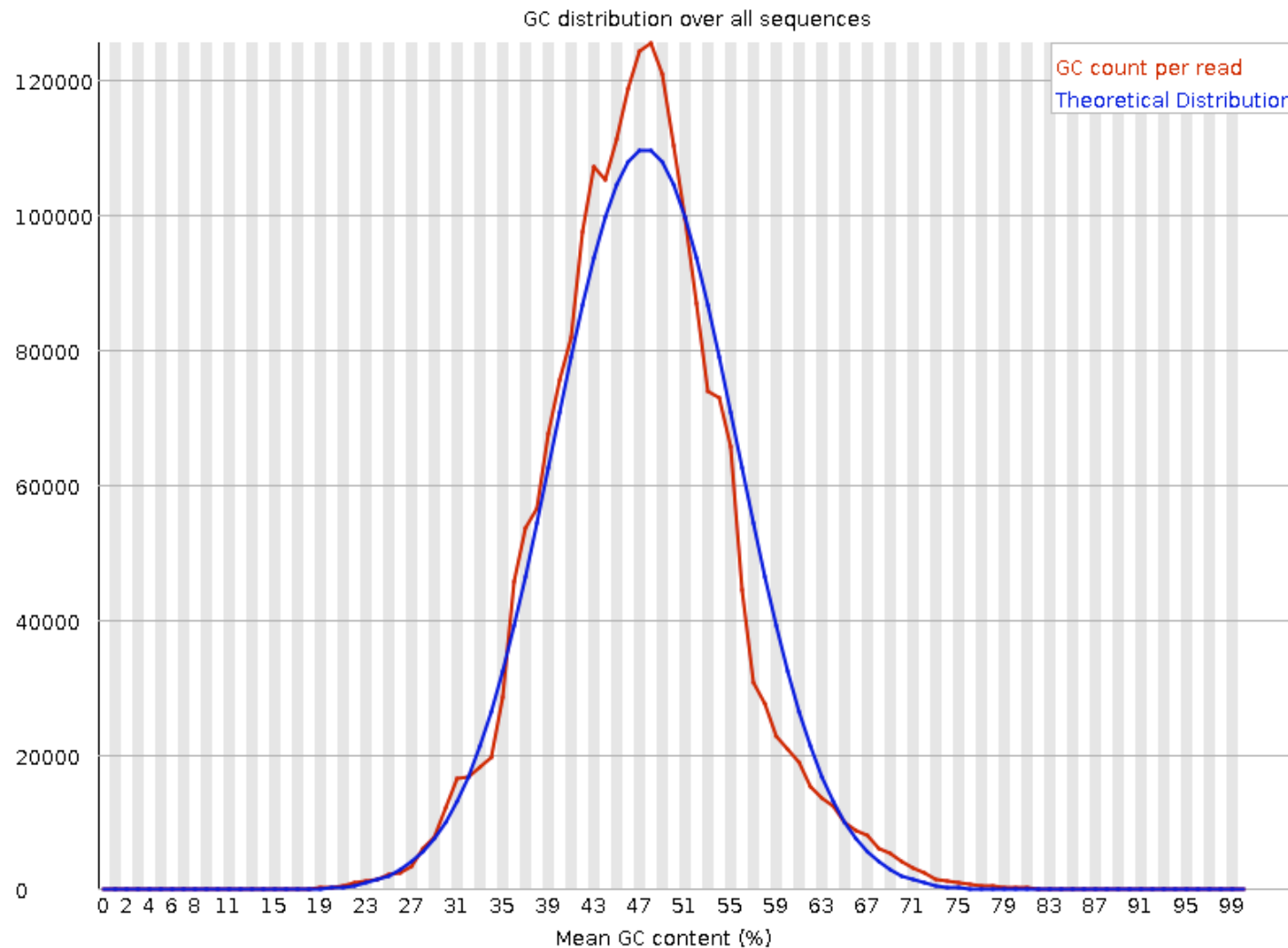
QUALITY CONTROL

Nucleotide content per position



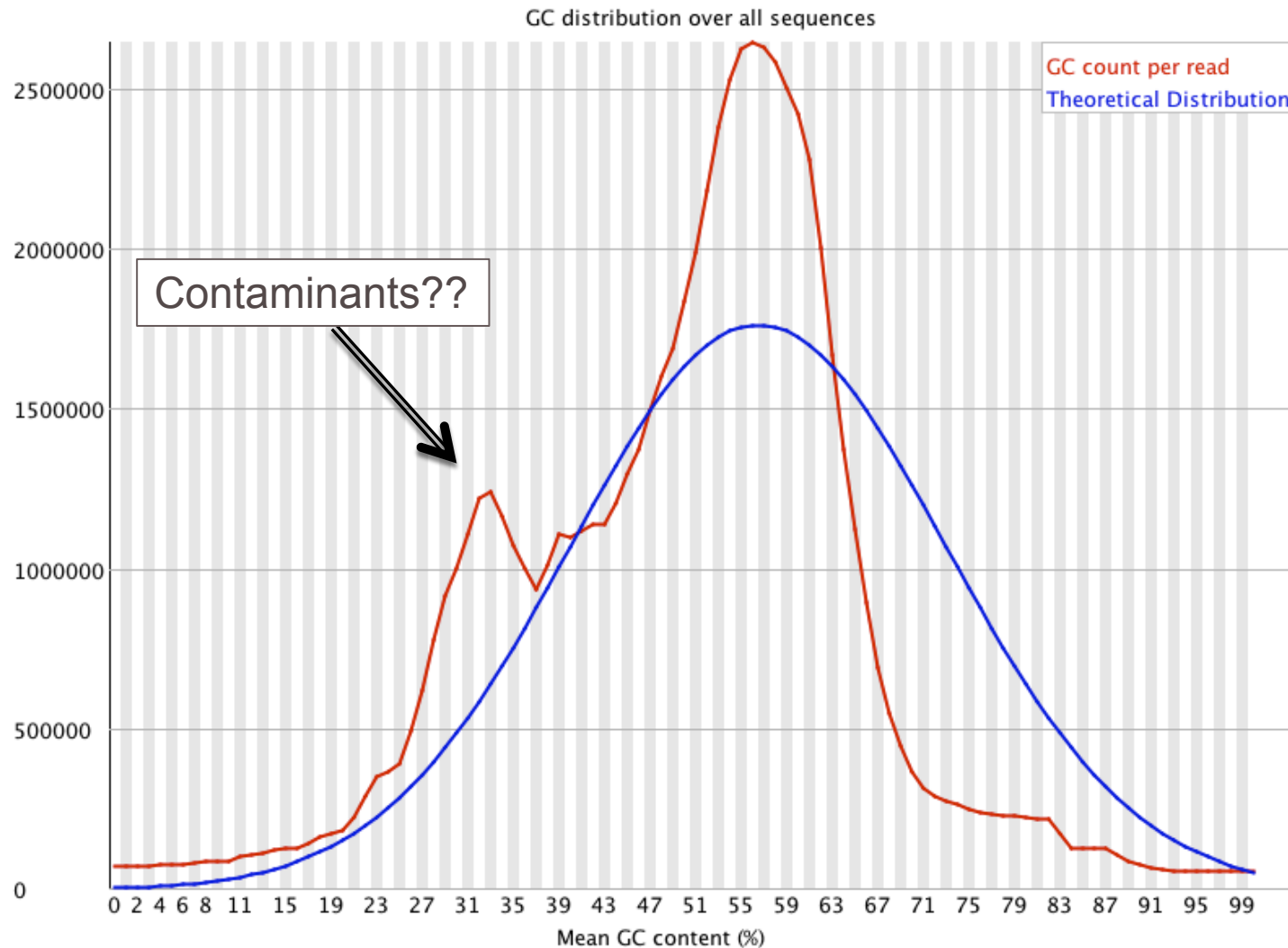
QUALITY CONTROL

GC content distribution per read



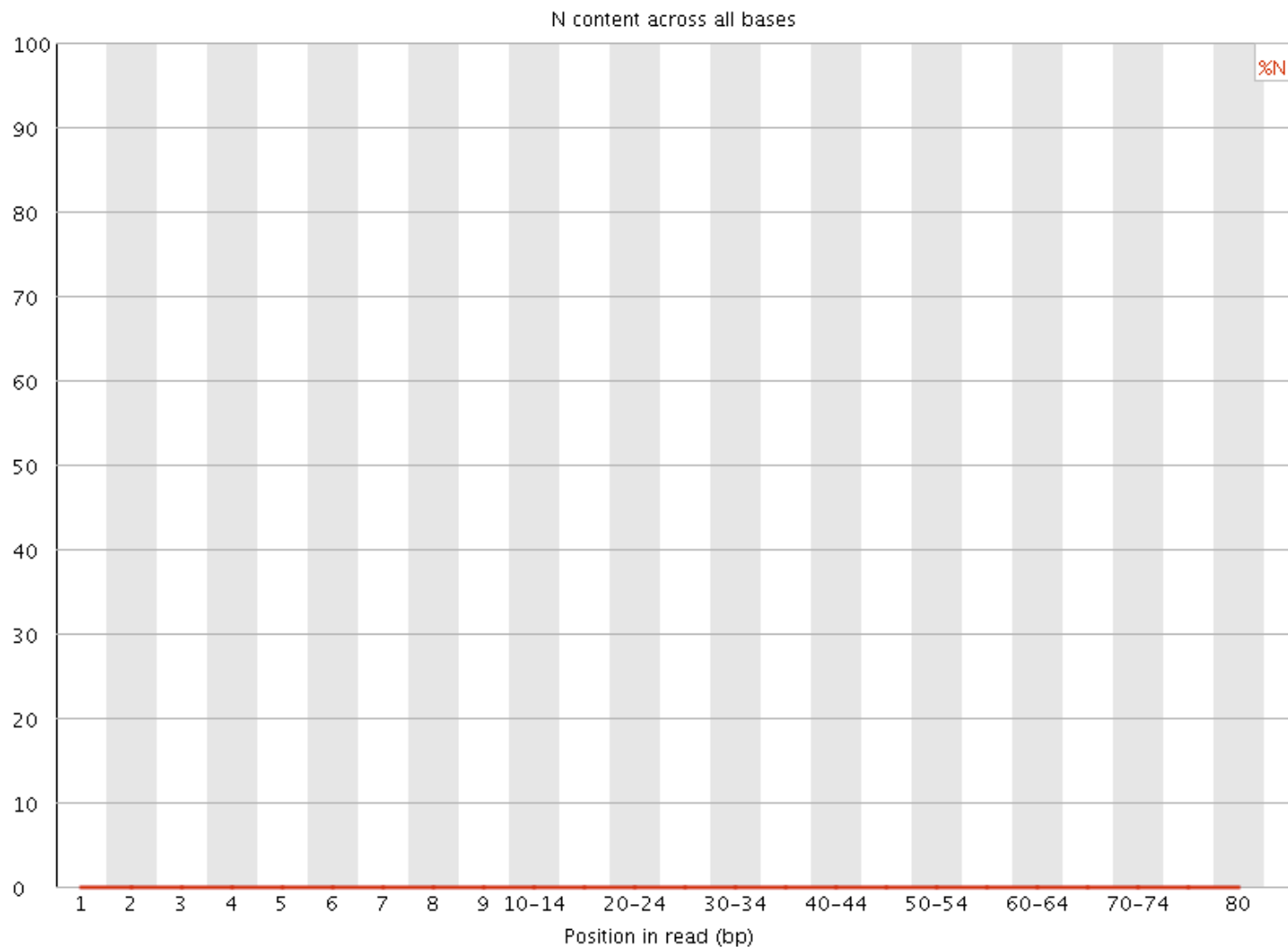
QUALITY CONTROL

GC content distribution per read



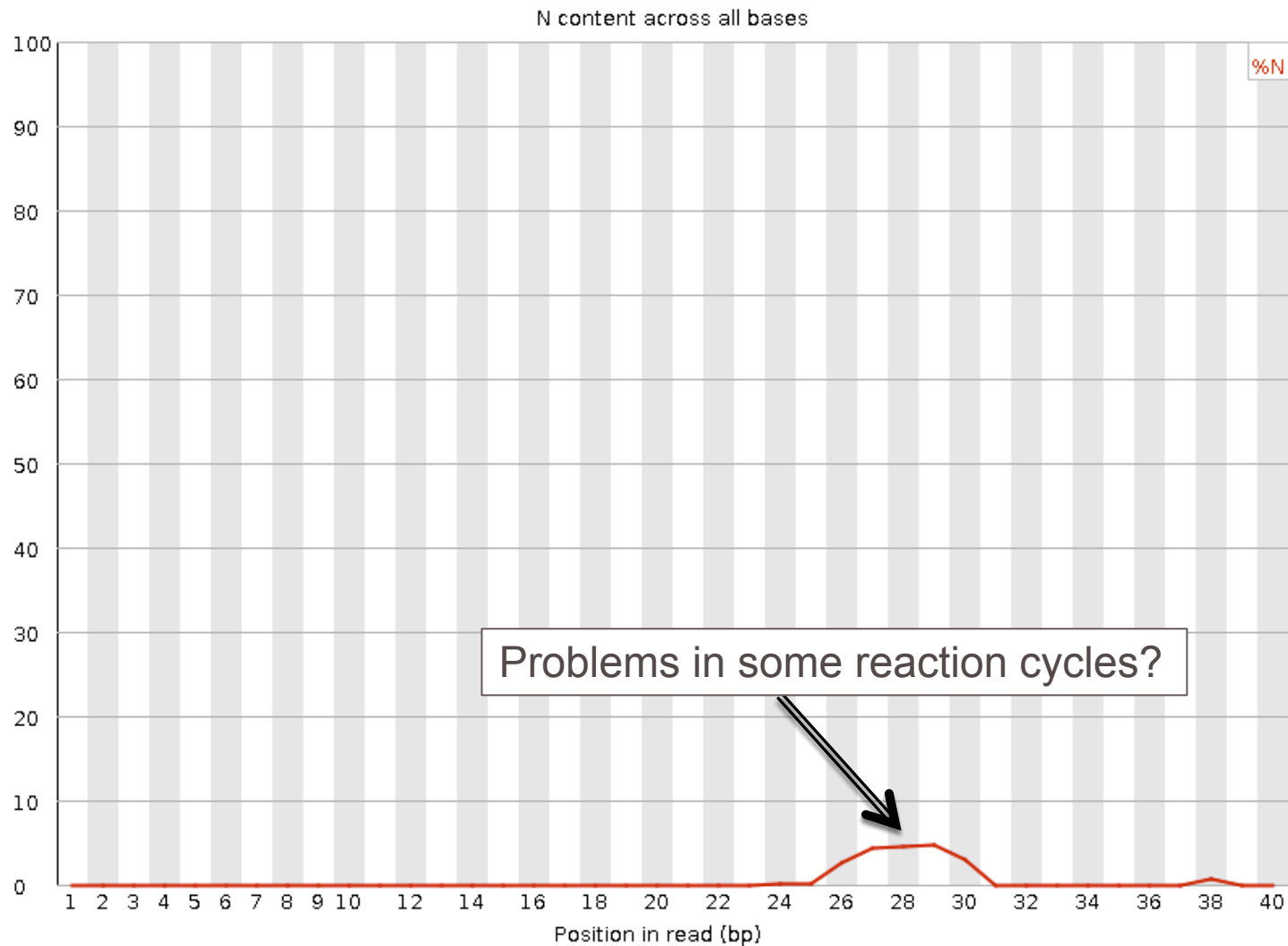
QUALITY CONTROL

Undetermined bases (N) per read position



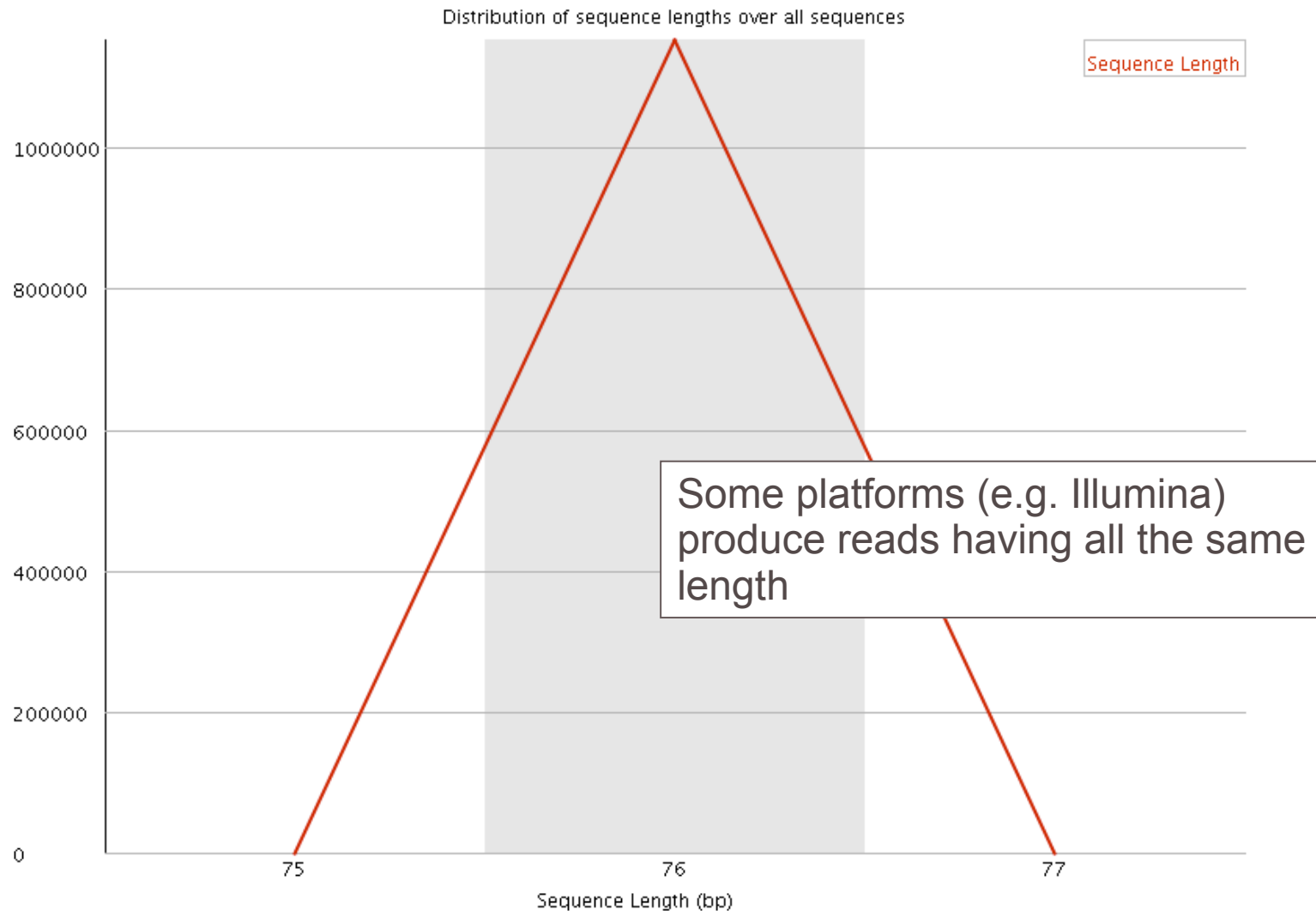
QUALITY CONTROL

Undetermined bases (N) per read position



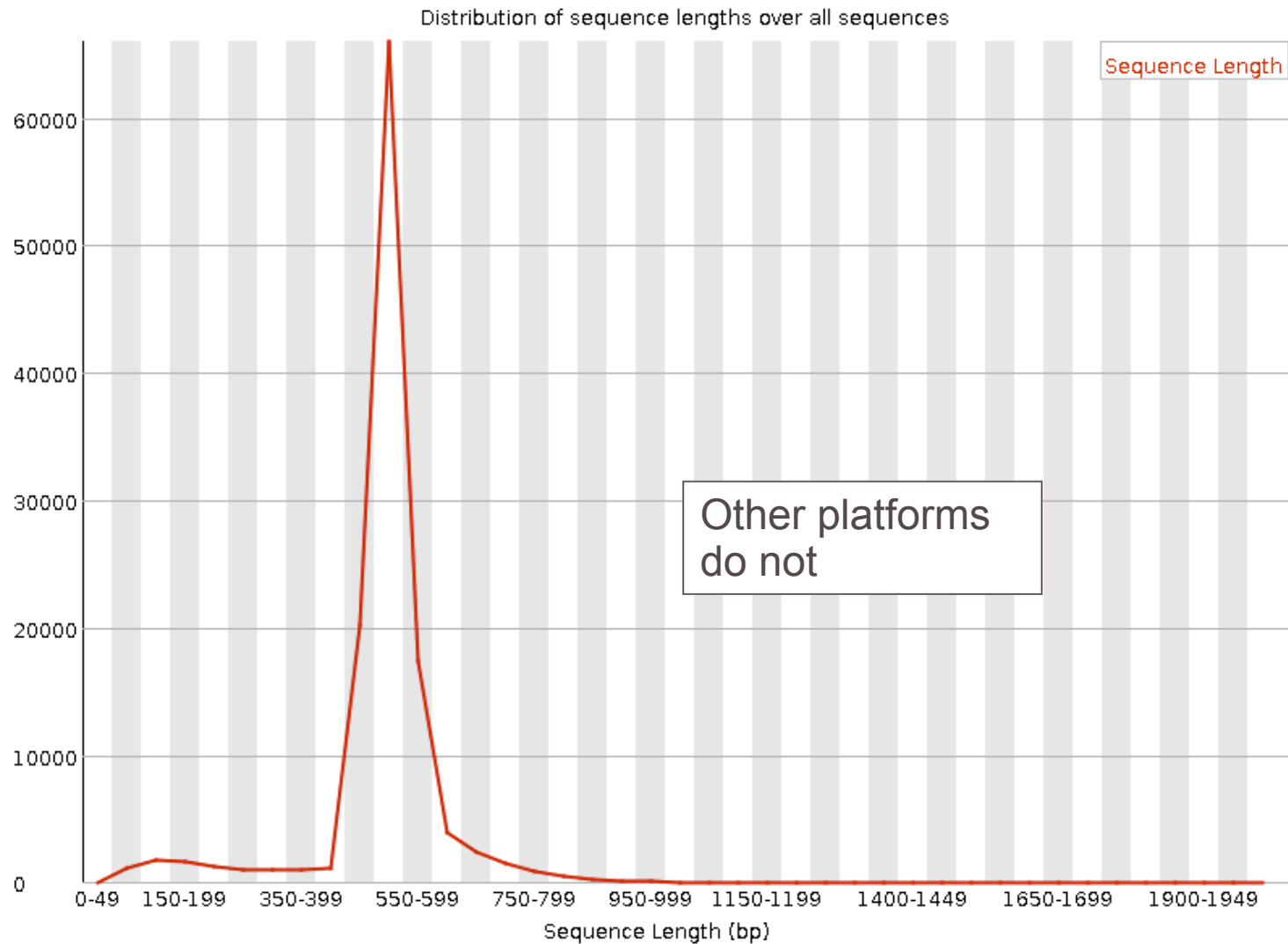
QUALITY CONTROL

Read length distribution



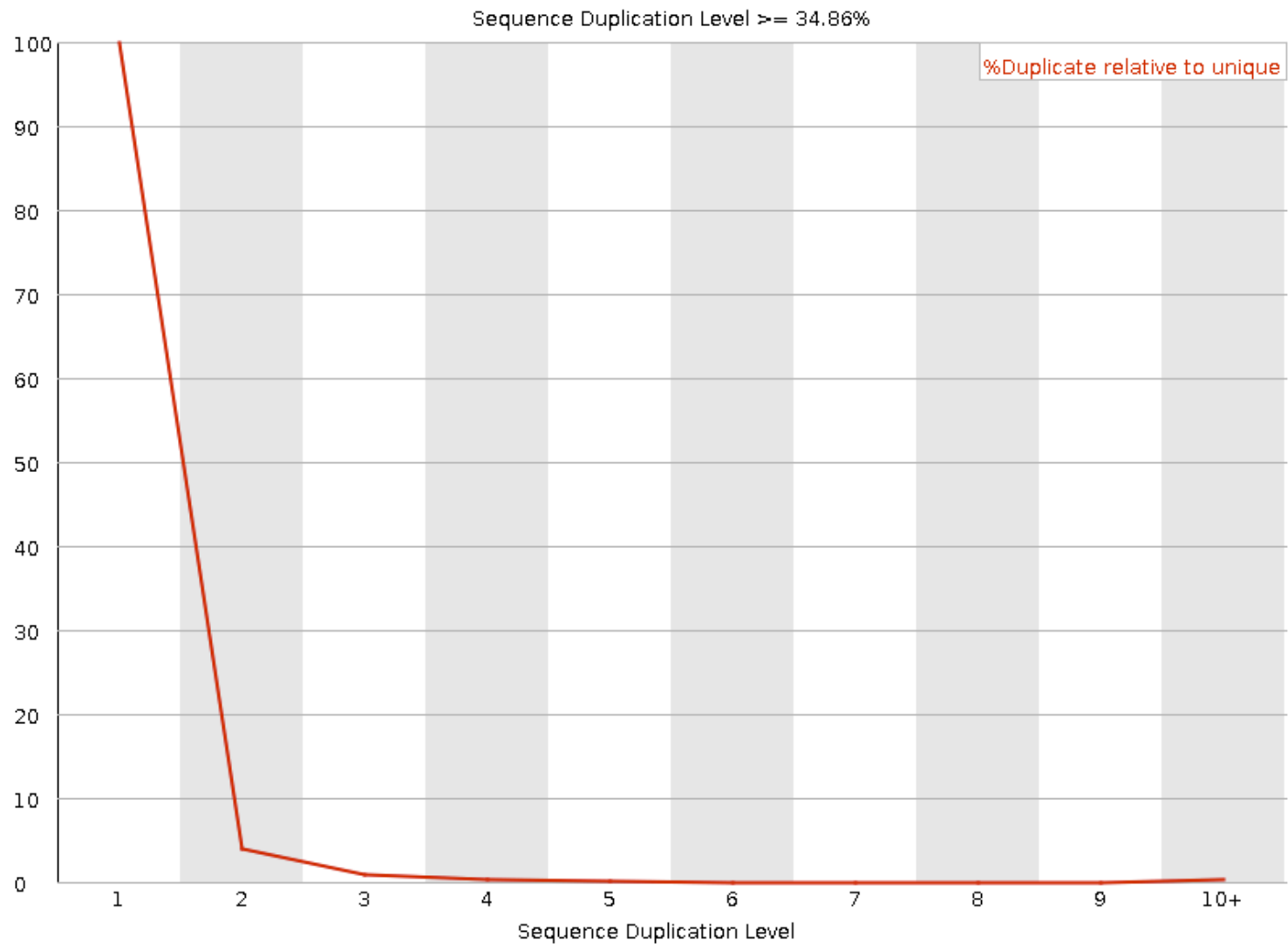
QUALITY CONTROL

Read length distribution



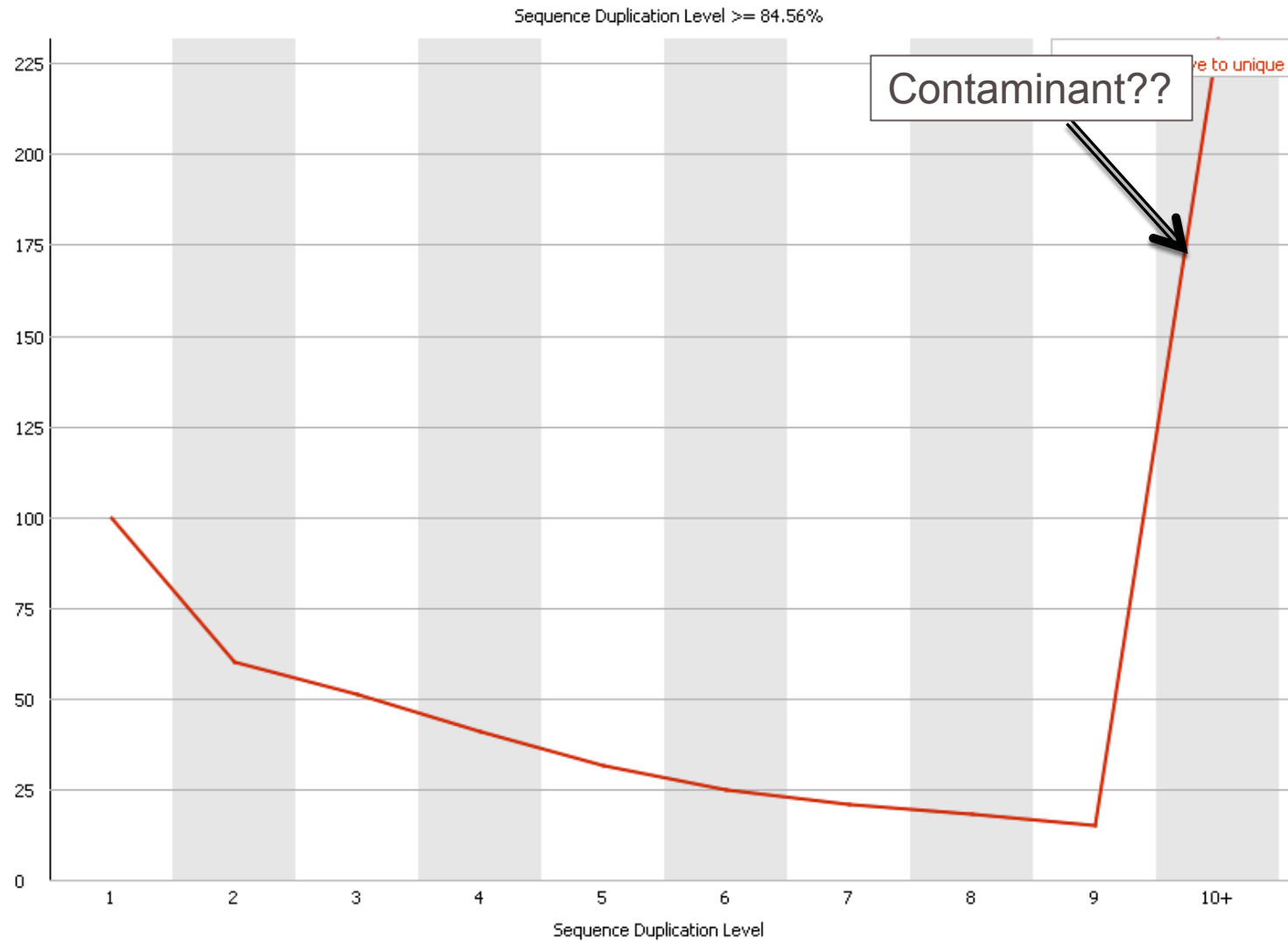
QUALITY CONTROL

Read duplication level



QUALITY CONTROL

Read duplication level



QUALITY CONTROL

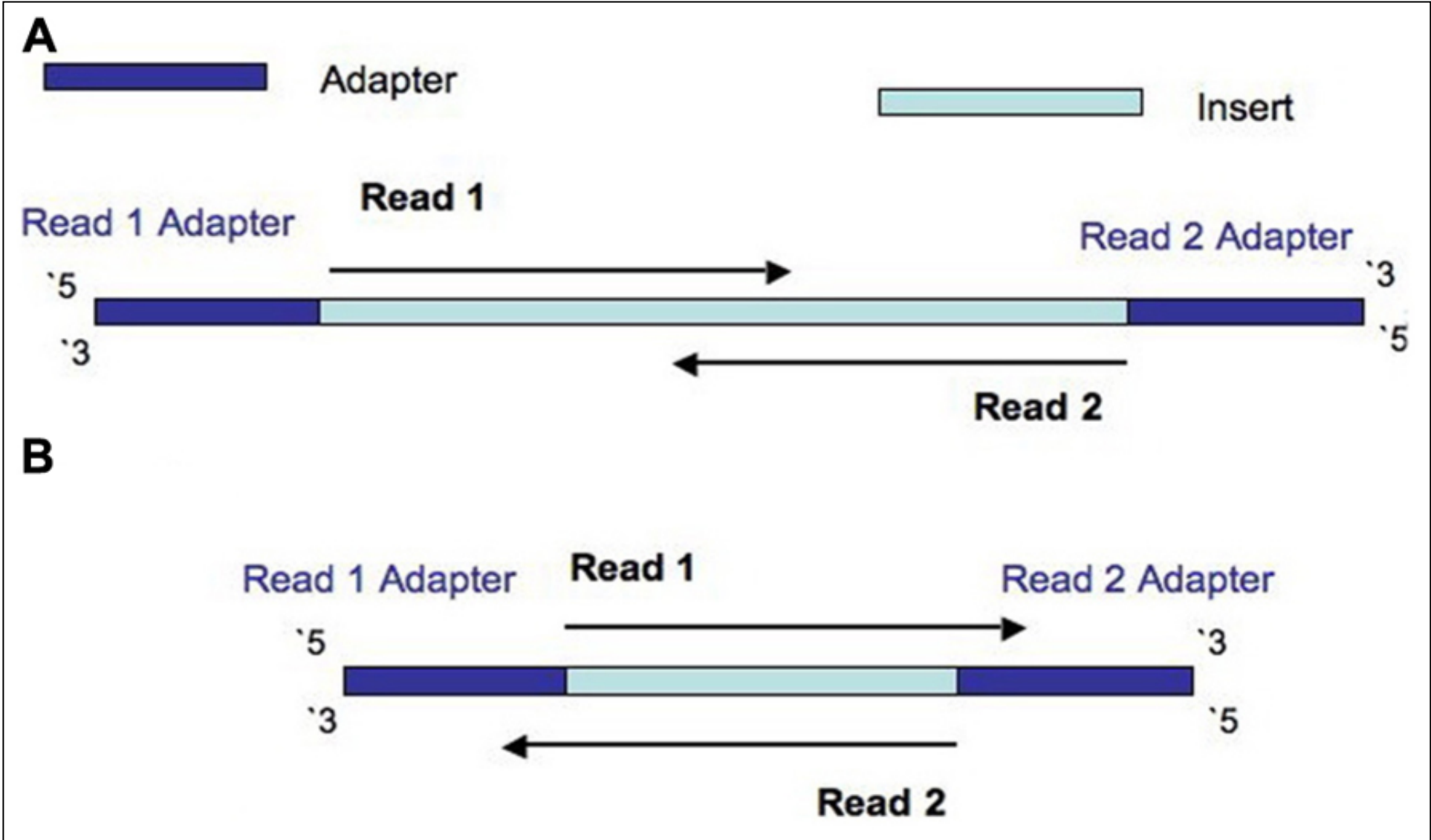
✖ Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACA	1060621	29.432719567181643	TruSeq Adapter, Index 5 (100% over 36bp)
GCTAACAAATACCCGACTAAATCAGTCAAGTAAATA	13630	0.37823875606902535	No Hit
NATCGGAAGAGCACACGTCTGAACTCCAGTCACACA	11728	0.3254573830651159	TruSeq Adapter, Index 5 (97% over 36bp)
GTTAGCTATTTACTTGACTGATTTAGTCGGGTATTT	10983	0.304783291115635	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACC	3658	0.10151117899490057	TruSeq Adapter, Index 1 (97% over 36bp)

✖ Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AACTTTCAACAACGGATCTCTGGCTCTCGCATCGATGAAGAACGCAGCGA	1443252	21.105778961197885	Search with Blastall+, more detail First hit on +100 : Hydrocybe nitida isolate DJL05NC65c5 voucher TENN:61894 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and large subunit ribosomal RNA gene, partial sequence Evalue=3.87318e-19, Ident=100%, QueryCovergap=0%
CAACTTTCAACAACGGATCTCTGGCTCTCGCATCGATGAAGAACGCAGCG	395417	5.782485525396801	Search with Blastall+, more detail First hit on +100 : Hydrocybe nitida isolate DJL05NC65c5 voucher TENN:61894 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and large subunit ribosomal RNA gene, partial sequence Evalue=3.87318e-19, Ident=100%, QueryCovergap=0%
AATGGCTTAAGGATCCGTAGAATGACATGTAATATAACATGTAAGAGTT	106935	1.5637923752856022	Search with Blastall+, more detail First hit on 8 : Issatchenkia orientalis strain NRRL Y-5396 small subunit ribosomal RNA gene, partial sequence; mitochondrial Evalue=3.87318e-19, Ident=100%, QueryCovergap=0%
			Search with Blastall+, more detail First hit on +100 : Candida sp. 147-2013a voucher Del. 12000 internal transcribed spacer 1

QUALITY CONTROL



QUALITY CONTROL

- When you have replicates for each biological condition, all replicates should show similar quality control reports, especially for GC content and k-mer content
- In case one replicate shows patterns remarkably different from the other replicates in the same group, it must be considered cautiously
- Analysis of the principal components (PCA) should cluster together biological replicates, and can be employed to detect anomalies
- In extreme cases, it is better to discard these peculiar replicates than to carry them forward in the pipeline

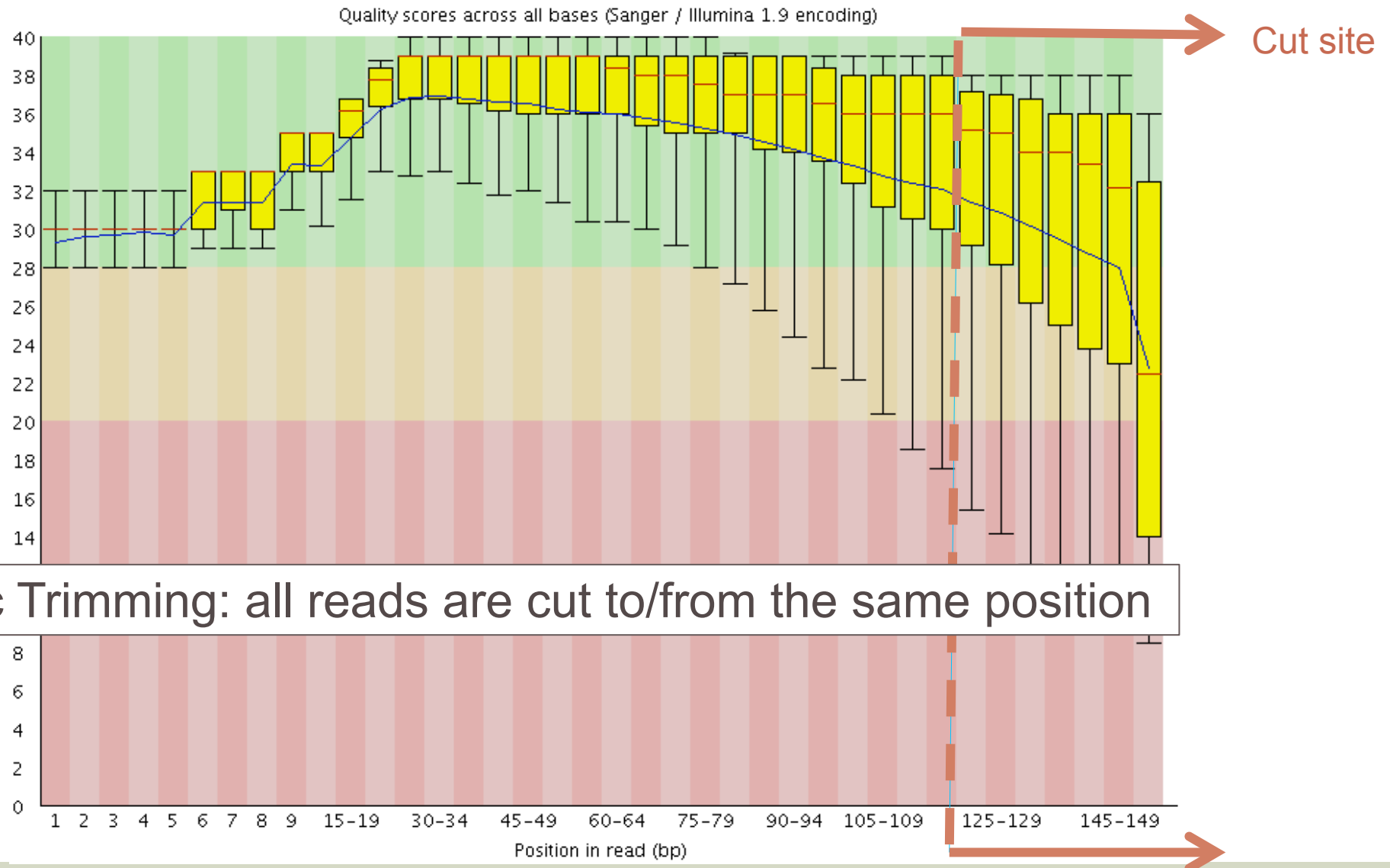
TRIMMING

Based on the quality control results, you could decide to clean up the dataset to facilitate and make more accurate the following steps of the analysis. You can:

- Discard read ends having low quality
- Discard reads that after this step remain too short to have an unambiguous alignment
- Discard whole reads having low average quality
- Remove adapter sequences

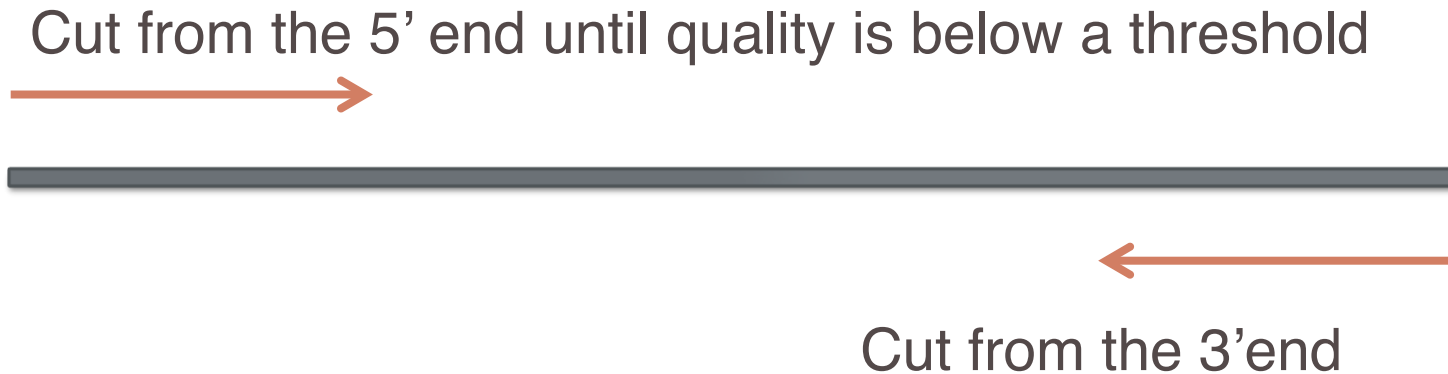
All these steps are performed using **trimming** algorithms

TRIMMING




TRIMMING

Dynamic Trimming



This way you will get some shortened reads but that have higher average quality. The reads in your dataset will not have the same length anymore (this is not a problem)

TRIMMING



[Overview](#) [Group](#) [Publications](#) [Supporting Info](#) [Teaching](#) [Software](#) [Internal](#)

☐ **Search**

Search:

☐ **News**

Page 1 of 2 [>](#) [>>](#)

Apr 4, 2013
[Sequencer](#)

Apr 4, 2013
[Moved](#)

Feb 26, 2012
[Fascination of Plant Days May 18](#)

Jan 7, 2012
[Pennellli N50 >500k](#)

Jan 5, 2012
[Seqanswers Wiki is wiki of the month](#)

Dec 24, 2011
[Forschungszentrum Jülich](#)

Nov 15, 2011
[Seqanswers Wiki published](#)

Sep 2, 2011
[New Location](#)

☐ **Contact Info**

You are here: [Supporting Info](#) » Trimmomatic

Trimmomatic: A flexible read trimming tool for Illumina NGS data

Citations

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

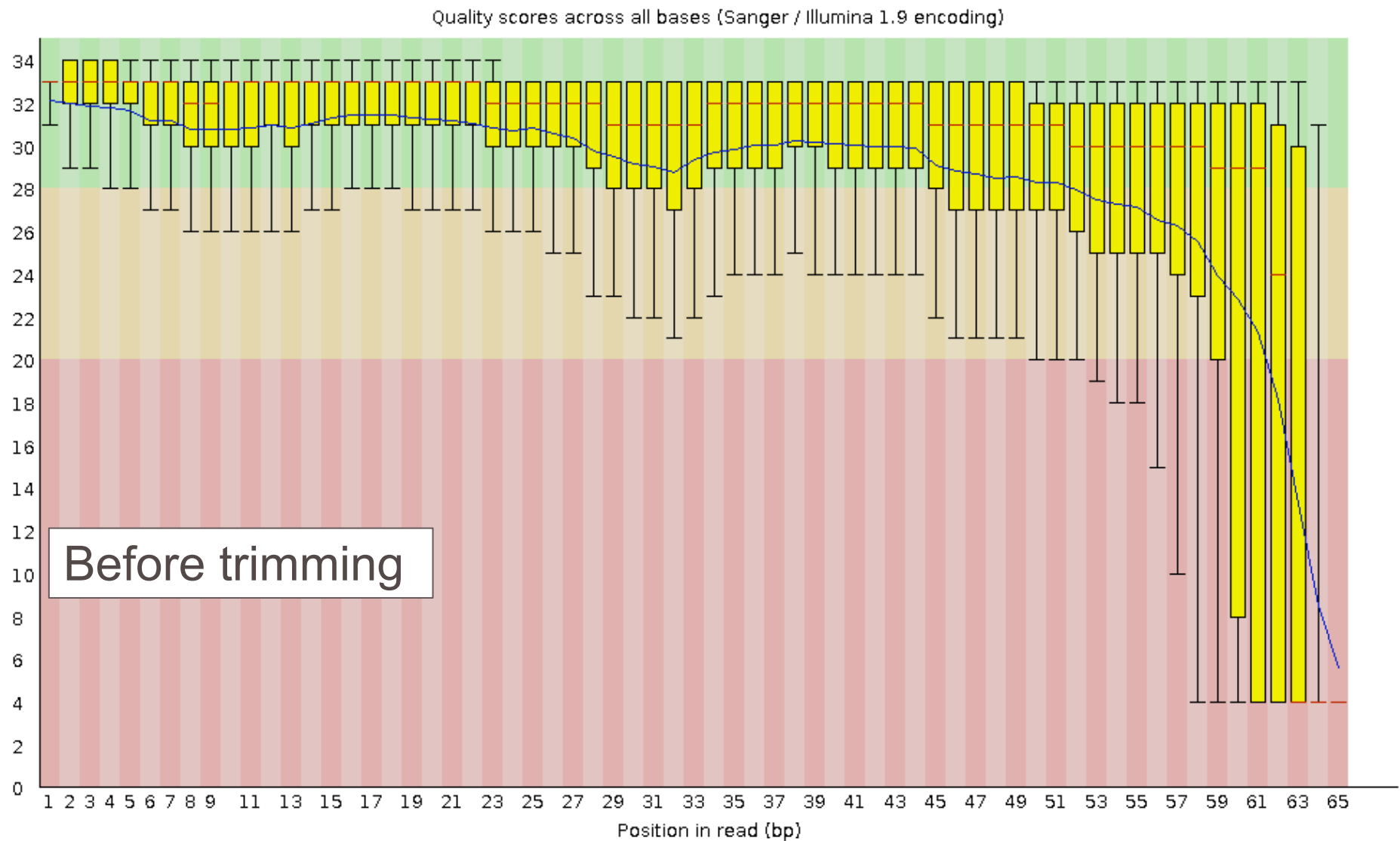
Description

Trimmomatic performs a variety of useful trimming tasks for illumina paired-end and single ended data. The selection of trimming steps and their associated parameters are supplied on the command line.

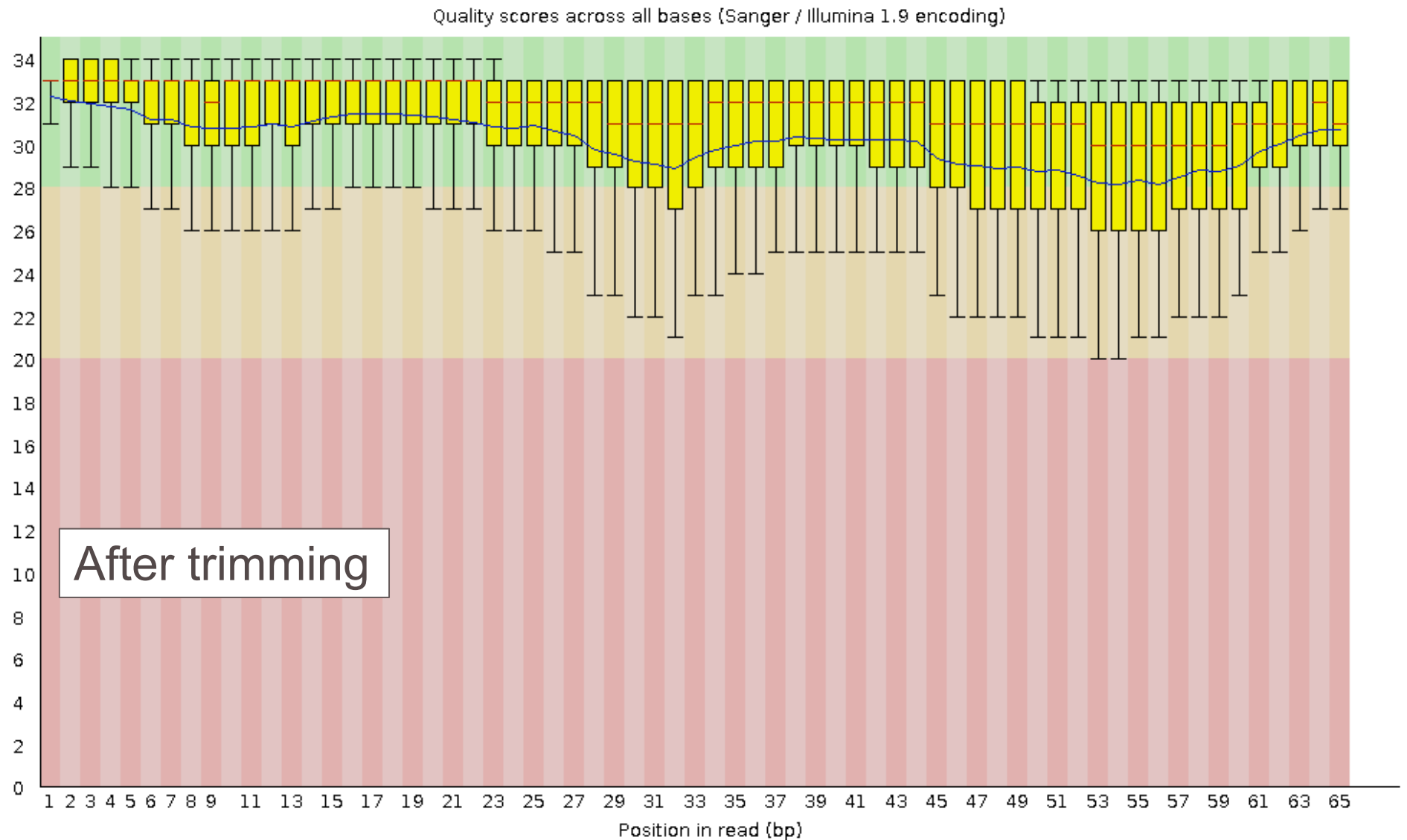
The current trimming steps are:

- ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.
- SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
- LEADING: Cut bases off the start of a read, if below a threshold quality
- TRAILING: Cut bases off the end of a read, if below a threshold quality
- CROP: Cut the read to a specified length
- HEADCROP: Cut the specified number of bases from the start of the read
- MINLEN: Drop the read if it is below a specified length
- TOPHRED33: Convert quality scores to Phred-33
- TOPHRED64: Convert quality scores to Phred-64

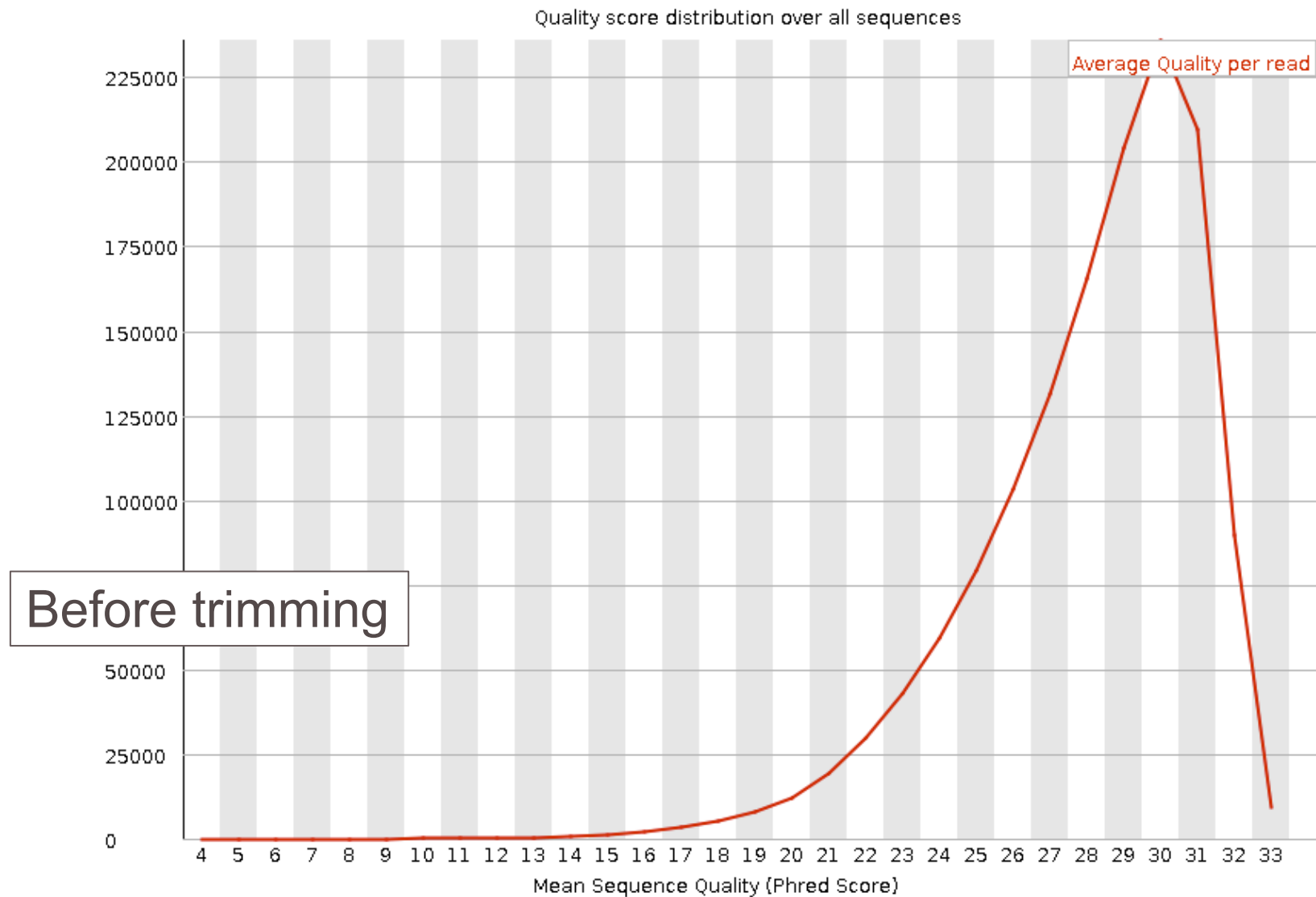
TRIMMING



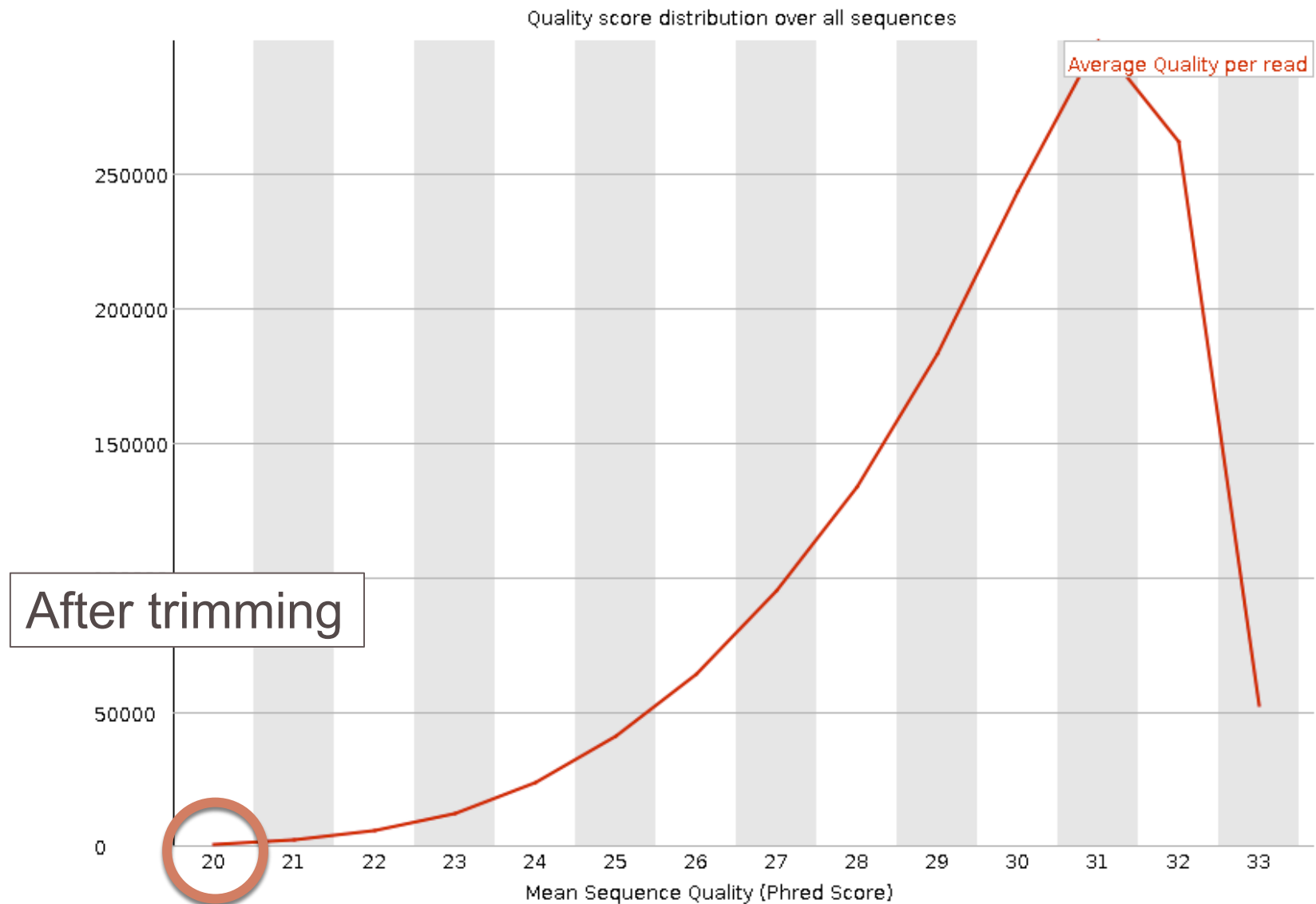
TRIMMING



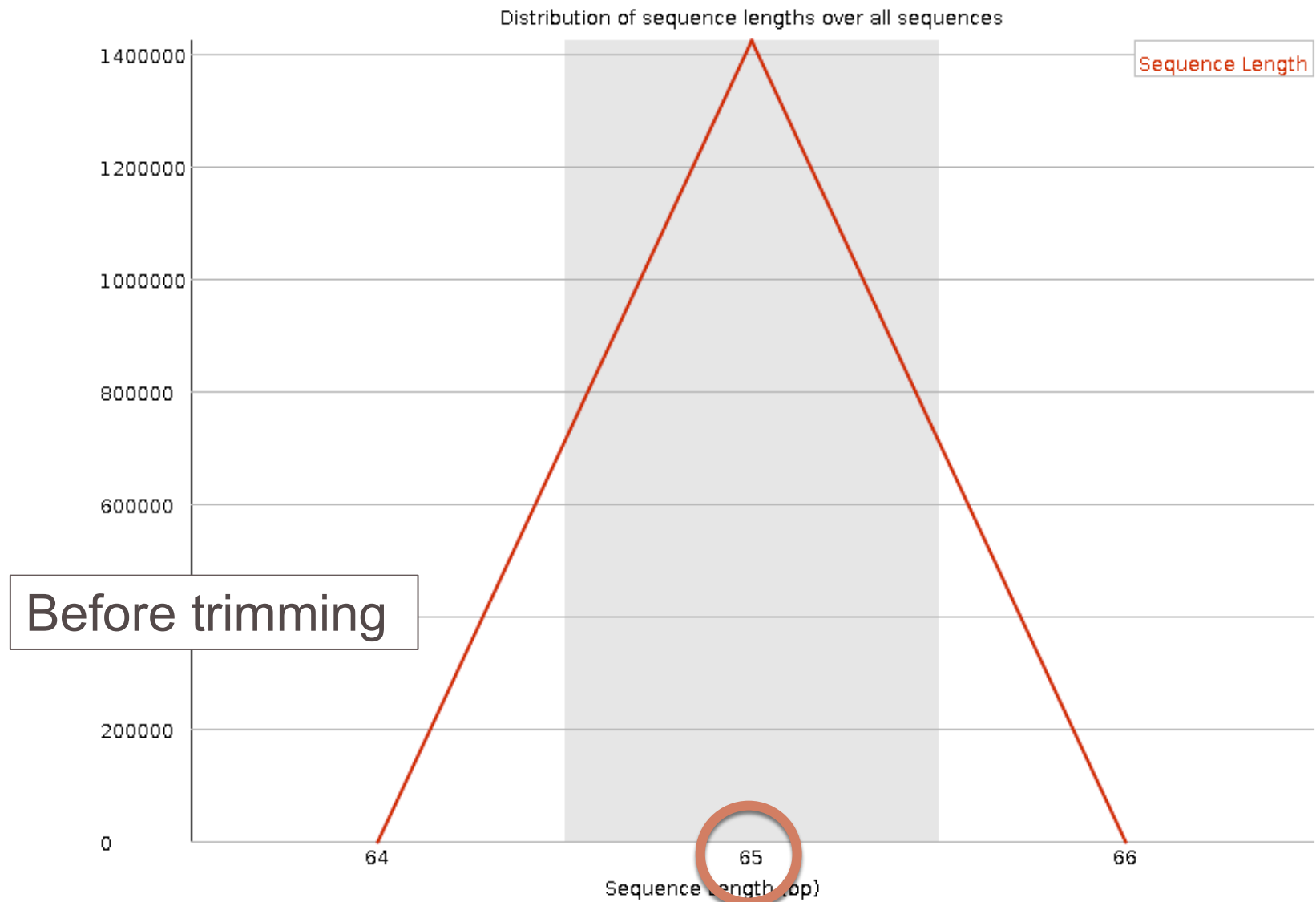
TRIMMING



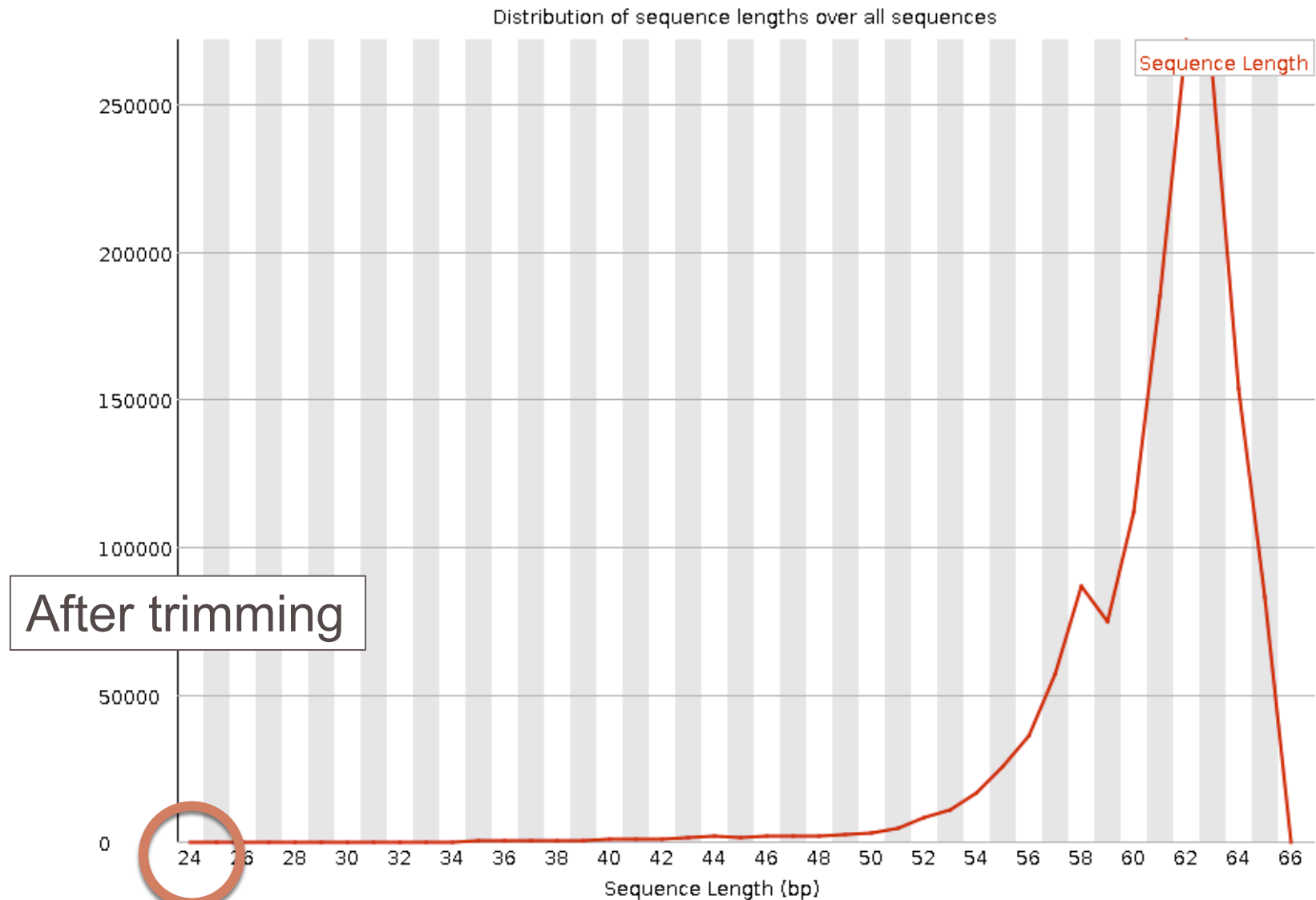
TRIMMING



TRIMMING



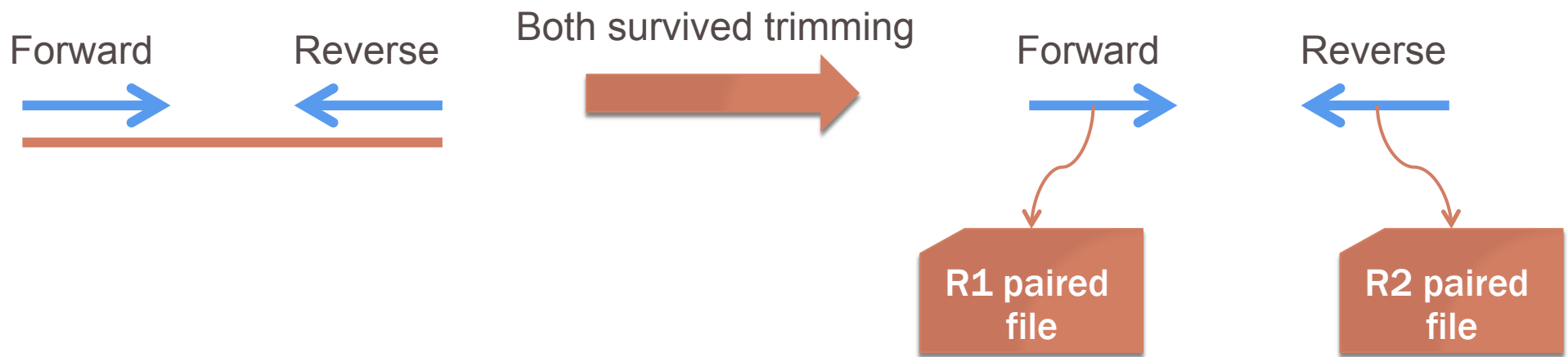
TRIMMING



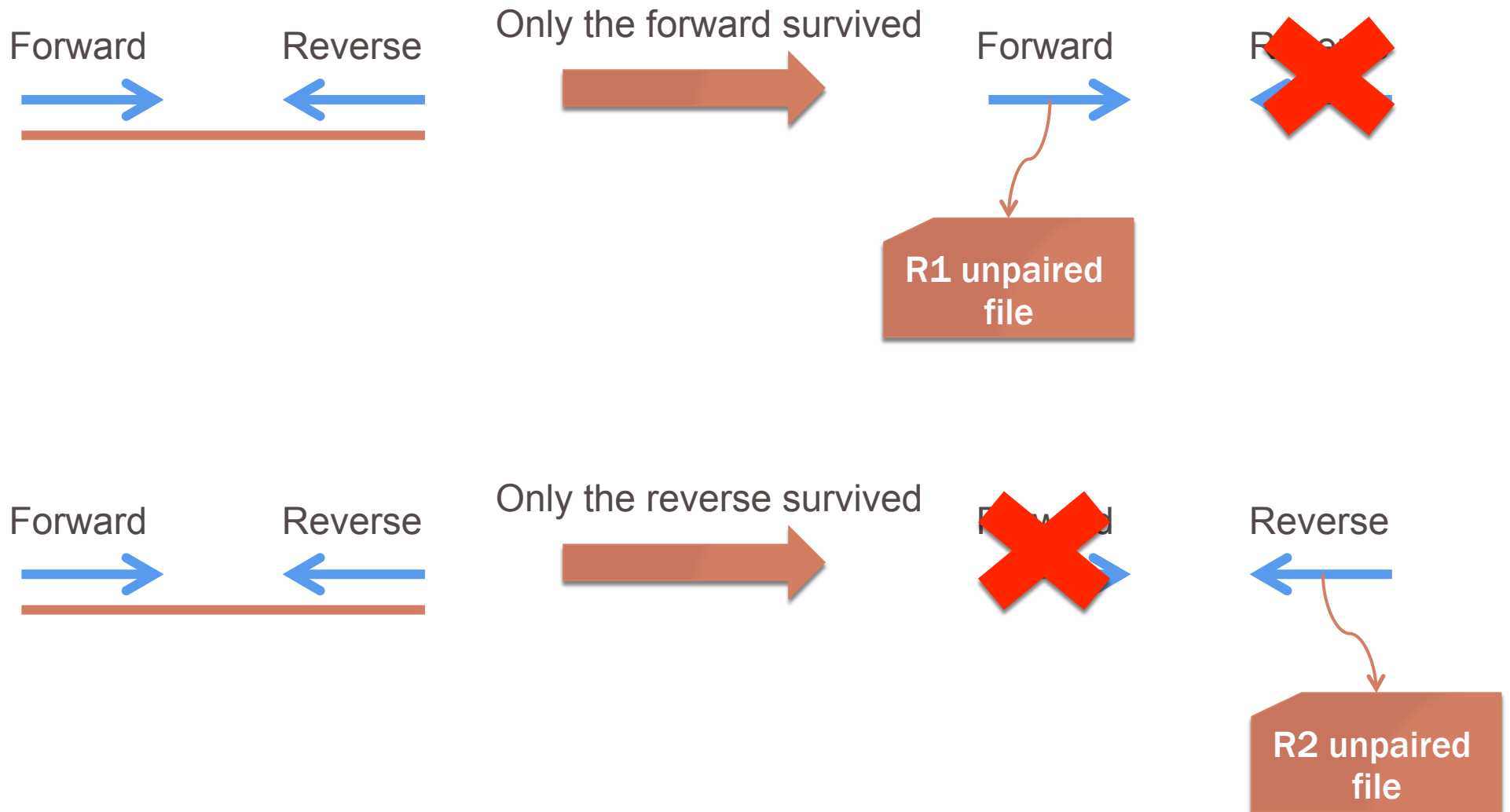
TRIMMING

- Adapter removal can be made only if a list of adapter sequences is provided to Trimmomatic
- This means that you need to know which adapter were used for sample preparation
- Illumina used different sets of adapter throughout the years; the most commonly used are the TruSeq and TruSeq2 series
- If you don't know which adapter were used, you might try to guess them by looking at over-represented sequences (for example in the FastQC output)

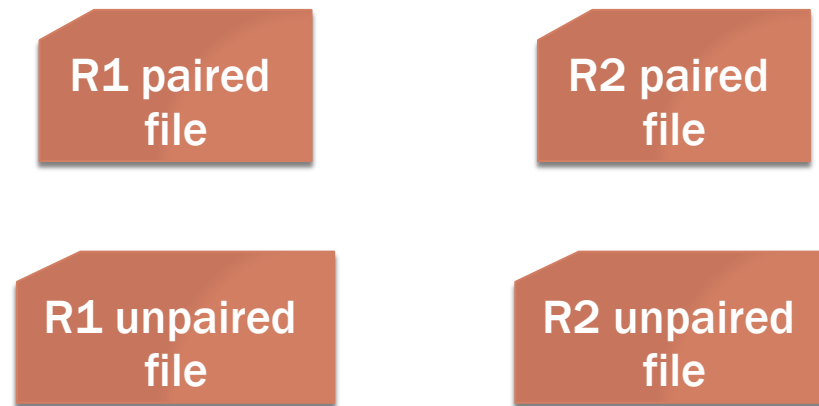
TRIMMING



TRIMMING



TRIMMING



The Trimmomatic procedure creates four output fastq files:

- R1 paired: forward reads for which their paired read survived the trimming and is found in the R2 paired file
- R2 paired: reverse reads for which their paired read survived the trimming and is found in the R1 paired file
- R1 unpaired: forward reads that lost their paired read
- R2 unpaired: reverse reads that lost their paired read

TRIMMING

Williams et al. *BMC Bioinformatics* (2016) 17:103
DOI 10.1186/s12859-016-0956-2

BMC Bioinformatics

RESEARCH ARTICLE

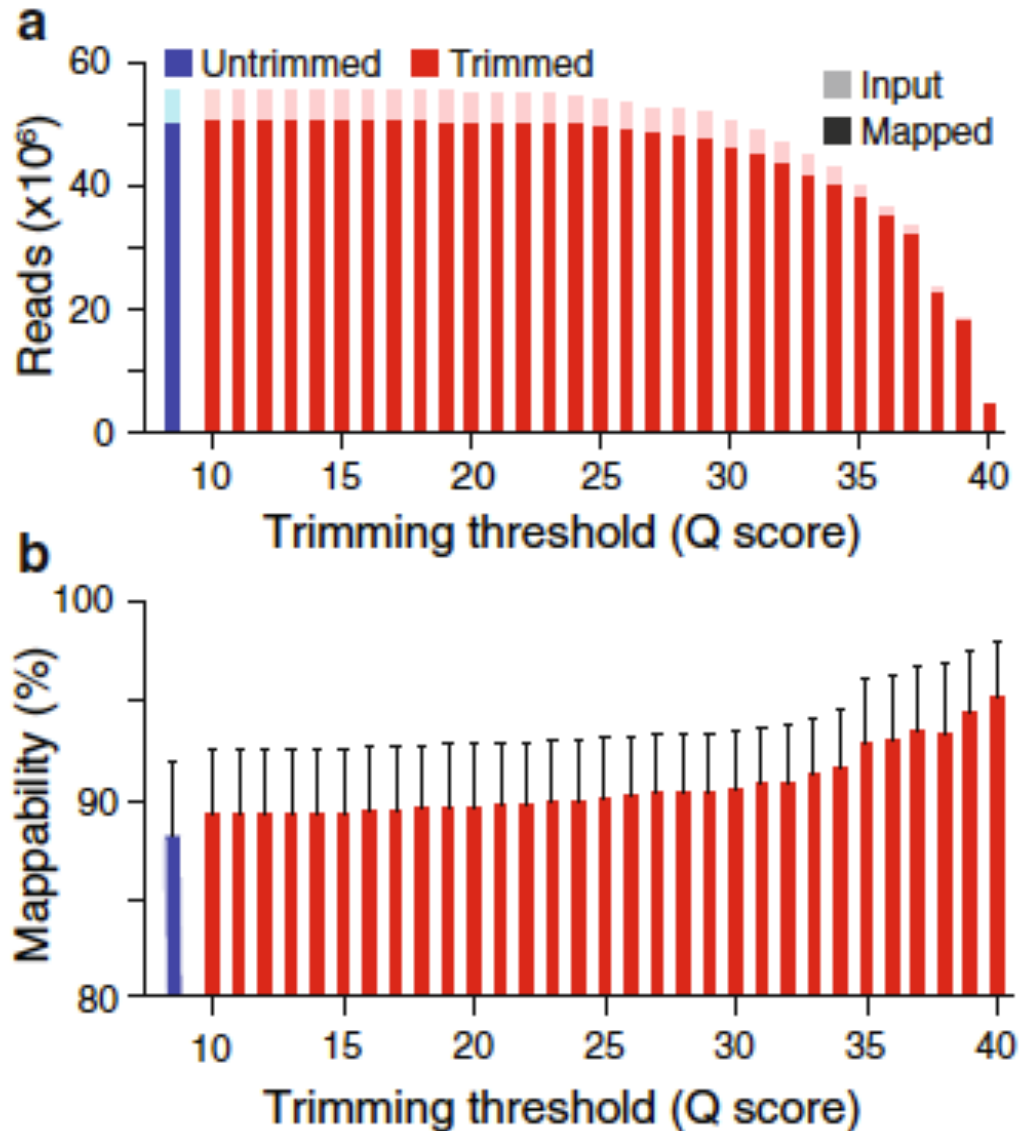
Open Access

Trimming of sequence reads alters RNA-Seq gene expression estimates



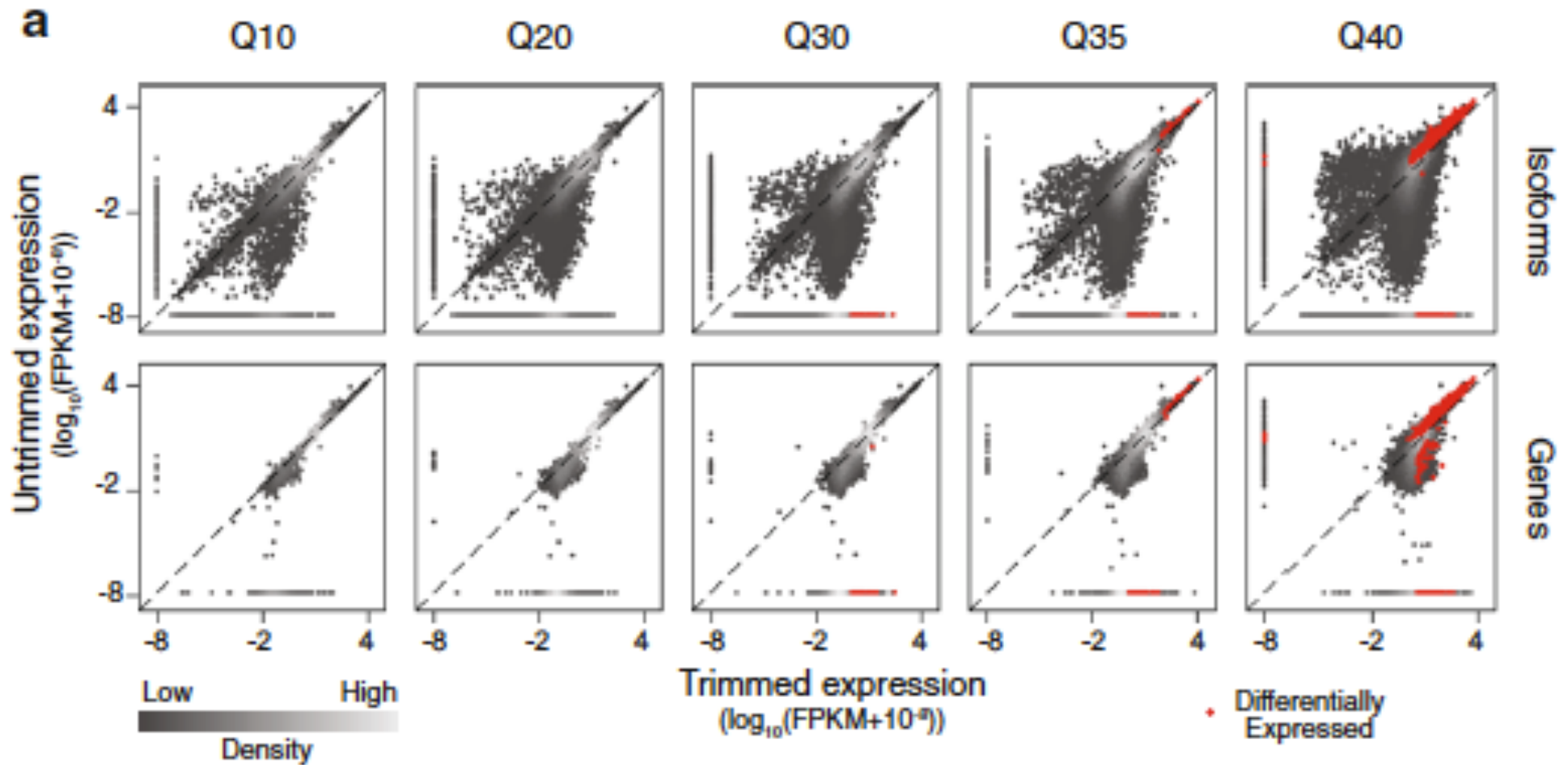
Claire R. Williams¹, Alyssa Baccarella², Jay Z. Parrish^{1*} and Charles C. Kim^{2,3*}

TRIMMING

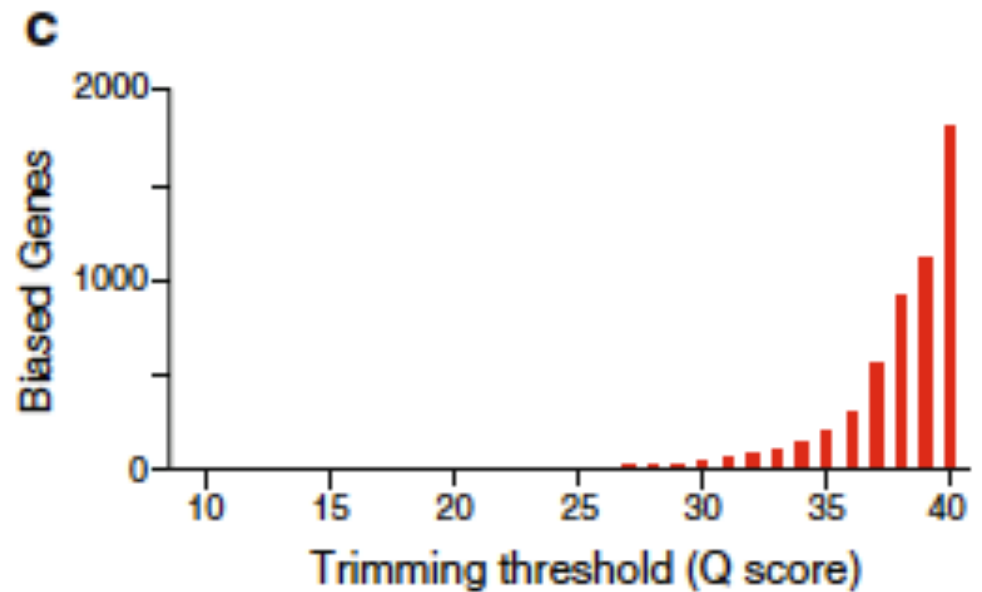
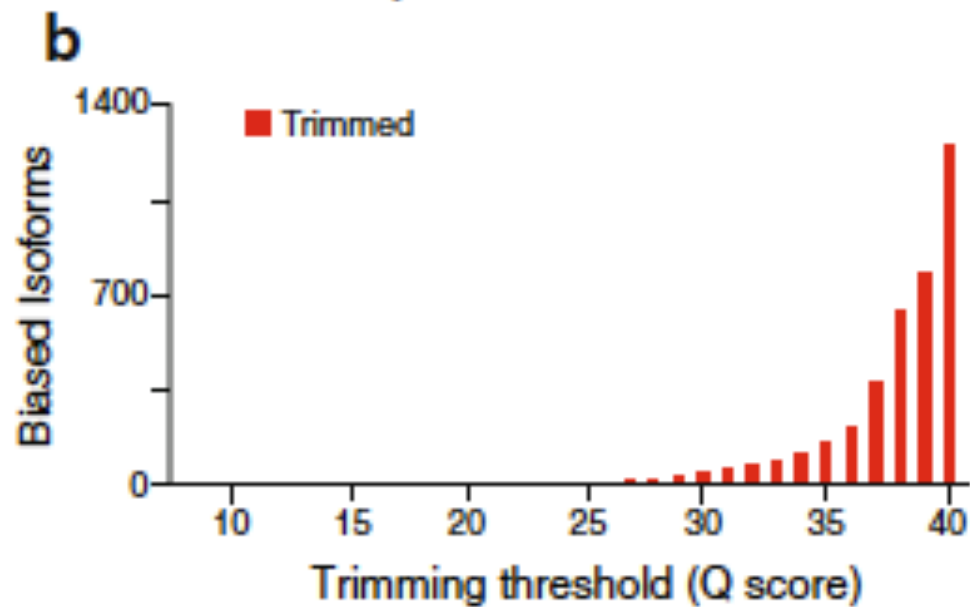


Shortening or discarding reads by quality score thresholds reduce the total number of reads and increase their mappability, i.e. the possibility of finding a good alignment with the reference genome

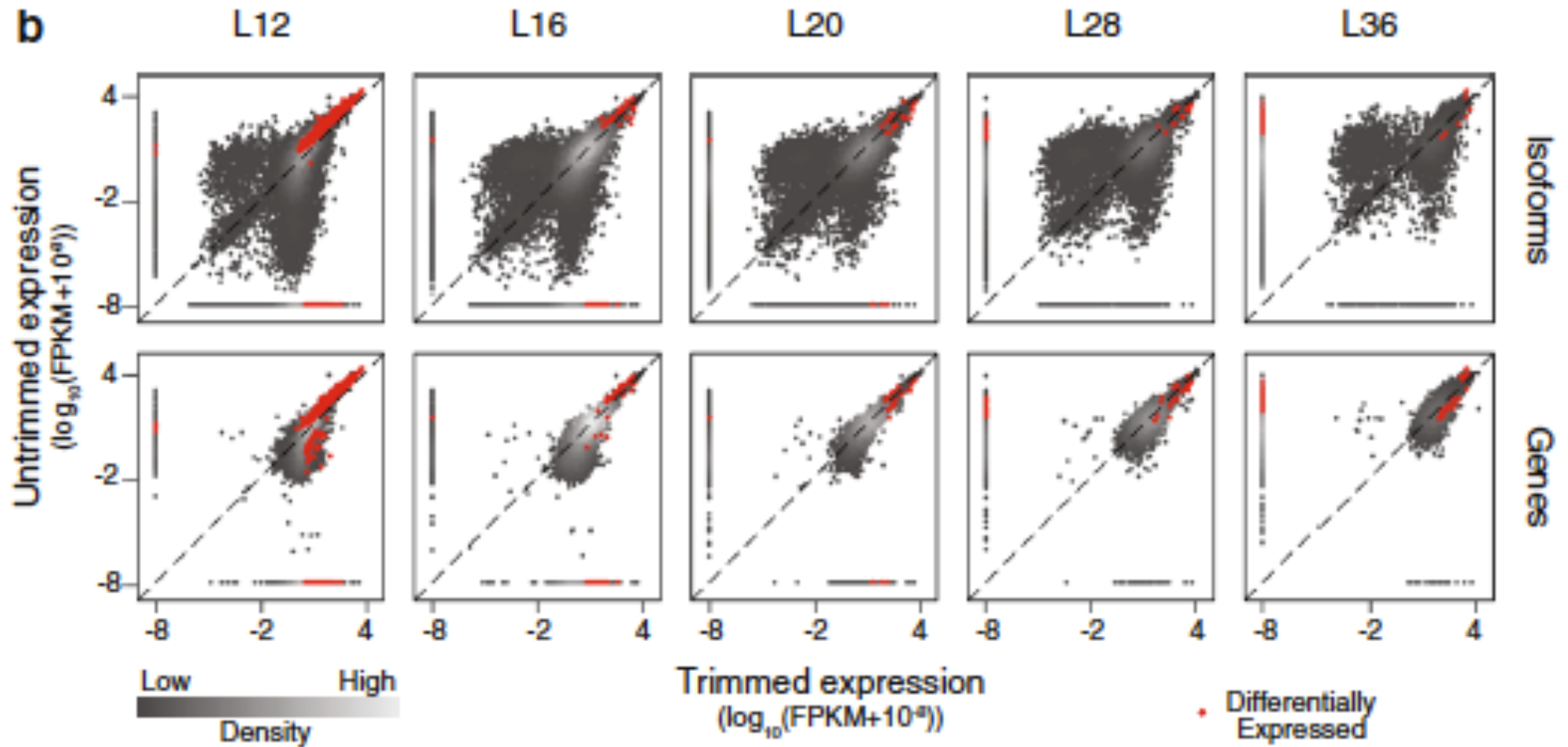
TRIMMING



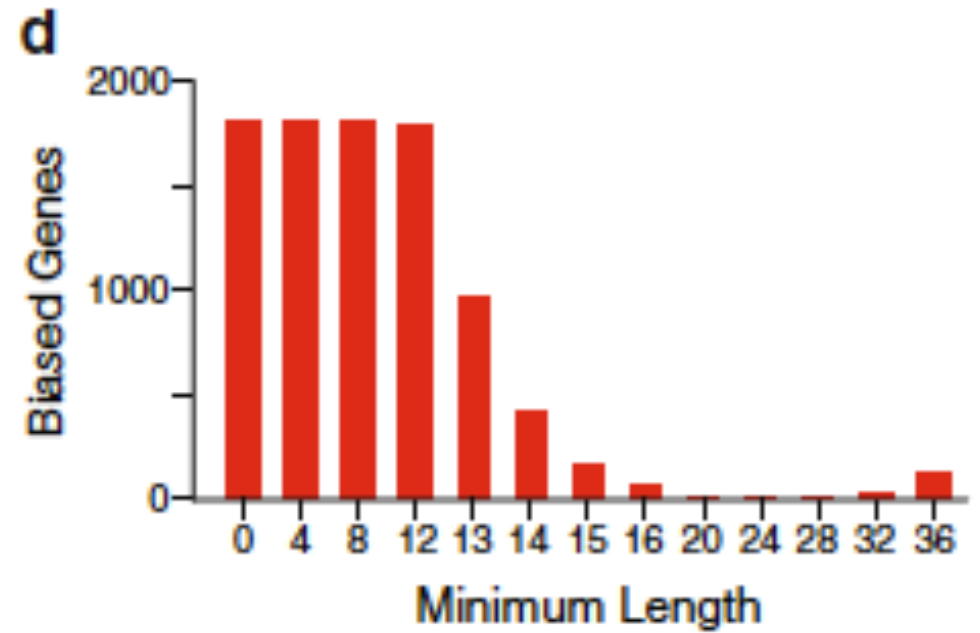
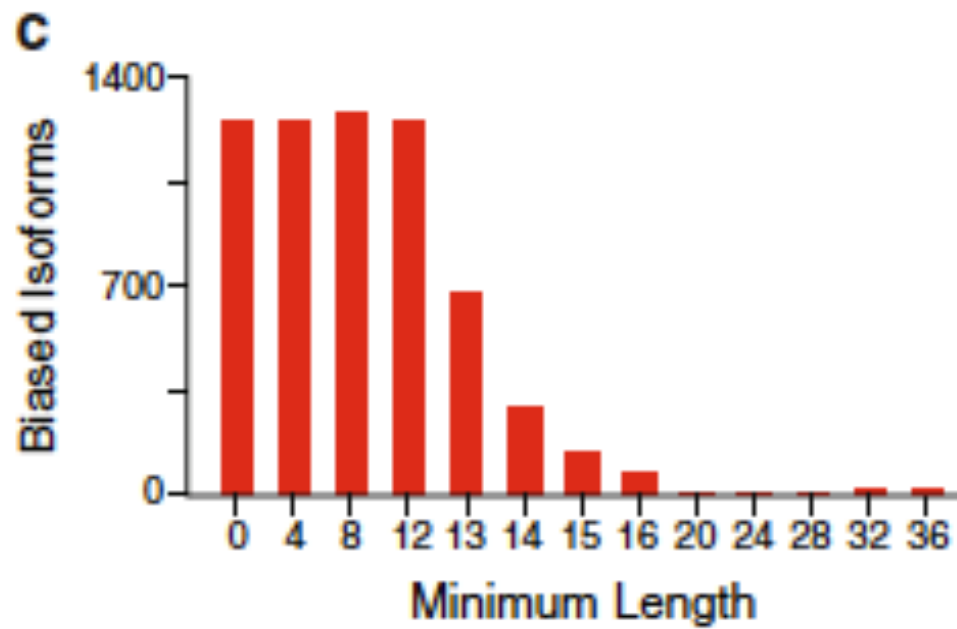
TRIMMING



TRIMMING



TRIMMING



TRIMMING

- Aggressive trimming by quality scores only can create biases affecting gene expression estimates and possibly leading to artificially inflated numbers of differentially expressed genes
- The reason is that by removing low quality regions, reads can become too short to have a reliable mapping to the genome
- Hence, it is advisable to keep quality score trimming at low stringency, and to filter all reads that after quality trimming become shorter than a threshold