BIO214 Lecture 3

Bioinformatics-II

Read Genome Mapping

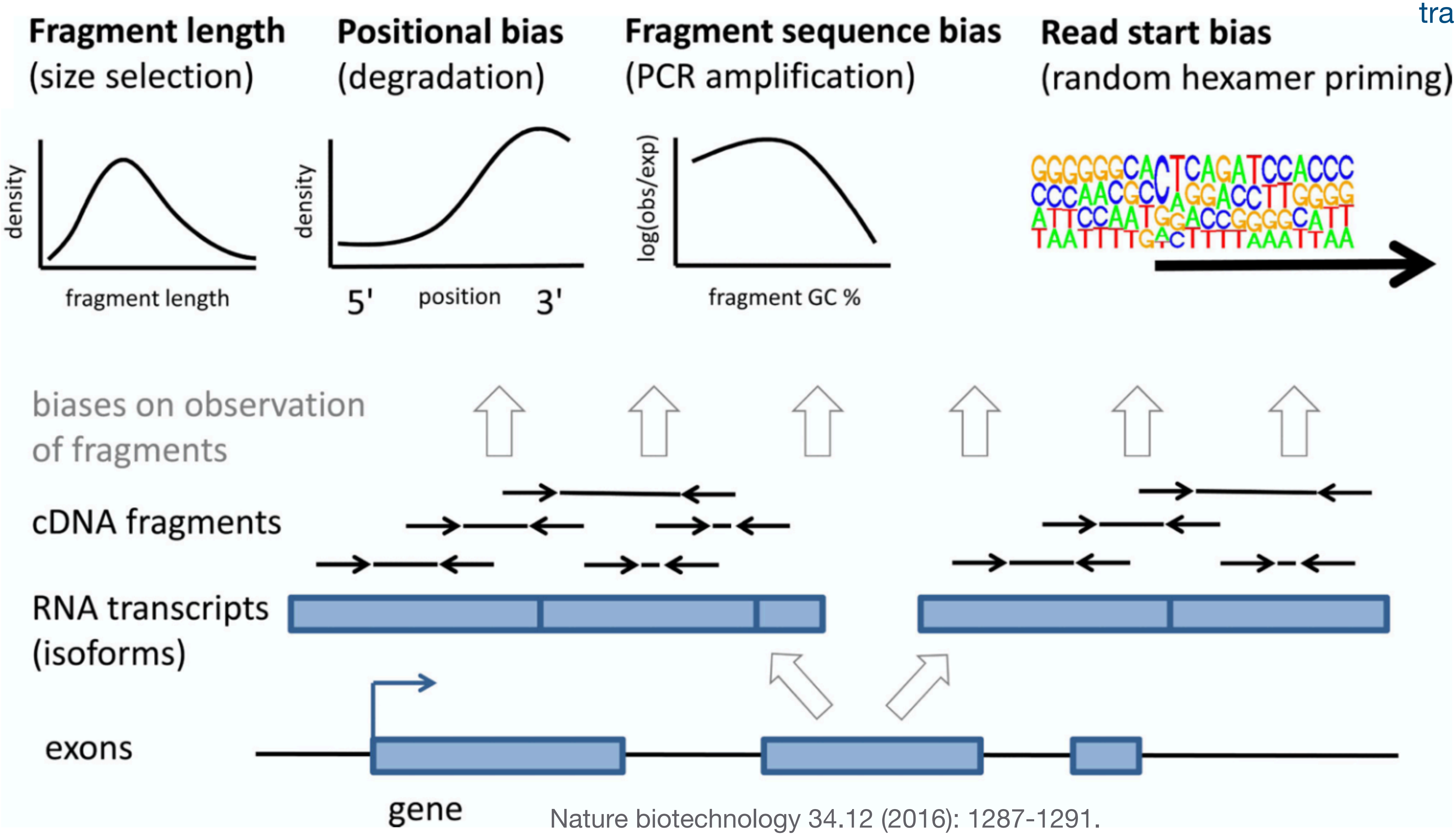
Zhen Wei; 2023-Feb-14

Outline

- Pre-mapping quality control
- Genome aligner
- Splice-awared genome aligner
- Alignment-free method
- Performances of different tools

Pre-mapping quality control

Reads quality control: what could go wrong?



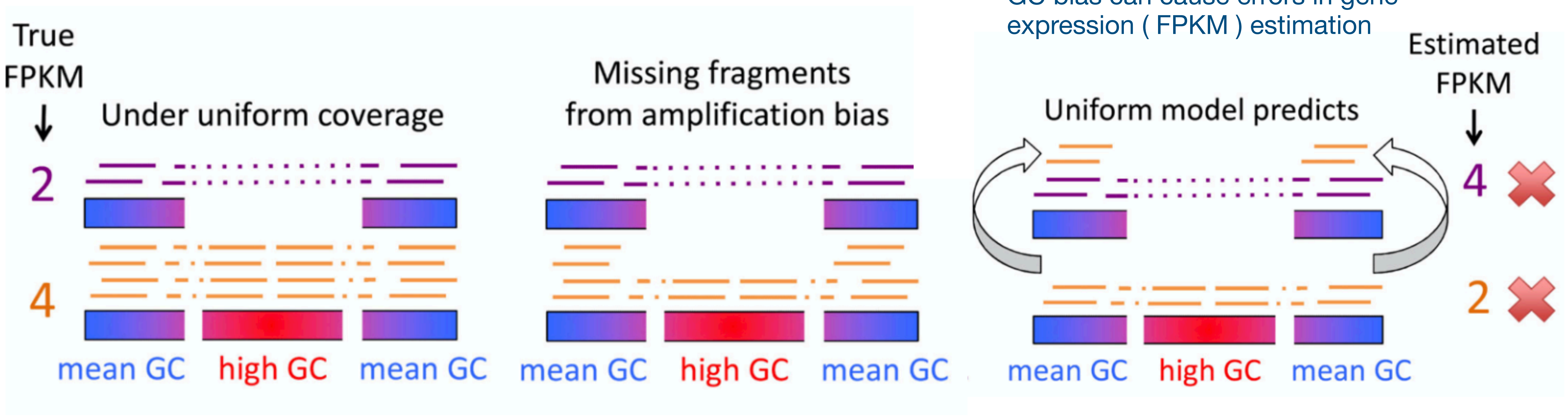
Nature biotechnology 34.12 (2016): 1287-1291.

Introduced in reverse transcription

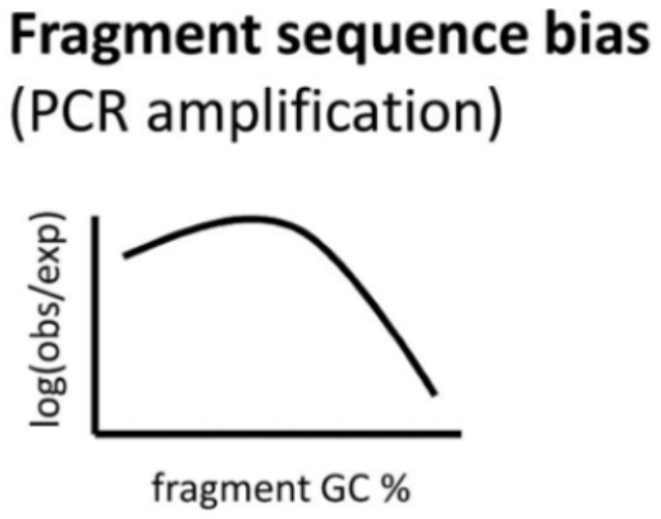
- In addition to base calling errors, NGS library preparation can introduce technical biases from multiple sources.
- These biases can lead to systematic error and batch effect in NGS data.

Fragment GC content bias

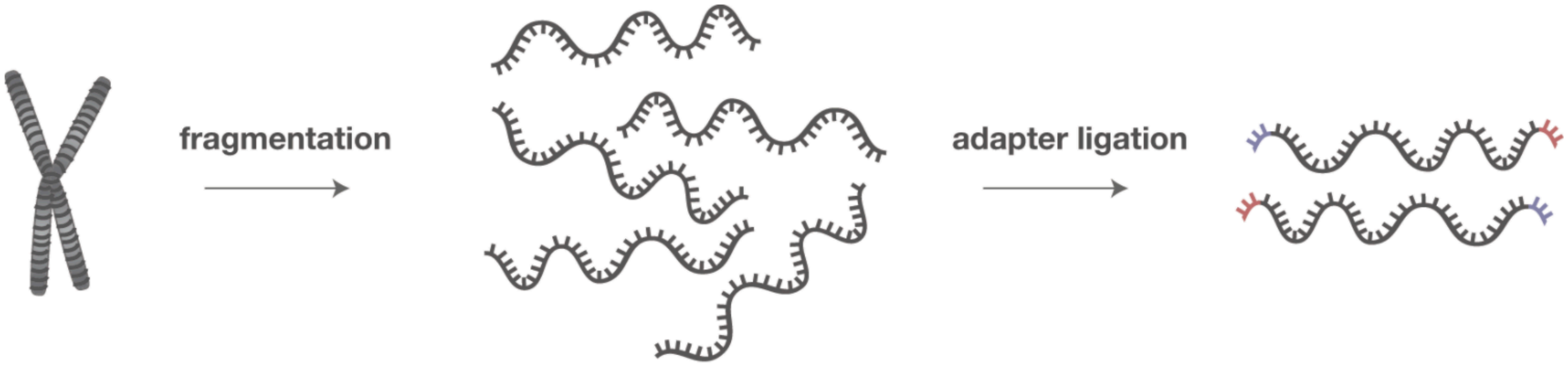
GC bias can cause errors in gene expression (FPKM) estimation

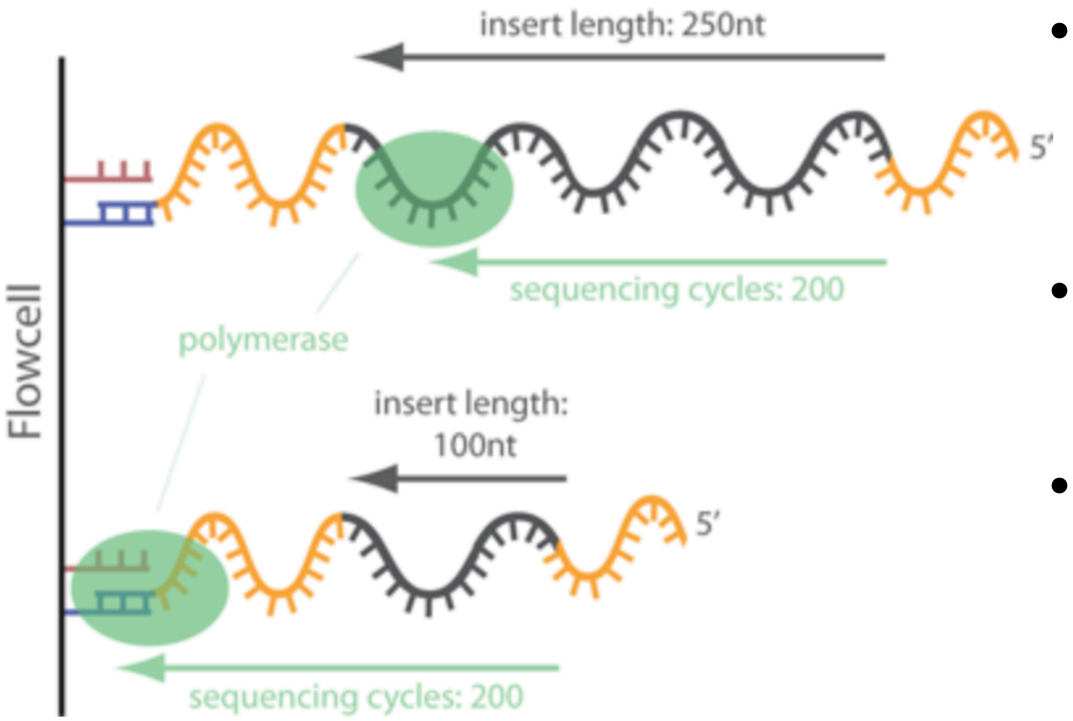


- PCR amplification of DNA/cDNA fragments introduces bias in 2nd generation sequencing-based techniques (e.g. DNA-Seq, RNA-Seq, Chip-Seq).
- This is typically the most severe type of technical bias for illumina sequencing.



Adaptor contamination



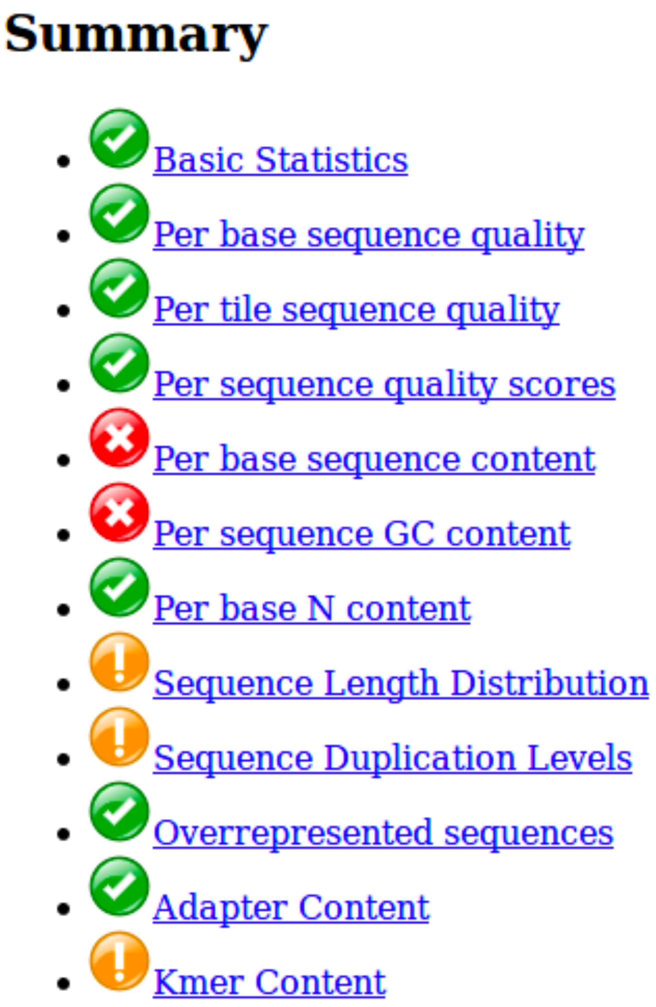


- Illumina sequencing uses adaptors, which are repeated sequences attached to both ends of DNA/cDNA fragments.
- Adaptors facilitate hybridization with probes (on the flow cell) and primers (in bridge PCR).
- Short fragments can lead to adaptor contamination at the 3' end of reads, especially when the read length exceeds the insert length.

How to detect read quality issues?

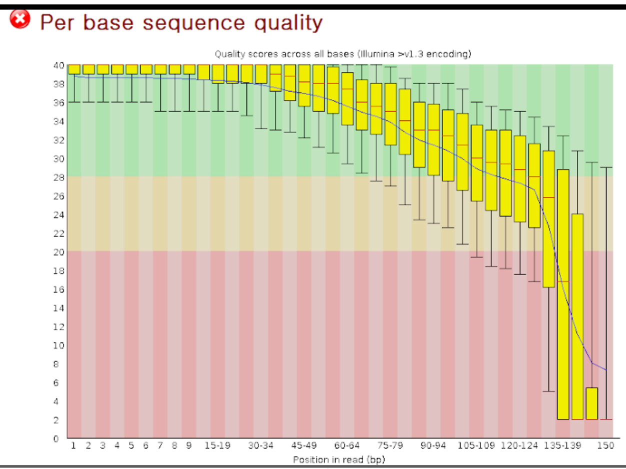
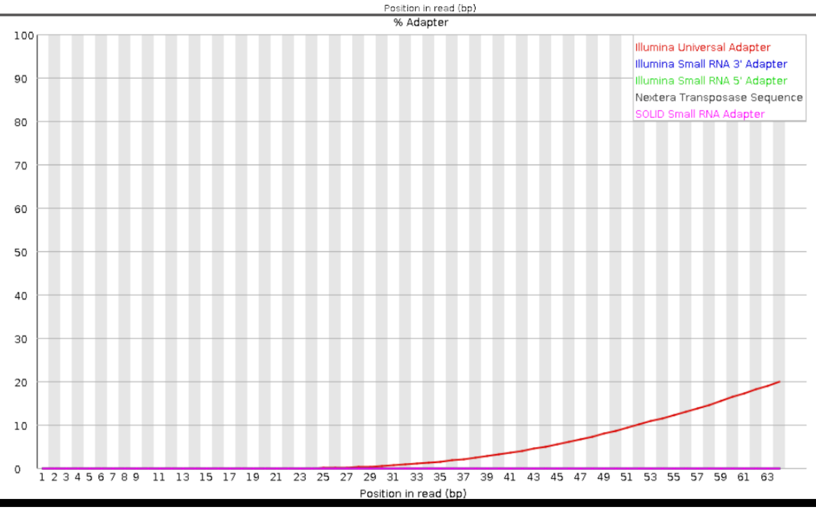
Read QC software

- ***fastqc*** is a command line tool on Linux/Unix system to generate quality report on fastq files.
- The output of *fastqc* includes an html report, which contains multiple QC statistics.
- It can be used on linux bash with a single line command.

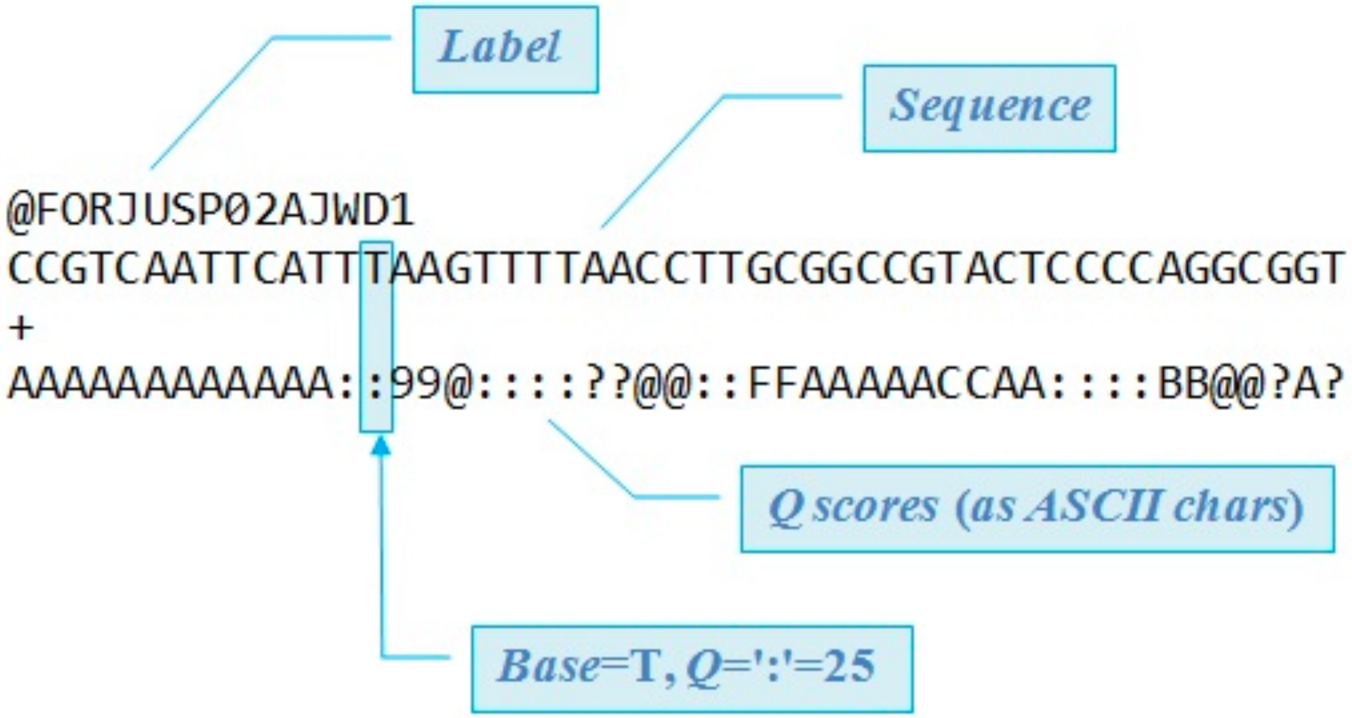


Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ! Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ! Kmer Content

example QC metrics	Diagram	Interpretation
Per base sequence quality		• A box plot of Phred scores for every positions of read. • If the IQR drop below the red line (< 20) near the 3'end, then quality trimming is needed.
Adapter Content		• Problematic if read 3' end contain adaptor contents. • Adaptor trimming can be used to remove adopters.

Fastq format



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

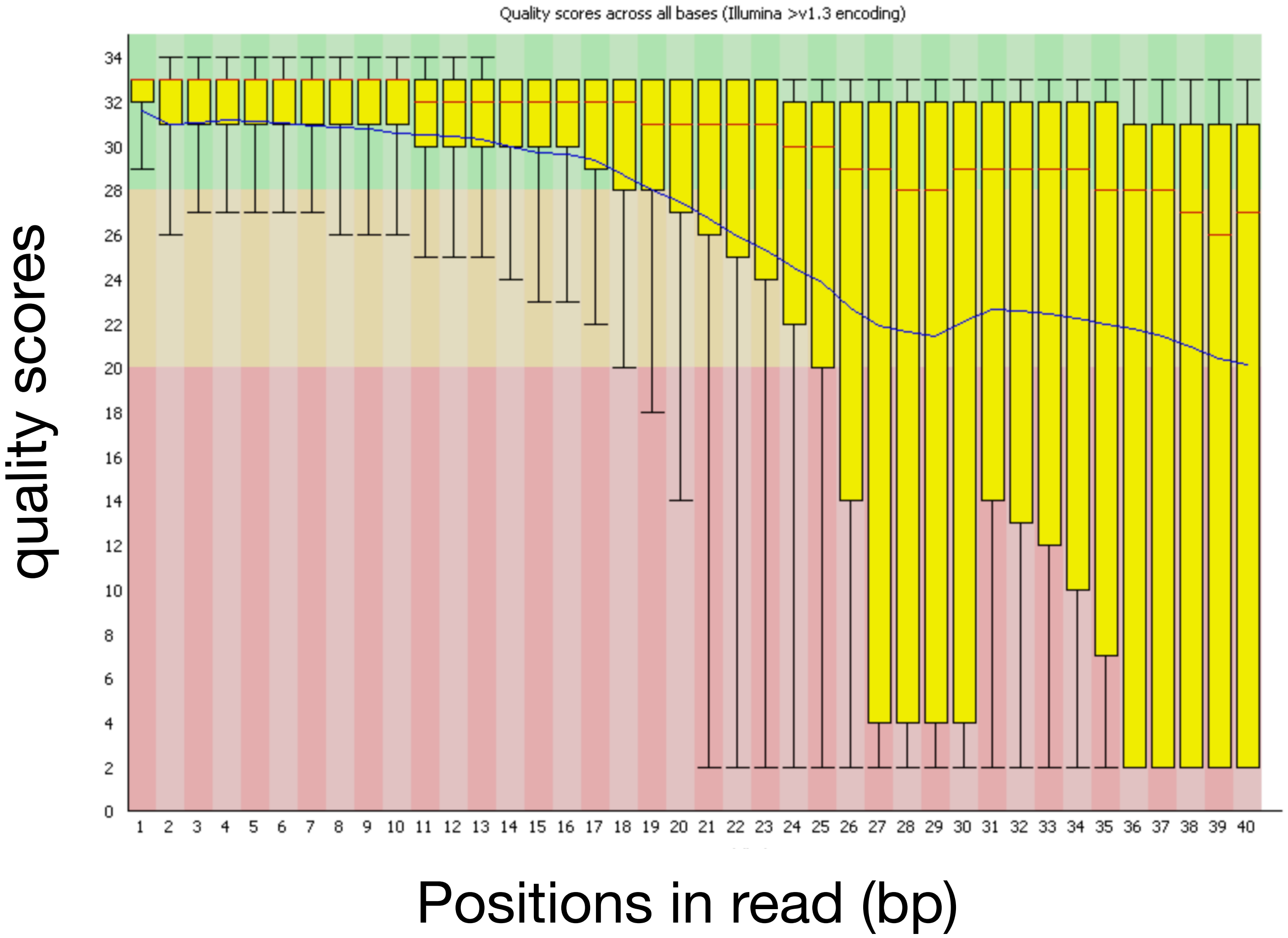
Fastq is a text-based format. It represents each raw read with 4 lines:

1. A sequence identifier with information about the sequencing run and the cluster.
2. The sequence or base calls in the order of 5’-3’; can be A, C, T, G and N.
3. A separator of a plus (+) sign.
4. Characters encoded base call quality scores (**Phred scores**). The Phred scores or Q scores have the following definition:

$$Q = -10 \times log_{10}(e)$$

where e is the estimated probability of the base call being wrong.

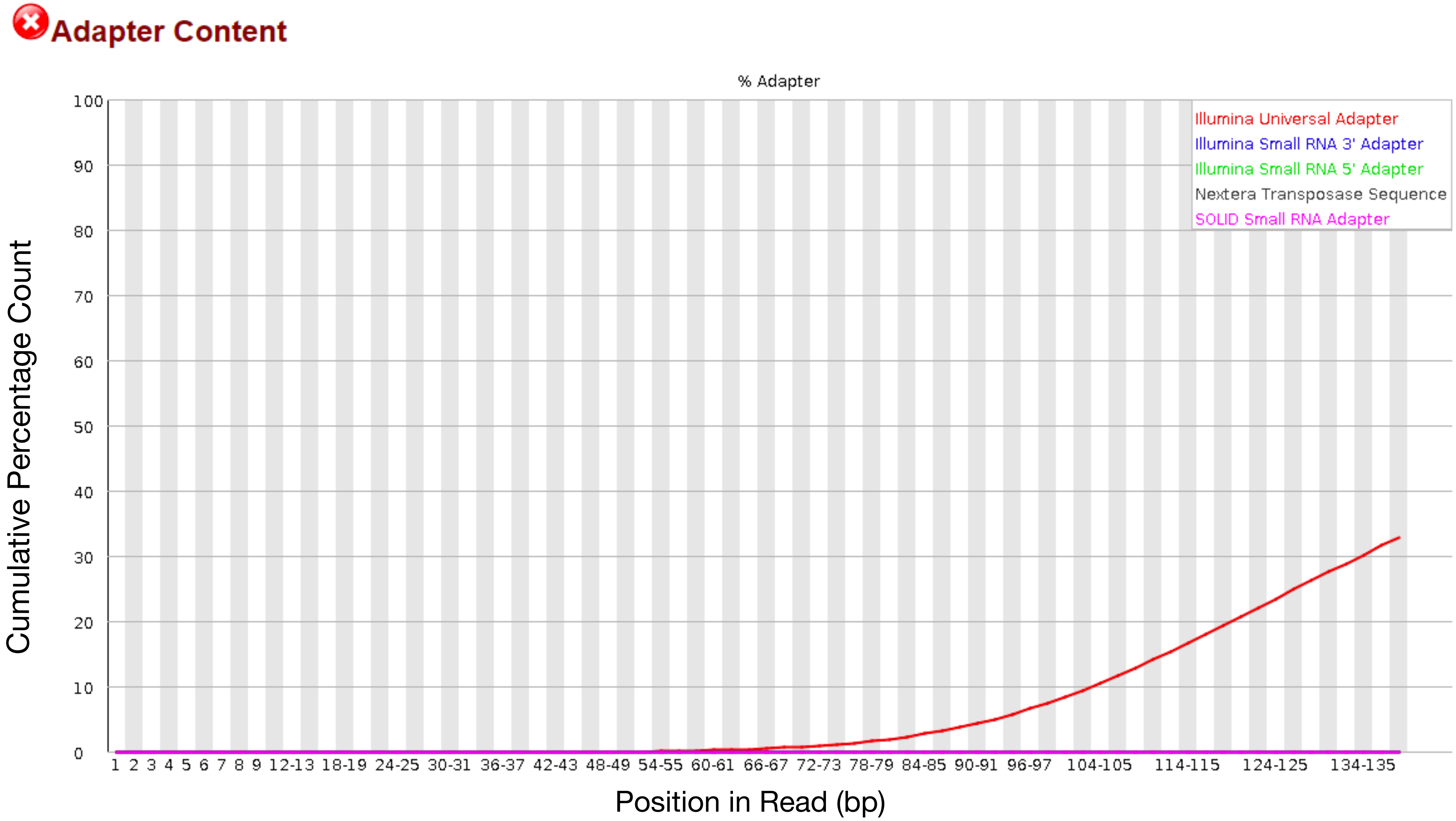
Per base sequence quality



Box plot of Phred scores

- The y-axis on the graph shows the Phred scores.
- The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red).
- Warning will be issued if the lower quartile for any bases fall below the red region.

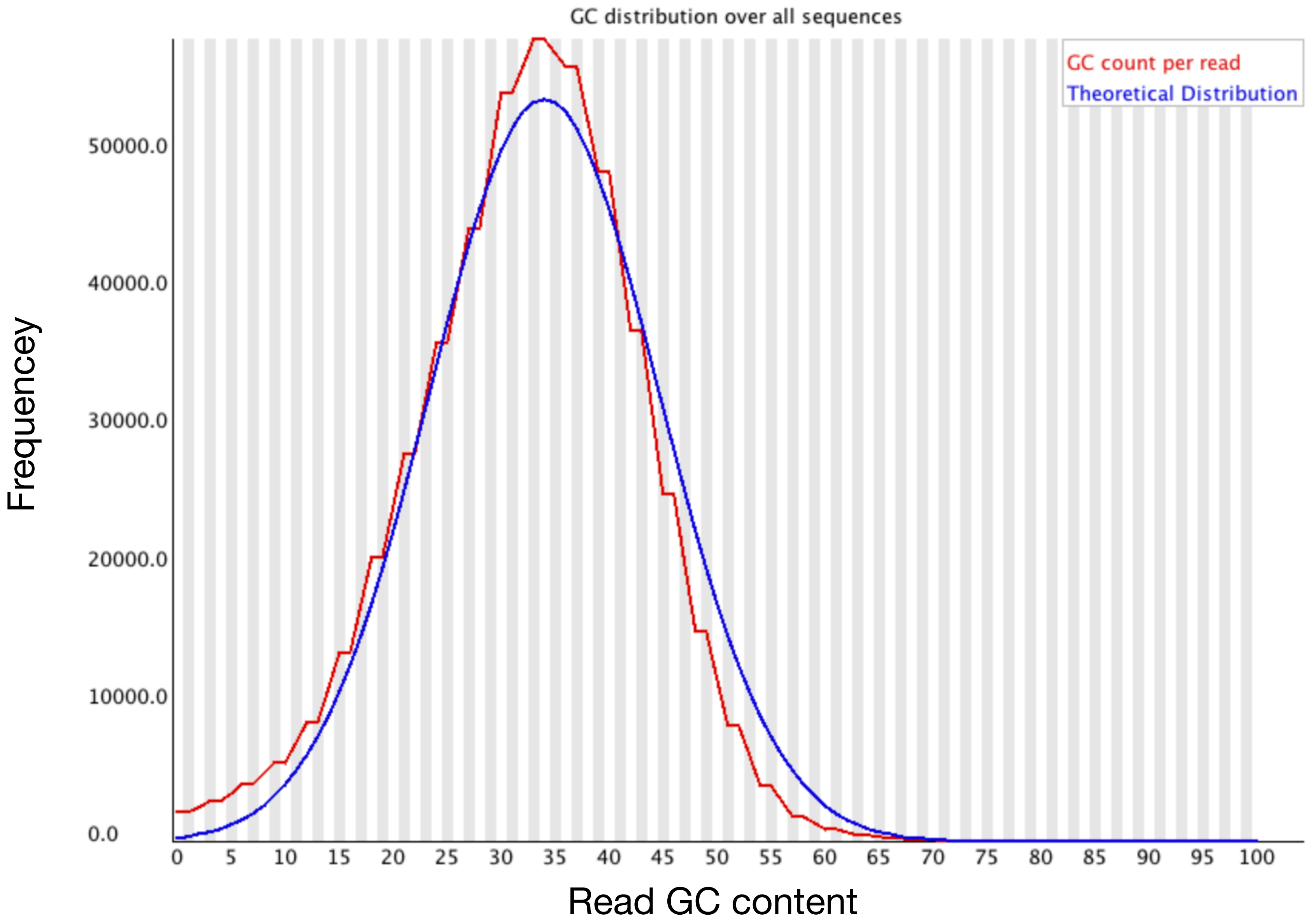
Adaptor content



- The plot shows a cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position.

- This module will issue a warning if any sequence is presented in more than 5% of all reads.

GC content distribution



- The graph displayed a histogram of GC content over all reads.

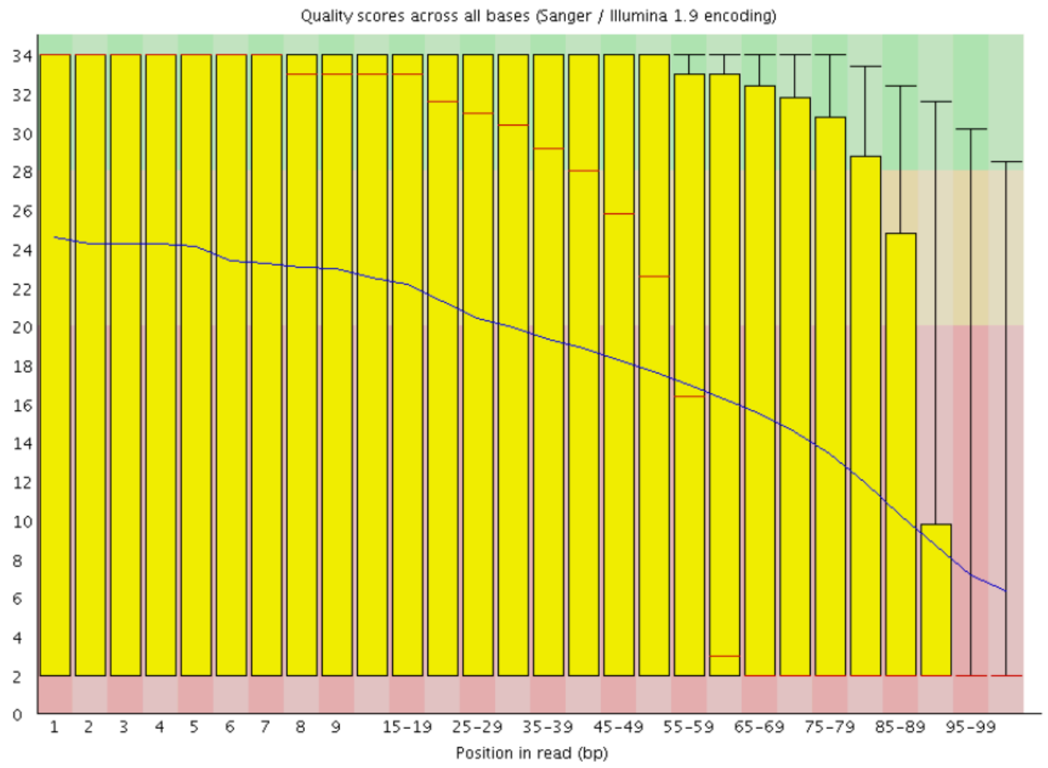
- Warning is issued when observed read GC content distribution (red) is significantly deviant from the expected normal distribution (blue).

How to fix the diagnosed issues?

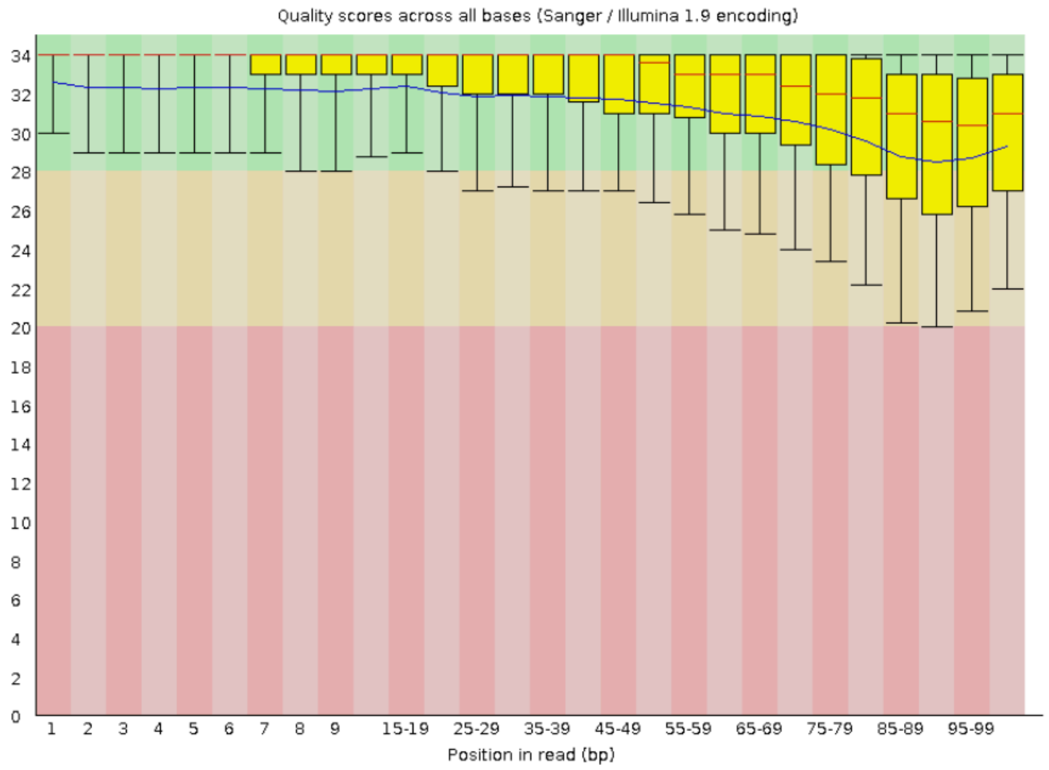
Trimming software

- The adaptor sequences and low quality ends can be removed via trimming.
- ***Trim Galore*** (a popular trimming software) can automatically scan & remove adaptors and low quality base calls from the read 3'end.
- Normalization methods are required to address other types of technical biases, such as GC content biases, in downstream analysis.

Before quality trimming



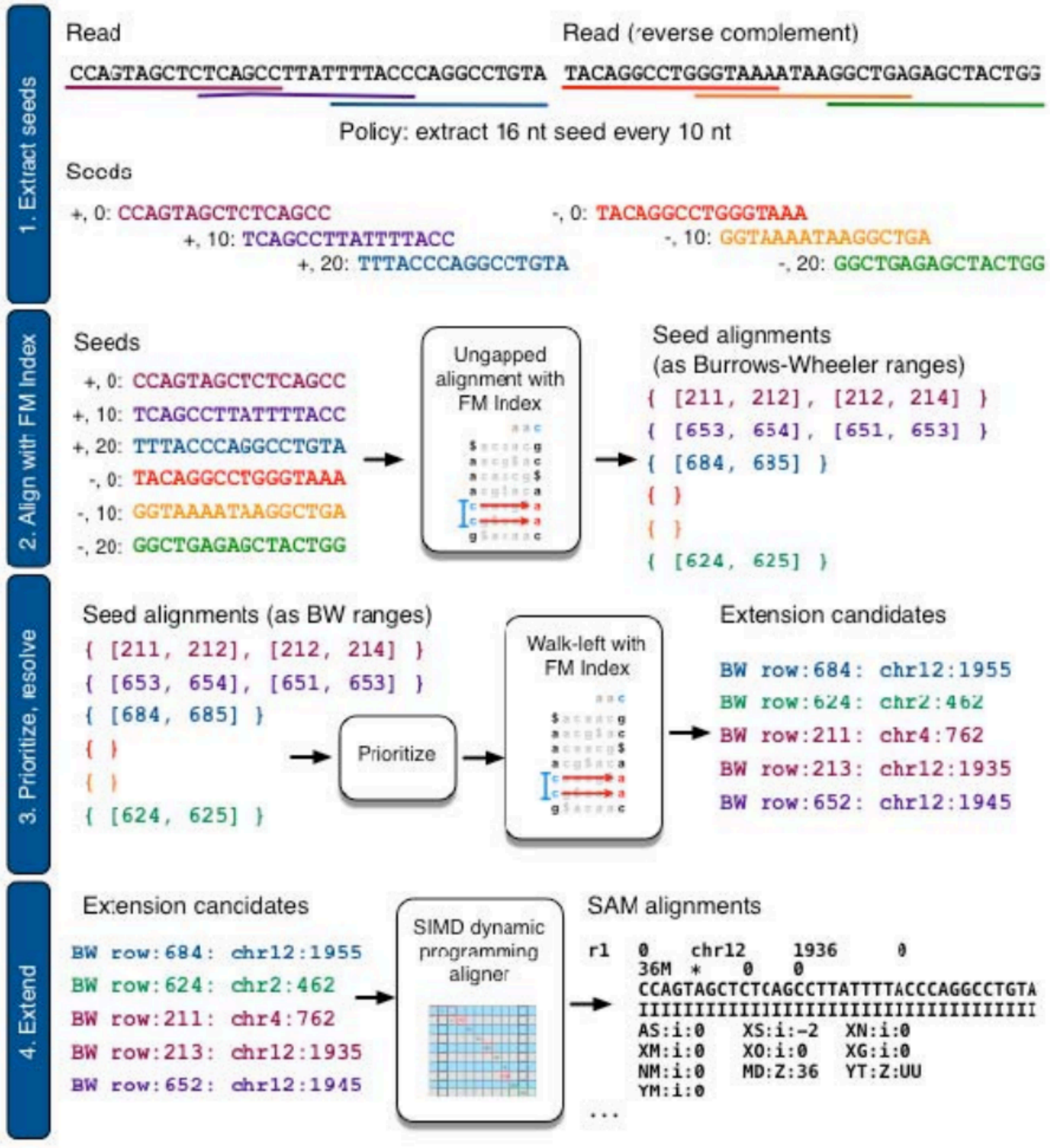
After quality trimming



Genome aligners

How to align short reads to genome efficiently?

Bowtie2



1. Extract seeds

Read

CCAGTAGCTCTCAGCCTTATTTTACCCAGGCCTGTA

Read (reverse complement)

TACAGGCCTGGGTAAAATAAGGCTGAGAGCTACTGG

Policy: extract 16 nt seed every 10 nt

Seeds

+, 0: CCAGTAGCTCTCAGCC
+, 10: TCAGCCTTATTTTACC
+, 20: TTTACCCAGGCCTGTA

-, 0: TACAGGCCTGGGTAAA
-, 10: GGTAAAATAAGGCTGA
-, 20: GGCTGAGAGCTACTGG

2. Align with FM Index

Seeds

+, 0: CCAGTAGCTCTCAGCC
+, 10: TCAGCCTTATTTTACC
+, 20: TTTACCCAGGCCTGTA
-, 0: TACAGGCCTGGGTAAA
-, 10: GGTAAAATAAGGCTGA
-, 20: GGCTGAGAGCTACTGG

Ungapped alignment with FM Index

Seed alignments (as Burrows-Wheeler ranges)

```
{ [211, 212], [212, 214] }
{ [653, 654], [651, 653] }
{ [684, 635] }
{ }
{ }
{ [624, 625] }
```

3. Prioritize, resolve

Seed alignments (as BW ranges)

```
{ [211, 212], [212, 214] }
{ [653, 654], [651, 653] }
{ [684, 685] }
{ }
{ }
{ [624, 625] }
```

Prioritize

Walk-left with FM Index

Extension candidates

```
BW row:684: chr12:1955
BW row:624: chr2:462
BW row:211: chr4:762
BW row:213: chr12:1935
BW row:652: chr12:1945
```

4. Extend

Extension candidates

```
BW row:684: chr12:1955
BW row:624: chr2:462
BW row:211: chr4:762
BW row:213: chr12:1935
BW row:652: chr12:1945
```

SIMD dynamic programming aligner

SAM alignments

```
r1  0    chr12    1936     0
    36M  *    0    0
    CCAGTAGCTCTCAGCCTTATTTTACCCAGGCCTGTA
    IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
    AS:i:0    XS:i:-2   XN:i:0
    XM:i:0    XO:i:0    XG:i:0
    NM:i:0    MD:Z:36   YT:Z:UU
    YM:i:0
...
```

- Bowtie 2 extracts seed substrings from the read and its reverse complement.

- Seeds are aligned to the reference genome with the help of the genome index.

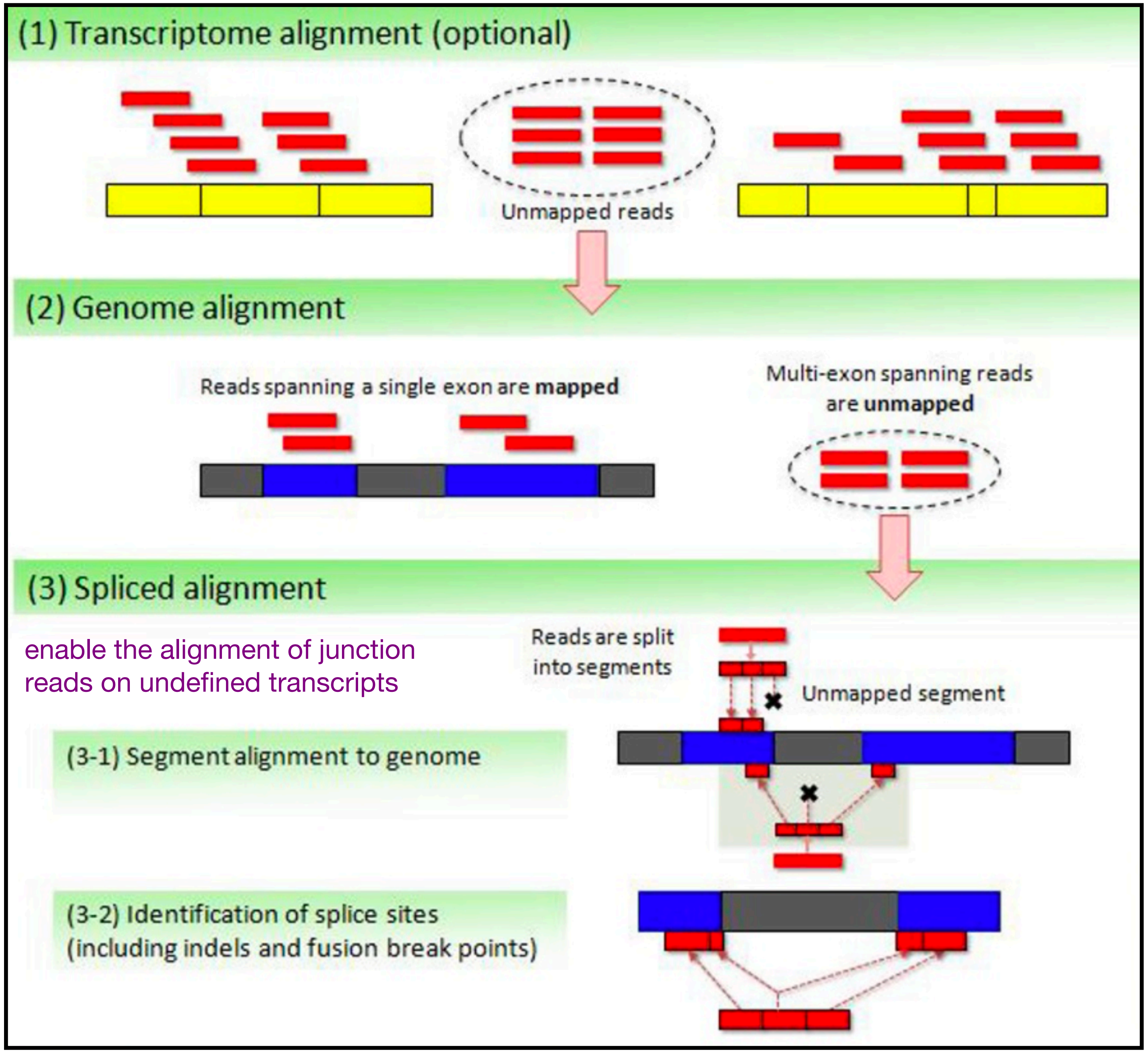
- The precise locations of seeds on the reference genome are calculated from the index.

- Seeds are extended into full alignments on the genome.

Langmead, Ben, and Steven L. Salzberg. "Fast gapped-read alignment with Bowtie 2." Nature methods 9.4 (2012): 357-359.

How to account splicing in RNA-Seq reads?

Tophat2 pipeline


(1) Transcriptome alignment (optional)
Unmapped reads
(2) Genome alignment
Reads spanning a single exon are mapped
Multi-exon spanning reads are unmapped
(3) Spliced alignment
enable the alignment of junction reads on undefined transcripts
Reads are split into segments
Unmapped segment
(3-1) Segment alignment to genome
(3-2) Identification of splice sites (including indels and fusion break points)

Genome biology 14.4 (2013): 1-13.

Tophat2 alignment pipeline:

- in step 1, reads are aligned against the transcriptome (defined in GTF).
- In step 2, unmapped reads from the previous step are aligned against the genome.
- In step 3, reads are split into smaller segments, and these segments are aligned to the genome using spliced alignment strategy.
- The alignment tool used by Tophat2 is Bowtie2.

Alignment-free method

Is it possible to map reads to transcripts without (precise) alignment?

Alignment:

	Transcript 1	Transcript 2	Transcript 3
Read 1	[57, 107]	Not align	Not align
Read 2	Not align	[12, 62]	Not align
Read 3	Not align	Not align	[134,184]
Read 4	[66, 116]	Not align	[85, 135]

Read is aligned at the specific range of [start, end]

Alignment-free:

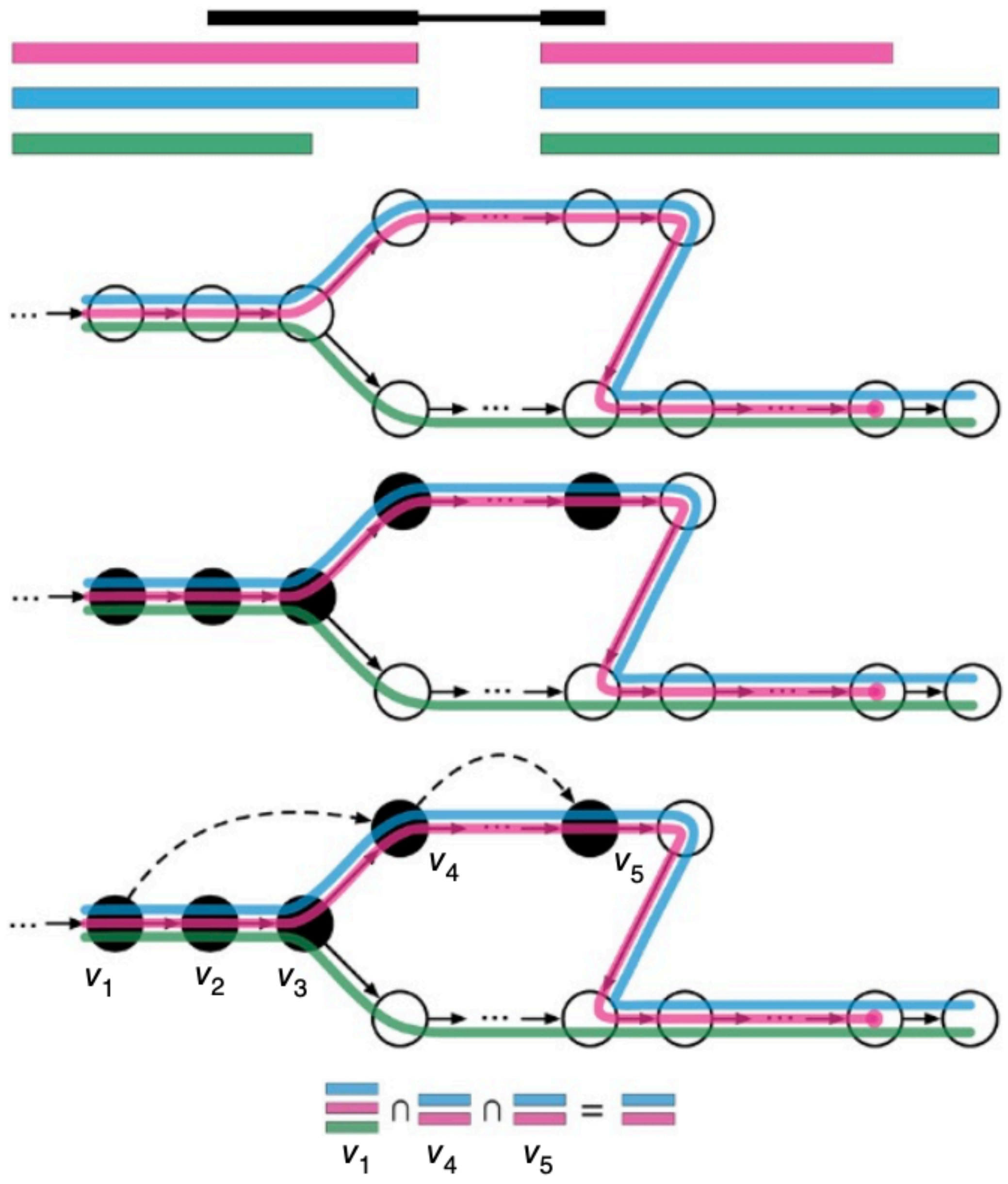
	Transcript 1	Transcript 2	Transcript 3
Read 1	1	0	0
Read 2	0	1	0
Read 3	0	0	1
Read 4	1	0	1

1: compatible
0: incompatible

- **Motivation**: Knowing the compatibility between reads and transcripts is enough to measure transcript expression levels, without needing to know the exact location of the reads on the transcripts.

Pseudo-alignment with TDB graph

Kallisto



- The input for Kallisto includes a reference transcriptome and RNA-Seq reads.
- Kallisto constructs a transcriptome de Bruijn graph (T-DBG) using k-mers as nodes.
- The T-DBG allows for the efficient identification of compatibility relationships between reads and transcripts, without requiring precise read mapping to the transcripts.
- Kallisto is able to quantify transcript expression levels based on these compatibility relationships.

Bray, Nicolas L., et al. "Near-optimal probabilistic RNA-seq quantification." Nature biotechnology 34.5 (2016): 525-527.

Performances of different tools

How different tools compared to each-other in accuracy?

Salmon (another alignment free method) & Kallisto

Tophat-Cufflinks & Tophat-HTSeq

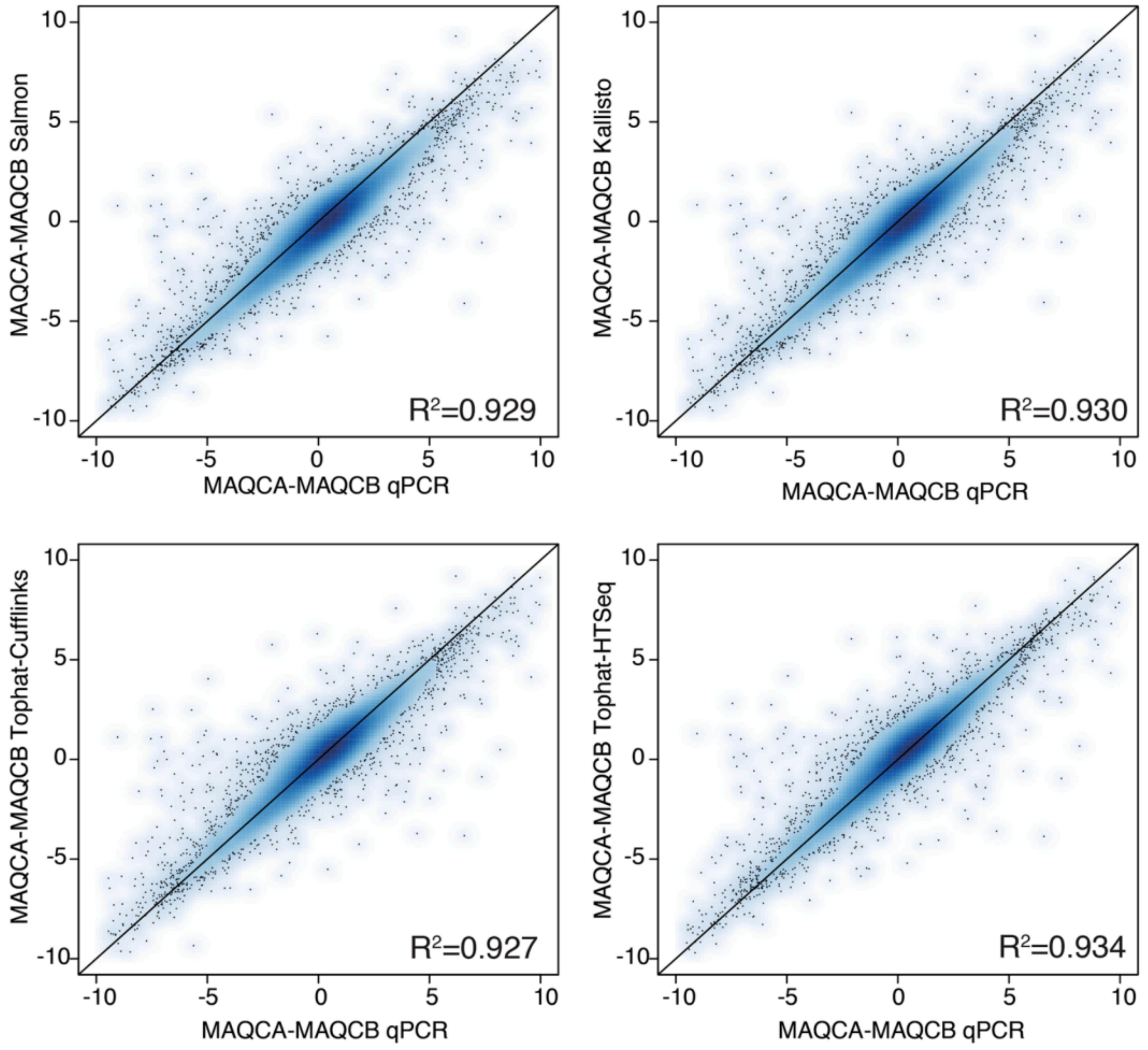
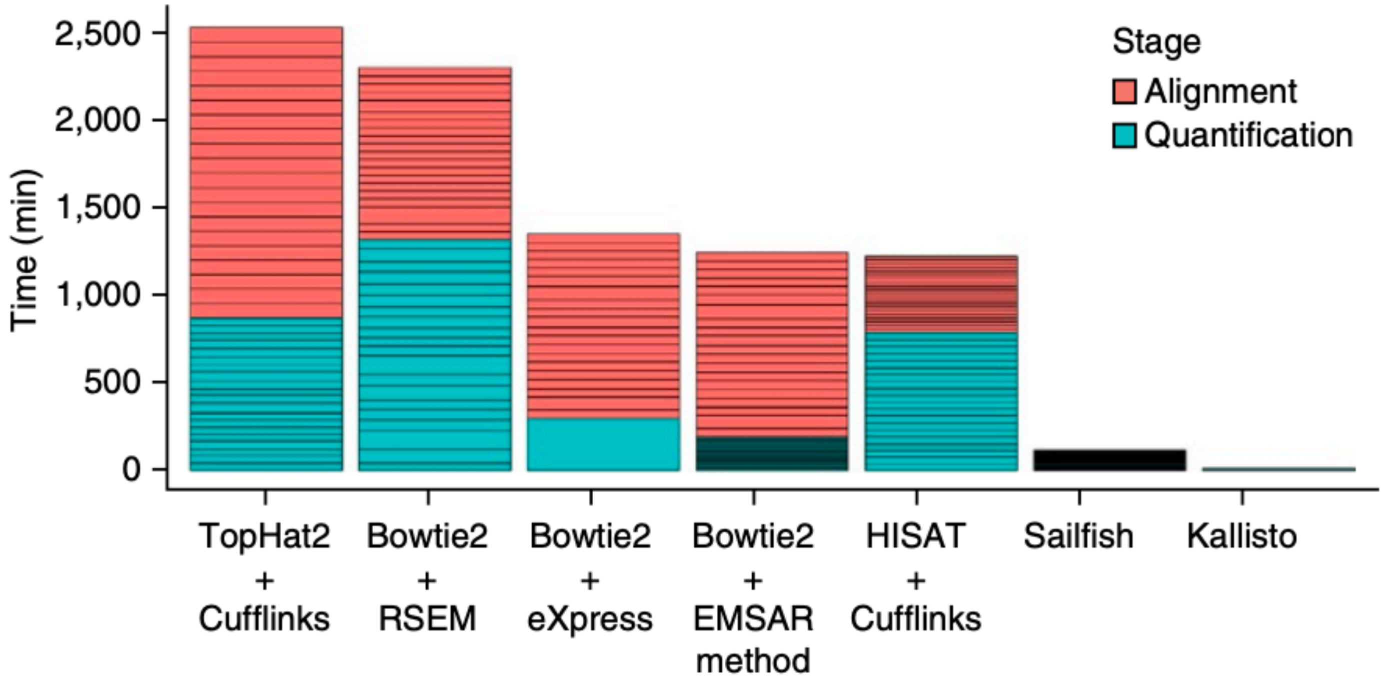


Figure 3. High fold change correlation between RT-qPCR and RNA-seq data for each workflow. The correlation of the fold changes was calculated by the Pearson correlation coefficient. Results are based on RNA-seq data from dataset 1.

BEveraert, Celine, et al. "Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data." Scientific reports 7.1 (2017): 1559.

- When RT-qPCR is used as a technically independent validation, all types of gene expression quantification workflows can explain approximately 93% of the variances (R^2).

Running time of different methods

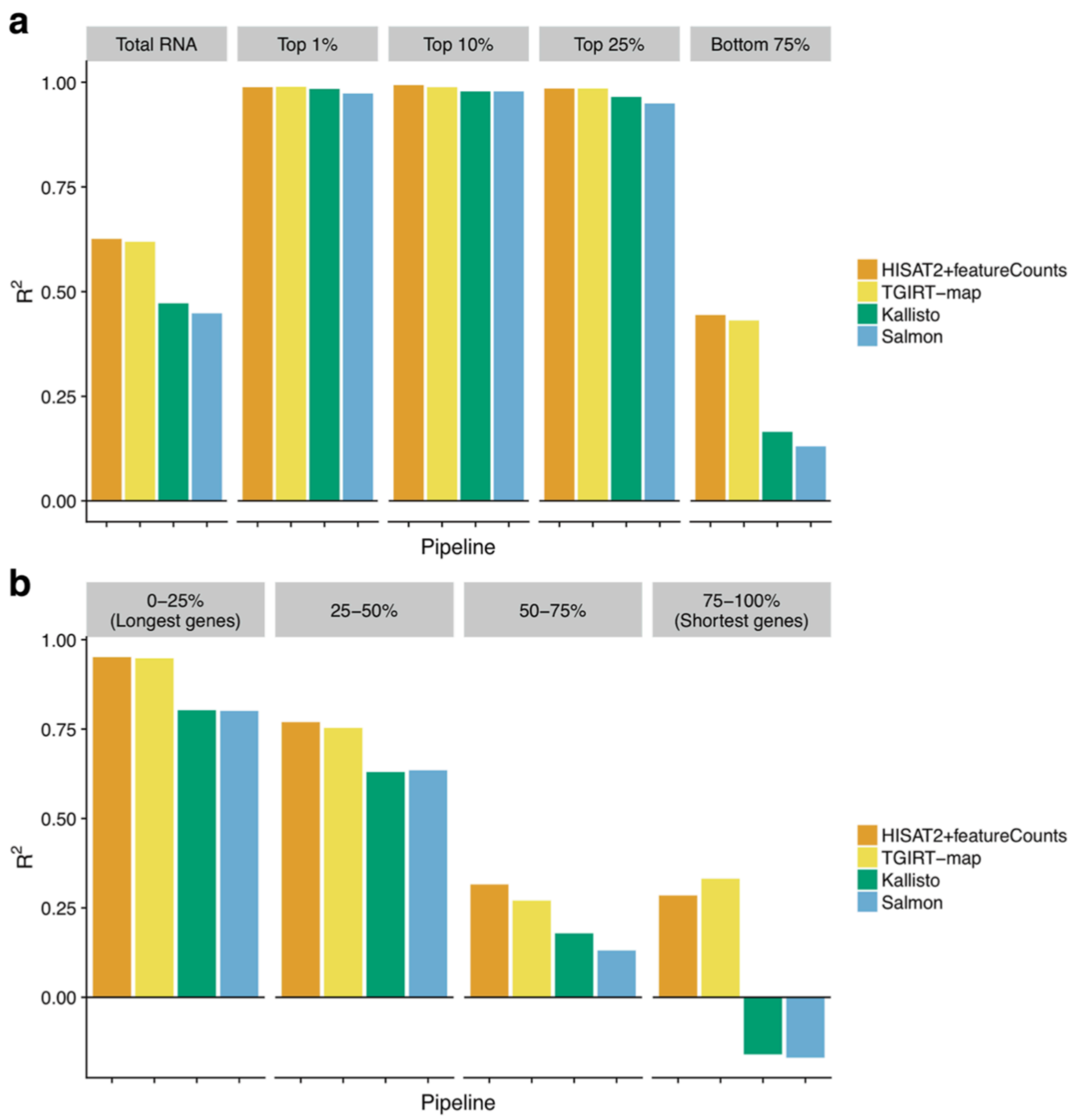


Nature biotechnology 34.5 (2016): 525-527.

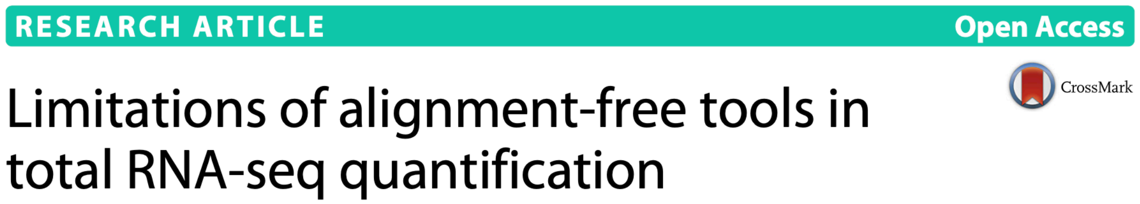
Aligner	Splice-awared	Pesudo-alignment	Speed	Memory demand
bowtie2	No	No	Fast	Small
STAR	Yes	No	Fast	Large
Tophat2	Yes	No	Slow	Large
Hisat2	Yes	No	Fast	Small
Kallisto	Yes	Yes	Altra fast	Very small
Salmon	Yes	Yes	Altra fast	Very small

- For DNA-Seq based assays, *bowtie2* is recommended.
- For RNA-Seq based assays, *Hisat2* or *Tophat2* is recommended.

Why not only use alignment-free methods?



- Alignment-free and traditional alignment-based quantification methods have similar performance for common gene targets such as protein-coding genes.
- However, alignment-free methods have limitations in analyzing and quantifying lowly-expressed genes and small RNAs, particularly when these small RNAs have biological variations.
- Therefore, sliding windows in peak calling cannot be reliably quantified using alignment-free methods due to their small feature (bin) size.


RESEARCH ARTICLE Open Access

CrossMark

Limitations of alignment-free tools in total RNA-seq quantification