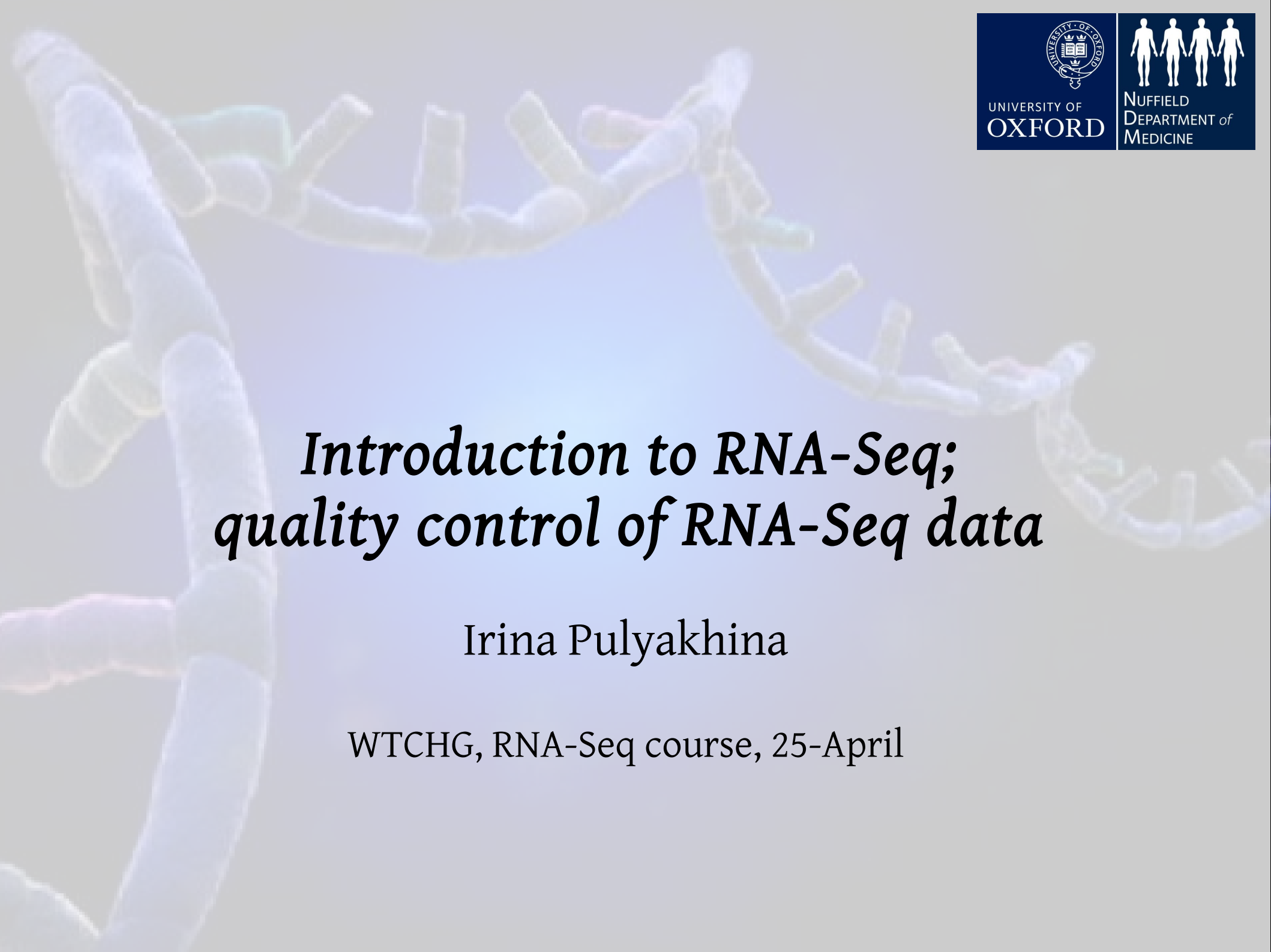


Introduction to RNA-Seq; quality control of RNA-Seq data

Irina Pulyakhina

WTCHG, RNA-Seq course, 25-April

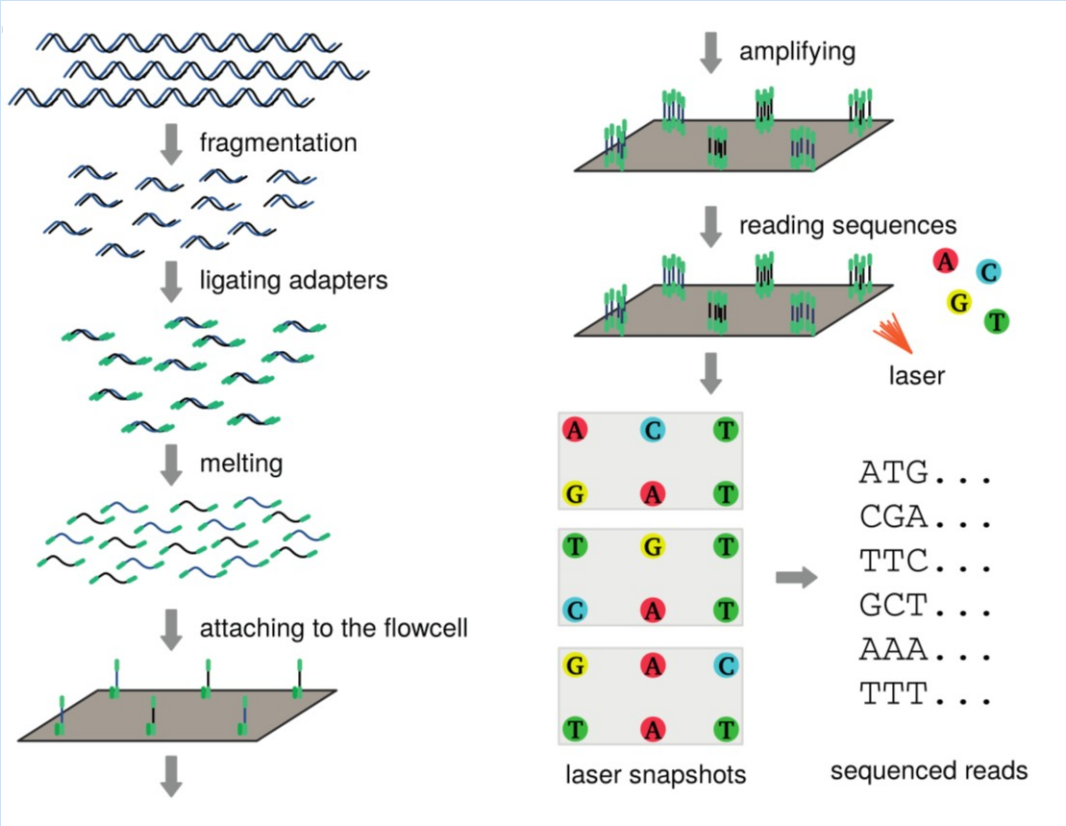
Concept of RNA-Seq

Massively parallel sequencing.

(usually) Whole-transcriptome scale.

Millions of short (100-500nt) reads.

Overview of sequencing workflow



RNA-Seq vs DNA-Seq

The main principle difference – one extra step is introduced for RNA-Seq:

cDNA is synthesized from RNA.

For this step, reverse transcriptase is applied to create DNA from the previously extracted RNA template using random hexamer primers.

The rest of the **sequencing** workflow is very similar to the one for DNA.

Protocols for mRNA sequencing

Enrichment for mRNA:

- oligo-dT capture (disadvantage – other polyA RNAs)

- rRNA depletion (disadvantage – contamination of non-rRNA RNA) is known rather as a **total RNA** sequencing library prep

Protocols for non-mRNA sequencing (1)

- total RNA (coding and non-coding RNA, removing cyt. and mit. rRNA)

Protocols for non-mRNA sequencing (2)

- total RNA (coding and non-coding RNA, removing cyt. and mit. rRNA)

- targeted RNA sequencing (probes)

Protocols for non-mRNA sequencing (3)

- total RNA (coding and non-coding RNA, removing cyt. and mit. rRNA)

- targeted RNA sequencing (probes)

- ribosome profiling (sequencing ribosome-protected mRNA fragments)

Protocols for non-mRNA sequencing (4)

- total RNA (coding and non-coding RNA, removing cyt. and mit. rRNA)

- targeted RNA sequencing (probes)

- ribosome profiling (sequencing ribosome-protected mRNA fragments)

- microRNA (total RNA isolation followed by size selection)

Protocols for non-mRNA sequencing (5)

- Nuclear RNA (pre-mRNA, chromatin associated RNA and nucleoplasmic RNA)

Protocols for non-mRNA sequencing (6)

- Nuclear RNA (pre-mRNA, chromatin associated RNA and nucleoplasmic RNA)

- CAGE (cap analysis; introducing a biotin group to the cap structure, capturing it with oligo primers)

Protocols for non-mRNA sequencing (7)

- Nuclear RNA (pre-mRNA, chromatin associated RNA and nucleoplasmic RNA)

- CAGE (cap analysis; introducing a biotin group to the cap structure, capturing it with oligo primers)

- SAGE (sequencing 3' end; chopping off 11 bp from 3' end)

Protocols for non-mRNA sequencing (8)

- Nuclear RNA (pre-mRNA, chromatin associated RNA and nucleoplasmic RNA)

- CAGE (cap analysis; introducing a biotin group to the cap structure, capturing it with oligo primers)

- SAGE (sequencing 3' end; chopping off 11 bp from 3' end)

- single cell RNA sequencing

Quality control of NGS and RNA data

- biases common for any NGS

- biases specific for RNA-Seq

NGS biases: amplification bias (1)

Amplification bias:

- pre-sequencing amplification to get enough input yield

- bridge amplification on the flow cell

NGS biases: amplification bias (2)

What happens:

- introduction of single-nucleotide errors

- amplification bias – depending of the sequence content fragments anneal/melt with variable efficiency

NGS biases: sequencing bias (1)

- Errors introduces by the polymerase during actual sequencing.

- Signal not strong enough for the CCD camera.

NGS biases: sequencing bias (2)

- Errors introduces by the polymerase during actual sequencing.

  Partial solution: number of reads containing this error.

- Signal not strong enough for the CCD camera.

  Partial solution: Phred quality score.

Sequencing quality score (1)

Phred score – per-base quality of the sequencer's base calling.

Phred score Q is related to the base calling error probability P.

$$Q = -10 \log_{10} P$$

Example: if Phred score of a base is 30, the chance of this base having been called incorrectly is 1 in 1000.

Sequencing quality score (2)

Illumina sequencing quality: > 99.9%

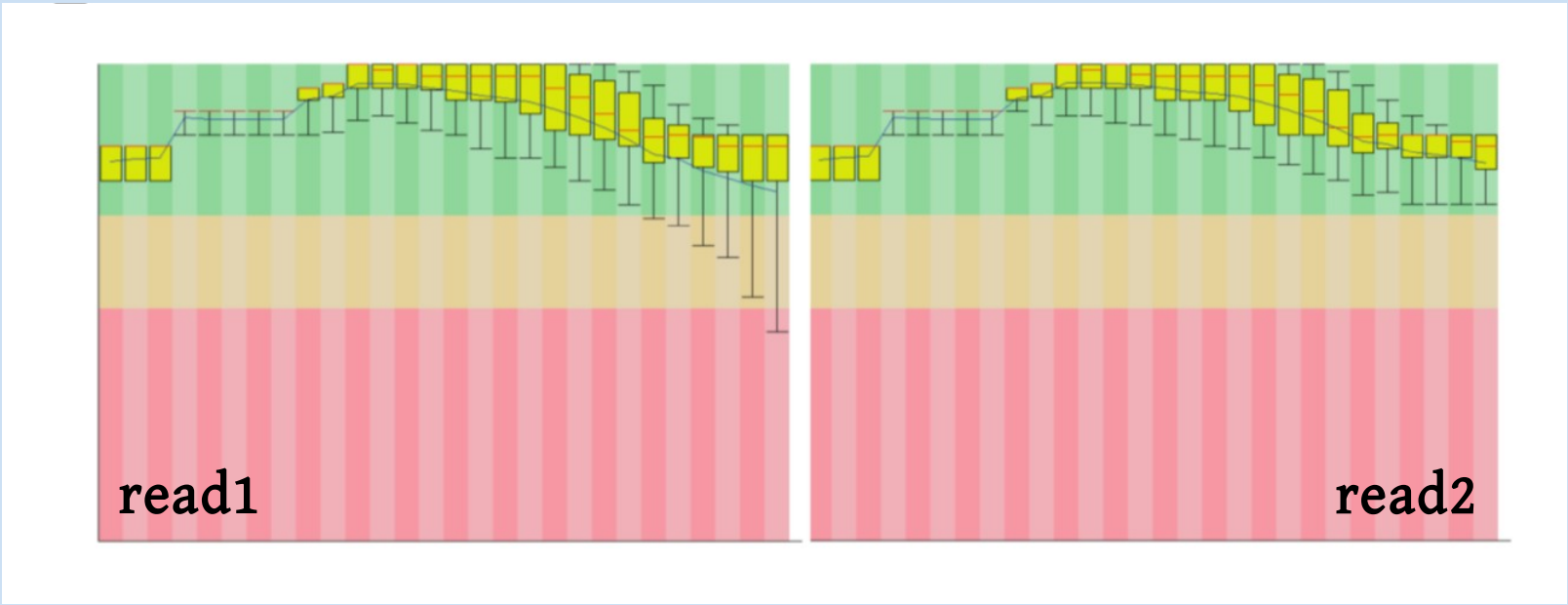
Illumina sequencing quality drops towards the end of the read.

This happens mainly due to phasing (blocking does not work perfectly).

Sequencing quality score (3)

Green area – 28...35

Red area – 0...20



Solution – trimming low-quality bases at the 3' end of each end.

RNA-Seq specific biases: random hexamers

Reverse transcriptase needs a primer to re-create cDNA from an RNA template.

Synthetic randomized 6nt-long oligonucleotides are used for this purpose.

However, transcriptome sequence is not random; therefore, random hexamers do not cover it evenly.

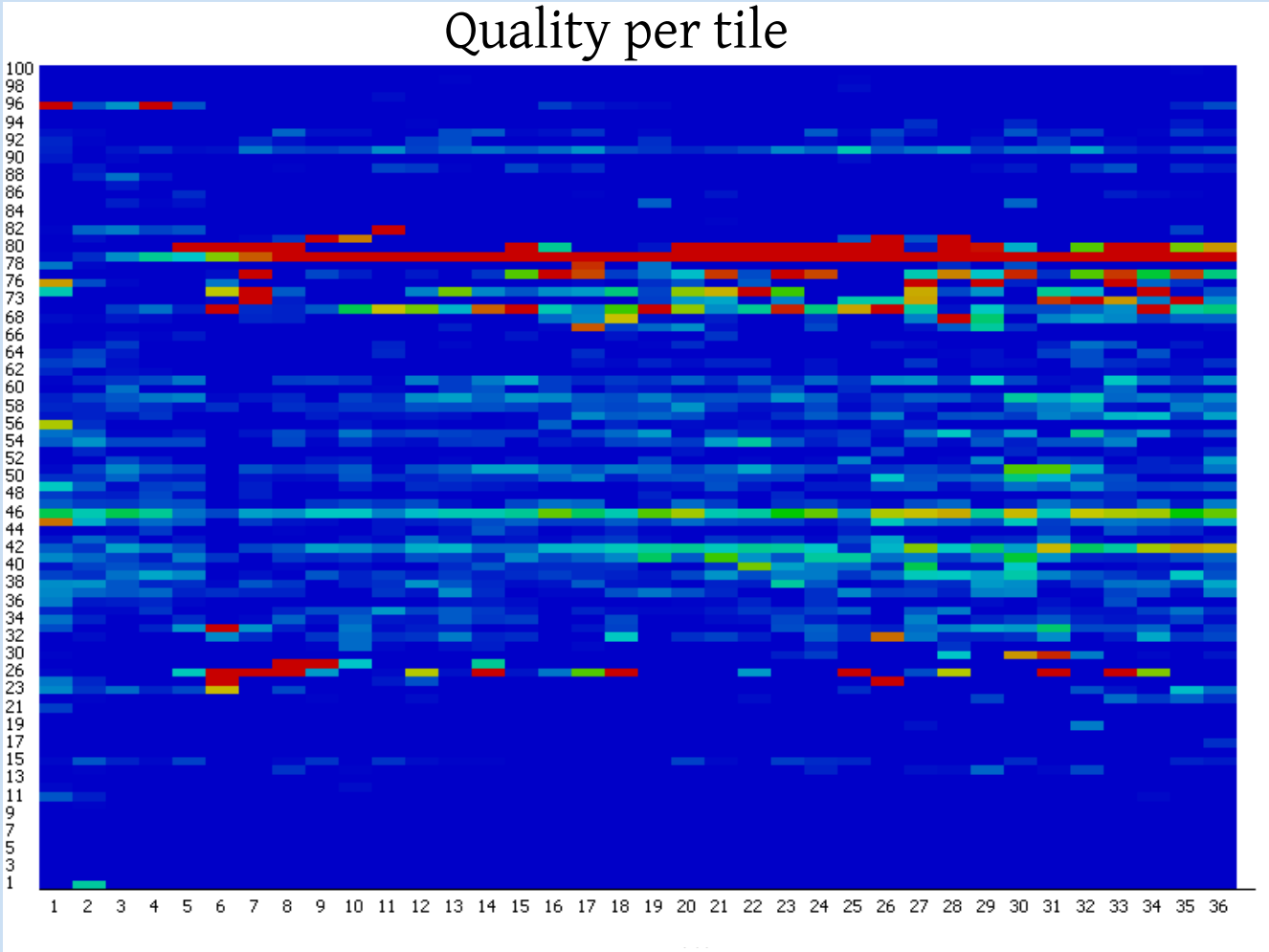
RNA-Seq specific biases: template switch

During the reverse transcription nascent DNA fragment can dissociate from RNA template.

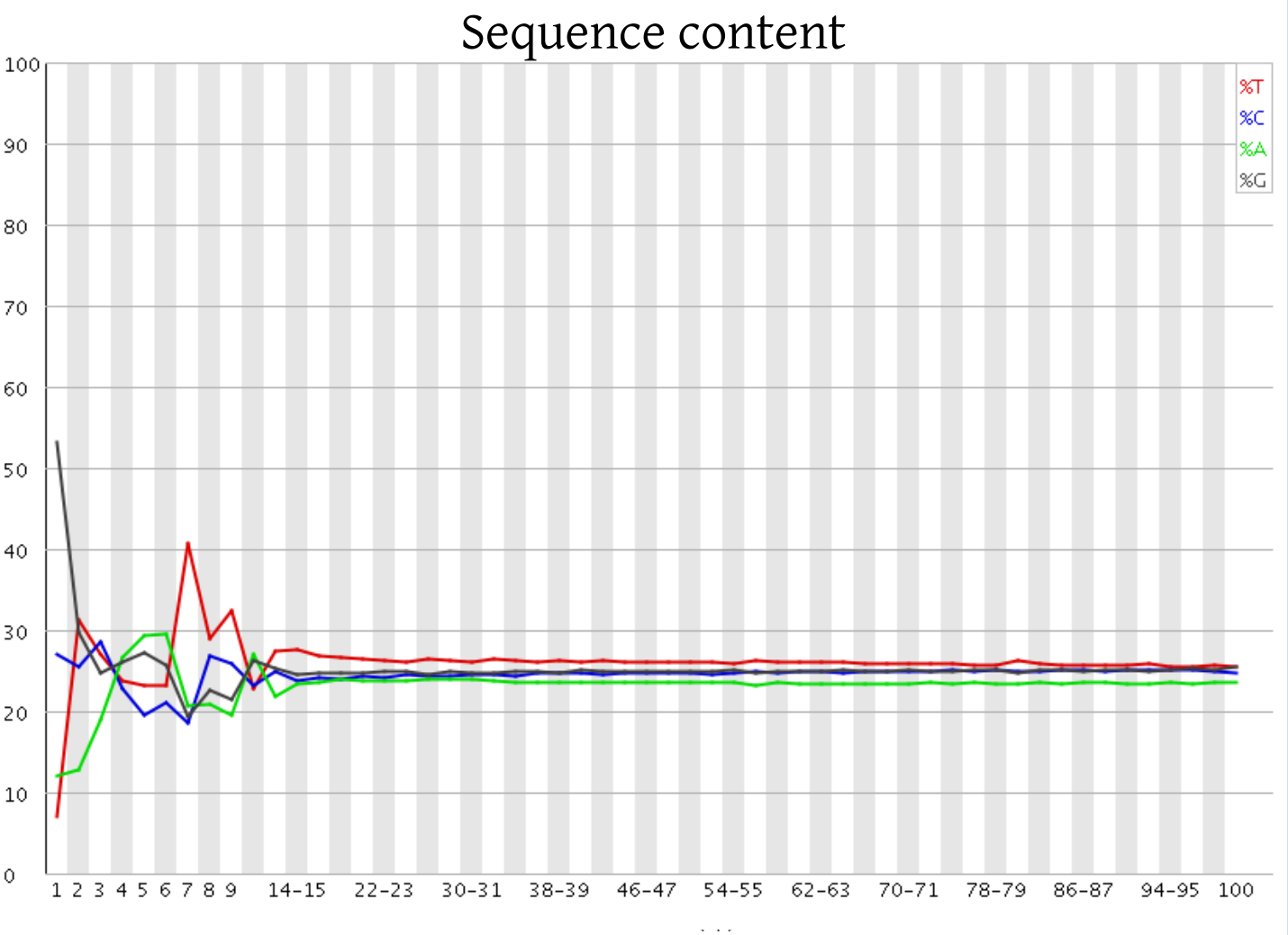
It then reanneals to a different region of RNA with similar sequence.

This leads to generating chimeric reads.

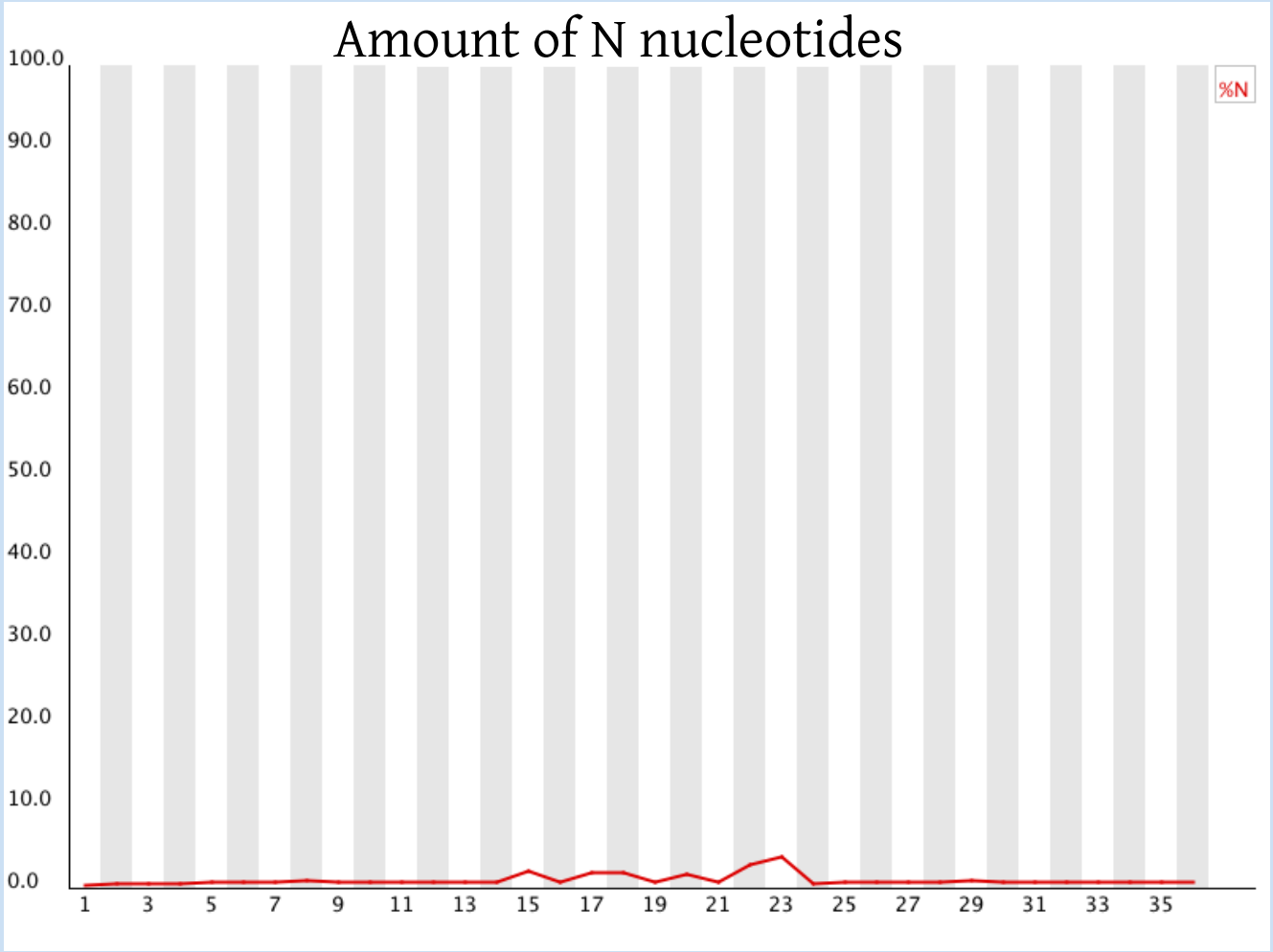
Pre-alignment QC: sequencer's tiles



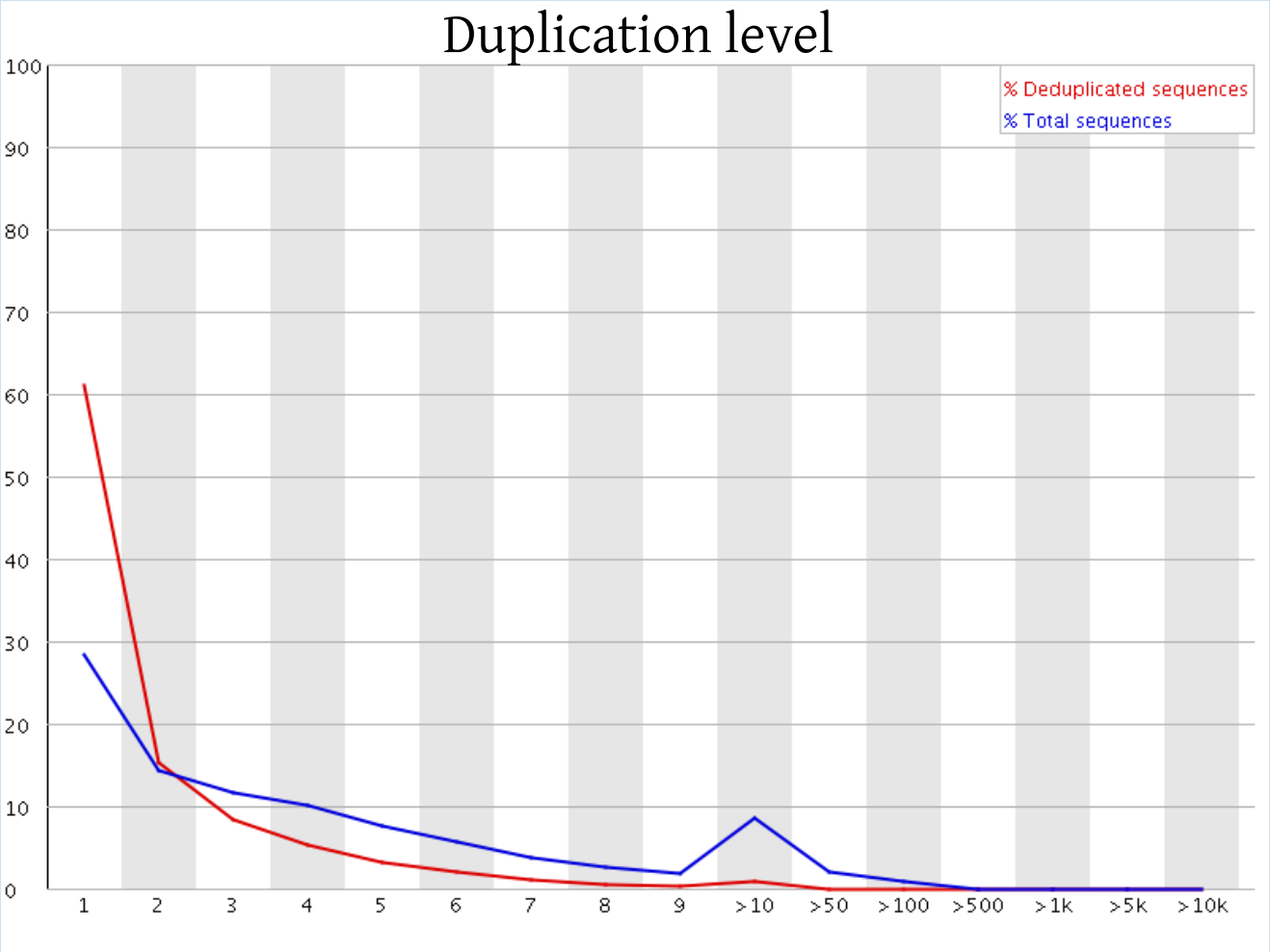
Pre-alignment QC: sequence content



Pre-alignment QC: N content



Pre-alignment QC: duplication level



Pre-alignment QC: adapter content