

Best practices for read trimming for Illumina Stranded mRNA and Total RNA workflows

Explore the impact of the T-overhang on sequence read quality and options for read trimming.

Introduction

RNA sequencing (RNA-Seq) with next-generation sequencing (NGS) is a powerful method for discovering, profiling, and quantifying RNA transcripts. Advances in the Illumina portfolio of RNA library preparation kits deliver the high-quality data researchers require, with a streamlined workflow. Illumina offers three RNA library prep kits:

- **Illumina Stranded mRNA Prep, Ligation** provides a cost-efficient option for coding RNA-focused analyses.
- **Illumina Stranded Total RNA Prep, Ligation with Ribo-Zero™ Plus** enables whole-transcriptome analysis, capturing coding and multiple forms of noncoding RNA.
- **Illumina RNA Prep with Enrichment** brings bead-linked transposome (BLT) technology to RNA enrichment.

Illumina Stranded mRNA and Total RNA Prep kits feature innovations to streamline the ligation-based library preparation chemistry, supporting increased throughput by multiplexing up to 384 unique dual indexes (UDIs) in a single reaction. After cDNA synthesis, double-stranded DNA fragments undergo “A-tailing”, in which a deoxyadenosine nucleotide is added to the 3’ end. This enables rapid ligation of sequencing adapters designed with 3’ T-overhangs, not present in previous ligation-based library prep kits (Figure 1A).

A by-product of this approach is that the first cycle of sequencing Read 1 and Read 2 will be derived from the T-overhang in the adapter and detected as a “T” (Figure 1B), and not from the DNA being sequenced. This may pose a challenge for Real-Time Analysis (RTA) software, as the first base for every cluster on the flow cell will be a “T”, resembling a low-diversity sequence. The presence of this low-diversity sequence within the first six cycles of a read may make it more difficult to define monoclonal clusters during image analysis, particularly for two-channel sequencing by synthesis (SBS) chemistry, which is used by the NextSeq™ 550 and NovaSeq™ 6000 Systems.

This technical note describes the effect of the T-overhang on sequencing data quality and recommends best practices for trimming the first base from sequencing reads to minimize the potential impact on downstream RNA-Seq data analysis.

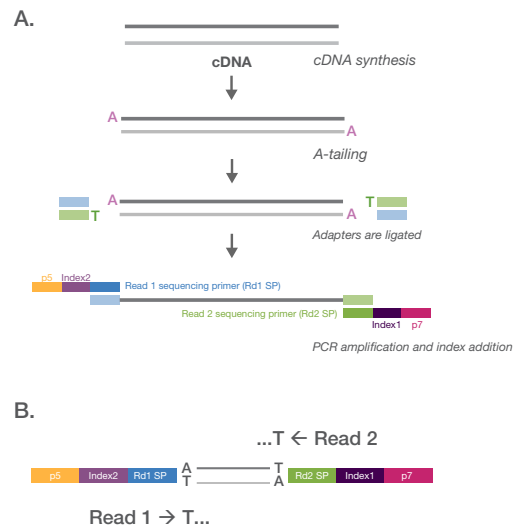


Figure 1: Illumina Stranded RNA library preparation—(A) After cDNA synthesis is complete, ligation of adapters and PCR amplification produces high-quality libraries. (B) The use of T-overhangs in sequencing adapters to facilitate rapid ligation results in all reads starting with a “T” in the first cycle.

Impact of T-overhang on read quality

To explore the impact of the T-overhang on read quality and alignment, a 9-plex pool of Illumina Stranded mRNA Prep libraries were run on the NextSeq 550 and NovaSeq 6000 systems. Analysis of the Q-scores for each cycle showed low-quality calls for the first cycle of Read 1 and Read 2 for the NextSeq 550 and NovaSeq 6000 Systems, as expected (Figure 2). However, analysis of performance metrics across the entire run showed minimal impact on data quality, as measured by the percent passing filter (PF), % ≥ Q30, and yield (Table 1).

Table 1: Impact of T-overhang on performance metrics

System	Read length	% PhiX ^a	% Q30	Yield (Gb)
NextSeq 550 System (v2 chemistry)	2 × 75 bp	0%	91.25%	77.82
NovaSeq 6000 System (S1 flow cell) ^b	2 × 75 bp	0%	91.79%	278.81
NovaSeq 6000 System (S4 flow cell)	2 × 75 bp	0%	94.50%	1890

- a. No PhiX was loaded in these sequencing runs to present the flow cell with only the T-overhang in the first read cycle.
- b. NovaSeq 6000 v1.0 reagents were used for this sequencing run.

* Recommendations presented in this technical note do not apply to the NextSeq 1000 and NextSeq 2000 Sequencing Systems, which use innovations to the two-channel SBS chemistry. A dark cycle custom recipe is recommended that will avoid the first base associated with the T-overhang. For more information, see the Illumina Stranded Total RNA Prep or Stranded mRNA Prep Reference Guides.

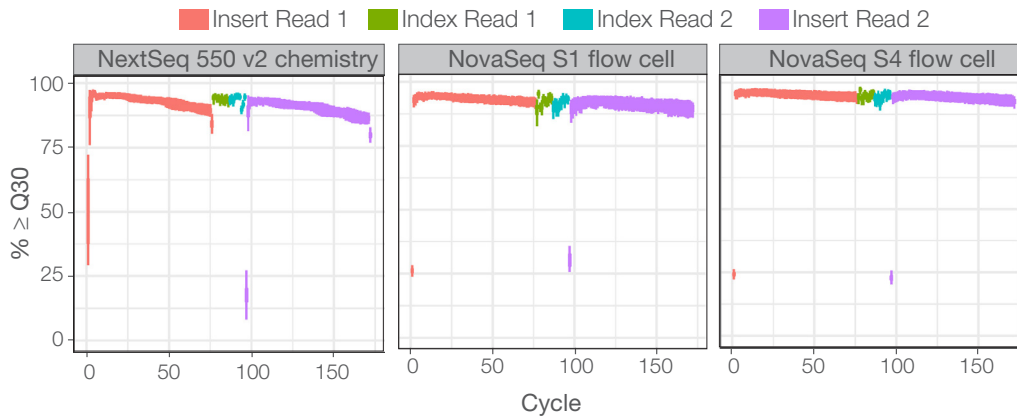


Figure 2: Impact of T-overhang on first cycle read quality—Including the T-overhang resulted in low read quality in the first cycle for Insert Read 1 and Insert Read 2 on the NextSeq 550 and NovaSeq 6000 Systems.

Impact of T-overhang on sequence alignment

Alignment to a reference sequence is a key step in RNA-Seq analysis. To evaluate the impact of the T-overhang on alignment, sequencing data were analyzed with or without trimming the T-overhang before alignment using the FASTQ Toolkit or the RNA-Seq Alignment app. Results show that key performance metrics, including percent duplicates, median coefficient of variation (CV) of coverage, and percent aligned, are largely unaffected whether the T-overhang is trimmed or not (Figure 3).

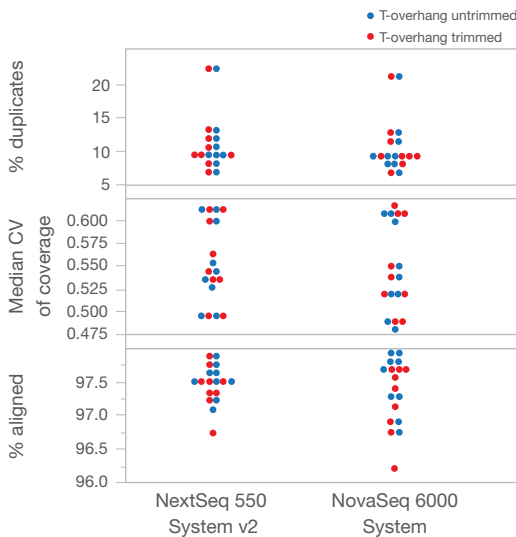


Figure 3: Impact of T-overhang on alignment performance metrics—Evaluation of inclusion of the T-overhang on RNA-Seq alignment shows minimal impact on key metrics for the NextSeq 550 and NovaSeq 6000 Systems.

To investigate the impact of the T-overhang on RNA-Seq alignment further, the percentage of reads with similar (± 1 bp) or identical alignment, whether the first cycle was trimmed bioinformatically or untrimmed, was determined. On average the untrimmed and trimmed data has 96.4% similar alignments (Figure 4A). Over 75% of reads

have identical alignment, whether they were trimmed or untrimmed for the first cycle (Figure 4B).

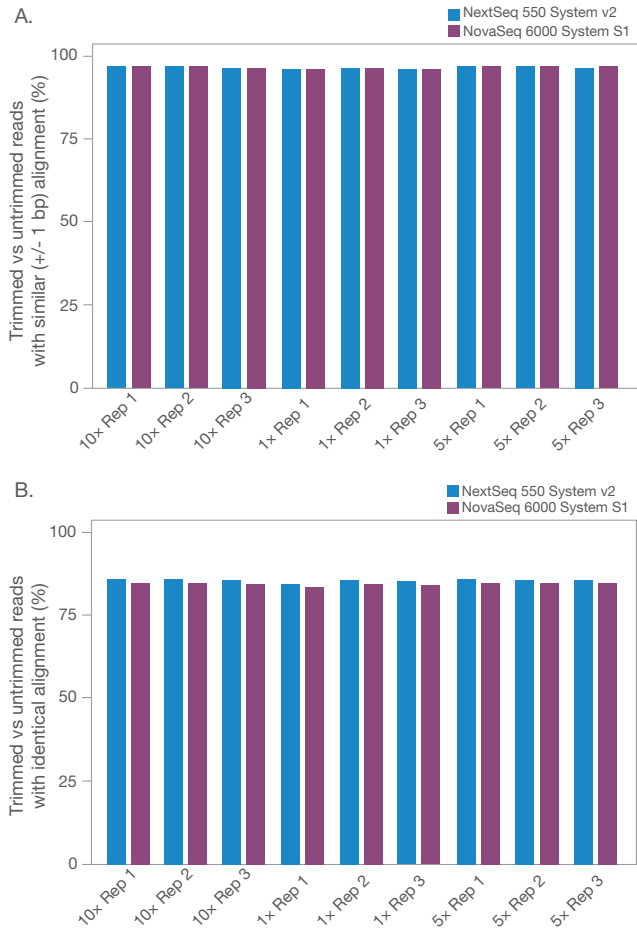



Figure 4: Impact of T-overhang on alignment—The percentage of reads with (A) similar (± 1 bp) or (B) identical alignment, whether the first cycle was trimmed or untrimmed is plotted, showing minimal impact across replicates of Illumina Stranded mRNA libraries run on the NextSeq 550 and NovaSeq 6000

Systems. The Y-axis represents the % of reads with either similar or identical alignment position.

Options for trimming the T-overhang

Although data presented in this technical note indicates that leaving the T-overhang untrimmed has minimal impact on overall read quality and sequence alignment, it is unclear how it might affect downstream analyses. Therefore, it is recommended as a best practice to trim the T-overhang from FASTQ files before proceeding with data analysis.

 For detailed guidance, refer to the Illumina Stranded mRNA Prep, Ligation Reference Guide† at support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/illumina_prep/RNA/illumina-stranded-mrna-reference-1000000124518-01.pdf

FASTQ Toolkit BaseSpace™ App

The FASTQ Toolkit App is used to trim the T-overhang from FASTQ files. In the Base Trimming section, set “Trim reads at the 5'-end by n positions” to 1 to remove the first base from input FASTQ files (Figure 5). This app is available in BaseSpace Sequence Hub.

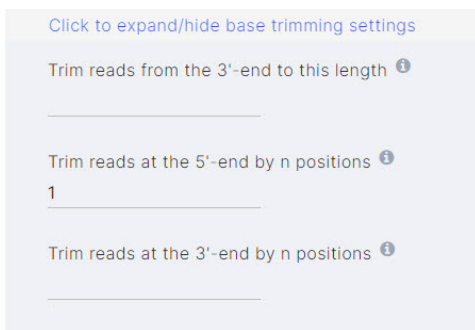


Figure 5: Base trimming using FASTQ Toolkit—Bases can be trimmed from either the 5'- or 3'-end by specifying the number of bases to be trimmed from each end.

Sample sheet configuration for bcl2fastq/bcl2fastq2 conversion software

If bcl2fastq or bcl2fastq2 is used for FASTQ generation, users can add the following settings to the [Settings] section of the SampleSheet.csv file. These settings configure the FASTQ generation to start from the second cycle, skipping the T-overhang.

```
[Settings]
Read1StartFromCycle,2
Read2StartFromCycle,2
```


If a sequencing run has multiple library types, one of which does not include a T-overhang, the libraries will need to be demultiplexed separately using distinct SampleSheet.csv files. This is because the FASTQ generation step cannot handle more than one setting per run.

Sample sheet configuration for BCL Convert software

BCL Convert is a standalone local software app that converts the Binary Base Call (BCL) files produced by Illumina sequencing systems to FASTQ files, performs adapter handling (through masking and trimming), and produces metric outputs. Users can add the following settings to configure the FASTQ conversion to start from the second cycle, skipping the T-overhang:

```
[BCLConvert_Settings]
OverrideCycles,N1Y75;I10;I10;N1Y75
```

In this example, the N1Y75 setting will skip the first cycle, but process the remaining 75 cycles of Read 1 and Read 2, given a read length of 76 bp. The I10 setting indicates an index read length of 10 bp. Settings will need to be adjusted for specific read lengths.

 For more details, refer to the BCL Convert Software Guide at support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/bcl_convert/bcl-convert-software-guide-1000000094725-00.pdf

Summary

Illumina Stranded mRNA and Total RNA Prep kits offer streamlined solutions for clear and comprehensive RNA-Seq analyses. To support increased sample throughput, these library prep kits feature adapter-ligation chemistry mediated by T-overhangs in the adapters and A-tailing of the inserts. As a consequence, the first cycle of sequencing Read 1 and Read 2 will be derived from the T-overhang and detected as a “T”. Results indicate that this will have minimal impact on overall data quality and sequence alignment. However, as a best practice, it is recommended that the T-overhang be trimmed before data analysis begins. There are several options for base trimming that can be executed quickly and easily.

† Options and guidance for trimming will be the same for Illumina Stranded Total RNA libraries.

