Improving transcriptome analysis by incorporating Unique Molecular Identifiers in RNA-sequencing libraries

Dora Posfai, Keerthana Krishnan, Chen Song, Pingfang Liu, Gautam Naishadham, Bradley W. Langhorst, Eileen T. Dimalanta, & Theodore B. Davis | New England Biolabs, Inc.

Introduction

RNA-sequencing is a powerful tool for the study of gene regulation and function. Applications of RNA-sequencing have vastly expanded with improvements to low RNA input library preparation methods. Although analyzing low-input samples is necessary to answer certain biological questions, generating these libraries introduces greater biases with additional PCR cycles that amplify certain transcripts disproportionately. To better gauge true transcript abundance levels, unique molecular identifiers (UMIs) can be used to distinguish PCR duplicates from reads that originate from distinct molecules. Accurate identification of such artifacts can significantly impact quantification of transcript abundance. To address this need in RNA-seq, we have developed barcoded adaptors incorporating UMIs.

We have developed 96 UMI-containing barcoded adaptors and optimized their use across various RNA inputs (10 ng to 1 µg). Ligation of barcoded adaptors followed by PCR enrichment produced high-quality libraries and sequencing metrics (e.g., yields and coverage across transcript length). However, duplication rates significantly differed when utilizing traditional computational approaches to identify duplicates based on mapping position compared to analysis incorporating UMIs. As many as 90% of reads identified as duplicates using read position alone were determined to in fact originate from unique molecules, increasing the total number of reads available for further analysis.

In this study we show that the incorporation of UMIs into RNA-sequencing analysis allows for a more accurate calculation of transcript abundance. Use of UMIs allows for identification of true read duplicates in RNA-seq, thereby increasing transcript detection accuracy and improving sensitivity of differential expression analysis.

Methods

Poly A enrichment → 1st strand cDNA synthesis → 2nd strand cDNA synthesis → End repair/ dAtailing → Adaptor ligation → USER → PCR enrichment

Sample preparation workflow

- Universal Human Reference RNA (Agilent) with ERCC RNA Spike-In Mix (ThermoFisher) was used for library preparation.
- mRNA was isolated using NEBNext Poly(A) mRNA Magnetic Isolation Module (NEBNext® #E7490).
- 1st strand cDNA was synthesized, followed by 2nd strand cDNA synthesis and end repair/dA tailing using NEBNext Ultra II Directional RNA Library Prep Kit (NEBNext® #E7760).
- Adaptors were ligated (NEBNext truncated adaptor, NEBNext UMI-containing barcoded adaptor, or IDT xGen Dual Index UMI Adapter) followed by PCR amplification of libraries. It is during the PCR enrichment that barcodes are incorporated for libraries constructed with NEBNext truncated adaptors.

Truncated adaptor (non-UMI)

UMI barcoded adaptor

Sequencing and analysis

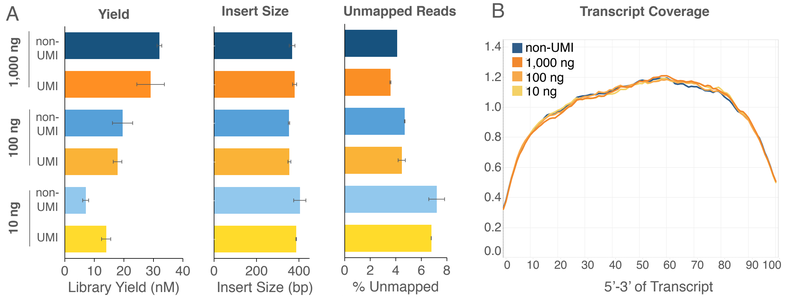
- Libraries were sequenced on an Illumina® NextSeq 500, 2x70 base paired reads.
- Reads were aligned to the human reference genome GRCh38[1,2]. Duplication rates were computed using the Picard MarkDuplicates[3] method or with the benefit of UMI information using fgbio AnnotateBamWithUmis[4].
- The Picard MarkDuplicates tool marks reads mapping to the same start coordinate in the genome as PCR duplicates, unable to distinguish whether the reads originated from the same molecule. Fgbio uses an alternate approach, only removing reads that share the same 11 base pair unique molecular identifier that was added before PCR amplification.

References

[1] Li, H.; Durbin, R. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* **2009**, *25* (14), 1754–1760.
[2] Langmead, B.; Salzberg, S. L. Fast Gapped-Read Alignment with Bowtie 2. *Nature Methods* **2012**, *9*, 357–359.
[3] https://github.com/fulcrumgenomics/fgbio
[4] Broad Institute (2015) Picard tools http://broadinstitute.github.io/picard/

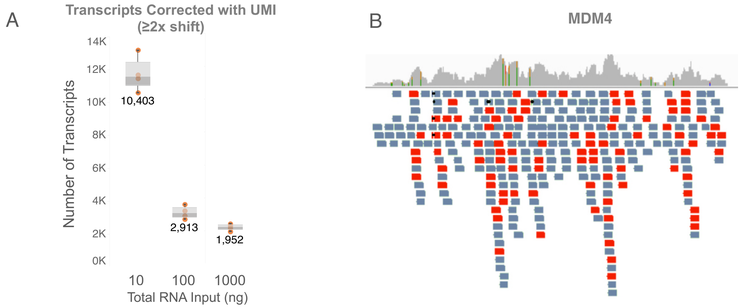
Results

Unique dual index UMI adaptors produce high-quality libraries and sequencing metrics comparable to truncated adaptors.



Libraries produced with UMI-containing barcoded adaptors produce high yields and high-quality sequencing metrics across various inputs (1µg, 100ng, 10 ng).
(A) Comparable yields, insert size, and unmapped reads are observed in UMI (*orange*) and non-UMI (*blue*) libraries at all inputs that were tested in technical triplicates. (B) UMI libraries (*orange*) have even coverage across transcripts, indistinguishable from non-UMI libraries (*blue*).

Removing duplicate reads that are detected by UMIs significantly shifts transcript abundance.

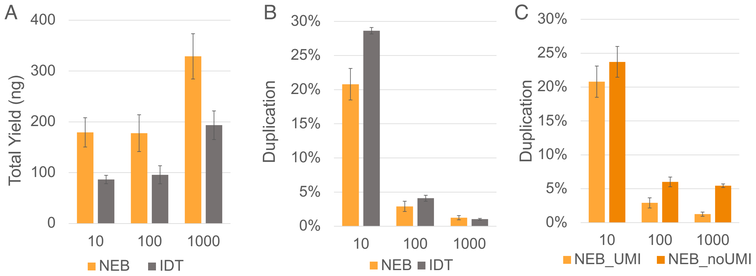


A significant number of transcripts have a ≥2x change in transcript counts when duplicates are removed.
(A) The average number of transcripts with a ≥2x shift in abundance is shown when duplicate reads are removed based on UMI analysis versus no removal of duplicates. An average of 4 technical replicates at three inputs (1,000 ng, 100 ng, and 10 ng) is shown before and after removal of duplicate reads. Each sample was downsampled to 10 million reads. Low-input libraries showed the greatest number of transcripts affected by PCR amplification. (B) MDM4 is an example of a gene with a high portion of mapped reads (*blue and red bars*) determined to be PCR duplicates (*red bars*) based on UMI analysis. Utilizing this information, it is possible to remove duplicate reads introduced by PCR amplification for downstream analysis while retaining reads that originate from unique molecules.

Acknowledgements

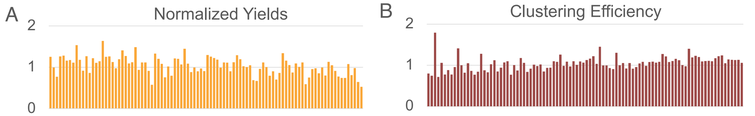
Thank you to the NEB Sequencing Core (Laurie Mazzola, Danielle Fuchs, Kirsten Augulewicz, and Harold Bell) for all their sequencing support.

Comparison of unique dual index UMI adaptors used for library preparation.



Comparison of library yields and duplication rates with various unique dual index UMI adaptors.
(A,B) During adaptor ligation either the NEBNext Unique Dual Index UMI Adaptors (UMI length = 11 bases) or IDT xGen Dual Index UMI Adapters (UMI length = 9 bases) were used. (A) The average library yield of triplicates is shown for three starting total RNA inputs: 10, 100, and 1,000 ng. Final library yields were quantified using the Agilent Tapestation 4200. (B) Libraries were sequenced on the Illumina Next-seq 500 and downsampled to 5 million reads. Duplication rate was determined utilizing the UMI sequence and mapping location. NEBNext Unique Dual Index UMI Adaptor libraries produced libraries with a lower percentage of read duplicates. (C) Duplication rate for libraries produced with NEBNext UMI adaptor libraries analyzed by two computational methods: utilization of UMIs (*light orange*) or read mapping position alone (*dark orange*).

All 96 NEBNext® unique dual index UMI adaptors perform consistently in library generation and sequencing.



NEBNext Unique Dual Index RNA UMI Adaptors have equal ligation, amplification, and subsequent clustering. (A) Library yields were quantified by Agilent TapeStation 4200 and normalized. Adaptor ligation efficiency was robust with uniformity across all 96 unique dual RNA UMI adaptors. Each bar represents the average of at least 2 technical replicates. (B) NEBNext Unique Dual Index UMI Adaptors have uniform clustering efficiency. 96 libraries were pooled and sequenced on the NextSeq 500. No clustering bias was observed across the 96 unique dual UMI adaptor libraries.

Conclusions

Incorporating our newly developed UMI barcoded adaptors into RNA-sequencing library preparation results in:

- Robust library yields and high-quality sequencing metrics
- No introduction of sequencing bias with UMIs
- Improved quantification of transcript abundance

A more accurate assessment of duplicate reads increases the number of reads that can be used for downstream analysis, having potential implications for:

- Improved detection of low frequency variants
- Increased sensitivity for differential expression of lowly expressed genes

NEBNext® Multiplex Oligos for Illumina (Unique Dual Index UMI Adaptors RNA Set1) were tested and are compatible with the following library prep kits:

- NEBNext® Poly(A) mRNA Magnetic Isolation Module
- NEBNext® rRNA Depletion Kit v2
- NEBNext® rRNA Depletion Kit (Bacteria)
- NEBNext® Globin and rRNA Depletion
- NEBNext® Ultra II RNA Library Prep Kit
- NEBNext® Ultra II Directional RNA Library Prep Kit