Incorporation of Unique Molecular Identifiers (UMIs) into Unique Dual Indexing (UDI) of Samples Improves the Accuracy of Quantitative Next Generation Sequencing

Keerthana Krishnan, Pingfang Liu, Chen Song, Dora Posfai, Karen McKay, Jian Sun, Gautam Naishadham,
Bradley W. Langhorst, Eileen T. Dimalanta and Theodore B. Davis
New England Biolabs, Inc.

Introduction

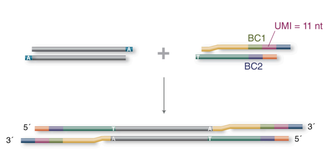
The use of Unique Molecular Identifiers (UMIs) have become increasingly popular and offer a multitude of advantages especially when paired with unique dual indexing (UDI).

We incorporate UMIs into UDI adaptors and assess their effect on the accuracy of quantitative sequencing assays. We studied the effectiveness of various computational methods to account for UMIs and remove base-calling errors introduced during sequencing. Using our NEBNext® Ultra™ II DNA Library Prep kit we demonstrate that the sensitivity of variant detection is improved with UMI consensus calling. To test the efficacy of UMIs in RNA-seq we introduced UMI-containing barcoded adaptors into our RNA-Seq workflow (NEBNext Ultra II Directional RNA Library Prep), optimized across various RNA inputs.

Our approach involves a simple new UMI-containing UDI adaptor design that can also be applied to other sequencing methods and platforms. We conclude that combining unique dual sample indexing with UMI molecular barcoding further improves data analysis accuracy, especially on patterned flow cells

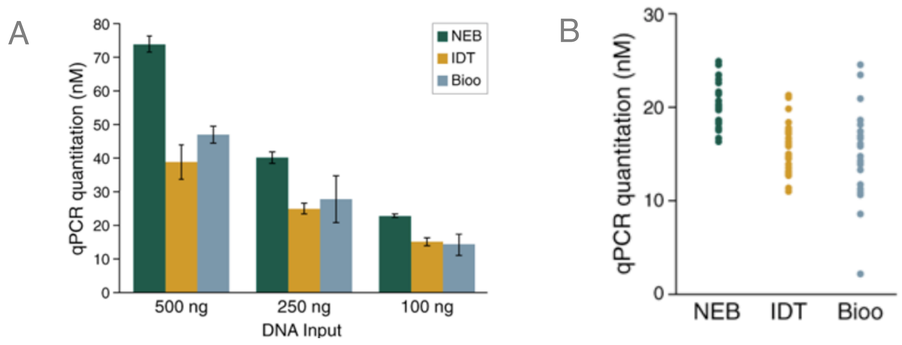
Methods and Results: DNA Workflow and Sequencing with UMI Adaptors

NEBNext Unique Dual Index UMI Adaptors have higher ligation efficiency



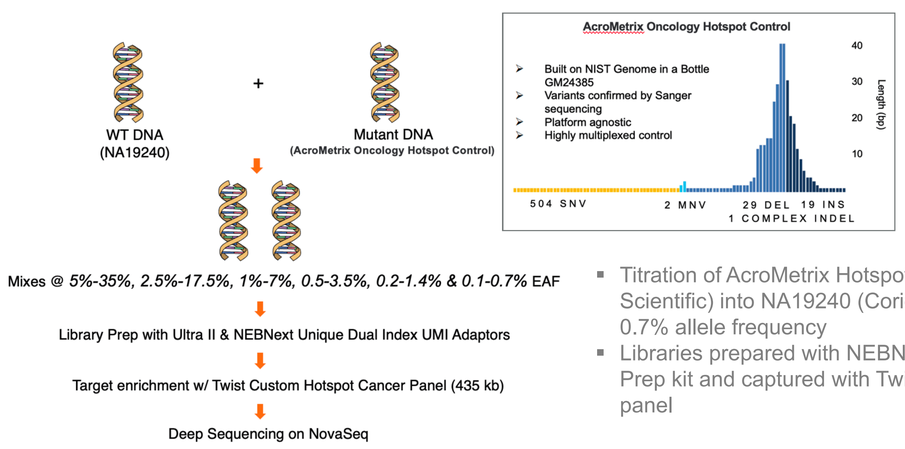
- NEBNext Unique Dual Index UMI Adaptor Design

- Test 100, 250, 500 ng DNA input amounts in NEBNext Ultra II FS PCR-free workflow
- Compare ligation efficiency of NEBNext Unique Dual Index UMI adaptors with other suppliers
- Quantify library yield using qPCR (NEBNext Library Quant Kit)

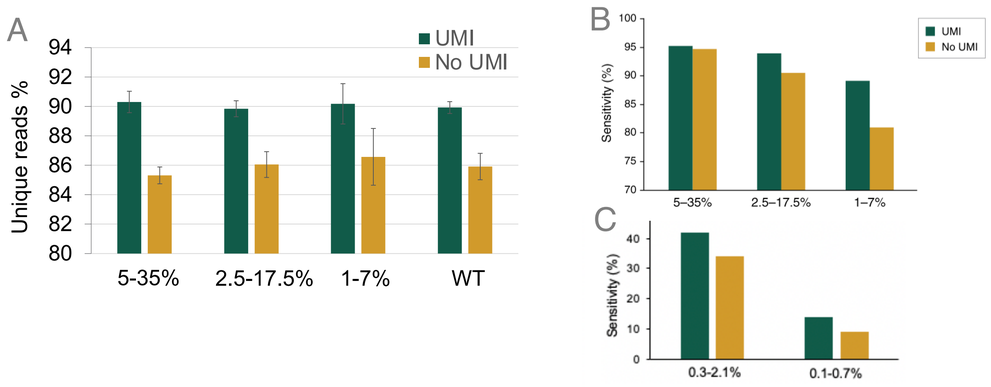


A) Libraries were prepared with 100, 250, and 500 ng inputs of human cell line NA19240 genomic DNA (Coriell Institute for Biomedical Research) & UDI adaptors from different suppliers using the Ultra II FS PCR-free DNA Library Prep Kit and libraries were quantified by qPCR (NEBNext Quant Kit). B) 90 Libraries were prepared with 100 ng NA19240 genomic DNA using the Ultra II FS PCR-free DNA Library Prep Kit in a 96 well plate. Of the 90 libraries, 30 different UDI adaptors each from NEB, IDT and Bioo were used for ligation and were quantified by qPCR.

NEBNext Unique Dual Index UMI Adaptors allow more sensitive low variant detection



- Titration of AcroMetrix Hotspot Control DNA[1] (Thermo Scientific) into NA19240 (Coriell Institute) down to 0.1-0.7% allele frequency
- Libraries prepared with NEBNext Ultra II DNA Library Prep kit and captured with Twist custom cancer hotspot panel



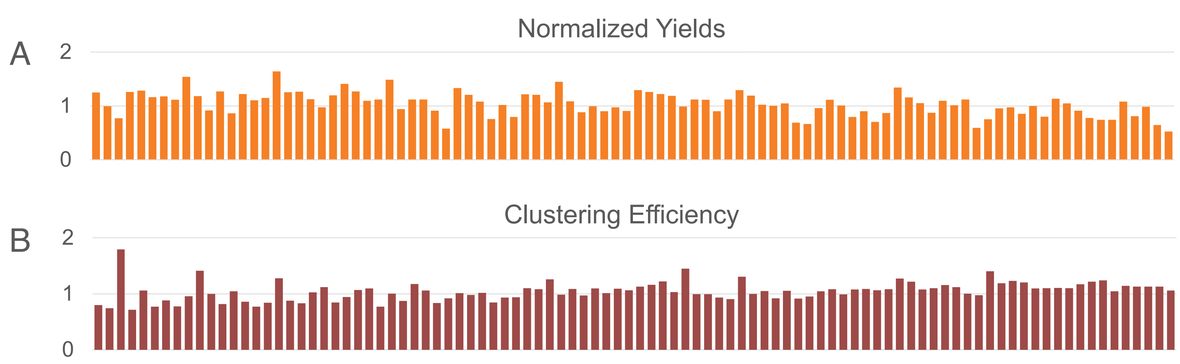
AcroMetrix Oncology Hotspot Control DNA (Thermo Scientific #969056) was used as mutation DNA source (>500 mutations with 5-35% allele frequency) and mixed with NA19240 DNA to generate a series of allele frequencies. Libraries were constructed with NEBNext Unique Dual Index UMI adaptors and multiplex hybrid capture was performed on all samples using a customized panel for 152 genes from Twist Bioscience. Libraries were sequenced on a NovaSeq™ 6000 (2x140) and downsampled to 110 million reads and mapped to hg38 with BWA MEM (0.7.17). Mapped reads were analyzed by MarkDuplicates (Picard 2.20.6)[2] without utilizing the UMI sequence or by building UMI consensus reads (Fgbio 0.8.1)[3]. The final BAM files were used to call somatic variants with Strelka2 (2.9.10). (A) Total and correct SNV calls increased when using UMI correction. (B) Variant detection sensitivity improved with UMI consensus calling. (C)The lower the allele frequency, the more benefit UMI produced in SNV detection.

Methods: RNA-Seq Workflow

- Universal Human Reference RNA (Agilent) with ERCC RNA Spike-in Mix (Thermo Scientific) was used for library input. mRNA was isolated using the NEBNext Poly(A) mRNA Magnetic Isolation Module and libraries were prepared using the Ultra II Directional RNA Library Prep Kit (NEB #E7760)
- Libraries were sequenced on a NextSeq™ 500 (2x70) and downsampled to a minimum of 1M reads
- Sequencing reads were aligned to GRCh38, duplication rates were computed using Picard MarkDuplicates or fgbio AnnotateBamWithUmis[2,3]

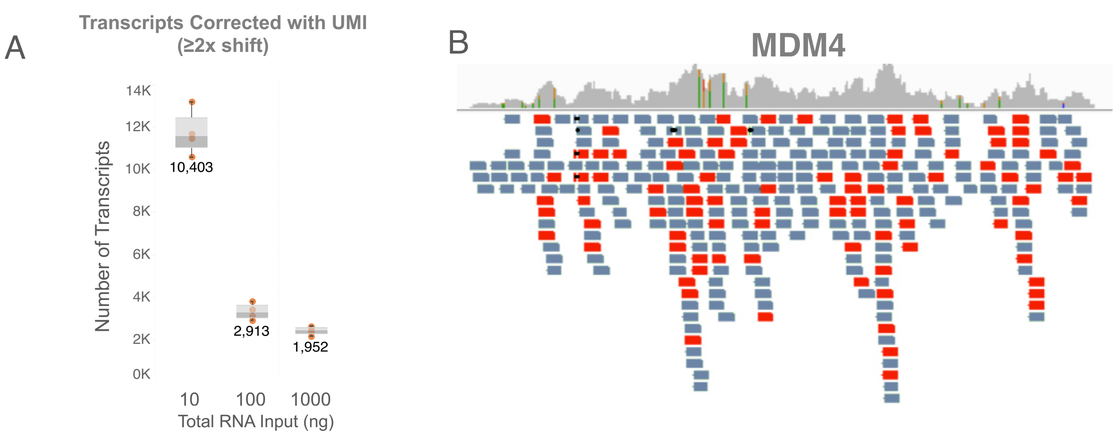
Results: RNA-Seq with UMI Adaptors

96 NEBNext Unique Dual Index UMI Adaptors perform consistently



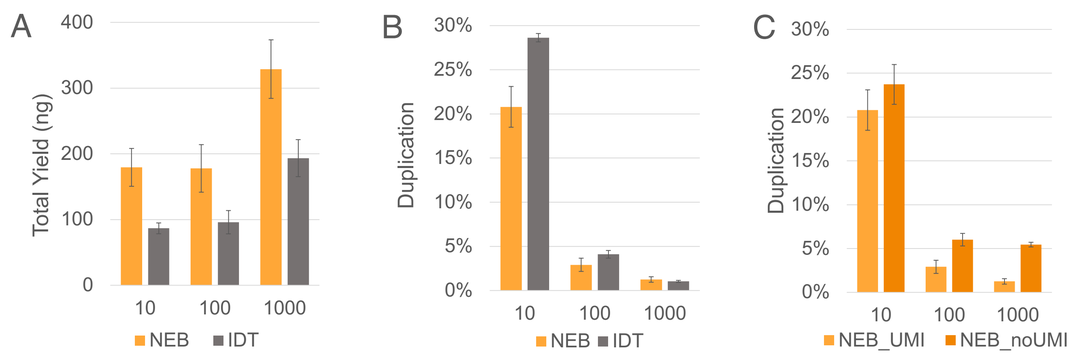
NEBNext Unique Dual Index RNA UMI Adaptors have equal ligation and subsequent amplification efficiency. (A) Library yields were quantified by Agilent TapeStation 4200 and normalized. Adaptor ligation efficiency was robust with uniformity across all 96 unique dual index UMI adaptors. Each bar represents 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.

Removing duplicate reads 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. Libraries prepared with a 10 ng input 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 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.

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 NextSeq 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).

Conclusions

NEBNext Unique Dual Index UMI Adaptors enable higher ligation efficiency and superior uniformity in library generation

Incorporating our NEBNext Unique Dual Index UMI adaptors into DNA sequencing results in:

- PCR-free library prep
- Improved detection of low frequency variants
- Error correction through consensus sequence building
- More accurate removal of duplicate reads

Incorporating our NEBNext Unique Dual Index UMI adaptors into RNA-sequencing results in:

- Robust library yields and high-quality sequencing metrics
- No introduction of sequencing bias with UMIs
- More accurate assessment of duplicate reads increases the number of reads that can be used for downstream analysis
- Improved quantification of transcript abundance

References:
[1]https://www.thermofisher.com/order/catalog/product/969056?us&en#/969056?us&en
[2]Broad Institute (2015) Picard tools (http://broadinstitute.github.io/picard/)
[3]https://github.com/fulcrumgenomics/fgbio

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