Increasing Power to Detect Low-frequency Variants Using Dual-Unique Molecular Indices

Chen Song, Jian Sun, Bradley W. Langhorst, Elizabeth A. Young, Pingfang Liu, and Eileen T. Dimalanta
New England Biolabs, Inc.

INTRODUCTION

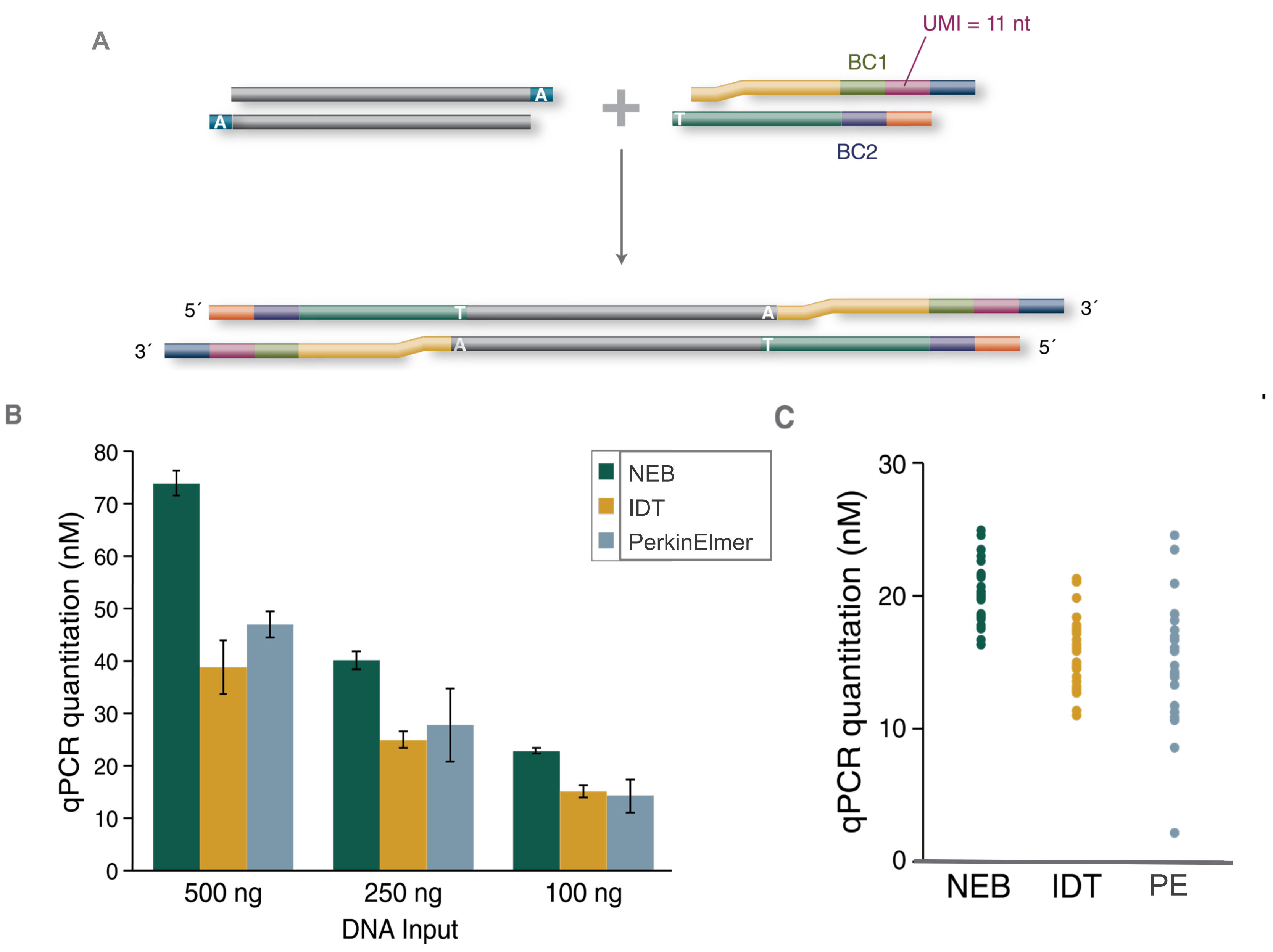
The use of Unique Molecular Identifiers (UMIs) has become increasingly popular – offering a multitude of advantages – particularly when paired with Unique Dual Indices (UDI). Two major factors affecting sequencing accuracy are 1) duplication, arising from PCR amplification of library molecules, and 2) errors introduced during library preparation and sequencing on the flow cell. UMIs, when incorporated into library preparation, can account for and mitigate the impacts of both of these factors.

We assessed the effect of inserting UMIs into UDI adaptors on the accuracy of low-frequency variant detection through duplicate removal and error correction. Using DNA with known allele frequencies, we mixed AcroMetrix™ Oncology Hotspot Control DNA (>500 mutations with 5-35% allele frequency) with NA19240 genomic DNA, at various ratios. Libraries were constructed using the NEBNext® Ultra™ II DNA Library Prep Kit with NEBNext Unique Dual Index UMI Adaptors DNA Set 1, and multiplex hybrid capture was performed on all samples using a customized panel of 152 genes from Twist Bioscience®. Libraries were sequenced on a NovaSeq® 6000 to an average of 2000x coverage after deduplication.

We compared our power to assess low-frequency variants using tumor/normal somatic variant calling, with and without use of UMI information, and observed a 25-200% increase in sensitivity to variants present at less than 1% frequency. In addition to use of UMIs, UDIs are important when using the patterned flow cells found on the NovaSeq to allow for *in-silico* filtering of mixed library clusters that could be interpreted as somatic mutations.

METHODS

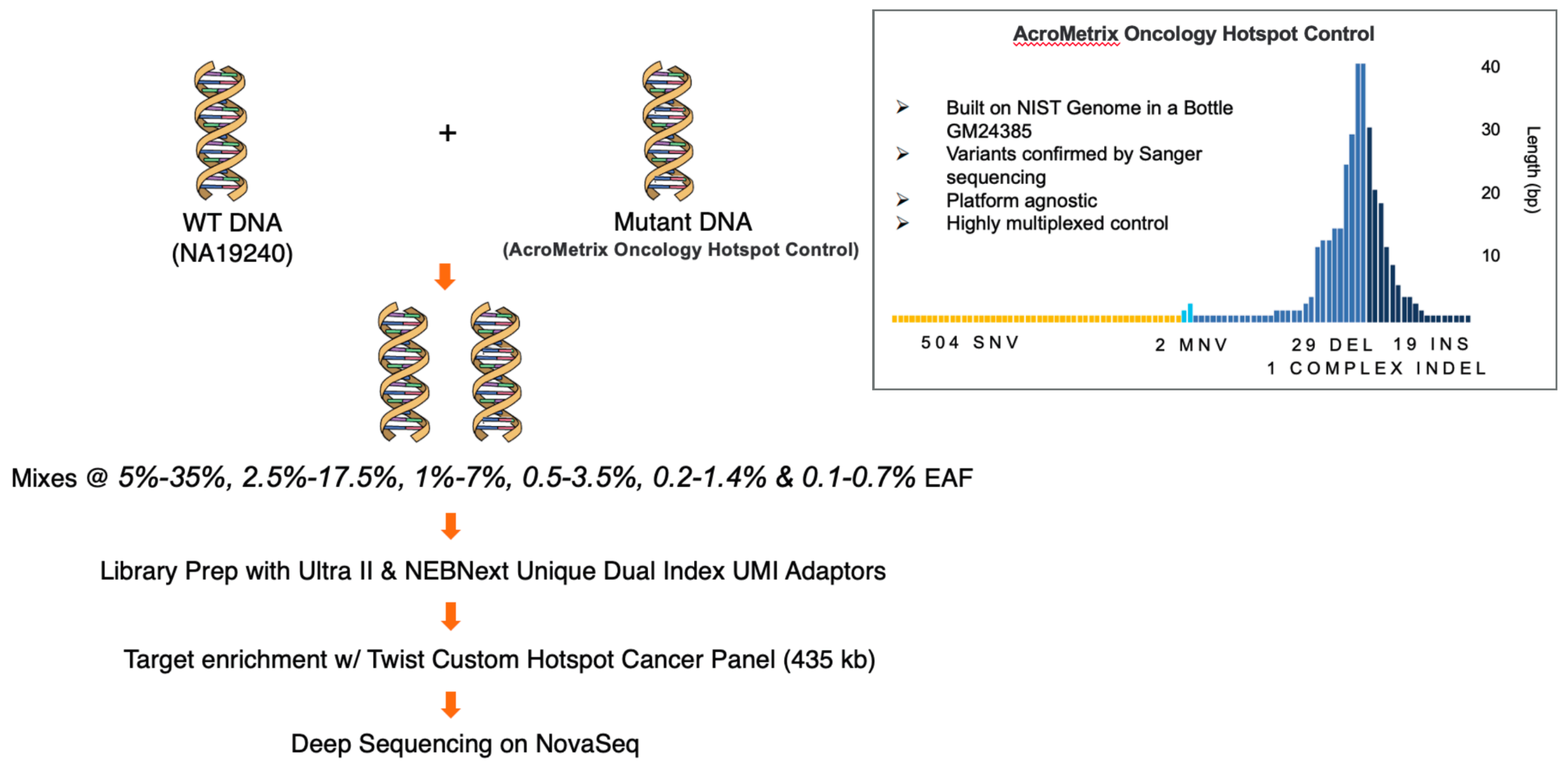
I. Library preparation using unique dual index UMI adaptors



NEBNext Unique Dual Index UMI Adaptors allow higher ligation efficiencies. A) The NEBNext Unique Dual Index UMI Adaptors include a pair of unique dual indices and a single unique molecular identifier. B) Libraries were prepared with 100, 250, and 500 ng inputs of human cell line NA19240 genomic DNA (Coriell Institute for Biomedical Research) and unique dual index adaptors from the suppliers shown with the NEBNext Ultra II FS DNA Library Prep Kit, without PCR amplification. After ligation, two rounds of SPRIselect clean-up steps were performed, and libraries were quantified using the NEBNext Library Quant Kit. C) 90 Libraries were prepared in a 96 well plate with 100 ng NA19240 genomic DNA using the Ultra II FS DNA Library Prep Kit, 30 different adaptors each from the suppliers shown and without PCR amplification. Two bead clean-up steps were performed, followed by qPCR quantification.

II. Low-frequency variant detection workflow

AcroMetrix Oncology Hotspot Control DNA (Thermo Scientific® #969056) was used as the mutated DNA source (>500 mutations with 5-35% allele frequency) and mixed with NA19240 genomic DNA at various ratios to generate a serial range of allele frequencies. Libraries were constructed with the NEBNext Ultra II DNA Library Prep Kit (NEB #E7645) with NEBNext Unique Dual Index UMI Adaptors (NEB #E7395), and multiplex hybrid capture was performed on all samples using a customized panel of 152 genes from Twist Bioscience. Libraries were sequenced on the Illumina NovaSeq, reads were downsampled to 110 million and mapped to hg38 with BWA MEM (0.7.17). Mapped reads were analyzed by MarkDuplicates (Picard 2.20.6) either without utilizing UMI sequences or by building UMI consensus sequences (Fgbio 0.8.1). The final BAM files were used to call somatic variants with Strelka2 (2.9.10).



WT DNA (NA19240)

Mutant DNA (AcroMetrix Oncology Hotspot Control)

AcroMetrix Oncology Hotspot Control

- Built on NIST Genome in a Bottle GM24385
- Variants confirmed by Sanger sequencing
- Platform agnostic
- Highly multiplexed control

Mixes @ *5%-35%, 2.5%-17.5%, 1%-7%, 0.5-3.5%, 0.2-1.4% & 0.1-0.7%* EAF

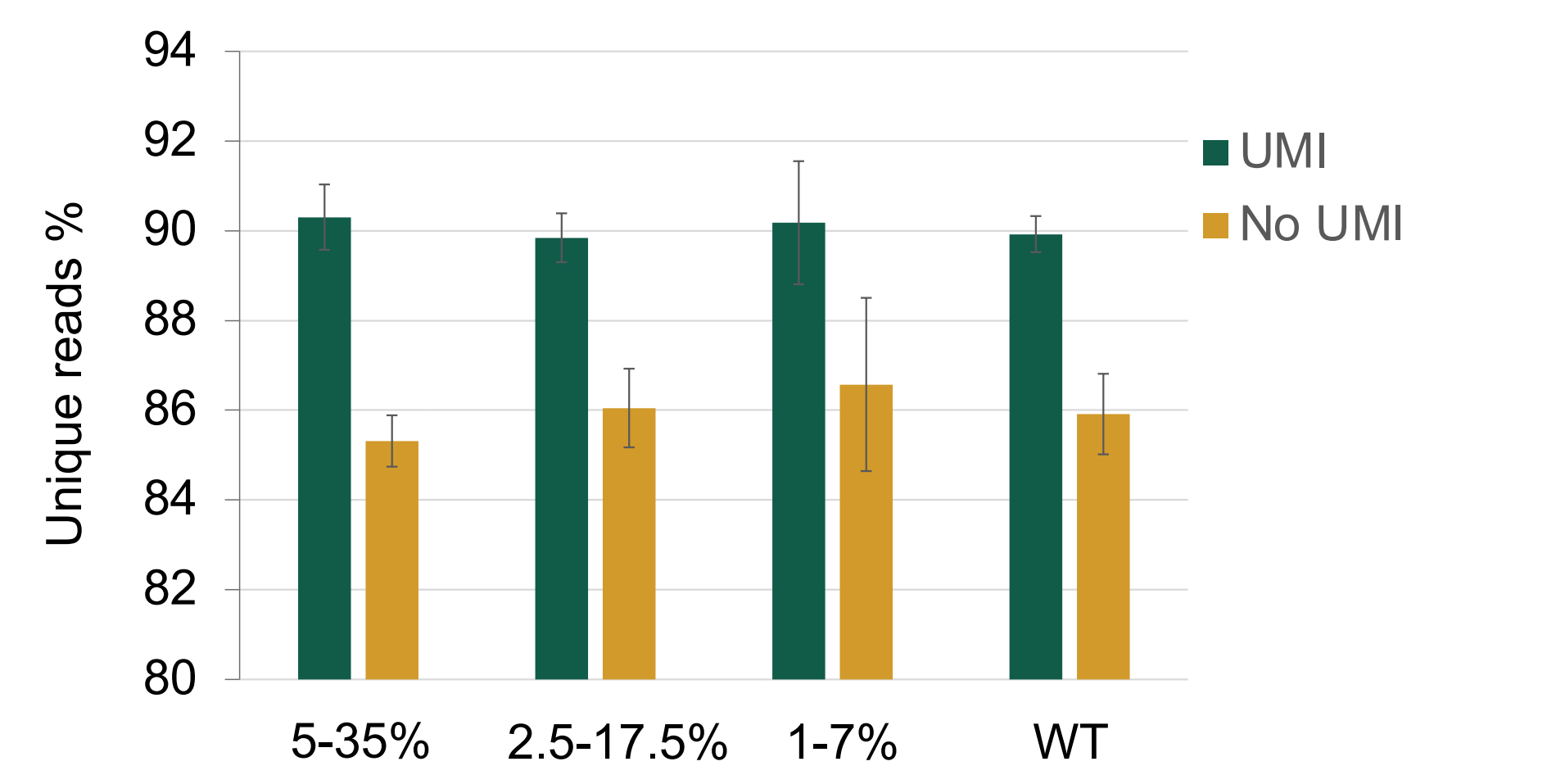
Library Prep with Ultra II & NEBNext Unique Dual Index UMI Adaptors

Target enrichment w/ Twist Custom Hotspot Cancer Panel (435 kb)

Deep Sequencing on NovaSeq

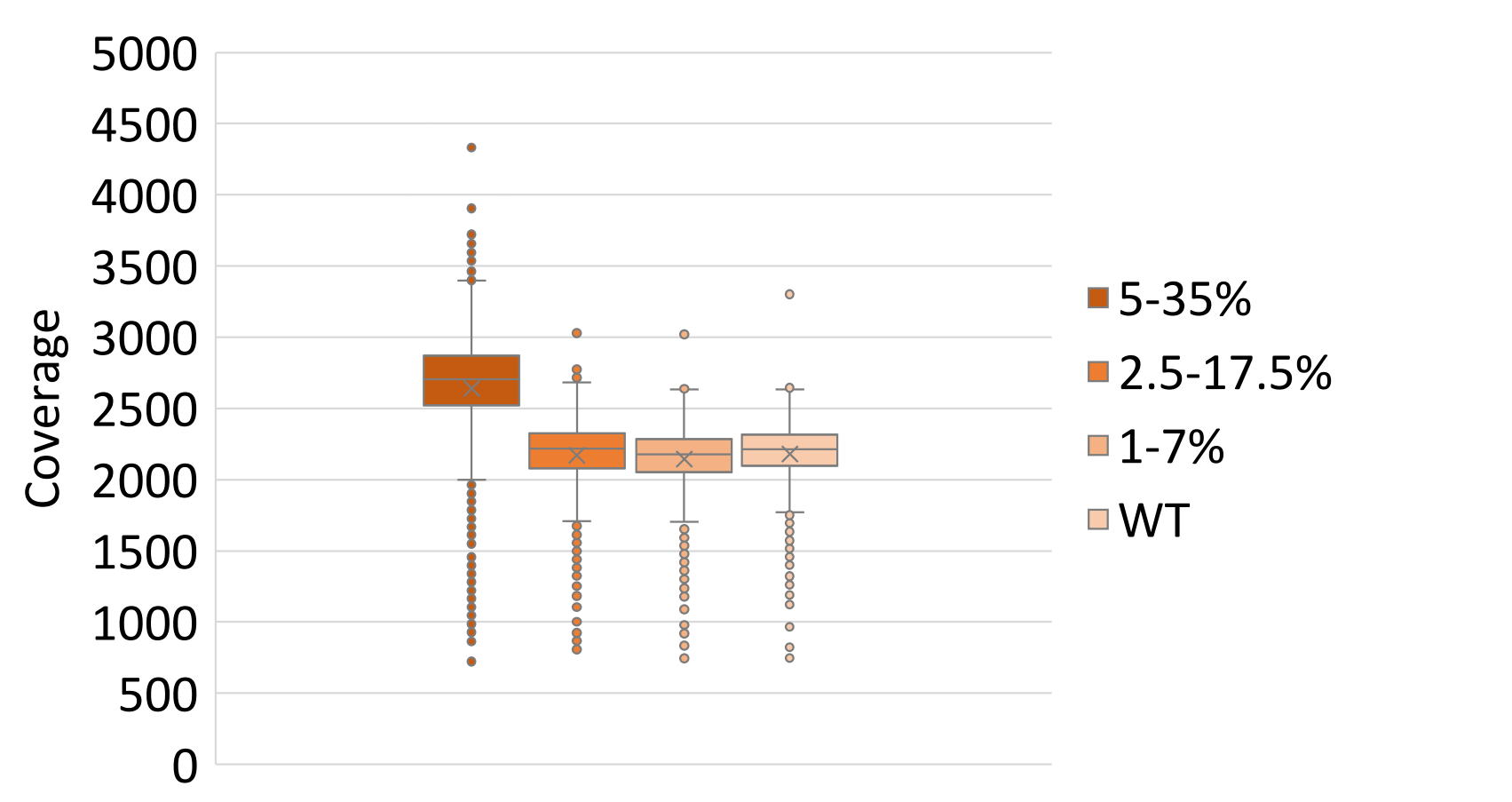
RESULTS

I. UMIs improve duplicate-rate calculation accuracy



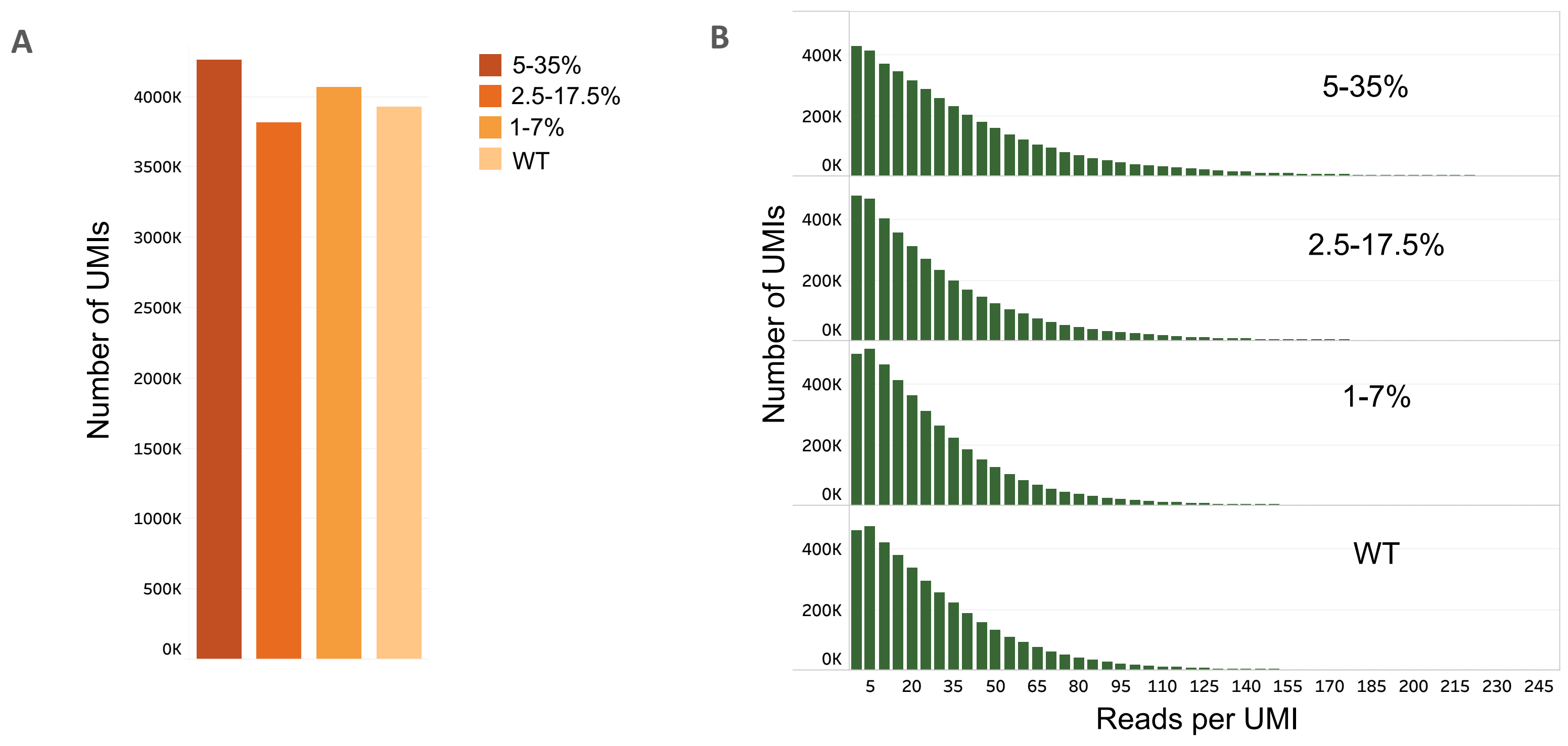
Each library had 300 million paired-end mapped reads from the NovaSeq 6000 output. Deduplication analysis using UMIs and without using UMIs was processed to compare the deduplicated unique reads. The unique reads percentages using UMIs are ~5% higher than without using UMIs, resulting in 10 million more usable reads for each library.

II. Coverage distribution



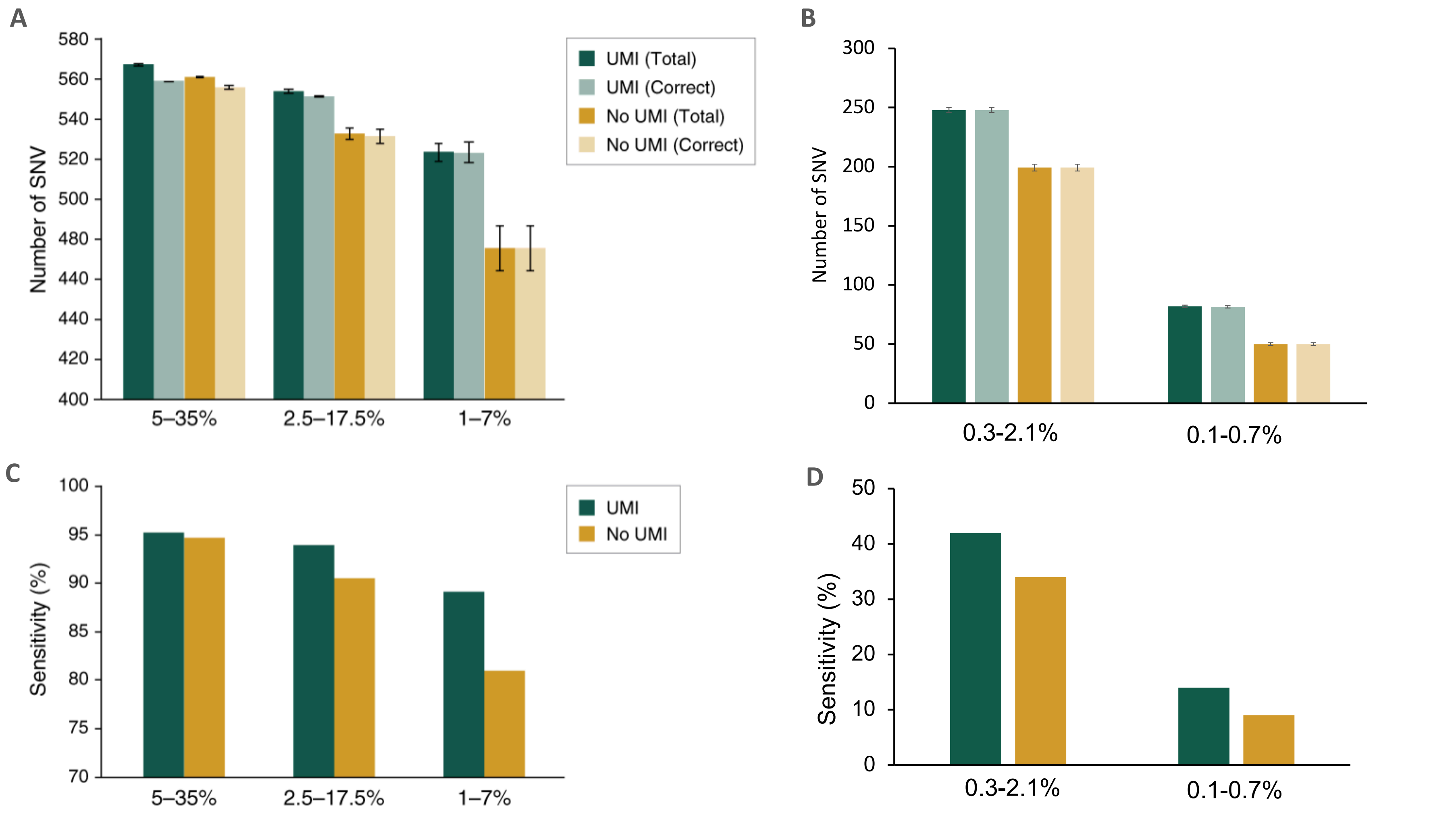
Coverage distribution was plotted using read depths from 2,500 regions with average sequence lengths of 168bp. Most of the regions had more than 2,000 coverage.

III. UMI number and distribution



(A) Number of unique UMIs in each library. (B) Distribution of reads per UMI. All the libraries have consistent UMI number and distribution.

IV. Unique Dual Index UMI Adaptors allow more sensitive low-frequency variant detection



The number of total and correct SNV calls increased when using UMIs for duplicate removal and consensus sequence-based error correction for (A) ≥ 1% variants and (B) <1% variants. The sensitivity of variant detection was improved with UMI consensus calling for (C) ≥ 1% variants and (D) <1% variants. The lower the allele frequencies, the more benefit provided by UMIs in SNV detection. Total is the total number of SNV calls using Strelka2 to call somatic variants against wild type library. Correct is the SNV calls in Total that are contained in the >500 mutations from AcroMetrix DNA.

CONCLUSIONS

- **NEBNext Unique Dual Index UMI Adaptors allow higher ligation efficiencies and enable the preparation of higher quality libraries for target enrichment and low-frequency variant detection.**
- **UMI utilization improves duplication calculation accuracy resulting in a higher number of usable reads.**
- **UMI-included error correction improves sensitivity of low-frequency variant detection.**