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

Genomics

Keerthana Krishnan (krishnan@neb.com), New England Biolabs, **Pingfang Liu**, New England Biolabs, **Chen Song**, New England Biolabs, **Dora Posfai**, New England Biolabs, **Karen McKay**, New England Biolabs, **Jian Sun**, NEB, **Gautam Naishadham**, New England Biolabs, **Bradley Langhorst**, New England Biolabs, **Eileen Dimalanta**, New England Biolabs, **Theodore Davis**, New England Biolabs

The use of Unique Molecular Identifiers (UMIs) have become increasingly popular and offer a multitude of advantages especially when paired with unique dual sample indexing (UDI). Two major factors affecting sequencing accuracy are 1) PCR duplication arising from amplification of library molecules and 2) errors introduced during library preparation and actual sequencing on the flow cell. We can account for these factors by incorporating UMIs into library preparation.

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. We analyzed the utility of UMIs for duplicate removal and error correction in low frequency variant detection in genomic sequencing and show that the sensitivity of variant detection is improved with UMI consensus calling. We demonstrate that combining unique dual sample indexing with UMI molecular barcoding further improves data analysis accuracy, especially on patterned flow cells.

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. Ligation of barcoded adaptors followed by PCR enrichment produced high-quality library and sequencing metrics. 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 were determined to in fact originate from unique molecules, increasing the total number of reads available for further analysis.

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.