Streamlining RNA and DNA library construction methods to meet the challenges of high throughput sequencing.

Karen R McKay, Chen Song, Chelsea Pinegar, Kaylinnette Pinet, Gautam Naishadham, Matthew Campbell, Dora Posfai, Bradley W. Langhorst, V K Chaithanya Ponnaluri, Pingfang Liu, and Keerthana Krishnan

New England Biolabs, Inc.

Introduction

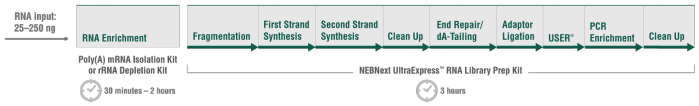
In response to user needs for faster, streamlined, and automatable library preparation protocols, we have developed three new NEBNext UltraExpress® workflows. Our development goal was to meet the criteria for processing samples of various input amounts, accommodating a range of sample types, streamlining the workflows to allow for automation-friendly protocols, and enabling scalability. Optimizations made throughout development included testing samples of various inputs with a single adaptor concentration and a single condition for PCR cycling, while maintaining high-quality sequencing metrics. The optimized and streamlined RNA and DNA workflows resulted in a single-tube solution incorporating master mixed reagents, reduced incubation times, and fewer clean-up steps, which in turn generates less plastic consumable waste.

Here, we present data obtained with the NEBNext UltraExpress® RNA, UltraExpress® DNA (mechanically sheared), and UltraExpress® FS DNA (enzymatically sheared) workflows. The RNA workflow is compatible with poly(A) enrichment or ribosomal RNA depletion and permits a single-day process for efficiently generating RNA-seq libraries from 25-250 ng of total RNA with a single condition. This offers a significant advantage over the standard RNA-seq workflow, enabling a much faster process with less hands-on time, minimal sample transfer, and fewer clean-up steps while maintaining high library yield and sequencing quality. The DNA workflows, with mechanical or enzymatic shearing, allow DNA samples of varied sources, including human genomic and microbial DNA from 10 ng to 200 ng, to be processed in a single tube without normalization or adjusting reaction conditions. Under all conditions, high library yield, minimal adaptor dimer, and high sequencing quality were achieved.

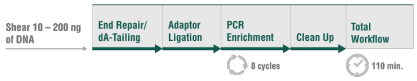
In conclusion, the new streamlined workflows serve as an optimal solution to meet high-throughput and automation demands. We expect these simplified library preparation methods will serve in a myriad of NGS applications, including clinical and academic settings.

Methods

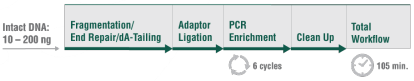
NEBNext UltraExpress RNA library preparation



NEBNext UltraExpress DNA library preparation

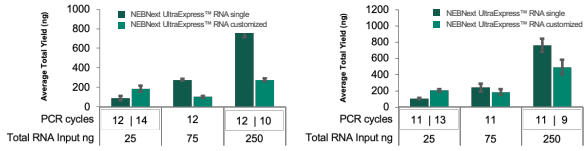


NEBNext UltraExpress FS DNA library preparation

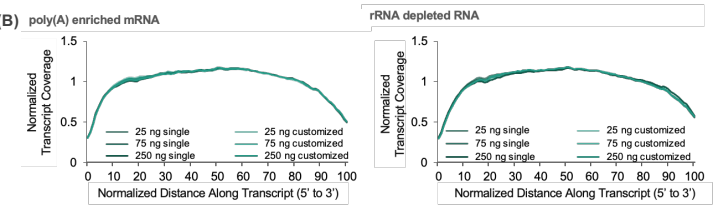


Results

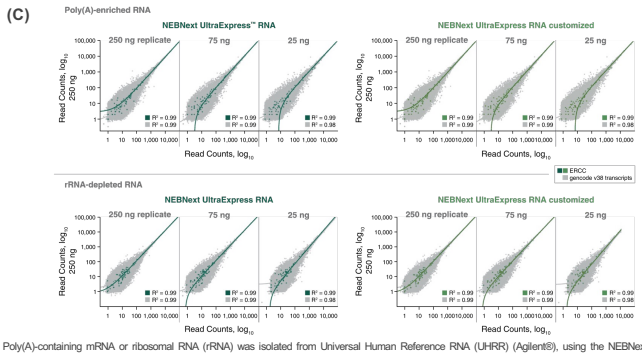
High library yields across a range of RNA inputs



Excellent transcript correlation between RNA inputs and replicates



Consistent transcript coverage with NEBNext UltraExpress RNA Library Prep

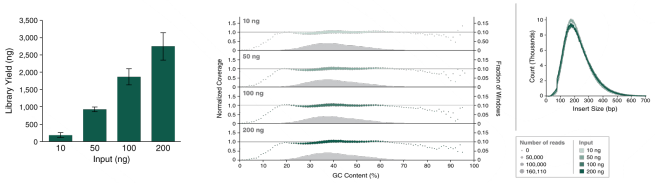


Poly(A)-containing mRNA or ribosomal RNA (rRNA) was isolated from Universal Human Reference RNA (UHRR) (Agilent®), using the NEBNext® Poly(A) mRNA Magnetic Isolation Module (NEB #E7490) or NEBNext rRNA Depletion Kit v2 (Human/Mouse/Rat – NEB #E7400). Libraries were prepared using the NEBNext UltraExpress RNA Library Prep with a single adaptor dilution (50X) and 12 PCR cycles for all inputs, or with customized adaptor dilutions and PCR cycle numbers (20X adaptor dilution with 10 PCR cycles for 250 ng input, 100X adaptor dilution with 12 PCR cycles (75 ng) and 14 PCR cycles (25 ng)). Libraries were sequenced on an Illumina NextSeq® 500 using paired-end mode (2x75 bp) and down sampled to 10M reads for analysis. **(A)** The total RNA input amount and number of PCR cycles are indicated. Library yields calculated from an average of three replicates are shown with error bar indicating the standard deviation between replicates. **(B)** Libraries were sampled and mapped to the hg38 reference genome using RNA STAR v2.7.8a and 5' to 3' transcript coverage was calculated from the top 1000 transcripts using the CollectRnaSeqMetrics (Picard) tool v2.18.2.2. **(C)** Libraries were correlated for transcript expression levels across inputs using Salmon v1.5.1 quantification of all gencode v38 transcripts and ERCCs. Each data point represents a transcript, with the log10 abundance in Number of Reads with 250 ng total RNA input on the y-axis compared to a replicate at 250 ng followed by 75 and 25 ng on the x-axis.

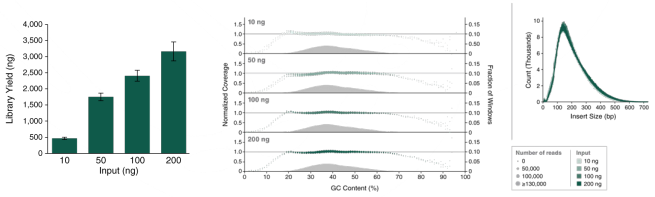
Results continued

Uniform GC coverage and insert size from a range of DNA input amounts

(A) NEBNext UltraExpress DNA



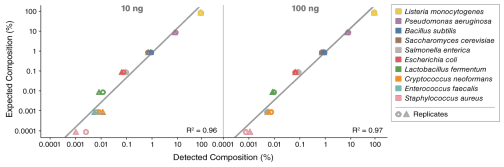
(B) NEBNext UltraExpress FS DNA



(A) Libraries were prepared from 10, 50, 100 and 200 ng of Human NA19240 genomic DNA (Coriell Institute for Medical Research) using the same adaptor amount and 8 PCR cycles. Libraries were pooled and sequenced on an Illumina® MiSeq® (2 x 75 bases). Library yields reflect input amounts given the single PCR cycle condition applied to all samples. Sequencing data showed consistent GC coverage and insert size. 2 million paired-end reads from each library were sampled (seqtk v1.3), adapter-trimmed (seqprep v0.1) and mapped to the GRCh38 reference (bowtie2 v2.4.5), and GC coverage information was calculated using Picard's CollectGCBiasMetrics (v 1.56.0). **(B)** Libraries were prepared from 10, 50, 100 and 200 ng of a 9:1 Human NA19240 genomic DNA (Coriell Institute for Medical Research) and Escherichia coli gDNA (Lofstrand Labs Limited) mixed sample, using the same adaptor amount and 6 PCR cycles. Library yields reflect input amounts given the single PCR cycle condition applied to all samples. Libraries were pooled and sequenced on an Illumina® NextSeq® 500/550 (2 x 75 bases). Sequencing data showed consistent GC coverage and insert size. 2 million paired-end reads from each library were sampled (seqtk v1.0), adapter-trimmed (seqprep v0.1) and mapped to a composite reference containing GRCh38 and E. coli MG1655 contigs (bowtie2 v2.5.0). GC coverage and insert size distributions were calculated using Picard's CollectGCBiasMetrics and Picard CollectInsertSizeMetrics (v1.56.0); Picard CollectGCBiasMetrics. (1.56) was run on human autosomes only due to the even copy number assumption of the tool.

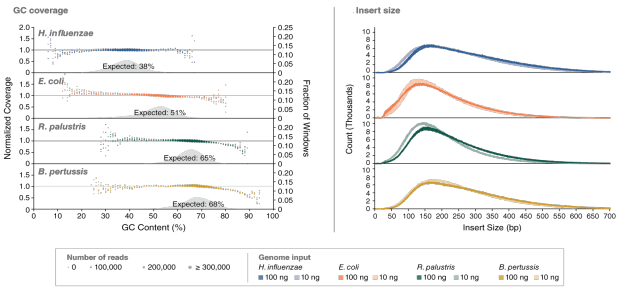
Accurate representation of diverse microbial DNA input types

(C) NEBNext UltraExpress DNA



(C) The NEBNext UltraExpress DNA Library Prep Kit generates libraries representative of input DNA. Libraries were prepared using a single protocol from 10 ng and 100 ng of the ZymoBIOMICS Microbial Community DNA Standard II (Log Distribution) (Zymo Research®, Catalog # D6311). Libraries were pooled and sequenced on an Illumina® MiSeq® (2 x 75 bases). 1.4 million reads from each library were sampled (seqtk v1.3), adapter-trimmed (seqprep v0.1), and aligned to a composite reference genome (bowtie2 v2.4.5). 1,000 bp windows of constituent genomes were counted (bedtools 2.30.0) and compared across expected and detected composition for both input levels. Consistent correlation between expected and detected composition was noted (R^2 = 0.96 for 10 ng and R^2 = 0.97 for 100 ng). Expected composition: *Listeria monocytogenes* 89.1%, *Pseudomonas aeruginosa* 8.9%, *Bacillus subtilis* 0.89%, *Saccharomyces cerevisiae* 0.89%, *Escherichia coli* 0.089%, *Salmonella enterica* 0.089%, *Lactobacillus fermentum* 0.0089%, *Enterococcus faecalis* 0.00089%, *Cryptococcus neoformans* 0.00089% and *Staphylococcus aureus* 0.000089%

(D) NEBNext UltraExpress FS DNA



(D) The NEBNext UltraExpress FS DNA Library Prep Kit produces representative GC coverage and insert size peaks for microbial genomic DNA over a broad range of GC composition. Libraries were prepared using the NEBNext UltraExpress FS DNA protocol for 10 ng and 100 ng of genomic DNA from *Haemophilus influenzae*, *Escherichia coli*, *Rhodopseudomonas palustris* and *Bordetella pertussis*. Data showed (A) representative GC coverage and (B) insert size peaks across samples with genome GC contents of 38%-68% GC. Libraries were pooled and sequenced on an Illumina NextSeq 500/550 (2 x 75 bases). 2 million paired-end reads from each library were sampled (seqtk v1.0), adaptor-trimmed (seqprep v0.1), and aligned to their respective reference genomes (bowtie2 v2.5.0). GC coverage and insert size distributions were calculated using Picard's CollectGCBiasMetrics and Picard CollectInsertSizeMetrics (v1.56.0). The horizontal grey line indicates the expected normalized coverage of 1.0, and the colored dots represent read numbers at each GC%. The grey area plot is a histogram representing the distribution of GC content in 100 bp windows of the reference genome for the 10 and 100 ng inputs.

Conclusions

- Shorter hands-on time
- Fewer steps and consumables
- Flexibility with sample types and protocols
- Automation friendly volumes

NEBNext UltraExpress RNA:

- Total RNA to library in 4-5 hours
- Ability to use a single condition for all inputs
- Compatibility with multiple RNA sample types
- Reaction volumes designed with automation in mind

NEBNext UltraExpress DNA and FS:

- Sample-in library-out < 2 hours
- Single-tube library prep to reduce consumables
- One protocol fits wide input range and diverse sample types
- Automation friendly pipetting/reaction volumes

Authors would like to acknowledge the technical assistance provided by Dora Posfai, Kristen Augulewicz, Harry Bell, and Rebecca Gawron at the New England Biolabs' Sequencing Core Facility.