Improving Transcriptome Profiling for Single Cell and Low Input RNA

Keerthana Krishnan[1], Yanxia Bei[1], Janine G. Borgaro[1], Shengxi Guan[1], Vaishnavi Panchapakesa[1], Karen Duggan[1], Lynne Apone[1], Timur Shtatland[1], Bradley W. Langhorst[1], Melissa Arn[1], Jonathan Sanford[1], Christine Sumner[1], Diwakar R Pattabiraman[2], Thomas C. Evans, Jr.[1], Eileen Dimalanta[1], Nicole M. Nichols[1] and Theodore Davis[1]

[1]New England Biolabs, Ipswich, MA 01938
[2]Norris Cotton Cancer Center, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756

INTRODUCTION

RNA sequencing has been widely used to determine gene expression profiles of diverse tissues, cell types, developmental stages and diseases. Most of these studies are based on population analyses using thousands of cells. Such studies, however, disguise the potentially significant biological variations among individual cells. To overcome this limitation, single-cell RNA-seq is emerging as a powerful approach to characterize gene expression heterogeneity within phenotypically identical or complex cell populations and in rare cell types.

We developed a simple and robust single-cell, low input RNA-seq workflow to generate full-length cDNAs that can easily be converted into sequencing-ready Illumina libraries when combined with the NEBNext® Ultra™ II FS DNA Library Preparation Kit which utilizes enzymatic fragmentation. Using this approach, we generated libraries from a variety of input material including Universal Human Reference (UHR) RNA (2 pg – 200 ng), single cells from cultured cell lines and mouse primary cells, and sequenced on the Illumina® NextSeq® 500.

High quality sequencing data was obtained from all samples. We observe excellent gene body coverage and high sensitivity as demonstrated by detection of a high number of transcripts and expected number of RNA spike-ins (ERCC), even at single copy numbers. The data showed strong gene expression correlation (Pearson r>0.9) between RNA inputs that span over five orders of magnitude and in cultured cells (single vs. hundreds). From the analysis of the primary cells, we could successfully distinguish two types of cells from 8-week old mouse mammary glands and were able to trace them back to the basal and luminal developmental lineages, highlighting the high sensitivity of the protocol.

We have developed a highly robust and sensitive method that consistently generates high quality sequencing data from single cell or low input RNA. It is streamlined and amenable to large-scale and high throughput automation. We envision this method facilitating novel discoveries in the area of low-input transcriptome applications and across various platforms.
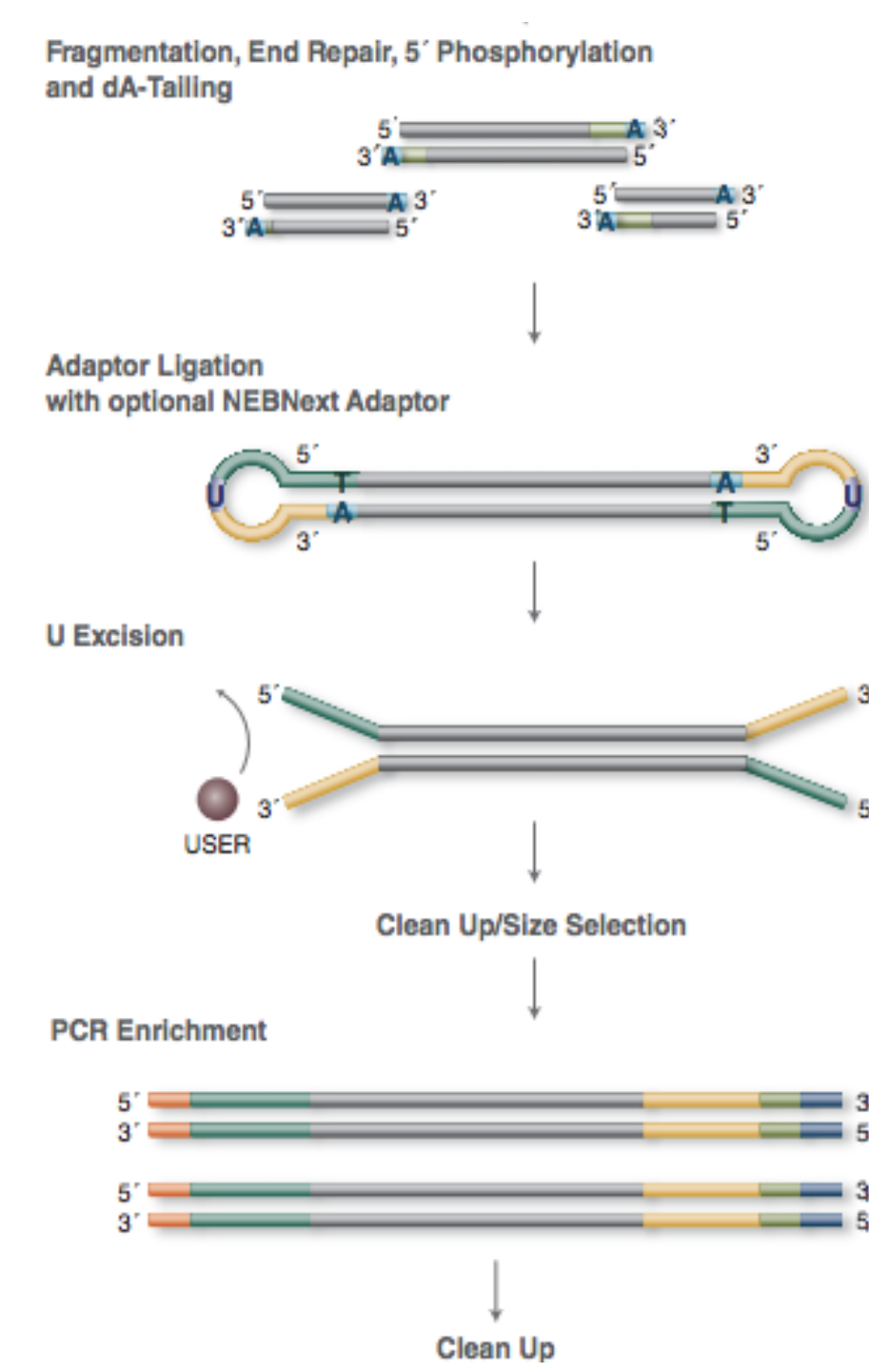
METHODS

A STEP I: cDNA Synthesis & Amplification
B STEP II: Library Generation
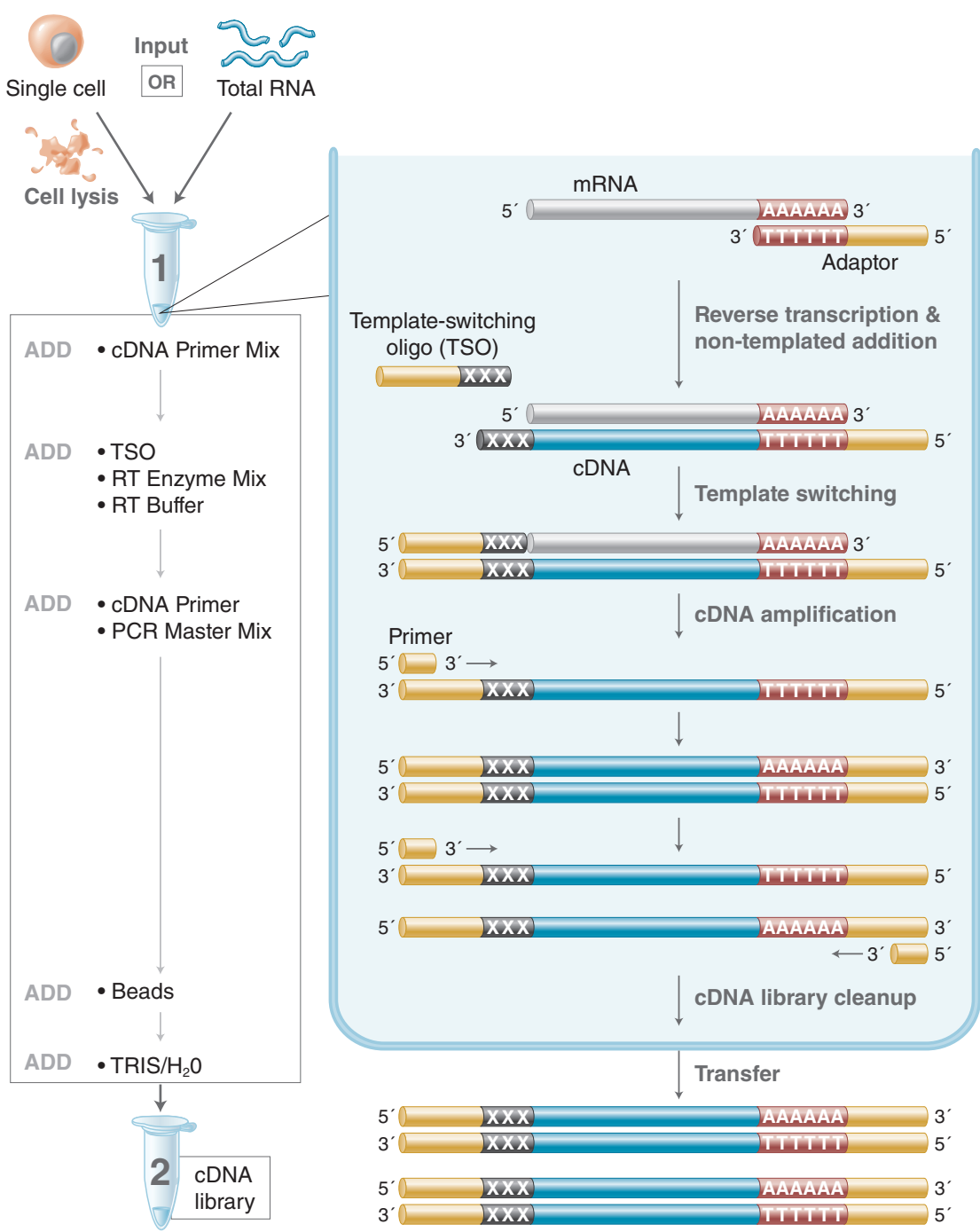
(A) cDNA libraries are prepared from intact cells or total RNA in a single-tube reaction. cDNA is synthesized by template-switching-mediated reverse transcription followed by amplification by PCR; (B) The full length cDNA is enzymatically fragmented, end repaired, dA-tailed, adaptors ligated and PCR amplified to generate final libraries to be sequenced on Illumina platforms; (C) Flowchart illustrating a streamlined Single Cell/Low Input RNA library prep workflow incorporating cDNA synthesis and library preparation with hands-on time of ~30 mins.
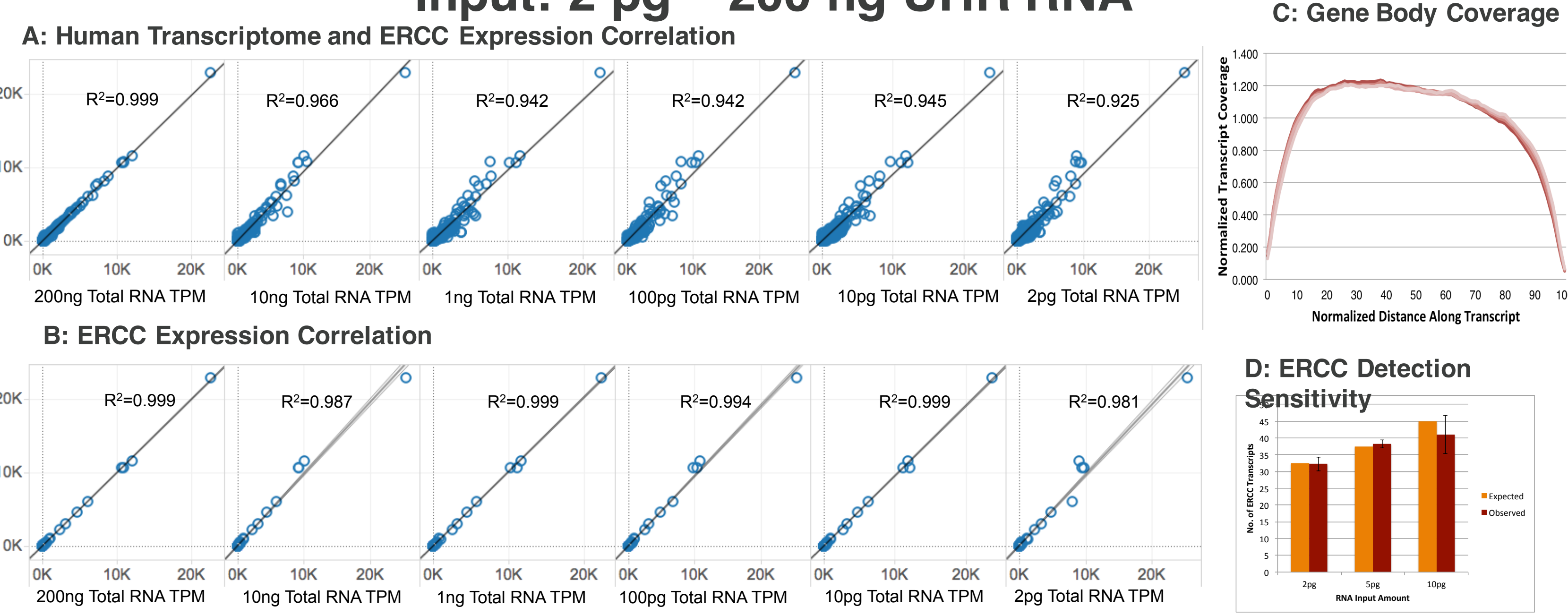
C Overview of Workflow

	NEBNext® Single Cell/Low Input cDNA Synthesis & Amplification Module				NEBNext® Ultra™ II FS DNA Library Prep Kit for Illumina®					
	Cell Lysis	Primer Annealing	Reverse Transcription	cDNA Amplification	Clean Up	Fragmentation, End Repair, & dA-Tailing	Adaptor Ligation	Clean Up	Amplification	Clean Up	Total Workflow
Hands-On	2 min.	1 min.	2 min.	2 min.	9 min.	1 min.	1 min.	5 min.	1 min.	5 min.	~30 min.
Total	7 min.	6 min.	102 min.	45–90 min.	43 min.	10 min.	16 min.	27 min.	27–36 min.	27 min.	5–6 hr.

RESULTS

Sensitive and Consistent Performance across different input amounts:

Input: 2 pg – 200 ng UHR RNA

A: Human Transcriptome and ERCC Expression Correlation
$R^2=0.999$ $R^2=0.966$ $R^2=0.942$ $R^2=0.942$ $R^2=0.945$ $R^2=0.925$

B: ERCC Expression Correlation
$R^2=0.999$ $R^2=0.987$ $R^2=0.999$ $R^2=0.994$ $R^2=0.999$ $R^2=0.981$

C: Gene Body Coverage
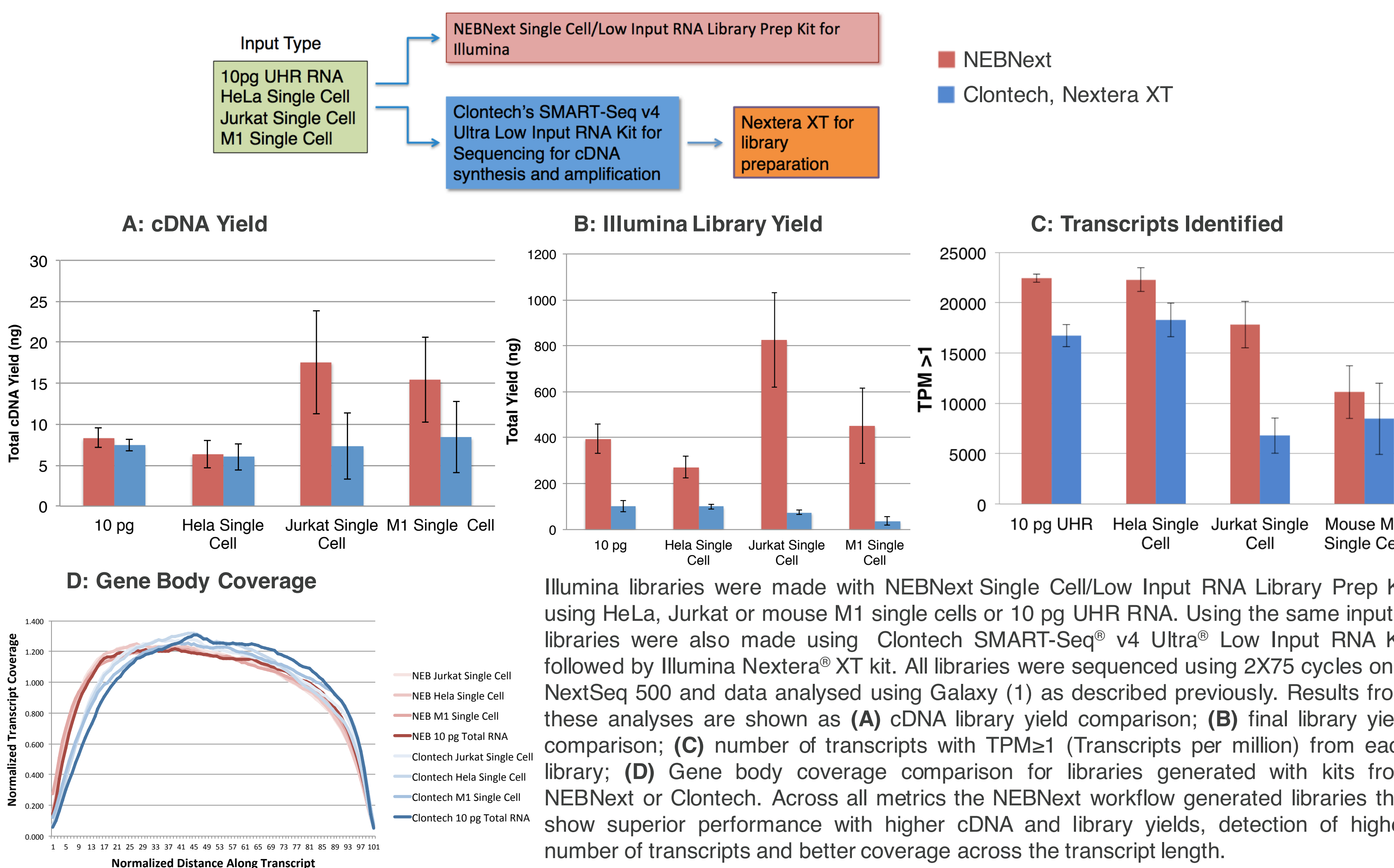D: ERCC Detection Sensitivity

Illumina libraries made from Total UHR RNA in the range of 2 pg to 200 ng were sequenced using 2X75 cycles on a NextSeq 500 and data analysed using Galaxy (1) on the GRCh38 transcriptome reference. Transcript abundance was quantified using Salmon v0.6 (2) on the GRCh38 transcriptome reference. Reads were mapped to hg19 Human Reference Genome by HISAT2 (3) and gene body coverage was calculated using Picard tools (4). Consistent performance was observed across libraries made using various inputs. Results are shown as (A) transcript expression correlation between replicates (200 ng vs. 200 ng) and across different RNA inputs, 200 ng, 10 ng, 1 ng, 100 pg, 10 pg, 2 pg; (B) ERCC transcript expression correlation plots from the same libraries described in A; (C) gene body coverage for libraries made from 2 pg– 200 ng of UHR RNA; (D) "Expected" vs. "Detected" number of ERCC RNA species in low input RNA samples (2 pg -10 pg). Expected: number of ERCC RNA species with at least 1 copy number in the RNA samples; Detected: number of ERCC RNA species with TPM (Transcripts per million) ≥1. Libraries show excellent correlation across all inputs and consistent gene body coverage over the entire transcript length

Input: HeLa Cells and HeLa Total RNA

A: Transcriptome Expression Correlation
$R^2=0.973$ $R^2=0.952$ $R^2=0.917$ $R^2=0.872$ $R^2=0.912$ $R^2=0.916$

B: Gene Body Coverage
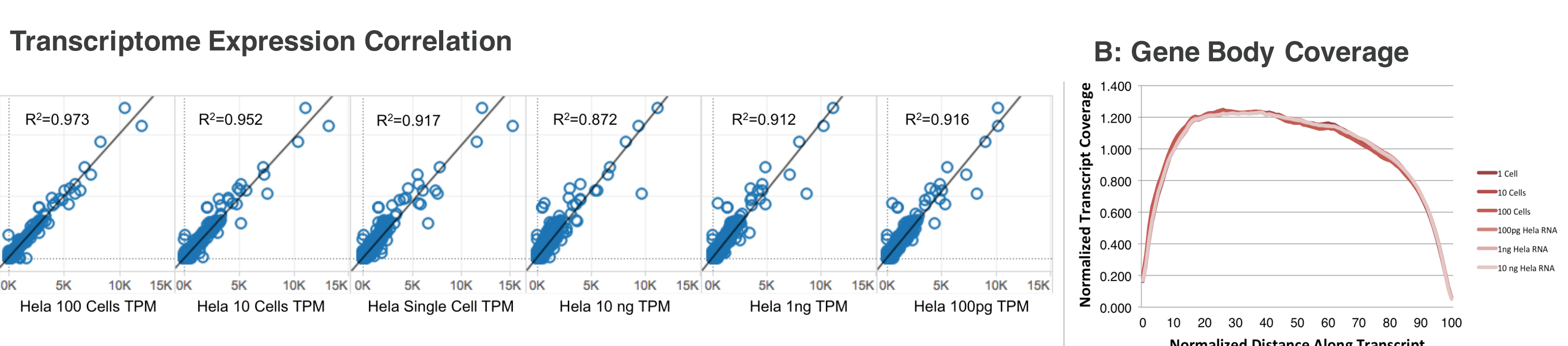
Illumina libraries made from Hela cells (single cell, 10 cells and 100 cells) and Hela Total RNA were sequenced using 2X75 cycles on a NextSeq 500. Results show (A) Consistent correlation of transcripts across 100 HeLa cells, 10 cells, and a single cell. Similar correlation is also seen with 10 ng, 1 ng, 100 pg of Total HeLa RNA; (B) Consistent gene body coverage across different inputs

Robust and Superior Performance across Different Sample Types: Input: 10pg UHR RNA, HeLa, Jurkat, M1 single cell

Input Type: 10pg UHR RNA, HeLa Single Cell, Jurkat Single Cell, M1 Single Cell → NEBNext Single Cell/Low Input RNA Library Prep Kit for Illumina / Clontech's SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing for cDNA synthesis and amplification → Nextera XT for library preparation

- NEBNext
- Clontech, Nextera XT

A: cDNA Yield
B: Illumina Library Yield
C: Transcripts Identified

D: Gene Body Coverage

- NEB Jurkat Single Cell
- NEB Hela Single Cell
- NEB M1 Single Cell
- NEB 10 pg Total RNA
- Clontech Jurkat Single Cell
- Clontech Hela Single Cell
- Clontech M1 Single Cell
- Clontech 10 pg Total RNA

Illumina libraries were made with NEBNext Single Cell/Low Input RNA Library Prep Kit using HeLa, Jurkat or mouse M1 single cells or 10 pg UHR RNA. Using the same inputs, libraries were also made using Clontech SMART-Seq® v4 Ultra® Low Input RNA kit followed by Illumina Nextera® XT kit. All libraries were sequenced using 2X75 cycles on a NextSeq 500 and data analysed using Galaxy (1) as described previously. Results from these analyses are shown as (A) cDNA library yield comparison; (B) final library yield comparison; (C) number of transcripts with TPM≥1 (Transcripts per million) from each library; (D) Gene body coverage comparison for libraries generated with kits from NEBNext and Clontech. Across all metrics the NEBNext workflow generated libraries that show superior performance with higher cDNA and library yields, detection of higher number of transcripts and better coverage across the transcript length.
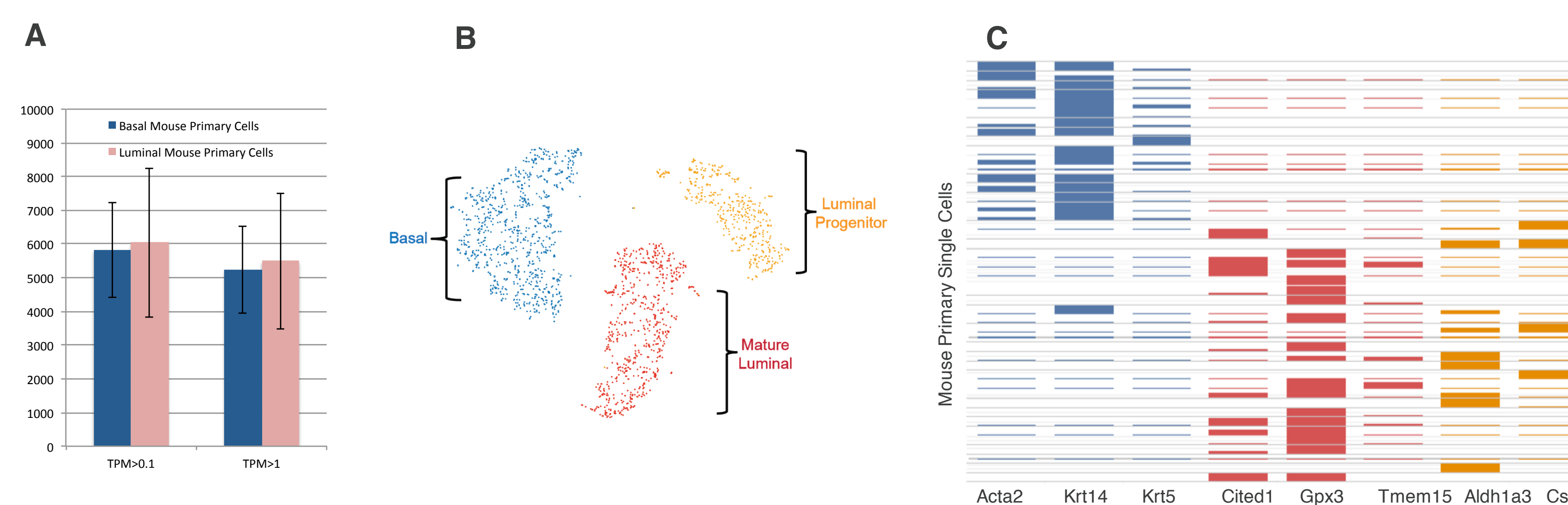
Sensitive and Superior Determination of Gene Expression: Input: Jurkat Cells
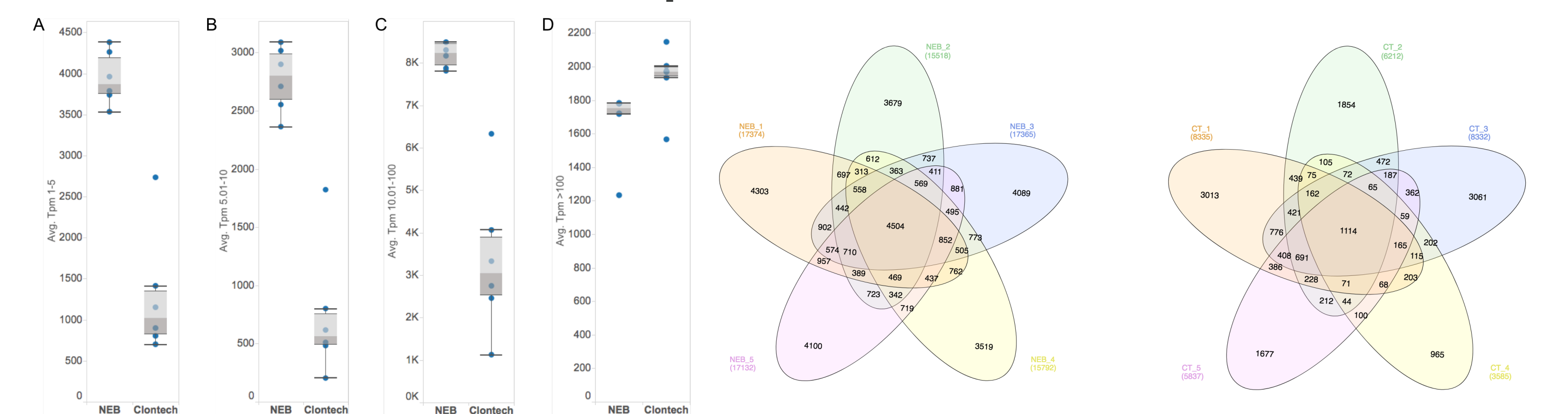
Illumina libraries were made with NEBNext Single Cell/Low Input RNA Library Prep Kit using Jurkat single cells. For comparison, Jurkat single cells were used to make libraries with Clontech SMART-Seq v4 Ultra Low Input RNA Kit followed by Illumina Nextera XT kit. All libraries were sequenced using 2X75 cycles on a NextSeq 500 and data analysed using Galaxy (1) as described previously. Figures A-D show the number of transcripts detected per Jurkat single cell (6 replicates) using different methods (NEBNext vs. Clontech) and different ranges of expression (grouped into 1-5, 5-10, 10-50 and > 50 TPM). TPM = Transcripts per Kilobase Million. The box plot shows the median, first and third quartiles per method, and range of expression. Libraries generated using the NEBNext workflow detect more transcripts, especially low abundance transcripts. (A) Number of transcripts detected within TPM 1-5; (B) Number of transcripts detected within TPM 5-10; (C) number of transcripts detected within TPM 10-50; (D) Number of transcripts detected >50 TPM; For the overlap, 5 replicates of Jurkat single cell libraries were chosen. (E) Overlapping transcripts detected with TPM ≥1 using the NEBNext Single Cell/Low Input RNA Library Prep Kit for Illumina; (F) Overlapping transcripts detected with TPM ≥1 using the SMART-Seq v4 Ultra Low Input RNA Kit followed by the Nextera XT DNA Library Prep Kit. NEBNext libraries consistently detect higher number and more overlapping transcripts across single Jurkat cells.

Sensitive Determination of Gene Expression Signature to Identify Subtypes: Input: Primary Mouse Mammary Epithelial Cells

A
- Basal Mouse Primary Cells
- Luminal Mouse Primary Cells

B
- Basal
- Luminal Progenitor
- Mature Luminal

C
- Mouse Primary Single Cells
- Acta2 Krt14 Krt5 Cited1 Gpx3 Tmem15 Aldh1a3 Csn3

Illumina libraries were generated from Mouse primary single cells (basal or luminal mammary cells) using the NEBNext Single Cell/Low Input RNA Library Prep, sequenced using 2X75 cycles on a NextSeq 500, and data analysed using Galaxy (1) as described previously. 10X Libraries were generated using the Chromium™ Single Cell 3' Reagent Kit and sequenced on an Illumina HiSeq® 2500. Data was aligned using Bowtie2 (5) and analysed on Loupe Cell Browser. (A) number of transcripts with TPM≥0.1 and TPM≥1 from each subtype of primary cells identified using the NEBNext workflow; (B) Clustering of >1800 cells using a 10X Chromium Single cell 3' Solution identifies the basal subtype, luminal progenitor or mature luminal subpopulation based on marker gene expression; (C) Similar signature of marker genes can be identified in single basal primary cells (17 cells) and two subpopulations of luminal progenitor or mature luminal can be identified in the single luminal primary cells (30 cells). NEBNext libraries detect expression of gene markers to identify subtypes of mouse mammary epithelial cells.

CONCLUSIONS

- The NEBNext Single Cell/Low Input RNA-seq workflow provides a streamlined and easy to use solution for NGS library preparation from single cells or a wide range of total RNA inputs from 2 pg up to 200 ng
- Minimal hands-on time and few handling steps to reduce errors, leading to high consistency and reproducibility
- High yields of cDNA and libraries are generated with uniform 5'-3' transcript coverage
- High transcript expression correlation between high and low input libraries is observed
- Sensitive detection of low abundance transcripts and consistent detection of transcripts across all inputs and different cell types

REFERENCES AND ACKNOWLEDGEMENT

1. https://academic.oup.com/nar/article/44/W1/W3/2499339
2. Patro et al., (2015). Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment. bioRxiv, 21592.
3. Kim D, Langmead B, and Salzberg SL (2015). HISAT: a fast spliced aligner with low memory requirements, Nature methods.
4. http://broadinstitute.github.io/picard
5. http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml