An Improved cDNA Library Generation Protocol for Transcriptome Analysis from a Single Cell

Rachel Fish, Sally Zhang, Magnolia Bostick, Cynthia Chang, Suvarna Gandlur, Andrew Farmer[1]
[1]Corresponding Author
Clontech Laboratories, Inc., 1290 Terra Bella Ave., Mountain View, CA 94043

Abstract

As Next Generation Sequencing (NGS) technologies and transcriptome profiling using NGS mature, they are increasingly being used for more sensitive applications that have only limited sample availability. The ability to analyze the transcriptome of a single cell consistently and meaningfully has only recently been realized.

SMART™ technology is a powerful method for cDNA synthesis that enables library preparation from very small amounts of starting material. Indeed, the SMARTer® Ultra™ Low RNA method allows researchers to readily obtain high quality data from a single cell or 10 pg of total RNA—the approximate amount of total RNA in a single cell. Recent studies have used this method to investigate heterogeneity among individual cells based on RNA expression patterns (1, 2).

A new SMARTer Ultra Low kit has been developed that is simpler and faster while improving the quality and yield of the cDNA produced. The full-length cDNA from this method may be used as a template for library sample preparation for Ion Torrent and Illumina® NGS platforms. Sequencing results for libraries created from single cells or from equivalent amounts of total RNA demonstrate that approximately 90% of the reads map to RefSeq, less than 0.5% of the total reads map to rRNA, and the average transcript coverage is uniform. Improvements in the protocol following first strand synthesis and during cDNA amplification show higher sensitivity with an increase in gene counts and improved representation from GC-rich genes. These data indicate that the improved SMART cDNA protocol is an ideal choice for single cell transcriptome analysis.

Introduction

The SMARTer Ultra Low Input RNA Kit for Sequencing - v3 is the latest in a series of Clontech® products targeted at generating cDNA for RNA-seq from very low inputs. This third generation oligo(dT)-primed kit has been especially designed for transcriptome profiling of 1–1,000 cells or 10 pg–10 ng of purified total RNA. The major changes in this kit include improvements in the cell lysis step, optimization of primers and oligonucleotides, elimination of the clean-up step prior to cDNA amplification, and using the more robust SeqAmp™ PCR polymerase for cDNA amplification.

Materials and Methods

cDNA libraries were generated according to the identified protocols. For UL-HV, cDNA was purified with AMPure XP beads and amplified with Advantage® 2 polymerase. For UL-v3, cDNA was not purified prior to amplification by SeqAmp polymerase.

Illumina adapters and indices were added using the Nextera® XT protocol with 100–250 pg input cDNA. cDNA libraries were sequenced on an Illumina MiSeq® platform with 1 x 50 reads.

Ion Torrent adapters and barcodes were added using the Ion Xpress Plus Fragment Library Preparation Kit and Ion Xpress Barcode Adapter Kit with 1–10 ng cDNA input. cDNA libraries were size selected for 200 base reads and sequenced on an Ion Torrent PGM platform.

Reads were trimmed by CLC Genomics Workbench and mapped to rRNA and the mitochondrial genome with CLC. Unmapped reads were then mapped with CLC to the human or mouse genome with RefSeq masking, producing uniquely mapped reads and % reads that mapped to RefSeq annotations. The number of genes identified in each library was determined by the number of genes with a RPKM ≥0.1. Expression levels of different genes were obtained using CLC Genomics Workbench after mapping to RefSeq.

Gene body coverage was determined using the geneBody_coverage.py module of RSeQC. Read coverage was normalized using Excel.

Conclusions

The SMARTer Ultra Low Input RNA Kit for Sequencing - v3 produces accurate, full-length, unbiased cDNA, even from single cells or single-cell levels of total RNA (10 pg). When used with Illumina or Ion Torrent NGS platforms, this kit generates robust libraries, accurate gene quantification, and reproducible, high-quality sequence data. This allows researchers to have greater confidence in their RNA-seq data.

- **Improved sensitivity**—The new kit results in higher mapping to RefSeq, a higher exon:intron ratio, a greater number of genes detected, lower mapping to rRNA (Table I), and more even gene body coverage (Figures 6 and 9).
- **Higher expression of GC-rich genes**—While the UL-HV and UL-v3 kits both generate highly reproducible data, the new PCR conditions in the UL-v3 kit result in higher expression of GC-rich genes (Figure 3).
- **Optimized for single-cell RNA-seq**—The UL-v3 kit produces cDNA of high yield and high quality from single cells and can be used for either Illumina or Ion Torrent library construction (Figure 4, Table II).
- **Contaminating sequences are minimized**—The new kit is optimized to maintain the representation of the original mRNA transcripts and prevent amplification of contaminating sequences (Figure 7, Table III).

References

1. Deng, Q., *et al.* (2014) *Science* **343**(193):193-196.
2. Shalek A.K., *et al.* (2013), *Nature* **498**(7453):236-240.
3. Mortazavi, A., *et al.* (2008) *Nature Methods* **5**(7):621-628.
4. Chenchik, A., *et al.* (1998) In RT-PCR Methods for Gene Cloning and Analysis. Eds. Siebert, P. & Larrick, J. (BioTechniques Books, MA), pp. 305–319.
5. Picelli, S., *et al.* (2013) *Nature Methods* **10**(11):1096–1098.
6. Picelli, S., *et. al.* (2014) *Nature Protocols* **9**(1):171–181.

PS603686

1 SMART Technology for cDNA Synthesis from Small Amounts of Total RNA

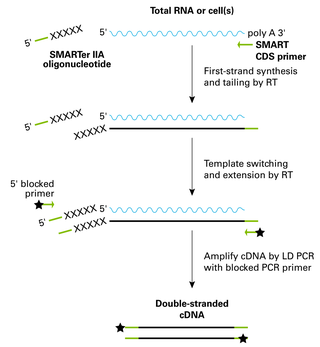
cDNA is generated with the SMARTScribe™ RT and SMART CDS primer.

The RT adds non-templated nucleotides to the 3′ end of the cDNA.

The SMARTer oligonucleotide pairs with these additional nucleotides and allows the RT to continue to generate cDNA (4).

The SMARTer oligonucleotide and primer sequences serve as universal priming sites for PCR amplification.

A blocked PCR primer prevents adapters from being ligated and sequenced.



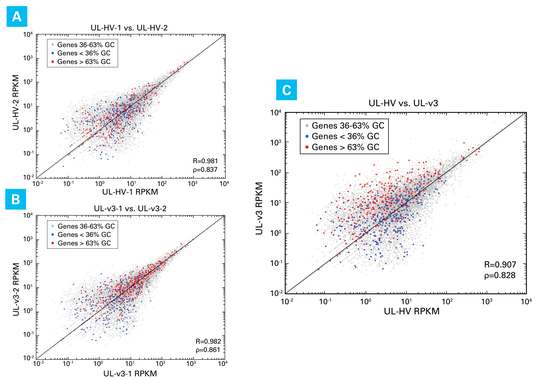
2 Primary Sequencing Metrics Are Improved with the UL-v3 Kit

Table I: Primary Sequencing Metrics from Very Small Amounts of Total RNA Comparing Two cDNA Synthesis Protocols

Input	100 pg HBR				10 pg MBR	
Protocol	UL-HV		UL-v3		UL-HV	UL-v3
Number of reads (millions)	1.9	2.2	2.0	2.0	3.8	4.2
Mapped to rRNA	4.4%	4.2%	1.3%	1.4%	2.7%	0.3%
Mapped to RefSeq	80%	79%	89%	89%	78%	86%
Mapped uniquely to RefSeq	70%	69%	79%	78%	68%	75%
Mapped to exons	57%	58%	75%	76%	67%	75%
Mapped to introns	43%	42%	25%	24%	33%	25%
Number of genes	11,539	11,631	11,992	12,003	8,518	9,136

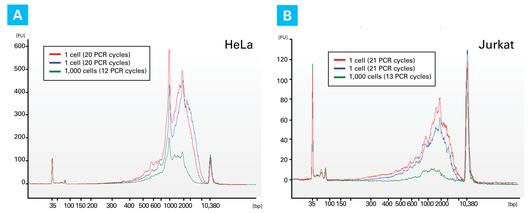
Sequencing metrics from very small amounts of total RNA comparing two cDNA synthesis protocols. Compared to UL-HV, the UL-v3 kit results in significant improvements in the number of genes detected, the exon:intron ratio, and mapping to RefSeq. The increased sensitivity gained using the UL-v3 protocol is most significant at the lowest RNA inputs.

3 Expression Levels of GC-Rich Genes Are Increased Using UL-v3



Expression levels by gene GC content comparing two cDNA synthesis protocols. The libraries made from 100 pg HBR using either the UL-HV protocol or the UL-v3 protocol were compared. Genes were binned by GC content and correlation plots were used to evaluate the two protocols. The average gene counts are very reproducible for replicate samples using UL-HV (**Panel A**) or UL-v3 (**Panel B**). Genes with a high GC content (shown in red) show higher expression with the UL-v3 protocol while genes with a median or low GC content (shown in gray and blue, respectively) show an even distribution when the two protocols are compared (**Panel C**). Pearson (R) and Spearman (ρ) correlations are indicated for all genes regardless of GC content.

4 UL-v3 Robustly Generates cDNA from Individual HeLa and Jurkat Cells



Electropherograms of amplified SMARTer cDNA. Individual cells or 1,000 cells (HeLa or Jurkat) were used as input for SMARTer cDNA synthesis using the UL-v3 kit. The cDNA samples were analyzed for purity and yield on an Agilent 2100 Bioanalyzer. The single main peak indicates the purity and yield of the cDNA (the additional peak at ~1 kb in the HeLa cell samples is the contribution of a single highly expressed gene, FTH1). The reproducibility of these traces shows the reliability and sensitivity of the UL-v3 kit for cDNA synthesis from whole cells.

5 Primary Sequencing Metrics Are Similar for a Single Cell and 1,000 Cells

Table IIA: Primary Sequencing Metrics from Illumina Sequencing

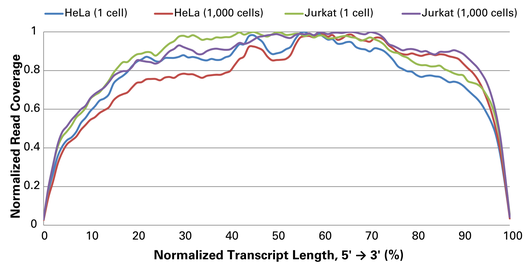
Sequencing platform	MiSeq			
Cell type	HeLa		Jurkat	
Input	1 cell	1,000 cells	1 cell	1,000 cells
Millions of reads	2.1	2.5	1.1	1.8
Mapped to rRNA	0.2%	0.4%	0.2%	0.3%
Mapped to RefSeq	94%	91%	83%	86%
Mapped uniquely to RefSeq	76%	70%	66%	68%
Mapped to exons	85%	92%	77%	84%
Mapped to introns	15%	8.4%	23%	16%
Number of genes	9,261	12,439	4,200	12,611
Pearson correlation (R)	0.973		0.876	

Table IIB: Primary Sequencing Metrics from Ion Torrent Sequencing

Sequencing platform	PGM	
Cell type	HeLa	Jurkat
Input	1 cell	1 cell
Millions of reads	2.3	2.4
Mapped to rRNA	0.2%	0.2%
Mapped to RefSeq	94%	80%
Mapped uniquely to RefSeq	83%	70%
Mapped to exons	85%	75%
Mapped to introns	15%	25%
Number of genes	9,256	4,199

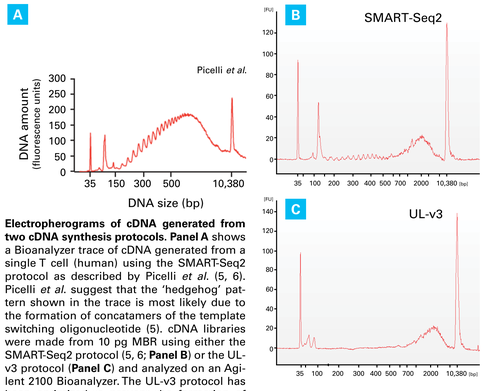
Sequencing metrics from cells. cDNA libraries made from 1 cell or 1,000 cells (HeLa or Jurkat) using the UL-v3 protocol were sequenced on an Illumina MiSeq platform (**Table IIA**). cDNA libraries made from 1 cell (HeLa or Jurkat) using the UL-v3 protocol were also sequenced on an Ion Torrent PGM platform (**Table IIB**). When using whole cells as input, the UL-v3 protocol maintains the high mapping to RefSeq, high exon:intron ratio, and low mapping to rRNA seen with low-input purified total RNA, even for a single cell.

6 Gene Body Coverage Is Even and Reproducible for Different Cell Types



Gene body coverage from different cell inputs. The gene body coverage of cDNA libraries made from 1 or 1,000 HeLa cells (blue and red lines, respectively), or 1 or 1,000 Jurkat cells (green and purple lines, respectively) using the UL-v3 kit is shown. Independent of the cell type used, there is good coverage without strong 5′ or 3′ biases. Coverage is similar for a given cell type regardless of the number of cells used as input for cDNA synthesis.

7 The UL-v3 Protocol Is Optimized for Accurate Generation of cDNA



Electropherograms of cDNA generated from two cDNA synthesis protocols. Panel A shows a Bioanalyzer trace of cDNA generated from a single T cell (human) using the SMART-Seq2 protocol as described by Picelli *et al.* (5, 6). Picelli *et al.* suggest that the 'hedgehog' pattern shown in the trace is most likely due to the formation of concatamers of the template switching oligonucleotide (5). cDNA libraries were made from 10 pg MBR using either the SMART-Seq2 protocol (5, 6; **Panel B**) or the UL-v3 protocol (**Panel C**) and analyzed on an Agilent 2100 Bioanalyzer. The UL-v3 protocol has been optimized to prevent the formation of concatamers; therefore no 'hedgehog' pattern is observed.

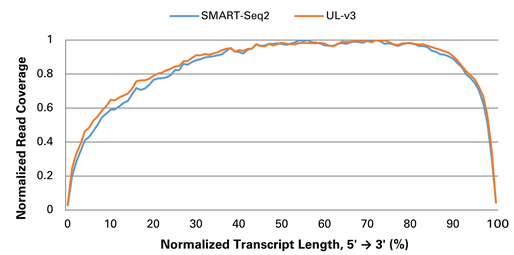
8 Primary Sequencing Metrics Are Similar Between UL-v3 and SMART-Seq2

Table III: Primary Sequencing Metrics from Two cDNA Synthesis Protocols

Input	10 pg MBR	
Protocol	SMART-Seq2	UL-v3
Number of reads (millions)	5.0	3.3
Reads discarded	**64%**	**3%**
Mapped to RefSeq	75%	83%
Mapped uniquely to RefSeq	64%	72%
Mapped to exons	71%	76%
Mapped to introns	29%	24%
Number of genes	7,776	7,421

Sequencing metrics for cDNA generated by two protocols from very small amounts of total RNA. Sequencing metrics generated using the Illumina MiSeq platform are similar between the two protocols. However, the % reads discarded is much higher for SMART-Seq2 generated libraries. It is likely that the concatamers in the SMART-Seq2 libraries (Figure 7B) contribute to this high fraction.

9 Gene Body Coverage Is Comparable for UL-v3 and SMART-Seq2



Comparison of gene body coverage for two cDNA synthesis protocols. The gene body coverage from libraries made with 10 pg MBR using either the SMART-Seq2 protocol (blue line) or the UL-v3 protocol (orange line) is shown. Both protocols give similar, unbiased gene body coverage profiles.

Abbreviations

RPKM—**R**ead **p**er **K**ilobase of Exon per **M**illion Reads (3)
UL-HV—SMARTer Ultra Low Input RNA Kit for Illumina Sequencing - HV
UL-v3—SMARTer Ultra Low Input RNA Kit for Sequencing - v3
RT—Reverse Transcriptase
HBR—Human Brain Total RNA
MBR—Mouse Brain Total RNA
SMART—**S**witching **M**echanism **a**t 5′ End of **R**NA **T**emplate