Achieving Unparalleled Sensitivity and Reproducibility in Single-Cell Transcriptomics

Rashmi Rajendra, Kazuo Tori, Cynthia Chang, Yevgeniy Gindin, Magnolia Bostick, and Andrew Farmer[1]

[1]Corresponding Author: Andrew_Farmer@clontech.com

Abstract

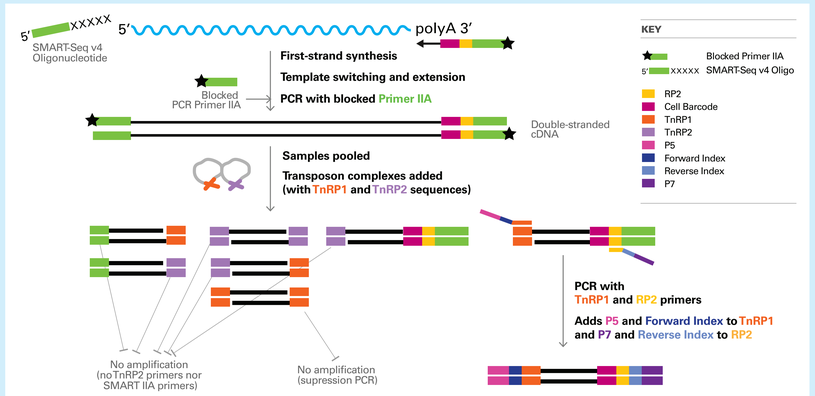
Gaining the ability to identify and quantify the mRNA from a single cell has been a substantial benefit to many scientific fields, especially those where homogeneous populations are elusive, such as cancer research, developmental biology, neurobiology, and immunology. SMART® (**S**witching **M**echanism **a**t the 5′ end of the **R**NA **T**emplate) single-cell technology has emerged as the best solution for processing the small amounts of mRNA present in each cell. With the SMART-Seq® v4 Ultra® Low RNA Input Kit for Sequencing, we have incorporated LNA technology in order to produce high-quality, reproducible sequencing data with superior identification of genes with low expression. Ideally, researchers working with single cells would like to analyze hundreds or thousands of individual cells. Unfortunately, many of the protocols for transcriptome library production from single cells are time consuming and not readily amenable to scaling experiment size. By modifying adapters, adding cellular indexes, and pooling, we have developed a simplified protocol that enables parallel library production for high-throughput mRNA quantification utilizing 3′ end capture.

Here we present this new approach with multiple validation experiments. We first confirm the validity of pooling by mixing single reactions from different sample types with ERCC spike-in controls and testing for cross contamination. Additionally, samples within pools are compared to samples done completely independently. We show limited cross contamination within pooled samples (0.02%) which is not present in independent samples. We further validate the SMART-Seq v4 Ultra Low RNA Input Kit for single-cell sequencing by pooling 12 individual cells. We observe a high number of transcripts identified and high correlations between the cells. This approach allows each pool of 12 cells to be tagged by one of the 96 Illumina HT barcode combinations, enabling up to 1,152 separate cell reads per run. Researchers using this kit can improve sensitivity while decreasing the cost and time required for discovery.

Introduction and Methods

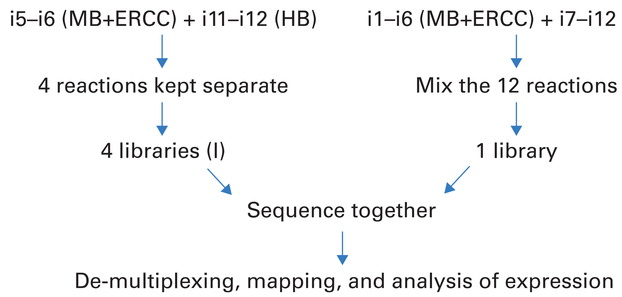
End-capture methods are appealing as they can decrease the amount of reads necessary to determine differential expression between cells. Additionally, samples can be pooled prior to library preparation when barcodes are added, decreasing the work and resources required and potentially increasing the multiplexing capabilities of a flow cell. In this poster, we demonstrate the use of SMART-Seq v4 to develop an end-capture method for low-input amounts down to the single-cell level. This method utilizes a modified RT primer which includes a barcode and a portion of the Illumina® read primer 2 sequence in order to accommodate a pooled library generation protocol.

Library preparation for differential expression



Overview of the end-capture method for analyzing differential expression. cDNA (black) is synthesized with a blocked (black star) and modified oligo(dT) primer that adds sequences for subsequent amplification and analysis—a cell barcode (magenta), part of the Illumina read primer 2 sequence (RP2, yellow), and the SMART IIA sequence (green). The SMART IIA sequence is used as a priming site during cDNA amplification, the Illumina RP2 sequence is used as a priming site during library amplification, and the cell barcode is used for demultiplexing pooled samples during analysis. The process works as follows: first, the template for SMARTScribe™ reverse transcriptase switches from the mRNA (blue wavy line) to the SMART-Seq v4 Oligonucleotide (green). After reverse transcription, the full-length cDNA is amplified by PCR with blocked Primer IIA oligonucleotides. After cDNA amplification, the presence of the barcode (magenta) allows for pooling up to 12 samples. The pooled samples are tagmented and Illumina Nextera® read primer 1 and 2 sequences are added by the Nextera Tn5 transposon (TnRP1 and TnRP2, orange and purple respectively). The 3′ ends of the original cDNA are captured by selective PCR with primers for the TnRP1 and RP2 sequences. Other products of the transposon-based reaction are not amplified, either because there are no primer sites for amplification or because of suppression PCR. Cluster generation (pink and dark purple) and indexing sequences (light blue and dark blue) are added during this PCR stage to generate a library ready for sequencing on an Illumina platform.

1 Testing the end-capture method



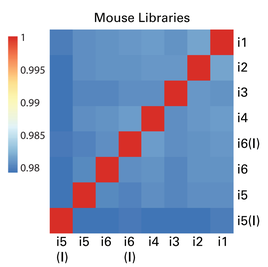
Experimental design to test the end-capture strategy with low-input amounts (10 pg RNA). To test the validity of the pooling strategy described in the Introduction and Methods (left panel), we designed an experiment to compare sequencing results from unpooled and pooled libraries. In this experiment, a total of 16 reactions were performed. Four unpooled reactions were performed: two reactions used Mouse Brain Total RNA (MB in the figure, Clontech product) with External RNA Control Consortium (ERCC) spike-in controls (Thermo Fisher Scientific) and priming oligos containing index5 (i5) or index6 (i6); and two reactions used human brain (HB) RNA (Clontech) (without spike-in controls) and priming oligos containing index11 (i11) and index12 (i12). These samples were kept unpooled until they were loaded on the sequencer to run. An additional twelve reactions were performed using all twelve of the priming oligos with all indexes (i1-i12). Of these twelve reactions, six used mouse brain (MB) RNA with ERCC spike-in controls (i1-i6), while six were made using human brain (HB) RNA (i7-i12). These reactions were pooled after PCR (according to the Methods figure in the Introduction and Methods), and a single Illumina index was added to the pooled samples. The five libraries (four unpooled and one pooled) were multiplexed and sequenced together on an Illumina MiSeq®. Figures 2–4 illustrate the results of this experiment; the four unpooled libraries are labeled independent (I).

2 Mapping statistics: total RNA

Samples	Individual reactions (not pooled)				Pooled samples											
RNA type	Mouse Brain + ERCC		Human Brain		Mouse Brain + ERCC						Human Brain					
In-Line barcode	i5 (I)	i6 (I)	i11 (I)	i12 (I)	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10	i11	i12
Overal Mapping (%)	80	84	80	77	84	88	89	87	84	86	82	77	76	74	77	74
mtRNA (%)	7	7	9	8	7	7	7	7	6	7	9	8	8	8	8	8
rRNA (%)	1.7	1.9	0.7	0.6	1.8	2.0	2.0	1.8	1.7	1.9	0.9	1.3	1.0	0.6	0.8	0.6
Mapped to ERCC (%)	4.1	4.1	0.0	0.0	3.9	4.0	4.2	3.9	4.0	3.9	0.015	0.015	0.015	0.017	0.013	0.015
Number of reads (M)	0.85	0.86	0.94	0.86	1.64	1.61	1.42	1.46	1.40	1.40	0.58	0.52	0.56	0.52	0.57	0.56
# of genes identified	7,999	8,349	6,045	5,041	8,983	8,861	8,866	8,736	8,782	8,672	6,010	6,213	5,801	5,879	6,059	6,110

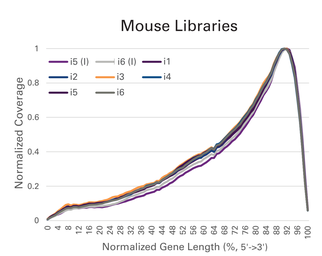
Mapping statistics for pooled and unpooled libraries. The five libraries (four unpooled and one pooled) were sequenced on an Illumina MiSeq® instrument with 150 bp for read 1 and 30 bp for read 2. The pooled libraries were demultiplexed based on the in-line barcode sequence from read 2. All libraries were mapped with TopHat v2.0.9 / Bowtie2 v2.1.0 (Langmead *et al.*, 2012 and Trapnell *et al.*, 2009) against the mouse genome (mm10) or human genome (hg19). The four unpooled libraries are labeled independent (I). In the case of the pooled samples, the reads map to the genome at a high rate (74–89%) with a small proportion mapping to rRNA or mitochondrial regions. There were no obvious differences between the libraries within or not within the pooled sample. The proportion of reads that map to the ERCC transcripts was also measured. The ERCC spike-in controls were added only to the mouse brain-derived libraries, and, as expected, these libraries contain reads that map to the ERCC transcripts. In the unpooled libraries, no ERCC reads were detected in the two human brain-derived libraries; however, a small percentage of the reads from the human brain libraries that were pooled mapped to the ERCC transcripts (<0.02%). The source of these transcripts is unknown, although it may be a result of crossover during PCR.

3 High library correlation: total RNA



Libraries are highly correlated regardless of pooling. The heat maps represent the Pearson correlations of regularized log-transformed read-counts (rlog) for the libraries derived from mouse brain samples. Unpooled libraries are labeled (I). For all comparisons, the correlation (R) was >0.9. There is greater variation among the human brain samples, where there were fewer genes identified. Overall, there is no evidence of a correlation bias, indicating that the unpooled samples (I) are not more similar to each other than the pooled samples and vice versa.

4 Efficient capture of 3′ cDNA ends



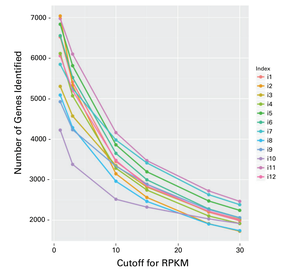
Gene body coverage analysis for unpooled and pooled libraries. Once the reads from each library were mapped to the mouse genome, gene body coverage analysis was performed to assess the ability of the methods to capture the 3′ ends of the cDNA. Across all transcripts (normalized in length to 100%), the majority of reads mapped to the last 20% of the transcripts. There was no difference in 3′ coverage between the pooled and unpooled methods, illustrating the accuracy of the selectivity of the second PCR reaction. The two unpooled libraries are labeled independent (I).

5 Mapping statistics: K562 single cells

In-Line	Pool	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10	i11	i12
mtRNA (%)	3.3	3.1	3.6	1.7	3.9	4.1	1.8	2.5	2.8	5.5	4.1	2.7	4.8
rRNA (%)	0.6	1.2	0.7	1.1	0.4	0.4	0.3	0.3	0.7	1.2	0.5	0.1	0.6
Uniquely Mapped reads (%)	69	68	68	70	68	71	71	72	68	69	69	74	69
Total Mapped reads (%)	97	96	98	97	97	98	98	98	96	97	96	98	98
Number of reads (M)	21.6	2.3	5.0	1.3	2.2	2.0	1.9	0.8	1.3	0.6	0.2	1.5	1.6

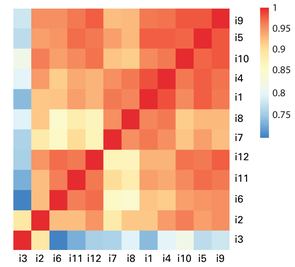
Mapping statistics for pooled libraries from K562 single cells. K562 (human immortalized myelogenous leukemia line) cells were diluted to one cell/µl in PBS buffer and twelve single cells were isolated, checked via optical microscopy, lysed, and subjected to cDNA synthesis. The pooled libraries were sequenced on an Illumina MiSeq instrument with 47 bp for read 1 and 26 bp for read 2. The pooled libraries were demultiplexed based on the in-line barcode sequence from read 2. All libraries were mapped with STAR v.2.3.0.1 (Dobin *et al.*, 2013) against the human genome (hg19). The reads map to the genome at a high rate (>96%) with a small proportion mapping to rRNA or mitochondrial regions.

6 Identification of a large number of genes



Number of genes identified from K562 single cells. Mapped libraries were analyzed with GenomicAlignments v1.4.2 (Lawrence *et al.*, 2013). The number of genes identified with different cutoffs (1, 3, 10, 15, 25, and 30) for Reads Per Kilobase of transcript per Million mapped reads (RPKM) are plotted. The amount of cDNA produced from each cell varies, leading to different read depths per cell. This affects the number of genes identified, as seen most obviously for sample i10 (read depth is shown in Figure 5 as Number of reads).

7 High library correlation: K562 single cells



Pearson correlation heat map matrix of K562 single cells. The heat map represents the Pearson correlations of regularized log-transformed read-counts (rlog) for the 12 single cell libraries. For all comparisons, the correlation (R) was >0.7, while the majority of single cell libraries are highly correlated (>0.9).

Conclusions

- End-capture technology based on SMART-Seq v4 chemistry can be used to pool samples
- Pooled libraries provide highly sensitive and reproducible data with low cross-contamination
- End-capture technology enables highly multiplexed experiments with single cells–up to 1,152 separate cells per run

References

1. Chenchik, A., *et al.* (1998) Generation and use of high-quality cDNA from small amounts of total RNA by SMART PCR. In RT-PCR Methods for Gene Cloning and Analysis. Eds. Siebert, P. & Larrick, J. (*BioTechniques* Books, MA), pp. 305–319.
2. Dobin, A., *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**(1):15-21.
3. Langmead, B. and Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**(4):357–359.
4. Lawrence, M., *et al.* (2013) Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology* **9**(8): e1003118.
5. Love, M. I., *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**(12):550.
6. Trapnell, C., *et al.* (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**(9):1105–1111.

For more information, please visit
http://info.clontech.com/ABRF_signup.html

1290 Terra Bella Ave., Mountain View, CA 94043
Orders and Customer/Technical Service: 800.662.2566
Visit us at www.clontech.com