Two methods for full-length RNA sequencing for low quantities of cells and single cells

Xinghua Pan[a,1], Russell E. Durrett[b,2], Haiying Zhu[a,c,2], Yoshiaki Tanaka[a,2], Yumei Li[a,d], Xiaoyuan Zi[a], Sadie L. Marjani[a], Ghia Euskirchen[e], Chao Ma[f,g], Robert H. LaMotte[f], In-Hyun Park[a], Michael P. Snyder[e], Christopher E. Mason[b], and Sherman M. Weissman[a,1]

Departments of [a]Genetics and [f]Anesthesiology, Yale University School of Medicine, New Haven, CT 06520; [b]Weill Cornell Medical College, New York, NY 10065; [c]Department of Cell Biology, Second Military Medical University, Shanghai 200433, China; [d]Jiangsu University Affiliated Hospital, Zhenjiang, Jiangsu 212001, China; [e]Department of Genetics, Stanford University, Palo Alto, CA 94305; and [g]Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing 100005, China

Contributed by Sherman M. Weissman, October 8, 2012 (sent for review August 22, 2012)

The ability to determine the gene expression pattern in low quantities of cells or single cells is important for resolving a variety of problems in many biological disciplines. A robust description of the expression signature of a single cell requires determination of the full-length sequence of the expressed mRNAs in the cell, yet existing methods have either 3′ biased or variable transcript representation. Here, we report our protocols for the amplification and high-throughput sequencing of very small amounts of RNA for sequencing using procedures of either semirandom primed PCR or phi29 DNA polymerase-based DNA amplification, for the cDNA generated with oligo-dT and/or random oligonucleotide primers. Unlike existing methods, these protocols produce relatively uniformly distributed sequences covering the full length of almost all transcripts independent of their sizes, from 1,000 to 10 cells, and even with single cells. Both protocols produced satisfactory detection/coverage of the abundant mRNAs from a single K562 erythroleukemic cell or a single dorsal root ganglion neuron. The phi29-based method produces long products with less noise, uses an isothermal reaction, and is simple to practice. The semirandom primed PCR procedure is more sensitive and reproducible at low transcript levels or with low quantities of cells. These methods provide tools for mRNA sequencing or RNA sequencing when only low quantities of cells, a single cell, or even degraded RNA are available for profiling.

RNA-seq | transcriptome amplification | full-length mRNA | single-cell analysis

Most populations of cells from higher eukaryotes are heterogeneous in ways that cannot be fully elucidated by bulk analysis. The causes of this heterogeneity include differentiation in subtly different ways, varying stages of the cell cycle, cellular senescence, and nonuniform RNA processing and degradation. Such cellular heterogeneity could be studied by robust techniques for single cell transcriptome analysis, particularly if the techniques analyze full-length transcripts. Reliable methods for transcriptome analysis are also required for situations where only low quantities (LQs) of cells are available, and where the RNA may be partly degraded.

Advances in high-throughput sequencing and innovations in biochemical techniques have revealed a complex picture of the mammalian transcriptome (1). Most genes that contain three or more exons give rise to alternatively spliced products that may vary with the cell type or state of differentiation (2), and these alternative splice forms often have different, even antagonistic functions. In an extreme case, the *Drosophila Dscam* gene has >30,000 alternative transcripts hypothesized to provide distinct identities to individual neuronal dendrites and to avoid self-interaction between the processes of a single neuron (3). Thousands of long, polyadenylated, intergenic noncoding RNAs (LINCs) have been discovered (4, 5) that may have diverse regulatory functions, including serving as scaffolds for proteins that interact with chromatin (6). A fraction of these LINC RNAs may be translated and encode short peptides (7). Cytoplasmic recapping of RNAs has been demonstrated enzymatically (8, 9). A number of genes use multiple promoters, and the position of the 5′ transcription start sites of RNAs may shift under different physiologic conditions. Finally, the mRNA 5′ UTR are now known to be translated frequently (10–13) and may produce biologically active peptides. More than half of the translation initiation sites used by a cell are not predicted from annotated genes, which include many that occur in the 5′ leader sequences of mRNAs, may use near-canonical UUG, CUG, or GUG start codons, and initiate from the internal region (13). These sites could generate proteins with altered functions (14). These events, as well as issues such as RNA editing and allele specific levels of expression (15), indicate the value of deep sequencing of full-length transcripts.

Several approaches have been proposed for obtaining transcriptome data from single cells. An early approach used RT and oligo-dT primers with a T7 phage RNA polymerase promoter sequence attached to the 5′ end of the oligo-dT run. The resulting cDNA was transcribed into multiple copies of RNA, which were then converted back to cDNA (16). This process often truncates the cDNA molecule, losing 5′ sequences of the original mRNA, especially for relatively long transcripts, and requires multiple rounds of processing when starting with LQ cells, further exacerbating cDNA truncation. A recent modification (17) enables multiplex analyses, but this is still 3′ end sequence biased. Other methods are based on PCR amplification of cDNA (18–26). However, these approaches may yield biased representations of sequences along the mRNA and fail to give complete sequences for long mRNAs because long DNA templates are discriminated against even when a long PCR is used.

We have explored two different methods for single and LQ cell cDNA amplification. One approach, Phi29 DNA polymerase-based mRNA transcriptome amplification [Phi29-mRNA amplification (PMA)], was adapted from our whole DNA-pool amplification procedure (WPA) (27), and the full-length mRNA-derived cDNA was circularized by intramolecular ligation before amplification. This method has the unique advantage that it potentially captured all end sequences. Previous analyses of Phi29 DNA polymerase-based whole genomic DNA showed that

Author contributions: X.P. and S.M.W. designed research; X.P., H.Z., Y.L., X.Z., S.L.M., G.E., and C.M. performed research; X.P., I.-H.P., M.P.S., C.E.M., and S.M.W. contributed new reagents/analytic tools; X.P., R.E.D., Y.T., I.-H.P., M.P.S., C.E.M., and S.M.W. analyzed data; and X.P., R.H.L., C.E.M., and S.M.W. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. E-mail: xinghua.pan@yale.edu or sherman.weissman@yale.edu.

[2]R.E.D., H.Z., and Y.T. contributed equally to this work.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1217322109/-/DCSupplemental.

the level of amplification of most regions of DNA varied within less than threefold (28), even though run-away regions of amplification have been noted by others (29, 30), and significant sequence underrepresentation was observed when applied to a single cell (28). Qiagen launched a related product (QuantiTect Whole Transcriptome) using cDNA ligation and phi29 DNA polymerase to generate products for qPCR, but this method has not been used with microarrays or sequencing. We initially implemented the Phi29 DNA polymerase method (27) with single-strand circularization of cDNA reverse transcribed with oligo-dT, and a somewhat similar procedure was recently demonstrated for a single bacterium (31). Using mammalian cells, we modified and improved the sensitivity and uniformity of our method. Also, when random primers were used for RT, a Phi29 DNA polymerase-based transcriptome amplification procedure [Phi29-transcriptome amplification (PTA)] was described. In the second approach, we developed a procedure called semirandom primed PCR-based mRNA transcriptome amplification [SRP-mRNA amplification (SMA)], by adapting a method for nano-ChIP-seq (32). After cDNA was generated, we used semirandom primed PCR to amplify the overlapping segments along the entire length of cDNAs for mRNA sequencing (mRNA-seq). When random primers were used for RT, a semirandom primed PCR-based whole transcriptome amplification method [SRP-transcriptome amplification (STA)] was also tested. Here we compare the relative advantages of each approach and demonstrate the applicability of cDNA sequencing from LQ or single cultured cells or neurons.

Results

Principle of the Methods. The goal of this work was to establish bench top methods for preparing cDNA libraries for high-throughput sequencing that require very limited cellular material and represent the full length of all cDNA molecules. To do this, we optimized the procedure for cDNA generation using a thermostable RT for the generation of cDNA. First-strand synthesis was carried out at 50 °C for efficient RT, in an effort to minimize effects of RNA secondary structure on the elongation of cDNA. Unless otherwise noted, the single-strand cDNA (sscDNA) was converted to the double-stranded form (dscDNA). We used two methods for amplification of very small amounts of cDNA from LQ or single cells. PMA (Fig. 1*A*) was based on previous whole-genome amplification methods (27), which depends on the high processivity and strand displacement properties of the Phi29 DNA polymerase that requires relatively long DNA templates (usually >3–4 kb) for efficient amplification. To avoid this size dependence, we circularized the full-length cDNA using CircLigase (Epicentre) for single-stranded DNA or T4 DNA ligase for double-stranded DNA before amplification. Small circles can be traversed more quickly by the polymerase, but this is largely compensated for by the presence of more primer binding sites on larger circles, such that the occupancy by the DNA polymerase per unit length cDNA is approximately independent of the circumference of the circle. When the DNA was sufficiently diluted such as in single cells or LQ cells, intramolecular circles are predicted to dominate. Thus, the sequence and orientation of the cDNA fragments is representative of the original pool of molecules.

The SMA method (Fig. 1*B*) uses oligonucleotides (SMA-p1) with random 3′ sequences for capture of the whole cDNA sequence and a universal 5′ sequence that serves as a priming site for uniform PCR amplification of all cDNA fragments. The cDNA before SMA remained intact, but the method produced similar results if the cDNA was fragmented into short pieces. After the amplicon was obtained, the oligonucleotide adapter was completely removed with a type IIs restriction enzyme, BciVI, whose recognition sequence was built into SMA-p1. The method uses linear dscDNA or sscDNA as a template and potentially may not capture a short region of sequence at the extreme ends of the cDNA molecules. However, in practice, this did not produce any significant sequence loss. Because of the semirandom priming, each sequence can be covered by multiple different lengths of PCR templates, and because all products are of similar length and amplified with the same primer, the amplification is not subject to the well-known biases of PCR that favor shorter fragments or certain primer sequences. This method enables an extensive and uniform coverage of all sequences. Also, we designed a set of alternative versions for whole transcriptome amplification, namely, the methods PTA and STA (Fig. S1 *A* and *B*) that use random primers on total RNA.

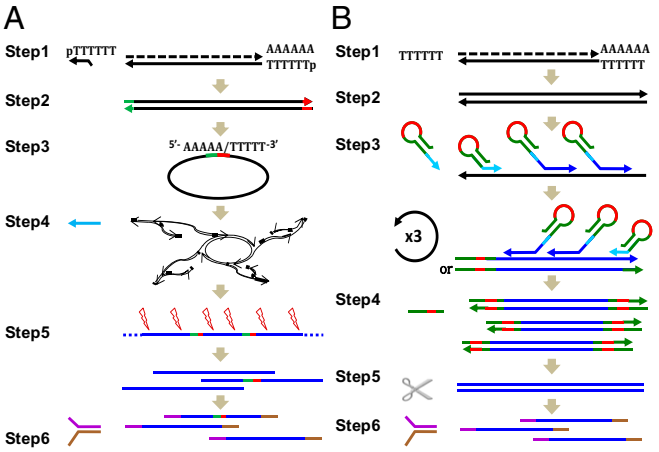

Fig. 1. Outline for PMA and SMA. (*A*) Flowchart for PMA with oligo-dT priming. Step 1: PMA using phosphorylated oligo-dT for mRNA selection with SuperScript Reverse Transcriptase III (SSRTIII) in RT. Step 2: dscDNA generation (for single-strand cDNA PMA, this step is omitted). Step 3: cDNA circularization, mostly intramolecular, possibly partial concatemers consisting of variants of cDNAs. The joint 5′ (green) and 3′ (red) end retain the direction of the transcript, which can be decoded in sequencing data analysis. Step 4: phi29 DNA polymerase-based whole DNA pool amplification for mRNA transcriptome. Step 5: random fragmentation of the amplicons to generate appropriate short sizes of fragments covering the whole mRNA transcriptome. Step 6: Illumina adapter ligation to the cDNA fragments and generation of NGS sequencing library. (*B*) Flowchart for SMA with oligo-dT priming. Step 1: RT for mRNA selection with SSRT III using oligo-dT. Step 2: dscDNA generation (when single-strand cDNA is applied, this step is omitted). Step 3: generation of library of overlapping cDNA fragments with tags on both ends using Sequenase with four cycles of priming (denaturing, annealing, Sequenase addition, and extension). The tricomponents SMA-p1 (semirandom primer) is shown: the blue line shows a 9-mer random portion for random priming; the red is the BciVI cutting site; the green tag helps to form a hairpin structure that minimizes SMA-p1 self-annealing. Step 4: the green and the red are defined universal sequences and used as SMA-p2 for library amplification via PCR. Step 5: BciVI is applied to remove the artificial universal sequence from both ends of all amplicons, leaving a 3′ end A-overhang. Step 6: an Illumina adapter is ligated to the cleared, naked dscDNA fragments, and an NGS sequencing library is generated by additional PCR.

General Amplification Characteristics. We observed that the PMA method did not demonstrate aberrant DNA products visualized by gel electrophoresis, unless a template was added to the reaction mixture. However, in the presence of very small amounts of template, a considerable amount of nonspecific product could be produced. Efficient ligation of the cDNA template was strictly required for the amplification to generate visible amounts of DNA from single cell equivalent of RNA input. With the SMA method, the negative control showed some primer-dimers, but these were obviously distinguishable from the amplicon derived from a template (Fig. S2). The primer-dimer can easily be cut into short pieces with BciVI and removed in downstream

processing. The amplicon yield with SMA, usually about 500 ng, was lower than with PMA (2–5 μg), but sufficient for quality evaluation and library construction.

To compare the two methods, we prepared a batch of K562 erythroleukemic cell RNA from 5×10^5 cells and diluted aliquots equivalent to 1,000 (k), 100 (h), and 10 (t) cells of RNA. cDNA was prepared from duplicate aliquots of each RNA concentration by either PMA or SMA, and the resulting amplified products were sequenced to a depth of at least 10 million total reads for each sample. For both PMA and SMA, the mappable reads and the fractions of sequencing reads that mapped to the coding sequence (CDS), the 3′ and 5′ UTRs, ribosomal RNA, and intergenic genomic DNA, were roughly similar (Fig. 2*A*; Tables S1 and S2). The average GC content profile in the transcripts was also similar, although in SMA it was slightly reduced (Fig. S3*A*). In addition, a significant number of reads of intron sequences was detected: 10- to 20-fold more than the number of reads of intergenic sequences (Fig. 2*A*; Tables S1 and S2). This result indicates that the intron sequences were derived from immature mRNA, rather than gDNA contamination. The majority of fragments were sequenced only once, and the number of fragments sequenced twice or more was less than one third of the total (Fig. 2*B*), suggesting that substantially fewer than one third of the fragments in the cDNA mixture had been identified, even with preparations from as few as 10 cells. In addition, the mappable read number was lower in amplicons from smaller numbers of cells: 10 cells < 100 cells ~ 1,000 cells < standard controls (STDs) (Tables S1 and S2), indicating that less cell input generated relatively more unmappable noise. Overall, the percentage of unique reads for SMA was comparable to STD, but for PMA with diluted RNA it was a little higher, and for PMA directly with low number of cells it was lower. PMA also showed more cDNA copies of ribosomal RNA when a sample with fewer cells was directly amplified (Fig. 2*B*).

When PTA was performed (Fig. 2*A*), the amplicons contained reduced rRNA representation compared with the original rRNA content of the preparation, and sequences of annotated RNA other than rRNA sequences represented up to ~25% of the total reads. This result makes it possible to apply PTA for partially degraded RNA samples and for the case when a whole transcriptome rather than just mRNA is desired. We were also able to perform SMA after only first-strand cDNA synthesis and obtained results that were close to those from double-stranded cDNA. Initial experiments with single-strand circle formation and PMA succeeded for low numbers of cells and single cells. Unfortunately later batches of the CircLigase enzyme (Epicentre) used for the ligation did not work well for larger circles, and thus the resulting data are not shown. However, it is apparent from the PTA data in this study that phosphorylated random primers would generate small single-stranded circles that could be ligated efficiently and used for single cell profiling. We also prepared STA amplicons from ~100-cell equivalent RNA (Fig. S3*E*), and PCR evaluation indicated a satisfactory detection of transcripts.

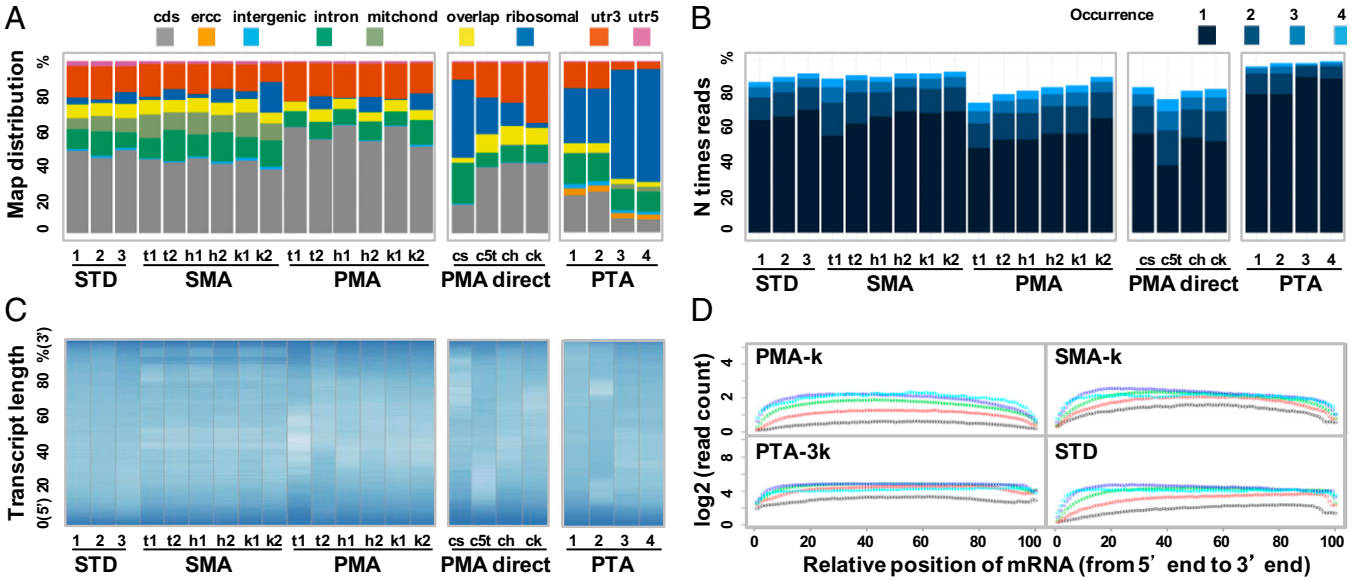

Fig. 2. Sequencing features and full-length coverage. (*A*) Mapping distribution for cDNA sequence reads from SMA and PMA amplicons and unamplified controls (STD). Different colors from the bottom to the top of each column represent fractions of the reads from coding sequences (cds), seeded synthetic oligonucleotides (ercc), intergenic sequences (intergenic), introns (intron), mitochondria (mitochond), sequences overlapping more than one type of target (overlap), ribosomal RNA sequences (ribosomal), and 3′ and 5′ untranslated regions (utr) of mRNA. The samples are labeled as follows (the same labeling is used in other figures hereafter if not specified otherwise): t, h, and k represent diluted total RNA equivalent to 10, 100, or 1,000 cells, respectively. Two replicates of each dilution and each method for amplification are shown. PMA direct corresponds to cDNA prepared from cell lysates (cs: single cell; c5t: 50 cells; ch: 100 cells; ck: 1,000 cells). K562 control (i.e., STD) was from RNA-seq with conventional protocol and represents three biologic replicates of K562 cell RNA prepared from a large culture (each >2 million cells), converted to cDNA, and sequenced without a preliminary amplification. PTA refers to the libraries obtained by random plus oligo-dT priming of RT beginning with total RNA equivalent to 3,000 (3k) cells of an acute promyelocytic leukemia cell line, NB4. PTA-a (PTA-1, 2) used less random primers than in PTA-b (PTA-3, 4). (*B*) Frequency of multiply sequenced cDNA fragments; 1×, 2×, 3×, and 4× sequenced fragments are shown in progressively lighter shades of blue. (*C*) Average coverage of transcripts from 5′ to 3′ ends (bottom to top). The transcripts were divided into 100 equal segments, and the relative intensity of each segment was displayed. Darker blue represents less coverage. (*D*) Distribution of reads across cDNAs of various lengths. These results were obtained from STD or amplified cDNA (labeled as in *A*) from RNA equivalent to 1,000 (k) K562 cells. Each of the SMA-k and PMA-k was the average of two replicates. PTA-3k was counted with the average of PTA-a and -b. *x* axis: the mRNA regions are divided into 100 bins, and RNA-seq reads are counted in each bin from the 5′ to 3′ end. *y* axis: average reads count. RefSeq genes are divided into five classes according to mRNA length. Black: 45–1,000 bp, 4,081 genes (transcripts); red: 1,001–2,000 bp, 5,483 genes; green: 2,001–5,000 bp, 9,634 genes; blue: 5,001–10,000 bp, 2,264 genes; cyan: 10,001–101,674 bp, 259 genes.

Full-Length Coverage. After sequencing, reads were mapped to the human genome (hg19) using TopHat (33). The result demonstrated that, in general, all lengths of transcripts were covered over their full lengths (Fig. 2 *C* and *D*; Figs. S3 *C* and *D* and S4). To display the general coverage of cDNAs, each annotated cDNA (including CDS and 5′ and 3′ UTRs) was divided into 100 parts. The relative intensity along each 1/100th for all cDNAs was summed and plotted (Fig. 2*C*). The results indicate that both methods were able to represent almost the entire length of the cDNA. To further evaluate the effect of cDNA length on coverage, we divided the cDNAs into five length categories according to their length and plotted the intensity of representation for each 1/100th of the cDNAs in each length category. The results (Fig. 2*D*; Fig. S4) show that there was good coverage of the full length of cDNAs independent of the size. Although the coverage for transcripts did drop off near the very ends of transcripts (in all cases at the 5′ end <10%, mostly <3–5% of the length including UTR sequence), we note that this range of drop-off is not significantly worse than all current sequencing RNA sequencing (RNA-seq) methods without amplification and is confounded by the limits of mapping of short reads to the transcripts, as well as other causes. In this aspect, PMA and SMA are superior to a recently reported method, which drops off for ∼40% of the sequences from the 5′ end of 15-kb transcripts (26). For the PMA protocol, one cause for this drop-off is the failure to map the reads derived from the poly-A tail and 5′ end chimerical sequences joined during circularization. This result can be improved through advances in bioinformatic analysis or genome indexing. Another cause of a loss of terminal sequences may be the shortening of the 5′ end of the cDNA during second-strand synthesis. This limitation presumably could also be potentially overcome by coupling the cDNA synthesis procedure with the incorporation of a switch mechanism at the 5′ end of reverse transcript (SMART) oligonucleotide at the 5′ end (34). For SMA, this may be followed by adding additional SMART and poly-dT oligonucleotides, separately incorporated with the universal sequence for capturing both 5′ and 3′ ends during the library generation step (Fig. 1*A*, step 3).

Gene Detection. To evaluate the efficiency of detection of expressed genes, we calculated reads per 1,000 bases of mRNA per million total reads (RPKM) values for annotated genes and scored the gene as present or absent based on various thresholds. The Venn diagrams in Fig. 3*A* show the decline in the numbers of genes detected as the amount of input RNA declined and the relative coverage by extensive sequencing of unamplified cDNA compared with the coverage by PMA and SMA. Both amplification procedures produce little background DNA fragments, but these signals increase as the RNA input amount decreases, and, as such, quantitative mapping is best done by only considering reads in known CDS/UTRs. SMA produces relative more spurious fragments that match genomic DNA at apparently random regions. One possibility for the cause of these spurious fragments is that more of these sequences appear as a result of the amplification of very short sequences of incompletely digested genomic DNA, although these also occur in standard RNA-seq (Fig. S3 *C* and *D*; Tables S1 and S2). Other unmappable reads contribute to the relatively lower mapping rate for LQ cells, especially single cells. These reads include any possible contamination of trace amounts of DNA from laboratory environment or reagents or some artificial DNA generated by the method, as has also been observed in other reports of RNA amplification methods. The consequence of these noise DNA fragments is that a progressively smaller number of reads map to cDNA sequences as the input template is decreased, and more sequencing runs are needed to obtain the desired coverage of cDNA sequences.

Reproducibility and Correlation. To compare the reproducibility and accuracy of both methods, we determined counts per kilobase of CDS/UTRs in the various amplified samples and STD (see Tables S1 and S2 for the read number). Within each method, amplicons were overall better correlated than were amplicons prepared from the same level of samples with the two different methods (Fig. 3B). Input RNA from as low as 10 cells in each method missed some of the weakly expressed cDNAs, but the cDNAs missed by the two methods were often divergent. More abundant cDNAs were generally well represented when either method was used for cDNA amplification (Fig. S5). This is similar to results shown in a recent report (26).

The general pattern of SMA is closer than PMA to STD. SMA also has better reproducibility (Fig. 3*B*). We also correlated several samples on the basis of the relative levels of RPKM for each gene (Fig. 3*C*). The correlation of the replicates and the various levels of starting materials within each method were much closer than that observed between different methods. In each group, 100-cell and 1,000-cell samples are closely related to each other, but 10-cell samples have slightly more variability,

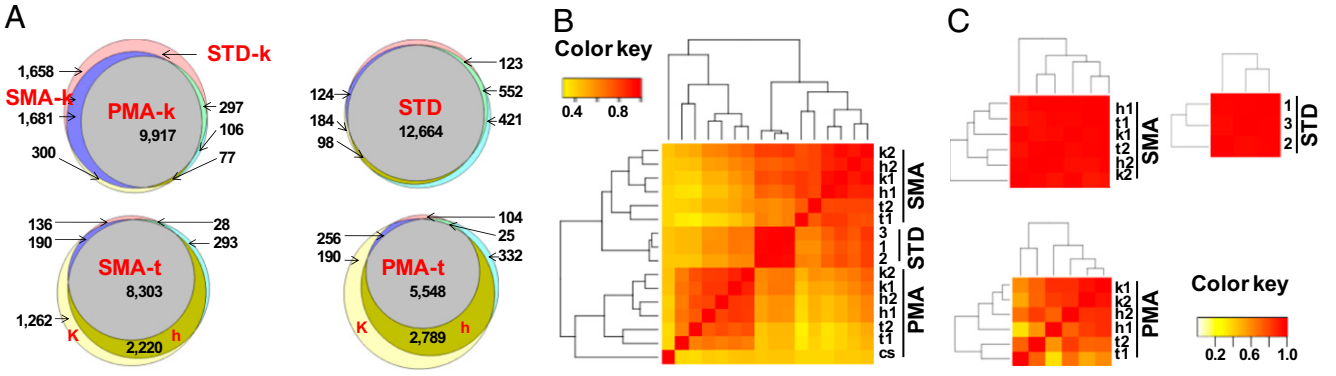

Fig. 3. Gene detection and expression profiling. (*A*) Venn diagram illustrating the number of genes detected at >0.1 RPKM in libraries from various numbers of cells (t, h, or h) with two methods: SMA and PMA. The gene number for each level is an average of two replicates. RNA-seq reads were assembled by Cufflinks with the RefSeq GTF file as a reference annotation. Numbers represent the numbers of genes detected in various overlapping subsets. (*B*) Heatmap comparing gene detection with libraries amplified with PMA and SMA. The results show comparison of replicates with each analysis at different levels of RNA inputs versus STD. The analysis was on the basis of the genes that were detected at >0.1 RPKM, regardless of the intensity of their signals. (*C*) Heatmaps for sequencing results with PMA and SMA amplicons (and reference STD) for comparison of RPKM profiles of genes detected (threshold: RPKM >0.1) in replicate experiments with amplified cDNA corresponding to t, h, or k cells, with 12,640 genes covered.

especially for PMA (Fig. 3*C*). Each method (SMA, PMA) produces reproducible profiles (Tables S3–S5). The Pearson correlation coefficient (r) was consistently >0.9 for SMA samples even when 10-cell RNA was amplified. When a 1,000-cell sample was amplified, the r (0.96) was comparable to the technical repeats of the standard RNA-seq without amplification. For PMA, the r was >0.925 for two 1,000-cell samples and 0.715 for the two 10-cell samples tested.

Single Cell Sequencing. We next explored the analysis of single cell transcriptomes using PMA (Figs. 2 and 4; Figs. S3–S6). It is worth noting that recovering signatures for the transcriptome of single cells is highly dependent on cell type. At one extreme, resting lymphocytes have little cytoplasm or RNA and may be poor candidates for single cell RNA amplification. To demonstrate the utility of PMA for single cells, we manually isolated single K562 erythroleukemic cells from suspension culture. Using one quarter of a lane of multiplex sequencing (75 bases paired end reads), ~5,000 transcripts were detected, and the more abundant genes were well represented with coverage of most or all exons (Fig. 4; Figs. S5 and S6). However, many less abundant genes were either not detected (Fig. S5) or incompletely represented at the depth of sequencing used. This single cell PMA sequencing also showed more unknown transcripts and unannotated transcripts than did amplicons (Fig. 4*A*) from 1000-cell equivalent diluted RNA. However, the mapped genes overall are similar to those when more cells were amplified. We applied the same analyses to a set of single murine dorsal root ganglion cell bodies. Each neuron was individually harvested by suction applied by a micropipette from an intact ganglion whose cells were loosened from their cellular neighbors by prior topical application of collagenase (35). A similar level of transcripts was also detected, as shown in an Integrated Genome Browser (IGB) screen shot (Fig. S6). The cDNA from these neurons was amplified by PMA after each neuron had been functionally classified as nociceptive by its action potential responses, electrophysiologically recorded in vivo, to noxious chemical, thermal, or mechanical stimuli delivered to its cutaneous receptive field (36). In addition, the application of SMA to single cells is also promising (Fig. S6*C*).

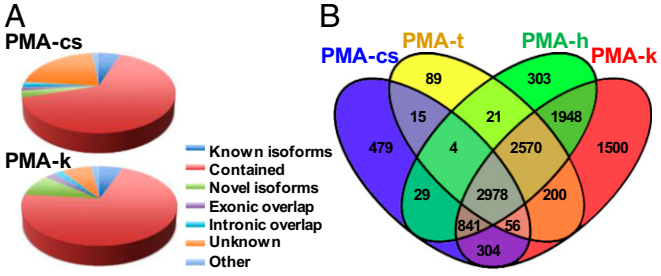

Fig. 4. Single cell mRNA transcriptome sequencing. (*A*) Mapping distribution of sequencing reads from PMA of a single K562 cell (PMA-cs). This is based on isoforms or CDS/UTRs compared with 1,000-cell PMA (PMA-k). PMA-cs totally covers 5,277 genes (each with at least one transcript), independent exons, introns, and other signals. Based on Cuffcompare commands, "novel isoforms" are multiexonic transcripts, which share at least one known exon. The "exonic overlap" and "intron overlap" are those signals that are not included in known or novel isoforms, whereas "exons" indicate single-exonic transcripts, which are overlapped with known exons. "Contained" means truncated isoforms where, for example, one exon is not detected, but the other exons completely match with known exons, so this category mostly contains known isoforms, but needs deeper sequencing to have a full coverage of all sequences/exons, although there may be some new isoforms. "Unknown" refers tothose not able to be classified into any of the above groups. (*B*) Venn diagram showing RNAs for PMA-cs versus PMA-k, PMA-h, and PMA-t for gene detection. All RPKM > 0.1 counts are considered. As in Fig. 3*A*, RNA-seq reads were assembled by Cufflinks with the RefSeq GTF file as a reference annotation.

Discussion

Each of the two procedures demonstrated full-length coverage of the RNA sequences, independent of the length of the transcripts, with cDNA as long as 23 kb. These procedures also covered the 3′ UTRs and 5′ UTRs. The reproducibility is higher within each method than between the two methods, and, for PMA, is also higher when more cells are used. When more starting material was used, the number of genes detected was increased. Because different procedures show somewhat different transcript patterns, for any given biological test, it is necessary to use a consistent procedure throughout the analyses.

The first step in the analysis of transcriptomes is the conversion of mRNA to cDNA. The efficiency of reverse transcription and other reactions depends on an adequate and rapid mixing of liquids may be a limiting factor in some protocols (37). Conversion from single- to double-stranded cDNA may also be a source of loss, particularly at the 5′ end of the mRNA. This should be at least partly avoided by the use of SMART oligonucleotides that attach a known primer binding sequence to the region corresponding to the 5′ end of the mRNA. However, our initial comparison of SMA with first-strand cDNA and with double-stranded cDNA suggests that the second-strand synthesis is not a major source of signal loss.

Overall, our results show that, in comparison with PMA, SMA detects more genes, gives a pattern closer to that obtained from RNA-seq of unamplified cDNA, and is more sensitive, given small amounts of starting material. In addition, SMA is probably more suitable for single cell RNA amplification on the bench top. When combined with magnetic capturing mRNA directly from cell lysate, followed by a direct RT, a high-throughput process of expression profiling should be practical. A similar semirandom PCR strategy of SMA is used in a commercial kit (Transplex Whole Transcriptome Amplification; Sigma-Aldrich). WTA performs well in microarray analysis compared with some other methods (20) but uses long artificial primer sequences not designed to be removed after amplification. Thus, its use has not been reported in conjunction with high-throughput sequencing.

With PMA, we observed incomplete representation of low abundance mRNAs when LQ cells were processed, sometimes with sequences missing from the 3′ end of the original mRNA. This suggests that the loss of these sequences occurs before or during cDNA circularization, perhaps due to exonuclease action during blunt end generation and ligation or incomplete ligation of segments of DNA from the second strand to give the full-length product. However, the sequence loss was at least partly random, as it was not consistent from sample to sample of the same cell type. Although PMA is somewhat less sensitive than SMA, PMA has certain advantages. In principle, it generates intact full-length copies of cDNAs that would be suitable for longer sequence runs as technology becomes available (38). These full-length cDNAs would be important for resolution of ambiguities in assigning splice isoforms. PMA has a particular advantage for application to closed microfluidic systems. This would allow a large number of single cells to be amplified in parallel. It is relatively simple in operation as the steps of manipulations and the number and range of changes of temperature are very limited. Alternatively SMA could be performed in microfluidic apparatus that has PCR capability (as Fluidigm). Carrying out reactions in nanoliter volumes has the potential to substantially improve single cell work (37, 39–42). Zhong et al. reported that the conversion of small amounts of mRNA to cDNA is more efficient in very small volumes and may reach 54% compared with conventional methods that yield as little as 12% (42). Also, the use of small volumes makes it possible to carry out reactions with amounts of enzyme that are more

proportionate to the amount of nucleic acid present. However, because the initial amplification is limited, when working in small volumes, a second-stage amplification may be needed to obtain enough material for some analyses.

In addition to the technical considerations, there is another level of complexity in evaluating the transcriptome of single cells, especially cells substantially smaller than oocytes or early blastocysts. The mRNA has a relatively short half life, and transcription may occur in bursts (43–45). Thus, at any one time, the mRNA content of a cell may be an incomplete representation of the total transcriptome during the cell cycle, as demonstrated here for single K562 cells. This phenomenon suggests that it is best to evaluate the transcriptome from several cells as nearly identical in nature as possible, such as cell cycle stage synchronized, to get the full signature of the transcriptome of a single cell type (46).

The present study demonstrates that rather similar overall results can be obtained for cDNA profiling from LQ cells or even single cells by either of the two amplification procedures described. Importantly, these methods can provide a relatively uniform representation of the full length of even very long cDNAs. At the single cell level, coverage is incomplete but adequate for the detection of the more abundant mRNA species and could be used to evaluate their relative use of different splice isoforms, as well as the detection of unannotated transcripts. In summary, these approaches offer considerable promise for applications in studies of a range of subjects, including development, nervous system structure, and normal and pathologic responses of the human immune system.

Materials and Methods

An overall description is provided in *Principle of the Methods*. K562 cells were cultured in suspension in DMEM. The amplified cDNA was converted into sequencing libraries according to Illumina protocols. Sequencing was performed on Illumina HiSeq instruments. PCR evaluation after transcriptome amplification was performed under standard conditions, and the primers used for PCR are shown in Table S6. Additional experimental and data analysis procedures are provided in *SI Materials and Methods*.

Note Added in Proof. While this manuscript was in preparation, the Smart-Seq method (26) was reported using a long PCR method that provided sequences for a substantial portion of even very long cDNAs, although the distribution of sequences was uneven and the sequences of the 5′ regions of many long mRNAs were significantly underrepresented.

ACKNOWLEDGMENTS. We thank Rong Fan, Shrikant Mane, and Milind Mahajan for support and comments during the project; Mei Zhong, John Overton, and Keith Bettinger for support in sequencing; and Jianfang Wu for help in generation of the flowchart. This work was supported by National Institutes of Health Grants 1P01GM099130-01, 1R21HD066457-01, R01 NS014624, 1R01NS076465-02, and 1R01NS076465-01; Ruth L. Kirschstein National Research Service Award F32GM087109 (to S.L.M.) from the National Institute of General Medical Sciences; and a Connecticut Innovations Stem Cell grant.

1. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet* 10(1):57–63.
2. Wang ET, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456(7221):470–476.
3. Hattori D, et al. (2009) Robust discrimination between self and non-self neurites requires thousands of Dscam1 isoforms. *Nature* 461(7264):644–648.
4. Guttman M, et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458(7235):223–227.
5. Carninci P (2010) RNA dust: Where are the genes? *DNA Res* 17(2):51–59.
6. Khalil AM, et al. (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci USA* 106(28):11667–11672.
7. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324(5924):218–223.
8. Schoenberg DR, Maquat LE (2009) Re-capping the message. *Trends Biochem Sci* 34(9):435–442.
9. Otsuka Y, Kedersha NL, Schoenberg DR (2009) Identification of a cytoplasmic complex that adds a cap onto 5′-monophosphate RNA. *Mol Cell Biol* 29(8):2155–2167.
10. Brar GA, et al. (2012) High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* 335(6068):552–557.
11. Oyama M, et al. (2007) Diversity of translation start sites may define increased complexity of the human short ORFeome. *Mol Cell Proteomics* 6(6):1000–1006.
12. Oyama M, et al. (2004) Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. *Genome Res* 14(10B):2048–2052.
13. Ingolia NT, Lareau LF, Weissman JS (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147(4):789–802.
14. Wethmar K, Smink JJ, Leutz A (2010) Upstream open reading frames: molecular switches in (patho)physiology. *Bioessays* 32(10):885–893.
15. Pastinen T (2010) Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev* 11(8):533–538.
16. Phillips J, Eberwine JH (1996) Antisense RNA amplification: A linear amplification method for analyzing the mRNA population from single living cells. *Methods* 10(3):283–288.
17. Hashimshony T, Wagner F, Sher N, Yanai I (2012) CEL-Seq: Single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* 2(3):666–673.
18. Liu M, Subramanyam YV, Baskaran N (1999) Preparation and analysis of cDNA from a small number of hematopoietic cells. *Methods Enzymol* 303:45–55.
19. Ozsolak F, et al. (2010) Digital transcriptome profiling from attomole-level RNA samples. *Genome Res* 20(4):519–525.
20. Gonzalez-Roca E, et al. (2010) Accurate expression profiling of very small cell populations. *PLoS ONE* 5(12):e14418.
21. Kanamori-Katayama M, et al. (2011) Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res* 21(7):1150–1159.
22. Islam S, et al. (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* 21(7):1160–1167.
23. Tang F, et al. (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 6(5):377–382.
24. Kurimoto K, et al. (2006) An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Res* 34(5):e42.
25. Qiu S, et al. (2012) Single-neuron RNA-Seq: Technical feasibility and reproducibility. *Front Genet* 3:124.
26. Ramsköld D, et al. (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* 30(8):777–782.
27. Pan X, et al. (2008) A procedure for highly specific, sensitive, and unbiased whole-genome amplification. *Proc Natl Acad Sci USA* 105(40):15499–15504.
28. Chitsaz H, et al. (2011) Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. *Nat Biotechnol* 29(10):915–921.
29. Baslan T, et al. (2012) Genome-wide copy number analysis of single cells. *Nat Protoc* 7(6):1024–1041.
30. Navin N, et al. (2011) Tumour evolution inferred by single-cell sequencing. *Nature* 472(7341):90–94.
31. Kang Y, et al. (2011) Transcript amplification from single bacterium for transcriptome analysis. *Genome Res* 21(6):925–935.
32. Adli M, Zhu J, Bernstein BE (2010) Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors. *Nat Methods* 7(8):615–618.
33. Trapnell C, et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7(3):562–578.
34. Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD (2001) Reverse transcriptase template switching: A SMART approach for full-length cDNA library construction. *Biotechniques* 30(4):892–897.
35. Ma C, Donnelly DF, LaMotte RH (2010) In vivo visualization and functional characterization of primary somatic neurons. *J Neurosci Methods* 191(1):60–65.
36. Ma C, Nie H, Gu Q, Sikand P, Lamotte RH (2012) In vivo responses of cutaneous C-mechanosensitive neurons in mouse to punctate chemical stimuli that elicit itch and nociceptive sensations in humans. *J Neurophysiol* 107(1):357–363.
37. Boon WC, Petkovic-Duran K, Zhu Y, Manasseh R, Horne MK, Aumann TD (2011) Increasing cDNA yields from single-cell quantities of mRNA in standard laboratory reverse transcriptase reactions using acoustic microstreaming. *J Vis Exp* 11(53):e3144.
38. Au KF, Underwood JG, Lee L, Wong WH (2012) Improving PacBio long read accuracy by short read alignment. *PLoS ONE* 7(10):e46679.
39. Lecault V, White AK, Singhal A, Hansen CL (2012) Microfluidic single cell analysis: From promise to practice. *Curr Opin Chem Biol* 16(3-4):381–390.
40. Boon WC, et al. (2011) Acoustic microstreaming increases the efficiency of reverse transcription reactions comprising single-cell quantities of RNA. *Biotechniques* 50(2):116–119.
41. Marcus JS, Anderson WF, Quake SR (2006) Microfluidic single-cell mRNA isolation and analysis. *Anal Chem* 78(9):3084–3089.
42. Zhong JF, et al. (2008) A microfluidic processor for gene expression profiling of single human embryonic stem cells. *Lab Chip* 8(1):68–74.
43. Suter DM, Molina N, Naef F, Schibler U (2011) Origins and consequences of transcriptional discontinuity. *Curr Opin Cell Biol* 23(6):657–662.
44. Hager GL, McNally JG, Misteli T (2009) Transcription dynamics. *Mol Cell* 35(6):741–753.
45. Voss TC, et al. (2011) Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism. *Cell* 146(4):544–554.
46. Wang D, Bodovitz S (2010) Single cell analysis: The new frontier in 'omics' *Trends Biotechnol* 28(6):281–290.