RNA sequencing by direct tagmentation of RNA/DNA hybrids

Lin Di[a,1], Yusi Fu[a,1,2], Yue Sun[a], Jie Li[b,c], Lu Liu[b,c], Jiacheng Yao[b,c], Guanbo Wang[d,e], Yalei Wu[f], Kaiqin Lao[f], Raymond W. Lee[f], Genhua Zheng[f], Jun Xu[g], Juntaek Oh[g], Dong Wang[g], X. Sunney Xie[a,3], Yanyi Huang[a,e,h,i,j,3], and Jianbin Wang[b,c,j,k,3]

[a]Beijing Advanced Innovation Center for Genomics (ICG), Biomedical Pioneering Innovation Center (BIOPIC), School of Life Sciences, Peking University, 100871 Beijing, China; [b]School of Life Sciences, Tsinghua University, 100084 Beijing, China; [c]Tsinghua-Peking Center for Life Sciences, Tsinghua University, 100084 Beijing, China; [d]School of Chemistry and Materials Science, Nanjing Normal University, 210046 Nanjing, China; [e]Institute for Cell Analysis, Shenzhen Bay Laboratory, 518132 Shenzhen, China; [f]XGen US Co, South San Francisco, CA 94080; [g]Department of Cellular and Molecular Medicine, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA 92093; [h]College of Engineering, Peking University, 100871 Beijing, China; [i]College of Chemistry and Molecular Engineering, Peking University, 100871 Beijing, China; [j]Chinese Institute for Brain Research (CIBR), 102206 Beijing, China; and [k]Beijing Advanced Innovation Center for Structural Biology, Tsinghua University, 100084 Beijing, China

Edited by David A. Weitz, Harvard University, Cambridge, MA, and approved December 31, 2019 (received for review November 11, 2019)

Transcriptome profiling by RNA sequencing (RNA-seq) has been widely used to characterize cellular status, but it relies on second-strand complementary DNA (cDNA) synthesis to generate initial material for library preparation. Here we use bacterial transposase Tn5, which has been increasingly used in various high-throughput DNA analyses, to construct RNA-seq libraries without second-strand synthesis. We show that Tn5 transposome can randomly bind RNA/DNA heteroduplexes and add sequencing adapters onto RNA directly after reverse transcription. This method, Sequencing HEteRo RNA-DNA-hYbrid (SHERRY), is versatile and scalable. SHERRY accepts a wide range of starting materials, from bulk RNA to single cells. SHERRY offers a greatly simplified protocol and produces results with higher reproducibility and GC uniformity compared with prevailing RNA-seq methods.

single cell | RNA-seq | Tn5 transposase

Transcriptome profiling through RNA sequencing (RNA-seq) has become routine in biomedical research since the popularization of next-generation sequencers and the dramatic fall in the cost of sequencing. RNA-seq has been widely used in addressing various biological questions, from exploring the pathogenesis of disease (1, 2) to constructing transcriptome maps for various species (3, 4). RNA-seq provides informative assessments of samples, especially when heterogeneity in a complex biological system (5, 6) or time-dependent dynamic processes are being investigated (7–9). A typical RNA-seq experiment requires a DNA library generated from messenger RNA (mRNA) transcripts. The commonly used protocols contain a few key steps, including RNA extraction, poly-A selection or ribosomal RNA depletion, reverse transcription, second-strand complementary DNA (cDNA) synthesis, adapter ligation, and PCR amplification (10–12).

Although many experimental protocols, combining chemistry and processes, have recently been invented, RNA-seq is still a challenging technology to apply. On one hand, most of these protocols are designed for conventional bulk samples (11, 13–15), which typically contain millions of cells or more. However, many cutting-edge studies require transcriptome analyses of very small amounts of input RNA, for which most large-input protocols do not work. The main reason for this incompatibility is because the purification operations needed between the main experimental steps cause inevitable loss of the nucleic acid molecules.

On the other hand, many single-cell RNA-seq protocols have been invented in the past decade (16–19). However, for most of these protocols it is difficult to achieve both high throughput and high detectability. One type of single-cell RNA-seq approach, such as Smart-seq2 (17), is to introduce preamplification to address the low-input problem, but such an approach is likely to introduce bias and to impair quantification accuracy. Another type of approach is to barcode each cell's transcripts and hence bioinformatically assign identity to the sequencing data that is linked to each cell and each molecule (20–23). However, the detectability and reproducibility of such approaches are still not ideal (24). An easy and versatile RNA-seq method is needed that works with input from single cells to bulk RNA.

Bacterial transposase Tn5 (25) has been employed in next-generation sequencing, taking advantage of the unique "tagmentation" function of dimeric Tn5, which can cut double-stranded DNA (dsDNA) and ligate the resulting DNA ends to specific adaptors. Genetically engineered Tn5 is now widely used in sequencing library preparation for its rapid processivity and low sample input requirement (26, 27). For general library preparation, Tn5 directly reacts with naked dsDNA. This is followed by PCR amplification with sequencing adaptors. Such a simple

Significance

RNA sequencing is widely used to measure gene expression in biomedical research; therefore, improvements in the simplicity and accuracy of the technology are desirable. All existing RNA sequencing methods rely on the conversion of RNA into double-stranded DNA through reverse transcription followed by second-strand synthesis. The latter step requires additional enzymes and purification, and introduces sequence-dependent bias. Here, we show that Tn5 transposase, which randomly binds and cuts double-stranded DNA, can directly fragment and prime the RNA/DNA heteroduplexes generated by reverse transcription. The primed fragments are then subject to PCR amplification. This provides an approach for simple and accurate RNA characterization and quantification.

Author contributions: Y.F., K.L., X.S.X., Y.H., and J.W. designed research; L.D., Y.F., Y.S., J.L., L.L., G.W., J.X., J.O., and D.W. performed research; L.D., Y.F., L.L., J.Y., Y.W., R.W.L., G.Z., Y.H., and J.W. analyzed data; and L.D., X.S.X., Y.H., and J.W. wrote the paper.

Competing interest statement: XGen US Co. has applied for a patent related to this work. X.S.X., K.L., and Y.W. are shareholders of XGen US Co. K.L., Y.W., R.W.L., and G.Z. are employees of XGen US Co.

This article is a PNAS Direct Submission.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

Data deposition: The sequence reported in this paper has been deposited in the Genome Sequence Archive database, http://gsa.big.ac.cn/ (accession no. CRA002081).

[1]L.D. and Y.F. contributed equally to this work.

[2]Present address: Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030.

[3]To whom correspondence may be addressed. Email: sunneyxie@pku.edu.cn, yanyi@pku.edu.cn, or jianbinwang@tsinghua.edu.cn.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1919800117/-/DCSupplemental.

Downloaded by guest on January 30, 2020

one-step tagmentation reaction has greatly simplified the experimental process, shortening the workflow time and lowering costs. Tn5 tagmentation has also been used for the detection of chromatin accessibility, high-accuracy single-cell whole-genome sequencing, and chromatin interaction studies (28–32). For RNA-seq, the RNA transcripts have to undergo reverse transcription and second-strand synthesis, before the Tn5 tagmentation of the resulting dsDNA (33, 34).

In this paper we present a RNA-seq method using Tn5 transposase to directly tagment RNA/DNA hybrids to form an amplifiable library. We experimentally show that, as an RNase H superfamily member (35), Tn5 binds to RNA/DNA heteroduplex similarly as to dsDNA and effectively fragments and then ligates the specific amplification and sequencing adaptor onto the hybrid. This method, named Sequencing HEteRo RNA-DNA-hYbrid (SHERRY), greatly improves the robustness of low-input RNA-seq with a simplified experimental procedure. We also show that SHERRY works with various amounts of input sample, from single cells to bulk RNA, with a dynamic range spanning six orders of magnitude. SHERRY shows superior cross-sample robustness and comparable detectability for both bulk RNA and single cells compared with other commonly used methods and provides a unique solution for small bulk samples that existing approaches struggle to handle. Furthermore, this easy-to-operate protocol is scalable and cost-effective, holding promise for use in high-quality and high-throughput RNA-seq applications.

Results

RNA-Seq Strategy Using RNA/DNA Hybrid Tagmentation. Because of its nucleotidyl transfer activity, transposase Tn5 has been widely used in recently developed DNA sequencing technologies. Previous studies (36, 37) have identified a catalytic site within its DDE domain (*SI Appendix*, Fig. S1*A*). Indeed, when we mutated one of the key residues (D188E) (38) in pTXB1 Tn5, its fragmentation activity on dsDNA was notably impaired (*SI Appendix*, Fig. S1 *A* and *B*). Increased amounts of the mutated enzyme showed no improvement in tagmentation, verifying the important role of the DDE domain in Tn5 tagmentation. Tn5 is a member of the ribonuclease H-like (RNHL) superfamily, together with RNase H and Moloney Murine Leukemia Virus (MMLV) reverse transcriptase (35, 39, 40); therefore, we hypothesized that Tn5 is capable of processing not only dsDNA but also RNA/DNA heteroduplex. Sequence alignment between these three proteins revealed a conserved domain (two Asps and one Glu) within their active sites, termed the RNase H-like domain (Fig. 1*A*). The two Asp residues (D97 and D188) in the Tn5 catalytic core were structurally similar to those of the other two enzymes (Fig. 1*B*). Moreover, divalent ions, which are important for stabilizing substrate and catalyzing reactions, also occupy similar positions in all three proteins (Fig. 1*B*) (39). We determined the nucleic acid substrate binding pocket of Tn5 according to charge distribution. We then docked double-stranded DNA and RNA/DNA heteroduplex into this predicted pocket and showed that the binding site had enough space

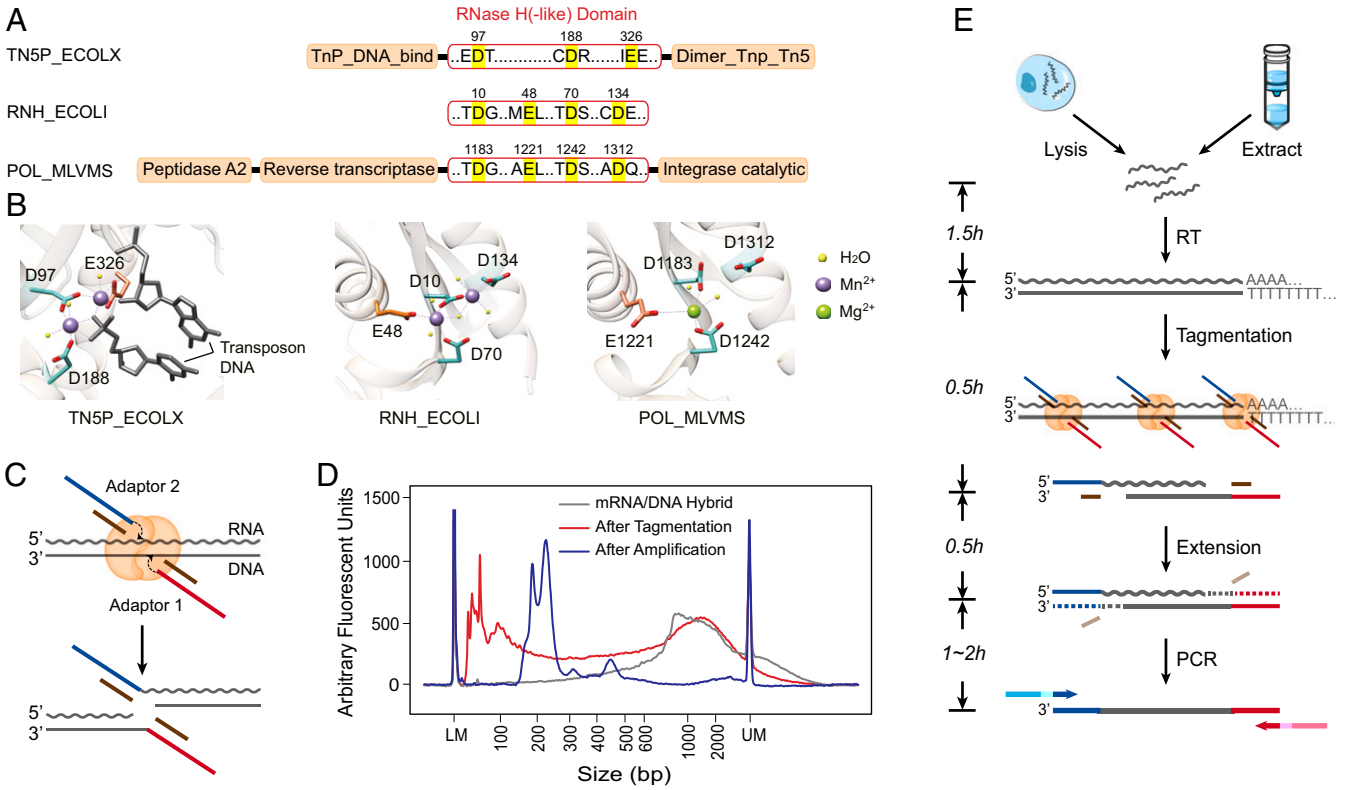

Fig. 1. Tn5 tagmentation activity on double-stranded hybrids and the experimental process of SHERRY. (*A*) RNase H-like (RNHL) domain alignment of Tn5 (TN5P_ECOLX), RNase H (RNH_ECOLI), and MMLV reverse transcriptase (POL_MLVMS). Active residues in the RNHL domains are labeled in bright yellow. Orange boxes represent other domains. (*B*) Superposition of the RNHL active sites in these three enzymes. Protein Data Bank IDs are 1G15 (RNase H), 2HB5 (MMLV), and 1MUS (Tn5). (*C*) Putative mechanism of Tn5 tagmentation of a RNA/DNA hybrid. Crooked arrows represent nucleophilic attacks. (*D*) Size distribution of mRNA/DNA hybrids with and without Tn5 tagmentation and after amplification with index primers. (*E*) Workflow of sequencing library preparation by SHERRY. The input can be a lysed single cell or extracted bulk RNA. After reverse transcription with oligo-dT primer, the hybrid is directly tagmented by Tn5, followed by gap-repair and enrichment PCR. Wavy and straight gray lines represent RNA and DNA, respectively. Dotted lines represent the track of extension step.

Downloaded by guest on January 30, 2020

for an RNA/DNA duplex (*SI Appendix*, Fig. S1*C*). These structural similarities among Tn5, RNase H, and MMLV reverse transcriptase and the docking results further supported the possibility that Tn5 can catalyze the strand transfer reaction on RNA/DNA heteroduplex (Fig. 1*C*).

To validate our hypothesis, we purified Tn5 using the pTXB1 plasmid and corresponding protocol. We prepared RNA/DNA hybrids using mRNA extracted from HEK293T cells. Using a typical dsDNA tagmentation protocol, we treated 15 ng of RNA/DNA hybrids with 0.6 μL pTXB1 Tn5 transposome. Fragment analysis of the tagmented RNA/DNA hybrids showed an obvious decrease (~1000 bp) in fragment size compared with that of untreated control, validating the capability of Tn5 to fragment the hybrid (Fig. 1*D*).

Based on the ability of the Tn5 transposome to fragment RNA/DNA heteroduplexes, we propose SHERRY (Sequencing HEteRo RNA-DNA-hYbrid), a rapid RNA-seq library construction method (Fig. 1*E*). SHERRY consists of three components: RNA reverse transcription, RNA/cDNA hybrid tagmentation, and PCR amplification. The resulting product is an indexed library that is ready for sequencing. Specifically, mRNA is reverse transcribed into RNA/cDNA hybrids using d(T)30VN primer. The hybrid is then tagmented by the pTXB1 Tn5 transposome, the adding of partial sequencing adaptors to fragment ends. DNA polymerase then amplifies the cDNA into a sequencing library after initial end extension. The whole workflow only takes approximately 4 h with hands-on time less than 30 min.

To test SHERRY feasibility, we gap-repaired the RNA/DNA tagmentation products illustrated in Fig. 1*D* (red line) and then amplified these fragments with library construction primers. Amplified molecules (Fig. 1*D*, blue line) were ~100 to ~150 bp longer than the tagmentation products, which matched the extra length of adaptors added by gap-repair and index primer amplification. (*SI Appendix*, Fig. S2). Thus, direct Tn5 tagmentation of RNA/DNA hybrids offers a strategy for RNA-seq library preparation.

Tn5 Has Ligation Activity on Tagmented RNA/DNA Hybrids. To further investigate the detailed molecular events of RNA/DNA hybrid tagmentation, we designed a series of verification experiments. First, we wanted to verify that the transposon adaptor can be ligated to the end of fragmented RNA (Fig. 2*A*). In brief, we prepared RNA/DNA hybrids from HEK293T RNA by reverse transcription. After tagmentation with the Tn5 transposome, we

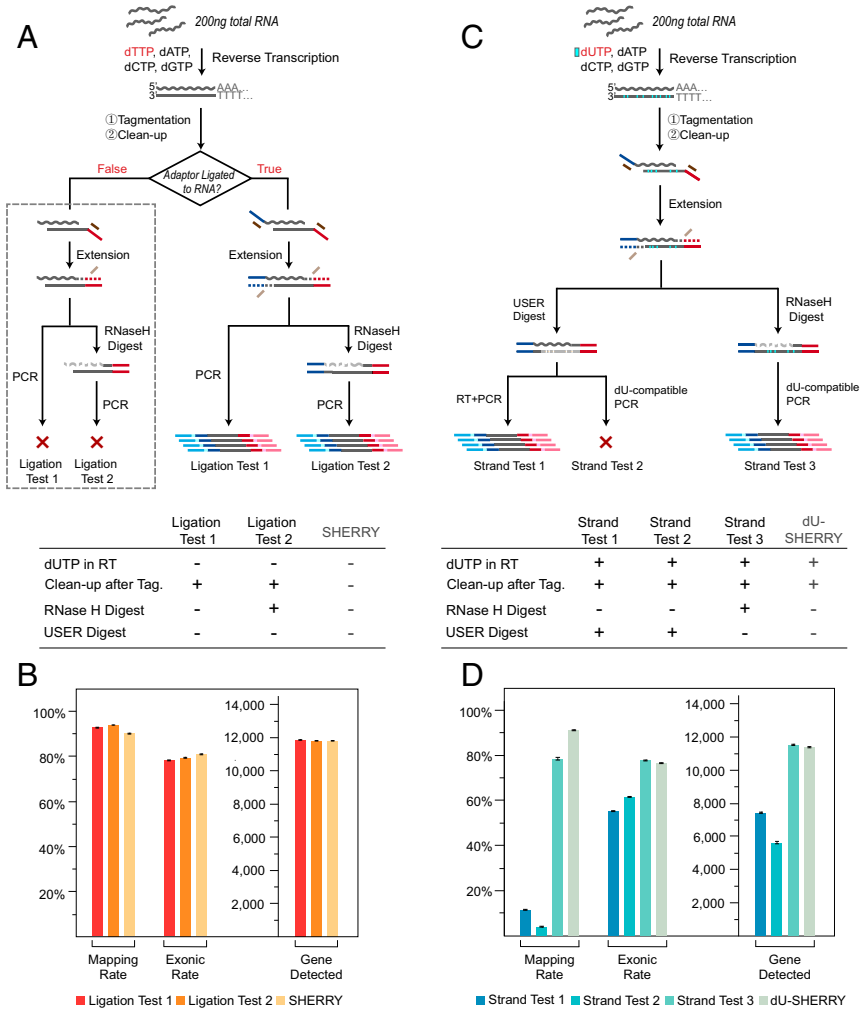

Fig. 2. Verification of Tn5 tagmentation of RNA/DNA heteroduplexes. (*A*) Procedures of two ligation tests. Gray dotted box indicates negative results. The table below lists key experimental parameters that are different from standard SHERRY. (*B*) Comparison of two ligation tests and standard SHERRY with respect to mapping rate, exonic rate, and number of genes detected. Each test consisted of two replicates of 200 ng HEK293T total RNA. (*C*) Strand test procedures. (*D*) Comparison among three strand tests and dU-SHERRY with respect to mapping rate, exonic rate, and number of genes detected. Each test consisted of two replicates of 200 ng HEK293T total RNA.

Downloaded by guest on January 30, 2020

purified the products to remove Tn5 proteins and free adaptors. We assumed that Tn5 ligated the adaptor to the fragmented DNA. At the same time, if Tn5 ligated the adaptor (Fig. 2*A*, dark blue) to the RNA strand, the adaptor could serve as a template in the subsequent extension step. After extension, the DNA strand should have a primer binding site on both 5′ and 3′ ends for PCR amplification. RNase H treatment should not affect production of the sequencing library. If Tn5 failed to ligate the adaptor to the RNA strand, neither strand of the heteroduplex would be converted into a sequencing library.

After PCR amplification, we obtained a high-quantity product regardless of RNase H digestion, indicating successful ligation of the adaptor to the fragmented DNA. Sequencing results from both reaction test conditions as well as from SHERRY showed >90% mapping rate to the human genome with ∼80% exon rate and nearly 12,000 genes detected, validating the transcriptome origin of the library (Fig. 2*B* and *SI Appendix*, Fig. S3*A*). The additional purification step after reverse transcription and/or RNase H digestion before PCR amplification did not affect the results, probably because of the large amount of starting RNA. We examined the sequencing reads with an insert size shorter than 100 bp (shorter than the sequence read length, also called "read through"), and 99.7% of them contained adaptor sequence. Such read-through reads directly proved ligation of the adaptor to the fragmented RNA (*SI Appendix*, Fig. S3 *B* and *C*). In summary, we confirmed that Tn5 transposome can tagment both DNA and RNA strands of RNA/DNA heteroduplexes.

Tagmented cDNA Is the Preferred Amplification Template. Next, we investigated whether RNA and DNA strands could be amplified to form the sequencing library (Fig. 2*C*). We replaced dTTP with dUTP during the reverse transcription and then purified the tagmented products to remove free dUTP and Tn5 proteins. Bst 2.0 WarmStart DNA polymerase was used for extension because it is able to use RNA as a primer and to process the dU bases in the template. The product fragments were then treated with either USER enzyme or RNase H to digest cDNA and RNA, respectively. We performed RT-PCR with the USER-digested product, to test the efficiency of converting tagmented RNA for library construction (Strand Test 1). To exclude interference from undigested DNA, we performed PCR amplification with the USER-digested fragments using dU-compatible polymerase (Strand Test 2). We also used dU-compatible PCR to test the efficiency of converting tagmented cDNA for library construction (Strand Test 3). For comparison, we included a control experiment with the same workflow as Strand Test 3 except that the RNase H digestion step was omitted (dU-SHERRY) to ensure that Tn5 can recognize substrates with dUTP.

Sequencing results of Strand Test 1 showed a low mapping rate and gene detection count that were only slightly higher than those of Strand Test 2. In contrast, Strand Test 3 demonstrated a similar exon rate and gene count to dU-SHERRY and SHERRY (Fig. 2*D* and *SI Appendix*, Fig. S3*A*). Based on these results, we conclude that the tagmented cDNA contributes to the majority of the final sequencing library, likely because of inevitable RNA degradation during the series of reactions.

SHERRY for Rapid One-Step RNA-Seq Library Preparation. We tested different reaction conditions to optimize SHERRY with 10 ng total RNA as input (*SI Appendix*, Fig. S4*A*). We evaluated the impact of different crowding agents, different ribonucleotide modifications on transposon adaptors, and different enzymes for gap filling. We also included purification after certain steps to remove primer dimers and carryover contaminations. Sequencing results showed little change in performance from most of these modifications, indicating that SHERRY is robust under various conditions.

We then compared the optimized SHERRY with NEBNext Ultra II, a commercially available kit, for bulk RNA library preparation. This NEBNext kit is one of the most commonly used kits for RNA-seq experiments, with 10 ng total RNA being its minimum input limit. We therefore tested the RNA-seq performance with 10 and 200 ng total RNA inputs, each condition having three replicates. SHERRY demonstrated comparable performances with NEBNext for both input levels (*SI Appendix*, Fig. S4*B*). For the 10 ng input tests, SHERRY produced more precise gene expression measurements across replicates (Fig. 3*A*), probably because of the simpler SHERRY workflow.

Next, we compared the ability to detect differentially expressed genes between HEK293T and HeLa cells using SHERRY and NEBNext. In all three replicates, SHERRY detected 11,269 genes in HEK293T cells and 10,774 genes in HeLa cells, with high precision (correlation coefficient R2 = 0.999) (Fig. 3*B* and *SI Appendix*, Fig. S5*A*). The numbers of detected genes and their read counts identified by SHERRY and NEBNext were highly concordant (Fig. 3*C* and *SI Appendix*, Fig. S5*B*). This excellent reproducibility of SHERRY ensured the reliability of subsequent analyses. Then we plotted a heatmap of the distance matrix (Fig. 3*D*) between different cell types and library preparation methods. Libraries from the same cell type were clustered together as expected. Libraries from the same method also tended to cluster together, indicating internal bias in both methods.

We then used DESeq2 to detect differentially expressed genes (P value $< 5 \times 10^{-6}$, |log2Fold change| > 1). In general, the thousands of differentially expressed genes detected by both methods were highly similar (*SI Appendix*, Fig. S5*C*) and their expression fold-change was highly correlated (correlation coefficient R2 = 0.977) between SHERRY and NEBNext (Fig. 3*E*). Examination of the genes that showed differential expression in only one method revealed the same trend of expression change in the data from the other method (*SI Appendix*, Fig. S5 *D* and *E*). We conclude that SHERRY provides equally reliable differential gene expression information as NEBNext, but with a much faster and less labor-intensive process, specifically saving around 2 h hand-on time (*SI Appendix*, Fig. S7*A*).

SHERRY Using Trace Amounts of RNA or Single Cells. We next investigated whether SHERRY could construct RNA-seq libraries from single cells. First, we reduced the input to 100 pg total RNA, which is equivalent to RNA from about 10 cells. SHERRY results were high quality, with high mapping and exon rates and nearly 9,000 genes detected (*SI Appendix*, Fig. S6*A*). Seventy-two percent of these genes were detected in all three replicates, demonstrating good reproducibility (Fig. 4*A*). The expression of these genes showed excellent precision with R2 ranging from 0.958 to 0.970. (Fig. 4*B*).

To further push the detection limit, we carried out single-cell SHERRY experiments (scSHERRY) using the HEK293T cell line. In contrast to the experiments with purified RNA, scSHERRY required several optimizations to the standard protocol (Fig. 4*C* and *SI Appendix*, Fig. S6*B*). Although we found no positive effect by replacing betaine in the standard protocol with other crowding agents during optimization, we found that addition of a crowding reagent with a higher molecular weight improved the library quality from single cells. Therefore, we used PEG8000 for the following scSHERRY experiments. For the extension step, the use of Bst 3.0 or Bst 2.0 WarmStart DNA polymerases detected more genes than the use of SuperScript II or SuperScript III reverse transcriptases. This is probably because of the stronger processivity and strand displacement activity of Bst polymerases, and better compatibility with higher reaction temperatures to open the secondary structure of RNA templates. We also tried to optimize the PCR strategy because extensive amplification can lead to strong bias. Compared to the continuous

Downloaded by guest on January 30, 2020

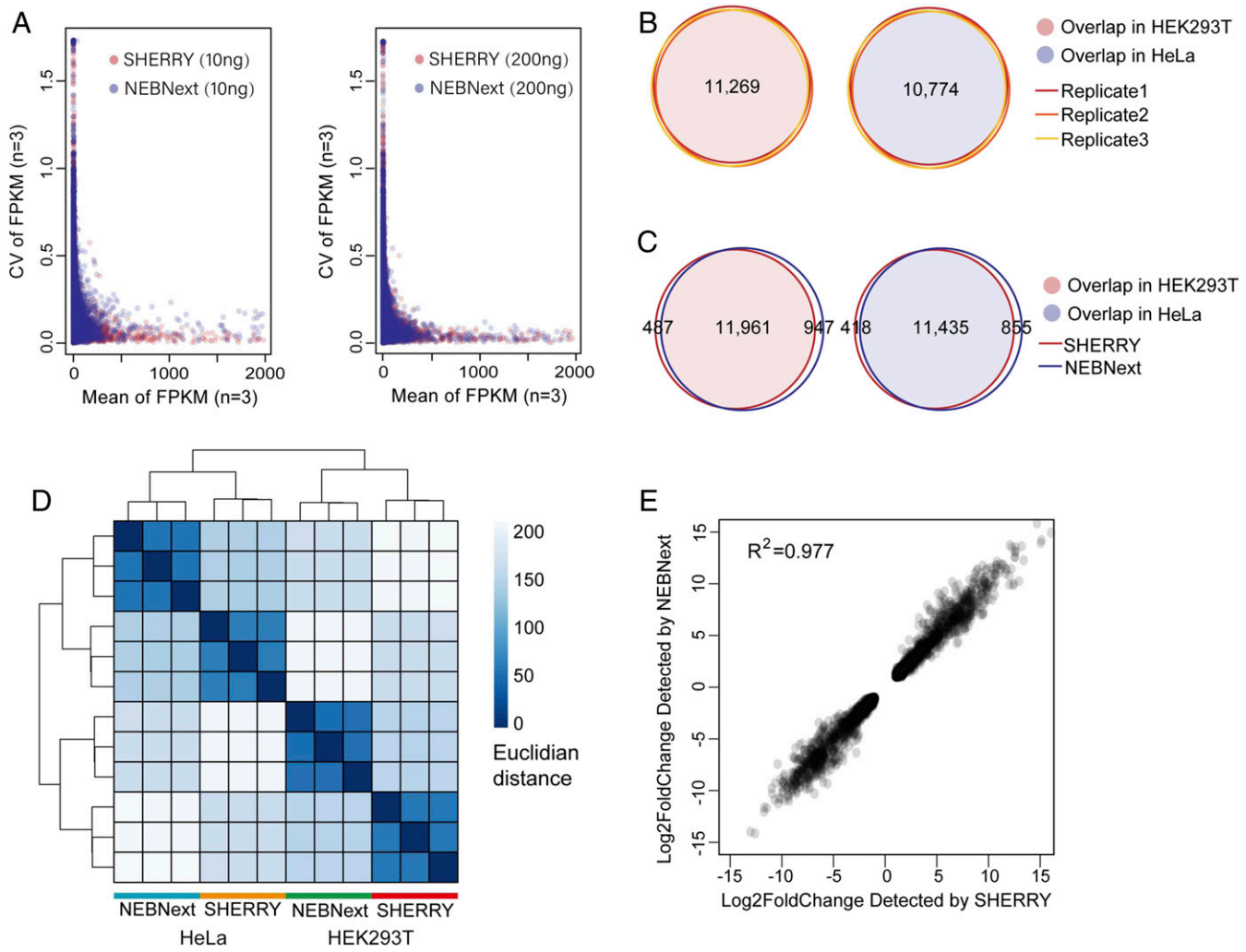

Fig. 3. Performance of SHERRY with large RNA input. (*A*) Coefficient of variation (CV) across three replicates was plotted against the mean value of each gene's FPKM (Fragments Per Kilobase of transcript per Million mapped reads). All experiments used HEK293T total RNA as input. (*B*) Genes detected by SHERRY in three replicates of 200 ng HEK293T or HeLa total RNA are plotted in Venn Diagrams. Numbers of common genes are indicated. (*C*) Common genes detected by SHERRY and NEBNext in the three replicates of 200 ng HEK293T or HeLa total RNA. (*D*) Distance heatmap of samples prepared by SHERRY or NEBNext for three replicates using 200 ng HEK293T or HeLa total RNA. The color bar indicates the Euclidian distance. (*E*) Correlation of gene expression fold-change identified by SHERRY and NEBNext. Involved genes are differentially expressed genes between HEK293T and HeLa detected by both methods.

28-cycle PCR, the incorporation of a purification step after 10 cycles, or simply reducing the total cycle number to 15 increased the mapping rate and the number of genes detected. Therefore, we performed 15-cycle PCR without extra purification to better accommodate high-throughput experiments.

The optimized scSHERRY was capable of detecting 8,338 genes with a 50.17% mapping rate (*SI Appendix*, Fig. S6*D*), and the gene read counts correlated well (correlation coefficient R2 = 0.600) with Smart-seq2, the most prevalent protocol in the single-cell RNA amplification field. Besides, scSHERRY showed better reproducibility than Smart-seq2 (Fig. 4*D*). Compared with Smart-seq2, the gene number and coverage uniformity of the scSHERRY-generated library was slightly inferior (*SI Appendix*, Fig. S6 *D* and *E*) because Smart-seq2 enriches for full-length transcripts via a preamplification step. However, this enrichment step of Smart-seq2 also introduced bias (Fig. 4*E*). We used 200 ng of HEK293T total RNA to construct a sequencing library using the NEBNext kit, expecting to capture as many genes as possible. Besides, the NEBNext protocol used short RNA fragments for reverse transcription and fewer cycles of PCR amplification, which should introduce less GC bias than the other protocols. We then compared the GC distribution of genes detected by scSHERRY or Smart-seq2 with NEBNext results. scSHERRY, which is free from second-strand synthesis and preamplification, produced a distribution similar to the standard. However, the library from Smart-seq2-amplified single-cell RNA (scRNA) showed clear enrichment for genes with lower GC content. Genes with high GC content were less likely to be captured by Smart-seq2, which may cause biased quantitation results. Overall, compared with Smart-seq2, scSHERRY produced libraries of comparable quality and lower GC bias. Moreover, the scSHERRY workflow spares preamplification and QC steps before tagmentation, saving around 4 h (*SI Appendix*, Fig. S7*A*), and the one tube strategy is promising for high-throughput application.

Discussion

We found that the Tn5 transposome has the capability to directly fragment and tag RNA/DNA heteroduplexes and, therefore, we have developed a quick RNA amplification and library

Downloaded by guest on January 30, 2020

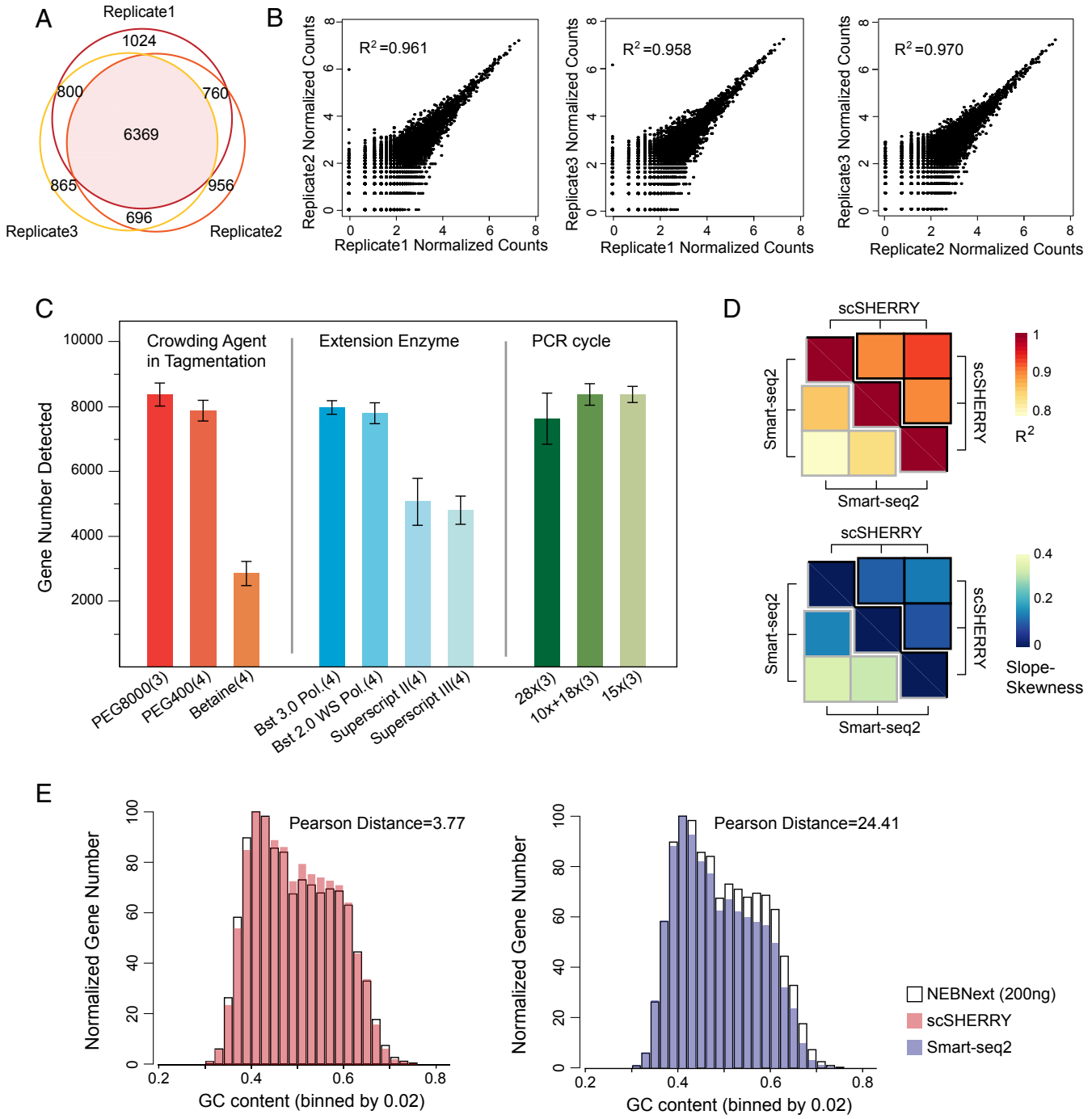

Fig. 4. Performance of SHERRY with microinput samples. (*A*) Genes detected by SHERRY in three replicates with 100 pg HEK293T total RNA. (*B*) Correlations of normalized gene read counts between replicates with 100 pg HEK293T total RNA. (*C*) Gene number detected by scSHERRY under various experimental conditions in single HEK293T cells. Each condition involved three to four replicates. (*D*) The heatmap of R2 calculated from scSHERRY and Smart-seq2 replicates, and slope deviation in a linear fitting equation for the two methods. (*E*) Normalized gene numbers with different GC content.

preparation method called SHERRY. The input for SHERRY could be RNA from single-cell lysate or total RNA extracted from a large number of cells. Comparison of SHERRY with the commonly used Smart-seq2 protocol for single-cell input or the NEBNext kit for bulk total RNA input, showed comparable performance for input amount spanning more than five orders of magnitude. Furthermore, the whole SHERRY workflow from RNA to sequencing library consists of only five steps in one tube and takes about 4 h, with hands-on time of less than 30 min. Smart-seq2, requires twice this amount of time and an additional library preparation step is necessary. The 10-step NEBNext protocol is much more labor-intensive and time-consuming (*SI Appendix*, Fig. S7*A*). Moreover, the SHERRY reagent cost is fivefold less compared with that of the other two methods (*SI Appendix*, Fig. S7*B*). For single cells, the lower mapping rate of SHERRY compared to Smart-seq2 could increase sequencing cost. However, both methods reached a plateau of saturation curve with 2 million total reads (*SI Appendix*, Fig. S8), which

Downloaded by guest on January 30, 2020

costs less than $5. Therefore, SHERRY has strong competitive advantages over conventional RNA library preparation methods and scRNA amplification methods.

In our previous experiments, we assembled a Tn5 transposome using home-purified pTXB1 Tn5 and synthesized sequencer-adapted oligos. To generalize the SHERRY method and to confirm Tn5 tagmentation of RNA/DNA heteroduplexes, we tested two commercially available Tn5 transposomes, Amplicon Tagment Mix (abbreviated as ATM) from the Nextera XT kit (Illumina) and TruePrep Tagment Enzyme (TTE) Mix V50 (abbreviated as V50) from the TruPrep kit (Vazyme). We normalized the different Tn5 transposome sources according to the enzyme processing 5 ng of genomic DNA to the same size under the same reaction conditions (*SI Appendix*, Fig. S9*A*). The tagmentation activity of our in-house pTXB1 Tn5 was 10-fold higher than V50 and 500-fold higher than ATM when using transposome volume as the metric. The same units of enzyme were then used to process RNA/DNA heteroduplexes prepared from 5 ng mRNA to confirm similar performance on such hybrids (*SI Appendix*, Fig. S9*B*). The RNA-seq libraries from all three enzymes showed consistent results, demonstrating the robustness of SHERRY (*SI Appendix*, Fig. S9*C*).

During DNA and RNA/DNA heteroduplex tagmentation, the Tn5 transposome reacted with these two substrates in different patterns. We tagmented 5 ng DNA or mRNA/DNA hybrids with 0.02, 0.05, or 0.2 μL pTXB1 Tn5 transposome (*SI Appendix*, Fig. S10). As the amount of Tn5 increased, dsDNA was cut into overall shorter fragments. While for the hybrid, Tn5 cut the template "one by one" because only hybrids above a certain size became shorter and most were too short to be cut. We supposed that such phenomenon might attribute to the different conformation of dsDNA and RNA/DNA hybrid, since diameter of the latter is larger. Thus, the binding pocket or catalytic site of Tn5 would be tuned to accommodate the hybrid strands and cause different tagmentation pattern.

Despite its ease-of-use and commercial promise, the library quality produced by SHERRY may be limited by unevenness of transcript coverage. Unlike the NEBNext kit, which fragments RNA before reverse transcription, or Smart-seq2, which performs preamplification to enrich full-length cDNAs, SHERRY simply reverse transcribes full-length RNA. Reverse transcriptase is well known for its low efficiency and, when using polyT as the primer for extension, it is difficult for the transcriptase to reach the 5′ end of the RNA template. This can cause coverage imbalance across transcripts, making the RNA-seq signal biased toward the 3′ end of genes (*SI Appendix*, Fig. S11*A*). In an attempt to solve this problem, we added template-switching oligo primer, the sequence and concentration of which was the same as Smart-seq2 protocol (17), to the reverse transcription buffer to mimic the Smart-seq2 reverse transcription conditions. The resulting hybrid was then tagmented and amplified following standard SHERRY workflow. This produced much improved evenness across transcripts (*SI Appendix*, Fig. S11 *A* and *B*), although some of the sequencing parameters dropped accordingly (*SI Appendix*, Fig. S11*C*). We believe that with continued optimization, SHERRY will improve RNA-seq performance.

Materials and Methods

Purification of pTXB1 Tn5 and D188E Mutation. The pTXB1 cloning vector, which introduced hyperactive E54K and L372P mutation into wildtype Tn5, was acquired from Addgene. The pTXB1 Tn5 and its mutant were expressed and purified mainly according to the protocol published by Picelli et al. (41)

Tn5 Transposome Tagmentation. As for RNA/DNA hybrid, tagmentation was performed in buffer containing 10 mM Tris-Cl (pH 7.6), 5 mM MgCl2, 10% N,*N*-Dimethylformamide, 9% PEG8000 (VWR Life Science, Cat. No. 97061), 0.85 mM adenosine 5′-triphosphate (ATP; NEB, Cat. No. P0756). In SHERRY library preparation, we used 0.05, 0.006, and 0.003 μL Tn5 transposome for input of 200 ng, 10 ng, and 100 pg total RNA, respectively. scSHERRY also used 0.003 μL Tn5 transposome.

The transposome could be diluted in 1× Tn5 dialysis buffer [50 mM Hepes (pH 7.2, Leagene, Cat. No. CC064), 0.1 M NaCl (Invitrogen, Cat. No. AM9759), 0.1 mM ethylenediaminetetraacetic acid (Invitrogen, Cat. No. AM9260G), 1 mM dithiothreitol, 0.1% Triton X-100, 10% glycerol]. The reaction was incubated at 55 °C for 30 min.

SHERRY Library Preparation and Sequencing. As for purified 10 or 200 ng total RNA input, the tagmentation product was firstly gap-filled with 100 units of SuperScript II and 1 × Q5 High-Fidelity Master Mix at 42 °C for 15 min, then SuperScript II was inactivated at 70 °C for 15 min. When inputting 100 pg total RNA, the extension enzyme was replaced with 4 units of Bst 2.0 WarmStart DNA Polymerase (NEB, Cat. No. M0538). Correspondingly, the reaction temperature was up-regulated to 72 °C and inactivation was performed at 80 °C for 20 min. After that, indexed common primers were added to perform PCR. We performed 12, 15, and 25 cycles of PCR for input of 200 ng, 10 ng, and 100 pg total RNA, respectively.

The resulting library was purified with 1:1 ratio by VAHTS DNA Clean Beads. Quantification was done by Qubit 2.0 and quality check was done by Fragment Analyzer Automated CE System. The sequencing platform we used was Illumina NextSEq 500 or HiSEq 4000.

A complete description of the materials and methods is provided in *SI Appendix, SI Materials and Methods*. The sequence reported in this paper has been deposited in the Genome Sequence Archive (accession no. CRA002081).

ACKNOWLEDGMENTS. We thank Dr. Yun Zhang and BIOPIC sequencing platform at Peking University for the assistance of high-throughput sequencing experiments. This work was supported by National Natural Science Foundation of China (21675098, 21525521), Ministry of Science and Technology of China (2018YFA0800200, 2018YFA0108100, 2018YFC1002300), 2018 Beijing Brain Initiative (Z181100001518004), Beijing Advanced Innovation Center for Structural Biology, and Beijing Advanced Innovation Center for Genomics.

1. Y. Liu *et al*., Peptidylarginine deiminases 2 and 4 modulate innate and adaptive immune responses in TLR-7-dependent lupus. *JCI Insight* **3**, e124729 (2018).
2. E. Schmidt *et al*., LincRNA H19 protects from dietary obesity by constraining expression of monoallelic genes in brown fat. *Nat. Commun.* **9**, 3622 (2018).
3. O. Wurtzel *et al*., A single-base resolution map of an archaeal transcriptome. *Genome Res.* **20**, 133–141 (2010).
4. A. A. Penin, A. V. Klepikova, A. S. Kasianov, E. S. Gerasimov, M. D. Logacheva, Comparative analysis of developmental transcriptome maps of Arabidopsis thaliana and Solanum lycopersicum. *Genes (Basel)* **10**, E50 (2019).
5. P. Civita *et al*., Laser capture microdissection and RNA-seq analysis: High sensitivity approaches to explain histopathological heterogeneity in human glioblastoma FFPE archived tissues. *Front. Oncol.* **9**, 482 (2019).
6. D. A. Jaitin *et al*., Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
7. Y. Chen *et al*., Single-cell RNA-seq uncovers dynamic processes and critical regulators in mouse spermatogenesis. *Cell Res.* **28**, 879–896 (2018).
8. M. Wang *et al*., Single-cell RNA sequencing analysis reveals sequential cell fate transition during human spermatogenesis. *Cell Stem Cell* **23**, 599–614.e4 (2018).
9. H. Hochgerner, A. Zeisel, P. Lönnerberg, S. Linnarsson, Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. *Nat. Neurosci.* **21**, 290–299 (2018).
10. U. Gubler, B. J. Hoffman, A simple and very efficient method for generating cDNA libraries. *Gene* **25**, 263–269 (1983).
11. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
12. P. Cui *et al*., A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics* **96**, 259–265 (2010).
13. N. Cloonan *et al*., Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5**, 613–619 (2008).
14. C. D. Armour *et al*., Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nat. Methods* **6**, 647–649 (2009).
15. D. Parkhomchuk *et al*., Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* **37**, e123 (2009).
16. F. Tang *et al*., mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
17. S. Picelli *et al*., Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
18. T. Hashimshony *et al*., CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-seq. *Genome Biol.* **17**, 77 (2016).
19. Y. Fu, H. Chen, L. Liu, Y. Huang, Single cell total RNA sequencing through isothermal amplification in picoliter-droplet emulsion. *Anal. Chem.* **88**, 10795–10799 (2016).

Downloaded by guest on January 30, 2020

20. G. X. Zheng *et al*., Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
21. T. M. Gierahn *et al*., Seq-Well: Portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395–398 (2017).
22. J. Cao *et al*., Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
23. A. B. Rosenberg *et al*., Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).
24. P. See, J. Lum, J. Chen, F. Ginhoux, A single-cell sequencing guide for immunologists. *Front. Immunol.* **9**, 2425 (2018).
25. I. Y. Goryshin, W. S. Reznikoff, Tn5 in vitro transposition. *J. Biol. Chem.* **273**, 7367–7374 (1998).
26. A. Adey *et al*., Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.* **11**, R119 (2010).
27. S. Picelli *et al*., Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).
28. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
29. C. Chen *et al*., Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science* **356**, 189–194 (2017).
30. S. Rohrback *et al*., Submegabase copy number variations arise during cerebral cortical neurogenesis as revealed by single-cell whole-genome sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 10804–10809 (2018).
31. L. Tan, D. Xing, C. H. Chang, H. Li, X. S. Xie, Three-dimensional genome structures of single diploid human cells. *Science* **361**, 924–928 (2018).
32. B. Lai *et al*., Trac-looping measures genome structure and chromatin accessibility. *Nat. Methods* **15**, 741–747 (2018).
33. J. Gertz *et al*., Transposase mediated construction of RNA-seq libraries. *Genome Res.* **22**, 134–141 (2012).
34. S. Brouilette *et al*., A simple and novel method for RNA-seq library preparation of single cell cDNA analysis by hyperactive Tn5 transposase. *Dev. Dyn.* **241**, 1584–1590 (2012).
35. K. A. Majorek *et al*., The RNase H-like superfamily: New members, comparative structural analysis and evolutionary classification. *Nucleic Acids Res.* **42**, 4160–4179 (2014).
36. D. R. Davies, I. Y. Goryshin, W. S. Reznikoff, I. Rayment, Three-dimensional structure of the Tn5 synaptic complex transposition intermediate. *Science* **289**, 77–85 (2000).
37. W. S. Reznikoff, Transposon Tn5. *Annu. Rev. Genet.* **42**, 269–286 (2008).
38. G. Peterson, W. Reznikoff, Tn5 transposase active site mutations suggest position of donor backbone DNA in synaptic complex. *J. Biol. Chem.* **278**, 1904–1909 (2003).
39. M. Nowotny, S. A. Gaidamakov, R. J. Crouch, W. Yang, Crystal structures of RNase H bound to an RNA/DNA hybrid: Substrate specificity and metal-dependent catalysis. *Cell* **121**, 1005–1016 (2005).
40. D. Lim *et al*., Crystal structure of the moloney murine leukemia virus RNase H domain. *J. Virol.* **80**, 8379–8389 (2006).
41. S. Picelli *et al*., Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).

Downloaded by guest on January 30, 2020