RNA Sequencing by Direct Tagmentation of RNA/DNA Hybrids

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

[a] Beijing Advanced Innovation Center for Genomics (ICG), Biomedical Pioneering Innovation Center (BIOPIC), School of Life Sciences, and Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China.
[b] School of Life Sciences, and Tsinghua-Peking Center for Life Sciences, Tsinghua University, Beijing 100084, China.
[c] School of Chemistry and Materials Science, Nanjing Normal University, Nanjing, Jiangsu Province, China
[d] Institute for Cell Analysis, Shenzhen Bay Laboratory, Shenzhen, Guangdong Province, China
[e] XGen US Co, South San Francisco, CA
[f] Department of Cellular and Molecular Medicine, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA 92093
[g] College of Engineering, Peking University, Beijing 100871, China

[#] These authors contributed equally to this work.
[1] Present address: Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA
* Corresponding authors: jianbinwang@tsinghua.edu.cn (J.W.), yanyi@pku.edu.cn (Y.H.) and sunneyxie@pku.edu.cn (X.S.X.)

Abstract

Transcriptome profiling by RNA sequencing (RNA-seq) has been widely used to characterize cellular status, but solely relies on second strand 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 analysis, to construct RNA-seq libraries without second strand synthesis. We discover that Tn5 transposome can randomly target the RNA/DNA heteroduplex 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 even single cells. SHERRY greatly reduces experimental difficulty, with higher reproducibility and GC uniformity than prevailing RNA-seq methods.

Introduction

Transcriptome profiling through RNA-seq has become a routine analysis in biomedicine since the popularization of next-generation sequencers, along with the dramatical decrease of the sequencing cost. RNA-seq has been widely used to facilitate quantitative analysis of various biological questions, from exploring pathogenesis of disease [1,2] to building up the transcriptome maps for various species [3,4]. RNA-seq provides informative assessments of samples, especially those that are related to heterogeneity in a complex biological system [5,6] or to time-dependent dynamic progresses [7-9]. A typical RNA-seq experiment requires a DNA library built upon the 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 cDNA synthesis, adapter ligation, and PCR amplification. [10-12]

Although many experimental protocols, combining new chemistries and processes, have been invented recently, we still face many challenges while performing RNA-seq. 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 using small amounts of RNA as inputs, for which most large-input protocols cannot work. The main reason of this incompatibility is due to the

purification operations needed between major experimental steps, making the whole process labor intensive and causing 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 such protocols it is difficult to achieve both high throughput and high detectability. One type of single-cell RNA-seq approaches, such as Smart-seq2 [17], is to introduce pre-amplification in the protocol to address the low-input problem but such an approach will probably introduce bias impairing quantification accuracy. Another type of approaches is to barcode each cell's transcripts and hence bioinformatically assign the sequencing data with the identity linked to each cell and each molecule [20-23], but the detectability and reproducibility are still not ideal for such approaches. We need an easy and versatile RNA-seq method that can cover input from single cells to bulk RNA.

Bacterial transposase Tn5 [24] has been introduced into 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 process and low sample input requirement [25,26]. For general library preparation, Tn5 directly reacts with naked dsDNA, followed by PCR amplification with sequencing adaptors. Such a simple one-step tagmentation reaction has been proved to greatly simplify the experimental process, shorten the time and lower the cost. Tn5-tagmentation has also been used for chromatin accessibility detection, high-accuracy single-cell whole-genome sequencing, and chromatin interaction studies [27-31]. For RNA-seq, the RNA transcripts have to undergo reverse transcription and second strand synthesis, before the Tn5 tagmentation of the resulting dsDNA [32,33].

In this paper we present a novel RNA-seq method using Tn5 transposase to directly tagment RNA/DNA hybrid to form an amplifiable library. We experimentally proved that, as an RNase H superfamily member [34], Tn5 can bind RNA/DNA heteroduplex similarly as to dsDNA, and effectively fragment and ligate the specific amplification and

sequencing adaptor onto the hybrid. This method, named Sequencing HEteRo RNA-DNA-hYbrid (SHERRY), can greatly improve the robustness of low-input RNA-seq with a simplified experimental procedure. We also showed that SHERRY can cover various amount of input samples, from single cells to bulk RNA, with a dynamic range spanning six orders of magnitude. SHERRY shows superior cross-sample robustness and comparable detectability to other prevalent methods for both bulk RNA and single cells and provides a unique solution for small bulk samples that existing approaches are struggling to handle. Furthermore, such easy-to-operate protocol is scalable and cost effective, holding promise for high-quality and high-throughput RNA-seq applications.

Results

A new RNA-seq strategy by RNA/DNA hybrid tagmentation. Because of its nucleotidyl transfer activity, transposase Tn5 has been widely used in various new DNA sequencing technologies. Previous studies [35,36] have identified its catalytic site within the DDE domain (**Fig. S1A)**. Indeed, when we mutated one of the key residues (D188E) [37] in the pTXB1 Tn5, we found drastic impairing of its fragmentation activity on dsDNA was notably impaired (**Fig. S1A, B**). An increase in the amount of enzyme used showed no improvement of tagmentation, verifying the important role of DDE domain in Tn5 tagmentation. Given the fact that Tn5 is a member of the ribonuclease H-like (RNHL) superfamily, together with RNase H and MMLV reverse transcriptase [34,38,39], we hypothesized that Tn5 is capable of processing not only dsDNA but also RNA/DNA heteroduplex. Sequence alignment between the three proteins revealed a conserved domain within the active sites, noted as RNase H like domain (**Fig. 1A**). Particularly, the two Asps (D97 and D188) in Tn5 catalytical core were structurally similar to the other two enzymes (**Fig. 1B**). Moreover, the divalent ions, which were important for stabilizing substrate and catalyzing reactions, also occupied similar positions in all of them (**Fig. 1B**) [38]. Furthermore, we worked out nucleic acid substrate binding pocket of Tn5 according to charge distribution and tried docking double-stranded DNA and RNA/DNA heteroduplex in this predicted pocket. It turned out that the binding site had enough space for RNA/DNA duplex (**Fig. S1C**). These structural similarities among Tn5, RNase H and MMLV reverse transcriptase and docking results further supported the possibility that Tn5

could catalyze the strand transfer reaction on RNA/DNA heteroduplex (**Fig. 1C**).

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

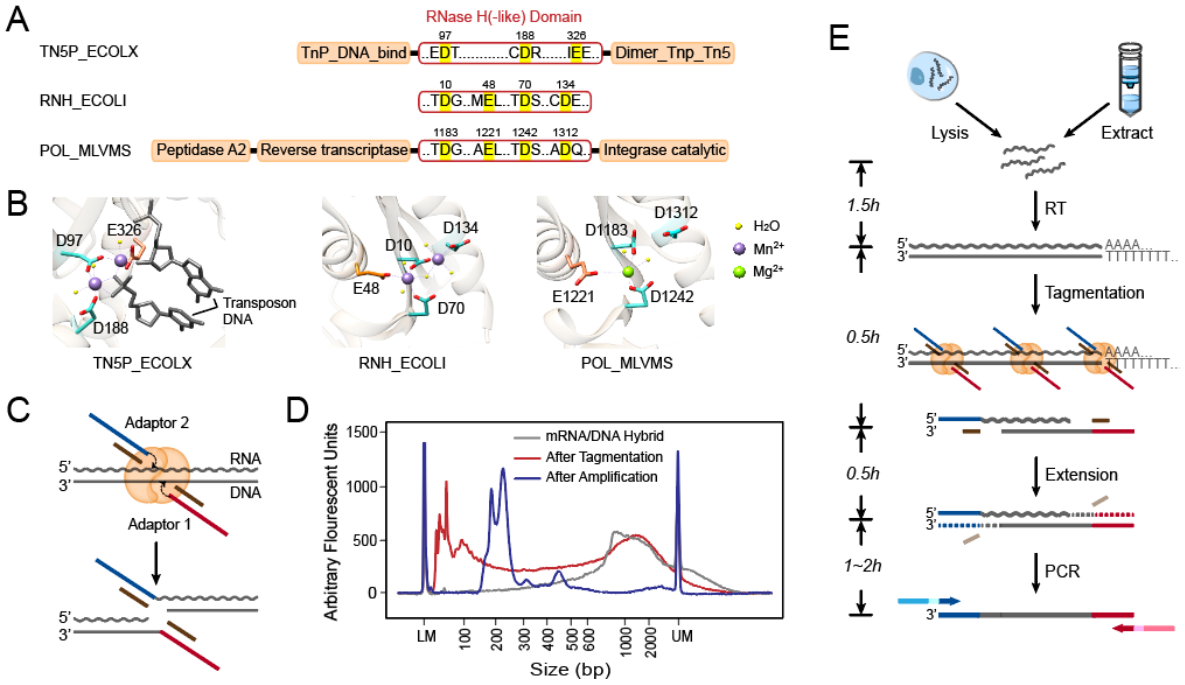


Fig.1 Proof of Tn5 tagmentation activity on double-stranded hybrids and the experimental process of SHERRY. (**A**) RNase H-like (RNHL) domain alignment among Tn5 (TN5P_ECOLX), RNase H (RNH_ECOLI) and M-MLV reverse transcriptase (POL_MLVMS). Active residues in RNHL domain are labeled in bright yellow. Orange boxes represent other domains in enzymes. (**B**) Superposition of RNHL active sites in these three enzymes. PDB IDs are 1G15 (RNase H), 2HB5 (M-MLV), 1MUS (Tn5), respectively. (**C**) Putative mechanism of Tn5 tagmentation on RNA/DNA hybrid. Arrows represent nucleophilic attacks. (**D**) Size distribution of mRNA/DNA hybrid with and without Tn5 tagmentation, and amplified after tagmentaion with index primers. (**E**) Workflow of sequencing library preparation through SHERRY. The input can be lysed single cell or extracted bulk RNA. After reverse transcription by oligo-dT primer, the hybrid is directly tagmented by Tn5, then proceeds to gap-repair and enrichment PCR. Gray wave and straight lines represent RNA and DNA, respectively.

Based on the fragmentation activity of Tn5 transposome on RNA/DNA heteroduplex, we

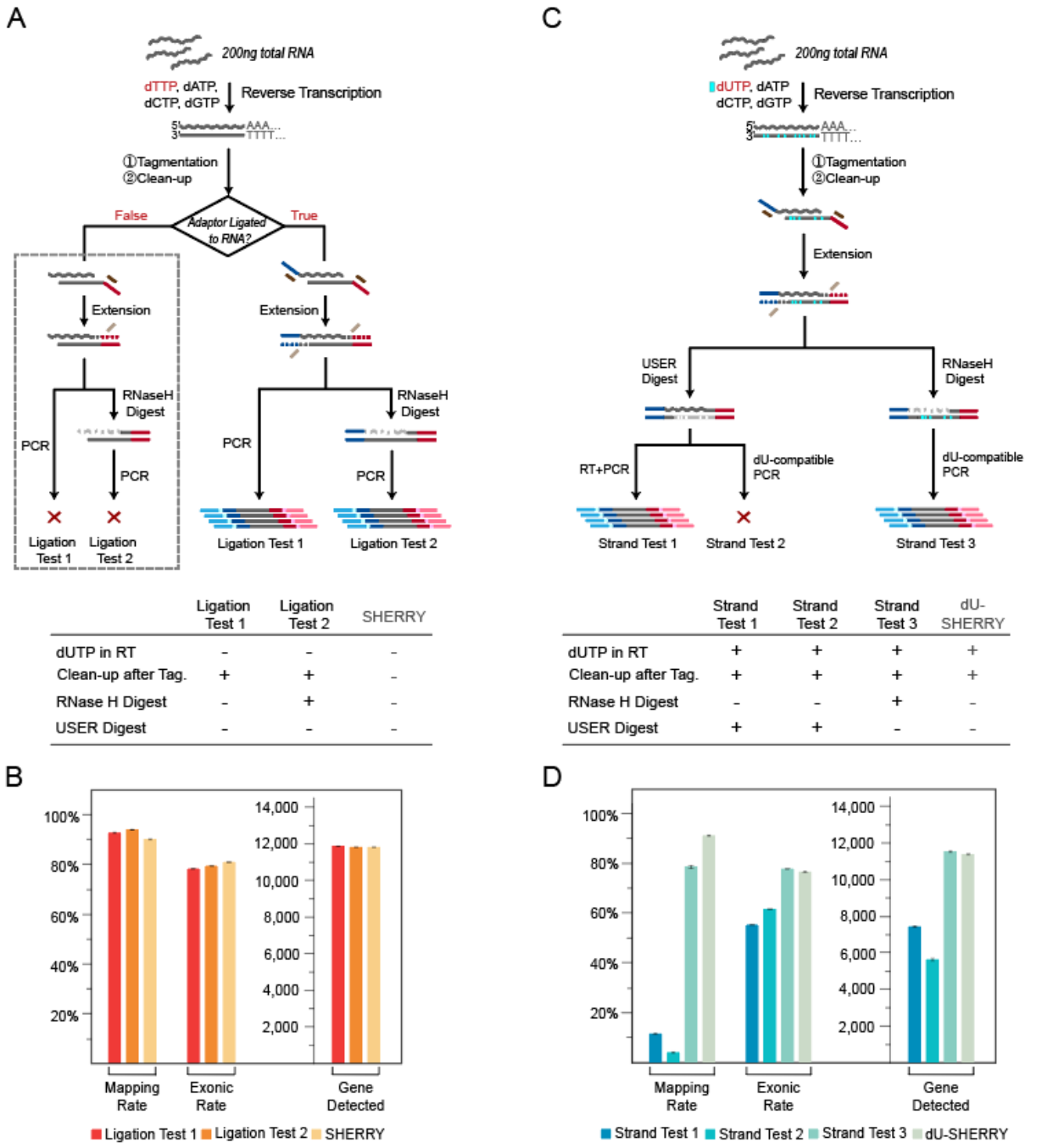
proposed SHERRY (Sequencing HEteRo RNA-DNA-hYbrid), a rapid RNA-seq library construction method (**Fig. 1E**). SHERRY consists of three experimental components: RNA reverse transcription, RNA/cDNA hybrid tagmentation, and PCR amplification. The resulting product is indexed library that is ready for sequencing. Specifically, mRNA is reverse transcribed into RNA/cDNA hybrid using $d(T)_{30}VN$ primer. The hybrid is then tagmented by pTXB1 Tn5 transposome, adding partial sequencing adaptors to fragment ends. DNA polymerase can be used to amplify the cDNA into a sequencing library after initial end extension. The whole workflow just takes around four hours.

To test the feasibility, we gap-repaired the RNA/DNA tagmentation products in **Fig. 1D** and amplified those fragments with library construction primers. We found that the size of amplified molecules was about 130bp longer than tagmentation product, which perfectly corresponded to the extended length of the indexed common primers. (**Fig. S2**). Thus, direct Tn5 tagmentation of RNA/DNA hybrid offers a new strategy for RNA-seq library preparation.

Tn5 has ligation activity on tagmented RNA/DNA hybrid. To better investigate the detailed molecular events during tagmentation of RNA/DNA hybrid, we designed a series of verification experiments. First, we wanted to verify if the transposon adaptor could be ligated to the end of fragmented RNA (**Fig. 2A**). In brief, we prepared RNA/DNA hybrid from HEK293T RNA using reverse transcription. After tagmentation with Tn5 transposome, we purified the products to remove Tn5 proteins and free adaptors. We assume that Tn5 can ligate the adaptor to the fragmented DNA. At the same time, if Tn5 ligated the adaptor (**Fig. 2A,** dark blue) to the RNA strand, the adaptor could serve as the template in the following extension step. After extension the DNA strand should have primer binding site on both 5' and 3' ends for PCR amplification. RNase H treatment should not affect the sequencing library production. If Tn5 failed to ligate the adaptor to the RNA strand, neither strand of the heteroduplex could be converted into sequencing library.

After PCR amplification, we obtained a high quantity of product regardless of the RNase

H digestion, suggesting the successful ligation of adaptor to the fragmented DNA. Sequencing results from both reaction test conditions as well as 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. 2B, Fig. S3A**). The additional purification step after reverse transcription and/or RNase H digestion before PCR amplification does not seem to affect the results, likely due to the large amount of starting RNA. We further examined the sequencing reads with insert size shorter than 100 bp (shorter than sequence read length), and about 99.7% of them contained adaptor sequence. Such read-through reads directly proved the ligation of adaptor to the fragmented RNA (**Fig. S3B and C**). In summary, our experiment proved that Tn5 transposome could tagment both strands of DNA and RNA in the RNA/DNA heteroduplex.



	Ligation Test 1	Ligation Test 2	SHERRY
dUTP in RT	-	-	-
Clean-up after Tag.	+	+	-
RNase H Digest	-	+	-
USER Digest	-	-	-

	Strand Test 1	Strand Test 2	Strand Test 3	dU-SHERRY
dUTP in RT	+	+	+	+
Clean-up after Tag.	+	+	+	+
RNase H Digest	-	-	+	-
USER Digest	+	+	-	-

Fig.2 Verification of Tn5 tagmentation activities on RNA/DNA heteroduplex. (**A**) Procedure of two Ligation Tests. Gray dotted box indicates negative results. The table below lists key experimental parameters different from standard SHERRY. (**B**) Comparison of two Ligation Tests and standard SHERRY in aspects of mapping rate, exon rate and gene number. Each test has two replicates. (**C**) Procedure of three Strand Tests. (**D**) Comparison of three Strand Tests and dU-SHERRY in aspects of mapping rate, exon rate and gene number. Each test has two replicates.

Tagmented cDNA is the preferred amplification template. Next, we tried to determine

if the RNA and DNA strand could be amplified to form the sequencing library (**Fig. 2C**). We replaced dTTP with dUTP during the reverse transcription, then purified the tagmented products to remove free dUTP as well as Tn5 proteins. Bst 2.0 Warmstart DNA polymerase was used to perform extension as it was able to use RNA as primer and to process the dU bases in the template. Then the product fragments were 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 the 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 for skipping the RNase H digestion step (dU-SHERRY) to ensure that Tn5 could recognize the substrates with dUTP.

Sequencing results of Strand Test 1 showed low mapping rate and gene detection count, which were only slightly higher than those of Strand Test 2. In contrast, Strand Test 3 demonstrated similar exon rate and gene count to dU-SHERRY as well as SHERRY (**Fig. 2D, Fig. S3A**). Based on these results, we conclude that the tagmented cDNA contributes to the majority of final sequencing library, likely due to inevitable RNA degradation during the series of molecular reactions.

SHERRY for rapid one-step RNA-seq library preparation. We further tested different reaction conditions to optimize SHERRY performance with 10 ng total RNA as input (**Fig. S4A**). Specifically, 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 different steps, which could remove primer dimers and carry-over contaminations. Sequencing results showed little performance change from most of these attempts, proving the robustness of SHERRY under various conditions.

Then we compared the optimized SHERRY to NEBNext® Ultra™ II, a commercially

available kit, for bulk RNA library preparation. This NEBNext kit is probably the most popular kit used in current RNA-seq experiments, with 10 ng total RNA as lowest input limit. We therefore tested the RNA-seq performance with 10 ng and 200 ng total RNA inputs, each condition having three replicates. SHERRY demonstrated comparable performance with NEBNext over the two input levels (**Fig. S4B**). For the tests with10 ng inputs, SHERRY showed higher precision of gene expression measurement across replicates (**Fig. 3A**), likely due to the simpler workflow of SHERRY.

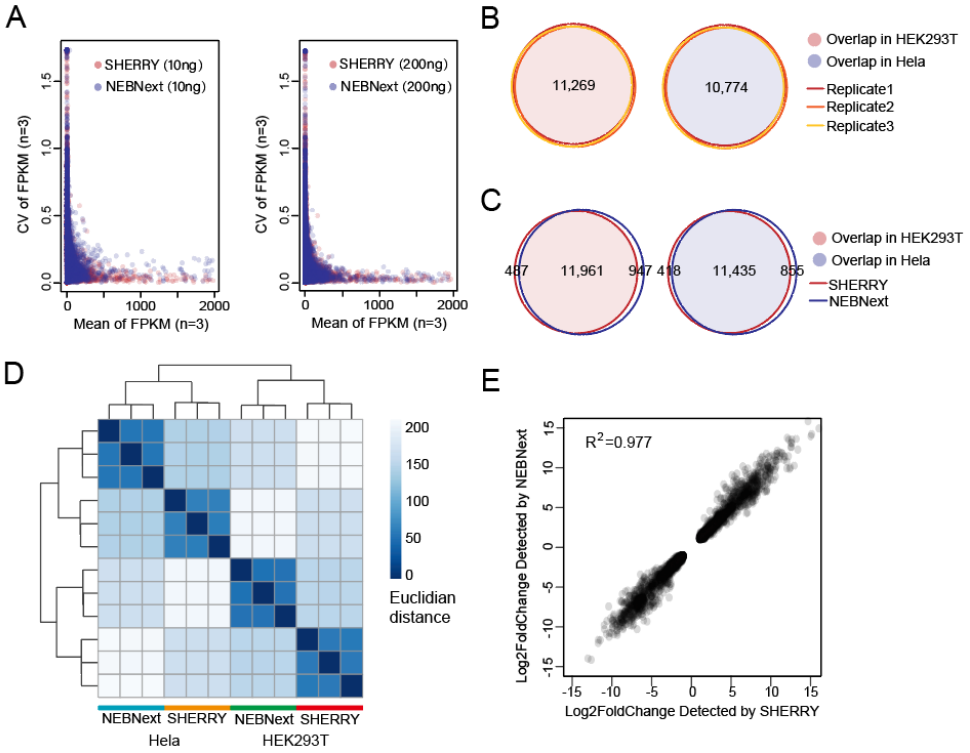


Fig.3 Performance of SHERRY on large input of RNA. (**A**) Coefficient of variation (CV) across three replicates was plotted against mean value for each gene's FPKM. (**B**) Genes detected by three replicates of SHERRY (200 ng input) are plotted as Venn Diagrams, with RNAs from HEK293T and Hela cells, respectively. Numbers of mutual genes are listed. (**C**) Mutual genes of three replicates detected by SHERRY and NEBNext (200 ng input), with RNAs from HEK293T and Hela cells, respectively. (**D**) Distance heatmap of samples prepared by SHERRY or NEBNext with input of 200 ng HEK293T or Hela RNA, three replicates each condition. The color bar indicates the Euclidian distance. (**E**) Correlation of fold change identified by SHERRY and NEBNext. Involved genes are differentially expressed genes detected by both methods.

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 and 10,774 genes in Hela, with high precision (correlation coefficient 0.999) (**Fig. 3B, Fig. S5A-C**). Besides, results from SHERRY and NEBNext

were highly concordant (**Fig. 3C**). This excellent reproducibility of SHERRY ensured the reliability of subsequent analysis. Then we plotted the heatmap of distance matrix (**Fig. 3D**) between different cell types and library preparation methods. Libraries from the same cell type were clustered together as expected. Libraries from the same method tended to cluster together as well, indicating internal bias in both methods.

We then used DESeq2 to detect differentially expressed genes (P-value<$5x10^{-6}$, $|log_2$Fold change| >1). In general, the thousands of differential expressional genes detected by both methods are highly similar (**Fig. S5D**). The fold change of these mutually detected differential expression genes had high correlation (correlation coefficient R^2=0.977) between SHERRY and NEBNext (**Fig. 3E**). After examining the genes that showed differentially expression in only one method, we discovered the same trend of expression change in the data from the other method (**Fig. S5E and F**). We conclude that SHERRY provides equally reliable differential gene expression information as NEBNext, with much less time and labor for the whole process.

SHERRY of trace amounts and single cells. We further set out to see if SHERRY could construct RNA-seq library from single cells. First, we reduced the input to 100 pg total RNA, which was equivalent to RNA from about 10 cells. SHERRY results showed high quality, with high mapping and exon rates and near 9,000 genes detected (**Fig. S6A**). 72% of these genes were mutually detected in all three replicates, demonstrating superior reproducibility (**Fig. 4A**). Expression of these mutually detected genes showed excellent precision with R^2 ranging from 0.958 to 0.970. (**Fig. 4B**).

To further push the detection limit, we carried out single-cell SHERRY experiments (scSHERRY). In contrast to the experiments with purified RNA, scSHERRY required more sophisticated tuning, and we made several optimizations over the standard protocol (**Fig. 4C, Fig. S6B**). Though we found no positive effect from replacing betaine with other crowding agents during standard protocol optimization, we found that addition of crowding reagent with higher molecular weight would improve the library quality from single cells, thus 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 due to the stronger processivity and strand displacement activity of Bst polymerases, and the compatibility with higher reaction temperature to open the secondary structure of RNA templates. We also tried to optimize the PCR strategy since extensive amplification could lead to strong bias. Compared to the continuous 28-cycle PCR, the adding of a purification step after 10 cycles, or simply the reduction of total cycle number to 15 increased the mapping rate as well as the gene number detected. Hence, we simply performed 15-cycle PCR without extra purification in order to better accommodate high-throughput experiments.

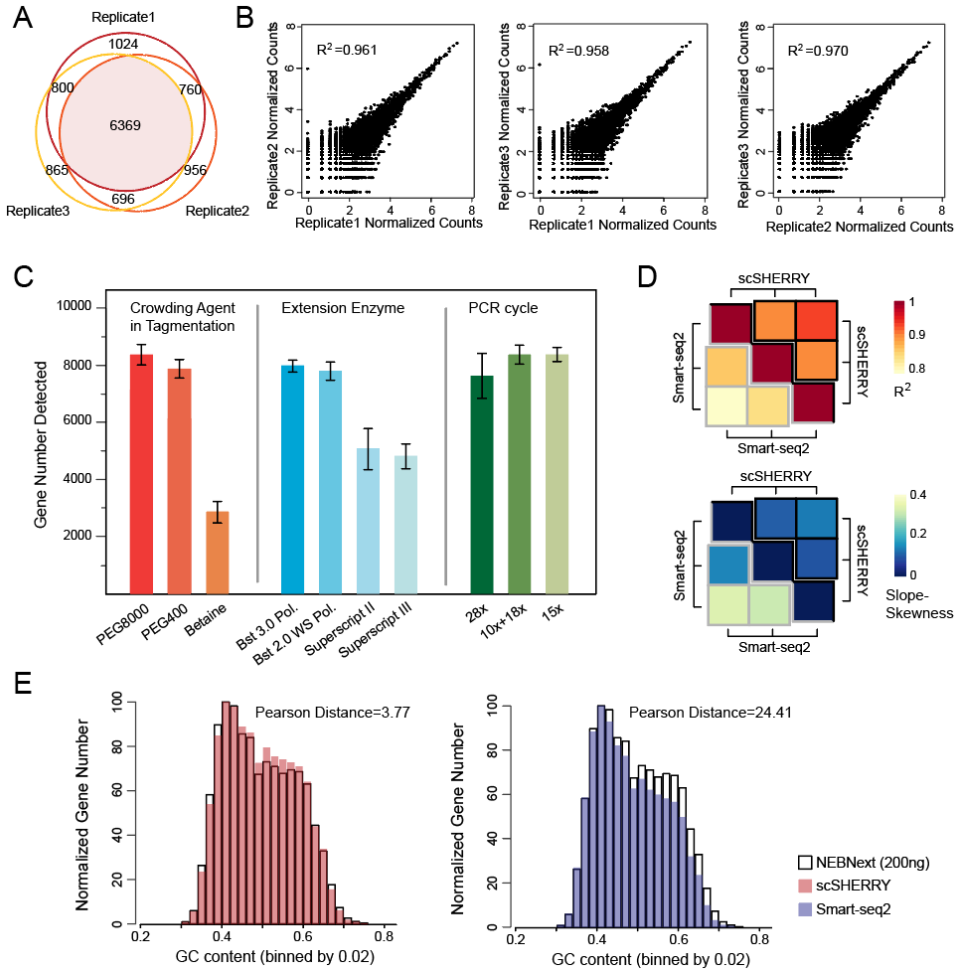


Fig.4 Performance of SHERRY on micro-input samples. (**A**) Genes detected by three replicates of SHERRY with 100 pg total RNA inputs. (**B**) Correlations of normalized gene counts between replicates with 100 pg inputs. (**C**) Gene number detected by scSHERRY under various experimental conditions. Each condition has 3-4 replicates. (**D**) The heatmap of R^2 calculated from replicates of scSHERRY

and Smart-seq2, and the deviation of slope and 1 in linear fitting equation of two methods. (**E**) The normalized gene numbers at different GC content.

The optimized scSHERRY was capable of detecting 8,338 genes with 50.17% mapping rate (**Fig. S6D**), and it showed better reproducibility than Smart-seq2, the most prevalent protocol in the single-cell RNA amplification field (**Fig.4D, Fig.S6C**). Compared to Smart-seq2, the gene number and coverage uniformity of scSHERRY resulted library was still a little inferior (**Fig. S6D, E**), as Smart-seq2 had enriched full-length transcript during preamplification step. However, this enrichment step of Smart-seq2 brought serious bias as well (**Fig. 4E**). We used 200 ng RNA to construct the sequencing library using NEBNext kit, and took the sequencing results as the GC bias free standard. Then we compared the GC distribution of genes detected by scSHERRY or Smart-seq2 with this standard. Results of scSHERRY, which is free from the second-strand synthesis and preamplification, showed a distribution similar to the standard. On the other hand, the library from Smart-seq2 amplified scRNA showed a 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 a biased quantitation result. Overall, compared with Smart-seq2, scSHERRY library had comparable quality and lower GC-bias. Moreover, scSHERRY workflow was convenient, time efficient, and promising for higher throughput.

Discussion

We found that Tn5 transposome has the capability to directly fragment and tag RNA/DNA heteroduplex and thus proposed a quick RNA amplification and library preparation method called SHERRY. The input of SHERRY could be RNA from single cell lysate or total RNA extracted from a large number of cells. By comparing with the prevalent Smart-seq2 protocol for single-cell input or commercial golden-standard kit (NEBNext) for bulk total RNA, SHERRY exhibited comparable performance for input amount spanning more than 5 orders of magnitude. What's more, the whole workflow of SHERRY consists only five steps in one tube and takes about four hours in total, with hands-on time less than 30 min, from RNA to sequencing library. For Smart-seq2, the time required is doubled and an additional library preparation step is necessary. For NEBNext, the protocol is much more labor-intensive with its time-consuming ten-step protocol (**Fig.S7A**). Moreover,

SHERRY offers five-fold reduction of the reagent cost compared to the other two methods (**Fig.S7B**). Therefore, SHERRY exhibits great advantages to compete with conventional RNA library preparation methods as well as scRNA amplification methods.

In our previous experiments, Tn5 transposome was assembled by home-purified pTXB1 Tn5 and synthesized sequencer-adapted oligos. To generalize SHERRY method, as well as to further verify Tn5 activity of tagmentation on RNA/DNA heteroduplex, we tested two commercially available Tn5 transposomes, Amplicon Tagment Mix (abbreviated as ATM) in Nextera XT kit (Illumina, USA) and TTE Mix V50 (abbreviated as V50) in TruPrep kit (Vazyme, China). We normalized different sources of Tn5 transposomes according to the enzyme unit processing 5 ng genomic DNA to the same size under the same reaction condition (**Fig. S8A**). The results indicated that the tagmentation activity of home-made pTXB1 Tn5 was 10-fold higher than V50 and 500-fold higher than ATM when using the volume of transposome as the metric. Same units of enzyme were then used to process RNA/DNA heteroduplex prepared from 5 ng mRNA, confirming that they had similar performance on such hybrid (**Fig. S8B**). RNA-seq results of libraries from all three enzymes showed consistent results, demonstrating the robustness of SHERRY (**Fig. S8C**).

During DNA and RNA/DNA heteroduplex tagmentation, we found that Tn5 transposome reacted with these two substrates in different pattern. We further titrated pTXB1 Tn5 transposome at three conditions (0.02 μl, 0.05 μl and 0.2 μl), then tagmented 5 ng DNA and mRNA/DNA hybrid separately with each condition (**Fig. S9**). As the amount of Tn5 increased, dsDNA was cut into shorter fragments as a whole. While for hybrid, it seemed that Tn5 cut the template ‘one by one’, since only part of hybrid size shifted shorter and most of them was too short to be cut.

Despite the easiness and commercial competence of SHERRY, the library quality of this method may be limited by its transcript coverage evenness. Unlike NEBNext kit which fragments RNA before reverse transcription, or Smart-seq2 which performs pre-amplification to select full-length cDNA, SHERRY simply reverse transcribes full-length

RNA. Reverse transcriptase is well known for its low efficiency and when using polyT as primer for extension, it would be difficult for the transcriptase to reach the 5' end of the RNA template and caused coverage imbalance across the transcripts, making the RNA-seq signal biased toward the 3' end of the gene (**Fig. S10A**). In an attempt to solve this problem, we added small amount of TSO primer in the reverse transcription buffer to mimic the Smart-seq2 RT condition. The result showed much improved evenness across the transcripts (**Fig. S10A and B**), though some of the sequencing parameters dropped accordingly (**Fig. S10C**). We believe, with continuous optimization, SHERRY will take RNA-seq to the next level.

Author contributions

Y.F., K.L., X.S.X., Y.H. and J.W. conceived the project; L.D., J.X., J.O. and D.W. performed structural analysis; Y.S., J.L. and G.W. performed Tn5 purification and characterization; L.D., Y.F. and L.L. conducted research; L.D., Y.F., L.L., J.Y., Y.W., R.L., G.Z., Y.H. and J.W. analyzed the data; L.D., X.S.X., Y.H. and J.W. wrote the manuscript with the help from all other authors.

Conflict of interest statement

X. Sunney Xie, Kaiqin. Lao and Yalei Wu are shareholders of XGen US Co. Kai Q. Lao, Yalei Wu, Raymond W. Lee and Genhua Zheng are employees of XGen US Co.

Data deposition

The sequence reported in this paper has been deposited in the Genome Sequence Archive (accession no. CRA002081).

Acknowledgement

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 Initiation (Z181100001518004), Beijing Advanced Innovation Center

for Structural Biology, and Beijing Advanced Innovation Center for Genomics.

References

1. Y. Liu, Y. L. Lightfoot, N. Seto, C. Carmona-Rivera, E. Moore, R. Goel, L. O'Neil, P. Mistry, V. Hoffmann, S. Mondal, P. N. Premnath, K. Gribbons, S. Dell'Orso, K. Jiang, P. R. Thompson, H. W. Sun, S. A. Coonrod, M. J. Kaplan, Peptidylarginine deiminases 2 and 4 modulate innate and adaptive immune responses in TLR-7-dependent lupus. JCI Insight **3**, e124729 (2018)

2. E. Schmidt, I. Dhaouadi, I. Gaziano, M. Oliverio, P. Klemm, M. Awazawa, G. Mitterer, E. Fernandez-Rebollo, M. Pradas-Juni, W. Wagner, P. Hammerschmidt, R. Loureiro, C. Kiefer, N. R. Hansmeier, S. Khani, M. Bergami, M. Heine, E. Ntini, P. Frommolt, P. Zentis, U. A. Ørom, J. Heeren, M. Blüher, M. Bilban, and J. W. Kornfeld, LincRNA H19 protects from dietary obesity by constraining expression of monoallelic genes in brown fat. Nat. Commun. **9**, 3622 (2018).

3. O. Wurtzel, R. Sapra, F. Chen, Y. Zhu, B. A. Simmons, R. Sorek, 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 **10**, 50 (2019).

5. P. Civita, S. Franceschi, P. Aretini, V. Ortenzi, M. Menicagli, F. Lessi, F. Pasqualetti, A. G. Naccarato, C. M. Mazzanti, 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, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, I. Amit, Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. Science **343**, 776-779 (2014)

7. Y. Chen, Y. Zheng, Y. Gao, Z. Lin, S. Yang, T. Wang, Q. Wang, N. Xie, R. Hua, M. Liu, J. Sha, M. D. Griswold, J. Li, F. Tang, M. H. Tong, Single-cell RNA-seq uncovers dynamic processes and critical regulators in mouse spermatogenesis. Cell Res. **28**, 879-896 (2018).

8. M. Wang, X. Liu, G. Chang, Y. Chen, G. An, L. Yan, S. Gao, Y. Xu, Y. Cui, J. Dong, Y.

Chen, X. Fan, Y. Hu, K. Song, X. Zhu, Y. Gao, Z. Yao, S. Bian, Y. Hou, J. Lu, R. Wang, Y. Fan, Y. Lian, W. Tang, Y. Wang, J. Liu, L. Zhao, L. Wang, Z. Liu, R. Yuan, Y. Shi, B. Hu, X. Ren, F. Tang, X. Y. Zhao, J. Qiao, Single-Cell RNA Sequencing Analysis Reveals Sequential Cell Fate Transition during Human Spermatogenesis. Cell Stem Cell **23**, 599-614 (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, Q. Lin, F. Ding, C. Xin, W. Gong, L. Zhang, J. Geng, B. Zhang, X. Yu, J. Yang, S. Hu, J. Yu, A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. Genomics **96**, 259-265 (2010).

13. N. Cloonan, A. R. Forrest, G. Kolle, B. B. Gardiner, G. J. Faulkner, M. K. Brown, D. F. Taylor, A. L. Steptoe, S. Wani, G. Bethel, A. J. Robertson, A. C. Perkins, S. J. Bruce, C. C. Lee, S. S. Ranade, H. E. Peckham, J. M. Manning, K. J. McKernan, S. M. Grimmond, Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat. Methods **5**, 613-619 (2008).

14. C. D. Armour, J. C. Castle, R. Chen, T. Babak, P. Loerch, S. Jackson, J. K. Shah, J. Dey, C. A. Rohl, J. M. Johnson, C. K. Raymond, Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. Nat. Methods **6**, 647-649 (2009).

15. D. Parkhomchuk, T. Borodina, V. Amstislavskiy, M. Banaru, L. Hallen, S. Krobitsch, H. Lehrach, A. Soldatov, Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. **37**, e123 (2009).

16. F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, M. A. Surani, mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods **6**, 377-382 (2009).

17. S. Picelli, O. R. Faridani, A. K. Björklund, G. Winberg, S. Sagasser, R. Sandberg, Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. 9, 171-181 (2014).

18. T. Hashimshony, N. Senderovich, G. Avital, A. Klochendler, Y. de Leeuw, L. Anavy, D. Gennert, S. Li, K. J. Livak, O. Rozenblatt-Rosen, Y. Dor, A. Regev, I. Yanai, 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).
20. G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, J. H. Bielas, Massively parallel digital transcriptional profiling of single cells. Nat. Commun. **8**, 14049 (2017).
21. T. M. Gierahn, M. H. Wadsworth, T. K. Hughes, B. D. Bryson, A. Butler, R. Satija, S. Fortune, J. C. Love, A. K. Shalek, Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nat. Methods **14**, 395-398 (2017).
22. J. Cao, J. S. Packer, V. Ramani, D. A. Cusanovich, C. Huynh, R. Daza, X. Qiu, C. Lee, S. N. Furlan, F. J.Steemers, A. Adey, R. H. Waterston, C. Trapnell, J. Shendure, Comprehensive single-cell transcriptional profiling of a multicellular organism. Science **357**, 661-667 (2017).
23. A. B. Rosenberg, C. M. Roco, R. A. Muscat, A. Kuchina, P. Sample, Z. Yao, L. T. Graybuck, D. J. Peeler, S. Mukherjee, W. Chen, S. H. Pun, D. L. Sellers, B. Tasic, G. Seelig, Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science **360**, 176-182 (2018).
24. I. Y. Goryshin, W. S. Reznikoff, Tn5 in vitro transposition. J. Biol. Chem. **273**, 7367–7374 (1998).
25. A. Adey, H. G. Morrison, Asan, X. Xun, J. O. Kitzman, E. H. Turner, B. Stackhouse, A. P. MacKenzie, N. C. Caruccio, X. Zhang, J. Shendure, Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. **11**, R119 (2010).
26. S. Picelli, A. K. Björklund, B. Reinius, S. Sagasser, G. Winberg, R. Sandberg, Tn5 transposase and tagmentation procedures for massively scaled sequencing projects.

Genome Res. **24**, 2033-2040 (2014).
27. 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).
28. C. Chen, D. Xing, L. Tan, H. Li, G. Zhou, L. Huang, X. S. Xie, Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). Science **356**, 189-194 (2017).
29. S. Rohrback, C. April, F. Kaper, R. R. Rivera, C. S. Liu, B. Siddoway, J. Chun, 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).
30. 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).
31. B. Lai, Q. Tang, W. Jin, G. Hu, D. Wangsa, K. Cui, B. Z. Stanton, G. Ren, Y. Ding, M. Zhao, S. Liu, J. Song, T. Ried, K. Zhao, Trac-looping measures genome structure and chromatin accessibility. Nat. Methods **15**, 741-747 (2018).
32. J. Gertz, K. E. Varley, N. S. Davis, B. J. Baas, I. Y. Goryshin, R. Vaidyanathan, S. Kuersten, R. M. Myers, Transposase mediated construction of RNA-seq libraries. Genome Res. **22**, 134-141 (2011).
33. S. Brouilette, S. Kuersten, C. Mein, M. Bozek, A. Terry, K. R. Dias, L. Bhaw-Rosun, Y. Shintani, S. Coppen, C. Ikebe, V. Sawhney, N. Campbell, M. Kaneko, N. Tano, H. Ishida, K. Suzuki, K. Yashiro, 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).
34. K. A. Majorek, S. Dunin-Horkawicz, K. Steczkiewicz, A. Muszewska, M. Nowotny, K. Ginalski, J. M. Bujnicki, The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. Nucleic Acids Res. **42**, 4160-4179 (2014).
35. 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).
36. W. S. Reznikoff, Transposon Tn5. Annu. Rev. Genet. **42**, 269-286 (2008).
37. 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).
38. 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)
39. D. Lim, G. G. Gregorio, C. Bingman, E. Martinez-Hackert, W. A. Hendrickson, S. P. Goff, Crystal structure of the Moloney murine leukemia virus RNase H domain. J. Virol. **80**, 8379-8389 (2006).