SHERRY2: A method for rapid and sensitive single cell RNA-seq

Lin Di,[1,2,3,4,#] Bo Liu,[5,6,#] Yuzhu Lyu,[1] Shihui Zhao,[2,7] Yuhong Pang,[2] Chen Zhang,[1] Jianbin Wang,[8,*] Hai Qi,[5,6*] Jie Shen,[1,*] Yanyi Huang[2,4,9,10,*]

[1] School of Basic Medical Sciences, Beijing Key Laboratory of Neural Regeneration and Repair, Advanced Innovation Center for Human Brain Protection, Capital Medical University, Beijing 100069, China
[2] Biomedical Pioneering Innovation Center, Beijing Advanced Innovation Center for Genomics, Peking University, Beijing 100871, China
[3] School of Life Sciences, Peking University, Beijing 100871, China
[4] Institute for Cell Analysis, Shenzhen Bay Laboratory, Guangdong 528107, China
[5] Laboratory of Dynamic Immunobiology, Institute for Immunology, Tsinghua University, Beijing 100871, China
[6] Department of Basical Medical Sciences, School of Medicine, Tsinghua University, Beijing 100871, China
[7] Peking University–Tsinghua University–National Institute of Biological Sciences Joint Graduate Program (PTN), Peking University, Beijing 100871, China
[8] School of Life Sciences, Beijing Advanced Innovation Center for Structural Biology, Tsinghua University, Beijing 100084, China
[9] College of Chemistry and Molecular Engineering, Beijing National Laboratory for Molecular Sciences, Peking University, Beijing 100871, China
[10] Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China

[#] These authors contributed equally.
* Corresponding authors: yanyi@pku.edu.cn (Y.H.), shenjie@ccmu.edu.cn (J.S.), qihai@tsinghua.edu.cn (H.Q.) and jianbinwang@tsinghua.edu.cn (J.W.)

Abstract

Many single cell RNA-seq applications aim to probe a wide dynamic range of gene expression, but most of them are still challenging to accurately quantify low-aboundance transcripts. Based on our previous finding that Tn5 transposase can directly cut-and-tag DNA/RNA hetero-duplexes, we present SHERRY2, an optimized protocol for sequencing transcriptomes of single cells or single nuclei. SHERRY2 is robust and scalable, and it has higher sensitivity and more uniform coverage in comparison with prevalent scRNA-seq methods. With throughput of a few thousand cells per batch, SHERRY2 can reveal the subtle transcriptomic differences between cells and facilitate important biological discoveries.

Introduction

Many experimental methods for transcriptome profiling by next generation sequencing (RNA-seq) have been developed to cover various scales of input samples, ranging from bulk samples[1,2] to single cells[3-5] or even subcellular components[6,7]. High quality single-cell RNA-seq (scRNA-seq) data can be used to reveal the kinetic details of gene expression and transitions between cell states or types[8-10]. Prevalent scRNA-seq methods mainly rely on template switching and pre-amplification of complementary DNA (cDNA). However, large-scale scRNA-seq techniques, commonly operated in micro-droplets or wells, have relatively low sensitivity.[11] Single-tube based scRNA-seq approaches can typically produce higher coverage for low-abundance genes, but they still suffer from quantification bias due to insufficient reverse transcription and GC imbalance during amplification. Besides, their complex experimental methods are generally unsuitable for large-scale studies.

We have reported a highly reproducible and rapid library preparation method for RNA-seq, SHERRY, which can be applied to minute amount of RNA samples[12]. The development of SHERRY was based on the recent discovery that Tn5 transposase can bind and cut RNA/DNA hetero-duplexes directly. With slight modifications, SHERRY could also be applied to various clinical metatranscriptome applications, such as identification of SARS-CoV-2 and other pathogens[13].

Although SHERRY was applied to process single cells and achieved less biased quantification of gene expression in comparison with other scRNA-seq methods, the results still exhibited clear coverage bias toward the 3′-ends of transcripts, relatively low sensitivity, and low tolerance to endogenous DNA. In this work, we present an optimized method, SHERRY2, which addresses the limitations of SHERRY and is fully compatible with single cells and single nuclei with low RNA content. In comparison with prevalent RNA-seq methods, SHERRY2 showed higher sensitivity, better concordance with reference data, greater reproducibility between replicates, and superior scalability, allowing the method to be used to process a few thousand single cells per batch and thus reducing the time required to conduct experiments.

Results

SHERRY2 provides high sensitivity and even coverage across gene bodies for scRNA-seq.

For scRNA-seq, both RNA degradation and incompleteness of reverse transcription (RT) are major factors that reduce gene detection sensitivity and coverage evenness. Although adding random RT primers facilitates the coverage of long transcripts, it requires removal of ribosomal RNA, which is incompatible with scRNA-seq[13]. Spiking template-switching oligonucleotides also provides more uniform coverage, but this strategy has limited detection sensitivity and specificity[12].

We altered various experimental parameters of the original SHERRY protocol for both bulk and single cell inputs (**Extended Data Fig. 1, 3**). To protect RNA from degradation, we lowered the concentration of free Mg^{2+}, either by reducing the amount of total Mg^{2+} or adding more dNTP to chelate Mg^{2+} ions[14], and observed significant improvement of the coverage evenness of RNA-seq. To facilitate cDNA synthesis, we screened different reverse transcriptases and found that SuperScript IV (SSIV), working at a relatively high temperature with a low Mg^{2+} concentration, could better overcome the secondary structure of RNA and hence simultaneously enhanced the sensitivity and uniformity of RNA-seq.

When RNA-seq was conducted using pictogram-level RNA inputs, sufficient amount of Tn5 transposome was important for high sensitivity, and Bst 3.0 DNA polymerase filled the gap left by Tn5 tagmentation more effectively than other enzymes. The protocol was insensitive to many experimental conditions, including the usage of single strand DNA binding proteins[15], Tn5 inactivation, the concentration of extension polymerase, and the usage of hot-start polymerase.

We named the optimized SHERRY protocol SHERRY2. Using RNA extracted from HEK293T cells as input, we compared the performance of SHERRY2 and the original SHERRY protocol. At the 10-ng level (**Extended Data Fig. 2**), both protocols identified more than 11,000 genes at saturation. At the 100-pg level, SHERRY2 performed better than SHERRY and detected 5.0% more genes at 0.6-million reads. In addition, SHERRY2 greatly diminished 3′-end coverage bias and increased the unique mapping rate for 10-ng and 100-pg inputs. We also constructed a bias-free RNA-seq library using 200-ng total RNA input via the conventional fragmentation-and-ligation method with the NEBNext E7770 kit (NEBNext). For 100-pg input, the gene overlap between NEBNext and SHERRY2 was greater than that between NEBNext and SHERRY (81.7% vs 78.4%), and the gene expression results of NEBNext and SHERRY2 were also more closely correlated (R=0.70 vs R=0.65).

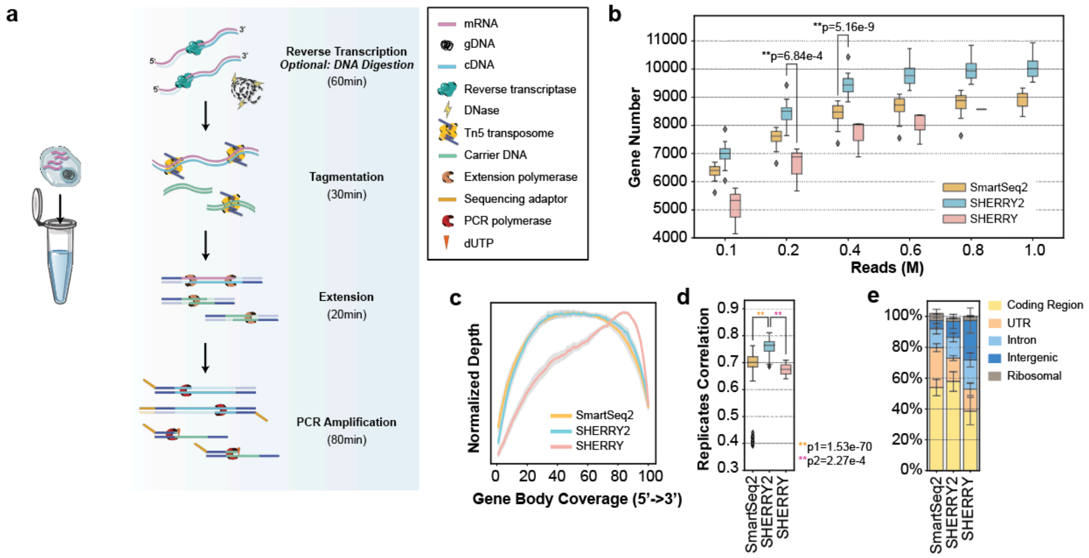


Fig 1. The workflow and general performance of SHERRY2 on single cell RNA-seq. a, The workflow of SHERRY2 for scRNA-seq. Poly(A) tailed RNA is firstly released from single cells and reverse transcribed. The resulting RNA/cDNA hetero-duplex is then tagmented by Tn5 transposome, followed by gap-repair and indexed PCR. Optionally, chromatin can be digested during lysis. The entire protocol is performed in one tube and takes 3 hours. **b**, Gene number (FPKM>1) with SmartSeq2, SHERRY2 and SHERRY when subsampling reads to 0.1, 0.2, 0.4, 0.6, 0.8 and 1 million reads. **c,** Gene body coverage detected by the three scRNA-seq protocols. The gray region represents the standard deviation of the normalized depth among replicates. **d,** Pairwise correlation of gene expression within replicates for the three scRNA-seq protocols. The correlation R-value was calculated by a linear fitting model with normalized counts of overlapped genes. **e,** Components of reads that were mapped to different regions of the genome using the three scRNA-seq protocols. The error bars show the standard deviation. The samples in **b-e** are single HEK293T cells. The p-values in **b** and **d** were calculated by the Mann-Whitney-U test.

The SHERRY2 protocol for scRNA-seq contains only four steps: reverse transcription, Tn5 tagmentation, gap-filling through extension, and PCR amplification. The entire SHERRY2 protocol can be completed within 3 hours, whereas the original SHERRY protocol requires 4 hours, and other high-sensitivity scRNA-seq method such as SmartSeq2 may require much more time to be completed[3] (**Fig. 1a**). The one-tube workflow of SHERRY2 is readily scalable to high-throughput applications. SHERRY2 was able to detect 10,024 genes (FPKM >1) within a single HEK293T cell at 1-million reads. When subsampling to 0.2-million reads, SHERRY2 still detected 8,504 genes, which was 1,622 (23.6%) more than SHERRY and 886 (11.6%) more than SmartSeq2 (**Fig. 1b**). In addition, the evenness of gene body coverage for SHERRY2 was significantly higher than that of the original SHERRY protocol (0.72 vs 0.84) and was comparable to that of SmartSeq2 (0.84) (**Fig. 1c**). Moreover, the reproducibility of SHERRY2 was significantly higher than that of SHERRY and SmartSeq2 (**Fig. 1d**) due to its simplified workflow and stable performance. The exonic rate of SHERRY2 was also improved in comparison with that of

SHERRY, likely due to the higher RT efficiency of the newly developed method (**Fig. 1e**).

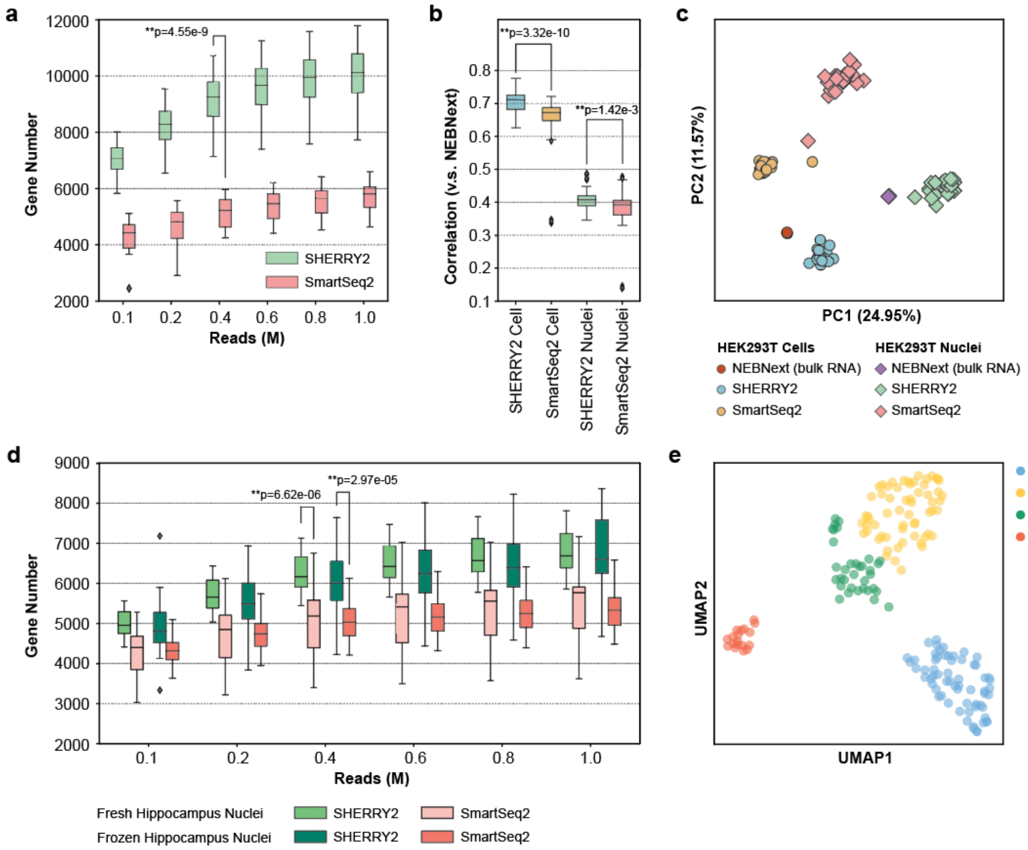


Fig 2. Sensitivity and accuracy of SHERRY2. a, Gene number (RPM>1) of single HEK293T nuclei detected by SHERRY2 and SmartSeq2 when subsampling reads to 0.1, 0.2, 0.4, 0.6, 0.8 and 1 million reads. b, Gene expression correlation between single HEK293T cells (or nuclei) and 200-ng RNA extracted from HEK293T cells (or nuclei). Single-cell (or nucleus) data were acquired by SHERRY2 and SmartSeq2. Bulk RNA results were acquired by the standard NEBNext protocol. The correlation R-value was calculated by a linear fitting model with normalized gene counts. c, Clustering of HEK293T cellular and nuclear RNA-seq data from SHERRY2, SmartSeq2 and NEBNext using principal component analysis. The analysis utilized differentially expressed genes (adjusted p-value < 1e-4 and fold change > 2) between cells and nuclei detected by NEBNext. d, Gene number (RPM>1) of single neuron nuclei detected by SHERRY2 and SmartSeq2 when subsampling reads to 0.1, 0.2, 0.4, 0.6, 0.8 and 1 million reads. The nuclei were isolated from mouse hippocampi that were freshly prepared or previously frozen at -80°C. e, Clustering of single hippocampal neuron nuclei visualized by UMAP plot. The snRNA-seq library was prepared by SHERRY2. The analysis utilized genes expressed (counts > 0) in more than 4 nuclei. The p-values in a, b, and d were calculated by the Mann-Whitney-U test.

SHERRY2 is highly accurate for scRNA-seq and snRNA-seq.

Single nucleus RNA-seq (snRNA-seq) has gained popularity since fresh and intact single cells are challenging to obtain in many applications. We tested the performance of SHERRY2 on snRNA-seq using single nuclei isolated from HEK293T cells. SHERRY2 detected 10,137 genes (RPM>1) at 1-million reads, which was 4,330 (74.6%) more than SmartSeq2, demonstrating that SHERRY2 had superior sensitivity for single nuclei (**Fig. 2a**). SHERRY2 also exhibited superior accuracy, as demonstrated by the significantly higher correlation between the SHERRY2 gene expression results and bias-free NEBNext libraries in comparison with that of SmartSeq2 (R=0.71 vs R=0.67 for scRNA-seq, and R=0.41 vs R=0.39 for snRNA-seq) (**Fig. 2b**).

In addition, scRNA-seq with SHERRY2 showed high tolerance to GC content and was insensitive to the length of transcripts (**Extended Data Fig. 4**). Unlike SmartSeq2, for which the gene overlap and expression correlation with bulk RNA-seq showed clear declines when GC-content was greater than 40%, SHERRY2 maintained these parameters at high and constant levels (82.6% overlap and R=0.76) until the GC content reached 60%. Transcript length did not influence the accuracy of SHERRY2 or SmartSeq2, although SmartSeq2 exhibited a small degree of intolerance for transcripts longer than 800 bases.

The high accuracy and sensitivity of SHERRY2 allowed better distinction between HEK293T cells and their nuclei, which had minimal differences. We performed principal component analysis (PCA) using RNA-seq data from NEBNext, SHERRY2 and SmartSeq2 (**Fig. 2c**). Single cells and nuclei prepared by SHERRY2 were much closer in distance to the bulk RNA results in comparison with those prepared with SmartSeq2. In addition, the expression pattern of the differential genes identified by SHERRY2 was more similar to that of NEBNext in comparison with SmartSeq2 (**Extended Data Fig. 5**).

snRNA-seq of mouse hippocampal neurons.

snRNA-seq is a popular method for studies of brain tissue due to the technical challenge of isolating intact single neurons. We constructed snRNA-seq libraries of hippocampal neurons with SHERRY2 and SmartSeq2. For both frozen and freshly prepared hippocampus, SHERRY2 detected significantly more genes than SmartSeq2 (6,600 vs 5,331 at 1-million reads for frozen samples, 6,686 vs 5,769 at 1-million reads for fresh samples) (**Fig. 2d**).

Next, we sequenced a small number of fresh hippocampal neurons (176 nuclei) with SHERRY2 and classified their cell types correctly. The nuclei were non-supervisedly clustered into 4 distinct groups (**Fig. 2e**), after which they were re-clustered using marker genes identified by sNuc-Seq[16] (**Extended Data Fig. 6**). The two clustering results were highly consistent. By comparing the expression pattern of sNuc-Seq marker genes with SHERRY2 expression data, three clusters in **Fig. 2e** were perfectly annotated: DG granule cells, CA1 pyramidal cells and GABAergic interneuron cells (**Extended Data Fig. 6**).

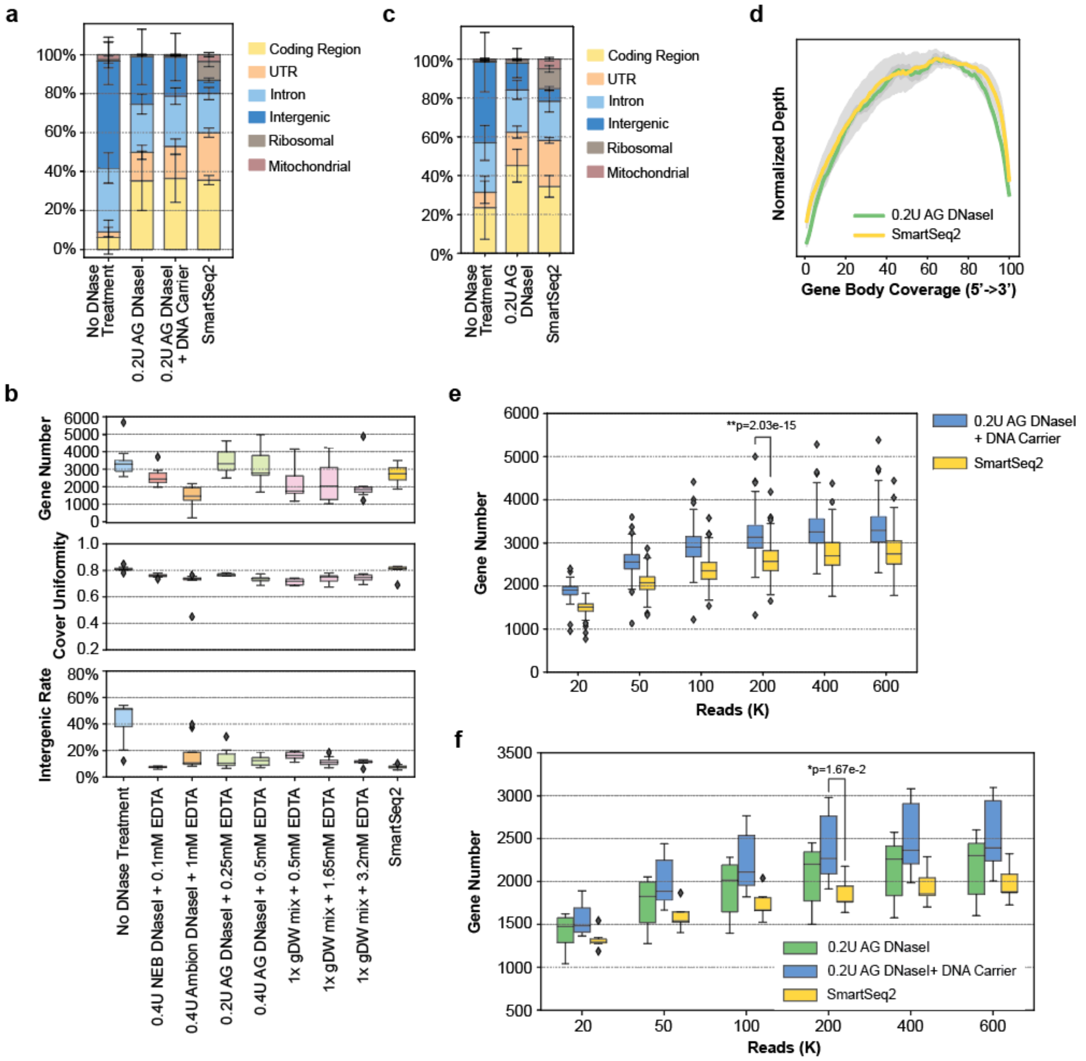


Fig 3. scRNA-seq of low RNA-content samples with SHERRY2. a, Proportions of genome regions covered by reads from SHERRY2 without DNase treatment, SHERRY2 with AG DNase I addition, SHERRY2 with AG DNase I and DNA carrier addition, and SmartSeq2. **b,** Library quality of SHERRY2, including gene number (FPKM>1) at 0.25-million reads, coverage uniformity across gene body and percentage of reads that were mapped to intergenic regions with different DNases. The labels below the figure indicate the amounts and names of the DNases, as well as the EDTA concentration that was added during DNase inactivation. SmartSeq2 was also performed as a reference. **c,** Components of reads covering different genome regions detected by SHERRY2 without DNase treatment, SHERRY2 with optimized AG DNase I, and SmartSeq2. **d,** Gene body coverage detected by SHERRY2 (with AG DNase I) and SmartSeq2. The gray region shows the standard deviation of the normalized depth among replicates. **e,** Gene number (FPKM>1) detected by SHERRY2 (with AG DNase I and DNA carrier) and SmartSeq2 when subsampling to 20, 50, 100, 200, 400, and 600 thousand reads. The samples in **b-e** were single B cells isolated from murine GC light zones. **f,** Gene number (FPKM>1) detected by SHERRY2 with AG DNase I addition, SHERRY2 with AG DNase I and DNA carrier addition, and SmartSeq2 when subsampling to 20, 50, 100, 200, 400, and 600 thousand reads. Samples with intergenic rate lower than 25% were counted. The samples in **a** and **f** were single lymphocyte cells from murine eyeball blood. The p-values in **e** and **f** were calculated by the Mann-Whitney-U test. The error bars in **a** and **c** show the standard deviation.

Cluster 2 was found to consist of cells in an intermediate state between CA1 and GABAergic cells. Such groups of cells were previously found in 10X scRNA-seq data but not able to be well charterized (**Extended Data Fig. 6**). The SHERRY2 results demonstrated that these clusters were reliable, while also revealing that they consisted of cells with relatively high expression of Tshz2 and Ptprd (**Extended Data Fig. 6**).

scRNA-seq using low RNA-content cells.

For low RNA-content cells, such as immune cells[17], we found that removal of intergenic DNA contaminations by DNase treatment was especially crucial for SHERRY2 scRNA-seq. In such cells, the open DNA regions of disassembled chromatin might be favored over RNA/DNA hybrids during Tn5 tagmentation. When DNase was omitted from the SHERRY2 protocol, more than 50% of reads sequenced from single mouse lymphocytes (**Extended Data Fig. 8**) were mapped to intergenic regions, and only around 10% of reads were exonic reads (**Fig. 3a**).

Different DNases performed differently in SHERRY2 scRNA-seq. We tested five DNases: NEB, Ambion, TURBO, AG, and gDW mix (short name). Three DNases (NEB, Ambion, and TURBO DNase I) that worked and inactivated at higher temperatures increased the intergenic rate unexpectedly, and this effect was probably due to RNA degradation at high temperatures with excess Mg^{2+} in the reaction buffer. In contrast, AG DNase I and gDW mix, which worked at room temperature, yielded ideal results.

We confirmed that all the five DNases could digest more than 99.5% of DNA (30-ng) by simply utilizing divalent ions of their respective storage buffer. Without adding extra divalent ions, the intergenic rates of single germinal center (GC) B cells for all DNases were less than 20% (**Fig. 3b**). Among the DNases, AG DNase I retained high sensitivity for gene detection, and more than 60% of reads were mapped to exon regions (**Fig. 3c**), while the evenness of coverage was not affected (**Fig. 3d**).

Next, dU-containing carrier DNA, which prevented being amplified by dUTP-intolerant polymerase, was added to further improve the efficiency of tagmentation of RNA/DNA hybrids. With carrier DNA, SHERRY2 detected more than 3,200 genes at saturation (0.6-million reads) for single GC B cells (**Fig. 3e**), and the number of detectable genes increased from 2,301 to 2,393 for single lymphocytes, with an exonic ratio comparable to that of SmartSeq2 (**Fig. 3a, f**). Based on these results, chromatin digestion and the addition of carrier DNA were included in the standard SHERRY2 protocol.

Selection dynamics in germinal centers profiled by SHERRY2.

SHERRY2 can be scaled easily to thousands of single cells per batch, owing to its simplified procedures. The GC is a transient structure that supports antibody affinity maturation in response to T cell-dependent antigens, and it contains diverse cell types with complex dynamics. Histologically, the GC can be separated into two micro-compartments, the dark zone and the light zone[18,19]. By surface phenotyping, cells in the two compartments can be distinguished based on CXCR4, CD83 and CD86 markers[20-22], with light-zone cells being $CXCR4^{lo}CD83^{+}CD86^{+}$ while dark-zone cells are $CXCR4^{+}CD83^{lo}CD86^{lo}$. GC cells cycle between the dark zone state and the light zone state. Dark zone cells are highly proliferative and undergo somatic hypermutation, which generates a range of affinities against antigens. In the light zone, these B cells compete with each other for survival factors and help signals, which are mainly derived from follicular helper T cells. Those B cells that have acquired higher-affinity B cell receptors are selected to differentiate into plasma cells (PC) or memory B cells (MBC) or cycle back to the dark zone[19,23-25]. Recently, gray zone, consisting of $CXCR4^{+}CD83^{+}$ cells with distinct gene expression patterns, was discovered and found to be involved in GC recycling[26]. The complex spatiotemporal dynamics of the GC and their underlying mechanisms are incompletely understood. To this end, sensitive scRNA-seq methods that can be used to detect gene expression with less bias are highly desirable.

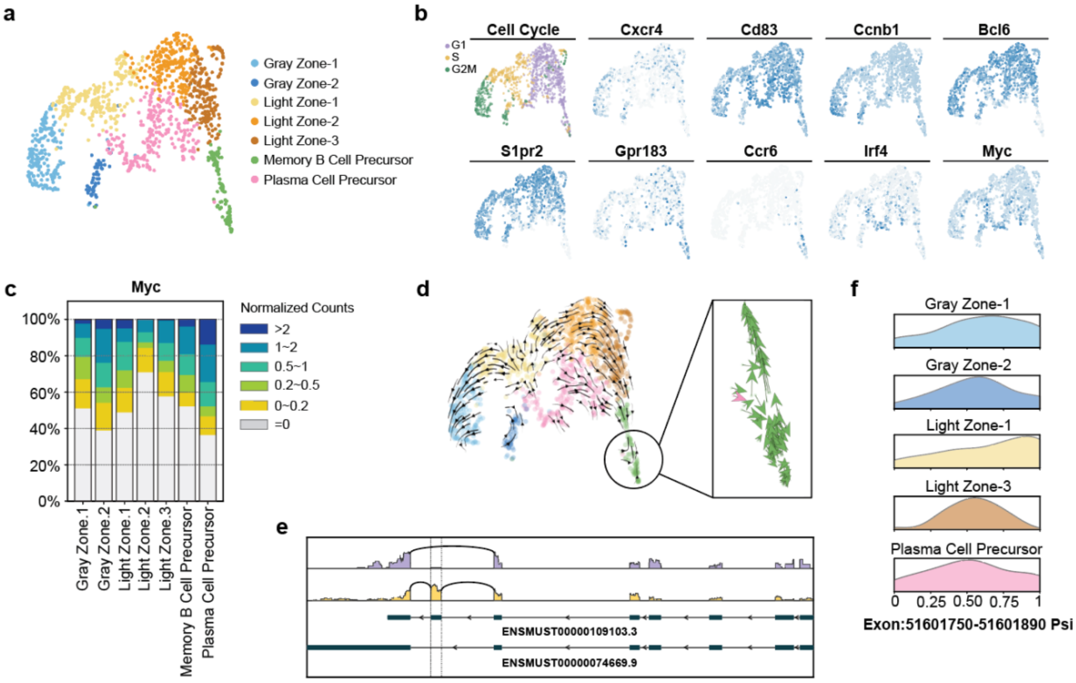


Fig 4. Mouse germinal center profiled by scRNA-seq through SHERRY2. a, Clustering of single B cells from murine GC light zones visualized by UMAP plot. The library was prepared by SHERRY2 (with AG DNase I and DNA carrier). Different colors indicate distinct cell types. **b,** Cell cycle and marker gene expression of different cell types marked on a UMAP plot. The gradient colors correspond to the normalized counts of a specific gene ranging from 0 (white) to 1 (blue). **c,** Distribution

of *Myc* gene expression in different cell types. Different colors indicate different intervals of normalized *Myc* counts. The percentages of cells within the clusters falling into corresponding intervals were counted. **d,** Dynamic process of the GC light zone indicated by vector fields of RNA velocity on a UMAP plot. The expanded region shows the velocity vector of each cell. The colors correspond to the same cell types as annotated in **a**. **e**, Isoforms of the *Hnrnpab* gene. The top two lines show isoforms from two example cells that rarely and preferentially used the highlighted exon in *Hnrnpab* transcripts. The bottom two lines show the isoform structures of *Hnrnpab* transcripts that include or exclude the exon. **f,** Inclusion ratio distribution of the highlighted exon in **e** in different cell types. Only cell types represented by more than 10 cells after filtering are shown.

We profiled 1,248 sorted CXCR4^{lo}CD86^{hi} GC light zone cells with SHERRY2, and 1,231 (98.6%) high-quality cells were retained for downstream analysis. The gene expression levels of *Cd19*, *Ccnd3*, *Fas*, *Cd86* and *Cxcr4* were consistent with flow cytometry gating, and no batch effect was observed (**Extended Data Fig. 8, 9**).

Unsupervised clustering identified seven clusters (**Fig. 4a**), two of which belonged to the gray zone, which was defined by co-expression of *Cxcr4* and *Cd83*, as well as on-going cell division (enriched *Ccnb1*)[26] (**Fig. 4b**). We observed the expected down-regulation of *Bcl6* and *S1pr2*, which are signature genes for GC B cells[27,28], in memory B cell precursors (MPs) and plasma cell precursors (PPs). Specifically, *Ccr6* was exclusively enriched in MPs[29], while *Irf4*, *Cd40*, *Icam1*, and *Slamf1* were up-regulated in PPs[30] (**Extended Data Fig. 9)**. The relatively low expression levels of *Prdm1* (not shown) and *Gpr183* in PPs were consistent with the early stage of plasma cell development. In total, 1,071 genes significantly up- or down-regulated in specific clusters were identified (**Extended Data Fig. 9**).

The high sensitivity of SHERRY2 enabled detection of *Myc* in 588 (47.8%) single GC light zone B cells. Using fluorescent protein reporting, Myc was found to mark light-zone cells destined for dark zone re-entry[31], although Myc expression *per se* has been difficult to identify in specific cell types by low-sensitivity scRNA-seq approaches[32]. Consistent with previous findings[26,30], Myc expression was significantly higher in PPs (**Fig. 4b, Extended Data Fig. 9**) and active in gray zones (**Fig. 4c**). Light Zone-1 also had a relatively higher portion of Myc$^+$ cells, which are probably those destined for cyclic reentry to the dark zone[31]. MPs also contained some cells that expressed Myc.

RNA velocity analysis (**Fig. 4d**) suggested that Light Zone-1 contained cells selected for dark zone reentry, which were migrating to the gray zone and had Myc expression characterized by burst kinetics (**Extended Data Fig. 9**). In addition, cells that appeared to have just entered the light zone were also identified. A few velocity vectors that moved to MPs were mixed in PPs, and these vectors were in the same direction with the down-regulation of Myc. According to the velocity analysis, the aforementioned Myc-expressing MPs seemed to have a tendency to cycle back to the GC, suggesting that some MPs with Myc up-regulation have the potential to re-participate in GC reactions.

We then assembled the BCR sequence for each cell to screen the usage of Igh variable sequences, which were assigned in 1,101 (89.4%) cells. As expected[33], IGHV1-72 dominated the NP-reactive GC response, and the coupled light chain was mainly IgL rather than IgK (**Extended Data Fig. 10**). In addition, we identified CDR1 and CDR2 regions in 269 (24.4%) and 493 (44.8%) cells in which Igh variable sequences were assigned, respectively.

SHERRY2 revealed differences in the usage frequencies of exons across cell types. The usage of a particular exon (chr11: 51,601,750-51,601,890) within the *Hnrnpab* transcript (**Fig. 4e**), which is widely expressed and encodes a protein that mainly functions in processing pre-mRNAs, was significantly biased among GC clusters. As shown in **Fig. 4f**, Light Zone-1 cells favored inclusion of this exon.

Summary

We present SHERRY2, an RNA-seq method designed for single cells and single nuclei. SHERRY2 is based on the direct tagmentation function of Tn5 transposase for RNA/DNA hetero-duplexes, and it provides high sensitivity, high accuracy, and high throughput for applications that require a large number of genes to be identified in each cell. The SHERRY2 method is relatively easy to perform, and it could be incorporated into other sequencing modalities for single cells.

Methods

Cell culture

HEK293T cell line was purchased from ATCC and incubated at 37°C with 5% CO_2 in Dulbecco's Modified Eagle Medium (DMEM) (Gibco, 11965092), which was supplemented with 10% fetal bovine serum (FBS) (Gibco, 1600044) and 1% penicillin-streptomycin (Gibco, 15140122). Cells were dissociated by 0.05% Trypsin-EDTA (Gibco, 25300062) at 37°C for 4min and washed by DPBS (Gibco, 14190136).

For DNA or RNA extractions, we took ~10^6 suspended cells, and followed the guideline of PureLink Genomic DNA Mini Kit (Invitrogen, K182002) or RNeasy Mini Kit (Qiagen, 74104). The extracted RNA was further dealt with 20U DNase I (NEB, M0303) for removal of DNA and re-purified by RNA Clean & Concentrator-5 kit (Zymo Research, R1015).

For single nuclei preparation, we followed the guideline of Nuclei EZ Prep kit (Sigma, NUC-101) and resuspended the nuclei into DPBS. Both single cells and single nuclei were sorted by FACSAria SORP flow cytometer (BD Biosciences).

Mice

For samples of hippocampus nuclei and lymphocytes, aged (2-months old) male C57BL/6 mice were used and obtained from Charles River Laboratories.

For samples of germinal center B cells, C57BL/6 mice were originally from the Jackson Laboratory. 6-12 week-old, age- and sex-matched mice were used for the experiments.

All mice were maintained under specific pathogen-free conditions and used in accordance of governmental, Capital Medical University, and Tsinghua University guidelines for animal welfare.

Hippocampal nuclei preparation and sorting

The isolated hippocampus tissue was transferred into a Dounce homogenizer (Sigma, D8938) containing 2ml of EZ Lysis Buffer (Sigma, NUC-101). The tissue was carefully dounced for 22 times with pestle A followed by 22 times with pestle B, then transferred to a 15ml-tube. Next, 1ml of EZ lysis buffer was added into the Dounce homogenizer to resuspend residual nuclei, then transferred to the same 15ml tube. The samples were centrifuged at 300g for 5 min. Supernatant was removed and the pellet was resuspended in 100μl of ice-cold PBS (Gibco, 10010023) with 1% BSA (NEB, B9000S) and 20U RRI (Takara, 2313). 40μm FlowMi cell strainers were firstly wetted with PBS and filtered the resuspended nuclei into 1.5 ml Eppendorf tubes. The nuclei were further washed by PBS (1% BSA).

To enrich neuron nuclei, 1,000-fold diluted mouse Anti-NeuN antibody (Millipore, MAB377) was added into 0.5ml nuclei suspension and incubated with the nuclei at 4°C for 30min. The nuclei were then stained with 1000-fold diluted Goat anti-Mouse IgG (H&L) antibody (Abcam, ab150113) and washed with PBS (1% BSA). The whole process was on ice. As gated in **Extended Data Fig. 6**, single neuron nuclei were sorted with FACSAria SORP flow cytometer.

For frozen samples, hippocampus tissues were previously flash frozen in liquid nitrogen, and stored in -80°C. Before single nuclei preparation, they were thawed on ice totally.

GC light zone B cells preparation and sorting

To generate T-cell dependent GC responses in B6 mice, 100μg NP-KLH (Biosearch Technologies, N-5060-5) plus 1μg LPS (Sigma, L6143) emulsified in 100μl 50% alum (Thermo, 77161) were utilized for intraperitoneal immunization.

Spleens isolated from 4 mice of 13-days post immunization were placed on a 70μm cell strainer (Falcon, 08-771-2), which was previously wetted with MACS buffer (1% FBS and 5mM EDTA in PBS), and minced by flat end of the plunger of 2ml syringes (Becton Dickinson, 301940). The splenocytes passed through the strainer with MACS buffer into a 50ml-tube. The mixed red blood cells were then lysed by ACK lysis buffer (Thermo, A1049201). The cell suspension was further incubated with biotinylated 4-Hydroxy-3-iodo-5-nitrophenylacetyl (NIP)15-BSA (Biosearch Technologies, N-1027-5) for 1.5h, and enriched by Anti-biotin cell isolation kit (Miltenyi Biotec, 130-090-485) to get NP-reactive cells.

The enriched cells were blocked with 20μg/ml 2.4G2 antibody (BioXCell, BE0307) and subsequently stained with APC-Cy7 (anti-B220, BD Biosciences, 552094), PE-Cy7 (anti-CD95, BD Biosciences, 557653), eF450 (anti-GL7, eBioscience, 48-5902-82), APC (anti-CD86, eBioscience, 17-0862-82) and PE (anti-CXCR4, BioLegend, 146505). Also, 7-AAD (Biotium, 40037) was stained to exclude dead cells. All staining reactions were incubated in MACS staining buffer (1% FBS and 5mM EDTA in PBS) for 30min on ice, followed by 3 times of washings. As gated in **Extended Data Fig. 8**, single GC Light Zone B cells (B220$^+$ GL7$^+$ Fas$^+$ CD86$^+$ CXCR4$^-$) were sorted into in lysis buffer using Aria **III** flow cytometer (BD Biosciences).

Lymphocyte cells preparation and sorting

The retro-orbital blood was taken from the eyeball of ether-anesthetized mice and dipped into K2EDTA tube (BD Vacutainer, 367525). PBS was added to dilute blood at ~50%. 1ml diluted blood was transferred into a clean 15ml-tube and incubated with 9ml 1x red blood cells lysing solution (BD Pharm Lyse, 555899) at room temperature for 15min avoiding light. The resulted cell suspension was washed twice by PBS containing 1% BSA at 200g for 5min, followed by staining with SYTOX green (Thermo, S7020) to identify intact cells. Single lymphocytes were sorted with FACSAria SORP flow cytometer according to the gates shown in **Extended Data Fig. 8**.

DNA carrier preparation

100-ng pTXB1 plasmids were firstly linearized by 10U XbaI (NEB, R0145S) at 37°C for 1h and purified by Zymo columns. Then we took 0.5-ng linearized plasmids for multiple displacement amplification (MDA), with all dTTPs replaced by dUTPs. Specifically, the 1μl DNA was mixed with 22μl reaction buffer containing 1x phi29 reaction buffer (NEB, M0269S), 20μM random primers (Thermo, SO181) and 1mM dNTP (NEB, N0446S and N0459S), then they were incubated at 98°C for 3min and immediately cooled down at 4°C for 20min. 2μl phi29 DNA polymerase was added and the amplification was carried out at 30°C for 5h, terminated at 65°C for 10min. The products were purified by Zymo columns.

Generation of RNA-seq library

We constructed NEBNext libraries with 200- and 10-ng RNA by following the guideline of NEBNext Ultra II RNA Library Prep Kit for Illumina kit (NEB, E7770). SmartSeq2 libraries with single cells were prepared following the protocol that was reported by Picelli, S. *et al*. 10X libraries of 10,000 single hippocampal nuclei were constructed by Chromium Single Cell 3' Reagent Kits (v3.1).

For scRNA-seq library of SHERRY2, single cells were sorted into 96-well plates containing 2μl lysis buffer which consisted of 0.5% Triton X-100 (Sigma, T9284), 2U SUPERaseIn RNase Inhibitor (Thermo, AM2694), 0.2U AG DNase I (Thermo, 18068015). The plates were immediately spun down and incubated at 20°C for 10min for DNA digestion. The plates could be stored at -80°C or proceeded with next step. 2μl inactivation buffer containing 5μM OligodTs (T_{30}VN, Sangon), 5mM dNTPs and 1mM EDTA (Thermo, AM9260G) was then added and the reaction was incubated at 65°C for 10min and 72°C for 3min to facilitate RNA denaturation at the same time. Next, RT was performed by adding 6μl RT mix (70U SuperScript IV (Thermo, 18090050), 1.7x SSIV buffer, 8.3mM DTT, 10U RRI, 1.7M Betaine (Sigma, B0300)) and incubated at 50°C for 50min, then inactivated the reverse transcriptase at 80°C for 10min. The resulted RNA/DNA hybrids mixed with 10-pg DNA carriers were tagmented by 0.05μl TTE Mix V50 (Vazyme, TD501) at 55°C for 30 min, through adding 10μl reaction mix containing 2x TD buffer (20mM Tris-HCl (ROCKLAND, MB-003), 10mM MgCl2 (Thermo, AM9530G), 20% N,N-Dimethylformamide (Sigma, D4551)), 16% PEG8000 (VWR Life Science, 97061), 0.5mM ATP (NEB, P0756), 8U RRI. 6U Bst 3.0 DNA polymerase (NEB, M0374M) within 1 x Q5 high-fidelity master mix were utilized to repair the gap left by V50 at 72°C for 15min, followed by 80°C for 5min to terminate the reaction. Finally, 3μl indexed primers mix (Vazyme, TD203) and 3μl Q5 mix were added to perform PCR amplification. PCR cycled as following: 98°C 30s for initial denaturation, 21 cycles of 20s at 98°C, 20s at 60°C and 2min at 72°C, 72°C 5min for final extension. The indexed products were merged and purified at 0.75x with VAHTS DNA Clean Beads (Vazyme, N411).

Libraries were quantified with Qubit 2.0 and their fragment length distributions were checked by Fragment Analyzer Automated CE System. Libraries were sequenced by Illumina NextSeq 500 or NovaSeq S4.

RNA-seq data analysis

Data quality. Adaptors, poly(A/T) sequences were trimmed, bases with quality less than 20 and reads shorter than 20 bases were removed from the raw sequencing data by Cutadapt (v1.15). Clean reads were mapped to indexed genome (human: Gencode.v31, mouse: Gencode.vM23) by STAR (2.7.1a). Only unique alignment was utilized for downstream analysis. The mitochondrial and ribosomal ratios were counted with samtools (v1.10). The ratios of coding region, UTR, intron and intergenic region were counted with Picard tools (v2.17.6). Exonic rate was defined as sum of coding region and UTR ratios. For cells, Cufflinks (v2.2.1) with exon annotations of protein coding genes were used to count gene number (FPKM>1). For nuclei, genes (RPM>1) were counted by featureCounts (v1.5.1) with transcript annotations. Coverage across gene body was calculated by RSeQC (v.2.6.4). The coverage uniformity was integral area between coverage curve and x-axis normalized by 100.

Clustering and marker genes. For scRNA-seq and snRNA-seq, clustering followed the basic tutorials of Scanpy (v1.8.1). The cell type annotations were through manually checking expression of well-known marker genes. Marker genes that identified by SHERRY2 should satisfy following conditions: 1) adjusted p-values calculated by Mann-Whitney-U test were less than 1e-3; 2) foldchanges were greater than 1.5 or less than 0.67; 3) The average normalized counts of up-regulated gene in the cell type, or down-regulated gene in the rest of cell types was greater than 0.3. For NEBNext, DESeq2 (v1.22.2) was utilized to identify the differentially expressed genes (adjusted p-value < 1e-4, foldchange > 2).

RNA Velocity. Splicing and unsplicing mRNA were quantified by Velocyto (v0.17.17) with unique alignment. The generated loom file was utilized by scVelo (v0.2.4) to profile velocity dynamics based on clustering results of Scanpy.

BCR assembly. BCR sequences of each cell was assembled by MIXCR (v3.0.13) with clean reads. The assembled BCR were realigned by IgBlast (v1.17.1) to determine clone types.

Exon usage. The frequency of exon usage in each cell was calculated by BRIE (v2.0.5). For each exon, cells satisfying following conditions were retained: 1) counts of gene which included the exon were greater than 10; 2) exon regions sided by the specific exon should be covered by greater than 50% with uniquely aligned reads; 3) at least one read should detect junctions involved in this exon splicing events. Pairwise comparison of exon usage frequency was made between cell types which contained greater than 10 cells using Mann-Whitney-U test. The exons with p-value less than 0.05 was further checked in IGV viewer to check whether transcript coverage was consistent with usage frequency. The passed ones were considered as significantly biased among cell types.

Acknowledgement

We thank Chenyang Geng and Yan Chen from the Peking University High-throughput Sequencing Center and Biomedical Pioneering Innovation Center for experimental assistance. This work was supported by the National Key

Research and Development Program of China (2018YFA0108100 to Y.H.), National Natural Science Foundation of China (22050002, 21927802 to Y.H.), Beijing Municipal Science and Technology Commission (Z201100005320016 to Y.H.), Beijing Advanced Innovation Center for Genomics, and Shenzhen Bay Laboratory.

Author contributions

Y.H. and J.W. conceived the study; L.D., B.L., Y.L., S.Z., Y.P., and J.S. performed exepriemtns; L.D. performed data analyses; C.Z. and J.S. provided samples; L.D., B.L., J.W., H.Q., J.S., and Y.H. wrote manuscript with input from all authors; Y.H., J.W., and H.Q. supervised all aspects of this study.

References

1 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5**, 621-628 (2008).

2 Cui, P. *et al.* A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics* **96**, 259-265 (2010).

3 Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods* **10**, 1096-1098 (2013).

4 Bagnoli, J. W. *et al.* Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nature communications* **9**, 1-8 (2018).

5 Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 1-12 (2017).

6 Perez, J. D. *et al.* Subcellular sequencing of single neurons reveals the dendritic transcriptome of GABAergic interneurons. *Elife* **10**, e63092 (2021).

7 Oguchi, Y., Ozaki, Y., Abdelmoez, M. N. & Shintaku, H. NanoSINC-seq dissects the isoform diversity in subcellular compartments of single cells. *Science Advances* **7**, eabe0317 (2021).

8 Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology* **37**, 38-44 (2019).

9 Trapnell, C. & Cacchiarelli, D. Monocle: Differential expression and time-series analysis for single-cell RNA-Seq and qPCR experiments. *Bioconductor. Fmrp. Usp. Br* **2** (2014).

10 La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494-498 (2018).

11 Wang, X., He, Y., Zhang, Q., Ren, X. & Zhang, Z. Direct comparative

analyses of 10X Genomics Chromium and smart-seq2. Genomics, Proteomics & Bioinformatics (2021).

12 Di, L. *et al.* RNA sequencing by direct tagmentation of RNA/DNA hybrids. *Proceedings of the National Academy of Sciences* **117**, 2886-2893 (2020).

13 Chen, C. *et al.* MINERVA: a facile strategy for SARS-CoV-2 whole-genome deep sequencing of clinical samples. *Molecular cell* **80**, 1123-1134. e1124 (2020).

14 Gerard, G. F., Collins, S. & Smith, M. D. Excess dNTPs minimize RNA hydrolysis during reverse transcription. *Biotechniques* **33**, 984-990 (2002).

15 Chandler, D. P., Wagnon, C. A. & Bolton Jr, H. Reverse transcriptase (RT) inhibition of PCR at low concentrations of template and its implications for quantitative RT-PCR. *Applied and environmental microbiology* **64**, 669-677 (1998).

16 Habib, N. *et al.* Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science* **353**, 925-928 (2016).

17 Yamawaki, T. M. *et al.* Systematic comparison of high-throughput single-cell RNA-seq methods for immune cell profiling. *BMC genomics* **22**, 1-18 (2021).

18 MacLennan, I. C. Germinal centers. Annual review of immunology 12, 117-139 (1994).

19 Victora, G. D. & Nussenzweig, M. C. Germinal centers. Annual review of immunology 30, 429-457 (2012).

20 Allen, C. D. *et al.* Germinal center dark and light zone organization is mediated by CXCR4 and CXCR5. *Nature immunology* **5**, 943-952 (2004).

21 Caron, G., Le Gallou, S., Lamy, T., Tarte, K. & Fest, T. CXCR4 expression functionally discriminates centroblasts versus centrocytes within human germinal center B cells. *The Journal of Immunology* **182**, 7595-7602 (2009).

22 Victora, G. D. *et al.* Germinal center dynamics revealed by multiphoton microscopy with a photoactivatable fluorescent reporter. *Cell* **143**, 592-605 (2010).

23 Victora, G. D. SnapShot: the germinal center reaction. *Cell* **159**, 700-700. e701 (2014).

24 Mesin, L., Ersching, J. & Victora, G. D. Germinal center B cell dynamics. *Immunity* **45**, 471-482 (2016).

25 De Silva, N. S. & Klein, U. Dynamics of B cells in germinal centres. *Nature reviews immunology* **15**, 137-148 (2015).

26 Kennedy, D. E. *et al.* Novel specialized cell state and spatial compartments within the germinal center. *Nature immunology* **21**, 660-670 (2020).

27 Green, J. A. *et al.* The sphingosine 1-phosphate receptor S1P 2 maintains the homeostasis of germinal center B cells and promotes niche confinement. *Nature immunology* **12**, 672-680 (2011).

28 Huang, C. & Melnick, A. Mechanisms of action of BCL6 during germinal center B cell development. *Science China Life sciences* **58**, 1226-1232 (2015).

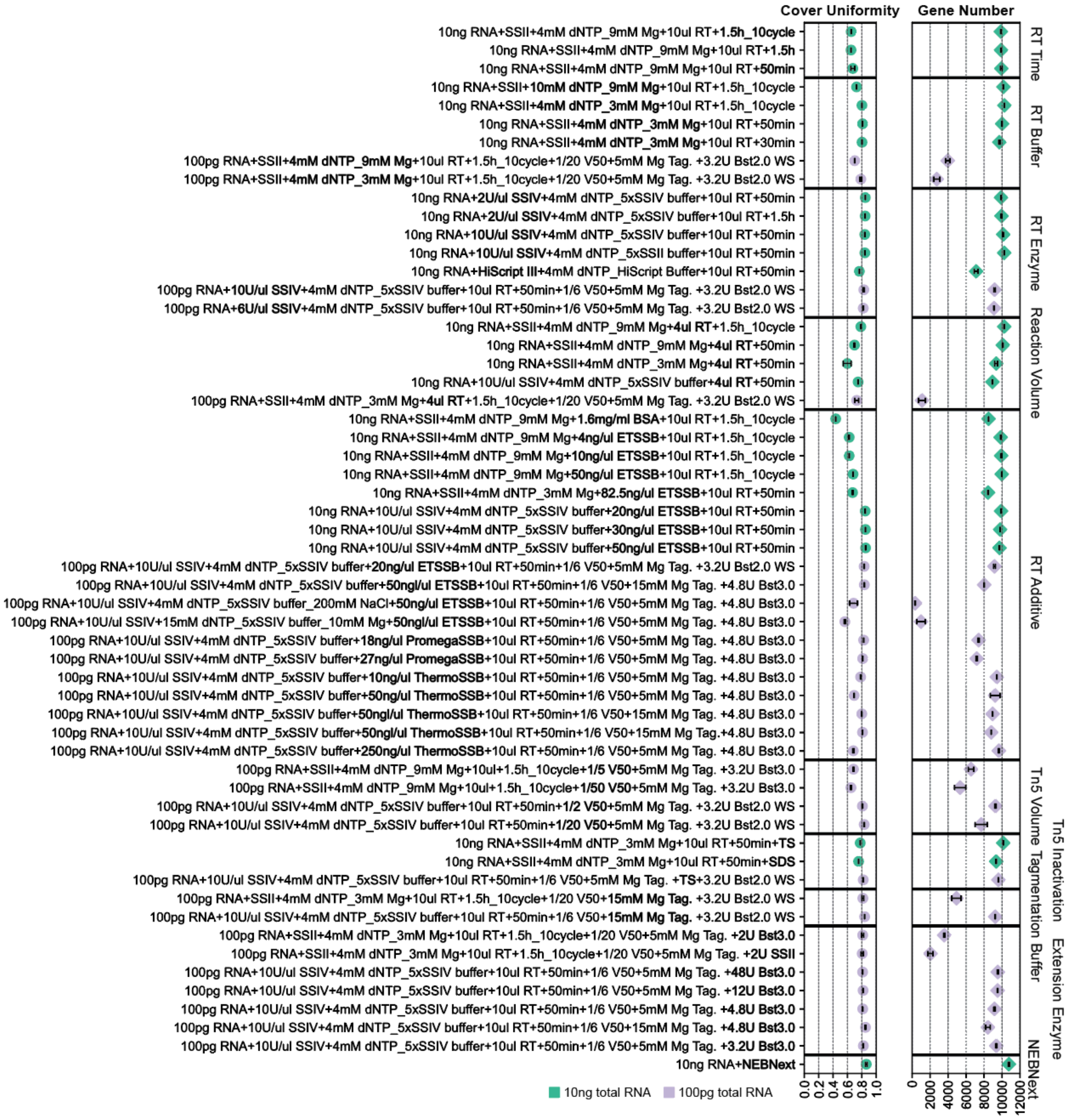
29 Suan, D. *et al.* CCR6 defines memory B cell precursors in mouse and human germinal centers, revealing light-zone location and predominant low antigen affinity. *Immunity* **47**, 1142-1153. e1144 (2017).

30 Ise, W. *et al.* T follicular helper cell-germinal center B cell interaction strength regulates entry into plasma cell or recycling germinal center cell fate. *Immunity* **48**, 702-715. e704 (2018).

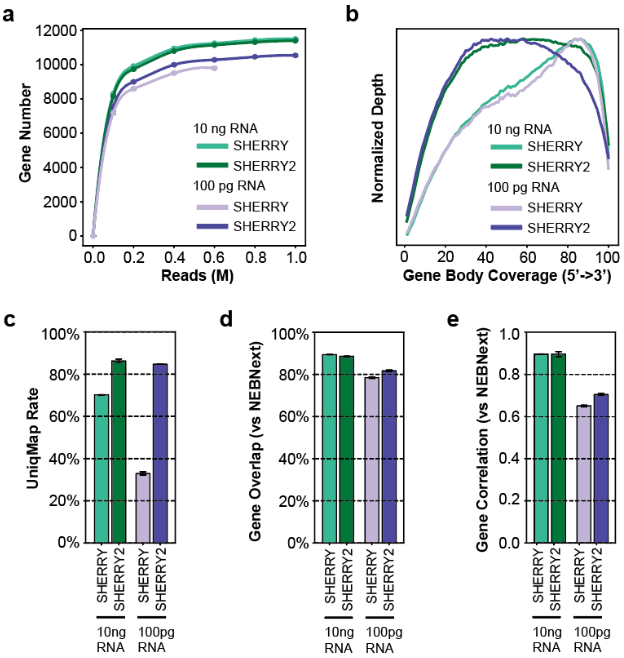
31 Dominguez-Sola, D. *et al.* The proto-oncogene MYC is required for selection in the germinal center and cyclic reentry. *Nature immunology* **13**, 1083-1091 (2012).

32 Laidlaw, B. J., Duan, L., Xu, Y., Vazquez, S. E. & Cyster, J. G. The transcription factor Hhex cooperates with the corepressor Tle3 to promote memory B cell development. *Nature immunology* **21**, 1082-1093 (2020).

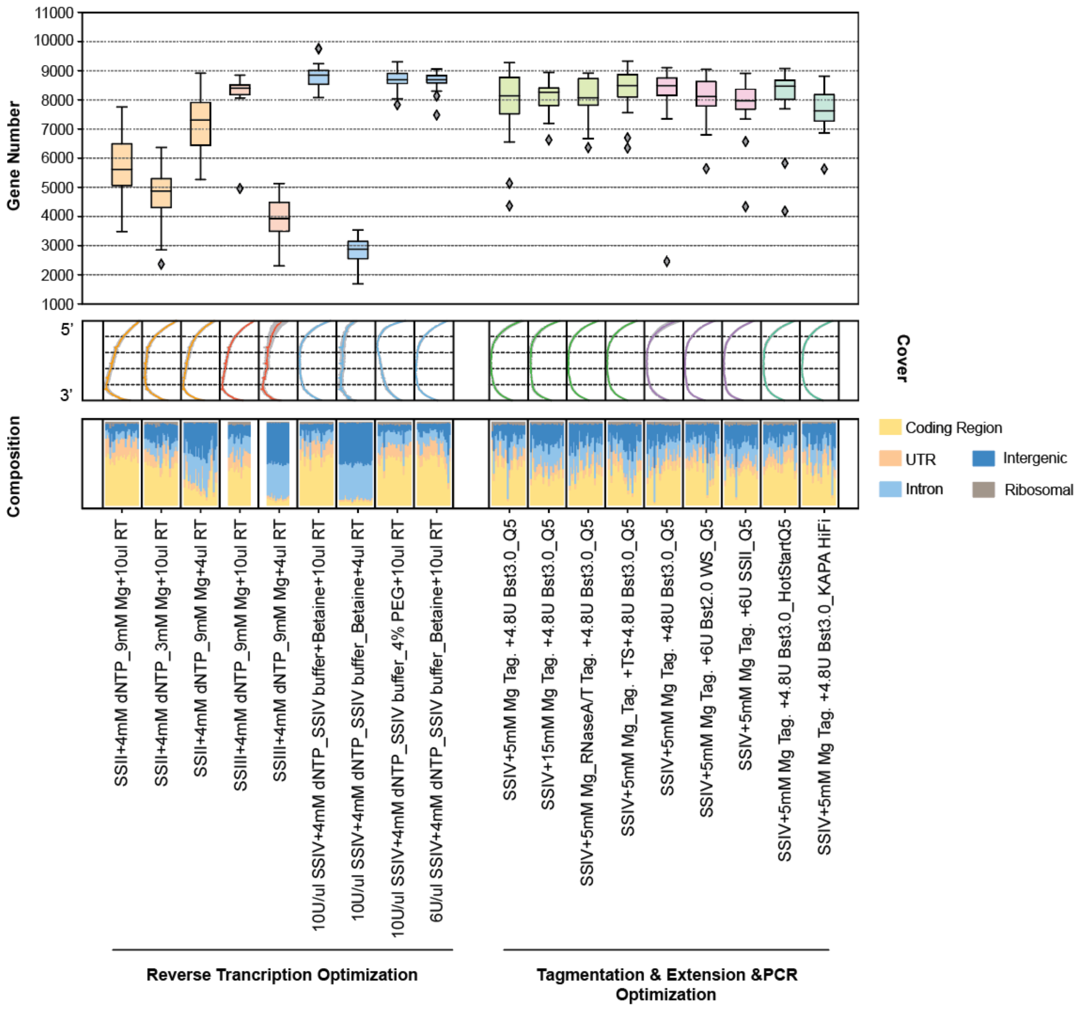
33 Jacob, J. & Kelsoe, G. In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl) acetyl. II. A common clonal origin for periarteriolar lymphoid sheath-associated foci and germinal centers. *The Journal of experimental medicine* **176**, 679-687 (1992).



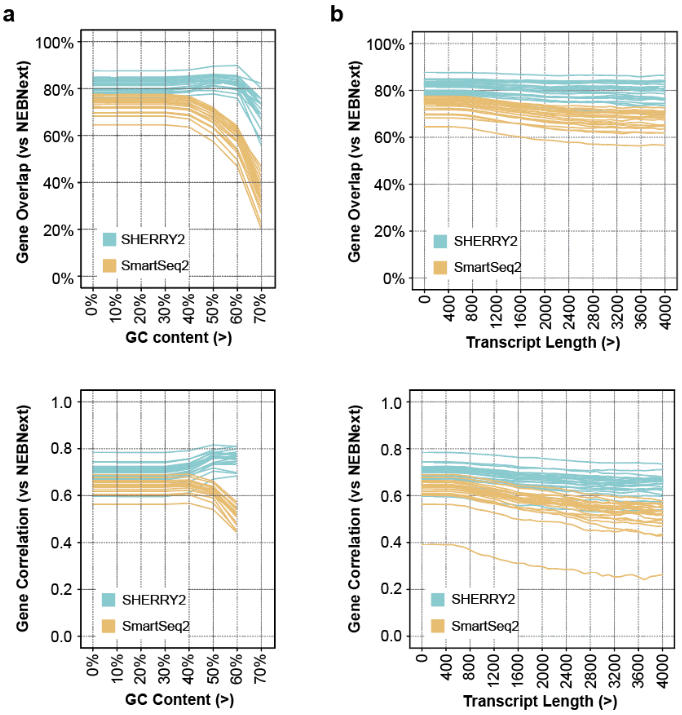
Extended Data Fig. 1 Conditions for bulk RNA. RNA-seq library quality of 10-ng (green) and 100-pg (purple) HEK293T RNA samples when testing different conditions of the SHERRY protocol. The coverage uniformity across gene body and gene number (FPKM>1) at 0.25-million reads are shown here. The bolded fonts of the labels indicate key variations compared to the basic protocol. The NEBNext standard protocol was performed with 10-ng HEK293T RNA as a reference. The error bars show the standard deviation.
]



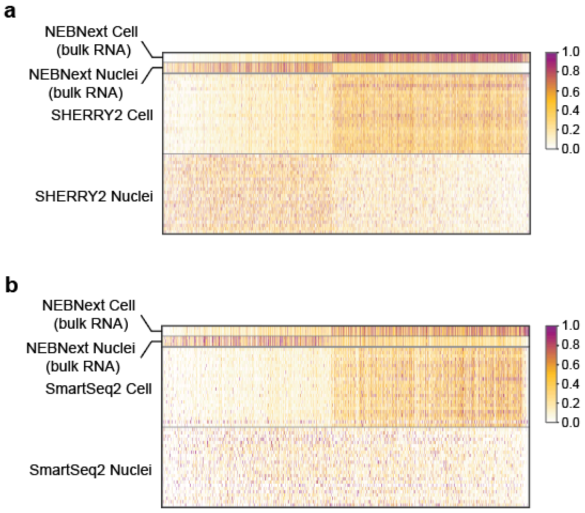
Extended Data Fig. 2 Improved performance of SHERRY2 on bulk RNA. a, Saturation curve of SHERRY and SHERRY2. The gene number (FPKM>1) is shown for sequencing at 0.2, 0.4, 0.6, 0.8, and 1.0 million reads. **b,** Gene body coverage detected by SHERRY and SHERRY2. **c,** Ratio of reads that uniquely mapped to the human genome. **d,** Ratio of expressed genes (FPKM>1) in the NEBNext results that overlapped the SHERRY and SHERRY2 results. **e,** Correlation of overlapped gene counts between NEBNext and the two protocols: SHERRY and SHERRY2. The inputs of SHERRY and SHERRY2 in **a-e** were 10-ng and 100-pg HEK293T RNA. The input of NEBNext in **d** and **e** was 200-ng HEK293T RNA. The error bars in **c-e** show the standard deviation.



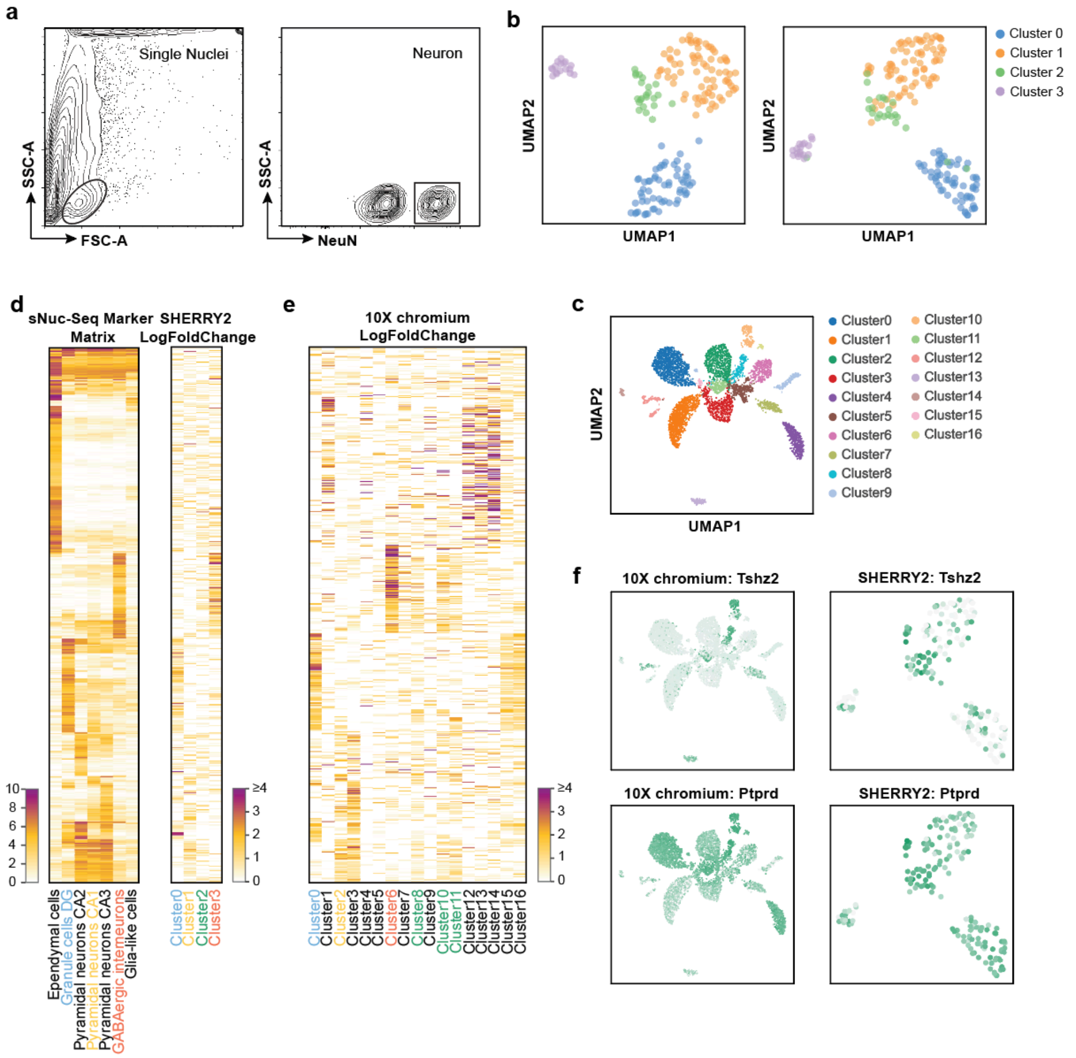
Extended Data Fig. 3 Conditions for single cells. scRNA-seq library quality of single HEK293T cells when testing different conditions of SHERRY protocol. The gene number (FPKM>1) at 0.25-million reads, gene body coverage and composition of reads mapped to different genome regions are shown. The colors of the boxplot and coverage plot indicate key variations in different steps. The gray region of the coverage plot shows the standard deviation of normalized depth among replicates.



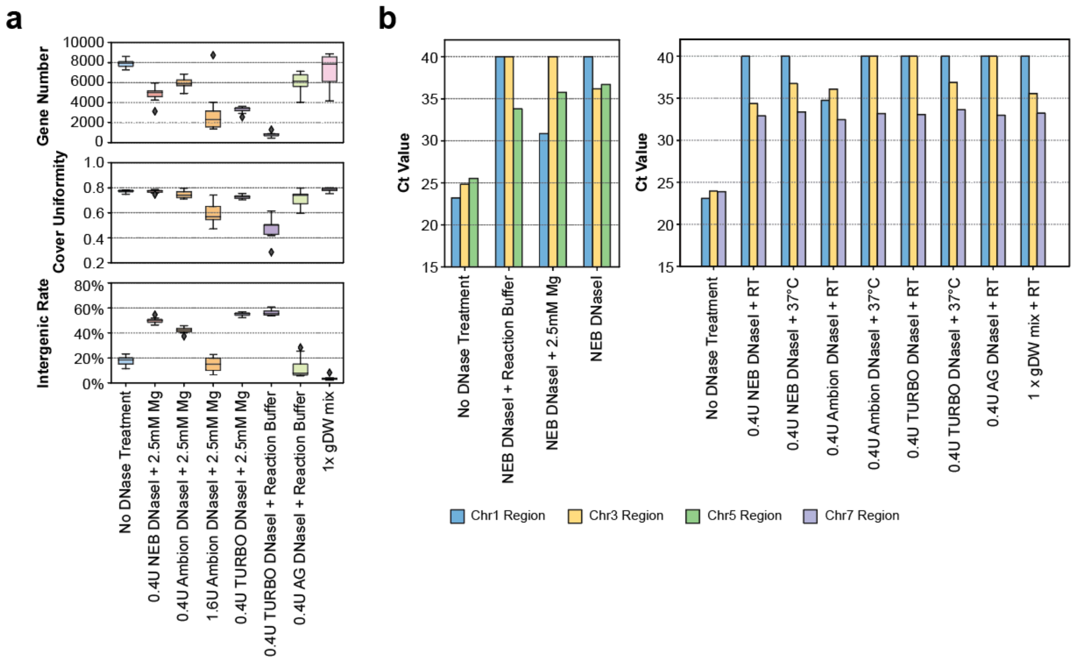
Extended Data Fig. 4 Accuracy of SHERRY2 for single cells. a, Ratio of expressed genes (FPKM>1) with different GC-content in the NEBNext results that overlapped by SHERRY2 and SmartSeq2, and the correlation of overlapped gene expression results between NEBNext and the two scRNA-seq methods. The x-axis indicates the minimum level of GC-content. Fewer than 100 genes had GC content >70%, and they were not included in the analysis. Each line represents one cell. **b,** Gene overlap ratios and correlation of gene counts between NEBNext and the two scRNA-seq methods, utilizing genes with different transcript lengths from the NEBNext results. The inputs of SHERRY2 and SmartSeq2 were single HEK293T cells. The input of NEBNext was 200-ng HEK293T RNA. The correlation R-value was calculated by a linear fitting model.



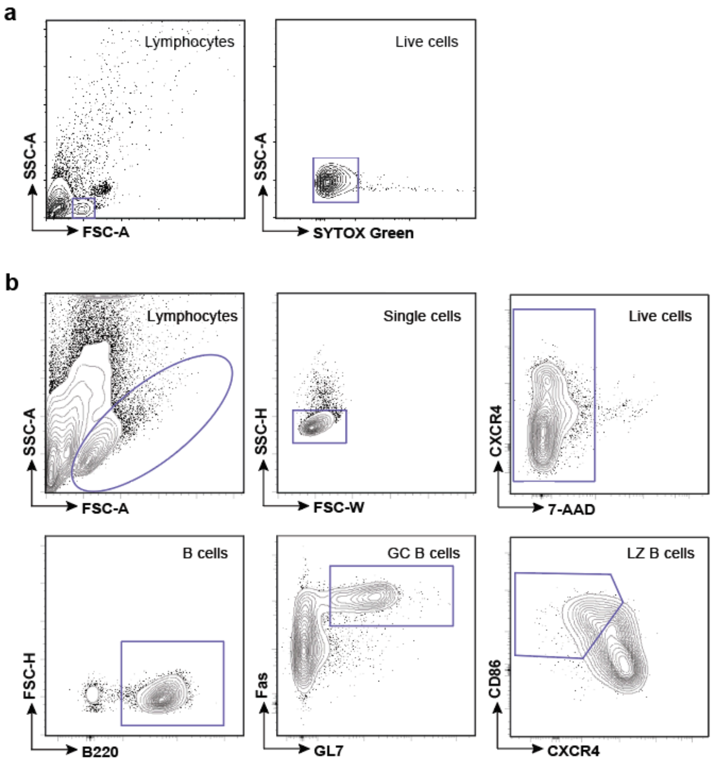
Extended Data Fig. 5 Expression pattern of differential genes between cells and nuclei in SHERRY2. Heatmap of the counts of differentially expressed genes between HEK293T cells and nuclei in the NEBNext, SHERRY2 (**a**) and SmartSeq2 (**b**) results. The differentially expressed genes were identified by the NEBNext method with 200-ng RNA input and sorted by log2 fold-change. The inputs of SHERRY2 and SmartSeq2 were single HEK293T cells and nuclei. Each line represents a replicate.



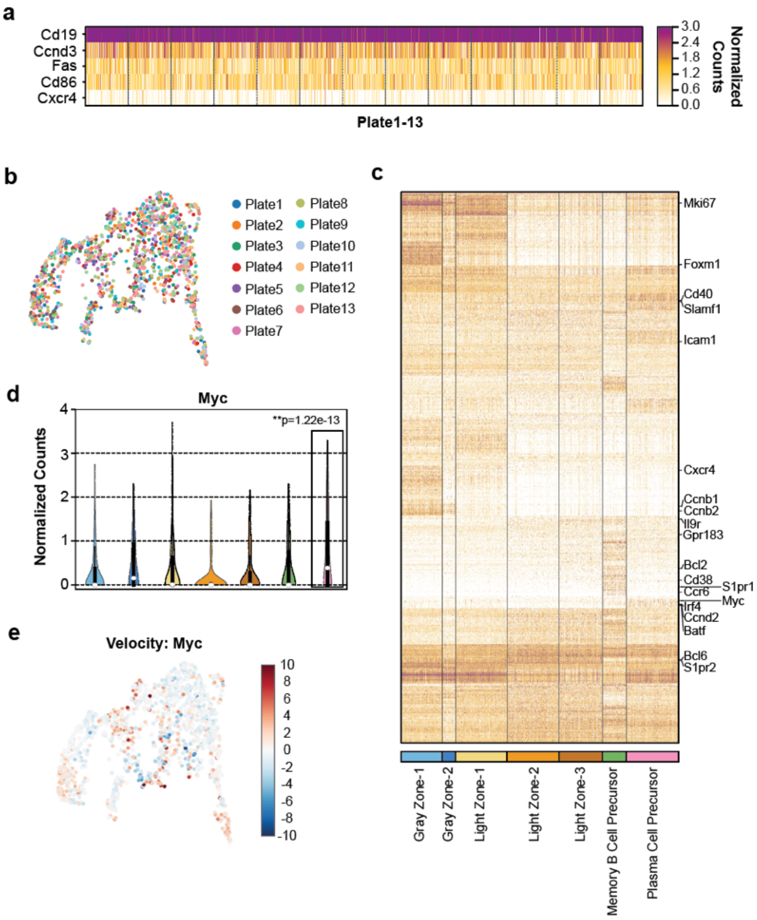
Extended Data Fig. 6 Clustering and annotation of hippocampal nuclei. a, Flow cytometry gating of single neuron nuclei that were isolated from mouse hippocampus. **b,** The left panel shows a UMAP projection of the neuron nuclei clustered using marker genes of hippocampal cell types reported by Habib, N. *et al*[15]. The right panel shows the nuclei positions in **Fig. 2e**, and each nucleus is colored using the same labeling scheme used in the left panel. The snRNA-seq libraries were prepared by SHERRY2. **c,** Clustering of single hippocampal nuclei visualized by UMAP plot. The libraries were prepared by 10X Chromium Single Cell 3′ Reagent Kits. **d,** Heatmaps of hippocampal maker gene counts acquired from Habib, N. *et al*[15] and their log fold change across clusters that were identified by SHERRY2 (**Fig. 2e**). Clusters with similar gene expression patterns are labeled with the same colors. **e,** Heatmaps of gene log fold change across clusters that were identified by 10X Chromium in **c**, with the same genes and sorting order shown in **d**. Clusters that exhibited an expression pattern similar to that of Cluster 2 in **Fig. 2e** are labeled in green. **f,** Potential marker genes of Cluster 2 in **Fig. 2e** and a UMAP plot of their expression in the SHERRY2 and 10X Chromium results.



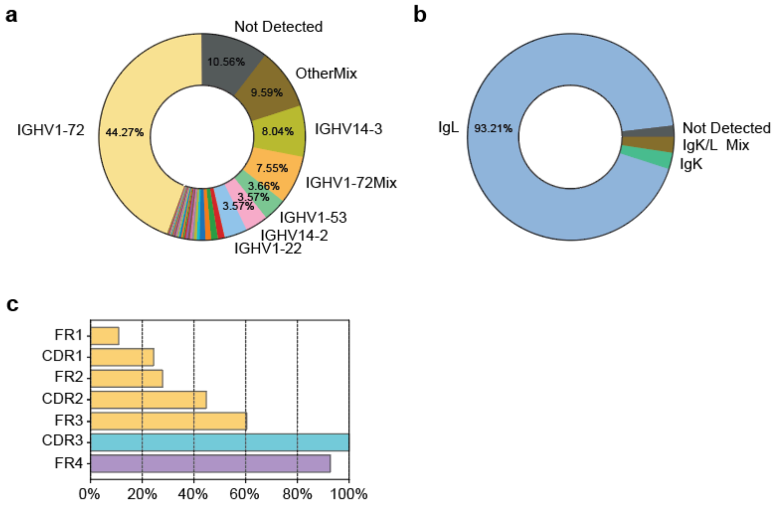
Extended Data Fig. 7 DNase activity and performance in SHERRY2 library construction. a, Library quality of SHERRY2 tested with single HEK293T cells and different DNases. The gene number (FPKM>1) at 0.25-million reads, gene body coverage uniformity and percentage of reads that were mapped to intergenic regions are shown. The x-axis labels indicate the amount of each DNase and its reaction buffer. **b,** qPCR quantification of 15-ng digested HEK293T genomic DNA by 0.4U NEB DNase I with the recommended reaction buffer, either containing 2.5 mM Mg^{2+} or without any divalent ions. The digested products were split into three parts, and ct values were measured using primers designed for the chr1, chr3 or chr5 region. **c,** qPCR quantification of 30-ng digested HEK293T genomic DNA. The x-axis labels indicate the amounts of each DNase and the reaction temperature (RT indicates room temperature). The digestions were all performed without any added divalent ions. The digested products were split into three parts, and ct values were measured using primers designed for the chr1, chr3 or chr7 region.



Extended Data Fig. 8 Flow cytometry gating of single lymphocytes and single GC B cells. Flow cytometry gating of single lymphocyte cells that were isolated from murine eyeball blood (**a**) and single B cells from murine GC light zones (**b**).



Extended Data Fig. 9 GC information gained from SHERRY2. a, Heatmap of the counts of the gating genes shown in **Extended Data Fig. 8b**, which were used to isolate single B cells from murine GC light zones. **b,** Single cells sorted in different 96-well plates were labeled with different colors on a UMAP plot showing clusters from the GC light zone in **Fig. 4a**. **c,** Up- or down-regulated genes (adjusted p-value < 1e-3, fold change > 1.5 or <0.67) identified by SHERRY2 across GC light zone cell types. The labeled genes were previously reported for the cell types. **d,** Violin plot of gene *Myc* counts within each GC light zone cell type. The colors correspond to the cell types annotated in **Fig. 4a**. The p-values between plasma cell precursors and other cell types were calculated by the Mann-Whitney-U test. **e,** RNA velocity of the *Myc* gene projected on a UMAP plot.



Extended Data Fig. 10 BCR sequences of single GC B cells identified by SHERRY2. a, Usage frequency of IgH variable genes that were assembled from the reads of each single GC light zone B cell. “Mix” indicates that more than 1 IgH variable gene was assigned, or more than 1 heavy chain sequence was assembled and also assigned to different genes, with a frequency less than 80%. “IGHV1-72Mix” indicates that one of the assigned genes was IGHV1-72. **b,** Usage frequency of light chain types that paired with IGHV1-72 sequences. **c,** Proportions of cells that covered different regions within IgH variable genes among cells for which IgH sequences could be assembled.