scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Short-read and long-read full-length transcriptome of mouse neural stem cells across neurodevelopmental stages

Chaoqiong Ding[1,4], Xiang Yan[1,4], Mengying Xu[2,4], Ran Zhou[2], Yuancun Zhao[2], Dan Zhang[2], Zongyao Huang[1], Zhenzhong Pan[1], Peng Xiao[1], Huifang Li[3], Lu Chen[2] ✉ & Yuan Wang[1] ✉

During brain development, neural stem cells (NSCs) undergo multiple fate-switches to generate various neuronal subtypes and glial cells, exhibiting distinct transcriptomic profiles at different stages. However, full-length transcriptomic datasets of NSCs across different neurodevelopmental stages under similar experimental settings are lacking, which is essential for uncovering stage-specific transcriptional and post-transcriptional mechanisms underlying the fate commitment of NSCs. Here, we report the full-length transcriptome of mouse NSCs at five different stages during embryonic and postnatal development. We used fluorescent-activated cell sorting (FACS) to isolate CD133$^+$Blbp$^+$ NSCs from C57BL/6 transgenic mice that express enhanced green fluorescent protein (EGFP) under the control of a Blbp promoter. By integrating short- and long-read full-length RNA-seq, we created a transcriptomic dataset of gene and isoform expression profiles in NSCs at embryonic days 15.5, 17.5, and postnatal days 1.5, 8, and 60. This dataset provides a detailed characterization of full-length transcripts in NSCs at distinct developmental stages, which could be used as a resource for the neuroscience community to study NSC fate determination, neural development, and disease.

Background & Summary

During mammalian brain development, neural stem cells (NSCs) give rise to major cell types in various brain regions, including neurons and glial cells (astrocytes and oligodendrocytes). Although embryonic and postnatal NSCs share molecular markers such as brain lipid-binding protein (Blbp, also known as fatty acid-binding protein 7, Fabp7) and CD133 (also known as Prominin-1), their cellular identities and fates vary significantly at different developmental stages. Embryonic NSCs are radial glial cells in the ventricular zone (VZ), which initially generate neurons in different cortical layers, and subsequently, undergo neuron-glia fate-switch to produce astrocytes and oligodendrocytes at late embryonic and perinatal stages[1]. After birth, a subset of radial glial cells transform into postnatal NSCs in the subventricular zone (SVZ) and subgranular zone (SGZ) in the hippocampus, which continue to generate interneurons and glia[2]. In the adult brain, the majority of NSCs in the SVZ and SGZ are committed to neuronal fate[3,4]. These fate switches in NSCs are driven by dramatic transcriptional alterations. Extensive efforts have been made to characterize human and mouse brain cells including NSCs at bulk and single-cell levels during neurodevelopment[5–7]. However, partly due to the scarcity of NSCs, full-length transcriptomic datasets of NSCs across different neurodevelopmental stages under similar experimental settings are lacking, which is essential for uncovering stage-specific transcriptional and post-transcriptional mechanisms underlying the fate commitment of NSCs.

Smart-seq2 is a powerful single-cell full-length sequencing protocol, which provides complete coverage across the genome allowing the detection of alternative transcript isoforms and SNPs[8,9].

[1]Department of Neurosurgery, State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, Sichuan University and National Collaborative Innovation Center, Chengdu, 610041, China. [2]Key Laboratory of Birth Defects and Related Diseases of Women and Children of MOE, State Key Laboratory of Biotherapy, West China Second Hospital, Sichuan University, Chengdu, 610041, China. [3]Core Facilities of West China Hospital, Sichuan University, Chengdu, China. [4]These authors contributed equally: Chaoqiong Ding, Xiang Yan, Mengying Xu. ✉e-mail: luchen@scu.edu.cn; wangyuan@scu.edu.cn

This protocol can also be adapted for full-length bulk RNA-seq of rare cell populations, such as NSCs. However, for the conventional 2nd-generation RNA-seq, cDNA generated from Smart-seq2 is fragmented before sequencing, resulting in accurate short-read raw data, which complicates the task of reconstructing and quantifying transcript isoforms. Long-read sequencing, or 3rd-generation sequencing, on the other hand, does not require cDNA fragmentation and provides a complete picture of the transcriptome at the cost of the sequencing accuracy. Combining short- and long-read sequencing can draw on their respective strengths.

In this study, we used fluorescence-activated cell sorting (FACS) to isolate CD133$^+$Blbp-EGFP$^+$ NSCs from C57BL/6 transgenic mice at five different stages of embryonic and postnatal development, including embryonic day 15.5 (E15.5, the peak of cortical neurogenesis), E17.5 (the transition to gliogenesis), postnatal day 1.5 (P1.5, neonatal stage), P8 (the peak of postnatal NSC proliferation and gliogenesis), and P60 (adult). We used the Smart-seq2 protocol to prepare the cDNA samples of NSCs at these stages, and performed a total of 20 short-read RNA-seq with at least three samples per stage, along with paired Oxford Nanopore long-read RNA-seq for each stage. The whole study design of the present study is present in Fig. 1a. The resultant dataset provides a detailed characterization of full-length transcripts in NSCs at distinct developmental stages in a similar experimental setting and could be used as a resource to study NSC fate determination, neural development, and disease.

Methods

Animals.

The Blbp-EGFP mice used in this study were initially generated by Anthony *et al.* at The Rockefeller University and obtained from Dr. Yuan Zhu's lab at Children's National Medical Center in Washington, DC under a material transfer agreement with Sichuan University. The mice were bred in the Experimental Animal Centre of Sichuan University and maintained on a C57BL/6 genetic background. Mice were housed in pressurized, individually ventilated cages (PIV/IVC) and maintained under specific-pathogen-free conditions, with free access to food and water in a 12 h light/dark cycle. All animal studies were approved by the Animal Care and Use Committee of Sichuan University. For timed pregnancies, the plug date was designated as E0.5 and the date of birth was defined as P0.5.

Sample collection and FACS.

To collect embryonic NSCs, embryonic brains were placed into pre-chilled 10% FBS solution (10% FBS in DPBS, Gibco), and the dorsal wall of the LV was dissected out under a dissecting microscope (Motic). The tissue was dissociated by pipetting, and the cells were filtered through a 40 μm nylon mesh cell strainer (BD Falcon) to prepare single-cell suspension.

Postnatal brains were placed into 10% FBS solution, cut into coronal slices, and the SVZ region was harvested, minced into small pieces, and dissociated with Accutase solution (Millipore) at 37 °C for 20 min. The resultant cells were filtered through a 40 μm nylon mesh cell strainer (BD Falcon) to prepare single-cell suspension.

NSCs were stained with viability marker fixable viability stain 510 (FVS510, BD Horizon, 564406) and NSC marking antibody CD133-APC (Abcam, ab19898), and subjected to FACS. 1000 FVS510$^-$/CD133$^+$/Blbp-EGFP$^+$ cells from each mouse were collected for subsequent RNA-seq (Fig. 1b). The FACS plots for all samples are presented in Figshare[10]. The cells for 10X single-cell RNA sequencing are also derived from FACS with the same sorting strategy as bulk RNA-seq.

NSC Culture.

Sorted FVS510$^-$/CD133$^+$/Blbp-EGFP$^+$ cells were cultured in NSC culture medium (1% N2, 2% serum-free B27, 20 ng/ml EGF and 20 ng/ml bFGF in DMEM/F12) in 6-well ultra-low binding plates (Corning). Neurospheres are visible after 4 days of culture. (Fig. 1c).

cDNA library construction and sequencing.

cDNA preparation. cDNA preparation was modified from a published protocol which was originally used for single-cell RNA sequencing[11]. Briefly, all the components of lysis buffer (TritonX-100, dNTP, Oligo-dT VN primer, and RNase inhibitor) were 2X except RNase inhibitor which was increased to 8X,resulting in a total volume of 8.8 μL. The components for RT-PCR reaction mix were increased accordingly. NSCs were collected in tubes containing lysis buffer and were immediately transferred onto dry ice. The lysate was vortexed vigorously for 1 min followed by incubation at 72°C for 3 min, and subjected to RT-PCR. Reverse transcription mixture was prepared by mixing 1.6 μL SuperScript II reverse transcriptase, 1.6 μL RNase inhibitor, 6.5 μL Superscript II first-strand buffer, 1.6 μL DTT, 6.5 μL betaine, 0.2 μL $MgCl_2$, 0.3 μL TSO and 0.9 μL nuclease-free H_2O to reach a total volume of 19.2 μL. Cell lysate was mixed with reverse transcription mixture and incubated at 42°C for 90 min, followed by 10X RT-PCR cycles: ① 50 °C for 2 minutes ②42 °C for 2 minutes. Afterwards, the reverse transcribed cDNA samples were incubated at 70 °C for 15 min. For additional PCR amplification, 33.5 μL cDNA was mixed with 33.5 μL KAPA HiFi HotStart ReadyMix and 0.7 μL ISPCR primers to a total volume of 67.7 μL. The mixture was first incubated at ③ 98 °C for 3 minutes, followed by 20X PCR cycles: ① 98 °C for 20 seconds, ② 67 °C for 15 seconds and ③ 72 °C for 6 minutes. The amplified cDNA samples were incubated at 72 °C for 5 minutes. cDNA purification was carried out with Ampure XP magnetic beads (0.8:1 ratio, Beckman Coulter, A63881). Before library construction, cDNA quality was checked with Agilent 2100 Bioanalyzer (Invitrogen). Library construction was performed with qualified cDNA for both short-read and long-read sequencing. The 10X single cell RNA-seq were prepared in the Chromium Single Cell Gene Expression Solution using the Chromium Single Cell 3′ Gel Bead, Chip and Library Kits v2 (10X Genomics) as per the manufacturer's protocol. 8000–10,000 total cells were added to each channel. The cells were then partitioned into Gel Beads in Emulsion in the Chromium instrument, where cell lysis and barcoded reverse transcription of RNA occurred, followed by amplification, shearing 5′ adapter, and sample index attachment. Libraries were sequenced on the Illumina NovaSeq 6000 platform at Novogene, Beijing, China[12].

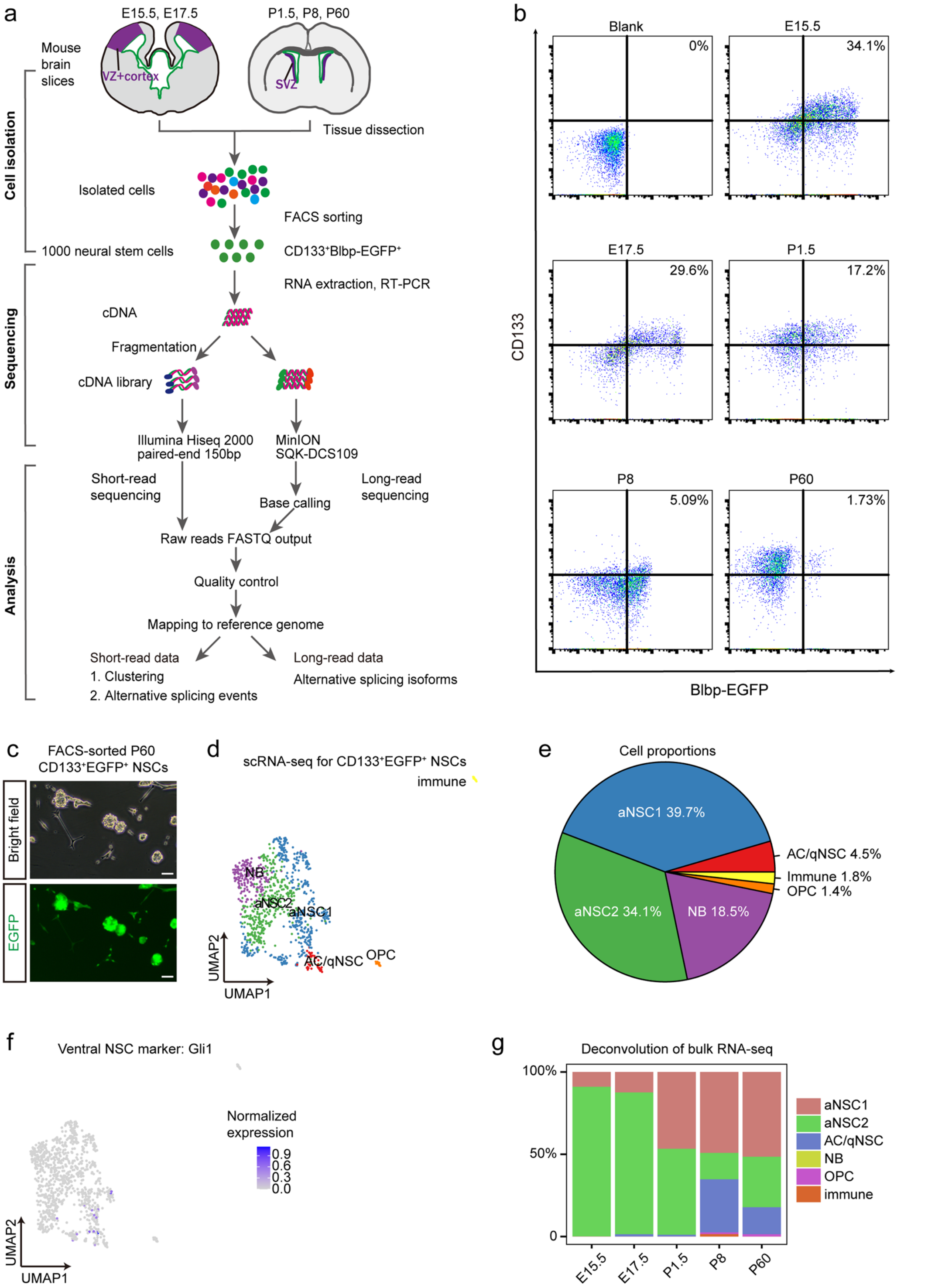

Fig. 1 Study design, NSC isolation and validation. (**a**) The study design includes cell isolation, sequencing, and analysis. (**b**) Representative flow cytometric results of NSC sorting include blank samples and samples at each stage. (**c**) P60 CD133$^+$Blbp-EGFP$^+$ cells from FACS were cultured into neurosphere after 4-day *in vitro* and carried green fluorescence EGFP. Scale bar, 100 µm. (**d**) UMAP analysis integrating high-quality single cells from sorted NSCs samples at P60. Clusters are separated by colours. (**e**) The Pie chart shows the proportion of each cell type by UMAP clusters. (**f**) UMAP shows the expression of Gli1, a marker for ventral NSCs, in the NSC populations. (**g**) Deconvolution of bulk RNA-seq. The percentage of cell types at five time points.

Short-read sequencing. Qualified cDNA samples were respectively taken for short-read library construction, including DNA fragmentation, end-repair, 3′ ends A-tailing, adapter ligation, PCR amplification, and library validation. cDNA library was subjected to quality inspection with PerkinElmer LabChip® GX Touch. Qualified libraries were then loaded on the Illumina Hiseq platform for PE150 sequencing.

Long-read sequencing. The cDNA samples for long-read sequencing were taken from the same pools for short-read sequencing. Equal content of cDNA samples of the same stage were mixed into one sample. ONT Ligation Sequencing Kit (SQK-LSK109) was used for library preparation according to the manufacturer's instructions except that DNA was not sheared before native barcode ligation. The library construction included DNA repair, end preparation, native barcode ligation, purification pooling, and sequencing adapter ligation. The cDNA libraries were pooled evenly at the amount of 80 ng from each stage. Sequencing was performed using MinKNOW (v20.06.4, Oxford Nanopore Technologies Ltd.). MinKNOW is the instrument control software that runs on the host computer to which the MinION equipped with an R9.4.1 flow cell is connected. The data output from MinKNOW consist of 4,000 sequence reads in an HDF5 format called FAST5.

Validation of splice junctions by PCR. To validate NSCs time points-specific splice junctions we designed exon-specific PCR primers for 8 stage-specific alternative splicing events, including two types of splice junctions: Alternative 3′splice site (A3SS) and Skipping exon (SE). Pooled cDNA libraries for NSCs from each time point (E15.5, E17.5, P1.5, P8 and P60) were mixed with the primers and subjected to 30X PCR cycles. Each PCR reaction contained 25 ng of library DNA template and 10 pmol of each gene specific primer in a PCR master mix (2X Phusion Plus Green, Thermo Scientific) at a total volume of 50 µl. PCR products were subjected to electrophoresis separation in 2% TBE/agarose gels. The images were captured and gel band intensity was calculated by Image J. The PCR PSI is calculated as the intensity of the long transcript divided by the total intensity of the long + short transcripts.

Gene Name	**Forward Primer (5′→3′)**	**Reverse Primer (5′→3′)**
Capzb	GCACGCTGAATGAGATCTAC	GCGTGGTCGATGCAAACTG
Dync1i2	CTATGTCTCCATCCTCCAAGTC	GGTCTGAGTTTCCTTTGTGTATG
Gkap1	CTCCCGCTCCAGAGCACAAC	GGACCGTCAGCTCGTTCTTG
Gpm6b	CATGTCCTATCACCTGTTCATTG	CAGTTCCTGCTCTTCCTTTGC
Hnrnpdl	CCAGAACAATTACCAGCCCTAC	GAGTCATCATAACACAGGTAGC
Nkain4	GTCTATGGTTGCTACGTGGTCAG	CTCACAGTTGTAGCCACCCTGTC
Abat	GAGAACGGTGGCTGGAATCATCG	GCAGGTCTTCCCGCTTGATGATG
Sox5	CACCAGGCTTAGGCCCACTC	CAGAGCTGGCATGTGAGGAGAG

Data processing.

Deconvolution of bulk RNA-seq. We used MuSiC to deconvolute the transcriptome of Bulk RNA-Seq samples into the likely constituent cell types, using scRNA-seq datasets from same samples as Bulk RNA-Seq as a reference. We calculated the predicted proportions of each cell type in bulk samples, and visualized these proportions with bar plot[12].

ScRNA-seq data analysis. We used Seurat (v3.1.0) for downstream analyses including data normalization (NormalizeData, LogNormalize method, scaling factor 10,000), data feature scaling (ScaleData), variable gene detection (FindVariableGenes with vst method) and PCA of variable genes (RunPCA). The statistically significant PCs were used for Harmony to remove the batch effect, and the two-dimension UMAP was calculated among the Harmony matrix[13]. Then the original Louvain algorithm (FindClusters) with clustering resolution 1.4 was performed to cluster the cells. We computed DEGs using the FindAllMarkers function in the Seurat package with default parameters. To determine the cell types, we used the list of DEGs and the published dataset of marker genes[10,14].

Base calling. The raw data generated by MinKNOW software were converted from.fast5 files to base-called. fastq files under high accuracy mode using the ONT basecaller Guppy software (v.4.0.14)[15]. In the meantime, the sample barcodes were trimmed off with the modes '--barcode_kits' and '--trim_barcode'.

Quality control. The quality of the short-read sequencing data was checked using FastQC software (v0.11.8) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and RSeQC package (v4.0.0)[16] (http://rseqc.sourceforge.net). For long-read sequencing data, the quality check was performed with NanoComp software (v1.33.1)[17].

Alignment. The paired-end reads of short-read sequencing were aligned to the mouse reference genome GRCm38 with annotation from ENSEMBLE release 93 using STAR (v2.7.1a)[18], and the reads of long-read sequencing were aligned to the same genome file using Minimap2 (v2.17-r974-dirty)[19].

Aligned reads distribution. Gene body coverage, reads' distribution over genome feature, and RNA integrity at cDNA level of short-read sequencing data were calculated by geneBody_coverage.py, read_distribution.py, and tin.py from RSeQC[16], respectively.

Gene expression quantification. For short-read sequencing data, the gene expression was quantified using the HTSeq (v0.11.2)[20]. The raw read counts were then normalized by their library size factors and were normalized to stabilize the variance across the samples using DESeq 2 (v1.28.1)[21] with variance stabilizing transformation (VST). The top 500 highly variable genes were utilized for unsupervised clustering analysis.

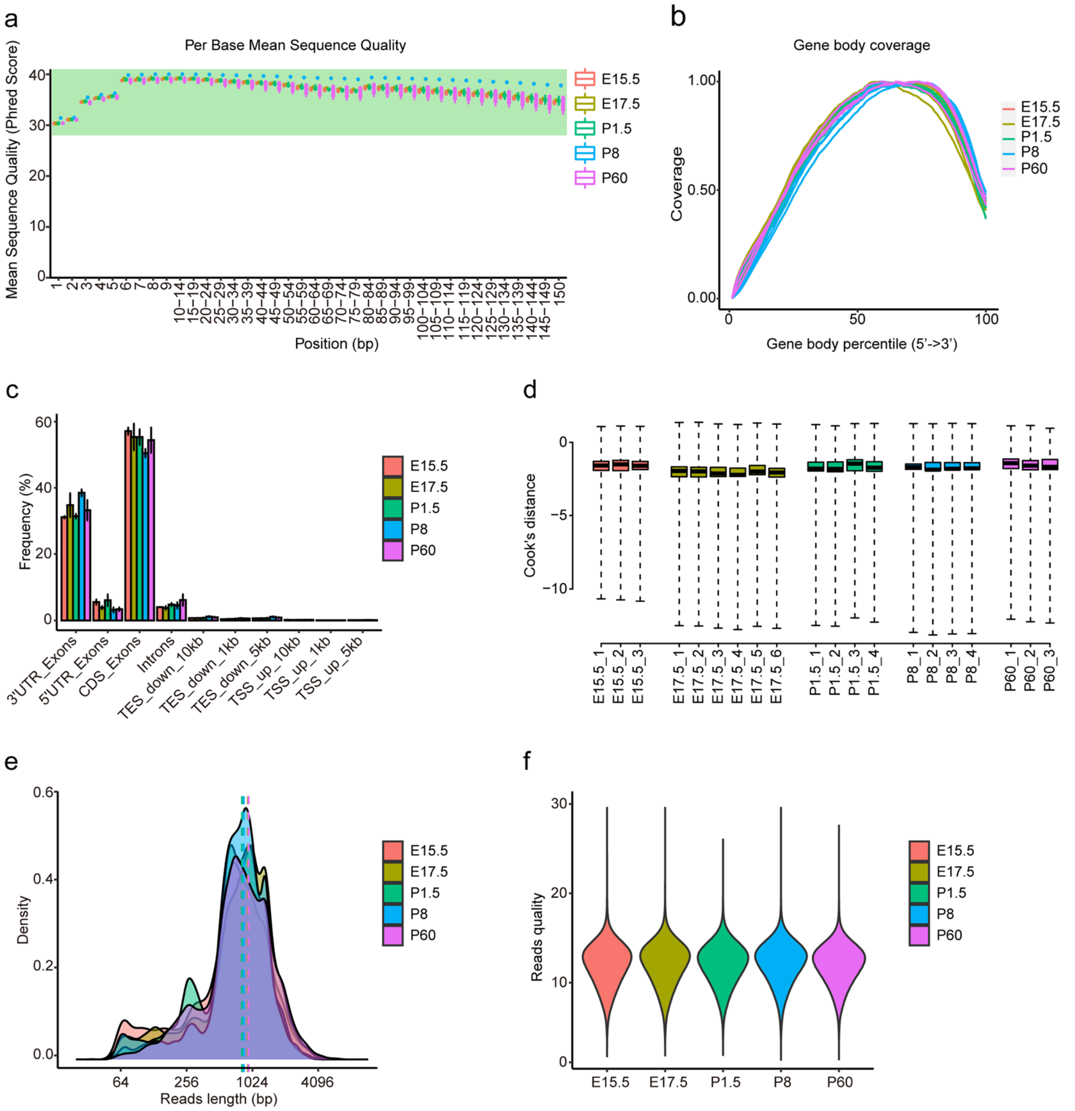

Fig. 2 Quality control of short-read and long-read sequencing data. (**a**) Average sequencing quality per base of each sample. (**b**) Reads distribution along the gene body. (**c**) The frequency of counts in various gene regions and the error bars depict the standard deviation. (**d**) Cook's distance was calculated for each sample. (**e**) Distribution of reads length of randomly selected 50,000 reads from all samples. (**f**) Base calling quality score of each sample.

Differential splicing usage. The percent-spliced-in (PSI, also denoted Ψ) value was calculated according to a previous study[22] and recapitulated here. The PSI metric was computed directly by counting reads that aligned to known or predicted splicing junctions (SJs) generated from STAR[18]. The significance of enrichment was tested by a two-tailed hypergeometric test[23]. The functional annotations of each AS event were performed by using the GenomicRanges R package (v1.40.0, findOverlaps module)[24] and mouse reference genome GRCm38 with annotation from ENSEMBLE release 93.

Identification of AS modes. The resulting alignments (in BAM format) of long-read sequencing data were used to build sample-specific transcriptome assembled with Stringtie2 (v2.1.4)[25]. The mode -G was specified to use the mouse reference, and the mode -L was specified in long reads format. GffCompare[26] was used to compare and evaluate the accuracy of Stingtie2[25] transcript assemblers. The consolidated set of accurate isoforms (GTF format) were used to obtain a list of all possible AS events, and SUPPA2 generateEvents mode[23] was used to generate all AS events with the parameter of "-f ioe -e SE SS MX RI FL".

Transcript visualization. For visualization of transcripts, samtools (v1.10.2)[27] was used to extract the gene region, bedtools (v2.30.0)[28] was used to convert bam file to bed format files, UCSC tools were used to convert bed format files to GTF format. Visualization was carried out by using R package ggbio (v1.36.0)[29]. Sashimi plots of short-read sequencing data were plotted using pysashimi (https://github.com/ygidtu/pysashimi).

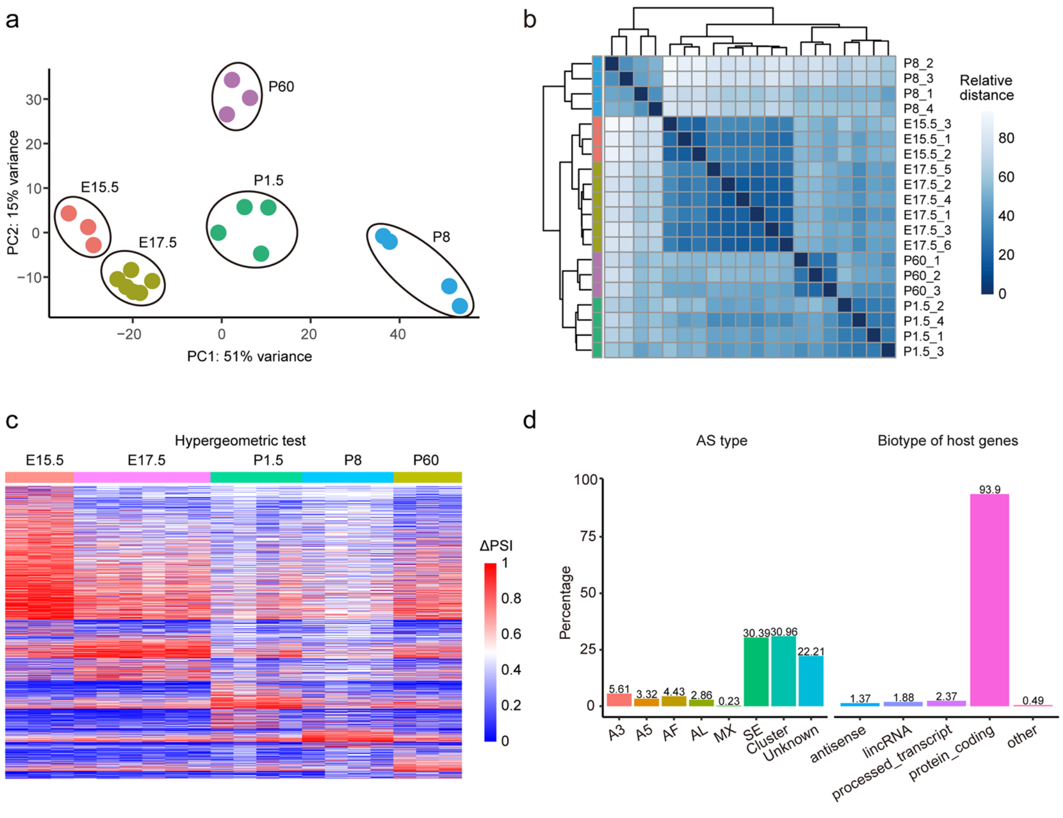

Fig. 3 Samples clustered by gene expression and the differential splicing usage based on short-read sequencing. (**a**) PCA analysis for all samples. (**b**) Hierarchical clustering analyses for each sample. (**c**) Heat-map visualizing the differential SJs' PSIs in samples from each stage. (**d**) The proportion of AS types and biotypes of the host genes for differential SJs.

Data Records

The raw fastq files were deposited at NCBI under accession number SRP321063[30]. The FACS data for individual samples, as well as the processed files, including the quantification of gene expression, isoforms, SJs from both short-read and long-read data were uploaded in Figshare[10]. We also included detailed gene expression matrix and cell type determination of a validation scRNA-seq dataset in Figshare[10].

Technical Validation

NSC purity. To confirm our sorted cells are indeed mainly composed of NSCs, we first performed neurosphere culture assay. Sorted cells readily formed Blbp-EGFP$^+$ neurospheres after 4-day non-adherent culture *in vitro*, indicating they are neural stem or progenitor cells (Fig. 1c). We further performed single-cell RNA-seq on CD133$^+$Blbp-EGFP$^+$ cells from P60 SVZ to identify individual cell types in these cells[10]. The majority of the cells are aNSCs (73.8%) (Fig. 1d,e). The qNSC population is relatively minor, and is transcriptionally hard to distinguish from astrocytes. Consistent with our sampling of dorsal SVZ regions, Gli1, a marker for ventral NSCs, is barely detected in the NSC populations (Fig. 1f). To determine the NSC percentage in bulk samples, we performed deconvolution of bulk RNA-seq using our scRNA-seq dataset as a reference. Consistently, the majority of the cells are aNSC-like (Fig. 1g). These data support that our sample collection method can enrich for NSCs.

RNA integrity. As the mRNA was reversely transcribed immediately after bulk NSCs were lysed, the mRNA integrity could not be measured directly. We performed the quality check by examining the fragment distribution of cDNA. It turned out that all peaks of the sample cDNA were longer than 1200 bp (Online-only Table 1). The RNA integrity at the transcript level was further evaluated using the Transcript Integrity Number (TIN) algorithm, which was calculated with the tin.py script from the RSeQC package[16]. TIN represents a score ranging from 0 to 100 for each expressed transcript, and the medTIN (median TIN score across all the transcripts) can be used to measure the RNA integrity at the sample level. The mean TIN score of all samples was 43.87 (Online-only Table 1).

Data quality. Biological replicates are fundamental to guarantee data reliability. In the present study, we took 3 E15.5 samples, 6 E17.5 samples, 4 P1.5 samples, 4 P8 samples, and 3 P60 samples for bulk transcriptome sequencing. The average depth was 25.99 M (SD = 11.78) for the short-read sequencing. The quality of each base generated was assessed using FastQC. There is no significant difference in the distribution of average quality score per base in samples from different stages, and the mean of Q30 is over 100% (Fig. 2a). The reads

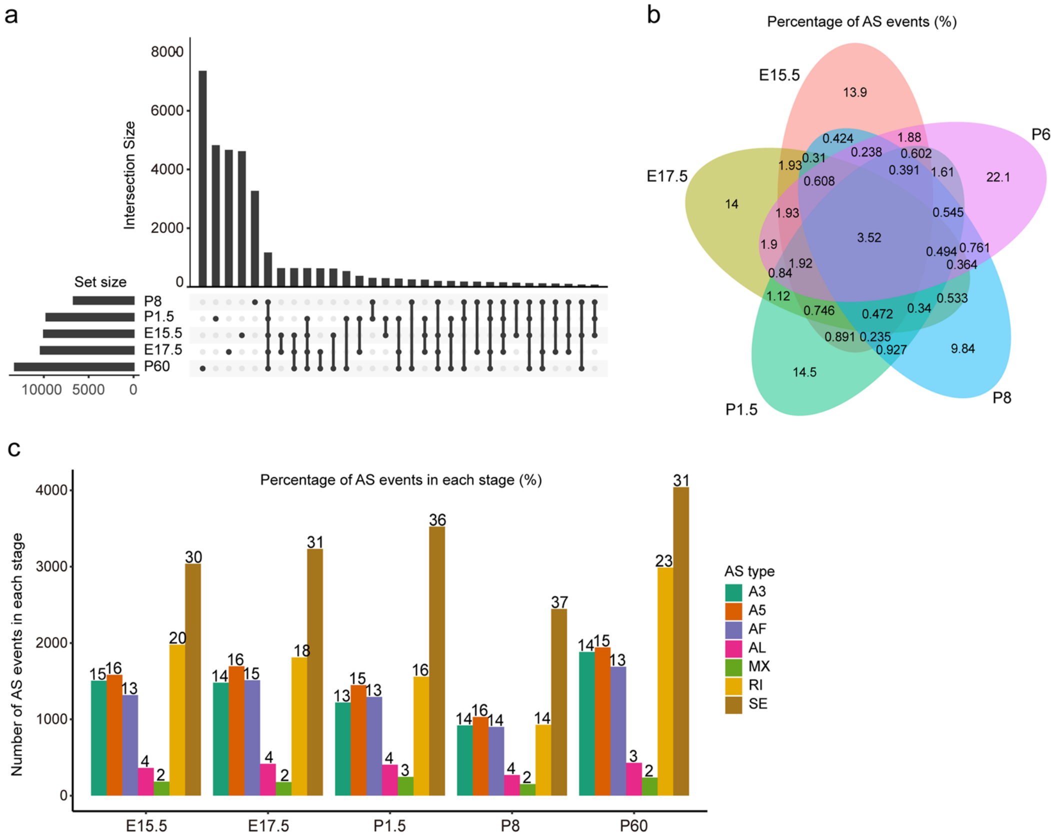

Fig. 4 AS event analysis based on long-read sequencing data. (**a**) The upset plot shows the numbers of specific and total AS events in all stages. (**b**) Venn diagram shows the percentages of specific and overlapped AS events in all stages. (**c**) The numbers (y-axes) and percentages (above the histogram) of detected AS types in each stage.

generated from all samples were distributed approximately uniform across the gene body (Fig. 2b). We further gathered the gene regions where the reads mapped to, and more than 90% of reads were mapped to exon regions (Fig. 2c). Moreover, Cook's distance was calculated to test for outliers, with none detected (Fig. 2d). All the samples have over 80% uniquely mapped reads (Online-only Table 1). Besides, Q30 of each sample is higher than 85% (Online-only Table 1).

For long-read sequencing data, we selected 20.6 Gb reads using Guppy[15] from 77 Gb raw data and identified 12,387,984 reads. The sequencing statistics were counted using NanoComp[17]. The mean read length of all the five stages was around 800 bp, and the P60 sample had the maximum mean length (Fig. 2e, Online-only Table 1). On the aspect of base calling quality, the mean read quality score was above 12 for each sample (Fig. 2f, Online-only Table 1).

To establish the congruency of short-read data among all stages, we carried out principal component analysis (PCA) (Fig. 3a) and hierarchical clustering (Fig. 3b) using the top 500 highly variable genes from normalized RNA-seq data with variance stabilizing transformation (VST) in DESeq 2[21]. The PC1 explained 51% of the variance, while the PC2 explained 15% variance. PCA revealed that stage E15.5 was close to E17.5, while stage P1.5 was close to P60 (Fig. 3a). Besides, the samples of P8 were distant from all the other stages (Fig. 3a). Hierarchical clustering showed a similar result to that of PCA (Fig. 3b).

Differential splicing usage. As an index of AS, the PSI value was calculated for the inclusion levels of internal exons, as described in a previous study[22]. Differential PSI was calculated via a two-tailed hypergeometric test. The heatmap (Fig. 3c) showed all the 4403 differential alternative SJs in 5 stages with the P-value < 0.01, $\Delta PSI > 0.2$, and these SJs were detected in more than 60% samples of a target group and existed in other groups in short-read sequencing data. To further understand the differential splicing usage, we analyzed the AS types and the host genes of differential SJs. The SJs with single skipping exon (SE) (30.39%) were much more than SJs with other singles, including alternative 3′ splice site (A3), alternative 5′ splice site (A5), alternative first exon (AF), alternative last exon (AL) and mutually exclusive exons (MX) (Fig. 3d). Besides, there were 22.21% SJs with unknown AS types and clusters (multi-AS types). Most host genes (93.9%) were protein-coding genes, with only 2.37% host genes expressing processed transcripts, 1.88% expressing lincRNA, and 1.37% as antisense genes (Fig. 3d). The left 0.49% host genes belong to other types (Fig. 3d).

Stage-specific AS. For long-read sequencing data, seven types of AS were quantified to analyse the relative contribution of AS at all five stages. The quantification of each type of AS event was performed by using the

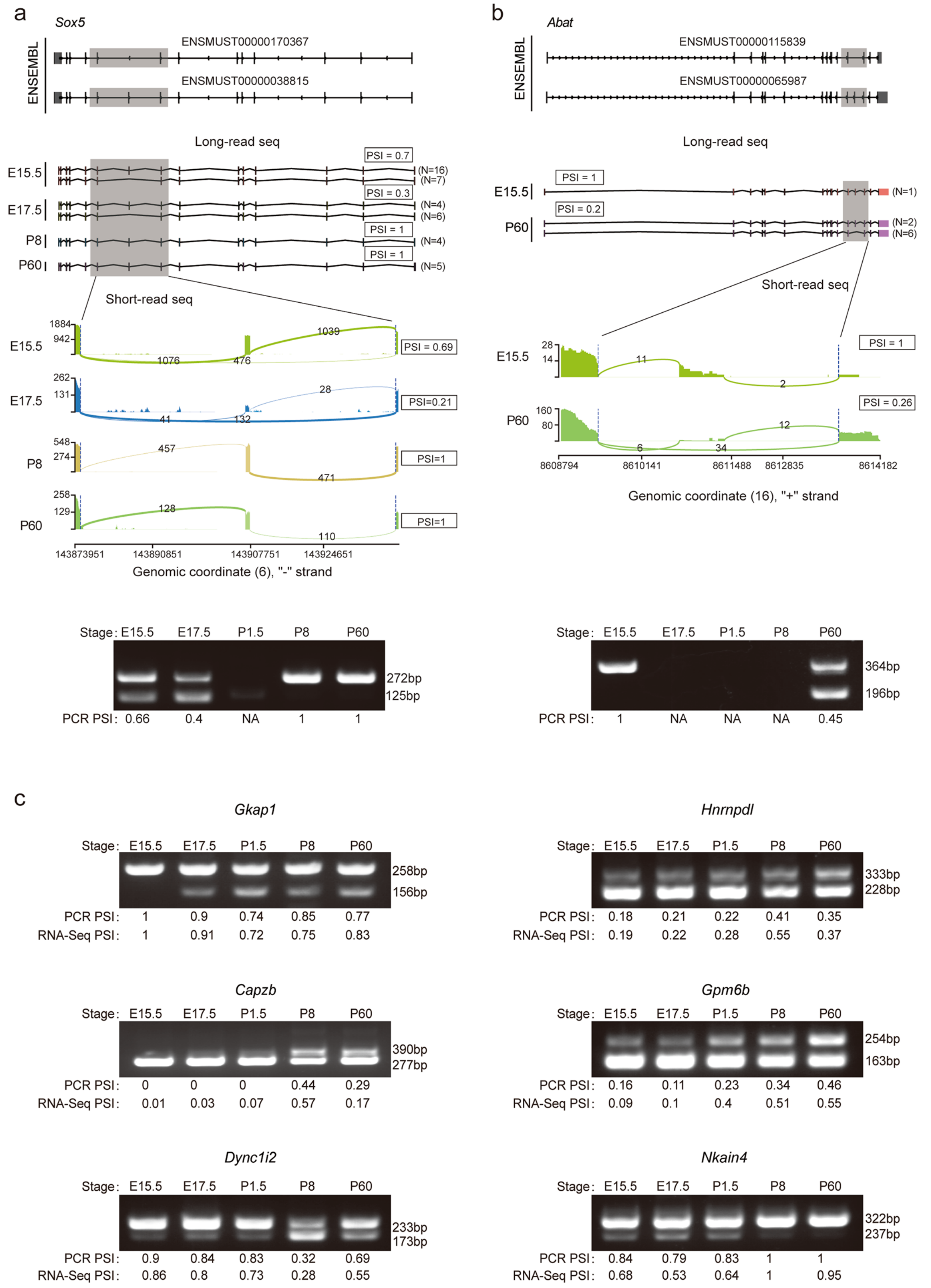

Fig. 5 Transcript visualization and PCR validation of AS. (**a**,**b**) Transcript visualization of gene *Sox5* and *Abat*. Upper panels, transcript annotations in ENSEMBL. Middle panels, reads from long-read and short-read sequencing by stage. PSI represents the percentage of splicing, and N represents the counts of reads. Lower panels, PCR validation of AS events. The PCR PSIs were calculated and shown. (**c**) PCR validation for 6 additional alternative splicing events (*Gkap1, Hnrnpdl, Capzb, Gpm6b, Dync1i2 and Nkain4*), and the PCR PSIs are consistently correlated with RNA-seq PSIs.

SUPPA2 program[31]. In total, 33,230 AS events were identified from all five stages. The P60 stage had the most 13211 AS events, while P8 had the least 6649 AS events (Fig. 4a). Moreover, both E15.5 and E17.5 had nearly 10,000 AS events (Fig. 4a). Stage E15.5 and E17.5 had the most overlapped AS events (Fig. 4a). The P60 stage had the largest percentage of specific AS events (22.1%), while P8 had the smallest percentage of specific AS events (9.84%) (Fig. 4b). The other 3 stages had around 14% specific AS events (Fig. 4b). The percentage of each AS type was also analyzed. SE was the most AS type, which constitutes 30%~37% AS events in each stage, whereas AL and MX were the least with each frequency less than 5% (Fig. 4c).

To further test the consistency between short-read and long-read sequencing data, we checked the SE of two neural development-associated genes, *Sox5* and *Abat* (Fig. 5a). We screened some reads according to their start and end sites within the range of 600 bp upstream and downstream of the annotated transcripts in the mouse reference genome GRCm38. For *Sox5*, Exon 7 skipping was annotated by Ensembl which involves 2 transcripts, ENSMUST00000170367 and ENSMUST00000038815 (Fig. 5a). These 2 transcripts were found out in the long-read sequencing data but were not evenly distributed among the five stages. These transcripts were highest in E15.5 samples and not expressed in P1.5 samples (Fig. 5a). Samples from embryonic stages (E15.5 and E17.5) but not postnatal stages contain the transcript that skips Exon 7 (Fig. 5a). The splicing junction analysis of short-read sequencing data showed that Exon 7 skipping occurred partially at E15.5 (PSI = 0.69) and E17.5 (PSI = 0.21) stages (Fig. 5a). However, no Exon 7 skipping occurred at P8 (PSI = 1) or P60 stages (PSI = 1) (Fig. 5a). For gene *Abat*, Exon 12 skipping was annotated by Ensembl which involves 2 transcripts, ENSMUST00000115839 and ENSMUST00000065987 (Fig. 5b). The former transcript with no Exon 12 was only found in long-read sequencing data of the P60 stage, while the later one harbouring Exon 12 was found in both E15.5 and P60 stages (Fig. 5b). Splicing junction analysis of short-read sequencing also found that Exon 12 skipping occurred at the P60 stage (PSI = 0.26) but not at E15.5 (PSI = 1) The exon retention ratios for *Sox5, Abat* are confirmed by PCR analysis (Fig. 5a,b). We performed PCR validation for 6 additional alternative splicing events (*Gkap1, Hnrnpdl, Capzb, Gpm6b, Dync1i2* and *Nkain4*), and the PCR PSIs are consistently correlated with RNA-seq PSIs (Fig. 5c). Detailed transcript information for these genes are available in Figshare[10].

Usage Note

The present study provides full-length transcriptomic profiles of mouse NSCs across embryonic and postnatal stages. As the dataset contains both short-read and long-read sequencing data, the profiles are reliable for related researches. The profiles are valuable for transcriptional and posttranscriptional mechanisms of neurodevelopment and fate commitment of NSCs, especially the stage-specific gene expression and alternative splicing. Besides, the profiles are also suitable for exploring the molecular mechanisms underlying diseases related to neurodevelopment.

Code availability

The codes used in this article were deposited in https://github.com/LuChenLab/Neuron.

Received: 26 July 2021; Accepted: 8 February 2022;
Published online: 02 March 2022

References

1. Kriegstein, A. & Alvarez-Buylla, A. The glial nature of embryonic and adult neural stem cells. *Annu Rev Neurosci* **32**, 149–184, https://doi.org/10.1146/annurev.neuro.051508.135600 (2009).
2. Bond, A. M., Ming, G. L. & Song, H. Adult Mammalian Neural Stem Cells and Neurogenesis: Five Decades Later. *Cell Stem Cell* **17**, 385–395, https://doi.org/10.1016/j.stem.2015.09.003 (2015).
3. Menn, B. *et al*. Origin of oligodendrocytes in the subventricular zone of the adult brain. *J Neurosci* **26**, 7907–7918, https://doi.org/10.1523/JNEUROSCI.1299-06.2006 (2006).
4. Suh, H. *et al*. *In vivo* fate analysis reveals the multipotent and self-renewal capacities of Sox2+ neural stem cells in the adult hippocampus. *Cell Stem Cell* **1**, 515–528, https://doi.org/10.1016/j.stem.2007.09.002 (2007).
5. Codega, P. *et al*. Prospective identification and purification of quiescent adult neural stem cells from their *in vivo* niche. *Neuron* **82**, 545–559, https://doi.org/10.1016/j.neuron.2014.02.039 (2014).
6. Rosenberg, A. B. *et al*. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182, https://doi.org/10.1126/science.aam8999 (2018).
7. Zhong, S. *et al*. Decoding the development of the human hippocampus. *Nature* **577**, 531–536, https://doi.org/10.1038/s41586-019-1917-5 (2020).
8. Picelli, S. *et al*. Smart-seq 2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* **10**, 1096–1098, https://doi.org/10.1038/nmeth.2639 (2013).
9. Ramskold, D. *et al*. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* **30**, 777–782, https://doi.org/10.1038/nbt.2282 (2012).
10. Chaoqiong, Ding. *et al*. Short-read and long-read full-length transcriptome of neural stem cells across different stages of mouse brain development, *Figshare*, https://doi.org/10.6084/m9.figshare.14658867.v1 (2021).
11. Picelli, S. *et al*. Full-length RNA-seq from single cells using Smart-seq 2. *Nature Protocols* **9**, 171–181, https://doi.org/10.1038/nprot.2014.006 (2014).
12. Wang, X. *et al*. Sequential fate-switches in stem-like cells drive the tumorigenic trajectory from human neural stem cells to malignant glioma. *Cell Res* **31**, 684–702, https://doi.org/10.1038/s41422-020-00451-z (2021).
13. Korsunsky, I. *et al*. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, 1289–1296, https://doi.org/10.1038/s41592-019-0619-0 (2019).
14. Borrett, M. J. *et al*. Single-Cell Profiling Shows Murine Forebrain Neural Stem Cells Reacquire a Developmental State when Activated for Adult Neurogenesis. *Cell reports* **32**, 108022, https://doi.org/10.1016/j.celrep.2020.108022 (2020).
15. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol* **20**, 129, https://doi.org/10.1186/s13059-019-1727-y (2019).
16. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185, https://doi.org/10.1093/bioinformatics/bts356 (2012).
17. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669, https://doi.org/10.1093/bioinformatics/bty149 (2018).
18. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21, https://doi.org/10.1093/bioinformatics/bts635 (2013).

19. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
20. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169, https://doi.org/10.1093/bioinformatics/btu638 (2015).
21. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq 2. *Genome Biol* **15**, 550, https://doi.org/10.1186/s13059-014-0550-8 (2014).
22. Pervouchine, D. D., Knowles, D. G. & Guigo, R. Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics* **29**, 273–274, https://doi.org/10.1093/bioinformatics/bts678 (2013).
23. Kachitvichyanukul, V. & Schmeiser, B. Computer-Generation of Hypergeometric Random Variates. *J Stat Comput Sim* **22**, 127–145, https://doi.org/10.1080/00949658508810839 (1985).
24. Lawrence, M. *et al*. Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9**, e1003118, https://doi.org/10.1371/journal.pcbi.1003118 (2013).
25. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* **20**, 278, https://doi.org/10.1186/s13059-019-1910-1 (2019).
26. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Res* **9**, https://doi.org/10.12688/f1000research.23297.2 (2020).
27. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079, https://doi.org/10.1093/bioinformatics/btp352 (2009).
28. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842, https://doi.org/10.1093/bioinformatics/btq033 (2010).
29. Yin, T., Cook, D. & Lawrence, M. ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol* **13**, R77, https://doi.org/10.1186/gb-2012-13-8-r77 (2012).
30. Chaoqiong, Ding. Short- and long-read RNA-seq of mouse neural stem cells across five developmental stages, *NCBI Sequence Read Archive*, https://identifiers.org/bioproject:PRJNA731598 (2021).
31. Trincado, J. L. *et al*. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol* **19**, 40, https://doi.org/10.1186/s13059-018-1417-1 (2018).

Acknowledgements

We thank Bin Chen for technical support. Y.W. is supported by the National Key Research and Development Program of China, Stem Cell and Translational Research (2017YFA0106500), the National Natural Science Foundation of China (31871376), the Distinguished Young Scientists Program of Sichuan Province (2019JDJQ0029), and the 1·3·5 project for disciplines of excellence, West China Hospital, Sichuan University (ZYYC20019). L.C. is supported by the National Key Research and Development Program of China, Stem Cell and Translational Research (2017YFA0106800 and 2017YFA0106500), and the National Science Fund for Excellent Young Scholars (81722004).

Author contributions

Y.W. and L.C. conceived and supervised the study, and finalized the manuscript. C.D., X.Y. and M.X. drafted the manuscript. C.D., X.Y., P.X., Y.Z. and H.L. performed the tissue preparation, FACS, RNA-seq, PCR experiments, and analyzed the data. M.X., assisted by R.Z., D.Z. and Z.H., performed most of the computational analyses, analyzed the data.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.C. or Y.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.