





OPEN

DATA DESCRIPTOR

Short-read and long-read RNA sequencing of mouse hematopoietic stem cells at bulk and single-cell levels

Xiuran Zheng^{1,3}, Dan Zhang^{1,3}, Mengying Xu^{1,3}, Wanqin Zeng¹, Ran Zhou¹, Yiming Zhang¹, Chao Tang¹, Li Chen¹, Lu Chen¹  & Jing-wen Lin^{1,2} 

Hematopoietic stem cells (HSCs) lie at the top of the differentiation hierarchy. Although HSC and their immediate downstream, multipotent progenitors (MPP) have full multilineage differentiation capacity, only long-term (LT-) HSC has the capacity of long-term self-renewal. The heterogeneity within the HSC population is gradually acknowledged with the development of single-cell RNA sequencing and lineage tracing technologies. Transcriptional and post-transcriptional regulations play important roles in controlling the differentiation and self-renewal capacity within HSC population. Here we report a dataset comprising short- and long-read RNA sequencing for mouse long- and short-term HSC and MPP at bulk and single-cell levels. We demonstrate that integrating short- and long-read sequencing can facilitate the identification and quantification of known and unannotated isoforms. Thus, this dataset provides a groundwork for comprehensive and comparative studies on transcriptional diversity and heterogeneity within different HSC cell types.

Background & Summary

Hematopoiesis starts with a population of self-renewing hematopoietic stem cells (HSCs), which produce a series of increasingly more lineage-committed progenitor cells, ultimately giving rise to all types of mature blood cells. In the conventional model, long-term (LT) HSCs differentiates into short-term (ST) HSCs and subsequently multipotent progenitors (MPPs). Although all three populations have full multilineage differentiation capacity, they gradually lose self-renewal ability. Heterogeneity is presented in both HSC and MPP populations with distinct lineage bias¹.

Both transcriptional and post-transcriptional regulations are pivotal in balancing the constitutive and low-level HSC turnover and downstream differentiation and hematopoietic reconstitution. In multicellular organisms, alternative splicing is a key post-transcriptional regulation mechanism that expands transcript diversity. Accumulating studies have revealed that alternative splicing patterns are essential during hematopoiesis. For instance, alternative splicing events specific in blood progenitors² or megakaryocyte and erythrocyte lineage commitment³ were identified. Alternative splicing patterns in key hematopoietic regulators, such as HMGA2 was found to impact HSC molecular identity⁴. Moreover, aberrant AS is a hallmark of various types of cancer⁵, including leukemia⁶.

RNA sequencing (RNA-seq) that utilises short-read next-generation sequencing (NGS) or long-read sequencing (such as PacBio and Oxford Nanopore Technologies sequencing) is a powerful tool in understanding transcriptional diversity and regulation during various biological processes, including hematopoiesis². While NGS is more reliable in expression quantification, the short reads can only provide limited information in AS events across the splicing joints. In contrast, the long-read sequencing methods offer a unique opportunity to examine alternative splicing isoforms with unprecedented accuracy in providing full-length information.

¹Key Laboratory of Birth Defects and Related Diseases of Women and Children of MOE, Department of Laboratory Medicine, State Key Laboratory of Biotherapy, West China Second University Hospital, Sichuan University, Chengdu, China. ²Biosafety Laboratory of West China Hospital, Sichuan University, Chengdu, Sichuan, 610041, China. ³These authors contributed equally: Xiuran Zheng, Dan Zhang, Mengying Xu. ✉e-mail: luchen@scu.edu.cn; lin.jingwen@scu.edu.cn

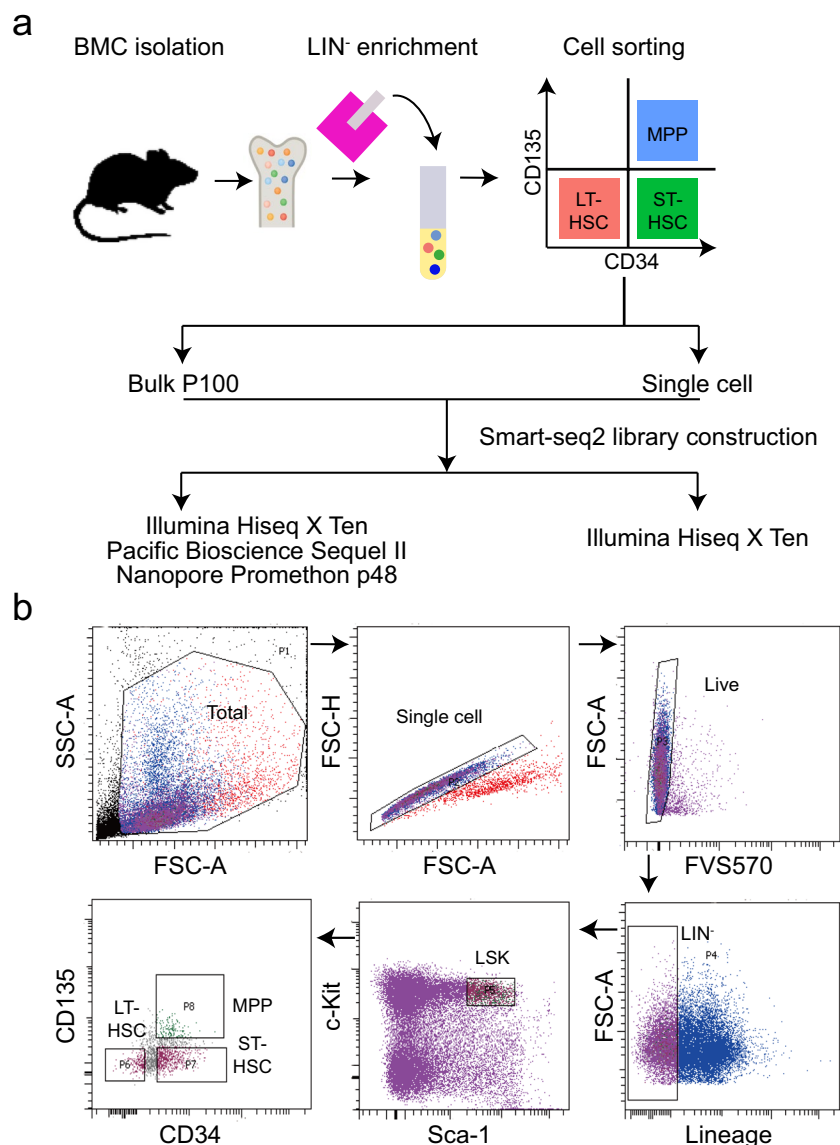


Fig. 1 Sample collection and workflow. **(a)** Sample collection and sequencing methods. Bone marrow cells (BMC) were isolated, and lineage negative (LIN^{-}) cells were enriched using magnetic beads. Long-term (LT) and short-term (ST) hematopoietic stem cells (HSC) and multipotent progenitors (MPP) were sorted according to their surface markers: LT-HSC ($Sca-1^{+}c-Kit^{+}CD34^{-}CD135^{-}$), ST-HSC ($Sca-1^{+}c-Kit^{+}CD34^{+}CD135^{-}$) and MPP ($Sca-1^{+}c-Kit^{+}CD34^{+}CD135^{+}$). Single-cell or 100 cells (bulk P100) were sorted for library construction following the Smart-seq2 protocol. The cDNA libraries were used for short-read (Illumina HiSeq) or long-read (Nanopore or PacBio) sequencing. **(b)** Gating strategy for cell sorting. The main population was gated via FSC-A (forward scatter) and SSC-A (side scatter), and single cells were gated via FSC-A and FSC-H. FSC, forward scatter; SSC, side scatter; A, area; H, height.

Here we present comprehensive transcriptomic profiling of mouse HSC and MPP using both short-read and long-read RNA sequencing at bulk and single-cell levels.

Methods

Mice. C57BL/6J mice were purchased from Beijing HFK Bioscience Co. Ltd (Beijing, China), housed and bred under SPF condition (Specific Pathogen Free) at Laboratory Animal Center of West China Second University Hospital and were allowed access to diet and water *ad libitum*. All animal experiments were carried out following the protocols approved by the Institutional Animal Care and Use Committee of West China Second University Hospital [(2018) Animal Ethics Approval No.004].

Sample collection. The sample preparation and the overall workflow are summarized in Fig. 1. Female adult mice between 8–9 weeks were anesthetized and sacrificed by cervical dislocation. Bone marrow cells were isolated from the femur and tibia through a 70 μ m cell strainer (BD Falcon). Hematopoietic stem and progenitor

cells (HSPCs) were first enriched using EasySep™ Mouse Hematopoietic Cell Isolation Kit (STEMCELL, Cat No. 19856) with a lineage cocktail (biotinylated-CD11b, B220, Gr-1, TER-119 and CD3e) to remove mature cells, according to the manufacturer's instructions.

To purify LT/ST-HSC and MPP, cells were sorted on a BD FACSAria SORP according to the following markers: LT-HSC [lineage (LIN)⁻Sca-1⁺c-Kit⁺CD34⁻CD135⁻], ST-HSC (LIN⁻Sca-1⁺c-Kit⁺CD34⁺CD135⁻) and MPP (LIN⁻Sca-1⁺c-Kit⁺CD34⁺CD135⁺). The following antibodies were used for staining (all purchased from BD): PerCP/Cyanine5.5 anti-CD11b Antibody (Clone M1/70, Cat No. 550993), PerCP/Cyanine5.5 anti-CD3e (Clone 145-2C11, Cat No. 551163), PerCP/Cyanine5.5 anti-TER119 (Clone TER119, Cat No. 560512), PerCP/Cyanine5.5 anti-Gr-1 (Clone RB6-8C5, Cat No. 552093), PerCP/Cyanine5.5 anti-B220 (Clone RA3-6B2, Cat No. 552771), BV421 anti-CD135 (Clone A2F10.1, Cat No. 562898), PE/Cyanine7 anti-Sca-1 (Clone D7, Cat No. 558162), APC/Cyanine7 anti-CD117 (c-Kit) (Clone 2B8, Cat No. 560185), FITC anti-CD34 (Clone RAM34, Cat No. 553733), Fixable Viability Stain 510 (BD Pharmingen, Cat No. 564406). Data were analyzed using FlowJoV10.7.1 software.

For single-cell RNA-sequencing (scRNA-seq), cells were individually sorted into 8-Strip PCR tubes containing the lysis buffer. For bulk RNA-seq, 100 cells (P100) were sorted into one PCR tube as a biological replicates. The batch information for cell sorting was included in Online-only Tables 1, 2.

Library construction and sequencing. The cDNA libraries were generated following the Smart-seq2 protocol⁷ and the Illumina Nextera XT DNA preparation kit was used. The batch information of cDNA generation and sequencing was included in Online-only Tables 1-3.

Short-read sequencing. cDNA was then sheared randomly by Bioruptor Pico sonication device (Diagenode) for Illumina library preparation protocol including DNA fragmentation, end repairing, 3' ends A-tailing, adapter ligation, PCR amplification and library validation. After library preparation, PerkinElmer LabChip® GX Touch and Step OnePlus™ Real-Time PCR System were introduced for library quality inspection. Qualified single-cell or P100 libraries were then loaded on Illumina HiSeq X Ten platform for PE150 sequencing. The average sequencing depth was 23.3 million reads (SD = 3.41) for P100 and 4.2 million reads (SD = 1.1) for single cells. Sequencing library construction and sequencing were done in Annoroad Gene Technology (Beijing, China).

PacBio sequencing. The full-length cDNA libraries were generated as described above, 5 µg of full-length cDNA were used for size selection using the BluePippin™ Size Selection System (Sage Science, Beverly, MA, USA). SMRTbell library was constructed using 1 µg size-selected (above 4 kb) cDNA with the Pacific Biosciences SMRTbell template prep kit. The binding of SMRTbell templates to polymerases was conducted using the Sequel II Binding Kit, and then primer annealing was performed. Sequencing was carried out on the Pacific Bioscience (PacBio) Sequel II platform. Sequencing library construction and sequencing were done in Annoroad Gene Technology (Beijing, China).

Oxford Nanopore Technologies cDNA sequencing. Similarly, the full-length cDNA was generated using the Smart-seq2 protocol as described above. About 250 ng cDNA were subjected to end-repair and dA-tailing ligation using NEBNext FFPE DNA Repair Mix (NEB, Cat No. M6630L) and NEBNext Ultra II End Repair/dA-Tailing Module (NEB, Cat No. E7645). PromethION library preparation was performed according to the manufacturer's instructions (Oxford Nanopore Technologies, SQK-LSK109 Kit, EXP-NBD104 and EXP-NBD114). Sequencing was done on the Promethion p48 platform. Sequencing library construction and sequencing were done in Biomarker Technologies Co., Ltd (Beijing, China).

Sequencing data analysis. All programs used default parameters unless stated otherwise in the corresponding section.

Single-cell short-read sequencing. Single-cell analysis was carried out using the Seurat R package (v3.2)⁸. We filtered cells using the following criteria: (a) feature counts below 500 or above 8000, (b) UMI counts below 70,000 or over 5,000,000, (c) identified as doublets using DoubletFinder (v2.0.3)⁹ with the parameter of "PCs = 1:6, pN = 0.25, pK = 0.09, nExp = 24". A total of 109 LT-HSCs, 98 ST-HSCs and 107 MPPs were kept for further analyses. The quality statistics were provided in Online-only Table 1. We used principal component analysis (PCA) with variable genes as input and identified the top 6 significant PCs that were used as input for tSNE (t-distributed stochastic neighbor embedding). Cell markers from previously published studies^{10,11} were used to verify the cell types. Differentially expressed genes in the cell types were determined using the FindAllMarkers function in Seurat R package (v3.2)⁸.

Bulk short-read sequencing. The quality of the raw sequence data was analyzed using FastQC software (v0.11.8) (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and RSeQC software¹² (v3.0.1) (<http://rseqc.sourceforge.net/>). The paired-end short-reads were aligned to the mouse reference genome GRCm38 with annotation from ENSEMBL release 93 using STAR (v2.7.1a)¹³ with the parameters of "—outSJfilterOverhangMin 12 12 12 12 —alignSJoverhangMin 3 —alignSJBoverhangMin 3 —chimSegmentMin 12 —chimScoreMin 2 —chimScoreSeparation 10 —chimJunctionOverhangMin 12 —outFilterMultimapNmax 1 —chimOutType Junctions SeparateSAMold".

Gene body coverage, distribution of aligned reads over genome feature and RNA integrity at transcript-level were calculated using RSeQC¹². The RNA integrity at the transcript-level was evaluated using the Transcript Integrity Number (TIN) algorithm¹⁴. These quality statistics were provided in Online-only Table 2. Gene expression was quantified using function summarizeOverlaps from GenomicAlignments (v1.24.0)¹⁵ with parameters of

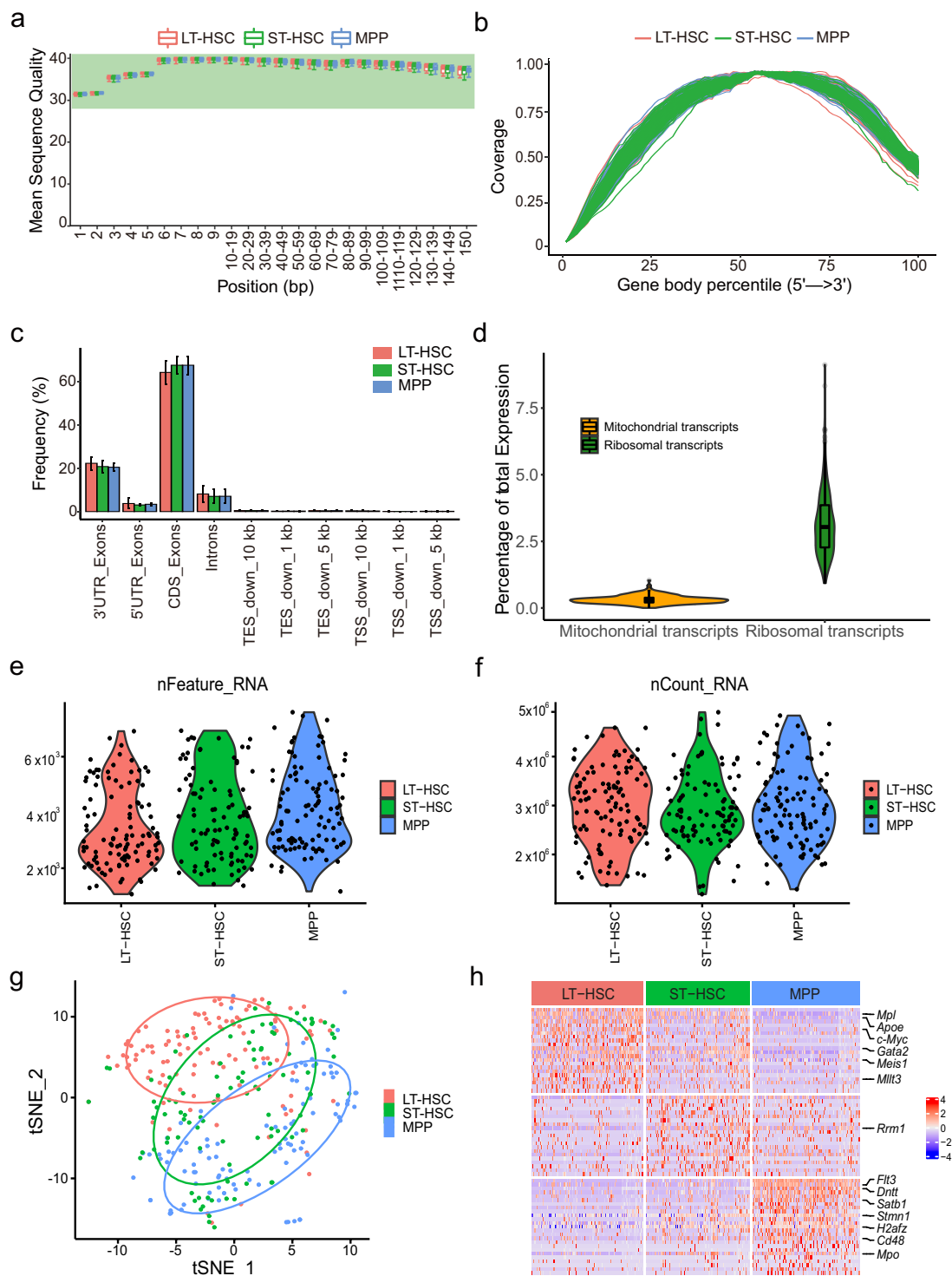


Fig. 2 Quality check of single-cell Illumina short-read sequencing. **(a)** Boxplot showing average sequence quality per base for LT-HSC (red), ST-HSC (green) and MPP (blue) single-cell samples. **(b)** Gene body coverage of all detected genes. **(c)** Frequency of counts mapped in various gene regions. Bar chart showing mean and standard deviation. CDS, coding sequence; TES, transcription end site; TSS, transcription start site. **(d)** Violin plots showing percentage of reads that mapped to mitochondrial (orange) or ribosomal (green) transcripts. Violin plots showing the number of detected genes per cell **(e)** or the number of counts per barcode **(f)**. **(g)** tSNE of three different cell types. **(h)** Heatmap showing the top 30 differentially expressed genes for each cell type (FDR < 0.05, log₂(fold change) > = 1.5). The color scale indicates the expressing level.

“mode = “Union”, singleEnd = FALSE, ignore.strand = TRUE, fragments = TRUE”. The read counts were then normalized by function rlog with the parameter of “blind = FALSE” in DESeq2 (v1.28.1)¹⁶. Principal component analysis (PCA) and sample distances were calculated using FactoMineR (v2.4)¹⁷ and DESeq2 (v1.28.1)¹⁶,

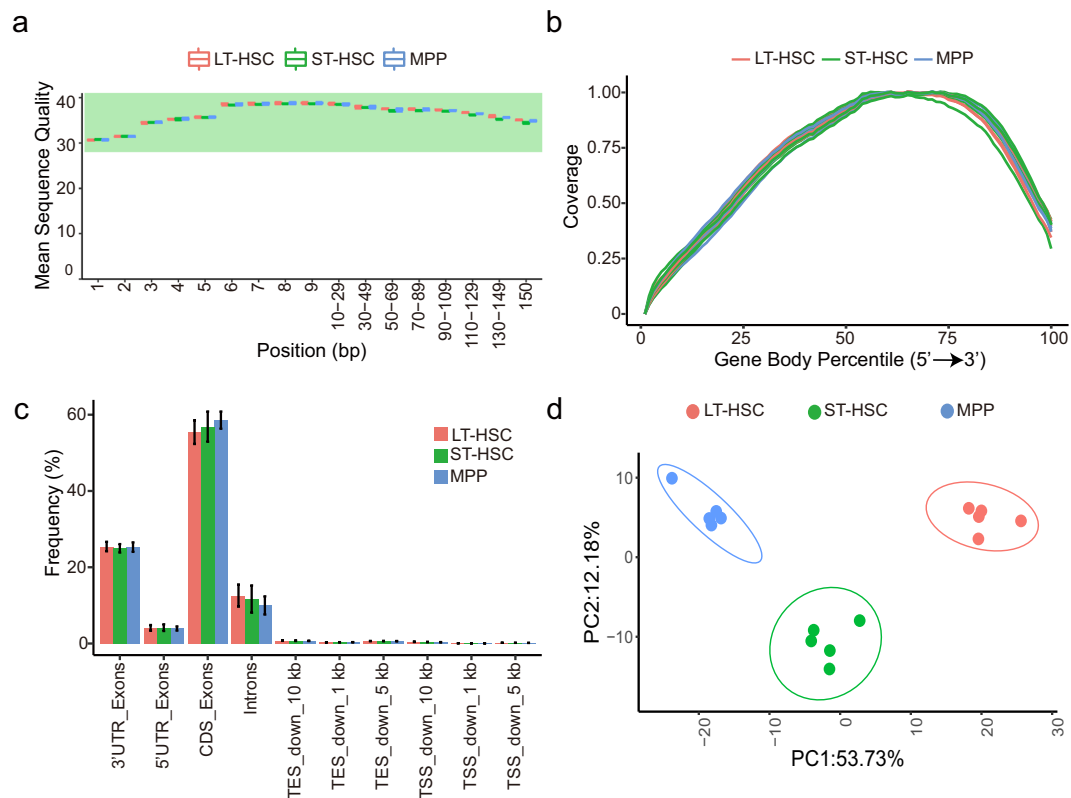


Fig. 3 Quality check of bulk Illumina short-read sequencing. **(a)** Boxplot showing average sequence quality per base for LT-HSC (red), ST-HSC (green) and MPP (blue), 5 replicates per cell type. **(b)** Gene body coverage of all detected genes. **(c)** Frequency of counts mapped in various gene regions. Bar chart showing mean and standard deviation. CDS, coding sequence; TES, transcription end site; TSS, transcription start site. **(d)** Principal Component Analysis (PCA) of all samples after batch correction.

respectively. The batch effect (provided in Online-only Table 2) was corrected using ComBat from the sva package (v3.36.0)¹⁸.

Bulk long-read sequencing. The alignment was performed using Minimap2 (v2.17-r974-dirty)¹⁹ with the parameter of “-ax splice”. The number of reads, read length and quality were analyzed using NanoComp software (v1.33.1)²⁰ with the parameter of “-raw-store-tsv_stats” and visualized using R package ggplot2 (v3.3.2)²¹ after filtering the reads with length below 200 or over 150,000 bp. These quality statistics were provided in Online-only Table 3.

Stringtie2 (v2.1.7) was used to assemble and quantify both short and long reads with or without a reference guiding. The parameter of the assembly process was “-p 20 -e -G -A”, and “-L” for long reads. The read counts of genes were extracted with a python script (prepDE.py3) with “-l 600” and normalized by function rlog with the parameter of “blind = FALSE” in DESeq2 (v1.28.1)¹⁶. The batch effects of gene expression from bulk Illumina, PacBio and Oxford Nanopore Technologies sequencing were corrected using ComBat from the sva package (v3.36.0)¹⁸. Principal component analysis (PCA) was calculated using plotPCA function in DESeq2 (v1.28.1)¹⁶. Correlation was calculated by function cor from R package stats (v4.0.2) with the parameter of “method = pearson”. GffCompare (v0.12.1)²² was used to compare and evaluate the accuracy of Stingtie2²³ transcript assembly.

SUPPA2 (v2.3)²⁴ was employed to classify the AS events with the parameter of “-f ioe -e SE SS MX RI FL”. Seven AS types were identified, including skipping exon (SE), retained intron (RI), alternative 5' splice site (A5), alternative 3' splice site (A3), mutually exclusive exons (MX), alternative first exon (AF), in which alternative first-exon use results in mRNA isoforms with distinct 5' UTRs, and alternative last exon (AL), in which alternative use of multiple polyadenylation sites results in distinct terminal exons.

To visualize long-reads for transcripts, the gene regions were extracted using Samtools (v1.10.2)²⁵. Bedtools (v2.30.0)²⁶ was used to convert all bam files to GTF format. Visualization was performed using R package ggbio (v1.36.0)²⁷. Sashimi plots of short-read sequencing data were plotted using pysashimi (<https://github.com/ygildtu/pysashimi>).

Data Records

The raw data were deposited at NCBI BioProject under accession number PRJNA706066²⁸. The accession number for individual samples were summarized in Online-only Tables 1-3. The processed files, including the quantifications of gene expression, isoform, splicing junction from both protocols were uploaded in figshare²⁹.

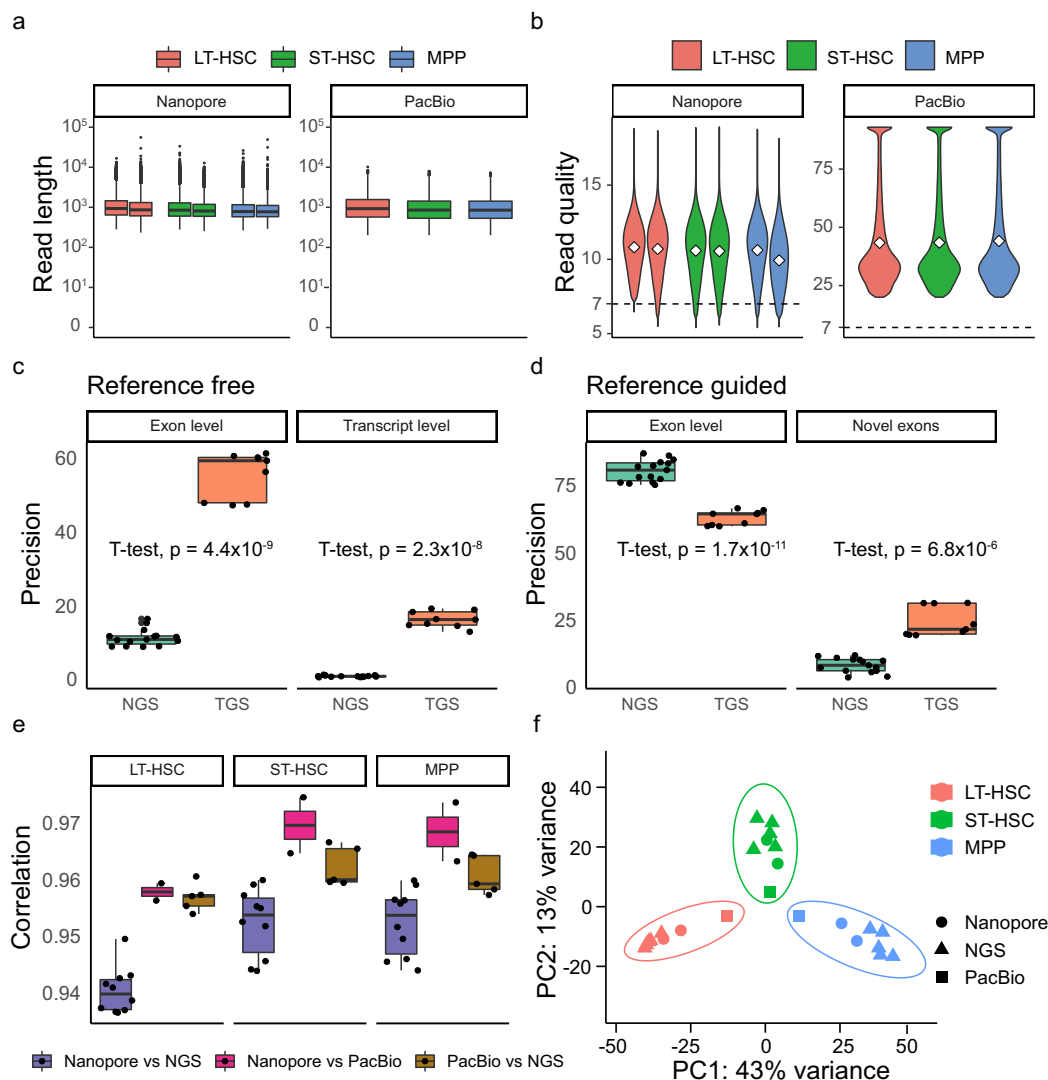


Fig. 4 Quality check of bulk long-read sequencing. (a) Boxplots showing the length of reads per sample for nanopore (left) or PacBio (right) sequencing. LT-HSC (red), ST-HSC (green) and MPP (blue). (b) Violin plots showing read quality per sample for nanopore (left) or PacBio (right) sequencing. The white dots represent the means and the dashed line indicates the quality value of 7. (c) Precision of assembly accuracy at exon and transcript levels of Stingtie2 without a reference. (d) Precision of assembly accuracy at exon and novel exons levels of Stingtie2 with a reference. (e) Correlation coefficients of gene expression among the three sequencing methods. (f) Principal Component Analysis (PCA) of all samples from bulk Illumina, PacBio and Nanopore sequencing methods after batch correction.

Technical Validation

Quality control for short-read Illumina sequencing data. Read quality was assessed using FastQC, the number of total and clean reads, percentage of unique mapped reads, the mapped read length, Q30 and TIN (Transcript Integrity Number) statistics were provided in Online-only Tables 1-2. TIN indicates the RNA integrity at transcript-level¹⁴. The bulk RNA-seq samples generally have higher TIN scores than the single-cell samples. All single-cell samples had median TIN scores ranging from 24.8 to 51.8, with an average value of 40.7. Similar ranges of TIN were observed for all cell types (Online-only Table 1). Whereas the bulk samples had median TIN scores ranging from 49.4 to 66.1, with an average value of 56.9 (Online-only Table 2).

There was no significant difference in the distribution of average quality score per base among the samples of different cell types for both single-cell (Fig. 2a) and bulk (Fig. 3a) samples, and the reads were distributed approximately uniformly across the gene body for both datasets (Figs. 2b, 3b), indicating the high integrity of the RNA. We further checked the gene regions where the reads were mapped to and found that all samples have significantly more reads mapped to the exonic regions, while less were mapped to the intronic regions (Fig. 2c and Fig. 3c), in line with the observation of the previous reports³⁰.

For single-cell sequencing data, we also examined the proportion of reads mapped to mitochondrial and ribosomal genes (Fig. 2d). The median percentage of mitochondrial genes per cell was 0.29 and the median

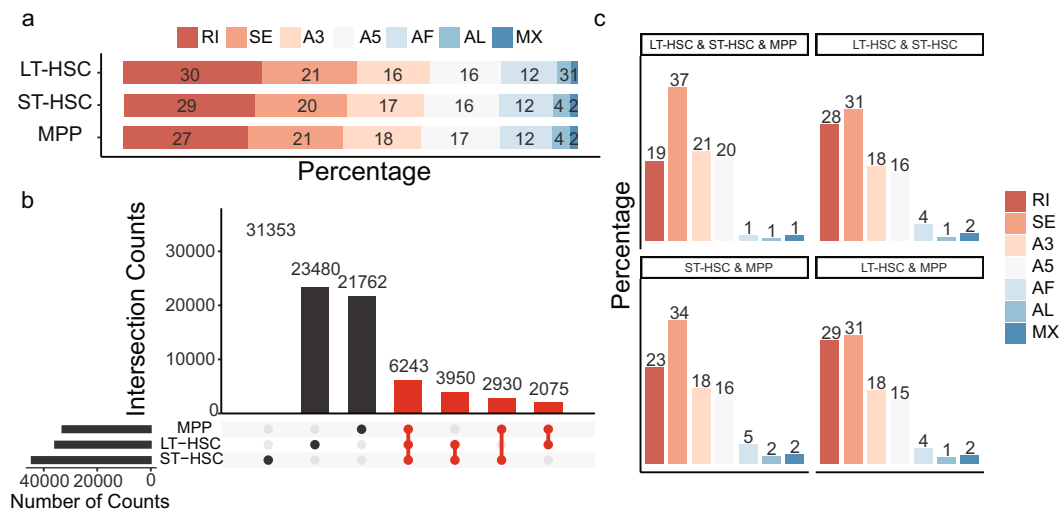


Fig. 5 Overall patterns of alternative splicing and cell type-specific/shared events of bulk long-read sequencing. **(a)** Bar plot of the percentages of alternative splicing types in three cell types based on SUPPA2. **(b)** Bar plot showing the numbers of specific and shared alternative splicing events in three cell types. **(c)** Percentage of alternative splicing types for the shared alternative splicing events among cell types. SE, skipping exon; RI, retained intron; A5, alternative 5' splice site; A3, alternative 3' splice site; MX, mutually exclusive exons; AF, alternative first exon; AL alternative last exon.

percentage of ribosomal genes per cell was 3.04. MPP had the highest average number of detected genes (Fig. 2e), significantly higher than LT-HSC ($p < 0.001$, T-test), while the number of UMI counts per cell were comparable amongst three cell types (Fig. 2f). UMAP graph also showed that ST-HSC lied in between LT-HSC and MPP (Fig. 2g). The differentially expressed genes were analyzed between the cell types. There were 62, 63 and 266 differentially expressed genes ($FDR < 0.05$, $\log_2(\text{fold change}) \geq 1.5$) in LT-HSC, ST-HSC and MPP, respectively. Furthermore, several known HSC signatures, including *Mpl*, *c-Myc*, *Mllt3*, *Gata2*, showed significantly higher expression in LT-HSC (Fig. 2h).

For bulk RNA sequencing, principal component analysis showed the three cell types separate on PC1 that accounts for 53.7% variation (Fig. 3d) after batch correction. Similar to the single-cell data, ST-HSC lied in between LT-HSC and MPP, corresponding to their biological status. Taken together, these results showed that the cells were correctly sorted and both the bulk and single-cell short-read RNA-seq data were of high quality.

Quality control and concordance for long-read sequencing data. The number of reads, read length, percentage of mapped reads and quality statistics were provided in Online-only Table 3. Nanopore sequencing data had a mean length of 1024 bp (Fig. 4a). Whereas PacBio sequencing data had the mean length of 946 bp (Fig. 4a). The quality score was higher for PacBio sequencing than nanopore sequencing (Fig. 4b), with an average of 47.57 and 10.53, respectively. The quality scores were comparable amongst all three cell types for both sequencing methods. Next, we compared the short- and long-read sequencing in terms of the precision of identifying exon and transcripts with or without reference. We found that long-read sequencing can provide relatively complete chains of exon including novel exons at transcript-level with or without reference (Fig. 4c,d), whereas short-read sequencing provides higher accuracy in defining exon junctions when a reference is available (Fig. 4d).

To assess the concordance between the replicates, we calculated the correlations of gene quantification between the short-read and long-read sequencing. The correlation coefficients were all over 0.93 (Fig. 4e), indicating high concordances among replicates. Moreover, the PCA illustrated that the samples were clustered by cell type between short-read and long-read sequencing replicates after batch correction (Fig. 4f). These results suggested that long-read sequencing data were of high quality, and the biological replicates are of high concordance. Moreover, the long-read sequencing enables the *de novo* identification and quantification of novel exon and transcripts.

Overall alternative splicing patterns. To investigate the overall alternative splicing patterns using this long-read dataset, we first identified the alternative splicing events and types using SUPPA2²⁴. Interestingly, the most prevalent type of alternative splicing was retained intron (RI) in all the cell types, followed by skipped exon (SE) and alternative 3' or 5' splice sites (Fig. 5a). Next, we revealed that more than 21,762 alternative splicing events were cell-type specific (Fig. 5b). The SE was the most-shared alternative splicing type among cell types (Fig. 5c), followed by RI and alternative 3' or 5' splice sites. These results suggested that the long-read sequencing facilitated the identification of large numbers of cell-specific or shared alternative splicing events that may be of potential function in hematopoiesis.

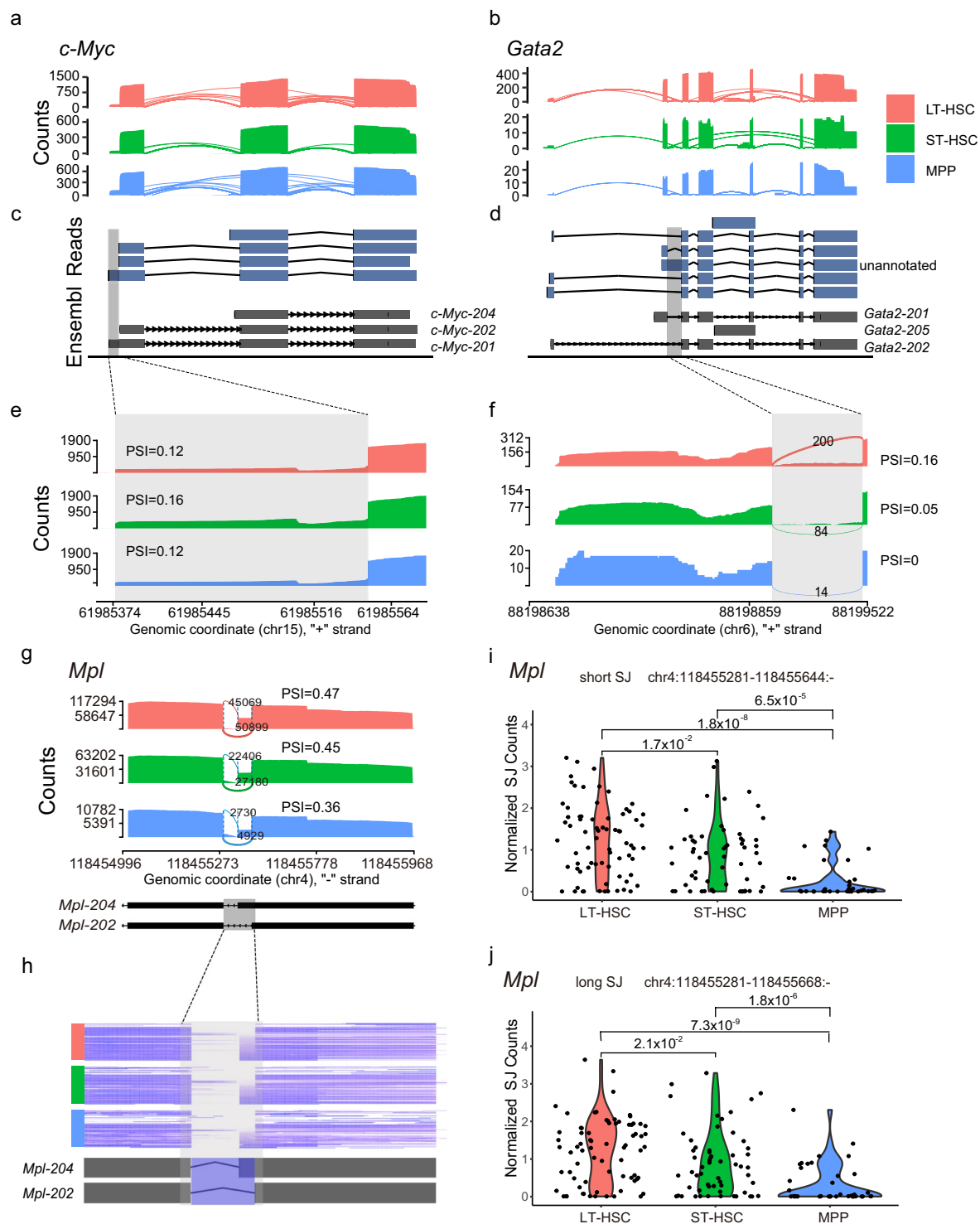


Fig. 6 Alternative splicing isoforms identified using the long-read sequencing and quantified using the short-read sequencing. Sashimi plots showing long-reads mapped to *c-Myc* (a) or *Gata2* (b) in LT-HSC (red), ST-HSC (green) and MPP (blue) by Nanopore or PacBio sequencing. The identified nearly full-length transcripts (upper panel) and annotated transcripts in Ensembl (lower panel) for *c-Myc* (c) and *Gata2* (d) were shown. Sashimi plots showing Illumina generated short reads that are mapped to the indicated (shaded by grey) region for *c-Myc* (e) or *Gata2* (f). (g) Sashimi plots of *Mpl* gene showing the decreased PSI of an alternative splicing event in LT-HSC (red), ST-HSC (green) and MPP (blue) by summing up the reads from all the single cells of each type. The percentage of spliced-in (PSI) for the indicated regions was calculated for each cell type. (h) Heatmap illustrates that the heterogeneity of alternative splicing of *Mpl* among single cells within each cell type. (i,j) Violin plots showing the normalized splicing junction counts of the short junction at chr4:118455281–118455644:- (i) and the long junction at chr4:118455281–118455668:- (j) in LT-HSC (red), ST-HSC (green) and MPP (blue).

Alternative splicing isoforms identification and quantification. To further confirm the quality of long-reads in identifying alternative splicing isoforms, we visualized the transcripts of known LT-HSC maker, *c-Myc* and *Gata2* (Fig. 2h) in the three cell types using nanopore and PacBio sequencing data. We screened all the reads mapped to the genes *c-Myc* and *Gata2* and their annotated transcripts. For *c-Myc*, we found 2915 reads, and LT-HSC had the highest number of reads (Fig. 6a). We visualized reads of full-length transcripts both in nanopore and PacBio sequencing and found that all annotated isoforms were identified (Fig. 6c). A 5' alternative start site was found in the first exon of *c-Myc*. We used the short-read sequencing data to quantify the percentage of spliced in (PSI) for this locus and found that the longer isoform has a similar PSI (percent of splice-in) in all three cell types with ST-HSC had a slightly higher proportion of reads containing the longer isoform of the first exon (Fig. 6e). For *Gata2*, we found 595 reads, and similarly, LT-HSC had nearly 20 times more reads than the other two cell types (Fig. 6b). By comparing the full-length reads and annotated transcripts, we found an unannotated in Ensembl transcript with intron retention in LT-HSC (Fig. 6d). We further verified this intron retention using short-read sequencing data and quantified PSI for this intron with the highest PSI in LT-HSC (Fig. 6f). Next, we showed an example of integrating long-read sequencing with the single-cell RNA-seq based on Smart-seq2. We identified an alternative 5' splicing site (A5) with a 24-bp alternative-spliced region in *Mpl* from our long-read sequencing data. Using our Smart-seq2 data, we identified that this event is differentially spliced with a decreasing PSI from LT-HSC (0.47), ST-HSC (0.45) to MPP (0.36) (Fig. 6g–j). The heterogeneity among single cells can be visualized using the Smart-seq2 data (Fig. 6h), and the involved long- and short-splicing junction (SJ) were significantly down-regulated during the hematopoiesis (Fig. 6i,j). These results indicated that integrating short- and long-read sequencing facilitates the identification of differentially regulated isoforms.

Usage Notes

The bulk short-read RNA-seq can be used to quantify gene expression and alternative exon usage with high accuracy in various tissues or cell types. Single-cell RNA sequencing is powerful in revealing heterogeneity in gene expression within cell types. Full-length RNA-seq protocols such as Smart-seq2 also enable measurements of heterogeneity in alternative splicing. However, the assembly of transcripts using short-read sequencing remains difficult.

Our dataset provides a unique opportunity in revealing transcript diversities in the HSC population. By integrating the short- and long-read bulk sequencing datasets, one can better identify and quantify (novel) alternative splicing isoforms as shown above. The scRNA-seq data can further provide information on how these transcripts varied within different HSC cell types. Furthermore, this dataset is useful in developing statistical models to reconstruct the isoforms and enables further investigation of the largely unexplored post-transcriptional regulations such as alternative splicing and RNA editing at single-cell levels.

Gene expression and alternative splicing are regulated in gender-, age- and strain-specific manners, we generated data from female adult C57BL/6J mice between 8–9 weeks in this study. Thus, users should take into consideration the gender, development stage and mouse strain when concluding. We expect that the accumulation of data will provide a more complete transcriptional and post-transcriptional landscape of hematopoiesis.

Code availability

The codes used in this article were deposited in <https://github.com/LuChenLab/hemato>.

Received: 1 June 2021; Accepted: 27 September 2021;

Published online: 29 November 2021

References

- Cheng, H., Zheng, Z. & Cheng, T. New paradigms on hematopoietic stem cell differentiation. *Protein and Cell* **11**, 34–44 (2020).
- Chen, L. *et al.* Transcriptional diversity during lineage commitment of human blood progenitors. *Science* **345**, 1251033 (2014).
- Edwards, C. R. *et al.* A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages. *Blood* **127**, e24–e34 (2016).
- Cesana, M. *et al.* A CLK3-HMGA2 Alternative Splicing Axis Impacts Human Hematopoietic Stem Cell Molecular Identity throughout Development. *Cell Stem Cell* **22**, 575–588.e7 (2018).
- Venables, J. P. Aberrant and alternative splicing in cancer. *Cancer Res.* **64**, 7647–7654 (2004).
- Crews, L. A. *et al.* RNA Splicing Modulation Selectively Impairs Leukemia Stem Cell Maintenance in Secondary Human AML. *Cell Stem Cell* **19**, 599–612 (2016).
- Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
- Stuart, T. *et al.* Comprehensive Integration of Single-. *Cell Data.* *Cell* **177**, 1888–1902 (2019).
- McGinnis, C. S., Murrow, L. M. & Gartner, Z. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst.* **8**, 329–337.e4 (2019).
- Chambers, S. M. *et al.* Hematopoietic Fingerprints: An Expression Database of Stem Cells and Their Progeny. *Cell Stem Cell* **1**, 578–591 (2007).
- Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
- Wang, L., Wang, S. & Li, W. RSeQC: Quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
- Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- Wang, L. *et al.* Measure transcript integrity using RNA-seq data. *BMC Bioinformatics* **17**, 1–16 (2016).
- Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.* **9**, 1–10 (2013).
- Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).
- Lé, S., Josse, J. & Husson, F. FactoMineR: An R Package for Multivariate Analysis. *J. Stat. Softw.* **25**, 1–18 (2008).
- Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
- Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

20. 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 (2018).
21. H, W. ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York, 2016).
22. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Research* **9** (2020).
23. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* **20**, 278 (2019).
24. Trincado, J. L. *et al.* SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.* **19**, 1–11 (2018).
25. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
26. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
27. Yin, T., Cook, D. & Lawrence, M. ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol.* **13**, R77 (2012).
28. Zheng, X. Short-read and long-read RNA sequencing of mouse haematopoietic stem cells at bulk and single-cell levels. *BioProject* <https://identifiers.org/bioproject:PRJNA706066> (2021).
29. Zheng, X. Short-read and long-read RNA sequencing of mouse haematopoietic stem cells at bulk and single-cell levels. *figshare* <https://doi.org/10.6084/m9.figshare.14569548.v7> (2021).
30. Bush, S. J., McCulloch, M. E. B., Summers, K. M., Hume, D. A. & Clark, E. L. Integration of quantitated expression estimates from polyA-selected and rRNA-depleted RNA-seq libraries. *BMC Bioinformatics* **18**, 1–12 (2017).

Acknowledgements

This work was supported by the National Key Research and Development Program of China, Stem Cell and Translational Research (2017YFA0106800) to Lu Chen and J.-w.L., and the National Science Fund for Excellent Young Scholars to Lu Chen (81722004).

Author contributions

J-w.L. and Lu Chen. designed and directed the study; X.Z. performed the experiments; D.Z., M.X., W.Z., R.Z., Y.Z., C.T., L.C. analyzed the data; D.Z., M.X., X.Z., W.Z., Lu Chen and J.-w.L. generated the figures, wrote and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Lu Chen or J.-w.L.

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.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2021