BMC Genomic Data

RESEARCH

Open Access

A dataset of single-cell transcriptomic atlas of Bama pig and potential marker genes across seven tissues

Long Chen[1,2], Xingyan Tong[1], Yujie Wu[1], Can Liu[1], Chuang Tang[1], Xu Qi[1], Fanli Kong[3], Mingzhou Li[1], Long Jin[1*] and Bo Zeng[1*]

Abstract

The use of single-cell sequencing technology for single-cell transcriptomics studies in pigs is expanding progressively. However, the comprehensive classification of cell types across different anatomical tissues and organs of pig in multiple datasets remains relatively limited. This study employs single-cell and single-nucleus sequencing technologies in Bama pig to identify unique marker genes and their corresponding transcriptomic profiles across diverse cell types in various anatomical tissues and organs, including subcutaneous fat, visceral fat, psoas major muscle, liver, spleen, lung, and kidney. Through detailed data analyses, we observed widespread cellular diversity across various anatomical tissues and organs of Bama pig. This work contributes a comprehensive dataset that supports physiological studies and aids in the identification and prediction of potential marker genes through single-cell transcriptomics of these tissues. The methodologies and data employed in this study are designed to improve the accuracy of cell type identification and ensure consistent cell type allocation.

Keywords Bama pig, Single-cell RNA sequencing (scRNA-seq), Single-nucleus RNA sequencing (snRNA-seq), Marker genes

Background

The continual advancements and breakthroughs in transcriptomic technology [1–3] have allowed researchers to conduct in-depth studies at the transcriptomic level in *Sus scrofa* (pig or swine). The main aim is to discover intrinsic information, such as gene expression regulation mechanisms [4], immune responses [5], functional genes [6], and metabolic pathways [7]. For example, one study completed transcriptome gene annotation across multiple species, including pigs [8]. A highly integrated resource of transcriptomic features was provided by detailed analysis of the swine transcriptomic landscape, laying a solid foundation for research at the transcriptomic level in pigs [9]. The rapid development of swine transcriptomic technology has significantly aided

*Correspondence:
Long Jin
longjin@sicau.edu.cn
Bo Zeng
apollobovey@163.com
[1]State Key Laboratory of Swine and Poultry Breeding Industry, College of Animal Science and Technology, Sichuan Agricultural University, Chengdu 611130, China
[2]Key Laboratory of Agricultural Bioinformatics, Ministry of Education, Sichuan Agricultural University, Chengdu 611130, China
[3]College of Life Science, Sichuan Agricultural University, Ya'an 625099, China

research on disease and improvements in breeding [10, 11].

Although traditional transcriptomic methods, such as bulk RNA sequencing, have significantly advanced the understanding of gene expression, they fail to capture cellular heterogeneity at the single-cell level [12]. Single-cell RNA sequencing technologies [13] have recently emerged as vital tools for investigating cellular heterogeneity and for single-cell research [14]. Recently, a comprehensive account of the immense potential and challenges of single-cell technologies was provided [15]. The feasibility of conducting systematic analyses and constructing comprehensive atlases at the single-cell, single-nucleus level even in tissues where endothelial cells constitute a minority has been demonstrated [16]. Notably, during the global outbreak of African swine fever in 2018, virus-regulated signaling pathways were discovered via single-cell technologies, providing crucial scientific evidence for disease control strategies [17]. Research on porcine single cells will profoundly impact the understanding of biological complexity and contribute to solving pressing issues such as disease [18, 19].

The Bama pig is a small pig breed that bears physiological, anatomical, nutritional, metabolic, and disease-related similarities to humans, making it extensively used in human disease research [20, 21]. Concurrently, extensive research has been conducted on Yucatan miniature pigs, particularly in cardiac [22], peripheral blood [23], colonic [24], and kidney [25] pig tissues, via single-cell sequencing. However, the application of single-cell sequencing technology in transcriptomic studies of Bama pig remains relatively limited. Specifically, research into systematically defining unique marker genes in the visceral tissues of Bama pig is limited.

Methods

Animals and sample collection

The research utilized an adult female purebred Chinese Bama pig (2 years old, 45 kg) provided by Hengshu Bio-Technology Co., Ltd., Yibin, Sichuan, China. The pig was maintained in a stably controlled laboratory environment, with the room temperature set at 25–28 °C and the humidity at 70%. The feed energy level required for animal maintenance was determined on the basis of the NRC (2012) and the Chinese Fatty Growing and Finishing Pig Feeding Standards (2004). After 12 h of fasting, the pig was placed in a restraint bag, and Zoletil®50 anesthetic was injected into the well-developed neck muscles using a 14-gauge needle at a dose of 5 mg/kg. Subsequently, the pig was slaughtered, and samples from seven organs or tissues, namely, the liver, spleen, lung, kidney, psoas major muscle (PMM), subcutaneous adipose tissue of the back (SAT), and greater omentum (GOM), were collected.

Tissue preparation for single-cell (scRNA-seq) libraries (liver, spleen, lung, kidney): Tissues intended for single-cell RNA sequencing library generation were obtained from a local slaughterhouse. Freshly collected tissues were immediately placed on ice and processed within 30 min. Each type of tissue was dissociated and digested separately. To ensure effective digestion and maintain cell viability, after tissue dissociation, the cell suspensions were passed through a cell strainer to remove debris and lyse red blood cells. Cell viability for each tissue type was assessed at the core facility of Novogene Co., Ltd. (Beijing, China) via a flow cytometer (Acea Bioscience, Inc., US), which employs Hoechst 33,342 (Invitrogen, Cat# H3570) and propidium iodide (Thermo Fisher Scientific, Cat# P3566), with a viability threshold of over 80% required for subsequent sequencing analyses.

Liver

Fresh livers were sampled from five distinct anatomical regions: the left lateral lobe (LLL), left medial lobe (LML), right medial lobe (RML), right lateral lobe (RLL), and quadrate lobe (QL). Each region provided 1 g of tissue, which was washed twice with cold PBS. The tissues were mixed, chopped into small pieces, and then transferred to a 50 mL tube, to which 20 mL of digestion solution was added. This mixture contained 0.5 mg/mL collagenase type II (Gibco, Cat#17101015), 1.25 mg/mL protease (Sigma, Cat#P5147–100MG), and 7.5 μg/mL DNase I (Sigma–Aldrich, Cat#D4527–10KU) in cold HBSS. The digestion was carried out at 37 °C for 15 min, with gentle shaking every 5 min. The reaction was terminated in cold MACS buffer containing 0.25% BSA (Sigma–Aldrich, Cat# 10735096001) and 2 mM EDTA, and the sample was filtered through a 100 μm cell strainer (Sigma–Aldrich, Cat# CLS431752-50EA), followed by flow cytometric cell staining.

Spleen

Fresh spleens were sampled from two anatomical sides, the visceral and parietal sides, with 1 g of tissue taken from each, and washed twice with cold PBS. The tissues were mixed, chopped into small pieces, and transferred to a 50 mL tube, where 10 mL of digestion solution was added. This mixture contained 20 mg/mL collagenase type IV (Gibco, Cat# 17104019), 1 U/mL dispase II (Gibco, Cat# 17105041), and 7.5 μg/mL DNase I in 10 mL of HBSS. The digestion was carried out at 37 °C for 15 min, with gentle shaking every 5 min. The reaction was terminated in cold MACS buffer containing 0.25% BSA and 2 mM EDTA, and the sample was sequentially filtered through 100 μm and 40 μm cell strainers, followed by flow cytometric cell staining.

Lung

Fresh lungs were sampled from seven different areas: the left apex, left medial, left main, right apex, right medial, accessory, and right main lobes. Each area provided 0.5 g of tissue, which was subsequently washed twice with cold HBSS. The tissues were chopped into small pieces and transferred to a 50 mL Falcon tube, to which 20 mL of digestion solution was added, containing 1 mg/mL collagenase type II, 2.5 mg/mL collagenase type IV, and 7.5 μg/mL DNase I. The digestion was conducted at 37 °C for 30 min, with gentle shaking every 5 min, and terminated in MACS buffer. The sample was diluted in cold HBSS and filtered through 100 μm and 40 μm cell strainers, followed by debris removal and flow cytometric cell staining.

Kidney

Fresh kidneys were sampled from four areas: the upper pole, lower pole, cortex, and medulla. Each area provided 1 g of tissue, which was subsequently washed twice with cold PBS. The tissues were chopped into small pieces and transferred to a 50 mL Falcon tube, to which 20 mL of digestion solution containing 1 mg/mL collagenase type II (Gibco, Cat# 17101015), 2 mg/mL collagenase type IV (Gibco, Cat# 17104019), 1 U/mL dispase II (Gibco, Cat# 17105041), and 7.5 μg/mL DNase I (Sigma–Aldrich, Cat# D4527–10KU) was added. The digestion was conducted at 37 °C for 20 min, with gentle shaking every 5 min. The reaction was terminated in 20 mL of cold MACS buffer containing 0.25% BSA (Sigma–Aldrich, Cat# 10735096001) and 2 mM EDTA, and the sample was filtered through 100 μm and 40 μm cell strainers, followed by flow cytometric cell staining.

Sample collection for single-nucleus sequencing (snRNA-seq)

The tissues used for single-nucleus sampling included visceral adipose (greater omentum fat, GOM), subcutaneous fat (SAT), and psoas major muscle (PMM) from a Bama pig, which were meticulously dissected while adhering to ethical guidelines. The collected tissues were washed with cold PBS, immediately frozen in liquid nitrogen, and stored at −80 °C until use. For nuclear extraction, the tissues were thawed, cut into small pieces, and transferred to homogenization buffer containing 20 mM Tris pH 8.0, 500 mM sucrose, 0.1% NP-40, 0.2 U/mL RNase inhibitor, 1% BSA, and 0.1 mM DTT. The tissues were homogenized via a pestle 15 times and filtered through a 40 μm strainer. The samples were then centrifuged at 4 °C for 10 min at 500×g, and the supernatant was carefully discarded. The pellet (nuclei) was resuspended in PBS containing 1% BSA and 20 U/μL RNase inhibitor and prepared for subsequent snRNA-seq library construction.

10× genomics library preparation and sequencing

The samples were subsequently transported on dry ice to Novogene Co., Ltd. (Beijing, China). where the cDNA libraries were constructed and sequenced. The 10× Genomics Chromium single-cell 3′ gene expression solution was used. Among the seven data samples collected under the designated conditions, four datasets were from single-cell sequencing, and the other three datasets were from single-nucleus sequencing. All the experimental procedures were conducted in accordance with the manufacturer's protocol (www.10xgenomics.com/support/single-cell-gene-expression). The quality control criteria for the preparation of single-cell suspension samples were a cell viability > 80%, a cell concentration of 700–1200 cells/μL, and a cell diameter of 5–30 μm. The cell nuclei were extracted from the PMM, SAT, and GOM samples at the time of sample preparation (10× Chromium Nuclei Isolation Kit) because of their excessively large cell diameters. Qualified cell and cell nuclei suspensions were loaded onto 10× Genomics Single-Cell 3.0 Chips. During this step, the cells were partitioned into gel beads-in-emulsion (GEMs) along with gel beads coated with 10× barcode oligonucleotides (including the 14-bp index and 10-bp UMIs (unique molecular identifiers)). After generating the GEMs, the samples were transferred into PCR tubes, and reverse transcription was performed via a T100 Thermal Cycler (Bio-Rad). cDNAs with both barcodes were amplified, and libraries were constructed via a single-cell 3′ Reagent Kit (v3) for each sample. The resulting libraries were sequenced on an Illumina NovaSeq 6000 System in PE150 mode.

Cell demultiplexing and gene counting

The raw sequencing data were used directly for sequence quality control and gene quantification via CellRanger (v7.1.0, 10× Genomics, using default parameters). The reference genome assembly was downloaded from Ensembl in FASTA format (Sscrofa11.1, GCA_000003025.6) together with the gene annotation GTF file (release 109). The cell metadata, which include barcodes.tsv, features.tsv, and gene expression matrix (*.mtx) files, were automatically generated via CellRanger. The initial data are available in additional files (Supplementary Table 1).

Quality control of cells and genes

We used R software (version 4.2.3, https://www.r-project.org/) and the Seurat R package [26] (version 4.4.0, https://satijalab.org/seurat/) for the downstream analyses. Initial mitochondrial quantification was conducted, and data quality control was performed according to the following criteria: genes expressed in < 10 cells were excluded; genes with exceedingly low or high overall expression (nFeature_RNA < 200, nFeature_RNA > 5000)

were filtered out; and in single-nucleus sequencing tissues (SAT, GOM, and PMM), cells with mitochondrial gene percentages greater than 10% were eliminated. In single-cell tissues (lung, kidney, and spleen), cells with mitochondrial gene percentages greater than 30% were eliminated. In single-cell tissues (liver), cells with mitochondrial gene percentages greater than 50% were eliminated. The R package DoubletFinder [27] (github.com/chris-mcginnis-ucsf/DoubletFinder) was subsequently used to remove doublet cells (DoubletRate = 0.075).

Data normalization and cell clustering

For the refined dataset, gene expression count information for each sample was normalized via the `NormalizeData` function, followed by feature selection of 3,000 variable genes via the `FindVariableFeatures` function with the "vst" method. The data were then scaled via the `ScaleData` function [28]. Principal component analysis (PCA) and UMAP projection were executed through the `RunPCA` and `RunUMAP` functions, respectively. For different tissues, we used either 30 (single-nucleus datasets) or 20 (single-cell datasets) principal components as determined by the inflection point of the elbow plot for each tissue. Cell clusters were identified via the `FindClusters` function and visualized via UMAP. For the accurate determination of tissue-specific resolution sizes for each dataset, clustree software was employed. Specifically, the resolutions were set as follows: 0.8 for the liver, 0.2 for the spleen, 0.2 for the lung, 0.1 for the kidney, 0.5 for the PMM, 0.5 for the SAT, and 0.2 for the GOM.

Differentially expressed genes and annotation

To identify differentially expressed genes (DEGs) across different cell clusters, we employed the FindAllMarkers function from the Seurat package, utilizing the Wilcoxon rank sum test. This function was configured to detect only genes positively marking clusters, with a minimum percentage of expressing cells (min.pct) set at 25% and a log fold change (logfc.threshold) threshold of 0.25. We subsequently filtered these DEGs to include only those with an adjusted p value less than 0.05. After all the DEGs were identified, the cell types were manually annotated by referencing tissue-specific marker genes that have been documented in previously published studies [10, 16, 29–43]. These known marker genes are listed in additional files (Supplementary Table 2), while the most typical marker genes used in our annotations are displayed in additional files (Supplementary Figures).

Predicting marker genes

On the basis of the differentially expressed genes (DEGs) identified, we annotated cell types using known marker genes. To uncover more potential marker genes for each cell type across the seven tissues, we further refined the criteria for gene selection: the gene's pct1 value must exceed 0.7; avg_log2FC must be greater than 0.5; and the gene must exhibit pronounced expression specificity, using the percentage of the gene's relative expression as one of the standards, such as a gene accounting for more than 70% of its total expression in a specific cell type. These criteria increase their potential as cell type-specific markers.

Data integration

We utilized the 'merge' function to integrate the postquality control single-cell datasets (liver, spleen, lung, and kidney) and single-nucleus datasets (PMM, SAT, and GOM) independently. We used Harmony [44], a method that relies on multidimensional scaling techniques, for the separate integration of single-cell and single-nucleus datasets. Harmony eliminates batch effects caused by technical and biological variations by harmonizing high-dimensional similarities between cells. Thus, Harmony addresses not only technical factors such as library preparation but also biological variations between cell subpopulations [45].

Results and discussion

Experimental workflow and initial data quality

We collected samples from seven different tissues and organs of Bama pig, including the liver, spleen, lung, kidney, psoas major muscle, subcutaneous back fat, and omental fat. These tissue samples were immediately refrigerated upon collection and underwent standardized cellular dissociation and digestion processes to ensure cell viability. The samples were subsequently sent to the laboratory for single-cell library construction and high-throughput sequencing. Using the 10x Genomics platform and CellRanger software for data processing, we achieved detailed cell typing and annotation for both whole and individual tissues. Batch effects were corrected via the Harmony algorithm during integrated analysis, ensuring data consistency and reliability (Fig. 1a). The proportion of reads uniquely mapped to the genome was above 58%, with the average number of reads per cell ranging from 56,339 to 184,315. Analysis via CellRanger revealed that the median gene expression per cell ranged from 1,000 to 2,696. The average sequencing saturation for the Bama pig samples was 77.10%, demonstrating high data utility. Furthermore, CellRanger analysis indicated that the number of viable cells typically met or exceeded the anticipated capture of 3,000 cells (Fig. 1b). The detailed raw data, such as the Q30 values for UMI sequences across all seven tissues exceeding 90%, are reported in additional files (Supplementary Table 1). Overall, these data highlight the high-quality sequencing depth and excellent coverage of the Bama pig reference genome, providing a solid basis for further analyses.

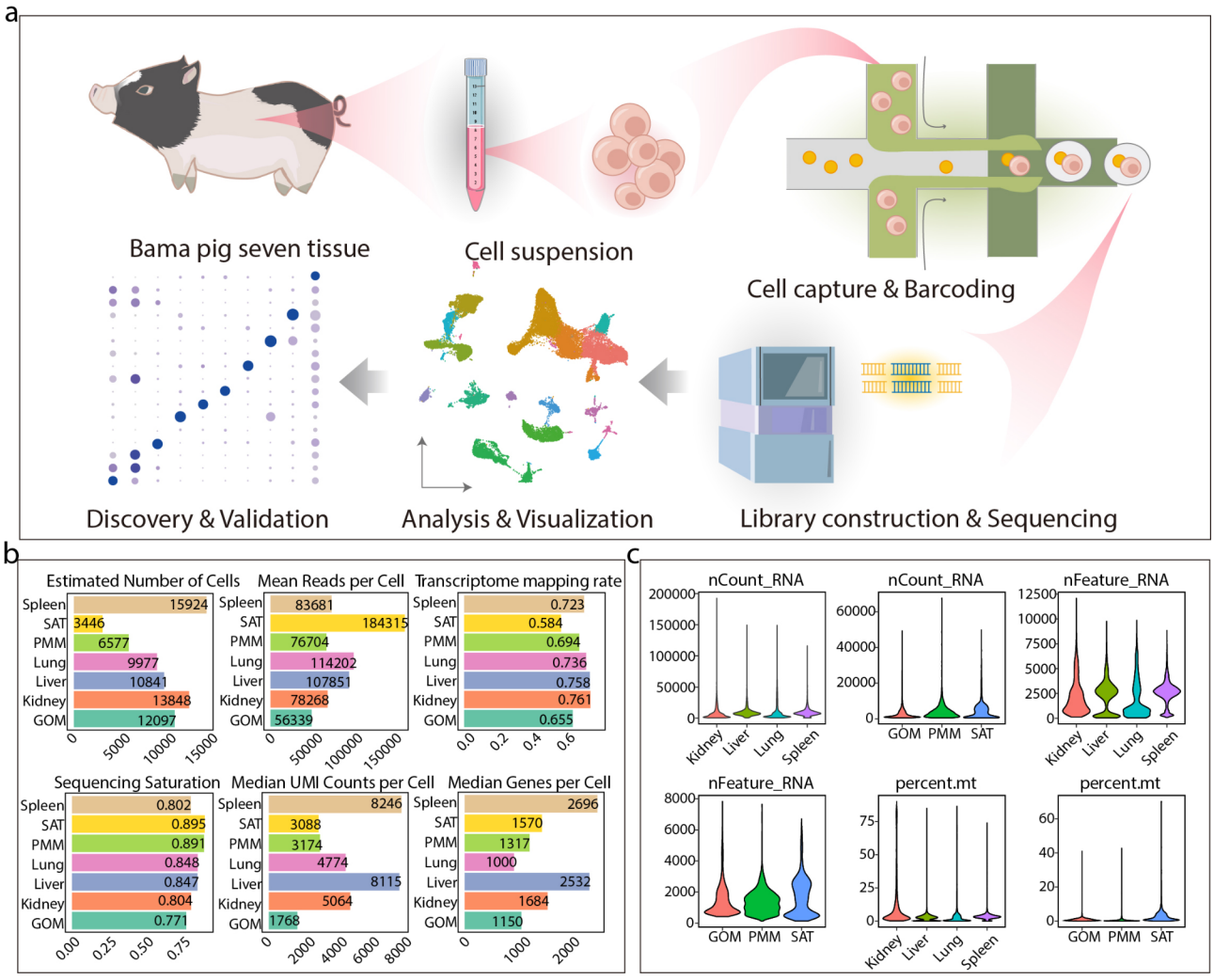

Fig. 1 Single-cell RNA sequencing workflow and data overview for seven bama pig tissues. **a** Schematic representation of the experimental workflow. Tissue samples from seven different parts of the Bama pig were processed for scRNA-seq and snRNA-seq, followed by high-throughput sequencing and downstream bioinformatics analysis. **b** Statistics of single-cell sequencing data for different tissues. Estimated Number of Cells: indicates the number of viable cells; Mean Reads per Cell: represents the average read coverage per cell; Transcriptome Mapping Rate: denotes the proportion of reads mapped exclusively to the genome; Sequencing Saturation: indicates the complexity index of the sequencing library; Median UMI Counts per Cell: median number of unique molecular identifiers per cell; Median Genes per Cell: median number of genes expressed per cell. **c** Variations in mitochondrial and ribosomal gene expression across tissues. nCount_RNA, nFeature_RNA, and percent.mt correspond to each of the seven samples. In the context of the snRNA-seq libraries, mitochondrial expression ratios in PMM, GOM, and SAT are lower than in the single-cell tissues (kidney, liver, lung, spleen)

The raw single-nucleus sequencing data (SAT, GOM, and PMM) represent RNA found only within cell nuclei. RNA from mitochondrial (percent.mt) genes is generally absent in nucleus-only sequencing data; hence, the percentage of mt is typically lower in snRNA-seq data. Moreover, since single-nucleus sequencing captures only nascent RNA in nuclei, the quantity of RNA (nCount_RNA) is also expected to be lower than that of scRNA-seq methods, with correspondingly lower numbers of genes and reads (nFeature_RNA) per cell sequenced (Fig. 1c).

Tissue-specific cell type classification

To independently verify the reliability of each dataset and identify tissue-specific cell types or subtypes, we conducted separate analyses through cell annotation and classification for each dataset. After CellRanger processed the raw data output, we used Seurat to construct single-cell matrices and then performed quality control on each dataset. As previously mentioned, the percentage of mitochondrial genes is used as a proxy for cell death, but mitochondrial gene RNA is generally absent in nuclear data; therefore, the percent.mt is lower in single-nucleus sequencing data. Notably, Hepatocytes may have very high mitochondrial content [46]; hence, a threshold

of 50% was set (liver, percent.mt < 50) to optimally preserve hepatocytes and remove dead or dying cells. In single-cell microfluidics processes, droplets containing two or more cells can occur, which are then mistakenly identified as single cells. To further remove cells with abnormal gene expression, we employed DoubletFinder [27], eliminating doublets. Finally, each cluster's identity was assigned on the basis of established cell-specific marker gene expression. As expected, each tissue exhibited specific cell types; subcutaneous back fat (SAT) and omental fat (GOM) uniquely featured "fibro-/adipogenic progenitors" and "adipocytes". The psoas major muscle (PMM) showed skeletal muscle fiber cells, "I myofiber", "II myofiber 2B", and "II myofiber 2X". The spleen contains "dendritic cells", which are crucial for initiating immune responses, particularly in antigen presentation [47]. The lung features "alveolar type 1 cells", which cover the surfaces of the alveoli for gas exchange, and "alveolar type 2 cells", which are responsible for secreting surfactants to prevent alveolar collapse [48]. The kidney displays "proximal tubule cells", which are crucial for filtering waste from the blood and absorbing water and nutrients [49]. Liver tissue includes "Hepatocytes", which are associated with the metabolism of proteins, carbohydrates, and fats [50], and specialized immune cells in the liver, "Kupffer cells" (Fig. 2).

Integrative analysis results of the ScRNA-seq data

After individually assessing and confirming the high quality of each tissue dataset, which featured a diverse array of cell types and tissue-specific cells, we found that integrative analysis was feasible. Batch effects between different datasets can compromise the authenticity of comparisons, even when the same sequencing technologies are used. To standardize comparisons, we merged the postquality control scRNA-seq samples from the liver, spleen, lung, and kidney and conducted dimensionality reduction clustering (resolution = 0.8). This approach effectively mitigated batch effects, yielding a consolidated dataset of 39,406 cells and 24,686 genes for comprehensive annotation and comparison. Typically, cells with more total RNA molecules also exhibit more diverse gene detection. The correlation coefficient between nCount_RNA (total RNA molecules in cells) and nFeature_RNA (number of distinct genes detected per cell) increased from 0.83 to 0.9 after quality control (Fig. 3a), indicating a significant increase in the quality of the subsequent datasets. After correction for batch effects, the integration of similar cell types across different tissues was significantly enhanced (Fig. 3b). In this integrated analysis, using a curated set of cell-specific marker genes (Fig. 3e), we annotated 12 cell types from 33 cell clusters (Fig. 3c). In addition to identifying most cell types previously detected in individual analyses (Fig. 3d), we also discovered four established marker genes common across tissues, namely, *CRTAM*, *GNLY*,

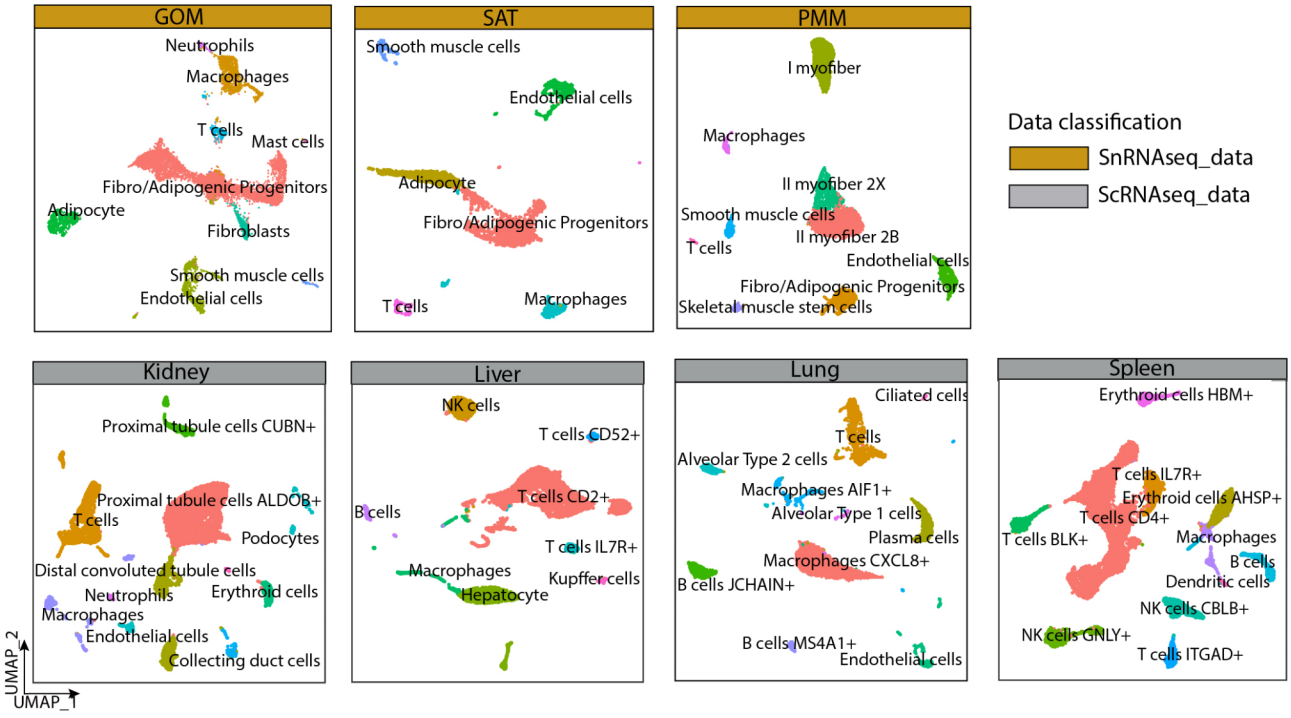

Fig. 2 UMAP plots of the seven datasets after individual clustering analysis

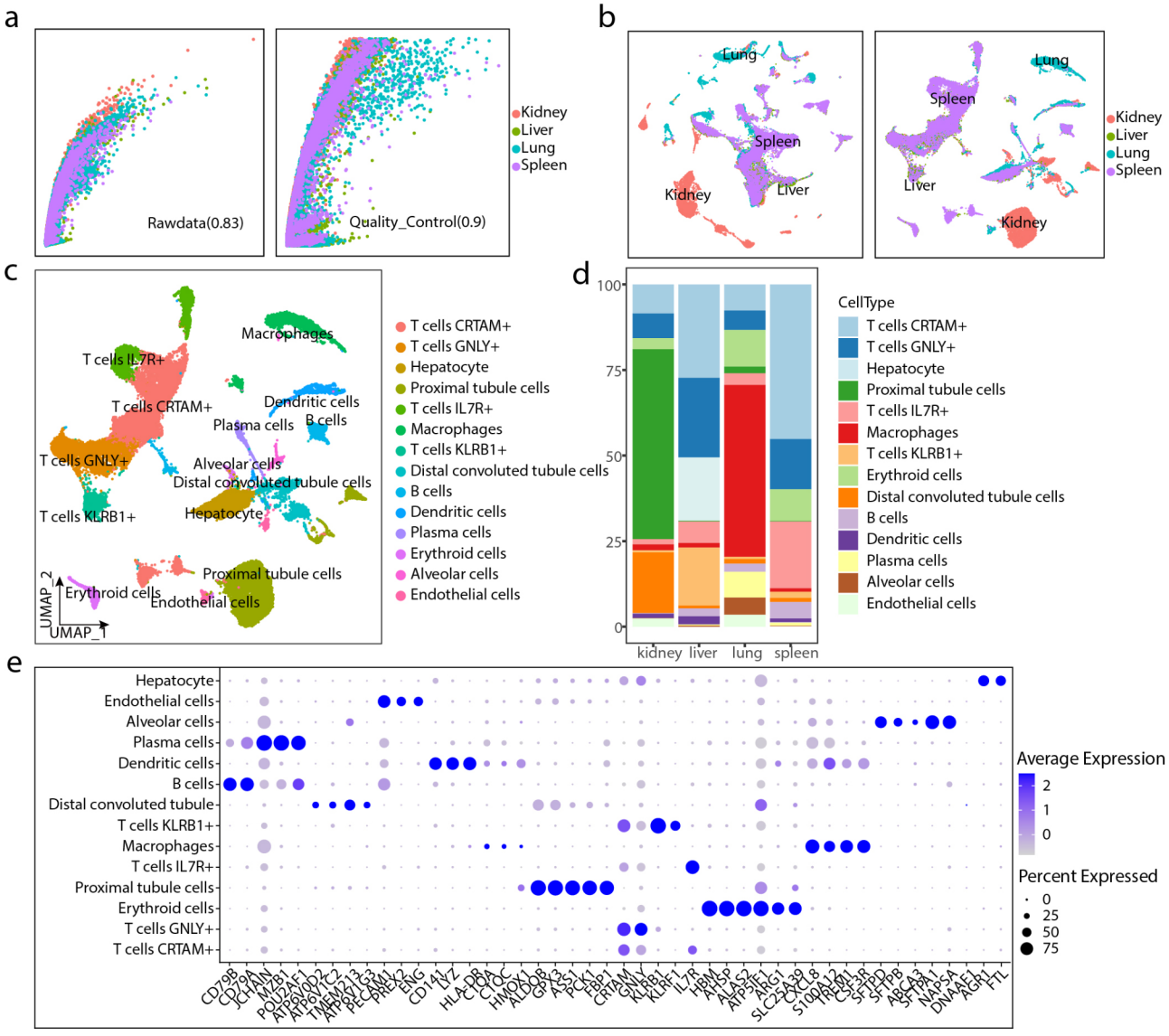

Fig. 3 Technical validation of integrated single-cell tissue datasets (kidney, liver, lung, and spleen). **a** Scatter plots showing the correlation between nCount_RNA (x-axis) and nFeature_RNA (y-axis). The correlation coefficient before integration (Rawdata) is 0.83, which increases to 0.9 after low-quality cells and doublets are removed. **b** Comparison of UMAP plots illustrating batch effect correction before and after integration across the four single-cell datasets under the same standard. **c** Annotation UMAP plot using established marker genes for the datasets. **d** Bar chart comparing the abundance of cell types across the four integrated single-cell tissue datasets. **e** Bubble chart of established marker genes for 12 cell types in the integrated single-cell data. Larger and darker bubbles indicate stronger specificity

IL7R, and KLRB, for identifying T cells in all four tissues. In terms of cell abundance, the spleen, which is predominantly composed of immune cells, represented 96.74%, whereas the kidney was characterized mainly by tubule cells. The liver and lung presented a more complex assortment of cell types.

Integrative analysis results of the SnRNA-seq data

Similarly, to compare snRNA-seq data under the same standard, snRNA-seq samples (SAT, GOM, and PMM) were merged and subjected to dimensionality reduction clustering (resolution = 0.4), resulting in a total of 20,090 nuclei and 22,991 genes after batch effect removal. The correlation coefficient between nCount_RNA (total RNA molecules in nuclei) and nFeature_RNA (number of distinct genes detected per nucleus) increased from 0.9 to 0.94 after quality control (Fig. 4a). Compared with the scRNA-seq dataset, the snRNA-seq dataset presented more pronounced batch effects (Fig. 4b), and the cells became more distinctly separated, with similar cell types clustering more closely after batch effect correction. In the snRNA-seq dataset, 12 cell types were annotated from 21 cell clusters via a curated set of cell-specific

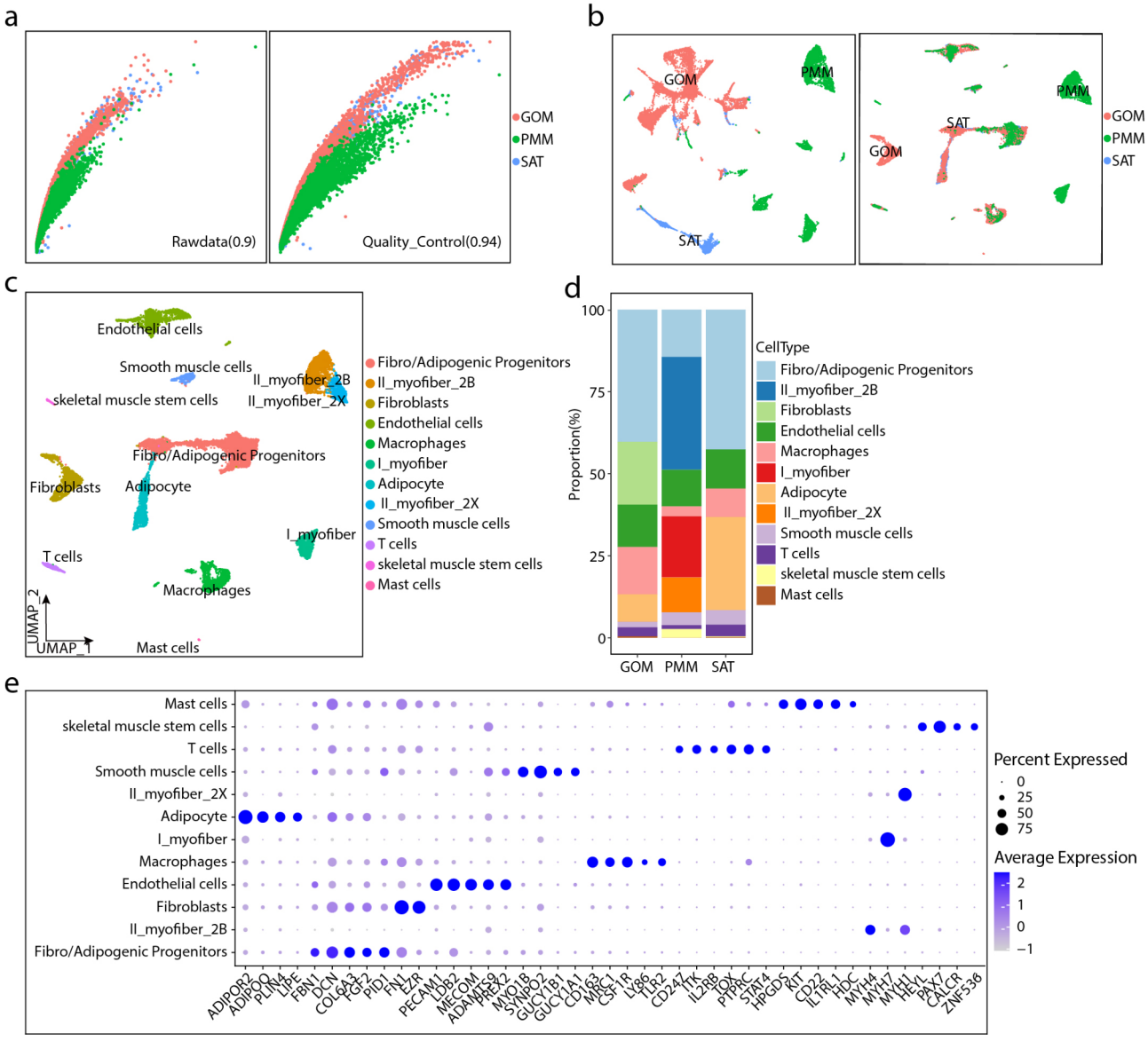

Fig. 4 Technical validation of integrated single-nuclei datasets (SAT, GOM, and PMM) **a** Correlations between nCount_RNA and nFeature_RNA before (0.9) and after (0.94) quality control. **b** UMAP plots comparing batch effects pre- and postcorrection across datasets under the same standard. **c** UMAP annotations using established marker genes. **d** Bar chart of cell type proportions in integrated snRNA-seq datasets. **e** Bubble chart showing marker gene specificity by cell type, with size and color intensity indicating specificity

marker genes (Fig. 4e, c). In terms of cell abundance, each tissue presented unique cell types with high cell type abundance (Fig. 4d). Overall, the data integration highlights the reliability of the data, not only providing a comparative abundance of cell types under a unified standard but also serving as a reference for future data integration and other research endeavors.

Potential marker genes

In single-cell transcriptomics, high-quality data are often closely associated with gene quality. To confirm the high quality of our data, we used a verification procedure that involved the careful selection of 0–5 potential marker genes for each single-cell type to assess their predictive power. Using the filtering criteria in our methodology, we found that the genes summarized in our I additional files.

(Supplementary Table 3) clearly exhibited strong marker gene representation. Remarkably, our results show robust concordance between the functions of the predicted annotated genes and the original annotations of the cell types, confirming a high level of correspondence. Together, the results highlight the effectiveness of these genes as marker genes, which aligns perfectly with our expectations regardless of their specificity or

expression levels. In summary, within a rigorous quality control framework, including meticulous batch effect removal, the final set of high-quality genes and cells provides additional potential marker genes for a thorough exploration of distinct cell types in miniature pig (Bama pig), offering resources for studying the heterogeneity between different tissues or organs at the transcriptomic level.

One limitation of this study is the lack of replicates. Although we have collected rich single-cell transcriptomic data from seven different tissues and organs, and ensured high data quality through strict quality control steps, the absence of replicates means that our results may be influenced by individual variations or sample processing inconsistencies. Additionally, the study is based solely on single-cell transcriptomic data from Bama pigs, so it would be beneficial to integrate and compare data from other similar or different pig breeds to further validate and expand the identified marker genes and cell types. Furthermore, while we have identified several potential marker genes, these still require experimental validation to confirm their biological significance across different tissues and cell types. Therefore, future studies should consider incorporating replicates, integrating data from different pig breeds or experimental conditions to enhance the generalizability and robustness of the findings, and further validating the potential marker genes.

Conclusions

In the present study, we gathered data from seven distinct tissues and organs of a Bama pig. This high-quality dataset comprises 72,710 cells and serves as a valuable resource that offers a comprehensive single-cell reference for miniature pigs. This resource will not only aid in exploring the broad heterogeneity across different pig tissues and organs but also become a valuable resource and reference for identifying potential biomarkers unique to Bama pig.

Abbreviations

scRNA-seq	Single-cell RNA sequencing
snRNA-seq	Single-nucleus RNA sequencing
PMM	Psoas major muscle
SAT	Subcutaneous adipose tissue
GOM	Omental fat tissue
UMAP	Uniform manifold approximation and projection
nCount_RNA	Total RNA molecules per cell or nucleus
nFeature_RNA	Distinct RNA features (genes) detected per cell or nucleus
GEO	Gene expression omnibus
UMI	Unique molecular identifier
Q30	Quality control metric for sequencing accuracy
Batch effects	Technical variations in data analysis
Cell-specific marker genes	Genes specific to certain cell types
Clustree	Tool for visualizing clustering hierarchy in single-cell data
Resolution	Granularity of clustering in single-cell analysis
Percent.mt	Mitochondrial gene percentage as a proxy for cell death

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12863-025-01308-3.

Supplementary Material 1: Supplementary Figures. Comparative Bubble Plots of Top Marker Genes Across Seven Tissues (Kidney, Liver, Lung, Spleen, PMM, SAT, and GOM) Highlighting the Most Signi cant Markers for Each Cell Type Based on Normalized Gene Expression.

Supplementary Material 2: Supplementary Table 1. CellRanger metrics and quality control cells for single-cell and single-nucleus datasets. Supplementary Table 2. Known marker genes for seven anatomical sites of Bama pigs with references from 17 studies. Supplementary Table 3. Detailed potential marker genes of distinct cell populations in each anatomical tissue, based on individual dataset analyses.

Acknowledgements

We thank the high-performance computing platform of Sichuan Agricultural University for providing data analysis support. This work was supported by the National Key R & D Program of China (2023YFD1300400 to F.K.), the Sichuan Science and Technology Program (2021ZDZX0008 to L.J. and 2021YFYZ0009 to M.L.), the National Natural Science Foundation of China (32272837 to L.J. and 32225046 to M.L.), and the Program for Pig Industry Technology System Innovation Team of Sichuan Province (SCCXTD-2024-8).

Authors' contributions

B.Z. and L.J. conceptualized the study and rigorously monitored and revised the manuscript. L.C. conducted the bioinformatic analyses, created the figures, and drafted the manuscript. X.Y.T. organized the tables and figures. Y.J.W. and C.L. assisted in manuscript writing. X.Q. and C.T. provided support for the bioinformatic analysis. M.Z.L. and F.L.K. offered critical assistance with the data samples.

Funding

This work was supported by the National Key R & D Program of China (2023YFD1300400 to F.K.), the Sichuan Science and Technology Program (2021ZDZX0008 to L.J. and 2021YFYZ0009 to M.L.), the National Natural Science Foundation of China (32272837 to L.J. and 32225046 to M.L.)

Data availability

The raw sequencing data in the FASTQ files constitute a reservoir of unprocessed information that will be vital for an in-depth exploration of the genomic landscape of pigs. The sequencing data for this study have been uploaded to the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) under accession number GSE24155546. This includes raw.fastq.gz files for scRNA-seq samples (4 samples, each with 6 fastq.gz files: index1, index2, and two sets of paired read 1 and read 2) and raw.fastq.gz files for snRNA-seq data (3 samples, each with 3 fastq.gz files: index1, read 1, and read 2). Additionally, we provide intermediate files for both the scRNA-seq and snRNA-seq datasets: filtered_feature_bc_matrix and raw_feature_bc_matrix output folders from CellRanger Count, as well as web summary reports (web_summary.html), which can be found under the GEO accession ID GSE241555 as supplementary files for each replicate (*_202303.tar.gz) [51]. The original fastq.gz files for each sequencing run can be found under the sample section via the SRA accession or by clicking the link labeled "SRA Run Selector" [51]. Our final Seurat objects with UMAP embeddings for scRNA-seq and snRNA-seq data can be found on figshare [52]. Related data, including seven raw and normalized count matrices, as well as metadata (QC metrics, sample IDs, doublet removal, batch effect removal and cell type annotations for the seven datasets), are provided as "RNA_rawcounts_matrix", "Normalized_QC_matrix", "Sc_Integrated_normalized_QC_matrix", and "Sn_Integrated_normalized_QC_matrix" files under the same figshare project.

Declarations

Ethics approval and consent to participate
All experiments were conducted in accordance with the Administration of Affairs Concerning Experimental Animals (Ministry of Science and Technology, China, revised in March 2017) and approved by the Animal Ethical and Welfare Committee (AEWC) of Sichuan Agricultural University under permit No. DKY-2021302126.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 26 November 2024 / Accepted: 6 March 2025
Published online: 12 March 2025

References

1. Chen A, Liao S, Cheng M, Ma K, Wu L, Lai Y, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. Cell. 2022;185(10):1777–e9221.
2. Ding J, Sharon N, Bar-Joseph Z. Temporal modelling using single-cell transcriptomics. Nat Rev Genet. 2022;23(6):355–68.
3. Rao A, Barkley D, Franca GS, Yanai I. Exploring tissue architecture using Spatial transcriptomics. Nature. 2021;596(7871):211–20.
4. Chen MC, Chang JP, Chang TH, Hsu SD, Huang HD, Ho WC, et al. Unraveling regulatory mechanisms of atrial remodeling of mitral regurgitation pigs by gene expression profiling analysis: role of type I angiotensin II receptor antagonist. Transl Res. 2015;165(5):599–620.
5. Ferguson SH, Foster DM, Sherry B, Magness ST, Nielsen DM, Gookin JL. Interferon-lambda3 promotes epithelial defense and barrier function against Cryptosporidium parvum infection. Cell Mol Gastroenterol Hepatol. 2019;8(1):1–20.
6. Wu YQ, Zhao H, Li YJ, Khederzadeh S, Wei HJ, Zhou ZY, et al. Genome-wide identification of imprinted genes in pigs and their different imprinting status compared with other mammals. Zool Res. 2020;41(6):721–5.
7. Nunez Y, Radovic C, Savic R, Garcia-Casco JM, Candek-Potokar M, Benitez R et al. Muscle transcriptome analysis reveals molecular pathways related to oxidative phosphorylation, antioxidant defense, fatness and growth in Mangalitsa and Moravka pigs. Anim (Basel). 2021;11(3):844.
8. Foissac S, Djebali S, Munyard K, Vialaneix N, Rau A, Muret K, et al. Multi-species annotation of transcriptome and chromatin structure in domesticated animals. BMC Biol. 2019;17(1):108.
9. Jin L, Tang Q, Hu S, Chen Z, Zhou X, Zeng B, et al. A pig bodymap transcriptome reveals diverse tissue physiologies and evolutionary dynamics of transcription. Nat Commun. 2021;12(1):3715.
10. Zhang L, Zhu J, Wang H, Xia J, Liu P, Chen F, et al. A high-resolution cell atlas of the domestic pig lung and an online platform for exploring lung single-cell data. J Genet Genomics. 2021;48(5):411–25.
11. Zhao Y, Hou Y, Xu Y, Luan Y, Zhou H, Qi X, et al. A compendium and comparative epigenomics analysis of cis-regulatory elements in the pig genome. Nat Commun. 2021;12(1):2217.
12. Cho CS, Xi J, Si Y, Park SR, Hsu JE, Kim M, et al. Microscopic examination of Spatial transcriptome using Seq-Scope. Cell. 2021;184(13):3559–e7222.
13. Jovic D, Liang X, Zeng H, Lin L, Xu F, Luo Y. Single-cell RNA sequencing technologies and applications: A brief overview. Clin Transl Med. 2022;12(3):e694.
14. Zhang J, Song C, Tian Y, Yang X, Single-Cell RNA. Sequencing in lung cancer: revealing phenotype shaping of stromal cells in the microenvironment. Front Immunol. 2021;12:802080.
15. Jiang H, Yu D, Yang P, Guo R, Kong M, Gao Y, et al. Revealing the transcriptional heterogeneity of organ-specific metastasis in human gastric cancer using single-cell RNA sequencing. Clin Transl Med. 2022;12(2):e730.
16. Wang F, Ding P, Liang X, Ding X, Brandt CB, Sjostedt E, et al. Endothelial cell heterogeneity and microglia Regulons revealed by a pig cell landscape at single-cell level. Nat Commun. 2022;13(1):3620.
17. Zheng Y, Li S, Li SH, Yu S, Wang Q, Zhang K, et al. Transcriptome profiling in swine macrophages infected with African swine fever virus at single-cell resolution. Proc Natl Acad Sci U S A. 2022;119(19):e2201288119.
18. Cai S, Hu B, Wang X, Liu T, Lin Z, Tong X, et al. Integrative single-cell RNA-seq and ATAC-seq analysis of myogenic differentiation in pig. BMC Biol. 2023;21(1):19.
19. Kimble KM, Dickinson SE, Biase FH. Extraction of total RNA from single-oocytes and single-cell mRNA sequencing of swine oocytes. BMC Res Notes. 2018;11(1):155.
20. Ma Z, Mao C, Chen X, Yang S, Qiu Z, Yu B, et al. Peptide vaccine against ADAMTS-7 ameliorates atherosclerosis and postinjury Neointima hyperplasia. Circulation. 2023;147(9):728–42.
21. Mo J, Lu Y, Xing T, Xu D, Zhang K, Zhang S, et al. Blood metabolic and physiological profiles of Bama miniature pigs at different growth stages. Porcine Health Manag. 2022;8(1):35.
22. Vaure C, Gregoire-Barou V, Courtois V, Chautard E, Degletagne C, Liu Y. Gottingen minipigs as a model to evaluate longevity, functionality, and memory of immune response induced by pertussis vaccines. Front Immunol. 2021;12:613810.
23. Moss A, Robbins S, Achanta S, Kuttippurathu L, Turick S, Nieves S, et al. A single cell transcriptomics map of paracrine networks in the intrinsic cardiac nervous system. iScience. 2021;24(7):102713.
24. Li T, Morselli M, Su T, Million M, Larauche M, Pellegrini M, et al. Comparative transcriptomics reveals highly conserved regional programs between Porcine and human colonic enteric nervous system. Commun Biol. 2023;6(1):98.
25. Firl DJ, Lassiter G, Hirose T, Policastro R, D'Attilio A, Markmann JF, et al. Clinical and molecular correlation defines activity of physiological pathways in life-sustaining kidney xenotransplantation. Nat Commun. 2023;14(1):3022.
26. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33(5):495–502.
27. McGinnis CS, Murrow LM, Gartner ZJ, DoubletFinder. Doublet detection in Single-Cell RNA sequencing data using artificial nearest neighbors. Cell Syst. 2019;8(4):329–37. e4.
28. McDavid A, Finak G, Gottardo R. The contribution of cell cycle to heterogeneity in single-cell RNA-seq data. Nat Biotechnol. 2016;34(6):591–3.
29. Chen L, Li H, Teng J, Wang Z, Qu X, Chen Z et al. Construction of a multi-tissue cell atlas reveals cell-type-specific regulation of molecular and complex phenotypes in pigs[Preprint]. 2023.
30. Xiong X, Kuang H, Ansari S, Liu T, Gong J, Wang S, et al. Landscape of intercellular crosstalk in healthy and NASH liver revealed by Single-Cell secretome gene analysis. Mol Cell. 2019;75(3):644–60. e5.
31. Li X, Li S, Wu B, Xu Q, Teng D, Yang T, et al. Landscape of immune cells heterogeneity in liver transplantation by Single-Cell RNA sequencing analysis. Front Immunol. 2022;13:890019.
32. Liang Y, Kaneko K, Xin B, Lee J, Sun X, Zhang K, et al. Temporal analyses of postnatal liver development and maturation by single-cell transcriptomics. Dev Cell. 2022;57(3):398–414. e5.
33. Liao J, Yu Z, Chen Y, Bao M, Zou C, Zhang H, et al. Single-cell RNA sequencing of human kidney. Sci Data. 2020;7(1):4.
34. Park J, Shrestha R, Qiu C, Kondo A, Huang S, Werth M, et al. Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. Science. 2018;360(6390):758–63.
35. Miao Z, Balzer MS, Ma Z, Liu H, Wu J, Shrestha R, et al. Single cell regulatory landscape of the mouse kidney highlights cellular differentiation programs and disease targets. Nat Commun. 2021;12(1):2277.
36. Madissoon E, Wilbrey-Clark A, Miragaia RJ, Saeb-Parsy K, Mahbubani KT, Georgakopoulos N, et al. scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. Genome Biol. 2019;21(1):1.
37. Dos Santos M, Backer S, Saintpierre B, Izac B, Andrieu M, Letourneur F, et al. Single-nucleus RNA-seq and FISH identify coordinated transcriptional activity in mammalian myofibers. Nat Commun. 2020;11(1):5102.
38. Machado L, Geara P, Camps J, Dos Santos M, Teixeira-Clerc F, Van Herck J, et al. Tissue damage induces a conserved stress response that initiates quiescent muscle stem cell activation. Cell Stem Cell. 2021;28(6):1125–35. e7.
39. Chemello F, Wang Z, Li H, McAnally JR, Liu N, Bassel-Duby R, et al. Degenerative and regenerative pathways underlying Duchenne muscular dystrophy revealed by single-nucleus RNA sequencing. Proc Natl Acad Sci U S A. 2020;117(47):29691–701.
40. Ma S, Sun S, Geng L, Song M, Wang W, Ye Y, et al. Caloric restriction reprograms the Single-Cell transcriptional landscape of Rattus Norvegicus aging. Cell. 2020;180(5):984–e100122.
41. Norreen-Thorsen M, Struck EC, Oling S, Zwahlen M, Von Feilitzen K, Odeberg J, et al. A human adipose tissue cell-type transcriptome atlas. Cell Rep. 2022;40(2):111046.

42. Sarvari AK, Van Hauwaert EL, Markussen LK, Gammelmark E, Marcher AB, Ebbesen MF, et al. Plasticity of epididymal adipose tissue in response to Diet-Induced obesity at Single-Nucleus resolution. Cell Metab. 2021;33(2):437–53. e5.
43. Whytock KL, Sun Y, Divoux A, Yu G, Smith SR, Walsh MJ, et al. Single cell full-length transcriptome of human subcutaneous adipose tissue reveals unique and heterogeneous cell populations. iScience. 2022;25(8):104772.
44. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with harmony. Nat Methods. 2019;16(12):1289–96.
45. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 2020;21(1):12.
46. MacParland SA, Liu JC, Ma XZ, Innes BT, Bartczak AM, Gage BK, et al. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. Nat Commun. 2018;9(1):4383.
47. Mailliard RB. Dendritic Cells and Antiviral Defense. Viruses. 2020;12(10):1152.
48. Juul NH, Stockman CA, Desai TJ. Niche cells and signals that regulate lung alveolar stem cells in vivo. Cold Spring Harb Perspect Biol. 2020;12(12):a035717.
49. Hall AM, Unwin RJ. A case of Drug-Induced proximal tubular dysfunction. Clin J Am Soc Nephrol. 2019;14(9):1384–7.
50. Meng D, Qin Y, Lu N, Fang K, Hu Y, Tian Z, et al. Kupffer Cells Promote the Differentiation of Adult Liver Hematopoietic Stem and Progenitor Cells into Lymphocytes via ICAM-1 and LFA-1 Interaction. Stem Cells Int. 2019;2019:4848279.
51. Chen L, Zeng B & Jin L. GEO. https://identifiers.org/geo/GSE241555. (2023).
52. Chen L, Zeng B, Jin L. A dataset of single-cell transcriptomic atlas of Bama pigs and potential marker genes across seven tissues. figshare, https://doi.org/10.6084/m9.figshare.c.7247752. (2024).

Publisher's note