SMARTdb: an integrated database for exploring single-cell multi-omics data of reproductive medicine

Authors: Zekai Liu[1,†], Zhen Yuan[1,†], Yunlei Guo[1,†], Yusheng Guan[1,3], Ruilin Wang[1], Zhanglian Wang[1], Yunan Chen[1], Tianlu Wang[1], Meining Jiang[1], Shuhui Bian[1,2,3]*

[1]State Key Laboratory of Reproductive Medicine and Offspring Health, Nanjing Medical University, Nanjing 211166, P.R. China

[2]Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing 211166, P.R. China

[3]School of Basic Medical Sciences, Nanjing Medical University, Nanjing 211166, P.R. China

***Correspondence:** bianshuhui@njmu.edu.cn (S. Bian)

†Co-first authors.

Abstract

Single-cell multi-omics sequencing has greatly accelerated reproduction research in recent years, and the data are continually growing. However, utilizing those data resources is challenging for wet-lab researchers, and therefore a comprehensive platform for exploring single-cell multi-omics data related to reproduction is urgently needed. Here we introduce SMARTdb (single-cell multi-omics atlas of reproduction), which is an integrative and user-friendly platform for exploring molecular dynamics of reproductive development, aging, and disease, covering multi-omics, multi-species, and multi-stages. We have curated and analyzed single-cell transcriptome and epigenome data of nearly 1.8 million cells from 5 species across whole lifespan. A series of powerful functionalities are provided, such as "Query gene expression", "DIY expression plot", "DNA methylation plot", and "Epigenome browser". With SMARTdb, we found that the male germ-cell-specific expression pattern of *RPL39L* and *RPL10L* is conserved between human and other model animals. Female germ cells also highly express RPL39L in fetal stage, but gradually decrease RPL39L expression in adult stage. Moreover, DNA hypomethylation and open chromatin may regulate the specific expression pattern of RPL39L collectively in both male and female germ cells. In summary, SMARTdb is a powerful platform for convenient data mining and gaining novel insights into reproductive development, aging and disease.

SMARTdb is publicly available at https://smart-db.cn.

Key words: single-cell multi-omics, reproduction, embryo, germ cell, gonad

Introduction

Single-cell multi-omics technique has greatly promoted our molecular understanding of early embryos[1–14], fetal germ cells and gonadal somatic cells[15–22], developing testis and ovary[23–30], adult testis and ovary[26,29,31–38], aging testis and ovary[39–41], endometrium[42,43], maternal-fetal interface[44–46], as well as reproductive diseases[47–52]. Multiple molecular layers, including gene expression patterns, DNA methylation dynamics, and open chromatin features, have been profiled with high-throughput sequencing technology, and the data have been growing continually. It is noted that previous single-cell multi-omics studies typically provide comprehensive atlas rather than digging into individual genes or genomic sites, but they offer valuable resources for more wet-lab researchers to perform further investigations into specific questions, genes, and genomic sites of interest. For example, it is important for researchers using animal models to know if the expression pattern of a specific gene is conserved in human, and whether knocking out of the gene could lead to important phenotypic changes.

However, analyzing the massive and complex single-cell multi-omics data requires experienced skills and a large number of computing resources, which is challenging and time-consuming for wet-lab researchers especially. For instance, as each of those studies typically focuses on a narrow time span, researchers who wish to query the expression levels of a particular gene in child, young adult, aging adult and infertile man need to download and process large amount of data from several separate studies, and they might find this process difficult and laborious due to a lack of analytical skills and computing resources. This inconvenience hinders the optimal utilization of the published data, thus limiting potential novel discoveries. Hence, integrated and user-friendly platforms are needed to bridge the gap between rich data resource and wet-lab researchers.

Endeavors such as Human Cell Atlas (HCA, https://www.humancellatlas.org/) and Human Protein Atlas (HPA, https://www.proteinatlas.org/) have served as excellent platforms for exploring massive single-cell data conveniently. However, these

databases do not contain much reproduction-related data. Currently, there are a few small-scale reproduction-related websites available, which typically only include data of their own study, are limited in data amount and time span, only provide simple function such as querying gene expression of a specific gene, and lack long-term maintenance. As a result, there is still an urgent need for an integrated and user-friendly platform which focuses on promoting efficient utilization of public single-cell multi-omics data related to reproduction across full life-cycle.

Here we introduce SMARTdb (single-cell multi-omics atlas of reproduction), which is a manually curated, integrative and interactive platform aiming to bridge the gap between mass data resources and wet-lab researchers, and promote exploring molecular dynamics of reproductive development, aging, and disease.

Data collection and processing

Data collection

We curated single-cell transcriptome and epigenome (DNA methylation and chromatin accessibility) data related to reproduction from published studies (Table S1). The gene expression matrix, DNA methylation, and chromatin accessibility data were downloaded from NCBI GEO (https://www.ncbi.nlm.nih.gov/geo), EMBL-EBI (https://www.ebi.ac.uk/), and Zenodo (https://zenodo.org/). In addition to data of developmental stages and normal physiological states, data of aging and human infertility (such as non-obstructive azoospermia, NOA) are also incorporated. In summary, the single-cell data of nearly 1.8 million individual cells across five species (human, monkey, mouse, pig, and buffalo) were curated, involving early embryos, fetal germ cells and gonadal somatic cells, developing testis and ovary, adult testis and ovary, aging testis and ovary, endometrium, maternal-fetal interface, and human infertility. These data cover the entire life cycle, and span >100 specific time points of 5 species (Table S1).

Processing of single-cell transcriptome data and cell clustering

To reproduce the cell clustering results of original studies as much as possible, for the datasets with metadata provided by the original study, only the cells with cell type

identities were used in downstream analyses. For the datasets without provided metadata, the cells were filtered according to the quality control standards mentioned in the original studies. For those studies not providing concrete cutoff values, only cells with >500 expressed genes and <20% mitochondrial gene ratios were retained. Seurat[53] (version 4.3.0) was used to perform the following processing steps. Briefly, the UMI count of each gene for each cell was first divided by the total counts for that cell and multiplied by 10,000 (or author-designated parameters), and was then natural-log transformed. The top highly variable genes among cells were selected with "FindVariableFeatures" function with default or author-designated parameters. The normalized data were then centered and scaled, resulting in that the mean expression of each gene across all cells was zero and the standard deviation was one. Principal component analysis (PCA) was performed on the scaled data with the previously determined highly variable genes used as input. Then the cells were clustered with Louvain algorithm. The uniform manifold approximation and projection (UMAP) was used to visualize the cell clustering in low-dimensional space. For the datasets with batch effects, the anchors of different samples were identified and then different samples were integrated to correct batch effects. For the datasets without obvious batch effects, no additional correction was performed.

Cell type annotation and differentially expressed gene analysis of single-cell transcriptome data

A series of analyses were performed to ensure the accuracy of cell identities for each dataset. For the datasets with metadata provided by the original study, the cells were annotated correspondingly, and classical marker genes for each cell type were checked to confirm accuracy. For the datasets without metadata, the top differentially expressed genes (DEGs) of each cell cluster were identified using the Wilcoxon rank sum test with the "FindAllMarkers" function in Seurat[53] (version 4.3.0) with default parameters (logfc.threshold=0.25, min.pct=0.1, return.thresh=0.01), and their consistency with classical marker genes for each cell type was carefully checked. After these steps, the cell type identities can be annotated accurately.

Visualization of gene expression levels

First, Seurat objects were converted into "anndata" objects with the R package sceasy (version 0.0.7). Then, the dot plots, heatmaps and violin plots were generated with Scanpy[54] (version 1.9.3). For dot plots and heatmaps, both the relative expression levels with and without standard scaling are provided. The standard scale normalization strategy in Scanpy was performed with parameters "standard_scale = 'var'", which standardized the expression level of a given gene between 0 and 1 by subtracting the minimum and dividing each by its maximum.

Identification of transcription factor motifs

The *de novo* motif discovery was performed with HOMER software (version 4.11)[55] with default parameters.

Processing of single-cell epigenome data

The processed single-cell DNA methylation and chromatin accessibility data provided by the original study were downloaded from GEO database, and were further processed to standard BigWig format with bedtools (version 2.30.0), UCSC bedGraphToBigWig (version 2.9) and bigWigMerge (version 2) for downstream usage. For the data generated by scBS-seq[56] technique, only CpG sites were retained. For the data generated by scCOOL-seq[8] technique, WCG (W=A, T) sites were used for DNA methylation analysis and GCH (H=A, C, T) sites were used for chromatin accessibility analysis.

Considering the sparse coverage of single-cell epigenome data, we merged single-cell epigenome data according to their cell type identities and stages (e.g. 2-cell embryo, 8-week female mitotic germ cells, etc.), and sequencing techniques (e.g. scBS-seq and scCOOL-seq) using custom scripts. Briefly, for the genomic sites only covered by one cell, they were directly retained in the merged files. For the genomic sites covered by multiple cells in the group, the mean values of covered cells were calculated to represent the final states of the sites. For single-cell DNA methylation data, only the genomic sites with DNA methylation levels <0.1 or >0.9 in each individual cell were used in downstream merging. The merged single-cell data of cell groups are more convenient and efficient for downstream exploration.

The DNA methylation "Tanghulu" plot was implemented by a custom Python script to visualize DNA methylation state of each CpG site in a narrow genomic region of interest. For single-cell DNA methylation data, only the genomic sites with values <0.1 or >0.9 in each individual cell will be plotted. The WashU epigenome browser[57] was incorporated into SMARTdb for convenient visualization of single-cell DNA methylation levels and chromatin accessibility states.

Implementation of the database website

SMARTdb was developed utilizing Python 3 in conjunction with the Django framework (https://djangoproject.com). The data housed within SMARTdb is systematically organized and accessible through PostgreSQL (https://www.postgresql.org). The user-friendly front-end interface was created using NG-ZORRO and Angular. The platform is hosted using the Nginx web server (http://nginx.org) and is freely accessible online at https://smart-db.cn without the need for registration or login. For optimal performance, SMARTdb is best used on major web browsers such as Google Chrome, Microsoft Edge, Safari, and Mozilla Firefox, with JavaScript enabled.

Database content and usage

Overall design

SMARTdb is an integrative and interactive platform for exploring single-cell multi-omics data related to reproduction. It was designed to have three main features (Figures 1 and S1). First, single-cell multi-omics data of nearly 1.8 million individual cells were manually curated and analyzed (Table S1). Users can access gene expression levels, DNA methylation levels and chromatin accessibility levels at single-cell resolution conveniently. Second, the main stages of full life-cycle were covered, including zygote, early embryo, fetus, infant, child, puberty, adult, and aging. More than 100 specific time points are included in total. Besides, data of human infertility are also incorporated. Users can not only explore molecular dynamics during full life-cycle, but also compare normal physiological state with disease. Third, considering cross-species conservation analyses are very important, data from five species were included, including human, non-human primate (monkey), mouse, pig, and buffalo. For example, the researchers

using mouse as animal model can find out if the expression pattern of the gene that they decide to investigate is conserved in human, which will help select more meaningful research targets.

As a user-friendly platform, SMARTdb offers multiple featured function modules to users for exploring the rich single-cell multi-omics data resources, such as "Query gene expression", "DIY expression plot", "DNA methylation plot" and "Epigenome browser". The featured function modules are described below.

"Overview" and "Query gene expression" modules

In the "Overview" module, SMARTdb facilitates users to get a quick overview of cell clustering, sample origins, cell counts, and DEGs of each cell type for each scRNA-seq dataset (Figure 2A). In the "Query gene expression" module, users can conveniently plot the expression levels of a selected gene across the cell types in the chosen dataset with UMAP and violin plots (Figure 2B). Additionally, a concise introduction for each human gene is also provided in this module.

"DIY expression plot" module

This module provides a very practical function for users to customize expression plots according to their personalized requirements (Figure 2C). Users can select multiple genes and multiple cell types of interest to plot simultaneously, and choose appropriate plot forms for visualization. Three main visualization plots are offered in this module, including dot plots, heatmaps, and violin plots. Moreover, the orders of genes and cell types displayed in the plots can be designated by users freely. Users can also choose whether to standardly scale the expression levels within each gene when using dot plots and heatmaps. The generated figures are of publication quality.

"DNA methylation plot" module

This module provides "Tanghulu" plots to visualize DNA methylation status of each cytosine site in CpG or WCG context (Figure 3A). Users can conveniently generate a customized "Tanghulu" plot by selecting datasets and cell types of interest, and designating specific genomic region range. In "Tanghulu" plot, each row represents an individual cell, and each circle represents a cytosine site in CpG or WCG context. Black circles stand for methylated sites, while the white circles refer to unmethylated sites.

"Epigenome browser" module

This module incorporates the WashU epigenome browser tool[57] and offers a state-of-the-art solution for browsing epigenomic datasets (Figure 3B). Users can select single-cell DNA methylation and chromatin accessibility datasets of various cell types based on their interest, such as oocytes, sperms, early embryos, fetal germ cells, and gonadal somatic cells. The DNA methylation and chromatin accessibility levels of selected samples can be directly and conveniently visualized in the browser. Especially, since single-cell data are typically sparse, the single-cell data of each cell type or sample type are merged for optimal visualization and enabling more efficient exploration. The genomic region locator allows users to navigate to a specific gene or designated genomic region, zoom in and out, and move to new regions by dragging the mouse left and right. Additionally, the browser interface is allowed to be customized to meet personalized requirements, such as display mode (bar plot or heatmap), RGB colors, axis scales, and text labels.

Examples for using SMARTdb

Here we illustrate how to make good use of SMARTdb to facilitate exploration of single-cell multi-omics data and promote new findings. A recent study has revealed that RPL39L is a male germ-cell-specific ribosome protein and plays critical roles in spermatogenesis and male fertility in mouse[58]. Deletion of *Rpl39l* in mice can lead to defective sperm formation and reduced fertility[58]. With high-purity enrichment and proteomic analyses for mouse spermatogonia cells, pachytene spermatocytes, round spermatids, elongated spermatids, and somatic cells (Sertoli cells and interstitial cells), the study showed that RPL39L was mainly expressed in germ cells, whereas the corresponding core ribosome protein RPL39 displayed decreased expression levels in germ cells during spermatogenesis and was mainly expressed in Sertoli cells and interstitial cells. The study also mentioned that the expression of RPL10L was testis-specific in bulk level, but its expression levels in specific subtypes of male germ cells were not revealed. Inspired by the study, we further asked:

(1) Whether the germ-cell-specific expression pattern is conserved in testis of human, non-human primates as well as other animal models?

(2) When does the germ-cell-specific expression pattern of RPL39L form? In fetal stage, peri-puberty or adulthood?

(3) Does epigenetic regulation function in regulating the male germ-cell-specific expression pattern? We demonstrate how to explore the three questions with SMARTdb below.

Cross-species conservation

To answer the first question, we utilized the "DIY expression plot" module to plot the expression levels of the ribosomal large subunit protein genes *RPL39L* and *RPL10L*, and the corresponding core ribosome protein genes *RPL39* and *RPL10* in human[36], as well as their homologous genes in mouse[37], monkey[32], and pig[28] (Figures 2C and 4A). For human adult testis[36], taking the advantage of accurate cell type identification with single-cell RNA-seq data, we found that the expression levels of *RPL39L* and *RPL10L* peak at spermatocyte (especially pachytene and/or diplotene spermatocyte), and then reduce in later stages of spermatogenesis at mRNA level (Figure 2, B and C). In contrast, *RPL39* and *RPL10* decreased their expression levels during spermatogenesis, and mainly expressed in somatic cells (Figure 2C). Moreover, the germ-cell-specific expression pattern of *RPL39L* and *RPL10L*, and decreasing of *RPL39* and *RPL10* during spermatogenesis are conserved in human and other species (Figure 4A).

Developing dynamics

To clarify the second question, we plotted the expression levels of *RPL39L* in fetal stage[18] and peri-pubertal stage[24] with the "DIY expression plot" module (Figure 4B).We found that the germ cells in fetal stages have started to highly express *RPL39L*, as early as in migrating primordial germ cells (PGCs). Meanwhile, although the somatic niche cells in fetal stage relatively lowly expressed *RPL39L*, their expressed percentages were still higher than those of postnatal stages (peri-puberty and adulthood), indicating that the germ-cell specific pattern of *RPL39L* had started to establish but had not been completely established in fetal stage at mRNA level. In human male peri-puberty, the germ-cell-specific pattern of *RPL39L* has been established, with high expression in germ cells and low expression in somatic cells (both in the expressed percentages and mean expression levels). Additionally, for the corresponding core ribosome protein gene *RPL39*, some subtypes of fetal germ cells still have relatively

high expression levels of *RPL39*, such as male migrating PGCs, and both male and female mitotic germ cells. Meanwhile, in human male peri-pubertal stage, the expression levels and expressed percentage of *RPL39* decreased gradually from spermatogonia to spermatid, leaving only somatic cells with relatively high expression levels.

Moreover, we noticed that *RPL39L* transcripts are highly expressed in both male and female germ cells in human fetal stage, but diverge between male and female in adulthood. The fully-grown, MI, and MII oocytes sharply decrease the expression levels of *RPL39L*, leaving only growing oocytes in human adult ovary[34] maintaining relatively high expression levels of *RPL39L* (Figure 5A). As a result, *RPL39L* is expressed in male germ cells and female growing oocytes in human adulthood. The divergence in the *RPL39L* expression among oocytes of various stages in human adult ovary was detected in mRNA level, indicating potential transcriptional regulation of the gene.

Regulation mechanism

Then we further explored the transcriptional regulatory mechanism using the "Epigenome browser" module. With epigenetic data of human fetal stage[15], we found that germ cells in fetal stage have hypomethylated and more accessible regions around the promoter of *RPL39L* compared with somatic cells (Figures 3B and S2), which is consistent with the relatively high expression of *RPL39L* of fetal germ cells (Figure 4B). In human adult female[34], growing oocytes displayed similar DNA methylation and chromatin accessibility pattern with fetal germ cells (Figure 5B). In contrast, fully-grown, MI, and MII oocytes up-regulate the DNA methylation levels around *RPL39L* promoter, and down-regulate the chromatin accessibility levels around *RPL39L* promoter (Figure 5B), which may explain the decreased expression level of RPL39L in fully-grown, MI, and MII oocytes. Hence, epigenetic regulation potentially plays important roles in shaping the expression pattern of *RPL39L* in both male and female human germ cells during both fetal and adult stages.

In summary, we validated that *RPL39L* and *RPL10L* are germ-cell-specific at mRNA level, which is consistent with previous studies[58]. Moreover, we obtained multiple

novel findings with SMARTdb. We found that the male germ-cell-specific expression pattern is conserved between human and other model animals, illustrating its importance during evolution. The germ-cell-specific pattern of *RPL39L* at mRNA level has started to establish but has not been completely established in fetal gonad. In human adulthood, *RPL39L* is relatively highly expressed in male germ cells and female growing oocytes, but lowly expressed in fully-grown, MI, and MII oocytes. Furthermore, DNA methylation and chromatin accessibility may cooperate in regulating the divergence of *RPL39L* expression in germ cells of human adult at mRNA level. These results demonstrate the power of SMARTdb, providing valuable clues for further investigations.

Discussion

The mystery of reproductive development, aging and disease has long been explored. Single-cell multi-omics sequencing technique has revolutionized researches in reproduction in recent years, and has led to accelerated accumulation of massive datasets. The public data provide valuable resources for researchers to perform further data mining. However, difficulties in processing big data for broad wet-lab researchers prevent them from being effectively utilized. To deal with the situation of growing data and promote a more efficient usage, we present SMARTdb – a single-cell multi-omics platform for reproduction research. To the best of our knowledge, SMARTdb is the first integrated platform with multiple powerful functionalities to explore single-cell multi-omics data in the field of reproductive medicine. Moreover, it will be continuously updated by adding high-quality single-cell multi-omics data from reproductive research.

Our database will promote reproductive medicine research in several aspects. For example, researchers using animal models are able to speculate whether knocking out of a specific gene could potentially cause important phenotypic changes. Users can know whether the expression pattern of a specific gene is conserved in human, which helps select more meaningful research targets. For researcher using human samples, cross-species conservation analyses provide important clues for selecting suitable animal models for further validation. Cross-stage dynamic analysis helps select the

right time window for further investigations. Moreover, epigenetics data resources will provide novel insights for potential regulatory mechanisms.

We showed examples for using SMARTdb and demonstrated the power of it by exploring germ-cell-specific *RPL39L* and *RPL10L*. We not only validated the germ-cell-specific expression of ribosome gene *RPL39L*, but also gained a deeper understanding of its cross-species conservation, cross-stage expression dynamics, and potential epigenetic regulation. Moreover, taking the advantages of single-cell transcriptome sequencing, the gene expression patterns in multiple cell types can be observed more accurately and in detail. For example, we found that *RPL39L* was not only specifically expressed in male germ cells in human adult testis, but also highly expressed in the growing oocytes (one subtype of oocytes) in human adult ovary.

In conclusion, we firmly believe that with the assistance of SMARTdb, researchers in the field of reproduction can easily access and explore massive single-cell multi-omics data, and hopefully, inspire novel ideas and new findings.

References

1. Guo, H. *et al.* The DNA methylation landscape of human early embryos. *Nature* **511**, 606–610 (2014).
2. Zhu, P. *et al.* Single-cell DNA methylome sequencing of human preimplantation embryos. *Nat. Genet.* **50**, 12–19 (2018).
3. Blakeley, P. *et al.* Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development* **142**, 3613–3613 (2015).
4. Yan, L. *et al.* Single-cell RNA-seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* **20**, 1131–1139 (2013).
5. Zhou, F. *et al.* Reconstituting the transcriptome and DNA methylome landscapes of human implantation. *Nature* **572**, 660–664 (2019).
6. Petropoulos, S. *et al.* Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* **165**, 1012–1026 (2016).
7. Guo, H. *et al.* Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.* **23**, 2126–2135 (2013).
8. Guo, F. *et al.* Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell Res.* **27**, 967–988 (2017).
9. Jiang, S. *et al.* Single-cell chromatin accessibility and transcriptome atlas of mouse embryos. *Cell Rep.* **42**, 112210 (2023).
10. Wang, Y. *et al.* Single-cell multi-omics sequencing reveals the functional regulatory landscape of early embryos. *Nat. Commun.* **12**, 1247 (2021).
11. Scialdone, A. *et al.* Resolving early mesoderm diversification through single-cell expression profiling. *Nature* **535**, 289–293 (2016).
12. Argelaguet, R. *et al.* Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**, 487–491 (2019).
13. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
14. Li, L. *et al.* Single-cell multi-omics sequencing of human early embryos. *Nat. Cell Biol.* **20**, 847–858 (2018).
15. Li, L. *et al.* Dissecting the epigenomic dynamics of human fetal germ cell development at single-cell resolution. *Cell Res.* **31**, 463–477 (2021).
16. Chitiashvili, T. *et al.* Female human primordial germ cells display X-chromosome dosage compensation despite the absence of X-inactivation. *Nat. Cell Biol.* **22**, 1436–1446 (2020).

17. Garcia-Alonso, L. *et al.* Single-cell roadmap of human gonadal development. *Nature* **607**, 540–547 (2022).
18. Li, L. *et al.* Single-cell RNA-seq analysis maps development of human germline cells and gonadal niche interactions. *Cell Stem Cell* **20**, 858-873.e4 (2017).
19. Guo, F. *et al.* The transcriptome and DNA methylome landscapes of human primordial germ cells. *Cell* **161**, 1437–1452 (2015).
20. Stévant, I. *et al.* Dissecting cell lineage specification and sex fate determination in gonadal somatic cells using single-cell transcriptomics. *Cell Rep.* **26**, 3272-3283.e3 (2019).
21. Chen, M. *et al.* Integration of single-cell transcriptome and chromatin accessibility of early gonads development among goats, pigs, macaques, and humans. *Cell Rep.* **41**, 111587 (2022).
22. Stévant, I. *et al.* Deciphering cell lineage specification during male sex determination with single-cell RNA sequencing. *Cell Rep.* **22**, 1589–1599 (2018).
23. Guo, J. *et al.* Single-cell analysis of the developing human testis reveals somatic niche cell specification and fetal germline stem cell establishment. *Cell Stem Cell* **28**, 764-778.e4 (2021).
24. Guo, J. *et al.* The dynamic transcriptional cell atlas of testis development during human puberty. *Cell Stem Cell* **26**, 262-276.e4 (2020).
25. Zhao, J. *et al.* Cell-fate transition and determination analysis of mouse male germ cells throughout development. *Nat. Commun.* **12**, 6839 (2021).
26. Ernst, C., Eling, N., Martinez-Jimenez, C. P., Marioni, J. C. & Odom, D. T. Staged developmental mapping and X chromosome transcriptional dynamics during mouse spermatogenesis. *Nat. Commun.* **10**, 1251 (2019).
27. Huang, L. *et al.* Single-cell RNA sequencing uncovers dynamic roadmap and cell-cell communication during buffalo spermatogenesis. *iScience* **26**, 105733 (2022).
28. Zhang, L. *et al.* Single-cell RNA-seq analysis of testicular somatic cell development in pigs. *J. Genet. Genomics* **49**, 1016–1028 (2022).
29. Sohni, A. *et al.* The neonatal and adult human testis defined at the single-cell level. *Cell Rep.* **26**, 1501-1517.e4 (2019).
30. Gu, C., Liu, S., Wu, Q., Zhang, L. & Guo, F. Integrative single-cell analysis of transcriptome, DNA methylome and chromatin accessibility in mouse oocytes. *Cell Res.* **29**, 110–123 (2019).
31. Chen, Y. *et al.* Single-cell RNA-seq uncovers dynamic processes and critical regulators in mouse spermatogenesis. *Cell Res.* **28**, 879–896 (2018).

32. Shami, A. N. *et al.* Single-cell RNA sequencing of human, macaque, and mouse testes uncovers conserved and divergent features of mammalian spermatogenesis. *Dev. Cell* **54**, 529-547.e12 (2020).
33. Fan, X. *et al.* Single-cell reconstruction of follicular remodeling in the human adult ovary. *Nat. Commun.* **10**, 3164 (2019).
34. Yan, R. *et al.* Decoding dynamic epigenetic landscapes in human oocytes using single-cell multi-omics sequencing. *Cell Stem Cell* **28**, 1641-1656.e7 (2021).
35. Zhang, Y. *et al.* Transcriptome landscape of human folliculogenesis reveals oocyte and granulosa cell interactions. *Mol. Cell* **72**, 1021-1034.e4 (2018).
36. Guo, J. *et al.* The adult human testis transcriptional cell atlas. *Cell Res.* **28**, 1141–1157 (2018).
37. Green, C. D. *et al.* A comprehensive roadmap of murine spermatogenesis defined by single-cell RNA-seq. *Dev. Cell* **46**, 651-667.e10 (2018).
38. Hermann, B. P. *et al.* The mammalian spermatogenesis single-cell transcriptome, from spermatogonial stem cells to spermatids. *Cell Rep.* **25**, 1650-1667.e8 (2018).
39. Nie, X. *et al.* Single-cell analysis of human testis aging and correlation with elevated body mass index. *Dev. Cell* **57**, 1160-1176.e5 (2022).
40. Huang, D. *et al.* A single-nucleus transcriptomic atlas of primate testicular aging reveals exhaustion of the spermatogonial stem cell reservoir and loss of Sertoli cell homeostasis. *Protein Cell* pwac057 (2022).
41. Wang, S. *et al.* Single-cell transcriptomic atlas of primate ovarian aging. *Cell* **180**, 585-600.e19 (2020).
42. Garcia-Alonso, L. *et al.* Mapping the temporal and spatial dynamics of the human endometrium in vivo and in vitro. *Nat. Genet.* **53**, 1698–1711 (2021).
43. Wang, W. *et al.* Single-cell transcriptomic atlas of the human endometrium during the menstrual cycle. *Nat. Med.* **26**, 1644–1653 (2020).
44. Vento-Tormo, R. *et al.* Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature* **563**, 347–353 (2018).
45. Arutyunyan, A. *et al.* Spatial multi-omics map of trophoblast development in early pregnancy. *Nature* **616**, 143–151 (2023).
46. Greenbaum, S. *et al.* A spatially resolved timeline of the human maternal–fetal interface. *Nature* **619**, 595–605 (2023).
47. Zhao, L. *et al.* Single-cell analysis of developing and azoospermia human testicles reveals central role of Sertoli cells. *Nat. Commun.* **11**, 5683 (2020).

48. Di Persio, S. *et al.* Single-cell RNA-seq unravels alterations of the human spermatogonial stem cell compartment in patients with impaired spermatogenesis. *Cell Rep. Med.* **2**, 100395 (2021).
49. Alfano, M. *et al.* Aging, inflammation and DNA damage in the somatic testicular niche with idiopathic germ cell aplasia. *Nat. Commun.* **12**, 5205 (2021).
50. Mahyari, E. *et al.* Comparative single-cell analysis of biopsies clarifies pathogenic mechanisms in Klinefelter syndrome. *Am. J. Hum. Genet.* **108**, 1924–1945 (2021).
51. Chen, Y. *et al.* Deciphering the molecular characteristics of human idiopathic nonobstructive azoospermia from the perspective of germ cells. *Adv. Sci.* 2206852 (2023).
52. Ferrero, H. *et al.* Single-cell RNA sequencing of oocytes from ovarian endometriosis patients reveals a differential transcriptomic profile associated with lower quality. *Hum. Reprod.* **34**, 1302–1312 (2019).
53. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
54. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
55. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
56. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820 (2014).
57. Li, D. *et al.* Washu epigenome browser update 2022. *Nucleic Acids Res.* **50**, W774–W781 (2022).
58. Li, H. *et al.* A male germ-cell-specific ribosome controls male fertility. *Nature* **612**, 725–731 (2022).

Acknowledgments

We thank professors Fuchou Tang, Jingchu Luo, Zhibin Hu, Jiahao Sha, and Ran Huo for discussion and suggestions. This work was supported by Young Elite Scientists Sponsorship Program by China Association for Science and Technology (CAST) (2020QNRC001) and research start-up funding from Nanjing Medical University (KY116RC20200007).

Author contributions

Shuhui Bian: Conceptualization, supervision, Writing - original draft, Writing - review & editing, Visualization, Data curation, Methodology, Formal analysis, Funding acquisition, Investigation, Project administration. **Zekai Liu:** Data curation, Formal analysis, Methodology, Software, Visualization, Writing - original draft. **Zhen Yuan:** Data curation, Formal analysis, Methodology, Software, Visualization, Writing - original draft. **Yunlei Guo:** Formal analysis, Methodology, Software, Visualization, Validation. **Yusheng Guan:** Data curation, Methodology, Validation.

Competing interests

The authors declare that they have no competing interests.

Data availability

SMARTdb is publicly available at https://smart-db.cn.

Figure 1

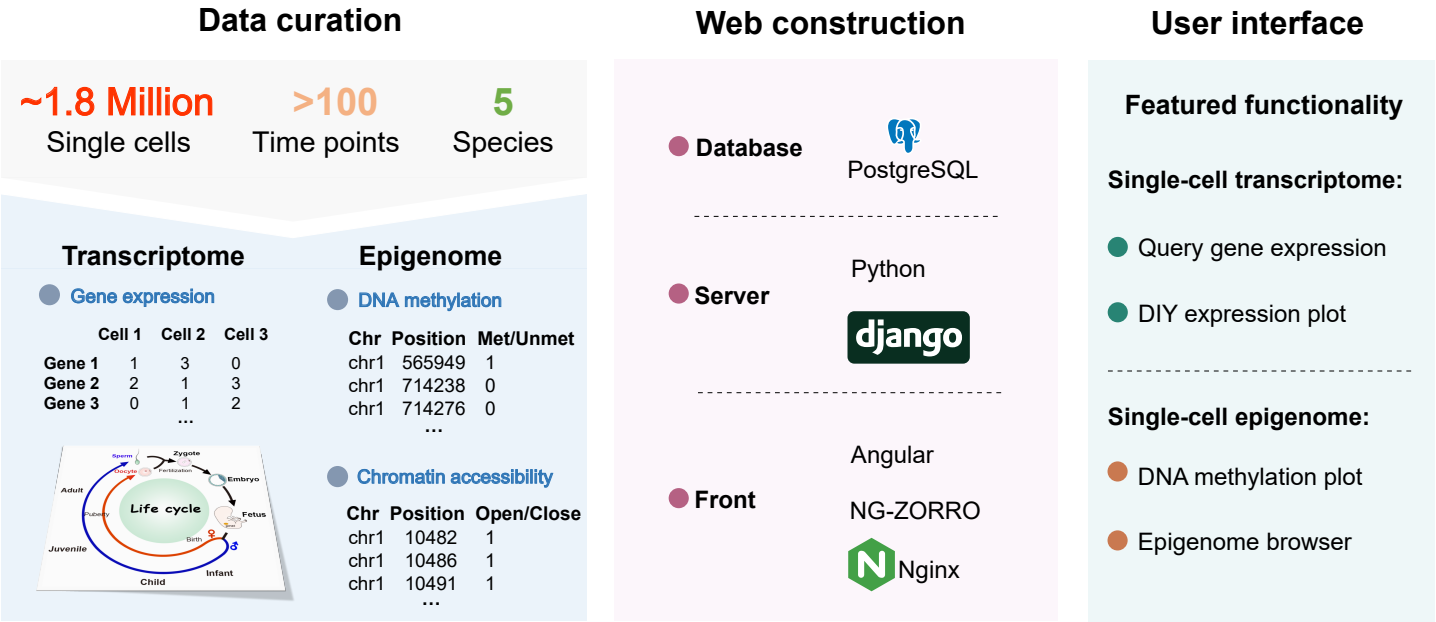

Data curation
~1.8 Million
Single cells
>100
Time points
5
Species
Transcriptome
Gene expression
Cell 1 Cell 2 Cell 3
Gene 1 1 3 0
Gene 2 2 1 3
Gene 3 0 1 2
...
Life cycle
Epigenome
DNA methylation
Chr Position Met/Unmet
chr1 565949 1
chr1 714238 0
chr1 714276 0
...
Chromatin accessibility
Chr Position Open/Close
chr1 10482 1
chr1 10486 1
chr1 10491 1
...
Web construction
Database
PostgreSQL
Server
Python
django
Front
Angular
NG-ZORRO
Nginx
User interface
Featured functionality
Single-cell transcriptome:
Query gene expression
DIY expression plot
Single-cell epigenome:
DNA methylation plot
Epigenome browser

Figure 1. A brief schema describing data curation, web construction and featured functionalities. The single-cell data of nearly 1.8 million individual cells across five species (human, monkey, mouse, pig, and buffalo) were curated, covering the entire life cycle and spanning >100 specific time points of 5 species. Multiple featured function modules are provided, such as "Query gene expression", "DIY expression plot", "DNA methylation plot" and "Epigenome browser".

Figure 2

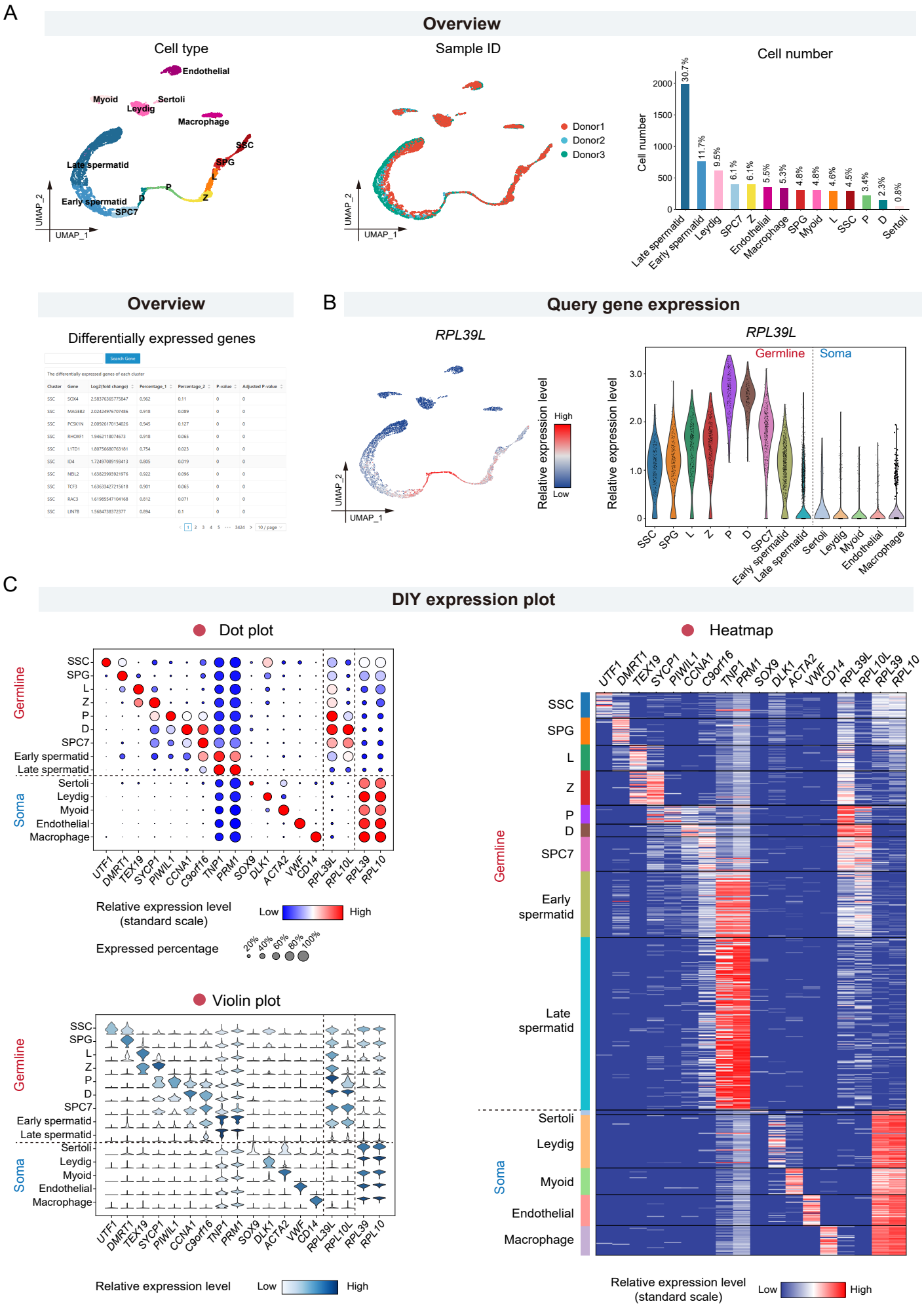

A
Overview
Cell type
Endothelial
Myoid
Sertoli
Leydig
Macrophage
SSC
SPG
L
Z
P
D
SPC7
Late spermatid
Early spermatid
UMAP_2
UMAP_1
Sample ID
Donor1
Donor2
Donor3
Cell number
Cell number
30.7%
11.7%
9.5%
6.1%
6.1%
5.5%
5.3%
4.8%
4.8%
4.6%
4.5%
3.4%
2.3%
0.8%
Late spermatid
Early spermatid
Leydig
SPC7
Z
Endothelial
Macrophage
SPG
Myoid
L
SSC
P
D
Sertoli
Overview
Differentially expressed genes
Search Gene
The differentially expressed genes of each cluster
Cluster | Gene | Log2(fold change) | Percentage_1 | Percentage_2 | P-value | Adjusted P-value
SSC | SOX4 | 2.58376365775847 | 0.962 | 0.11 | 0 | 0
SSC | MAGEB2 | 2.02424976707486 | 0.918 | 0.089 | 0 | 0
SSC | PCSK1N | 2.00926170134026 | 0.945 | 0.127 | 0 | 0
SSC | RHOXF1 | 1.94621118074673 | 0.918 | 0.065 | 0 | 0
SSC | L1TD1 | 1.80756680763181 | 0.754 | 0.023 | 0 | 0
SSC | ID4 | 1.72497089193413 | 0.805 | 0.019 | 0 | 0
SSC | NEIL2 | 1.63823993921976 | 0.922 | 0.096 | 0 | 0
SSC | TCF3 | 1.63633427215618 | 0.901 | 0.065 | 0 | 0
SSC | RAC3 | 1.61985547104168 | 0.812 | 0.071 | 0 | 0
SSC | LIN7B | 1.56847383723377 | 0.894 | 0.1 | 0 | 0
1 2 3 4 5 ··· 3424 > 10 / page
B
Query gene expression
RPL39L
UMAP_2
UMAP_1
Relative expression level
High
Low
RPL39L
Germline
Soma
Relative expression level
SSC
SPG
L
Z
P
D
SPC7
Early spermatid
Late spermatid
Sertoli
Leydig
Myoid
Endothelial
Macrophage
C
DIY expression plot
Dot plot
Germline
Soma
SSC
SPG
L
Z
P
D
SPC7
Early spermatid
Late spermatid
Sertoli
Leydig
Myoid
Endothelial
Macrophage
UTF1
DMRT1
TEX19
SYCP1
PIWIL1
CCNA1
C9orf16
TNP1
PRM1
SOX9
DLK1
ACTA2
VWF
CD14
RPL39L
RPL10L
RPL39
RPL10
Relative expression level
(standard scale)
Low
High
Expressed percentage
20%
40%
60%
80%
100%
Violin plot
Germline
Soma
SSC
SPG
L
Z
P
D
SPC7
Early spermatid
Late spermatid
Sertoli
Leydig
Myoid
Endothelial
Macrophage
UTF1
DMRT1
TEX19
SYCP1
PIWIL1
CCNA1
C9orf16
TNP1
PRM1
SOX9
DLK1
ACTA2
VWF
CD14
RPL39L
RPL10L
RPL39
RPL10
Relative expression level
Low
High
Heatmap
UTF1
DMRT1
TEX19
SYCP1
PIWIL1
CCNA1
C9orf16
TNP1
PRM1
SOX9
DLK1
ACTA2
VWF
CD14
RPL39L
RPL10L
RPL39
RPL10
Germline
Soma
SSC
SPG
L
Z
P
D
SPC7
Early spermatid
Late spermatid
Sertoli
Leydig
Myoid
Endothelial
Macrophage
Relative expression level
(standard scale)
Low
High

Figure 2. Featured functionalities for exploring single-cell transcriptome datasets. (A) The "Overview" module provides UMAP plots showing cell type and sample identity, bar plots showing cell number and proportion of each cell type, and tables showing differentially expressed genes for each single-cell dataset. The single-cell RNA-seq data of human adult testis (GSE112013) was shown as an example.

(B) The "Query gene expression" module enables visualization of the expression levels of each selected gene with UMAP plot (left) and violin plot (right). The expression level of a germ-cell specific ribosomal gene *RPL39L* in human adult testis was shown.

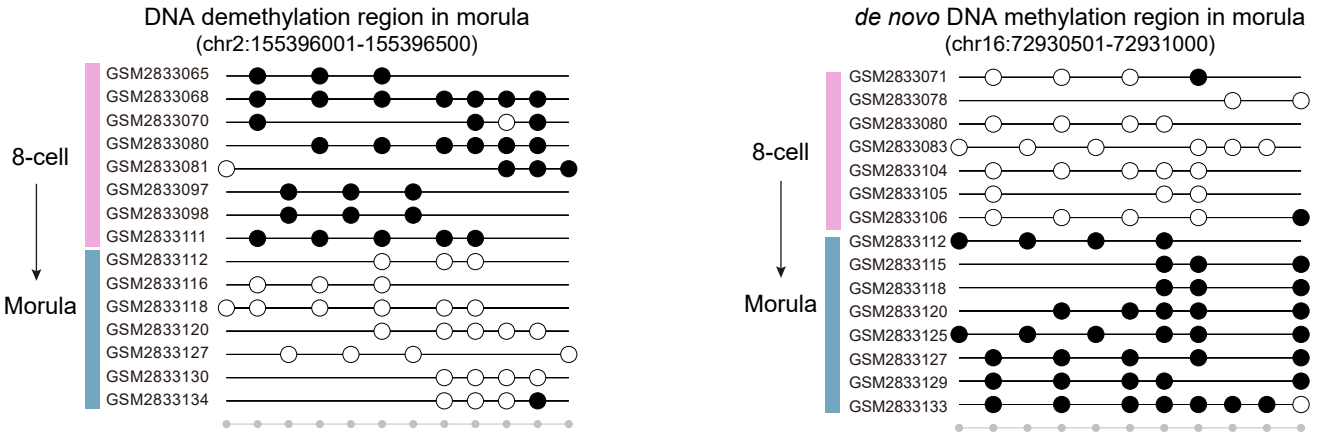
(C) The "DIY expression plot" module allows users to create customized plots displaying the expression levels of selected genes in selected cell types. Three types of plots (dot plot, violin plot and heatmap) are provided. The expression levels of representative marker genes and four ribosomal genes in major cell types of human adult testis were shown.

The description of cell type abbreviation are listed below. SSC, spermatogonia stem cell. SPG, spermatogonia. L, leptotene spermatocyte. Z, zygotene spermatocyte. P, pachytene spermatocyte. D, diplotene spermatocyte. SPC7, a mixture of diakinesis, metaphase, anaphase, telophase, and secondary spermatocytes.

Figure 3

A

DNA methylation plot

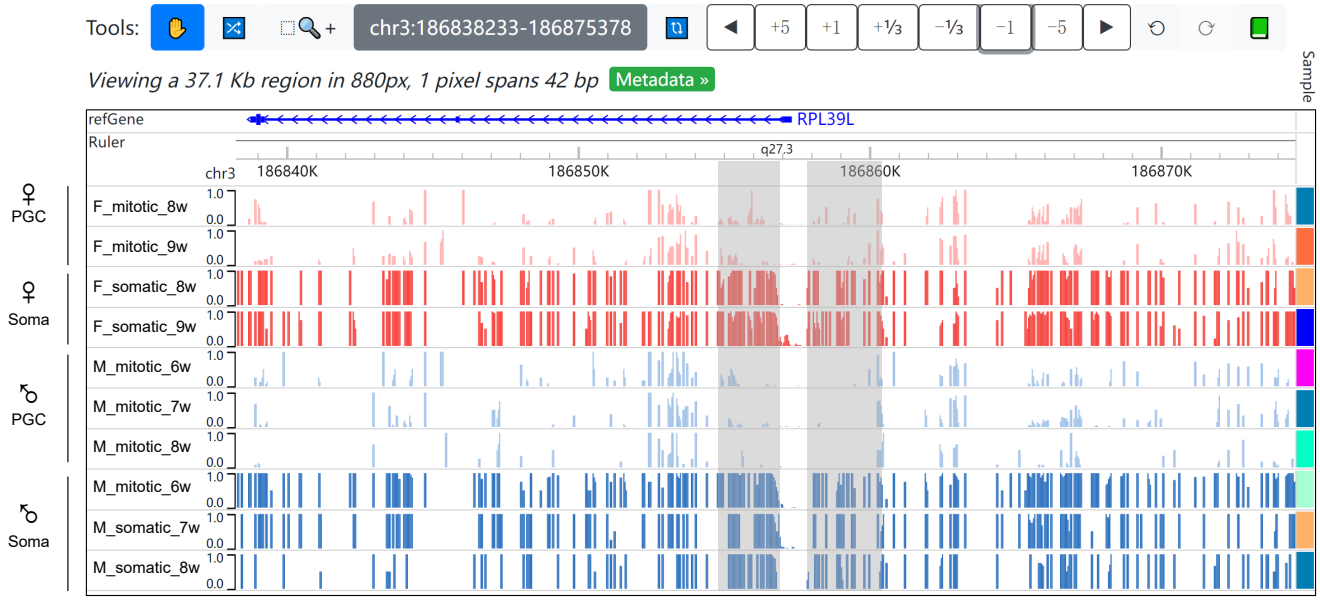


B

Epigenome browser

- DNA methylation

PGC and soma (before 11 weeks)



- Chromatin accessibility

PGC and soma (before 11 weeks)

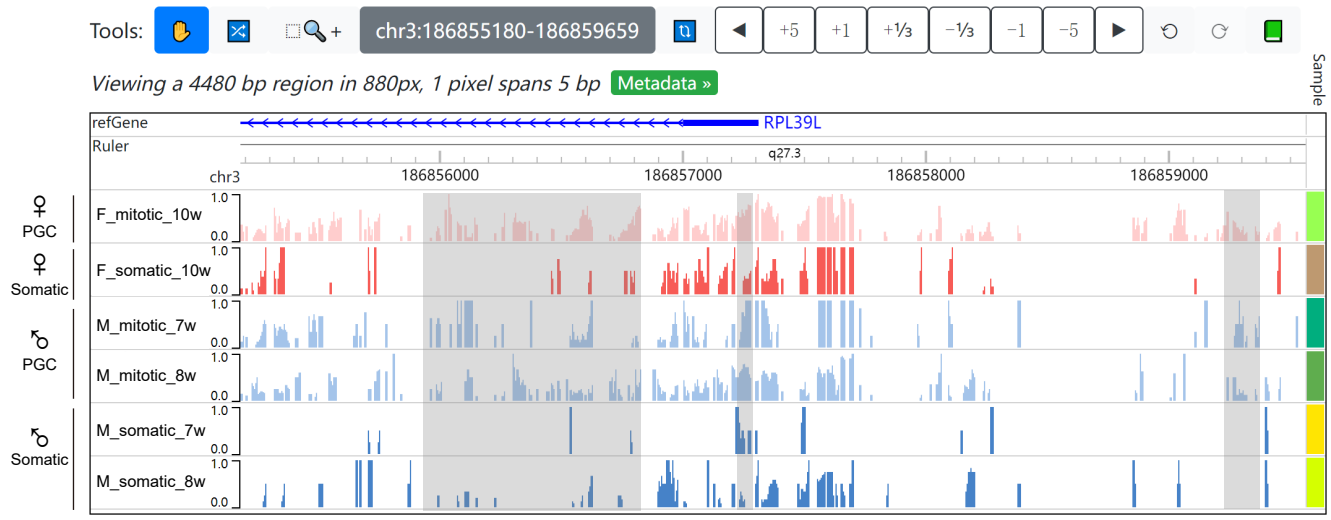


Figure 3. Featured functionalities for exploring single-cell epigenome datasets.

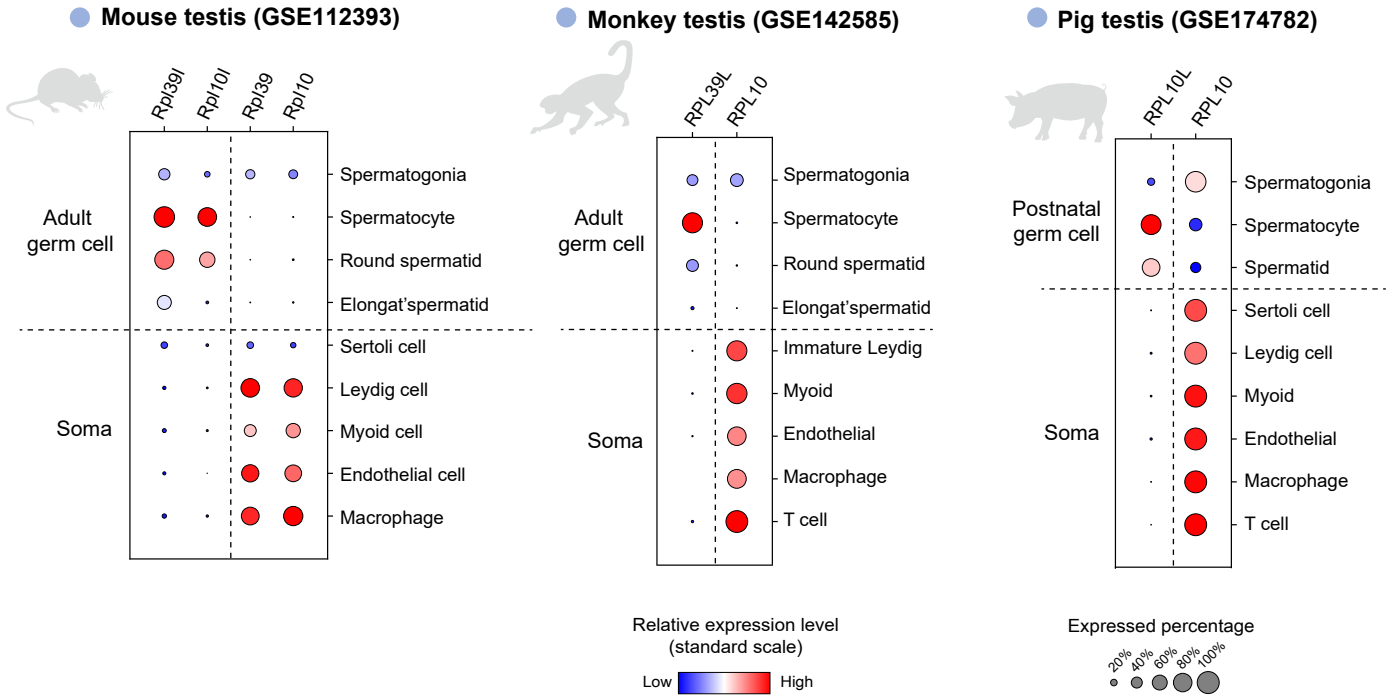
(A) The representative DNA demethylation and *de novo* methylation regions between human 8-cell embryo and morula were shown. Each row in the "Tanghulu" plot corresponds to an individual cell, and each circle represents a CpG site. Methylated CpG sites are represented by black dots, while unmethylated CpG sites are indicated by white dots. The grey dots at the bottom indicate covered CpG sites in selected cells. The CpG sites which were not covered by any cell were omitted. The single-cell DNA methylation data of human early embryo (GSE100272) was used.

(B) The screenshots of epigenome browser displaying the DNA methylation levels and chromatin accessibility levels around the promoter region of *RPL39L* in primordial germ cells and somatic cells. The grey rectangles highlight regions showing significant differences between somatic niche cells and germline cells. PGC, primordial germ cell.

Figure 4

A

Cross-species conservation of model animals



B

Cross-stage analysis

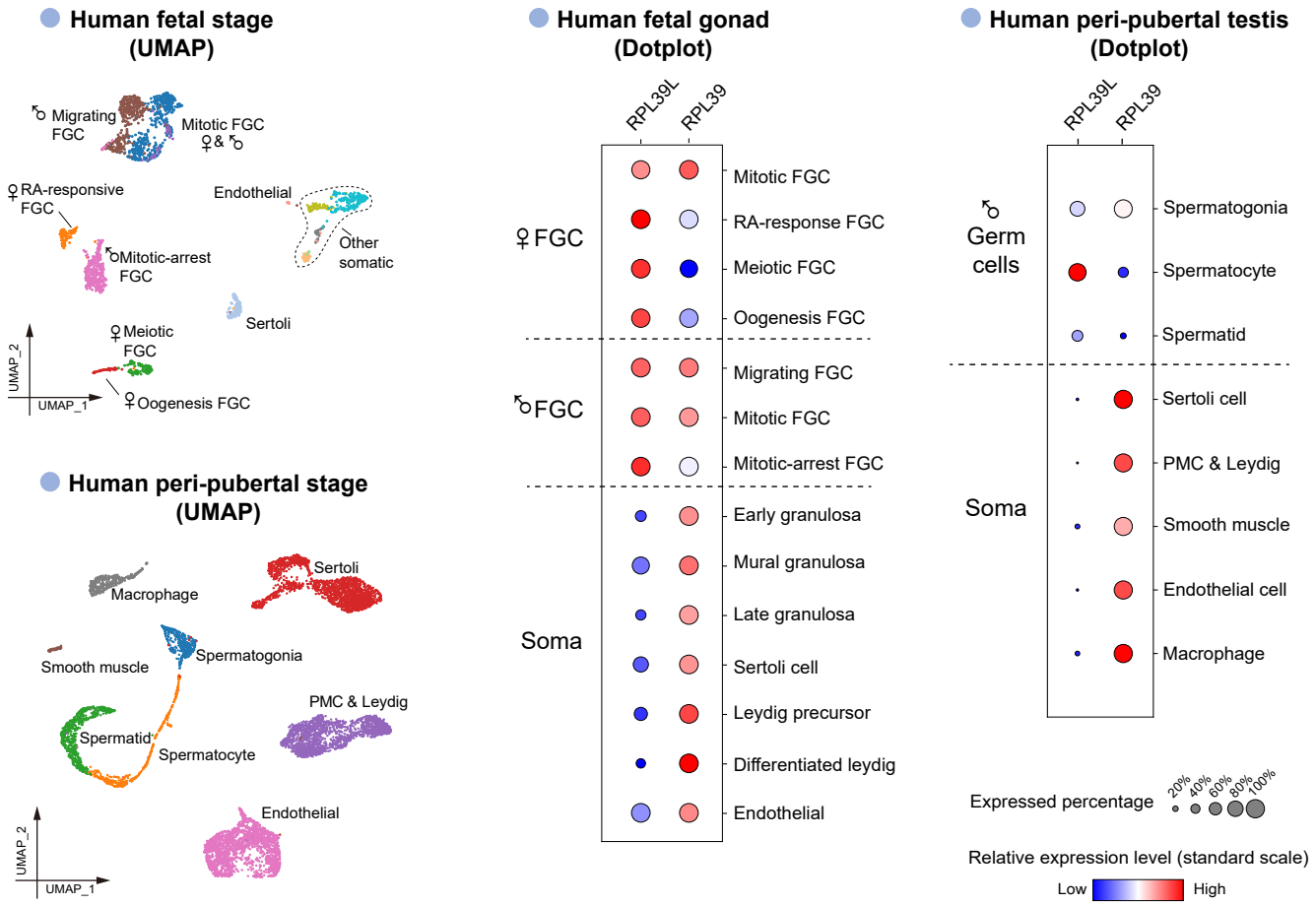


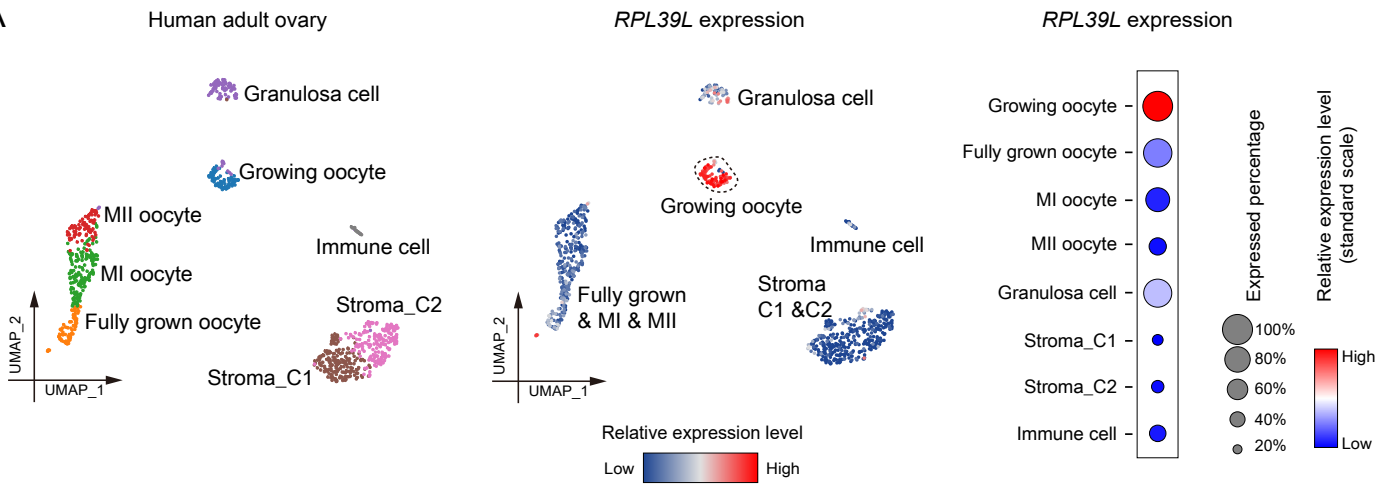
Figure 4. Cross-species and cross-stage analyses of ribosomal genes using SMARTdb.

(A) The dot plots show the expression levels of four ribosomal genes in mouse, cynomolgus monkey, and pig. PMC, peritubular myoid cell. IP, interstitial progenitor cell.

(B) The UMAP plots show the cell type identities of each cell cluster. The expression levels of two ribosomal genes (*RPL39L* and *RPL39*) in human germline and somatic niche cells of fetal stage and peri-pubertal stage were shown in dot plots. FGC, fetal germ cell. PMC, peritubular myoid cell.

Figure 5

A



B

- DNA methylation

- Chromatin accessibility

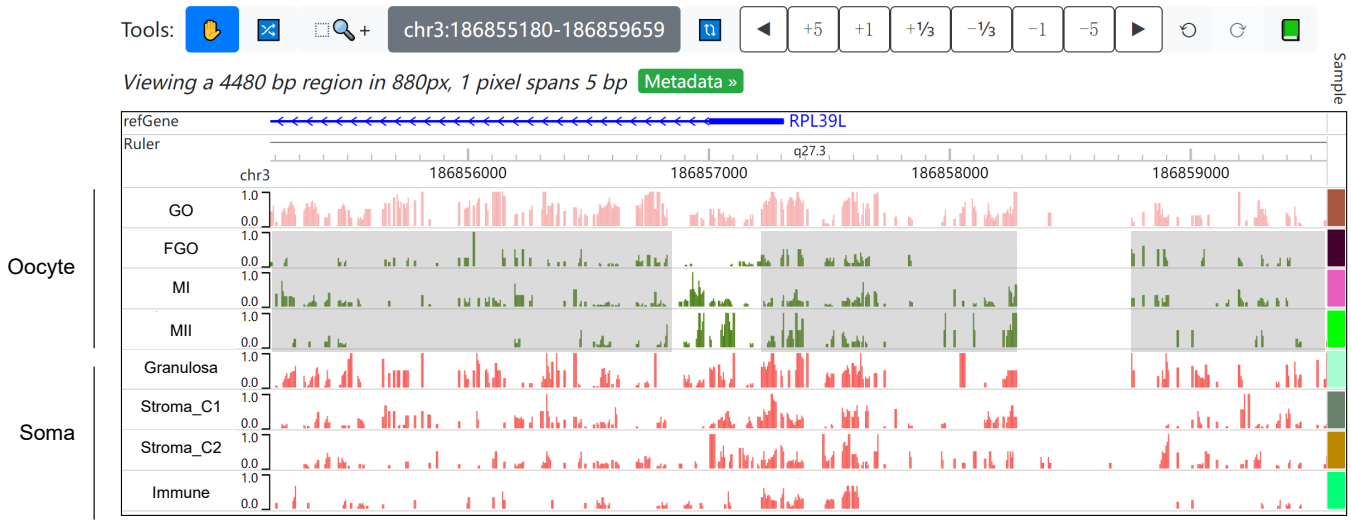


Figure 5. The gene expression, DNA methylation and chromatin accessibility patterns of *RPL39L* in human ovary.

(A) The leftmost UMAP plot shows the cell type identities of human adult ovary. The expression levels of *RPL39L* in various cell types of human adult ovary are shown on the middle UMAP plot and the rightmost dot plot.

(B) The screenshots of epigenome browser displaying the DNA methylation levels and chromatin accessibility levels around the promoter region of *RPL39L* in human adult ovary. The grey rectangles highlight regions showing changes in fully grown, MI, and MII oocytes compared with growing oocytes. GO, growing oocyte. FGO, fully grown oocyte. MI, metaphase I. MII, metaphase II.

Figure S1

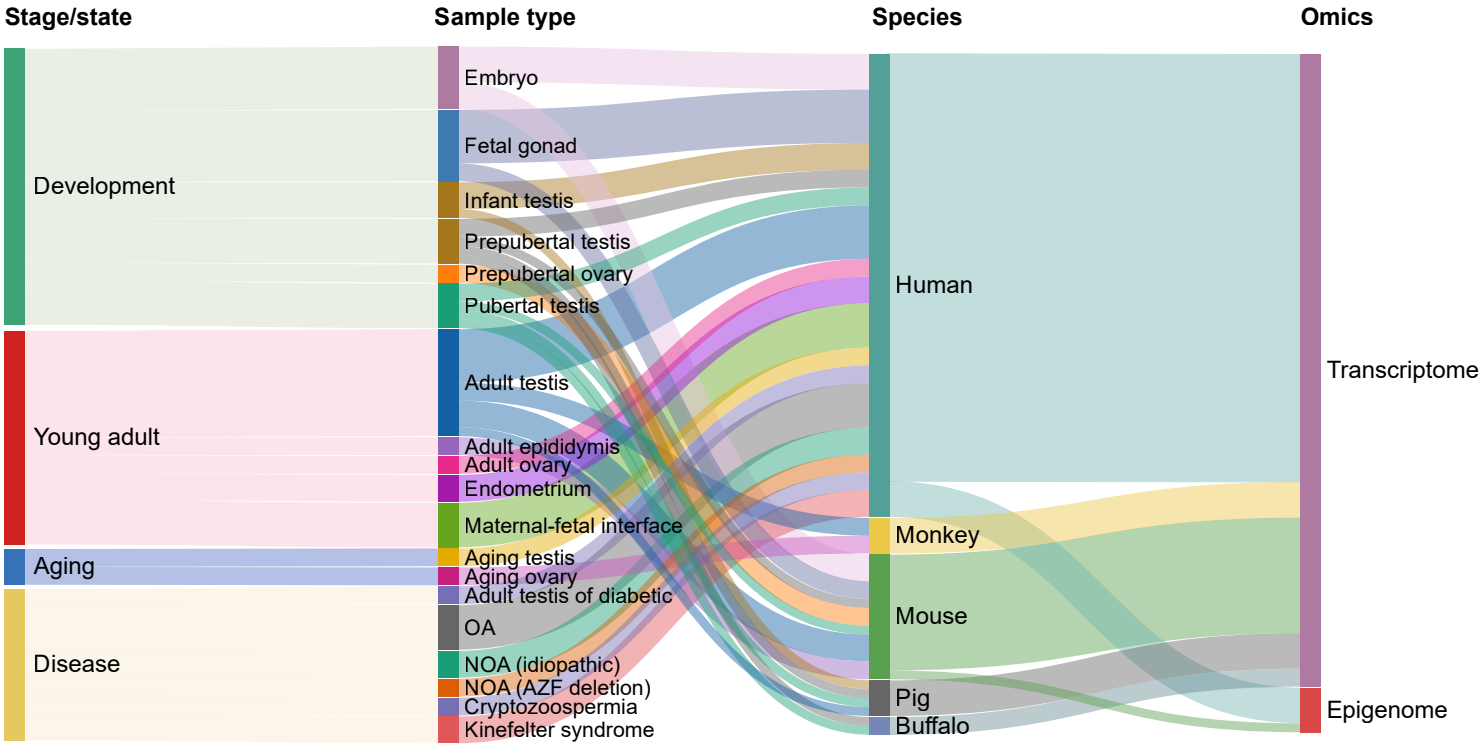

Stage/state
Sample type
Species
Omics
Development
Young adult
Aging
Disease
Embryo
Fetal gonad
Infant testis
Prepubertal testis
Prepubertal ovary
Pubertal testis
Adult testis
Adult epididymis
Adult ovary
Endometrium
Maternal-fetal interface
Aging testis
Aging ovary
Adult testis of diabetic
OA
NOA (idiopathic)
NOA (AZF deletion)
Cryptozoospermia
Kinefelter syndrome
Human
Monkey
Mouse
Pig
Buffalo
Transcriptome
Epigenome

Figure S1. The Sankey plot shows the composition of datasets included in SMARTdb. OA, obstructive azoospermia. NOA, Non-obstructive azoospermia.

Figure S2

- DNA methylation

- Chromatin accessibility

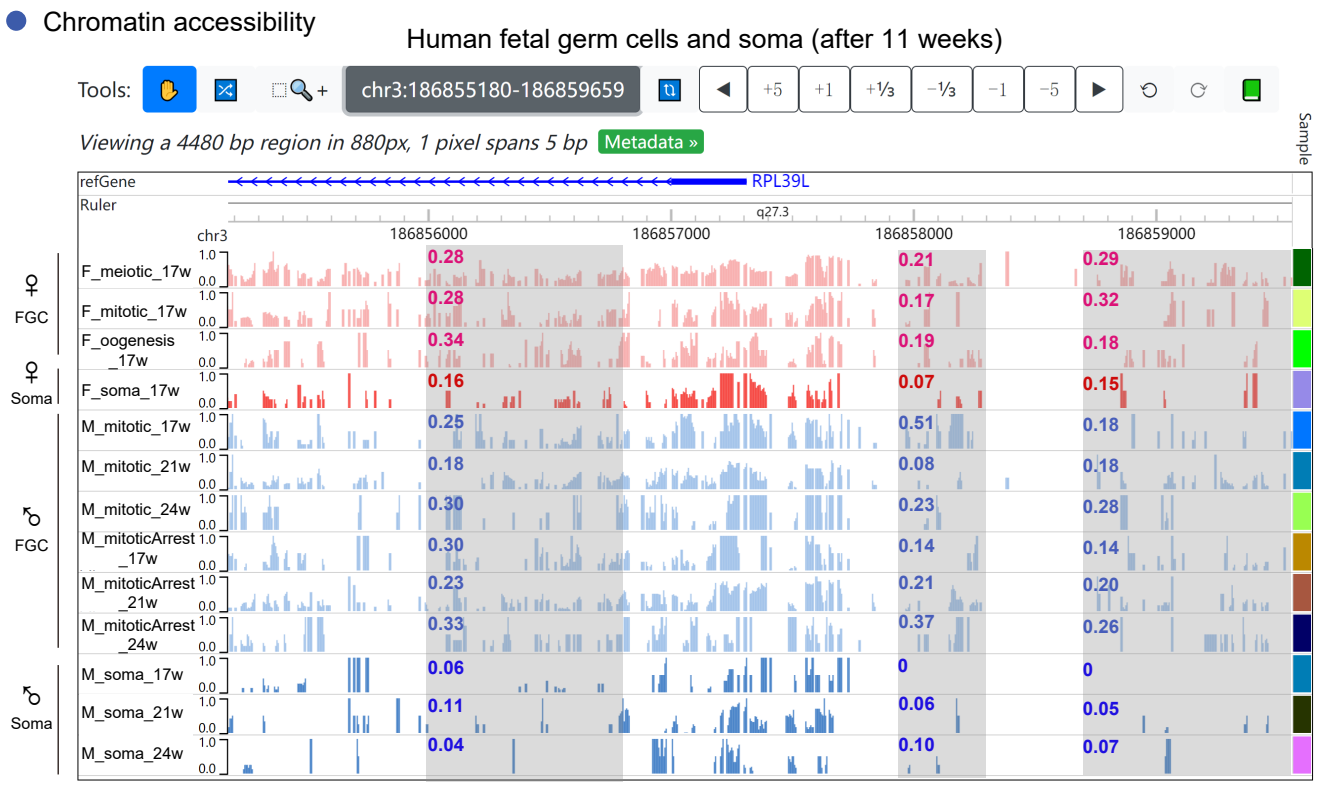


Figure S2. The DNA methylation levels and chromatin accessibility levels around the *RPL39L* promoter in fetal germ cells after 11 weeks.

The screenshots of epigenome browser displaying the DNA methylation levels and chromatin accessibility levels around the *RPL39L* promoter in fetal germ cells after 11 weeks post fertilization. The grey rectangles highlight regions showing significant differences between germ cells and somatic niche cells. The mean chromatin accessibility levels of the regions which are highlighted by grey rectangles are labelled.