Briefings in Bioinformatics, 2025, **26(6)**, bbaf586
https://doi.org/10.1093/bib/bbaf586
Problem Solving Protocol

ProjectSVR: mapping single-cell RNA-seq data to reference atlases by supported vector regression

Jianing Gao[1,2,3,‡], Jinman Fang[1,2,‡], Qizhi Zhu[1,4], Guoshu Li[2], Ziran Bi[2], Yue Hu[2], Bo Hong[1,2], Yuanwei Zhang[1,2,*], Shipeng Guo[5,*], Hongzhi Wang[1,2,*]

[1]Science Island Branch of Graduate School, University of Science and Technology of China, 350 Shushanhu Road, Shushan District, Hefei, Anhui 230031, China
[2]Anhui Province Key Laboratory of Medical Physics and Technology, Hefei Cancer Hospital of CAS, Institute of Health and Medical Technology, Hefei Institutes of Physical Science, Chinese Academy of Sciences (CAS), 350 Shushanhu Road, Shushan District, Hefei, Anhui 230031, China
[3]Westlake Laboratory of Life Sciences and Biomedicine, School of Life Sciences, Westlake University, 18 Shilongshan Road, Xihu District, Hangzhou, Zhejiang 310024, China
[4]Department of Oncology, The Second Affiliated Hospital of Anhui Medical University, 678 Furong Road, Shushan District, Hefei, Anhui 230032, China
[5]Department of Breast and Thyroid Surgery, The First Affiliated Hospital of Chongqing Medical University, 1 Youyi Road, Yuzhong District, Chongqing 400016, China
[*]Corresponding authors. Yuanwei Zhang, Institute of Health and Medical Technology, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, Anhui, China. E-mail: zyuanwei@cmpt.ac.cn; Shipeng Guo, Department of Breast, and Thyroid Surgery, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China. E-mail: guoshipeng2008@126.com; Hongzhi Wang, Institute of Health and Medical Technology, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, Anhui, China. Fax: 86-551-65591270; E-mail: wanghz@hfcas.ac.cn
[‡]Jianing Gao and Jinman Fang contributed equally to this work.

Abstract

Mapping the query cells onto a well-constructed reference atlas, known as reference mapping, enables robust, reproducible interpretation of new single-cell RNA-seq data in the context of curated and annotated cell subtypes and states. However, existing methods often rely on complex integration frameworks or require re-access to raw data, limiting their applicability and reproducibility. To address this, we introduce ProjectSVR, a machine learning-based framework that formulates reference mapping as a multi-target regression task. By leveraging ensemble support vector regression (SVR) to learn the relationship between gene set activity scores and low-dimensional reference embeddings, ProjectSVR enables platform-agnostic and integration-independent mapping. Benchmarking across diverse biological contexts—including immune responses, developmental trajectories, and disease states—demonstrates that ProjectSVR achieves comparable accuracy and robustness to state-of-the-art methods, with reduced dependence on data-specific preprocessing. Our findings demonstrate that ProjectSVR is a valuable tool for reference mapping, considerably simplifying the analysis of scRNA-seq data when well-constructed reference atlases are available.

Keywords: scRNA-seq; reference mapping; cell atlas; reproducible data analysis

Introduction

Advances in single-cell RNA sequencing (scRNA-seq) have revolutionized the study of cellular heterogeneity in development and disease. As the technology becomes more accessible and affordable, vast scRNA-seq datasets have accumulated, creating challenges in reproducibility and interpretation. Most current analyses adopt unsupervised pipelines—normalization, dimension reduction, clustering, and manual cell-type annotation—which are often subjective and time-consuming due to the lack of standardized references. Recently developed integration algorithms have enabled the construction of organ-scale atlases [1], offering biological context for consistent annotation [2–4], yet mapping novel data onto these references remains difficult.

Recently, several supervised methods have been developed to address the issues above. These methods are primarily based on two main strategies: classification and projection [5]. While the classification-based strategy has already demonstrated good performance in discrete cell type prediction and reached a certain level of maturity [6–8], it still has some limitations. These limitations include: (i) classification models cannot be used for dealing with continuous cell states; (ii) models must be retrained when the reference cell labels are changed; and (iii) it is challenging to balance the model performance and granularity of predicted labels, considering the recently published atlases using a multi-layer annotation system [9]. Projection-based methods, also known as reference mapping were conceived to remedy the limitations of classification methods. Since dimension reduction (such as Uniform Manifold Approximation and Projection) embeddings reflect the nuanced phenotypes of diverse cells within a reference cell atlas [10], projecting query cells onto these reference embeddings facilitates an unbiased comparison of data from different groups by inspecting the 2D density plots of projected cells, akin to in silico flow cytometry, without regard for the granularity of cell types [5]. Recently, several reference mapping

Received: June 15, 2025. **Revised:** August 21, 2025. **Accepted:** September 6, 2025

algorithms have been developed under the frame of corresponding integration methods [11–15]. However, dependency on particular integration methodologies restricts their applicability. Besides, the published reference cell atlas often adopted a data-tailored integration strategy, exampled by the recently published pan-cancer tumor-infiltrating T cell landscape [16]. The requirement for such tailored strategies hinders the application of existing reference mapping methods.

To overcome these limitations, we proposed ProjectSVR, a strategy that directly predicts the dimension reduction embeddings from the gene signature score matrix using the supported vector regression (SVR) model. ProjectSVR sets itself apart from existing methodologies by severing the dependency of the reference mapping process on any specific integration technique. Through benchmarking and application to immunological and reproductive datasets—including the construction of a mouse testicular cell atlas (mTCA)—we demonstrate that ProjectSVR offers an accurate, extensible, and integration-agnostic framework for reference mapping in scRNA-seq analysis.

Methods

Data collection and pre-processing for reference atlas

All data used in this manuscript were available publicly (Table S1).

The DISCO blood atlas (PBMC)

We downloaded the raw count matrix of the blood atlas (version 1.0) from the DISCO database. This dataset consists of 167,594 cells of PBMC from 100 samples of 30 studies. These cells were annotated into 14 major cell types and 24 fine-grained cell subtypes. The Uniform Manifold Approximation and Projection (UMAP) embeddings were generated using the scVI algorithm across different studies using the top 2000 highly variable genes with the default parameters. Details of other reference atlas, including MFI, Tumor-infiltrated T-cell landscape, and *mTCA* are provided in Supplementary material.

Data collection and pre-processing for query dataset

PBMCs from three 10× protocols

The PBMCs query datasets were each sequenced with different 10x protocols: 3′v1 (n = 4809 cells), 3′v2 (n = 8380 cells), and 5′ (n = 7697 cells). The raw count matrix and cell labels were obtained from GitHub (https://github.com/immunogenomics/harmony2019/). Details of other query datasets, including decidual immune cells, breast cancer and melanoma-infiltrating T cells, as well as germ cells from *Zfp541*/*Ythdc2* knock-out and NRRA induction experiments, are provided in Supplementary material.

Benchmark the reference mapping algorithms

To benchmark ProjectSVR, we used DISCO, MFI, and mTCA as reference datasets and selected batch-defined subsets as queries: GSE175499 for PBMC, FCA7196224, FCA7196225, and FCA7511884 for MFI, and GSM2928505, do17825, and GSM3744444 for Mtca. Accuracy and ARI were computed after label transfer via k-NN (k = 10). ProjectSVR was compared with Harmony, ProjecTILs, Seurat Anchored-rPCA, iNMF, scArches, and SCALEX (Table 1). Detailed parameters and scripts are available on GitHub (https://github.com/JarningGau/ProjectSVR-benchmark/).

Reference mapping by ensemble-supported vector regression

ProjectSVR framework

ProjectSVR is a computational framework to perform reference mapping—projecting new data into a reference atlas without altering the reference space—from the gene expression matrix of scRNA-seq data. ProjectSVR comprises two steps: model building and reference mapping (Fig. 1). Briefly, ProjectSVR takes the gene set score matrix generated by a gene set scoring method (e.g. AUCell) and reference cell embeddings from a non-linear dimension reduction technique (e.g. UMAP) as input to train a regression model from gene set scores to cell embeddings based on supported vector regression. Then the gene expression matrix of query data is transformed into a gene set score matrix and the cell embeddings in reference space were inferred by a pre-trained SVR model.

Identification of gene signatures relevant to cell identity

We first identify a set of co-expressed genes representing the cell states to calculate the gene set score matrix. When cell cluster information is available, the cluster-specific markers are selected using the Wilcox test via a 'one vs. rest' strategy. When there is no cell cluster information or when dealing with continuous cell states, the consensus non-negative matrix factorization (cNMF) was adopted on the logarithmic normalized gene expression matrix. Then the highly ranked genes of each latent factor are selected as co-expressed gene sets (Fig. 1A). The advantages of using gene set score as features rather than gene expression level are: (i) gene set score is more robust than gene expression across different batches (Fig. S1). (ii) Gene set score acts like a 'meta gene' that can recover the technical zeros (also known as 'drop-outs') of gene expression levels in single cells. (iii) The gene set score matrix is much smaller than the gene expression matrix in the number of features which significantly reduces the runtime during the model training and inference, considering the time-consuming is linearly related to model training and prediction of the SVR model. (iv) When adopting the AUCell algorithm, the gene set score is only relevant to the gene ranking in the individual cells rather than the absolute gene expression level, which makes it flexible to map query data with different preprocessing steps onto reference embeddings.

Ensemble-SVR model

We define the reference mapping task in ProjectSVR as predicting the reference embedding coordinates of each cell from its gene set score vector. We compute the gene set score matrix $\boldsymbol{X} \in \mathbb{R}^{Nc\times Nf}$ from the gene expression matrix $\mathbf{E} \in \mathbb{R}^{Nc\times Ng}$ using AUCell algorithm and the predefined gene sets described above, where Nc is the number of cells, Nf is the number of gene sets, and Ng is the number of genes. The resulting $\boldsymbol{X}$ and the reference embeddings $\mathbf{Y} \in \mathbb{R}^{Nc\times Nd}$, are then used to train the SVR models for mapping, where Nd is the embedding dimensionality.

$$f\left(\mathrm{x_i}\right) = \boldsymbol{w} \bullet \phi\left(\mathrm{x_i}\right) + \boldsymbol{b}$$

The objective function consists of the L2 regularization and the ϵ-insensitive loss:

$$\min_{\boldsymbol{w},b,\xi,\xi^*}\left\{\frac{1}{2}|\boldsymbol{w}|^2 + \mathrm{C}\sum_{\mathrm{i}=1}^{\mathrm{Nc}}\left(\xi_{\mathrm{i}} + \xi_{\mathrm{i}}^*\right)\right\}$$

Table 1. Overview of reference mapping methods.

Method	Methodology	Programming Language	Reference model data size (~0.1 M cells)	Projection Quality Metrics	Fixed Reference Coordinates	Independent to Specific Integration Method	Coupled Integration Algorithm
ProjectSVR	SVR	R	~10 Mb	√	√	√	No
Symphony	Batch alignment	R	~100 Mb	√	√		Harmony
scArches	VAE	Python	~100 Mb	√			VAE-based model
SCALEX	VAE	Python	~1 Gb				VAE
Seurat-RM	Batch alignment	R	~1 Gb		√		Anchored-CCA/rPCA
ProjectTILs	Batch alignment	R	~1 Gb		√		STACAS
online iNMF	NMF	R	~1 Gb				Liger

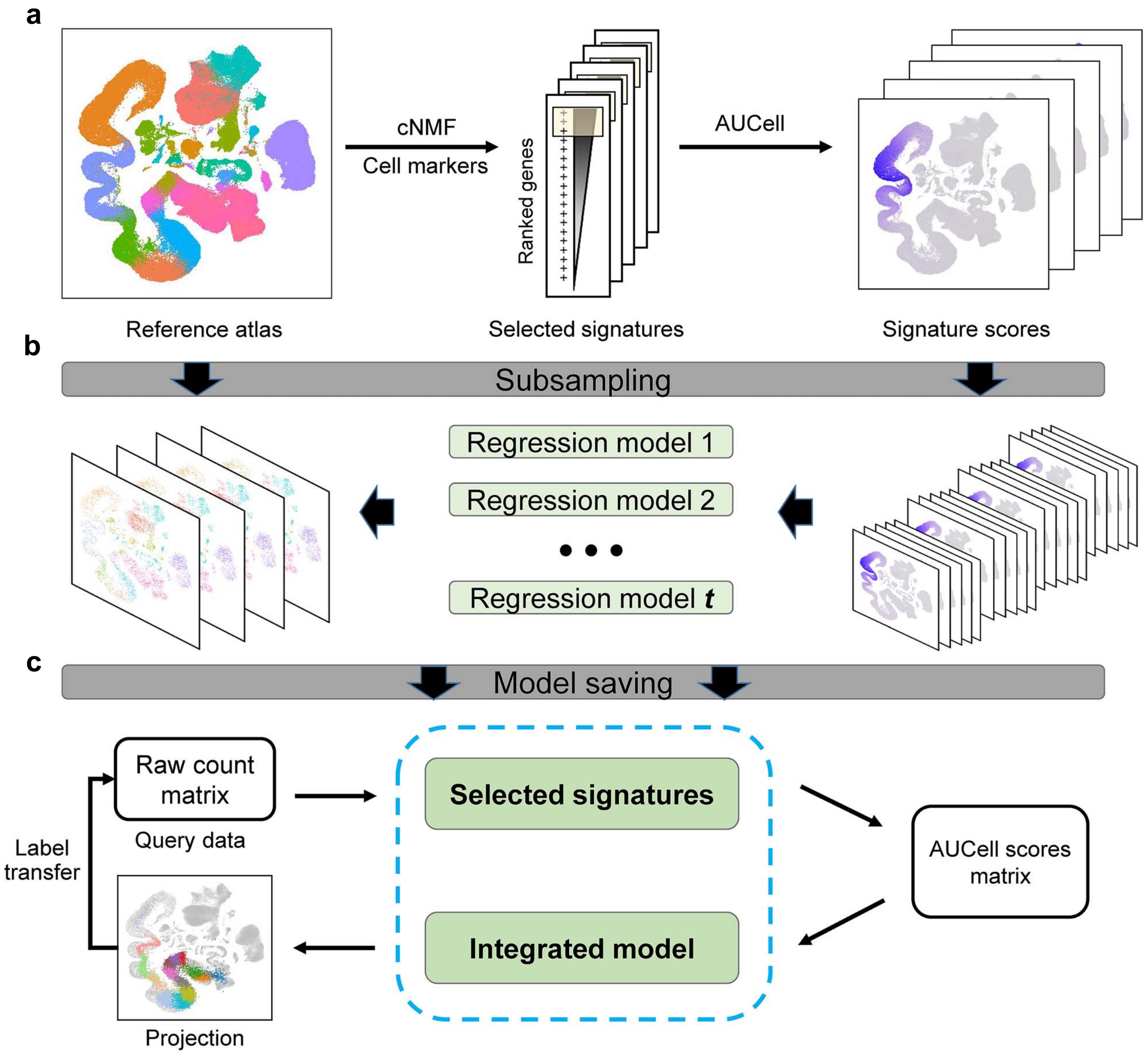

Figure 1. Workflow of ProjectSVR. (a) The ProjectSVR takes integrated embeddings (e.g. UMAP) and gene set scores calculated via the UCell algorithm as inputs. (b) Then a regression model is fitted by supported vector regression (SVR). (c) For mapping, the query count matrix is transformed into signature scores, and the trained models predict query embeddings. Final predictions are aggregated via median. Cell type labels are transferred using a k-nearest neighbors (KNN) classifier.

With the following constraints:

$$y_i - \boldsymbol{w} \bullet \phi(\mathbf{x}_i) - b \le \epsilon + \xi_i$$

$$\boldsymbol{w} \bullet \phi(\mathbf{x}_i) + b - y_i \le \epsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \ge 0$$

Here $\mathbf{x}_i \in \mathbb{R}^{Nf}$represents the gene set scores of the i^{th} cell, $y_i \in \mathbb{R}$ represents the reference embedding of a given dimension of the i^{th} cell. The SVR model is parameterized by a set of weights and biases, denoted by $\boldsymbol{w}$ and b. The hyperparameters C, ϵ, and kernel function ϕ can be tuned to yield sensible results. Here we use the default parameters setting in the *svm* function implemented in R package e1071 (C = 1, ϵ = 0.1, and ϕ is radial kernel). For each dimension of reference embeddings, we build a separate SVR model for it. Considering that the worst complexity of SVR training process is $O(Nc^2 \bullet Nf)$ and reference atlas always contains hundreds of thousands of cells, we introduced an ensemble SVR model to remit the time-consuming (Fig. 1b). Specifically, we randomly subsample *m* cells from the input gene set score matrix *X* for *t* times and train *t* SVR models for predicting each dimension of reference embeddings. The final predicted embeddings are the median value of predictions from the *t* models. Based on robustness analysis, we set *m* = 8000 and *t* = 10 by default.

Robust analysis

The subsampling procedure introduces two hyperparameters: the sample times *t* and sample size *m*. We performed a systematic robustness analysis on our mouse testicular cell atlas (mTCA) dataset. We find that the root mean square error of the predicted reference embeddings is reduced for a larger *m* and t. A larger sample size *m* and larger sampling times *t* benefits ProjectSVR's performance. However, when sampling times *t* exceed 10, more sampling times will not much improve ProjectSVR's performance when the sampling size is less than 8000. Overall, the performance of ProjectSVR is robust to these two hyperparameters in a wide range (Fig. S2). To balance the calculation efficiency and accuracy, the default values of *m* and *t* were set to 8000 and 10.

Reference mapping

Given query data, ProjectSVR first transforms its gene expression matrix into a gene set score matrix by the AUCell algorithm using the pre-defined gene signatures during the model training step. Then the cell embeddings in reference dimension reduction space were inferred by the pre-trained ensemble-SVR model (Fig. 1c). The running time of prediction using the SVR model is $O(Nc^2 \bullet Nf)$, where *Nc* representing the number of cells and *Nf* representing the number of features.

Metrics for mapping quality

To measure the mapping quantity of query cells, we defined the mean k-NN distance. The main idea is that a good mapping procedure should keep the local topological relationship after projection. We construct a k-NN network for query cells using the AUCell score matrix and calculate the average distance of the *K* neighbors in projected space for each cell. A larger mean k-NN distance for a query cell means a worse projection (Fig. S3a–c). To determine the cut-off of mean k-NN distance, the mean distance of *K* randomly selected cells was calculated and repeated $\boldsymbol{n}$ times to construct the null distribution. Then empirical p values were estimated using this null distribution to measure the significance of the mean k-NN distance is lower than the background (Fig. S3d–f). Then p values were corrected through the Benjamini-Hochberg Procedure. Users can filter the low-quality projected cells according both the mean k-NN distance or adjusted p value (Fig. S3g–i).

Label transfer

Once query cells are projected in the same embeddings as the reference, reference labels can transfer to query cells using any classification model. We use a simple k-NN classifier to make predictions via majority vote (Fig. 1c).

Results

Mapping the PBMC data across technologies to the harmonized reference space

To evaluate the feasibility of ProjectSVR for reference mapping, we designed a proof-of-concept task using a PBMC dataset. We first constructed a PBMC reference atlas by integrating annotated blood samples from the DISCO database [17] with the scVI algorithm [18] and visualized cells by UMAP using scVI-generated latent space (Fig. 1a). (Fig. 2a). Subsequently, we mapped 20,571 PBMCs obtained from three different 10x technologies (3'v1, 3'v2, and 5′) to this reference atlas.

To assess projection quality, we used UMAP visualization and the local inverse Simpson's Index (LISI) [19], which measures the diversity of cell labels in local neighborhoods. Without batch correction, query PBMCs clustered by platform rather than cell type (mean LISI = 1.06; Fig. 2b–d), indicating strong technical effects. ProjectSVR effectively corrected this, aligning cells by biological identity (mean LISI = 2.30), with performance comparable to leading reference mapping methods (Fig. 2d). Projected embeddings also distinguished known cell subtypes, suggesting biologically meaningful projections (Fig. 2c).

To further evaluate mapping quality, we calculated mean k-nearest neighbor (k-NN) distances in the reference space. Only 2.6% of query cells had a mean distance >1.5 and were defined as low-quality projections (Fig. S3g–i). These cells lay at cluster margins and showed elevated ambient RNA contamination (Fig. S3i, j), likely reflecting low cell identity signals. These were excluded before label transfer.

We used a k-NN classifier to transfer DISCO annotations to the query and compared predicted labels to those from Korsunsky et al. [19] (Fig. 2e). After harmonizing labels (Table S2), most subtypes, including CD16 NK, pDC, and naïve B cells, were predicted with high specificity (>0.95; Figs. 2F, S4a). Platelets and regulatory T cells showed low sensitivity (0.20 and 0.21). For instance, only 16.7% of megakaryocytes were labeled as platelets, with most misclassified as CD14 monocytes or T cells (Fig. S4b). Marker gene analysis confirmed these cells expressed CD14 and CD3D instead of PPBP (Fig. S4c), suggesting misannotations or cell-type mixing. A similar trend was observed for regulatory T cells (Fig. S4d, e). These findings highlight ProjectSVR's ability to resolve purer clusters, likely due to its biologically driven and independent projection framework. Moreover, ProjectSVR identified finer substructures—81.4% of effector CD8+ T cells were further divided into three subtypes based on marker expression (Fig. S4f, g).

ProjectSVR is comparable to the performance of the SOTA reference mapping methods

We benchmarked ProjectSVR against six state-of-the-art methods—Symphony [13], ProjecTILs [11], Seurat [20], scArches [14],

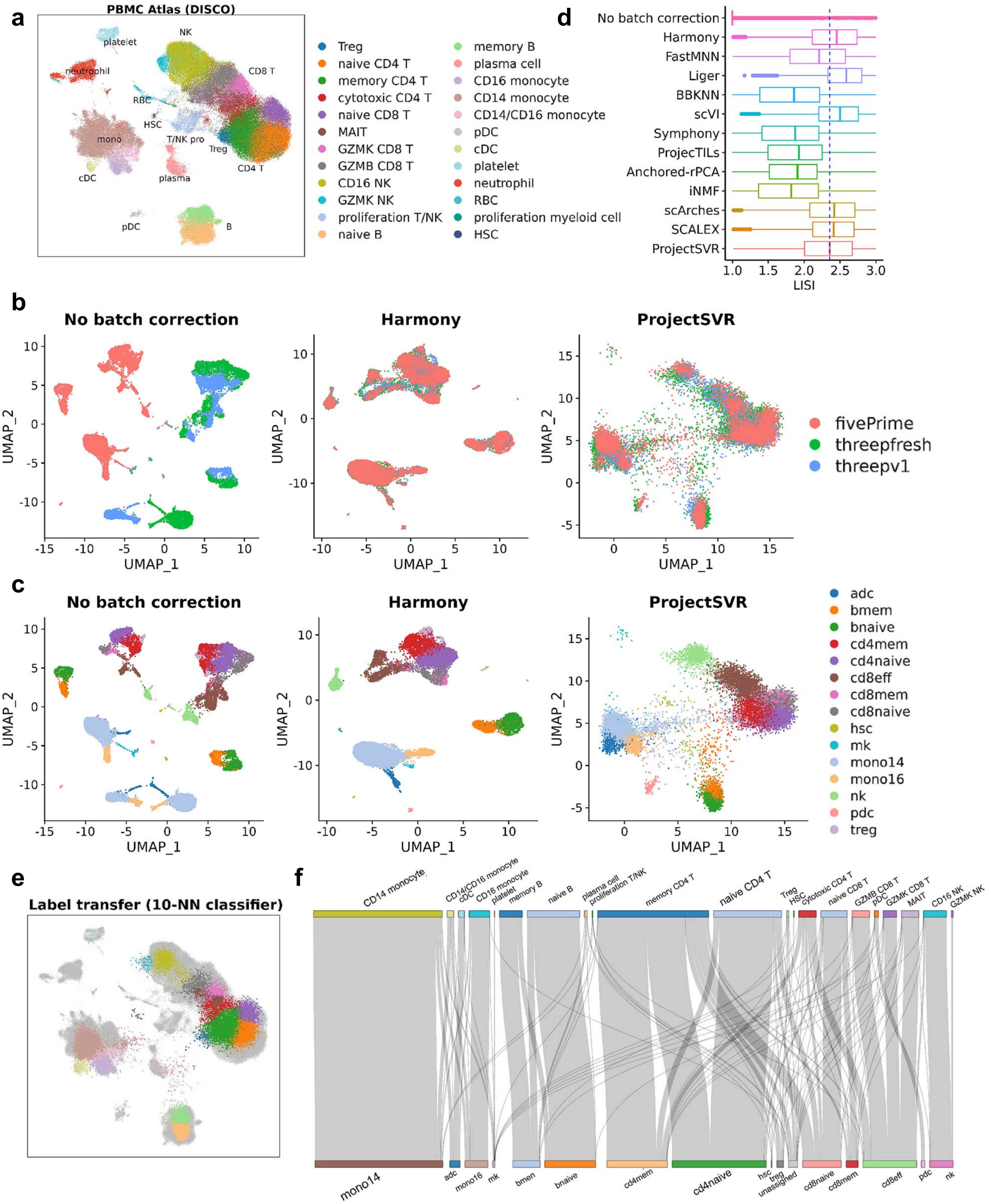

Figure 2. Mapping PBMC query data from different technologies to DISCO blood atlas. (a) Uniform manifold approximation and projection (UMAP) plots showing the blood atlas in DISCO database. Each dot represents a single cell, colored by cell type. (b, c). Projection of PBMC datasets from three 10x protocols onto DISCO blood atlas in (a) left, UMAP plots without batch correction; middle, UMAP plots by harmony; right, query cells projected by ProjectSVR. Dots were colored by protocols (b) and published cell types identified by Korsunsky et al. (d) Boxplot showing the distribution of LISI (local inverse Simpson's index) across different integration or reference mapping methods. (e) Transfer cell labels from blood atlas (reference) to query cells using a 10-NN classifier on the ProjectSVR embedding. Query cells were colored by cell type defined in (a), and grey dots represent reference cells. (f) Sankey diagram shows the consistency between predicted cell labels in (e) and published cell labels in (c).

Table 2. Benchmark results of ProjectSVR and other reference mapping algorithms.

Task	Method	Accuracy	ARI	Model building (seconds)	Reference mapping (seconds)
PBMC	De novo integration (scVI)	0.791	0.649	NA	NA
	ProjectSVR	0.733	0.585	**70.24**	21.35
	symphony	0.694	0.577	304.04	**3.52**
	projectTILs	0.487	0.412	4943.19	1106.34
	Seurat	0.753	0.600	1351.65	677.13
	scArches	**0.797**	**0.689**	373.53	156.81
	iNMF	0.640	0.532	100.20	3.53
	SCALEX	0.685	0.527	568.68	99.07
MFI	De novo integration (fastMNN)	0.895	0.848	NA	NA
	ProjectSVR	0.697	0.556	77.07	10.89
	symphony	0.626	0.492	85.97	**1.95**
	projectTILs	0.107	0.259	1454.18	326.04
	Seurat	0.740	**0.577**	88.46	862.62
	scArches	**0.739**	0.553	316.77	110.17
	iNMF	0.566	0.394	**74.90**	2.54
	SCALEX	0.523	0.401	493.47	61.11
mTCA	De novo integration (scANVI)	0.993	0.995	NA	NA
	ProjectSVR	**0.930**	0.952	663.98	83.82
	symphony	0.782	0.772	**261.56**	**3.26**
	projectTILs	0.690	0.750	4192.72	1081.52
	Seurat	0.916	0.949	960.07	1705.35
	scArches	0.895	**0.955**	329.19	135.79
	iNMF	0.726	0.671	200.96	5.26
	SCALEX	0.778	0.906	498.32	192.24

Note: ARI: Adjusted rand index, Bold values indicate results from the present method; italic values indicate the best result among all comparison methods.

iNMF [12], SCALEX [15]) on three different tasks (Table 1). A well-performing method should place query cells near reference cells of the same type. Thus, we used a k-NN classifier to assign labels and evaluated mapping quality via accuracy and adjusted Rand index (ARI). De novo integration results (scVI for PBMC/mTCA, fastMNN for MFI) served as gold standards. Across all datasets, ProjectSVR performed comparably to top methods, especially on the mTCA dataset, where it achieved accuracy and ARI above 0.92—close to the gold standard (Figs. 3a, S5, Table 2). In addition, ProjectSVR showed favorable time and memory efficiency relative to scArches and Seurat (Fig. 3b, c, Table 2). These results suggest that ProjectSVR is a competitive and scalable solution for reference mapping, particularly in settings where other tools may be impractical.

ProjectSVR facilitated the interpretation of the decidual immune microenvironment in RPL patients

Analyzing paired healthy and diseased samples is a common strategy in scRNA-seq studies to identify shifts in cellular composition and transcriptional states. A reliable reference mapping method should accurately assign query cells to a reference atlas, even in the presence of disease-related transcriptional perturbations. In this study, we applied ProjectSVR to project decidual immune cells from healthy controls and recurrent pregnancy loss (RPL) patients [21] onto a previously reported maternal-fetal interface atlas [2]. We first integrated the reference using fastMNN to correct for batch effects arising from different library protocols (Fig. 4a). Then, we trained an ensemble SVR model using ProjectSVR and projected query cells into the reference space. ProjectSVR successfully removed batch effects and performed comparably to state-of-the-art (SOTA) integration or mapping approaches (Fig. 4b, Fig. S6a–c). Query cells were annotated using marker genes from relevant studies [2, 21], which we treated as true labels (Fig. S6d). Using a k-NN classifier, we transferred reference labels to the query set and observed strong agreement with true labels (Fig. 4c). Notably, we detected a query-specific cluster (cluster 8) expressing heat shock protein (HSP) genes (Fig. S6e). These signatures likely originated from stress responses during tissue dissociation and storage, not from biological variation [22]. ProjectSVR effectively corrected this bias and enabled accurate identity prediction for affected cells (Fig. S6f, g). (Fig. S6f, g). To assess biological validity, we analyzed decidual NK (dNK) cells between healthy and RPL groups. Consistent with prior findings [21], n, we observed a decrease in the dNK1 population and an increase in dNK3 cells in RPL patients (Fig. 4e–f). This result underscores ProjectSVR's robustness in accurately projecting cells affected by both biological (disease) and non-biological (technical) transcriptome perturbations, facilitating reliable interpretation of immune alterations in disease-control study designs.

ProjectSVR enables accurate identification of the tumor-infiltrated T cell heterogeneity in cancer immunology study

The complexity and plasticity of T cells pose challenges for studying adaptive responses in the context of cancer [11]. A recently published pan-cancer T cell atlas [16] integrated 390,000 T cells from 316 patients across 21 cancer types into two meta-cell atlases, providing a valuable reference for defining tumor-infiltrated T cell states. However, due to complex integration techniques, this atlas is incompatible with most conventional reference mapping methods [11, 13, 14]. To address this, we used ProjectSVR to build reference models from the CD4 and CD8 T cell components of the atlas. Regression models were trained using gene set scores derived from cluster-specific genes provided by the original study [16]. We then projected a breast

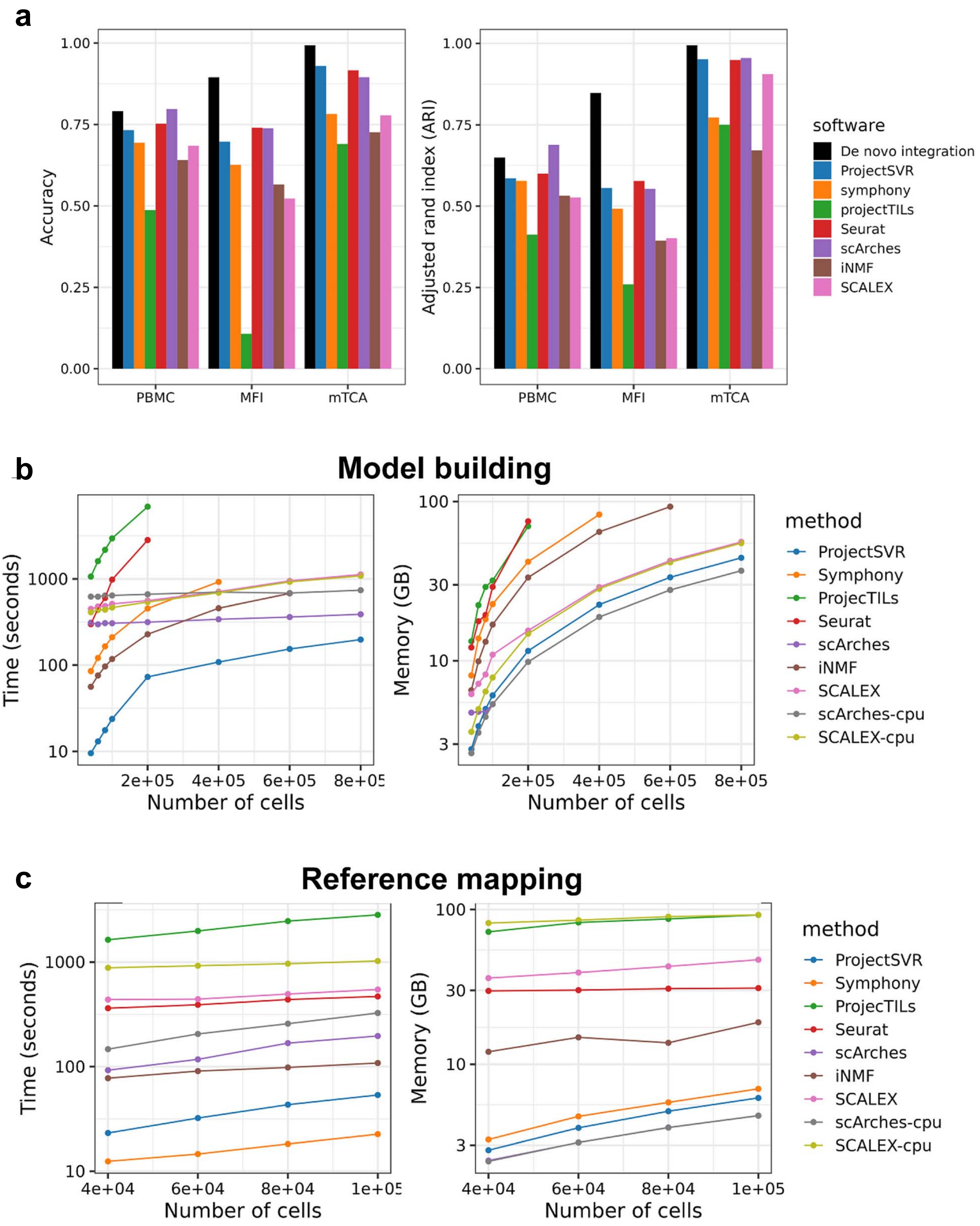

Figure 3. Benchmark the performance of ProjectSVR compared to other reference mapping methods. (a) Comparing the mapping accuracy of ProjectSVR and other reference mapping methods via projecting a query dataset to the large reference dataset. (b) Time and memory usage comparison between different reference mapping methods for building models on different-sized reference datasets. (c) Time and memory usage comparison between different reference mapping methods for mapping different-sized query data onto reference.

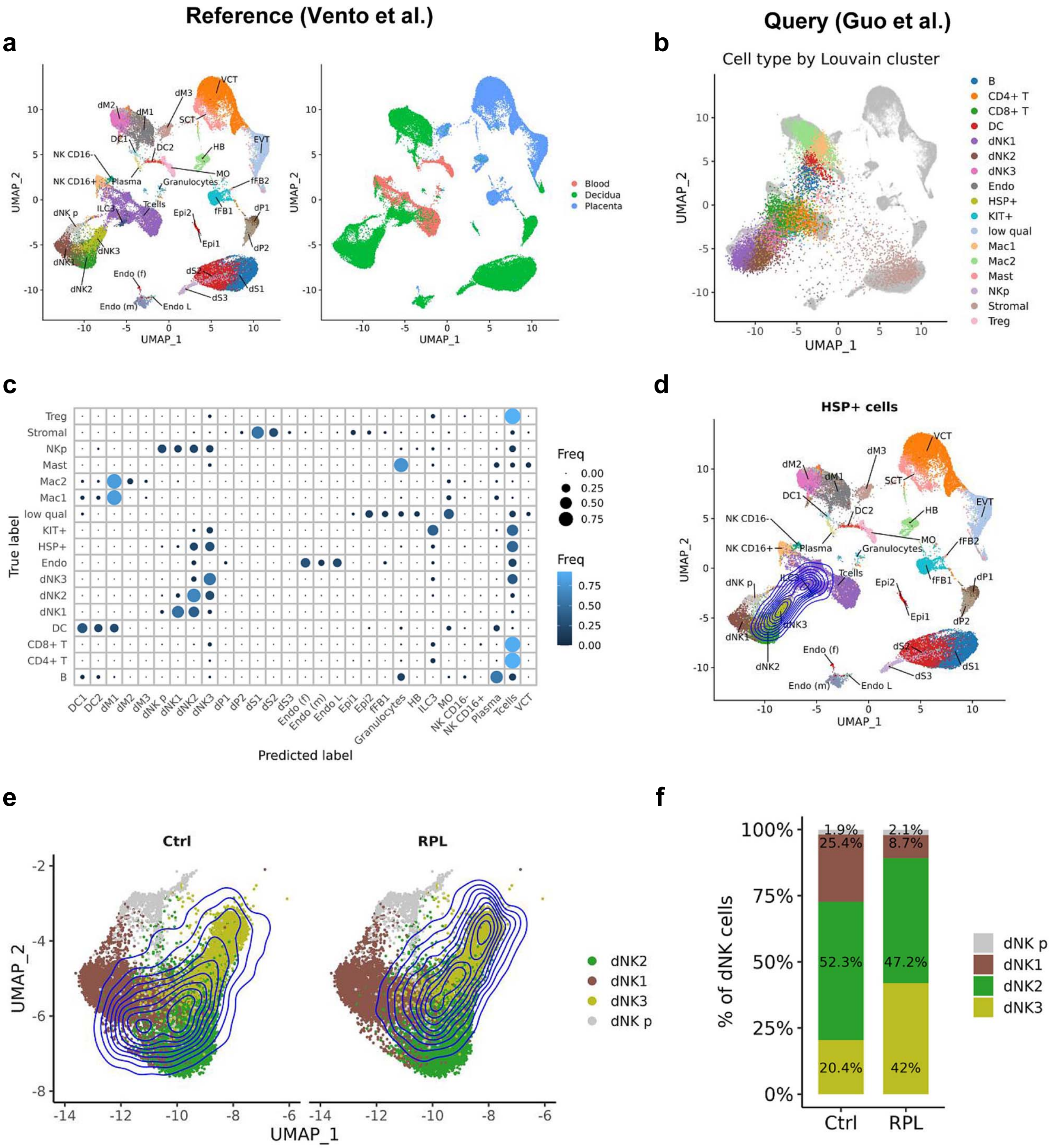

Figure 4. Projection analysis of decidual immune cells between recurrent pregnancy loss (RPL) patients and healthy controls. (a) UMAP of maternal-fetal interface atlas, colored by cell type and tissue source. (b) Projection of decidual cells (query) from RPL patients and controls onto reference. Grey = reference, colors = manually annotated query cells. (see also Fig. S6). (c) Dot plot comparing true cell labels (columns), defined in (b), to predicted cell labels (rows) via k-NN label transfer using ProjectSVR. The dot size represents the frequency of predicted labels in each true label (the sum of rows is normalized to 1). (d) 2D density plot showing the projection of cells highly expressing heat shock proteins (HSP+ cells) onto the reference. Each dot represents a reference cell, colored by reference cell type. (e) Density of the predicted decidual NK (dNK) cells from controls and RPL patients mapped to dNK reference. (f) The distribution of dNK cells from controls and RPL patients corresponding to panel (e).

cancer scRNA-seq dataset stratified by response to anti-PD1 therapy to examine differences in T cell composition. Consistent with the original findings [23], we observed an enrichment of T follicular helper (Tfh), T helper 1 (Th1), and regulatory T (Treg) cells in pre-treatment tumors from responders (Fig. 5a). Further analysis revealed that IFNG+ Tfh/Th1 (c17) and IL21+ Tfh (c16) were the most significantly enriched CD4 clusters in responders based on differential analysis (Fig. 5b, c). ProjectSVR projection followed by label transfer also identified elevated proportions of GZMK+ (c11) and terminally exhausted (c12) CD8 T cells in the responder group (Fig. 5d–f), consistent with immune activation signatures observed in effective responses to ICB therapy.

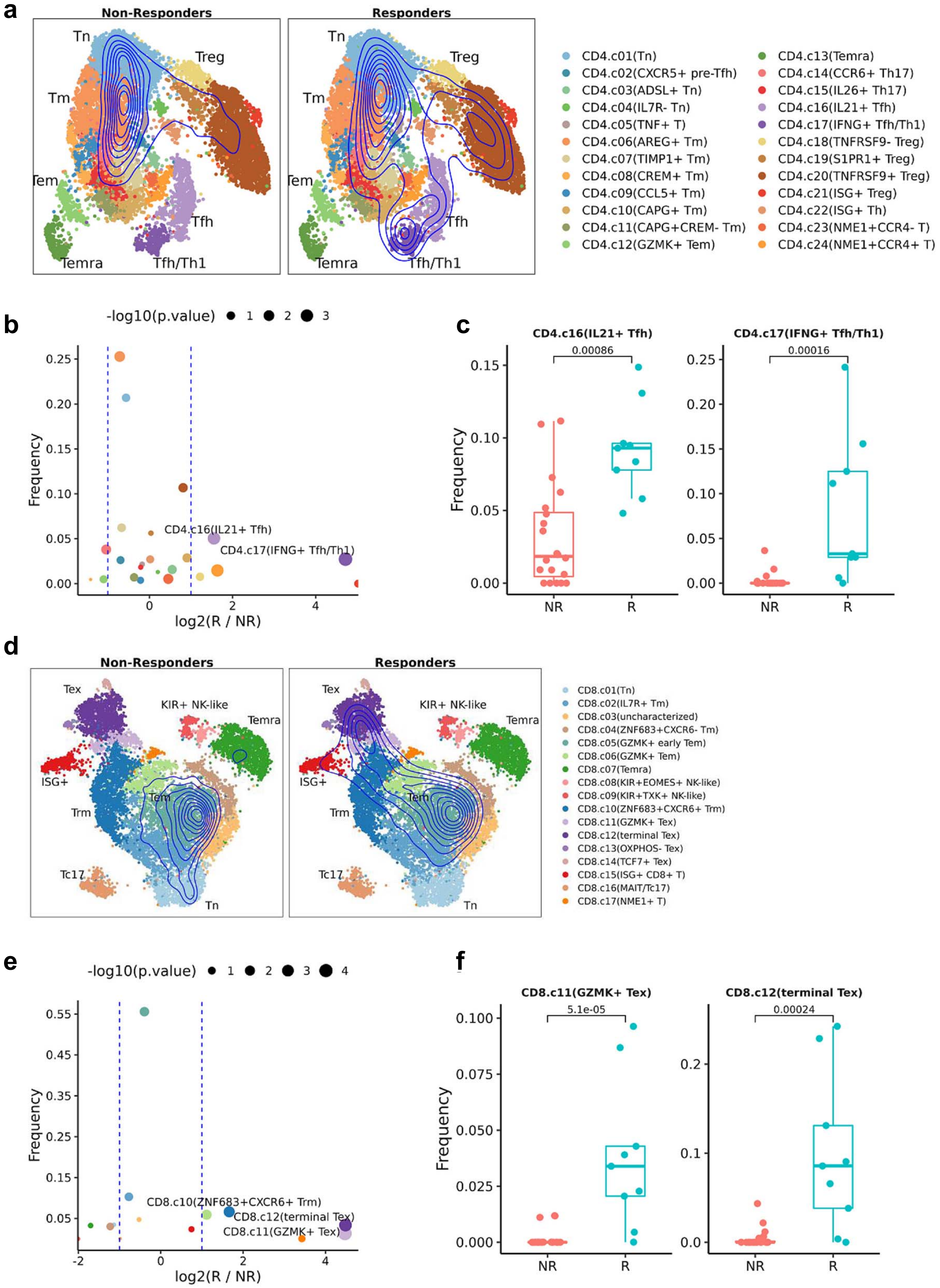

Figure 5. Interpreting the tumor-infiltrated T cells via ProjectSVR. We projected tumor-infiltrated CD4+ (a–c) and CD8+ T (d–f) cells from patients with breast cancer responding or non-responding to anti-PD1 treatment on the pan-cancer T cell atlas using ProjectSVR. And then transferred the reference T cell subtypes to query cells using a 10-NN classifier on reference embeddings. A. Projection of CD4+ T cells from responders and non-responders of anti-PD1 treatment by ProjectSVR. Dots represent reference metacells colored by cell type obtained from the original study. (b) Volcano plot of differential cell subtypes of CD4+ T cells between responders (R) and non-responders (NR). The P values were calculated by Wilcoxon test. Dots colored by cell subtypes are shown in (a). (c) Boxplots comparing the frequency of two CD4 + T cell subtypes between responders (R) and non-responders (NR). The P values by Wilcoxon test are shown. (d–f) The same plot as in (a–c) was applied to CD8+ T cells. Tfh, follicular helper T cells; Th1, T helper 1 cells. Treg, regulatory T cells. Tex, exhausted T cells; ISG, interferon-stimulated genes; Temra, terminally differentiated effector memory or effector; Tem, effector memory T cells; Trm, tissue-resident memory T cells; Tn, naïve T cells; and KIR, killer cell immunoglobulin-like receptors.

To further validate ProjectSVR, we reanalyzed scRNA-seq data from PD1-antibody-treated melanoma patients reported by Feldman et al. [24]. Our findings again reflected prior conclusions: responders exhibited more naïve CD8 T cells (c01) and fewer terminally exhausted CD8 T cells (c12) (Fig. S7). Notably, we also identified an increased proportion of Tc17 cells in responders, as described in the pan-cancer T cell atlas [16] (Fig. S7).

Hence, these two examples provide further evidence of the efficacy of ProjectSVR in accurately projecting and characterizing T cell states in cancer immunology studies.

Mapping the genetically perturbed germ cells to mouse testicular cell atlas

Organ-scale cell atlases, made possible by the growth of public scRNA-seq datasets and data integration tools, provide reference frameworks for future analyses [25]. We constructed an mTCA to evaluate whether ProjectSVR supports organ-scale reference mapping. We collected testicular cells from nine scRNA-seq datasets, covering stages from E6.5 embryos to adults (Fig. S8a, b, Table S3), and integrated them using the scVI algorithm [18]. Public cell type labels were harmonized, and unlabeled cells were classified using an SVM trained on gene module scores derived from cNMF (Supplementary materials). Cluster-level label refinement was done by majority voting over Leiden clusters [26] (Fig. S8c). The resulting mTCA comprises 34 cell types, including 17 germ cell types spanning five developmental stages, and 12 somatic types (Fig. 6a, b). Germ cells formed a continuous developmental trajectory on UMAP (Fig. 6a).

To assess ProjectSVR, we projected 17,253 germ cells from WT and *Zfp541*$^{-/-}$ mouse testes onto mTCA [27]. A k-NN classifier achieved 0.758 accuracy against original annotations (Fig. S9a, Table S4). Pseudotime inferred via ProjectSVR correlated strongly with published values ($r = 0.973$, Fig. S9b), confirming that *Zfp541*$^{-/-}$ cells failed to reach diplotene, as reported [27] (Fig. 6c, d).

We next projected WT and *Ythdc2*$^{-/-}$ germ cells from 10 dpp mouse testes onto mTCA [28]. Consistent with earlier cytological findings [29, 30], *Ythdc2*$^{-/-}$ cells arrested at the pre-leptotene stage (Fig. 6e, f). These examples show that ProjectSVR accurately maps continuous developmental trajectories onto organ-scale atlases, enabling interpretation of genotype-induced perturbations.

Evaluating *in vitro* induced meiosis using ProjectSVR

We further tested ProjectSVR's applicability in assessing *in vitro* development. Germ cells from 6-day-old mouse testes undergoing meiosis induction by nutrient restriction and retinoic acid (NRRA) were projected onto mTCA [31]. After NRRA treatment, spermatogonia progressed along the mTCA trajectory (Fig. 7a). Cell type transfer showed complete spermatogonial differentiation by Day 2 and meiosis initiation by Day 3 (Fig. 7b, c). Meiosis markers emerged at the pre-leptotene stage, while pre-meiotic markers were repressed (Fig. 7d), confirming NRRA's capacity to induce meiosis *in vitro* without Sertoli cells [31].

However, pachytene cells were not identified. Marker genes (e.g. *Piwil1*, *Tdrd6*) were absent (Fig. 7b–d). To test if pachytene stage was reached, we assessed meiotic sex chromosome inactivation (MSCI) by analyzing X chromosome UMI percentages along pseudotime, grouped by *in vivo* versus *in vitro* meiosis (Fig. 7e). Unlike *in vivo* meiosis, NRRA-induced cells failed to repress X-linked genes (Fig. 7f). Additionally, pachytene programs were inactive, and leptotene programs remained on in advanced *in vitro* cells (Fig. 7g). These results suggest that although pachytene-like cells appeared in cytological spreads, transcriptionally, NRRA-induced meiosis failed to reach the pachytene stage [31].

Discussion

With the rapid development of single-cell technology and decreasing sequencing costs, scRNA-seq has become essential for cell atlas construction and profiling cellular diversity under genetic or disease perturbation. However, inconsistent annotations and clustering across studies hinder comparisons between new and existing datasets. Integrating these data into biologically coherent atlases provides useful context and facilitates analysis. As shown in our case studies, ProjectSVR offers an unbiased method to compare query cells across technologies, genotypes, disease states, and treatments within a unified reference space.

Existing mapping tools typically treat reference mapping as a subset of data integration, reusing frameworks designed for aligning large, annotated datasets with smaller queries. However, many published atlases (e.g. tumor-infiltrated T cell landscape) lack mapping interfaces due to custom strategies or missing tools. In contrast, ProjectSVR frames mapping as a regression task, allowing direct prediction of reference embeddings using only the expression matrix and final embedding coordinates, making it suitable for cases lacking a dedicated mapping solution.

Feature engineering is key in machine learning, improving generalization and reducing training costs [8]. To address data sparsity in scRNA-seq, we used AUCell to compute gene activity scores (0–1) based on rank expression, robust to normalization and compatible with datasets lacking raw counts. Cell type-specific gene sets were derived either from overexpression testing (for discrete types) or cNMF (for continuous states). Although gene set size is a hyperparameter, results were consistent across different settings; we used the top 25 genes by default.

UMAP was chosen for reference embedding because it preserves global and local structure better than t-SNE, providing continuity across subsets [10]. While any meaningful low-dimensional representation can be used with ProjectSVR, UMAP was optimal for our use cases. Additionally, ProjectSVR supports pseudotime mapping: in mTCA, query cells were correctly placed along pseudotime (Fig. S9b), and in some cases (e.g. *Ythdc2*-KO), this representation was more sensitive to developmental arrest. We aim to support reproducibility in scRNA-seq analysis by providing five pre-trained reference models as part of ProjectSVR's initial release.

Compared to existing reference mapping methods such as Seurat, scArches, Symphony, and ProjecTILs, ProjectSVR offers a distinctive trade-off between generalization, flexibility, and computational efficiency. Our results (Fig. 3, Table 2) show that ProjectSVR achieves comparable or superior accuracy and ARI across multiple datasets, while also exhibiting favorable runtime and memory profiles. Notably, ProjectSVR operates without requiring access to the original integration models or latent space transformations, which are often unavailable for published references. This decoupling makes it particularly suitable for scenarios where researchers wish to reuse published atlases without replicating complex alignment pipelines. Furthermore, ProjectSVR's reliance on interpretable gene set scores instead of raw expression provides additional robustness across platforms and conditions. These characteristics position ProjectSVR as a lightweight, modular alternative for reference mapping—particularly valuable in situations involving inaccessible reference pipelines, cross-platform projections, or the need for scalable model deployment.

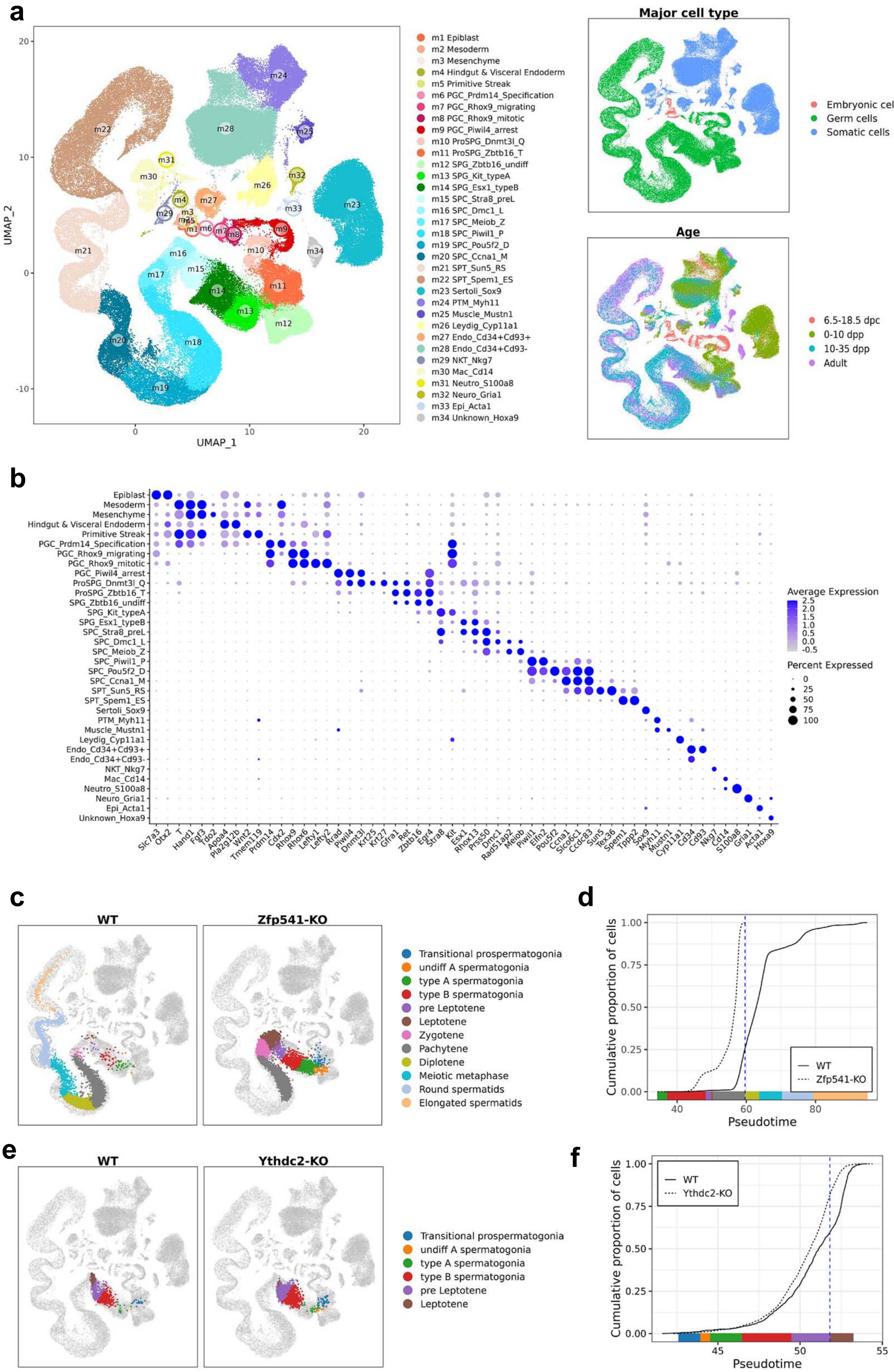

Figure 6. Mapping the genetically perturbed germ cells to mouse testicular cell atlas (mTCA). We built a reference of wild-type (WT) mouse testicular cells from 6.5 days post coitum (dpc) to adults and then mapped two datasets consisting of paired samples from WT and knock-out (KO) mice. (a) UMAP of WT mouse testis reference (n = 188,862 cells), colored by major cell types (upper right), age (lower right), and cell subtypes (left). Each dot represents a single cell. (b) Dot plot for expression of representative signature genes of testicular cells. The dot size represents the percentage of cells expressing the indicated genes in each stage and the dot color represents the average expression level of the genes. (c) Projection of WT and Zfp541-KO germ cells onto the testis reference. Grey dots represent reference cells, query cells colored by cell types predicted via ProjectSVR using a 10-NN classifier on reference embeddings. **D**. The cumulative distribution of cells along pseudotime from each mouse strain. The pseudotime was predicted via ProjectSVR from mTCA. (e, f) The same plots as in (c, d) were applied to *Ythdc2*-KO germ cells.

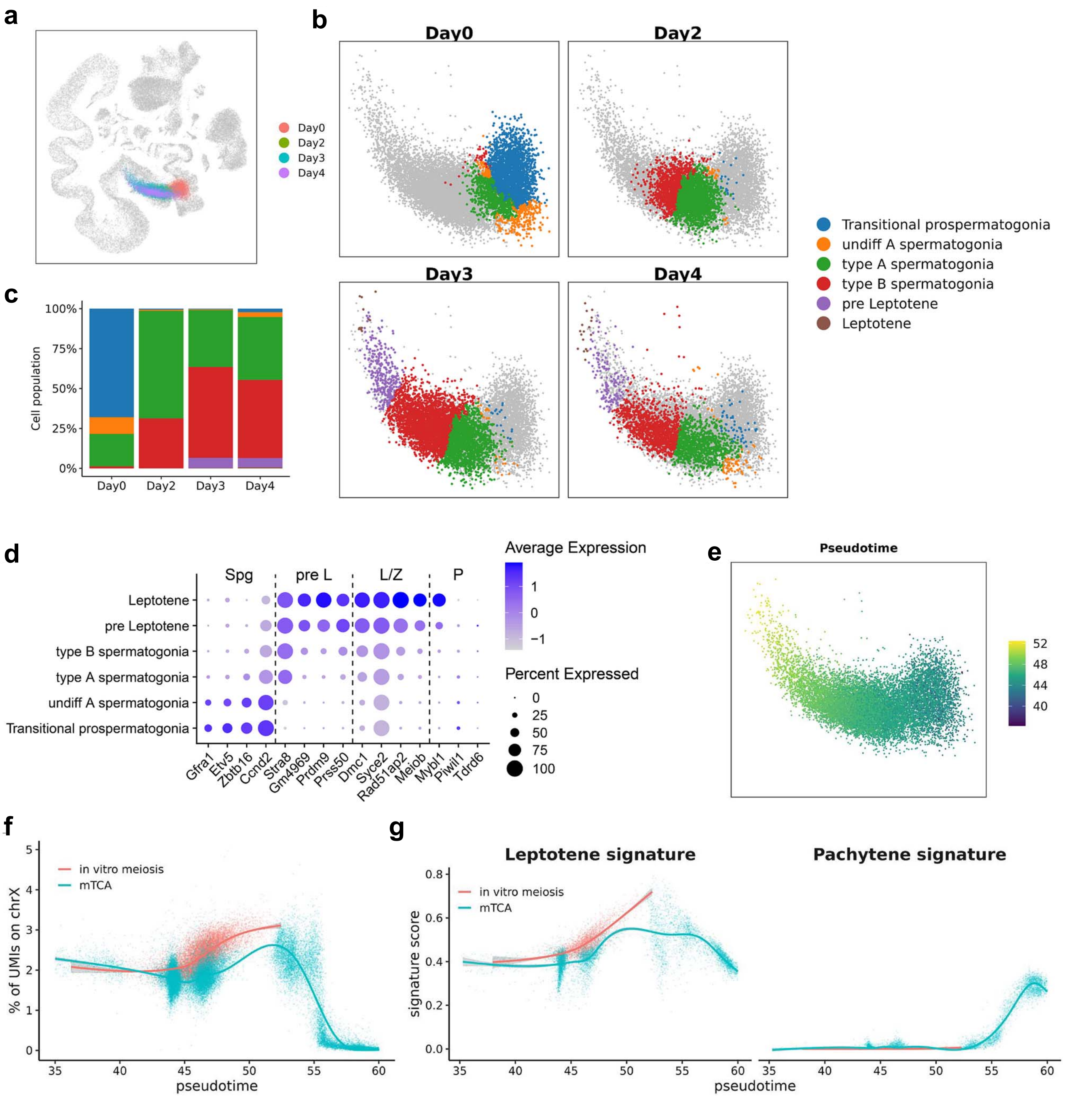

Figure 7. Comparing the *in vitro* induced meiotic cells to mTCA. (a) Projection of germ cells treated with nutrient restriction and retinoic acid (NRRA). Query cells were colored by days after NRRA treatment, and grey dots represent the reference cells of mTCA. (b) Cell labels transferred from mTCA using a 10-NN classifier on reference embeddings. Colored dots, cells collected from the indicated day after NRRA treatment colored by predicted cell labels; grey dots, cells not belonging to the indicated day. (c) The distribution of germ cell stages from different days after NRRA treatment. (d) Dot plot for expression of representative signature genes of indicated cell types. The dot size represents the percentage of cells expressing the indicated genes in each stage and the dot color represents the average expression level of the genes. Spg, spermatogonia; pre L, pre leptotene; L, leptotene; Z, zygotene; and P, pachytene. (e) Pseudotime predicted by ProjectSVR from mTCA. (f, g) Pseudotime analysis of MSCI and gene program activity (leptotene/pachytene) in NRRA-induced versus natural meiosis.

To further enhance reproducibility and usability, We plan to expand ProjectSVR into a scalable reference mapping ecosystem by releasing more pre-trained models and developing a web-based interface with an API and a growing database of curated reference models, focusing on system-specific rather than universal atlases like MCA [32]. Advances in machine learning and deep learning have inspired strategies for improving feature extraction, robustness, and scalability in high-dimensional biological data analysis [33–40]. ProjectSVR currently uses classical SVR algorithms with hand-crafted features. Recent advances in deep regression models—such as multilayer perceptrons (MLPs) or Transformer-based architectures—could enhance performance by learning richer representations from gene signature scores. These models may improve generalization across diverse data types and capture nonlinear dependencies missed by traditional SVR. In future work, we will therefore investigate these deep regression frameworks as alternative backends and incorporate model selection strategies tailored to specific biological contexts.

Despite its promising performance, ProjectSVR has limitations. (i) Its current modeling approach trains independent SVR models for each UMAP dimension. This strategy neglects potential correlations among embedding dimensions, which may limit the model's ability to learn joint structures. Future work could explore multi-output SVR or other multi-target regression methods to model the full embedding space more holistically. (ii) Reference building time is ~O(Nc^2), with Nc as the number of cells. For scalability, we subsampled to 4000–10,000 cells and used ensemble sub-models with median outputs, balancing performance and efficiency. (iii) If query cells contain novel states absent in the reference, fixed-model projections may be misleading. Users should compare unsupervised clustering to mapping results to detect inconsistencies. (iv) Performance declines when projecting across platforms (e.g. Smart-seq2 to 10x), due to differing UMI distributions and sequencing depth [41] (Table S5). Overall, we demonstrate the feasibility of reframing reference mapping as a multi-target regression task.

Key Points

- ProjectSVR addresses the limitations of existing reference mapping tools by decoupling mapping from data integration, enabling broad applicability even when integration interfaces are unavailable in existing atlases.
- ProjectSVR formulates reference mapping as a regression task, allowing the model to directly predict low-dimensional embeddings from gene activity scores without needing raw count matrices or pre-integrated datasets.
- A feature engineering strategy based on meta-gene activity scoring enhances robustness, allowing ProjectSVR to effectively handle data sparsity and generalize across technologies, genotypes, and biological conditions.
- ProjectSVR supports Uniform Manifold Approximation and Projection and pseudotime for context-aware cell mapping.

Author contributions

Jinman Fang and Jianing Gao conceived the idea, implemented and tested the software, and performed data analyses. Shipeng Guo, Guoshu Li, Ziran Bi, and Yue Hu provided advice on code optimization and visualization. Yuanwei Zhang, Bo Hong, and Hongzhi Wang supervised the project; Jianing Gao and Qizhi Zhu wrote the initial draft of the manuscript and substantially contributed to its revision and refinement during the review process. All authors reviewed and approved the final version of the manuscript.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (No. 82171601), the Anhui Provincial Key Research and Development Program (2022e07020054) and the Special Project for the Reform of the President of the Chinese Academy of Sciences (E3BA0G44131).

Conflict of interest: None declared.

Funding

None declared.

Data availability

The source code of ProjectSVR is available on GitHub: https://github.com/JarningGau/ProjectSVR. The reference cell atlases and query datasets involved in this paper are listed in Table 2 and available on Zenodo: https://zenodo.org/record/8350746 and https://zenodo.org/record/8350748. A step-by-step guide tutorial is provided on the GitHub page: https://github.com/JarningGau/ProjectSVR.

References

1. Luecken MD, Buttner M, Chaichoompu K. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* 2022;**19**:41–50. https://doi.org/10.1038/s41592-021-01336-8
2. Vento-Tormo R, Efremova M, Botting RA. *et al.* Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* 2018;**563**:347–53. https://doi.org/10.1038/s41586-018-0698-6
3. Han X, Zhou Z, Fei L. *et al.* Construction of a human cell landscape at single-cell level. *Nature* 2020;**581**:303–9. https://doi.org/10.1038/s41586-020-2157-4
4. Luecken M, Sikkema L, Strobl D. *et al. An Integrated Cell Atlas of the Human Lung in Health and Disease*. Preprint, Version 1, 15 March 2022. Research Square. https://doi.org/10.21203/rs.3.rs-1438584/v1
5. Lahnemann D, Koster J, Szczurek E. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol* 2020;**21**:31. https://doi.org/10.1186/s13059-020-1926-6
6. Abdelaal T, Michielsen L, Cats D. *et al.* A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol* 2019;**20**:194.
7. Huang Q, Liu Y, Du Y. *et al.* Evaluation of cell type annotation R packages on single-cell RNA-seq data. *Genomics Proteomics Bioinformatics* 2021;**19**:267–81. https://doi.org/10.1016/j.gpb.2020.07.004
8. Pasquini G, Rojo Arias JE, Schafer P. *et al.* Automated methods for cell type annotation on scRNA-seq data. *Comput Struct Biotechnol J* 2021;**19**:961–9. https://doi.org/10.1016/j.csbj.2021.01.015
9. Galdos FX, Xu S, Goodyer WR. *et al.* devCellPy is a machine learning-enabled pipeline for automated annotation of complex multilayered single-cell transcriptomic data. *Nat Commun* 2022;**13**:5271.
10. Becht E, McInnes L, Healy J. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology* 2019;**37**:38–44. https://doi.org/10.1038/nbt.4314
11. Andreatta M, Corria-Osorio J, Muller S. *et al.* Interpretation of T cell states from single-cell transcriptomics data using reference atlases. *Nat Commun* 2021;**12**:2965.
12. Gao C, Liu J, Kriebel AR. *et al.* Iterative single-cell multi-omic integration using online learning. *Nat Biotechnol* 2021;**39**:1000–7. https://doi.org/10.1038/s41587-021-00867-x
13. Kang JB, Nathan A, Weinand K. *et al.* Efficient and precise single-cell reference atlas mapping with symphony. *Nat Commun* 2021;**12**:5890.

14. Lotfollahi M, Naghipourfar M, Luecken MD. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat Biotechnol* 2022;**40**:121–30. https://doi.org/10.1038/s41587-021-01001-7
15. Xiong L, Tian K, Li Y. *et al.* Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. *Nat Commun* 2022;**13**:6118.
16. Zheng L, Qin S, Si W. *et al.* Pan-cancer single-cell landscape of tumor-infiltrating T cells. *Science* 2021;**374**:abe6474.
17. Li M, Zhang X, Ang KS. *et al.* DISCO: a database of deeply integrated human single-cell omics data. *Nucleic Acids Res* 2022;**50**:D596–602. https://doi.org/10.1093/nar/gkab1020
18. Lopez R, Regier J, Cole MB. *et al.* Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;**15**:1053–8. https://doi.org/10.1038/s41592-018-0229-2
19. Korsunsky I, Millard N, Fan J. *et al.* Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods* 2019;**16**:1289–96. https://doi.org/10.1038/s41592-019-0619-0
20. Hao Y, Hao S, Andersen-Nissen E. *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:e3529.
21. Guo C, Cai P, Jin L. *et al.* Single-cell profiling of the human decidual immune microenvironment in patients with recurrent pregnancy loss. *Cell Discov* 2021;**7**:1.
22. Denisenko E, Guo BB, Jones M. *et al.* Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol* 2020;**21**:130.
23. Bassez A, Vos H, Van Dyck L. *et al.* A single-cell map of intratumoral changes during anti-PD1 treatment of patients with breast cancer. *Nat Med* 2021;**27**:820–32. https://doi.org/10.1038/s41591-021-01323-8
24. Sade-Feldman M, Yizhak K, Bjorgaard SL. *et al.* Defining T cell states associated with response to checkpoint immunotherapy in melanoma. *Cell* 2018;**175**:e1020.
25. Sikkema L, Ramírez-Suástegui C, Strobl DC. *et al.* An integrated cell atlas of the lung in health and disease. *Nat Med* 2023;**29**:1563–77. https://doi.org/10.1038/s41591-023-02327-2
26. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 2019;**9**:5233.
27. Xu J, Gao J, Liu J. *et al.* ZFP541 maintains the repression of pre-pachytene transcriptional programs and promotes male meiosis progression. *Cell Rep* 2022;**38**:110540. https://doi.org/10.1016/j.celrep.2022.110540
28. Li L, Krasnykov K, Homolka D. *et al.* The XRN1-regulated RNA helicase activity of YTHDC2 ensures mouse fertility independently of m(6)a recognition. *Mol Cell* 2022;**82**:e1612.
29. Bailey AS, Batista PJ, Gold RS. *et al.* The conserved RNA helicase YTHDC2 regulates the transition from proliferation to differentiation in the germline. *Elife* 2017;**6**:e26116. https://doi.org/10.7554/eLife.26116
30. Jain D, Puno MR, Meydan C. *et al.* ketu mutant mice uncover an essential meiotic function for the ancient RNA helicase YTHDC2. *Elife* 2018;**7**:e30919. https://doi.org/10.7554/eLife.30919
31. Zhang X, Gunewardena S, Wang N. Nutrient restriction synergizes with retinoic acid to induce mammalian meiotic initiation *in vitro*. *Nat Commun* 2021;**12**:1758.
32. Han X, Wang R, Zhou Y. *et al.* Mapping the mouse cell atlas by microwell-Seq. *Cell* 2018;**172**:e1017.
33. Bing Z, Lemke C, Cheng L. *et al.* Energy-efficient and damage-recovery slithering gait design for a snake-like robot based on reinforcement learning and inverse reinforcement learning. *Neural Netw* 2020;**129**:323–33. https://doi.org/10.1016/j.neunet.2020.05.029
34. Hao X, Wang R, Guo Y. *et al.* Multimodal self-paced locality-preserving learning for diagnosis of Alzheimer's disease. *IEEE Transactions on Cognitive and Developmental Systems* 2022;**15**:832–43.
35. Deng R, Chen Z-M, Chen H. *et al.* Learning to refine object boundaries. *Neurocomputing* 2023;**557**:126742. https://doi.org/10.1016/j.neucom.2023.126742
36. Xie Z, Xa F, Chen X. Subsampling for partial least-squares regression via an influence function. *Knowledge-Based Systems* 2022;**245**:108661. https://doi.org/10.1016/j.knosys.2022.108661
37. Wang J, Li X, Ma Z. Multi-scale three-path network (MSTP-net): a new architecture for retinal vessel segmentation. *Measurement* 2025;**250**:117100. https://doi.org/10.1016/j.measurement.2025.117100
38. Zhao Y, Li X, Zhou C. *et al.* A review of cancer data fusion methods based on deep learning. *Information Fusion* 2024;**108**:102361. https://doi.org/10.1016/j.inffus.2024.102361
39. Mohammadzadeh-Vardin T, Ghareyazi A, Gharizadeh A. *et al.* DeepDRA: drug repurposing using multi-omics data integration with autoencoders. *PloS One* 2024;**19**:e0307649. https://doi.org/10.1371/journal.pone.0307649
40. Norouzi R, Norouzi R, Abbasi K. *et al.* DFT_ANPD: a dual-feature two-sided attention network for anticancer natural products detection. *Comput Biol Med* 2025;**194**:110442. https://doi.org/10.1016/j.compbiomed.2025.110442
41. Svensson V. Droplet scRNA-seq is not zero-inflated. *Nat Biotechnol* 2020;**38**:147–50. https://doi.org/10.1038/s41587-019-0379-5

Briefings in Bioinformatics, 2025, **26(6)**, bbaf586
https://doi.org/10.1093/bib/bbaf586
Problem Solving Protocol