Building a single cell transcriptome-based coordinate system for cell ID with SURE

Feng Zeng[1,2,4,5,*] and Jiahuai Han[2,3,*]

1. National Institute for Data Science in Health and Medicine, State Key Laboratory of Cellular Stress Biology, School of Aerospace Engineering, Xiamen University, Xiamen 361102, China
2. Research Unit of Cellular Stress of Chinese Academy of Medical Sciences, Cancer Research Center of Xiamen University, School of Medicine, Faculty of Medicine and Life Sciences, Xiamen University, Xiamen 361102, China
3. School of Life Sciences, Xiamen University, Xiamen 361102, China
4. Xiamen Key Laboratory of Big Data Intelligent Analysis and Decision, Xiamen 361102, China
5. Lead contact

[*] Correspondence: Feng Zeng, zengfeng@xmu.edu.cn; Jiahuai Han, jhan@xmu.edu.cn

Summary

The surge in single-cell transcriptome data necessitates a unified cell identification system. We propose a strategy that uses metacells as landmarks, coupled with a data-driven method called SURE, to create a cell ID coordinate system. This system streamlines cell atlases, enhances molecular signals, and ensures accurate cell identification while mitigating sequencing noise. We demonstrated the effectiveness of SURE by assembling similar cell types from datasets with cell type imbalance issues. We employed SURE to establish a unified coordinate system for the human hematopoietic system (HBMCA), which enables the direct, zero-shot mapping of single-cell datasets from 13 different diseases, showcasing its value as a common foundation for investigating the diverse disease effects on human blood cells. SURE has successfully addressed the long-standing challenge of positioning cells within a common multidimensional space, paving the way for a convergent understanding of cell identity across diverse experimental landscapes.

Introduction

Revealing and classifying the type and subtype of cells lie at the core of life sciences[1–4]. For centuries, defining the identity of a given cell (cell ID) has relied on discrete, concrete, and qualitative descriptors of features such as morphology and function. The advent of single-cell techniques has given rise to numerous cell atlas datasets, presenting an unprecedented opportunity for systematically defining cell IDs. Studies of these cell atlases have revealed that cell types and cell subtypes/states vary along a continuum, challenging the traditional notions of discrete cell identities. Furthermore, the single-cell measurement is highly sparse, and a variety of technical and biological noise can interfere with accurate characterization. There is no doubt that a unified, single cell transcriptome-based coordinate system for cell ID is timely needed, and a comprehensive characterization tool for cell populations has to be developed for constructing this reference system.

Currently, two strategies exist for building reference systems for transcriptome-based cell ID. The first relies on cell type annotations mostly obtained from bulk sequencing of purified cell populations of known types. The other involves cell manifolds[5]. Although cell manifolds are based on single cell sequencing data, the reference systems yielded by manifolds are inherently local and potentially differing from one another due to their constructions being made by different datasets. Both approaches are very valuable and had greatly facilitated research in the field. However, the former is

constrained by the availability and accuracy of prior annotations and the latter has inherent limitations of unsupervised manifold learning techniques. Both of them are unable to fully meet the need for the establishment of a data-driven, unified reference system for cell ID in analyzing single cell transcriptomes.

In this work, we propose leveraging metacells to establish a coordinate system that provides a reference framework for cell ID. The concept of metacell was initially introduced to address the challenges of sparsity and noise in single-cell genomics data[6]. A metacell represents small, homogeneous groups of cells identified by the minimum unit of homogeneity. Employing metacells should make it feasible to construct a coordinate system that can serve as a reference for cell ID. First, since a set of metacells should represent the distribution of cell types and subtypes in a given single cell transcriptome dataset, metacells can serve as landmarks for mapping, aligning, and assembling different datasets. This lays the foundation for establishing a global, standardized coordinate system for all cell types. Notably, this process is data-driven, circumventing the restriction of prior annotation of known cell types. Second, metacells pool the information from homogeneous cells and reduce the impact of single-cell sequencing noise and sparsity, providing a more precise representation of cell ID. Third, metacells provide a compressed representation of cell atlas. By summarizing the high-dimensional single-cell data into a smaller set of metacells, metacells enable the efficient analysis of large-scale single-cell datasets, opening up new avenues for data exploration and hypothesis generation.

Several computational methods for metacell calling have evolved from cell clustering, employing discriminative approaches to partition cell data based on similarity, including Metacell[7], Metacell2[8], SuperCell[9], and SEACells[10]. However, these clustering-based strategies are not well-suited for constructing a unified coordinate system. The objective of defining metacells for the coordinate system differs from metacells defined by cell clustering, as the former should capture the inherent continuity of cell states, transcending the discrete boundaries imposed by clustering algorithms. As a common problem, accurately defining cell similarity remains challenging due to the extreme sparsity inherent to single-cell genomics data, and existing metacell calling methods have not modeled and eliminated the impact of technical variations, especially batch effects, as well as the random differences among samples, rendering it difficult to project other single-cell data onto the coordinate system formed by these metacells. MCProj[11] attempts to achieve the projection of single-cell data by iteratively eliminating the differences between the query data and the metacell atlas, but this iterative algorithm struggles to meet the requirements of a unified coordinate system in terms of consistency, accuracy, and efficiency when dealing with large-scale, noisy data. Therefore, constructing a unified coordinate system requires new tools to overcome the limitations of existing metacell calling approaches.

In quest to define metacells for the coordinate system, we draw inspiration from vector quantization (VQ), a classic data compression technique from information theory[12]. VQ revolves around a codebook, a collection of codewords serving as data prototypes that capture the dataset's characteristics with minimal information loss. The VQ-based generative approach diverges from the traditional cluster-based definition of metacells, conceptualizing the codebook as a collection of metacells, each representing a fundamental cell type or subtype. By employing VQ and deep generative modeling, we have achieved a principle for defining metacells to establish a coordinate system for cell ID. Leveraging the codebook's compact representation, we discover metacells with enhanced representational power,

comprehensively capturing the distribution of cell states and overcoming technical noise, i.e., batch effects. These metacells establish a unified coordinate system that accurately reveals salient features and underlying patterns, even when single-cell data quality is suboptimal. Metacells based on the VQ generative model effectively addresses cell type imbalance and enable the assembly of multiple single-cell atlas datasets through a hierarchical strategy. Using this divide-and-conquer approach, we have established a coordinate system for the cells from human hematopoietic system using single-cell transcriptome data. This metacell-based coordinate system supports the mapping of query data without requiring batch correction or model fine-tuning, making it a completely zero-shot process. Our method for establishing a coordinate system provides a technical foundation for constructing a unified reference for cell ID, paving the way for advancing single-cell data integration and analysis.

Results

A generative model for metacell calling

A key characteristic of the VQ codebook is its compact nature, with its size typically being significantly smaller than that of the original data. The primary objective of VQ is to optimally represent the original data using the codewords within the codebook, thereby achieving efficient compression. VQ encompasses two fundamental processes: encoding and decoding. The encoding process involves mapping each data point to a corresponding codeword within the codebook. This mapping is achieved by utilizing the index of the codeword to represent the data point, effectively compressing the original information. When the index is transmitted to the destination, the decoding process comes into play. By leveraging the codebook and the received index, the decoding process can optimally recover the encoded data point, reconstructing the original information with minimal loss. The encoding-decoding paradigm of VQ serves as a powerful framework for defining metacells through generative modeling (**Figure S1**).

We introduce SURE (SUccinct REpresentation of cells), an advanced approach that seamlessly integrates VQ into a probabilistic generative model to define metacells (**Fig. 1a**). SURE consists of two main components: a generative process based on a VQ strategy and a universal encoder. In the metacell-centric generative modeling, we propose the existence of a cell codebook that can be applied to decode single-cell transcriptomic expression profiles. This codebook is composed of k metacells, which form the coordinate system, or backbone, supporting the latent space of cells. The codebook stores the standard cell states of these k metacells under conditions free from noise interference, such as batch effects. To generate a single-cell data point, we first randomly select a metacell. We then query the codebook to obtain the state variable of the selected metacell and use it as the initial state for the single cell. Taking into account the batch effects prevalent in atlas data, we augment this single-cell state variable by incorporating a factor variable indicating the data batch, creating a batch-specific state variable. Subsequently, we input this batch-specific state variable into a zero-inflated negative binomial (ZINB) model to simulate the sparse single-cell data generation process[13]. This completes the generation of a single-cell data point.

Having established a robust reference coordinate system, we envision the ability to project the data submitted by users onto this coordinate system, akin to mapping sequencing reads to a reference genome. Users' data may exhibit various variations, with batch effects being unpredictable. Therefore, we propose the projection onto this coordinate system to be a zero-shot means, eliminating the need

for fine-tuning to remove batch effects between users' data and the reference database. Instead, it directly identifies the true cell states within the users' data and projects it onto the reference coordinate system. Subsequent analysis results demonstrate that fine-tuning with different data input sequences will yield divergent outcomes. Furthermore, in genomic analysis, the mapping of sequencing reads is an entirely zero-shot process. Consequently, compared to fine-tuning, zero-shot will be a preferred strategy, enabling the capture of genuine cell information with greater fidelity.

To achieve zero-shot cell mapping, we draw inspiration from the asymmetric encoder-decoder structure of SCALEX[14] and propose an asymmetric variational Bayesian model. During the simulation of the single-cell data generation process, we incorporate batch information, as described above. However, in the encoding process, which projects cell data onto the coordinate system composed of metacells, we solely utilize the single-cell expression profile data without incorporating batch information. In both SCALEX's practice and our own, this asymmetric encoder-decoder architecture exhibits remarkable universal encoding capabilities, exemplifying zero-shot performance. In our implementation, the parameters of this metacell-based generative model and the codebook are jointly learned using stochastic variational inference techniques. Please refer to the online methods for more detailed information.

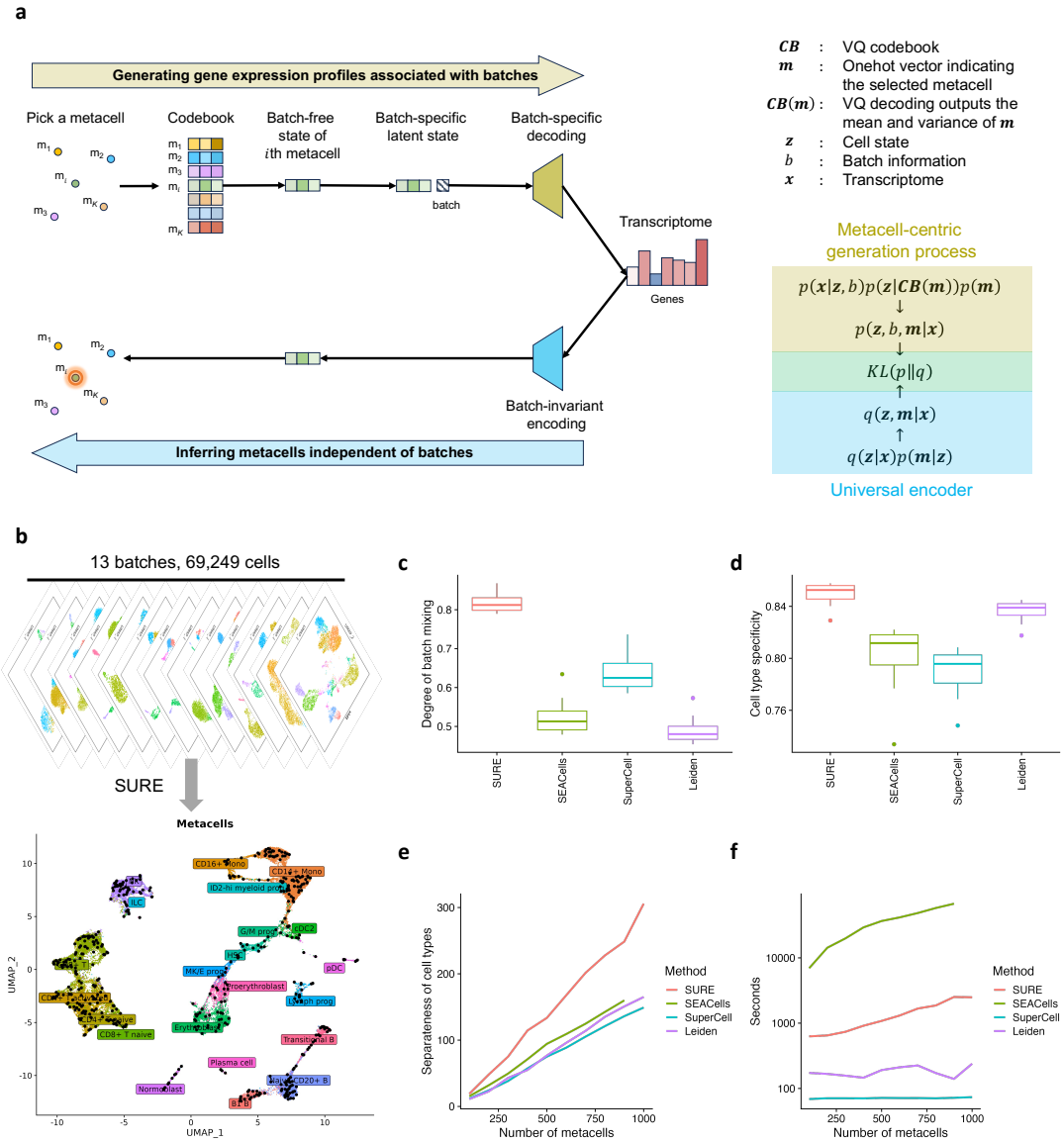

Fig. 1 Model architecture and evaluation of metacell calling algorithms. **a**, The generation progress commences with the selection of a metacell, retrieving its state from the codebook as the initial state of the cell being generated. After augmenting the initial state with batch factor information, a neural network decoder translates the batch-specific state to the gene expression profile of the cell. Conversely, neural network encoders map a cell to a point within the cell state space characterized by the learned codebook, and assign it to a metacell. **b**, Visualization of metacells called from the testing dataset comprising 13 batches. **c-f**, Performance evaluation metrics for metacell calling algorithms include degree of batch mixing (**c**), cell type specificity (**d**), separateness of cell types (**e**), and runtime in seconds (**f**).

Evaluation of metacell calling algorithms

First, we evaluate SURE, existing metacell calling algorithms, and clustering methods. We utilized the healthy human BMMC dataset provided by the NeurIPS 2021 Open Problems in Single-cell Analysis competition for the evaluation[15]. This dataset comprises 69,249 cells from 13 batches, generated using 10X Multiome technology, including both RNA and chromatin accessibility data (**Fig. 1b**). In this study,

we focus solely on the RNA data for evaluation. We selected three methods for comparison: two metacell calling algorithms, SEACells and SuperCell, and a cell clustering algorithm Leiden[16]. We did not include Metacell and Metacell2 in the evaluation, primarily because these algorithms do not allow users to set the number of metacells, precluding systematic testing across different parameter settings.

We employ these algorithms to call metacells and then evaluate their ability to represent the original data. For metacells to effectively and comprehensively capture the cell variations present in the data, they should exhibit outstanding performance in the following three key aspects. First, metacells should be independent of batch variations, a critical requirement for robust data representation. Second, each metacell should exhibit a distinct and homogeneous cell type identity, enabling the investigation of cell heterogeneity. Third, metacells should accurately capture the distinctions between various cell types, a crucial aspect of characterizing the continuous landscape of cell variations. Consequently, we employ three complementary metrics to quantify the performance of called metacells: (1) the degree of batch mixing, which quantifies the heterogeneity of batch origins for the cells assigned to each metacell; (2) cell type specificity, which quantifies the homogeneity of cell type identities for the cells assigned to each metacell; (3) the separateness of cell types, defined by the Variance Ratio Criterion[17], which is the ratio of the separation between different cell types to the separation within the same cell type. This metric, derived from principal component analysis (PCA) on the expression profiles of the metacells, quantifies the ability of metacells in representing the continuous cell landscape.

SURE exhibits exemplary performance in metacell calling. Firstly, the metacells identified by SURE demonstrate a high degree of representing consistent information across different data batches (**Fig. 1c**). SURE's average degree of batch mixing score exceeds 0.8, indicating its ability to mitigate batch effects and capture batch-independent cell variations. The scores of other methods do not surpass 0.65. Secondly, the metacells identified by SURE exhibit a high degree of representing unique and distinct cell types (**Fig. 1d**). SURE's average cell type specificity score reaches approximately 0.85, significantly outperforming all other evaluated methods. This result underscores SURE's ability to capture the inherent cell heterogeneity and identities within the data. Thirdly, the metacells obtained by SURE effectively capture the distinctions between various cell types (**Fig. 1e**). Notably, as the number of metacells increases, SURE's ability to characterize the continuous cell landscape improves, enabling a more nuanced representation of the data.

SURE not only effectively represents the variations in single-cell data but also demonstrates impressive computational scalability (**Fig. 1f**). Although SuperCell and clustering algorithms exhibit rapid computation times, their ability to define informative metacells lags behind SURE and SEACells. Notably, SURE's computation speed is orders of magnitude faster than SEACells. Particularly, as the number of metacells and the scale of the data increase, SURE's computational advantage over SEACells becomes more pronounced. These results underscore SURE’s suitability for the computational analysis of large-scale cell atlas datasets. Above all, SURE excels in capturing batch-independent, cell type-specific, and biologically meaningful metacells, while maintaining computational efficiency.

Geometric interpretation of metacells

From a geometric perspective, SURE defines metacells that constitute a landmark set, representing a collection of reference points within the continuous cell state space. By establishing a discrete set of

landmarks (metacells), SURE offers a principled framework for systematically organizing and navigating the intricate landscape of cell identities and phenotypes. The codebook, acting as a reference system, enables the consistent positioning of cells relative to these landmarks, facilitating comparisons and analyses across diverse datasets and experimental conditions. The encoder component of SURE allows for the projection of individual cells onto the defined reference system, effectively mapping their multidimensional expression profiles to specific coordinates within the cell state space. Conversely, the decoder component models the inverse mapping, translating the coordinates of a landmark (metacell) back to the corresponding expression profile. This inverse mapping offers a means to interpret and characterize the molecular signatures associated with each discrete cell state. This geometric interpretation of SURE highlights its potential to serve as a unifying framework for harmonizing and integrating single-cell data from various sources. By bridging the gap between the continuous cell state landscape and discrete representations, SURE provides a critical foundation for cell positioning, querying, and annotation. This framework holds the promise of advancing our comprehension of cell identity, heterogeneity, and fate decisions.

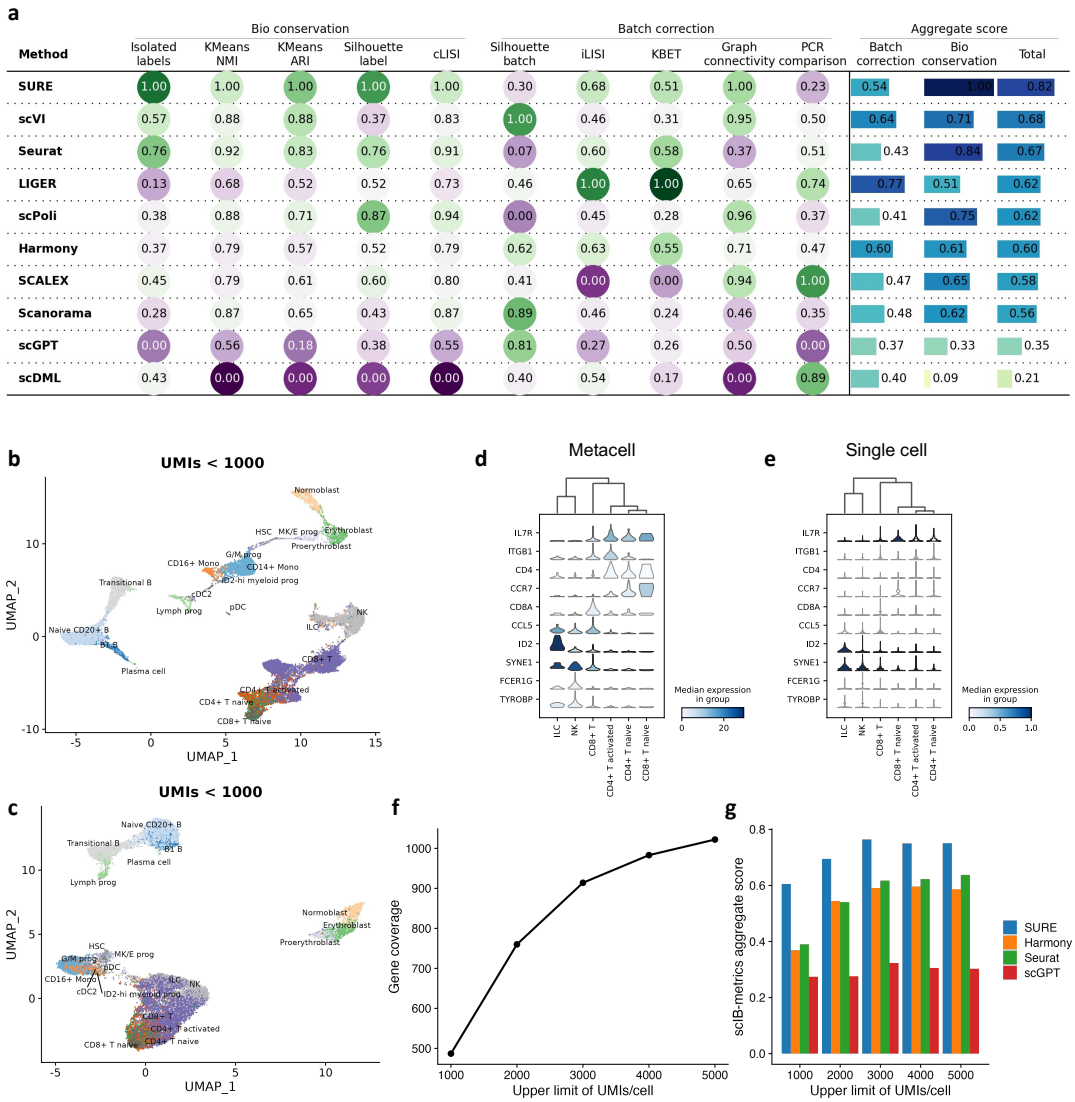

Fig. 2 Evaluation of the cell coordinate system defined by SURE. **a**, Performance of batch correction and biological signal conservation across various methods (with min-max scaling). **b-c**, Visualization

of integrated single cell data when SURE (**b**) and Harmony (**c**) were applied to low-quality data. **d-e**, Expression levels of T/NK/ILC-associated markers in metacells (**d**) and single cells (**e**). **f**, Gene coverage increases along with the improvement of data quality. **g**, Performances of SURE and other methods in batch correction and bio-conservation vary with data quality improvements.

Metacells establish a cell coordinate system unraveling the inherent biological structure precisely

We pose the following pivotal question: Can metacells serve as an effective cell coordinate system? A viable cell coordinate system should provide a consistent representation of cell ID across datasets. Hence, we compared SURE with a comprehensive set of low-dimensional representation learning methods for single-cell data, including eight unsupervised algorithms (Seurat[18], LIGER[19], Harmony[20], Scanorama[21], SCALEX[14], scVI[22], scPoli[23], and scDML[24]) and a large language model (scGPT[25]). We employed the scIB-metrics package[26] to assess the performance of these methods in batch correction and biological signal preservation, two critical benchmarks for evaluating the effectiveness of a cell coordinate system. Results are depicted in **Fig. 2a**.

The results demonstrate that SURE's batch effect removal capability achieves a competitive overall score. Specifically, SURE's aggregate score for batch correction is 0.54, trailing behind LIGER (0.77), Harmony (0.60), and scVI (0.64), but outperforming Seurat (0.43), Scanorama (0.48), and others. These findings underscore SURE’s competitive performance in mitigating batch effects, a crucial prerequisite for establishing a reliable cell coordinate system.

Furthermore, our comprehensive comparison results confirm that leveraging a cell coordinate system aids in preserving the biological structure inherent to the data. For instance, by utilizing metacells, SURE obtains the highest bio-conservation score, significantly outperforming all other evaluated algorithms. Among the remaining algorithms, Seurat performs the best with a score of 0.84, lower than that of SURE by 16%. The performance of scPoli (0.75) and scVI (0.71) follow closely behind。 Large language models like scGPT aim to extract intrinsic patterns, or a unified cell manifold, from massive amounts of single-cell data. However, the comparison results reveal a significant gap between scGPT's performance and the expectations, underscoring the challenges associated with unsupervised manifold learning approaches.

These findings demonstrates that unsupervised learning algorithms utilizing the cell coordinate system defined by metacells can substantially outperform unsupervised algorithms that do not leverage a cell coordinate system. Moreover, unlike cell type annotation, which is constrained by the availability and accuracy of prior knowledge, metacells are learned directly from the data itself, offering a more flexible and unbiased approach to define a cell coordinate system.

The metacell-defined cell coordinate system is stable with varying data quality

The cell coordinate system defined by metacells exhibits robustness to data quality variations. Even with suboptimal data quality, this reference system can accurately recover the inherent structure of the data. We utilize the UMI count per cell as an indicator of cell quality. By setting different upper limits on the UMI count per cell, we constructed five datasets representing varying degrees of single-cell sequencing quality. When sequencing quality is poor—characterized by low UMI counts and sparse gene coverage—detect cell subpopulations becomes challenging. For instance, when the UMI count per cell is less than

1,000, the number of genes detected in each cell is only 487 ± 144. In the UMAP visualizations generated by Harmony, T cell subtypes (CD4+ T naïve, CD8+ T naïve, CD4+ T activated) and monocyte cell subtypes (CD14+ Mono, CD16+ Mono, ID2-hi myeloid prog) are indistinguishable, appearing as intermixed populations. In contrast, SURE's results demonstrate its ability to better delineate T cell and monocyte cell subtypes, even under suboptimal data quality conditions (**Figs. 2b and c**). Moreover, metacells accurately capture the expression of marker genes characteristic of various cell subtypes. For example, CD4+ T activated cells highly express IL7R and ITGB1, CD4+ T naïve cells highly express CD4 and CCR7, ILCs highly express ID2 and SYNE1, and NK cells highly express FCER1G and TYROBP (**Fig. 2d**).

As the UMI count per cell increases, reflecting improved data quality, the gene coverage within cells also improves (**Fig. 2f**). With higher data quality, SURE's characterization of the inherent data structure becomes more precise, enabling finer delineation of cell subpopulations, such as the separation of the CD14+ Mono and CD16+ Mono subtypes of monocytes. We use the scIB-metrics aggregate score, which combines batch effect removal and biological signal preservation, to evaluate performance (**Fig. 2g**). As cell data quality improves, the performance of most methods increases. Among the compared methods, SURE's performance is the most notable.

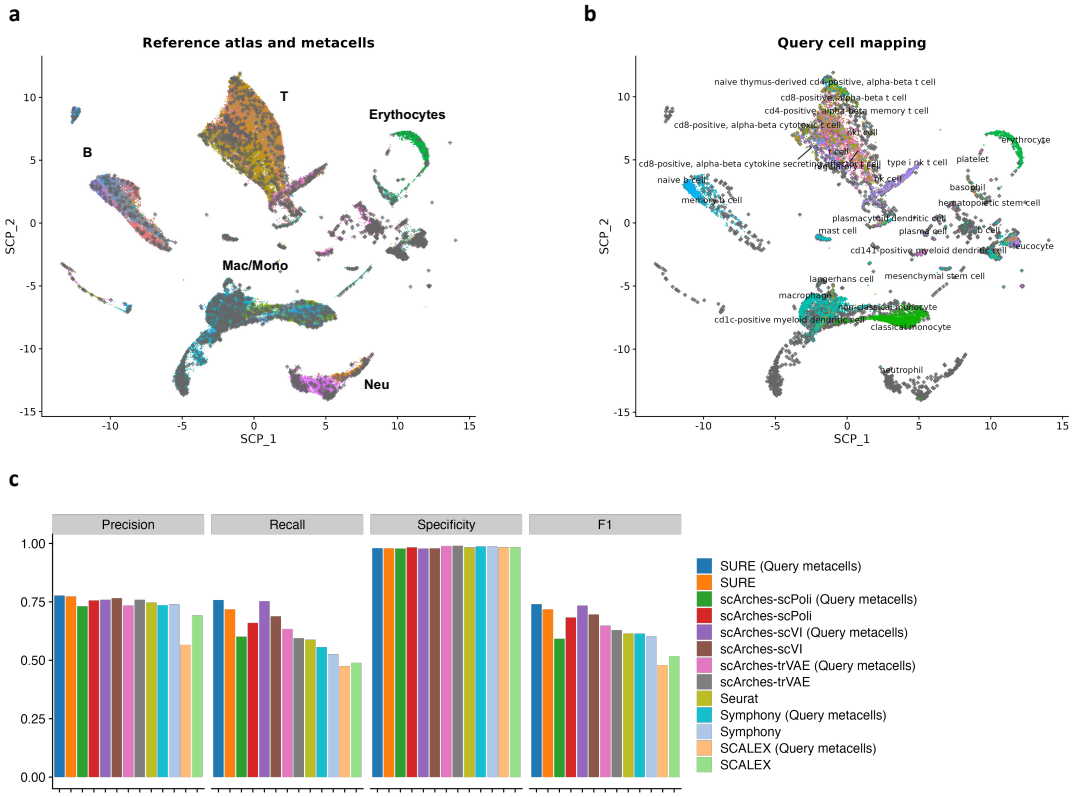

Fig. 3 The positioning of query cells within the cell coordinate system. **a**, Visualization of cells in the reference atlas, colored by cell types, and metacells called by SURE, displayed as gray points representing landmarks within the cell coordinate system. **b**, Visualization of query cells when mapped to the reference atlas according to the landmarks characterized by metacells. **c**, Evaluation metrics of query cell mapping methods, including precision, recall, specificity, and F1 score.

Accurate positioning of query cells

As an application, we employ SURE to define the reference system of single-cell atlas datasets and achieve their compression. An atlas dataset containing millions of cells can be effectively compressed into thousands of metacells. Notably, these metacells retain the essential functionalities of the original single-cell atlas dataset, such as the positioning of query cells.

Mapping query cells is one of the critical functions of single-cell atlas datasets, enabling the integration and interpretation of new data within the context of a comprehensive cell landscape. However, query datasets and atlas datasets may differ in their data generation processes, and these batch differences pose significant challenges for accurate query cell mapping. Employing batch correction algorithms to eliminate batch differences between query datasets and atlas datasets is not a feasible solution. Given the massive scale of atlas datasets, the computational cost of correcting batch differences is prohibitively high. Whenever a user submits a query dataset, re-correcting the batch differences between the query dataset and the atlas dataset would incur a computational cost that would severely hinder the practical application of the atlas dataset.

Currently, the main approaches for solving query cell mapping are the model fine-tuning strategy proposed by scArches[27] and the atomic sketch integration and TransferData/MapQuery methods introduced by Seurat[28]. These approaches represent two different strategies. One strategy leverages the computational efficiency of deep generative models. scArches fine-tunes model parameters based on the query dataset, using the model trained on the atlas dataset as a starting point, to eliminate batch differences between the query and the atlas datasets. The alternative strategy utilizes small-scale representative sketched cells, which is a cell subset conserving data properties[29], to improve the computational efficiency of query cell mapping. Metacells represent a third strategy, combining the advantages of the previous two approaches. By defining the reference system of the atlas dataset, metacells enable query cells to be directly mapped onto this reference system without the need for fine-tuning. This approach circumvents the computational overhead associated with model fine-tuning while avoiding the potential information loss inherent to cell sketching.

To demonstrate the effectiveness of our metacell-based strategy, we utilize the immune compartment of the human tissue cell atlas dataset Tabula sapiens[30] for evalulation (**Fig. 3a**). We extract data from one individual (TSP10) from this atlas as the query dataset and employ the remaining data for model training. We evaluate three distinct computational strategies. The first strategy is model fine-tuning. We train base models for the atlas dataset using scVI, trVAE[31], and scPoli, and then utilize scArches to fine-tune these models based on the query dataset. The second strategy involves the use of sketched cells. We integrate the atlas dataset using Seurat's atomic sketch integration and then employ MapQuery for mapping the query dataset. For the third strategy, we use SCALEX and Symphony[32] as baseline methods, which do not incorporate fine-tuning or algorithmic sketching capabilities. After obtaining the joint representation of the query dataset and the atlas dataset, cells in the query dataset should ideally be mapped to the corresponding cell type locations in the atlas dataset. Therefore, we train a logistic regression-based cell type classifier using the atlas dataset and evaluate the performance of query cell mapping by assessing the cell type annotation results of the query dataset.

Through metacells, SURE positions the various cell types in the query dataset at their corresponding locations within the cell landscape of the atlas dataset, and then leverages this cell positioning to

annotate the query cells (**Fig. 3b**). We employ metrics such as precision, recall, specificity, and F1 score to evaluate the performance of query cell positioning (**Fig. 3c**). Among these four metrics, SURE outperforms other methods in precision, recall, and F1 score. Both the model fine-tuning strategy and the query strategy based on sketched cells perform better than the baseline methods. Notably, the model fine-tuning strategy outperforms the query strategy based on sketched cells, indicating that the model fine-tuning strategy effectively mitigates the impact of batch effects on query data positioning. Although the query strategy based on sketched cells may have a computational advantage in handling large-scale data, batch effects can still adversely affect the performance of transfer learning. SURE's superior performance surpasses these methods, demonstrating its ability to effectively address these challenges and accurately locate query cells within the atlas dataset.

By employing metacells as robust landmarks within the cell coordinate system, SURE enables the precise positioning of query cells within the comprehensive cell landscape represented by the atlas dataset. This precise positioning is a critical step for annotating and interpreting new query data in the context of the established cell architecture. The exceptional performance of SURE, as demonstrated by superior precision, recall, and F1 metrics, underscores its unparalleled ability to accurately map query cells to their corresponding cell identities and states within the atlas. SURE outperforms alternative strategies, such as model fine-tuning and sketched cell-based approaches, in this crucial task.

Query in the form of metacells

We also made a compelling discovery that computing metacells from query cell datasets and utilizing these query metacells as representations of query cells, instead of the original query cells, significantly improves the performance of query cell mapping. For algorithms such as SURE, scVI, and Symphony, employing query metacells led to performance gains (**Fig. 3c**). However, for SCALEX, query metacells resulted in a performance decline. This finding indicates that the metacell approach effectively reduces the impact of noise on query cell data, thus optimizing query cell mapping and enabling more accurate integration and interpretation of new data within the established cell coordinate system.

Metacell-centric hierarchical assembly of cell atlas datasets

Next, we demonstrate how to leverage metacells to assemble different single-cell atlas datasets and construct a reference cell coordinate system. Assembling single-cell atlas datasets poses two main challenges (**Figure S2**). The first pertains to data scale. Since atlas datasets are typically massive, most integration algorithms are unable to handle the assembly of these large-scale datasets. The second challenge arises from data imbalance[33], as the cell atlas datasets being assembled may exhibit dramatic differences in cell type composition and distributions. Here, we show that metacells can effectively overcome both of these challenges.

We propose a hierarchical strategy for assembling single-cell atlas datasets (**Fig. 4a**). Assume we have n single-cell atlas datasets, we first use SURE to compute the metacells for each atlas dataset separately. We refer to these metacells as primary metacells. This initial computation of primary metacells achieves efficient compression of each individual dataset, reducing the computational burden. Then, using the expression profiles of the primary metacells, we again apply SURE to compute metacells, which we term secondary metacells. These secondary metacells, derived from the primary metacells, establish a unified cell coordinate system. Finally, based on the coordinates of the primary metacells in the unified

cell coordinate system, we assemble the n single-cell atlas datasets together. This hierarchical strategy enables the consistent positioning and assembly of cells across datasets, regardless of differences in cell type composition or distributions.

To demonstrate the integration of cell atlas datasets, we selected three human cell atlas datasets that exhibit striking differences in cell composition (**Fig. 4b**). These datasets encompass five major cell types: endothelial cells, epithelial cells, glial cells, stromal cells, and immune cells. However, the proportions and subtypes of these cell types vary significantly across the three datasets. The ImmuneCrossTissue atlas dataset[34] primarily comprises 21 immune cell subtypes derived from various tissue contexts. In contrast, the snCrossTissue atlas dataset[35] focuses on non-immune cell types, featuring 10 stromal cell subtypes, 10 epithelial cell subtypes, and a limited number of immune cells. Distinct from the first two datasets, the TabulaSapiens atlas dataset[30] offers a comprehensive representation of cell diversity, covering the majority of human cell types. Notably, these three atlas datasets employ different cell type annotation systems. This inconsistency in annotation poses a significant challenge for integrating cell atlas datasets from diverse sources.

We employed the proposed hierarchical strategy to assemble three cell atlas datasets. Remarkably, SURE demonstrated robust cell atlas assembly results in the face of cell type imbalance, outperforming conventional data integration methods (**i.e.**, Scanorama and Harmony), model fine-tuning strategies (**i.e.**, scArches-scVI), and large language models (**i.e.**, scGPT) (**Fig. 4c**). To illustrate this, we examined the integration results for epithelial cells, stromal cells, and immune cells. The ImmuneCrossTissue dataset lacked epithelial and stromal cells, while both the snCrossTissue and TabulaSapiens datasets contained these cell types, albeit with potential variances in cell subtype composition. Leveraging the cell coordinate system, SURE accurately assembled the epithelial and stromal cell types from the snCrossTissue and TabulaSapiens datasets. Among the comparison methods, only Harmony achieved a certain degree of data integration for epithelial and stromal cells. In contrast, Scanorama, scArches-scVI, and scGPT failed to integrate the epithelial and stromal cells from the scCrossTissue and TabulaSapiens datasets, indicating the impact of cell type imbalance on these methods. Interestingly, in SURE's results, we observed that some epithelial and stromal cells were mapped to the same metacells, a phenomenon also noted in Harmony's results. This reflects the similarity between these two cell types and the challenge in accurately characterizing their differences. Moreover, the epithelial and stromal cells from these two datasets rarely mapped to metacells closely associated with immune cells. Regarding immune cell types, SURE, Harmony, scGPT, and Scanorama successfully integrated them, with only scArches-scVI failing in this task. To find the optimal performance result of scArches-scVI, we conducted an intriguing experiment by adjusting the input order of datasets during fine-tuning. Surprisingly, we found that the input order had a significant impact on scArches-scVI, with different orders yielding completely different fine-tuning results (**Figures S3-S8**). This finding raises concerns about the reliability and robustness of single-cell analysis methods relying on fine-tuning strategies.

Furthermore, we observed that utilizing secondary metacells as anchoring points enabled precise alignment of cell type annotations across different atlas datasets (**Fig. 4d**). The naming conventions varied substantially among the three atlas datasets, posing a challenge for harmonization. By employing secondary metacells, these cell type annotations could be accurately aligned within the unified cell coordinate system. For instance, macrophages were annotated as "Immune (DC/macrophage)" in the

snCrossTissue dataset, while they were labeled as "macrophage" and "Macrophages" in the TabulaSapiens and ImmuneCrossTissue datasets, respectively. Within the reference cell coordinate system defined by secondary metacells, these differently named macrophages were mapped to the same coordinate points, effectively harmonizing their representation.

These results demonstrate that our proposed hierarchical strategy, based on primary and secondary metacells, can effectively assemble cell atlas datasets. This computational task fundamentally differs from single-cell data integration, which typically involves datasets with relatively balanced and similar cell type compositions. In contrast, cell atlas dataset assembly often involves datasets with significant variations in cell type distributions, presenting unique challenges for integration. SURE effectively addresses this challenge, which is crucial for constructing a comprehensive and unified cell coordinate system. Moreover, metacells serve as a compressed representation of cell atlas datasets, significantly enhancing computational efficiency. By leveraging this compressed representation, our method circumvents the computational limitations that often hinder the integration of large-scale datasets. In this application case, we demonstrated the assembly of three atlas datasets. Notably, our proposed strategy can be generalized to assemble all available cell atlas datasets, offering broad application prospects in the pursuit of a comprehensive cell landscape. This hierarchical approach represents a significant advancement in the field of single-cell genomics, paving the way for the creation of a unified and comprehensive reference database that encompasses the full spectrum of cell diversity.

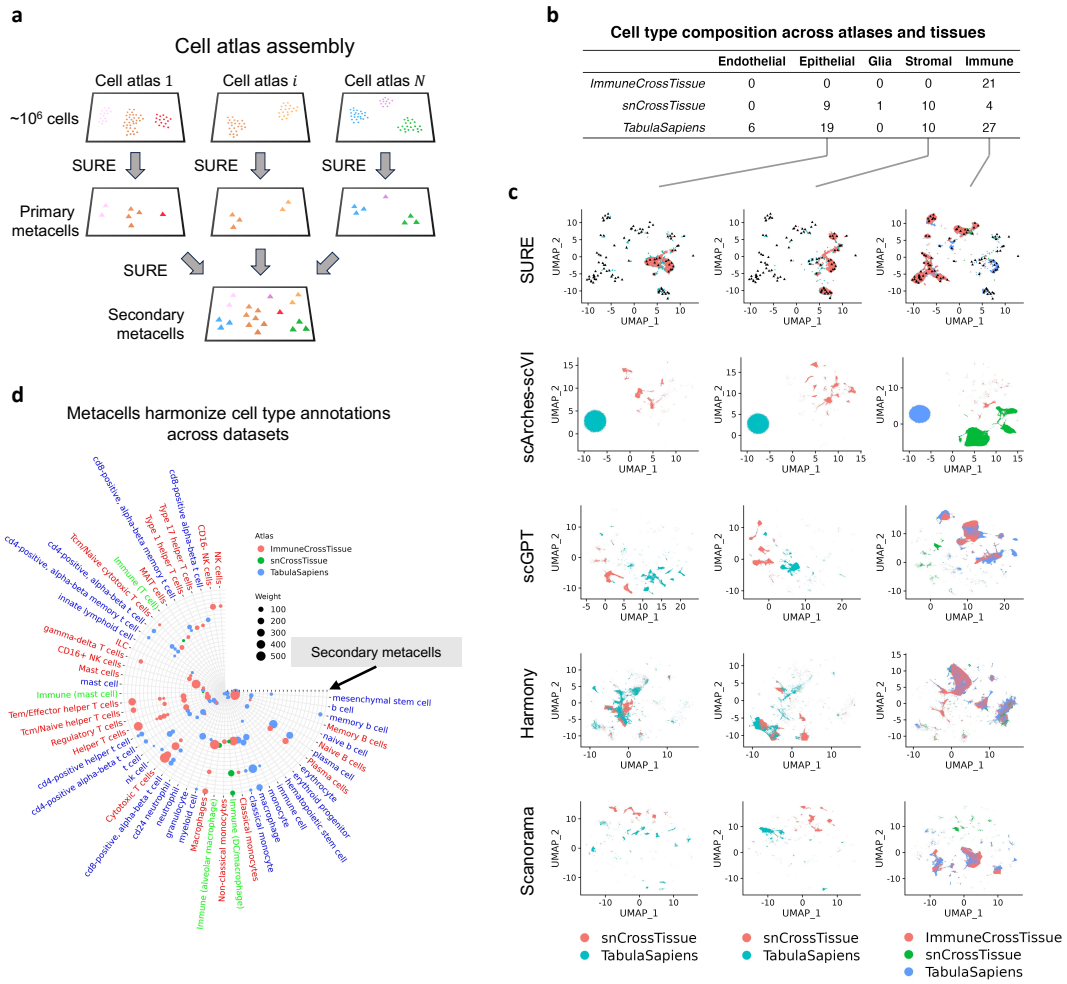

Fig. 4 Assembly of cell atlas datasets. **a,** Metacell-centric hierarchical assembly strategy. **b**, Cell type compositions of three cell atlas datasets being assembled, highlighting the imbalance in cell type distributions. **c**, Batch mixing information in the assembly results of SURE, scArches-scVI, scGPT, Harmony, and Scanorama. **d**, Harmonization of cell type annotations across three atlas datasets achieved by SURE.

HBMCA: Human Blood MetaCell Atlas

As an application, we utilized SURE to establish the Human Blood MetaCell Atlas (HBMCA), a cell coordinate system for the human hematopoietic system. We collected 29 public datasets, covering different types of cell suspensions (i.e., single cells and single nucleus) processed by various single-cell sequencing technologies, including 10x 5' (v1, v2), 10x 3' (v2, v3), Smart-seq2, Seq-Well, Microwell-seq, BD Rhapsody, and ScaleBio. This comprehensive dataset encompasses 1,906 individuals, with a total of 5,310,750 blood cells sequenced (**Fig. 5a**).

Using SURE, we established a reference cell coordinate system consisting of 500 metacells from these five million blood cells based on the overall performance of cell type purity, coverage, and pre-training complexity (**Figures S9-S11**). As shown in **Fig. 5b**, this coordinate system effectively integrates the cell states from the 29 datasets. To evaluate whether this reference coordinate system is sufficient to capture the cell variations in the human hematopoietic system, we conducted a saturation analysis using the following approach. We first built a coordinate system using 10,000 cells and then incrementally increased the number of cells to construct new coordinate systems until all cells were utilized. After each coordinate system was established, we projected the five million cells onto it and calculated the Silhouette score to measure the separation of curated cell types within the coordinate system, thereby assessing its representational power. Through this analysis, we discovered that the representational power of the coordinate system reached its upper limit when the number of cells used for its construction exceeded 850,000. In other words, the reference cell coordinate system we established using over five million cells is capable of capturing the vast majority of blood cell types.

To further analyze the similarities and differences between the metacell-based human blood cell type atlas and existing reference databases for blood cell types, we performed comparisons using MonocoImmuneData and ImmGenData[36]. The former utilized cell sorting techniques to isolate 29 cell types, while the latter isolated 20 cell types, both employing bulk RNA sequencing. Additionally, we utilized single-cell data of sorted immune cells provided by 10x Genomics[37]. We mapped these reference databases onto HBMCA through SURE (**Figs. 5d** and **e**, **Figures S12-S14**). Based on these mapping results, we observed that the four major cell populations in HBMCA corresponded to T/NK cells, B cells, monocytes, and neutrophils in these databases. The verification using bulk RNA sequencing data and single-cell sequencing data demonstrated the accuracy and reliability of our established cell coordinate system. Furthermore, this result also indicated that HBMCA supports the mapping of not only single-cell data but also other types of transcriptomic data (i.e., bulk RNA sequencing).

Simultaneously, we observed that the data-driven HBMCA exhibits clear advantages over traditional cell type reference systems in delineating cell types. On one hand, HBMCA provides a more comprehensive characterization of cell types. Traditional cell type reference systems can only capture a small portion of the blood cell atlas. As depicted in **Fig. 5f**, the 29 sorted human immune cells from MonocoImmuneData

465 mapped to 43 metacells, and based on the proportion of metacells, they are estimated to cover 8.6%
466 of the blood cell atlas. On the other hand, HBMCA offers a more refined characterization of cell types.
467 The resolution of traditional reference systems is limited; among the 29 sorted human immune cells, 10
468 (approximately 34.5%) mapped to 2 metacells, and 3 (approximately 10.3%) mapped to 3 metacells,
469 respectively. Therefore, the data-driven cell coordinate system possesses a more comprehensive ability
470 to define cells.
471

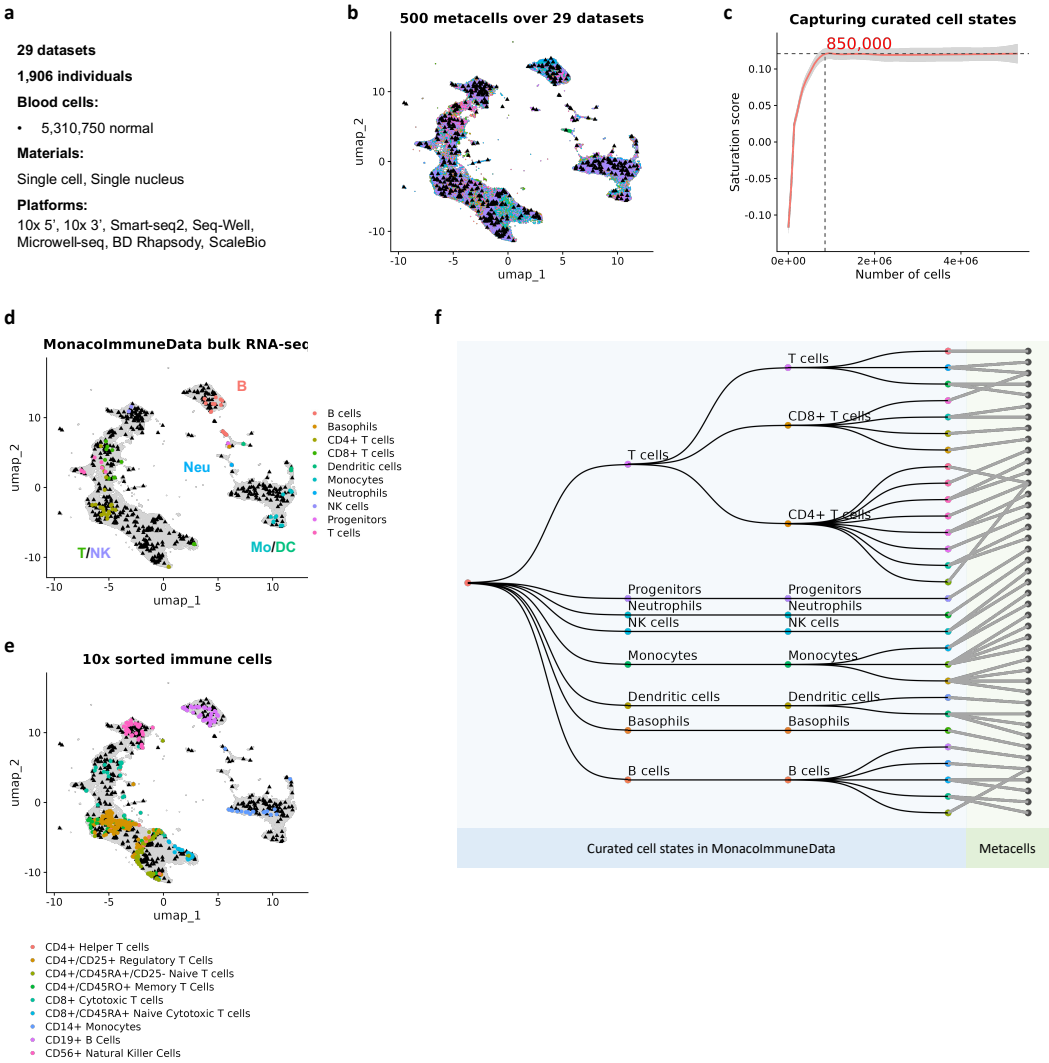

Fig. 5 Human blood metacell atlas, HBMCA. **a**, Statistics of the integrated dataset. **b**, Blood cell coordinate system constructed from metacells, with black triangles representing metacells. **c,** Saturation analysis. **d,** Projection of immune cell states derived from cell sorting and bulk sequencing onto HBMCA, with different cell types identified by different colors. **e**, Projection of sorted immune cells sequenced by 10x Genomics onto HBMCA, with different cell types identified by different colors. **f**, Mapping of curated cell states in the reference database MonacoImmuneData to metacells.

472

473 **HBMCA enables the visualization of variations across studies and illnesses**

474 The establishment of HBMCA represents a significant milestone in single-cell research, enabling the

comparative analysis of single-cell data from diverse studies and diseases within a unified reference coordinate system. HBMCA boasts several notable advantages. Firstly, it stands as the most comprehensive human blood single-cell atlas database to date, providing an unparalleled characterization of the distribution of cell states in the normal human hematopoietic system (**Fig. 6a**). Secondly, HBMCA's exceptional zero-shot capability allows for the direct projection of single-cell data from other studies onto its reference framework, eliminating the need for batch correction or model fine-tuning. This feature renders HBMCA an invaluable resource for investigating the impact of various diseases on the hematopoietic system.

To showcase this functionality, we employed two single-cell datasets focusing on hepatitis B virus (HBV) infection (**Fig. 6b**). The first dataset comprised single cells primarily derived from Asian HBV patients[38], while the second dataset contained single-cell data from European HBV patients[39]. These two studies were conducted independently by different research teams using distinct single-cell technology platforms. By leveraging SURE's universal encoder, we successfully projected these two HBV datasets onto HBMCA. The results unequivocally demonstrated that the single-cell data from these independent studies revealed remarkably similar effects of HBV on the hematopoietic system, with their single-cell findings exhibiting high reproducibility and corroboration.

To further illustrate the versatility of HBMCA, we curated single-cell data from 12 different illnesses, encompassing a wide spectrum of conditions such as dengue virus infection[40], sepsis[41], acute myocardial infarction[42], checkpoint inhibitor-induced colitis[43], juvenile dermatomyositis[44], multiple sclerosis[45], myasthenia gravis[45], systemic lupus erythematosus[45], various types of SARS-CoV-2 infection (including abortive, transient, and sustained infection)[46], and HIV infection[47] (**Fig. 6c**). Utilizing SURE, We directly projected these datasets onto HBMCA, circumventing the need for batch correction or model fine-tuning. Within the unified coordinate system provided by HBMCA, we were able to observe and analyze the changes in cell states of the human hematopoietic system under these diverse disease conditions.

The implications of these findings are profound and exciting. In the realm of genomic research, the existence of a well-established genomic coordinate system has enabled the comparison and replication of genomic studies, fostering rapid advancements and consensus in the field. However, single-cell research has long been hindered by the absence of a comparable reference such as a cell coordinate system, rendering discoveries largely study-specific and limiting the ability of achieve consensus scientific breakthroughs. The establishment of HBMCA, with its robust reference coordinate system and powerful zero-shot cell mapping functionality, has a potential to revolutionize the field of single-cell research. By providing a unified framework for the integration and comparison of single-cell data across diverse studies and diseases, HBMCA paves the way for groundbreaking discoveries and advancements akin to those witnessed in genomic research.

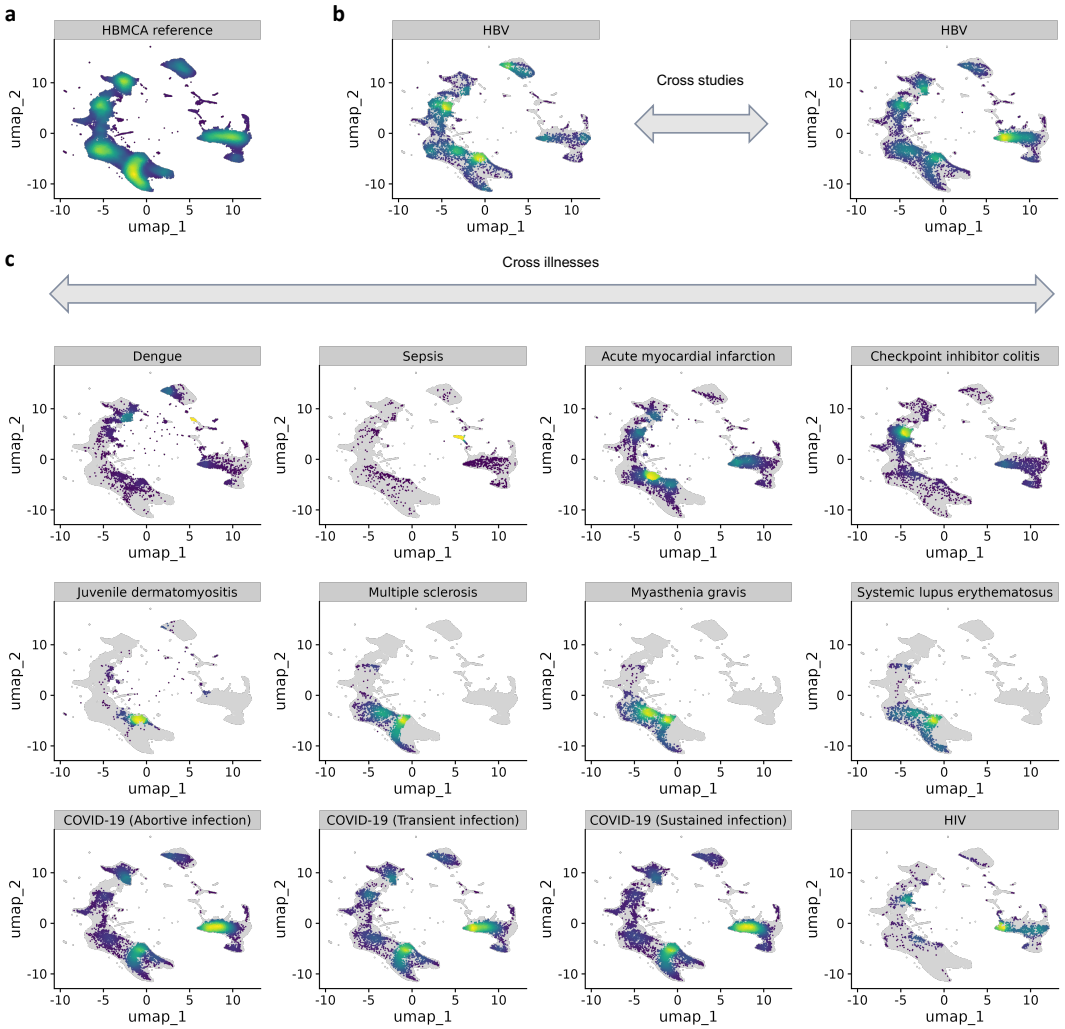

Fig. 6 Zero-shot mapping of cells across studies and illnesses onto HBMCA. **a**, Distribution of cells collected from normal individuals. **b**, Illustration of cell distributions across two HBV studies. **c**, Illustration of cell distributions across twelve disease conditions.

Discussion

The genome coordinate systems serve as a vital foundation for genomic research, enabling the direct mapping and comparison of genomic data generated by different laboratories worldwide. This zero-shot mapping process, which does not require fine-tuning, has greatly facilitated the advancement of genomics. Similarly, single-cell omics research urgently requires a coordinate system to provide a unified reference for various research efforts, holding profound implications for understanding and leveraging cell atlas datasets.

In this study, we proposed the utilization of metacells to establish a unified coordinate system for single cell IDs. We developed SURE, which employs vector quantization and deep generative modeling techniques to learn a set of metacells from single-cell transcriptome data. These metacells accurately characterize the distribution of cell states in the data, forming a coordinate system with excellent precision. This coordinate system provides a universal encoder that supports zero-shot mapping of single-cell query data without requiring batch correction or model fine-tuning, even being used by low

quality query data.

Prior to our work, reference systems for single-cell omics research primarily relied on two approaches: cell type annotation databases and cell manifolds. Cell type annotation represents an explicit cell coordinate system with curated reference cell types but faces limitations such as incomplete representations, uncharacterized cell types, lack of unified nomenclature, and fixed resolution in investigating the cell state space. Cell manifolds, employed in unsupervised algorithms, provide an implicit definition of the cell coordinate system learned from the data. However, cell coordinate systems represented by manifolds reconstructed from different datasets are inherently local and may differ from one another. Moreover, the presence of technical artifacts and variations in cell type compositions renders the identification of precise corresponding landmarks across distinct cell manifolds extremely challenging. Compared to these reference systems, the metacell-based reference coordinate system proposed in this study offers clear advantages. It is a data-driven coordinate system unaffected by the limitations of existing cell type annotations and effectively addresses the impact of batch effects and cell type imbalance. Even in extreme cases with low sequencing quality, the metacell-based reference coordinate system accurately reveals the cell type distribution in the data. We demonstrate the influence of cell type imbalance on existing single-cell integration methods, model fine-tuning strategies, and large models, highlighting the advantages of our proposed strategy.

There are a few computational methods for defining metacells. Due to different purpose in developing the tools, it is obvious that they are unsuitable for establishing a coordinate system for the cell ID. These methods do not handle batch effects and single-cell sequencing noise and do not support the projection of new single-cell data to the set of metacells. The detailed comparison between SURE and these metacell calling algorithms shows that SURE is an excellent tool in establishing a reference system for cell ID.

Our work may have limitations, primarily in learning the codebook for neural vector quantization. The stochastic strategy employed may result in the learning process failing to allocate sufficient weight to rare cell types, potentially overlooking or underestimating their composition in the original data. Which might be partially overcome by incorporating prior annotation information and employing supervised methods to avoid such occurrences.

In conclusion, SURE represents a powerful tool for exploring the intricate landscape of cell identities, allowing the establishment of an unified reference system for cell IDs using single cell transcriptome datasets. When metacells were set as landmarks, SURE offers exceptional performance, scalability, and the ability to harmonize datasets with varying cell type compositions, and enables consistent positioning and integration of cells across diverse datasets. Our approach has built a coordinate system for cell ID, which shall pave the way for elucidating the regulatory networks that shape cell phenotypes, ultimately contributing to the development of more effective diagnostic and therapeutic strategies.

Acknowledgements

This work is supported in part by the Natural Science Foundation of Xiamen, China (No: 3502Z202373020), National Natural Science Foundation of China (82388201, 61503314), Natural Science Foundation of Fujian Province (2019J01041), the National Key R&D Program of China

(2020YFA0803500), the CAMS Innovation Fund for Medical Sciences (2019-I2M-5-062).

Author contributions

F.Z. and J.H. conceived the concept and supervised the project. F.Z. designed the model. F.Z. wrote the code. F.Z. performed the experiments. F.Z. and J.H. analyzed the experimental results. F.Z and J.H. drafted the manuscript.

Declaration of interests

All the authors declare no competing interests.

Supplemental information

Figures S1-S14 and Table S1

References

1. Fleck, J.S., Camp, J.G., and Treutlein, B. (2023). What is a cell type? Science (80-.). *381*, 733–734. https://doi.org/10.1126/science.adf6162.
2. Zeng, H. (2022). What is a cell type and how to define it? Cell *185*, 2739–2755. https://doi.org/10.1016/j.cell.2022.06.031.
3. What Is Your Conceptual Definition of “Cell Type” in the Context of a Mature Organism? (2017). Cell Syst. *4*, 255–259. https://doi.org/10.1016/j.cels.2017.03.006.
4. Theriot, J.A., Simonsen, A., Tolić, I., Leonetti, M.D., Mayor, S., Bassereau, P., Paluch, E.K., Han, J., Covert, M.W., Mizushima, N., et al. (2024). Cell biology is…. Cell *187*, 219–224. https://doi.org/https://doi.org/10.1016/j.cell.2023.12.020.
5. Kharchenko, P. V. (2021). The triumphs and limitations of computational methods for scRNA-seq. Nat. Methods *18*, 723–732. https://doi.org/10.1038/s41592-021-01171-x.
6. Bilous, M., Hérault, L., Gabriel, A.A.G., Teleman, M., and Gfeller, D. (2024). Building and analyzing metacells in single-cell genomics data. Mol. Syst. Biol., 1–23. https://doi.org/https://doi.org/10.1038/s44320-024-00045-6.
7. Baran, Y., Bercovich, A., Sebe-Pedros, A., Lubling, Y., Giladi, A., Chomsky, E., Meir, Z., Hoichman, M., Lifshitz, A., and Tanay, A. (2019). MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biol. *20*, 206. https://doi.org/10.1186/s13059-019-1812-2.
8. Ben-Kiki, O., Bercovich, A., Lifshitz, A., and Tanay, A. (2022). Metacell-2: a divide-and-conquer metacell algorithm for scalable scRNA-seq analysis. Genome Biol. *23*, 100. https://doi.org/10.1186/s13059-022-02667-1.
9. Bilous, M., Tran, L., Cianciaruso, C., Gabriel, A., Michel, H., Carmona, S.J., Pittet, M.J., and Gfeller, D. (2022). Metacells untangle large and complex single-cell transcriptome networks. BMC Bioinformatics *23*, 336. https://doi.org/10.1186/s12859-022-04861-1.
10. Persad, S., Choo, Z.N., Dien, C., Sohail, N., Masilionis, I., Chaligné, R., Nawy, T., Brown, C.C., Sharma, R., Pe’er, I., et al. (2023). SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. Nat. Biotechnol., 2022.04.02.486748. https://doi.org/10.1038/s41587-023-01716-9.
11. Ben-Kiki, O., Bercovich, A., Lifshitz, A., Raz, O., Brook, D., and Tanay, A. (2023). MCProj: metacell projection for interpretable and quantitative use of transcriptional atlases. Genome

Biol. *24*, 220. https://doi.org/10.1186/s13059-023-03069-7.

12. Gray, R. (1984). Vector quantization. IEEE ASSP Mag. *1*, 4–29. https://doi.org/10.1109/MASSP.1984.1162229.

13. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. Nat. Commun. *9*, 284. https://doi.org/10.1038/s41467-017-02554-5.

14. Xiong, L., Tian, K., Li, Y., Ning, W., Gao, X., and Zhang, Q.C. (2022). Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. Nat. Commun. *13*, 6118. https://doi.org/10.1038/s41467-022-33758-z.

15. Lance, C., Luecken, M.D., Burkhardt, D.B., Cannoodt, R., Rautenstrauch, P., Laddach, A., Ubingazhibov, A., Cao, Z.-J., Deng, K., Khan, S., et al. (2022). Multimodal single cell data integration challenge: Results and lessons learned. In Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track Proceedings of Machine Learning Research., D. Kiela, M. Ciccone, and B. Caputo, eds. (PMLR), pp. 162–176.

16. Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. *9*. https://doi.org/10.1038/s41598-019-41695-z.

17. Caliński, T., and Harabasz, J. (1974). A dendrite method for cluster analysis. Commun. Stat. *3*, 1–27. https://doi.org/10.1080/03610927408827101.

18. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888-1902.e21. https://doi.org/https://doi.org/10.1016/j.cell.2019.05.031.

19. Welch, J.D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E.Z. (2019). Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell *177*, 1873-1887.e17. https://doi.org/https://doi.org/10.1016/j.cell.2019.05.006.

20. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296. https://doi.org/10.1038/s41592-019-0619-0.

21. Hie, B., Bryson, B., and Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat. Biotechnol. *37*, 685–691. https://doi.org/10.1038/s41587-019-0113-3.

22. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058. https://doi.org/10.1038/s41592-018-0229-2.

23. De Donno, C., Hediyeh-Zadeh, S., Moinfar, A.A., Wagenstetter, M., Zappia, L., Lotfollahi, M., and Theis, F.J. (2023). Population-level integration of single-cell datasets enables multi-scale analysis across samples. Nat. Methods *20*, 1683–1692. https://doi.org/10.1038/s41592-023-02035-2.

24. Yu, X., Xu, X., Zhang, J., and Li, X. (2023). Batch alignment of single-cell transcriptomics data using deep metric learning. Nat. Commun. *14*, 960. https://doi.org/10.1038/s41467-023-36635-5.

25. Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. (2024). scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat. Methods. https://doi.org/10.1038/s41592-024-02201-0.

26. Luecken, M.D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M.F., Strobl, D.C., Zappia, L., Dugas, M., Colomé-Tatché, M., et al. (2022). Benchmarking atlas-level data integration in single-cell genomics. Nat. Methods *19*, 41–50. https://doi.org/10.1038/s41592-021-01336-8.
27. Lotfollahi, M., Naghipourfar, M., Luecken, M.D., Khajavi, M., Büttner, M., Wagenstetter, M., Avsec, Ž., Gayoso, A., Yosef, N., Interlandi, M., et al. (2021). Mapping single-cell data to reference atlases by transfer learning. Nat. Biotechnol. https://doi.org/10.1038/s41587-021-01001-7.
28. Hao, Y., Stuart, T., Kowalski, M.H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., et al. (2023). Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat. Biotechnol. https://doi.org/10.1038/s41587-023-01767-y.
29. Hie, B., Cho, H., DeMeo, B., Bryson, B., and Berger, B. (2019). Geometric Sketching Compactly Summarizes the Single-Cell Transcriptomic Landscape. Cell Syst. *8*, 483-493.e7. https://doi.org/https://doi.org/10.1016/j.cels.2019.05.003.
30. Jones, R.C., Karkanias, J., Krasnow, M.A., Pisco, A.O., Quake, S.R., Salzman, J., Yosef, N., Bulthaup, B., Brown, P., Harper, W., et al. (2022). The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. Science (80-.). *376*, eabl4896. https://doi.org/10.1126/science.abl4896.
31. Lotfollahi, M., Naghipourfar, M., Theis, F.J., and Wolf, F.A. (2020). Conditional out-of-distribution generation for unpaired data using transfer VAE. Bioinformatics *36*, i610–i617. https://doi.org/10.1093/bioinformatics/btaa800.
32. Kang, J.B., Nathan, A., Weinand, K., Zhang, F., Millard, N., Rumker, L., Moody, D.B., Korsunsky, I., and Raychaudhuri, S. (2021). Efficient and precise single-cell reference atlas mapping with Symphony. Nat. Commun. *12*, 5890. https://doi.org/10.1038/s41467-021-25957-x.
33. Maan, H., Zhang, L., Yu, C., Geuenich, M.J., Campbell, K.R., and Wang, B. (2024). Characterizing the impacts of dataset imbalance on single-cell data integration. Nat. Biotechnol. https://doi.org/10.1038/s41587-023-02097-9.
34. C., D.C., C., X., B., J.L., B., R.D., B., W.S., T., G., K., H.S., O., S., K., P., W., K.H., et al. (2022). Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science (80-.). *376*, eabl5197. https://doi.org/10.1126/science.abl5197.
35. Eraslan, G., Drokhlyansky, E., Anand, S., Subramanian, A., Fiskin, E., Slyper, M., Wang, J., Wittenberghe, N. Van, Rouhana, J.M., Waldman, J., et al. (2021). Single-nucleus cross-tissue molecular reference maps to decipher disease gene function. bioRxiv, 2021.07.19.452954. https://doi.org/10.1101/2021.07.19.452954.
36. Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. *20*, 163–172. https://doi.org/10.1038/s41590-018-0276-y.
37. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049. https://doi.org/10.1038/ncomms14049.
38. Zhang, C., Li, J., Cheng, Y., Meng, F., Song, J.-W., Fan, X., Fan, H., Li, J., Fu, Y.-L., Zhou, M.-J., et al. (2023). Single-cell RNA sequencing reveals intrahepatic and peripheral immune

705 characteristics related to disease phases in HBV-infected patients. Gut *72*, 153 LP – 167.
706 https://doi.org/10.1136/gutjnl-2021-325915.
707 39. Beudeker, B.J.B., Osmani, Z., van Oord, G.W., Groothuismink, Z.M.A., de Knegt, R.J.,
708 Hoogenboezem, R.M., Bindels, E.M.J., van de Werken, H.J.G., and Boonstra, A. (2024).
709 Association of HBsAg levels with differential gene expression in NK, CD8 T, and memory B cells
710 in treated patients with chronic HBV. JHEP Reports *6*.
711 https://doi.org/10.1016/j.jhepr.2023.100980.
712 40. Ghita, L., Yao, Z., Xie, Y., Duran, V., Cagirici, H.B., Samir, J., Osman, I., Rebellón-Sánchez, D.E.,
713 Agudelo-Rojas, O.L., Sanz, A.M., et al. (2023). Global and cell type-specific immunological
714 hallmarks of severe dengue progression identified via a systems immunology approach. Nat.
715 Immunol. *24*, 2150–2163. https://doi.org/10.1038/s41590-023-01654-3.
716 41. Kwok, A.J., Allcock, A., Ferreira, R.C., Cano-Gamez, E., Smee, M., Burnham, K.L., Zurke, Y.-X.,
717 Novak, A., Darwent, M., Baron, T., et al. (2023). Neutrophils and emergency granulopoiesis
718 drive immune suppression and an extreme response endotype during sepsis. Nat. Immunol.
719 *24*, 767–779. https://doi.org/10.1038/s41590-023-01490-5.
720 42. van Blokland, I. V, Oelen, R., Groot, H.E., Benjamins, J.W., Pekayvaz, K., Losert, C.,
721 Knottenberg, V., Heinig, M., Nicolai, L., Stark, K., et al. (2024). Single-Cell Dissection of the
722 Immune Response After Acute Myocardial Infarction. Circ. Genomic Precis. Med. *0*, e004374.
723 https://doi.org/10.1161/CIRCGEN.123.004374.
724 43. Thomas, M.F., Slowikowski, K., Manakongtreecheep, K., Sen, P., Samanta, N., Tantivit, J.,
725 Nasrallah, M., Zubiri, L., Smith, N.P., Tirard, A., et al. (2024). Single-cell transcriptomic analyses
726 reveal distinct immune cell contributions to epithelial barrier dysfunction in checkpoint
727 inhibitor colitis. Nat. Med. *30*, 1349–1362. https://doi.org/10.1038/s41591-024-02895-x.
728 44. Rabadam, G., Wibrand, C., Flynn, E., Hartoularos, G.C., Sun, Y., Madubata, C., Fragiadakis, G.K.,
729 Ye, J., Kim, S., Gartner, Z.J., et al. (2024). Coordinated immune dysregulation in Juvenile
730 Dermatomyositis revealed by single-cell genomics. JCI Insight.
731 https://doi.org/10.1172/jci.insight.176963.
732 45. Yasumizu, Y., Takeuchi, D., Morimoto, R., Takeshima, Y., Okuno, T., Kinoshita, M., Morita, T.,
733 Kato, Y., Wang, M., Motooka, D., et al. (2024). Single-cell transcriptome landscape of
734 circulating CD4+ T cell populations in autoimmune diseases. Cell Genomics *4*, 100473.
735 https://doi.org/https://doi.org/10.1016/j.xgen.2023.100473.
736 46. Lindeboom, R.G.H., Worlock, K.B., Dratva, L.M., Yoshida, M., Scobie, D., Wagstaffe, H.R.,
737 Richardson, L., Wilbrey-Clark, A., Barnes, J.L., Polanski, K., et al. (2023). Human SARS-CoV-2
738 challenge resolves local and systemic response dynamics. medRxiv, 2023.04.13.23288227.
739 https://doi.org/10.1101/2023.04.13.23288227.
740 47. Kazer, S.W., Aicher, T.P., Muema, D.M., Carroll, S.L., Ordovas-Montanes, J., Miao, V.N., Tu,
741 A.A., Ziegler, C.G.K., Nyquist, S.K., Wong, E.B., et al. (2020). Integrated single-cell analysis of
742 multicellular immune dynamics during hyperacute HIV-1 infection. Nat. Med. *26*, 511–518.
743 https://doi.org/10.1038/s41591-020-0799-2.
744
745

STAR Methods

SURE

SURE provides an efficient method for integrating various single-cell datasets into a comprehensive cell atlas while defining a set of standard cell states or "metacells". These metacells serve as robust reference landmarks, constituting a unified cell coordinate system for consistent cell positioning and annotations across a variety of datasets. SURE offers remarkable computational efficiency, enabling scalable analyses of large-scale single-cell data. Furthermore, metacells assure robust representations of cell diversity and effectively counteract data imbalance. Also, by leveraging the unified coordinate system, SURE systematically eliminates technical noises like batch effects, simplifying data integration and enabling meaningful biological comparisons across multiple datasets.

The underpinning of SURE lies in a variational Bayesian framework, whereby it integrates neural vector quantization and generative modeling techniques. These crucial aspects will be addressed in the following sections.

Asymmetric Variational Bayesian Framework. Bayesian analysis forms an effective strategy for uncovering the true cell states in single-cell data. According to our model, the expression data of a cell is x, batch information is b, and its cell state is z. SURE's objective is to create a unified cell coordinate system for the unified cell state space. Given the data x, SURE can accurately deduce the cell state z in a unified context through the inference model $q(z|x)$. Notably, the omisson of batch information b in $q(z|x)$ significantly widens the method's applicability, indicating that, once adequately trained, the model can be universally deployed on new datasets without the necessity of batch information. Consequently, the inference of cell states is made batch-independent, aligning $q(z|x)$ with $p(z|x,b)$, and ensuring consistent, robust identification of cell types across diverse datasets.

The quest to realize a universal $q(z|x)$ methodology presents considerable challenges. Single-cell data, being continuous and high-dimensional, results in an exponentially vast solution space for z. Additionally, the high sparsity of the data complicates inference, and batch effects often obscure true biological distinctions. These challenges are tackled by borrowing insights from the scVI's generative model, implementing the Bayesian equation:

$$p(z|x,b) = \frac{p(x|z,b)p(z)}{p(x|b)}.$$

This equation provides a robust framework for deciphering z by contemplating the observed data x and batch information b, then going through the forward modeling of data generation $p(x|z,b)$.

To successfully establish a universal $q(z|x)$, we employ it as a variational approximation of $p(z|x,b)$. This adaptation separates the inference and the generation models into two individual components: the generation model processes the cell state z and batch information b to produce x, thereby estimating $p(z|x,b)$, while the inference model $q(z|x)$ aims to approximate $p(z|x,b)$ as closely as possible, i.e., $q(z|x) \approx p(z|x,b)$. We propose that this approximation is validated when the training data provide sufficient information about batch variations. Fortunately, over the past decade, tremendous amounts of data have been generated globally through various single-cell sequencing technologies, covering all possible batch variations.

Thus, SURE exhibits a distinct and asymmetric structure. While batch information b remains a deterministic variable in the generation process, SURE trains the inference model $q(z|x)$ using purely x, thereby sidelining batch information and reinforcing the model's universal applicability.

Neural Vector Quantization. To manage the high-dimensional, continuous nature of the cell state z, we implement neural vector quantization to simplify the complexity of the solution space. We hypothesize a codebook containing K landmarks within the cell state space $Z = \{z\}$, designed to comprehensively cover Z with adequate resolution. Each cell's state z can then be described by its proximity to the nearest landmark i, using the probabilistic model $p(i)p(z|i)$:

$$
\begin{aligned}
Codebook &= NeuralNetwork(I) \\
i &\sim OneHotDirichlet(\alpha) \\
\mu_i, \Sigma_i &= Codebook(i) \\
z &\sim MultivariateNormal(\mu_i, \Sigma_i)
\end{aligned}
$$

Here, I is a unit matrix denoting all landmarks, and α is a prior distribution over K landmarks. In our implementation, α is the noninformative prior. In SURE, a multi-layer neural network functions as the codebook for vector quantization, capable of outputting the mean vector and covariance matrix for the multivariate normal distribution associated with each landmark.

Our approach diverges from that of VQ-VAE in several significant ways: Firstly, instead of a static data structure, our codebook is represented through a dynamic multi-layer neural network, offering enhanced flexibility and representational power. Secondly, while VQ-VAE uses discrete points in its codebook to form a discrete representation of data distribution, our neural network-based codebook describes the data probability density through a superposition of K discrete normal distributions centered at these landmarks. Lastly, we incorporate Bayesian inference for learning the codebook, as opposed to the EM algorithm typically employed in VQ-VAE, enhancing our model's capacity to capture complex data structures in a probabilistically coherent framework.

Generative Modeling. Combining the Bayesian framework and neural vector quantization technique, we can outline the generative model of SURE as follows:

1. A metacell, i.e., an element from the codebook, is randomly selected as the initial state i for the cell.
2. The cellular state z, devoid of batch effects, is generated based on the probability distribution of the chosen metacell.
3. The cellular state $[z, b]$ with batch effects is obtained by integrating the batch factor b and augmenting z.
4. The cellular data x is generated based on the cellular state $[z, b]$ through the zero-inflated negative binomial (ZINB) model.

The above process can be encapsulated in the following probabilistic model:

$$p(i)p(z|i)p(x|z, b).$$

The variational model aligned with the above generative process is:

$$q(z|x)q(i|z).$$

We apply the Pyro probabilistic programming language to implement both the generative and

variational models. To optimize model parameter learning, we utilize the stochastic variational technique.

Zero-shot cell mapping

SURE capitalizes on metacells to define a unified cell positioning system for cell atlases, proposing two different modes of cell positioning depending on the desired applications.

The first mode uses metacells as reference points, positioning cells based on their similarity to these metacells, quantified by $q(i|z)$. This positioning strategy presumes the cell atlas already encompasses all cell states potentially present in the query data, making it well-suited for tasks like cell positioning and annotations, as opposed to discovering novel cell populations.

Conversely, the second mode employs SURE's encoder, $q(z|x)$, for cell positioning. As previously discussed, once sufficient batch variation data has been included during training, SURE's encoder can effectively distinguish genuine biological differences. Thus, $q(z|x)$ is adept at extracting relevant biological signals from the query data, rendering it as an ideal tool for biological discovery scenarios, such as discerning differences between normal and disease states.

Cell atlas assembly

To enable efficient integration and assembly of large-scale cell atlas datasets, we introduce a hierarchical metacell-based assembly method. This method employs a divide-and-conquer computational strategy, with the following steps:

1. First, we use SURE to compute metacells for each individual dataset, obtaining a collection of primary metacells.
2. We then apply SURE again to compute secondary metacells from the collection of primary metacells.

In this study, we successfully leveraged this hierarchical assembly method to assemble imbalanced atlas-scale single-cell datasets and achieve large-scale integration and assembly of the adult human hematopoietic cell atlas.

Public single-cell datasets used in the study

In this work, we exclusively utilized public datasets that are readily accessible and downloadable from the internet, ensuring transparency and reproducibility of our findings. To evaluate the performance of metacell calling algorithms, we primarily employed two datasets. The first dataset was provided by the single-cell multimodal data integration competition organized by the 2021 NeurIPS conference. This dataset, generated using the cutting-edge CITE-seq technology, can be obtained from the NCBI GEO database under the accession number GSE194122. The dataset comprises two modalities: gene expression and protein expression. In our work, we focused solely on the gene expression data, as it aligns with our research objectives and allows for a more targeted analysis. To assess the query cell mapping, we utilized the cross-tissue cell atlas dataset TabulaSapiens, provided by the Human Cell Atlas project. This comprehensive dataset serves as a valuable resource for understanding cell heterogeneity across different tissues. Specifically, we focused on the immune compartment data from TabulaSapiens for this evaluation task. In the subsequent atlas assembly task, we employed the entire TabulaSapiens dataset, which primarily consists of RNA transcripts from individual cells, enabling us to explore the full

spectrum of cell diversity. We incorporated two additional atlas datasets in the atlas assembly tasks. One of these datasets is scCrossTissue, another cross-tissue cell atlas dataset that primarily comprises RNA transcripts from individual cell nuclei. This dataset offers a complementary perspective on cell heterogeneity and allows for the exploration of nuclear RNA profiles. The other dataset, ImmuneCrossTissue, focuses specifically on profiling immune cells across multiple human tissues.

For the establishment of HBMCA, a comprehensive resource for studying the human blood mononuclear cell landscape, the atlas datasets utilized, along with their detailed data information and acquisition methods, are provided in the **Table S1**.

To validate the utility and effectiveness of HBMCA, we employed data from 13 different diseases, spanning a wide range of pathological conditions. These datasets were obtained from the latest research articles and are openly accessible in public databases, promoting collaboration and facilitating further investigations by the scientific community. Notably, the data from HBV studies on Asian and European populations can be accessed from the NCBI GEO database under the accession numbers GSE182159 and GSE247322, respectively. In conducting the zero-shot mapping and visualization analysis of single-cell data across multiple diseases, we obtained the datasets from various reputable sources to ensure the reliability and quality of the data. The datasets for dengue virus infection and checkpoint inhibitor-induced colitis were acquired from the NCBI GEO database. These datasets can be accessed using the accession numbers GSE220969 and GSE206298, respectively. To obtain the sepsis dataset, we utilized the Zenodo platform. The dataset can be retrieved using the following DOI: https://doi.org/10.5281/zenodo.7723202. For the acute myocardial infarction data, we accessed the dataset through the eQTLGen Consortium's website, which can be found at https://eqtlgen.org/sc/datasets/blokland2024.html. The juvenile dermatomyositis dataset was obtained from the CZ CELLxGENE Discover resource, a platform that facilitates the exploration and analysis of single-cell data. The dataset is available at https://cellxgene.cziscience.com/collections/c672834e-c3e3-49cb-81a5-4c844be4a975. To acquire the datasets for multiple sclerosis, myasthenia gravis, systemic lupus erythematosus, and HIV infection, we utilized the Single Cell Portal, a centralized repository for single-cell genomics data. The specific datasets can be accessed using the accession numbers SCP1963 and SCP256. Lastly, the COVID-19 dataset was obtained from the COVID-19 Cell Atlas website (https://www.covid19cellatlas.org/), which serves as a valuable resource for researchers studying the cellular and molecular mechanisms of SARS-CoV-2 infection.

Comparison methods

In this study, we employed a comprehensive suite of computational tools to conduct a rigorous comparative analysis, categorizing them based on the specific analytical tasks they address. Firstly, for the task of metacell calling, we utilized state-of-the-art algorithms such as SuperCell and SEACells. These tools have been specifically designed to identify and delineate metacells, which represent distinct cell states or subtypes within a heterogeneous population. Additionally, we leveraged the Leiden clustering algorithm provided by the widely-used Python package Scanpy to compute pseudocells, enabling a complementary approach to cell subpopulation identification. By employing multiple metacell calling methods, we ensure a robust and comprehensive analysis of the cell landscape.

Secondly, for the task of data integration and batch effect removal, we employed several commonly used algorithms that have demonstrated their effectiveness in harmonizing datasets from various sources while mitigating technical variations. These algorithms include scVI, Seurat, LIGER, scPoli, Harmony, SCALEX, Scanorama, scGPT, and scDML. Each of these tools employs unique strategies to align and integrate datasets, accounting for batch effects and enabling the identification of shared and dataset-specific cell states. By utilizing a diverse range of integration methods, we aim to provide a comprehensive and unbiased assessment of their performance and suitability for different datasets and biological questions.

Thirdly, for the task of query cell mapping, we utilized advanced reference mapping tools such as scArches, Symphony, and SCALEX. These tools enable the accurate mapping of query cells to reference datasets, facilitating the identification and characterization of cell types and states in the query data. By leveraging the information from well-annotated reference datasets, query cell mapping allows for the transfer of knowledge and the discovery of novel cell phenotypes.

Lastly, for the complex task of atlas assembly, we primarily employed scArches, scGPT, Harmony, and Scanorama. These tools have been specifically designed to integrate and harmonize multiple datasets, enabling the construction of comprehensive cell atlases that capture the full spectrum of cell diversity across different conditions, tissues, or species.

To ensure a fair and unbiased evaluation of all computational tools, we strictly adhered to the default settings provided in their respective official tutorials. This approach guarantees that our comparative analysis reflects the performance of each tool under standardized conditions, minimizing the influence of user-specific optimizations or subjective parameter tuning.

Evaluation metrics

To assess the performance of metacell calling and data integration algorithms, we utilized the evaluation functions provided by the Python packages scikit-learn and scIB-metrics. Scikit-learn is a widely-used machine learning library that offers a comprehensive set of tools for data preprocessing, feature selection, model evaluation, and performance metrics. By leveraging the evaluation functions from scikit-learn, we were able to quantitatively measure the quality and accuracy of the metacell calling results. These functions allow for the computation of various metrics, such as homogeneity score, silhouette score, and Davies-Bouldin score, which provide insights into the coherence and distinctiveness of the identified metacells.

In addition to scikit-learn, we employed scIB-metrics, a specialized library designed for evaluating the performance of single-cell data integration methods. scIB-metrics provides a suite of evaluation functions specifically tailored to assess the quality of data integration results in the context of single-cell genomics. These functions take into account the unique challenges and characteristics of single-cell data, such as high dimensionality, sparsity, and biological variability. By utilizing scIB-metrics, we were able to comprehensively evaluate the effectiveness of different data integration algorithms in terms of their ability to align and harmonize datasets from multiple sources, preserve biological variation, and reduce technical noise.

The combination of scikit-learn and scIB-metrics enables a rigorous and multi-faceted evaluation of the computational results obtained from metacell calling and data integration algorithms. By leveraging these well-established and widely-used libraries, we ensure that our assessments are based on standardized and reliable metrics, facilitating the comparison and interpretation of results across different methods and datasets. The use of these evaluation functions not only provides quantitative measures of performance but also allows for the identification of strengths and limitations of each algorithm, guiding the selection and refinement of computational approaches for single-cell data analysis.

Code availability

SURE are available for academic use on GitHub at https://github.com/ZengFLab/SURE.