Deciphering Cell Types by Integrating scATAC-seq Data with Genome Sequences

Yuedong Yang

yangyd25@mail.sysu.edu.cn

Sun Yat-sen University https://orcid.org/0000-0002-6782-2813

Yuansong Zeng

Chongqing University

Mai Luo

Sun Yat-sen University

Ningyuan Shangguan

Sun Yat-sen University

Peiyu Shi

Sun Yat-sen University

Junxi Feng

Sun Yat-sen University

Jin Xu

Sun Yat-sen University

Ken Chen

Sun Yat-sen University

Yutong Lu

Sun Yat-sen University

Weijiang Yu

Sun Yat-sen University

Article

Keywords:

Posted Date: March 18th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3539732/v1

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Computational Science on April 10th, 2024. See the published version at https://doi.org/10.1038/s43588-024-00622-7.

Deciphering Cell Types by Integrating scATAC-seq Data with Genome Sequences

Yuansong Zeng[1,2#], Mai Luo[2,#], Ningyuan Shangguan[2], Peiyu Shi[3], Junxi Feng[2], Jin Xu[3], Ken, Chen[2], Yutong Lu[2], Weijiang Yu[2], and Yuedong Yang[2,4*]

[1]School of Big Data & Software Engineering, Chongqing University, Chongqing 400000, China.

[2]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510000, China.

[3]State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, Guangzhou, 510275, China.

[4]Key Laboratory of Machine Intelligence and Advanced Computing (MOE), Guangzhou 510000, China.

* Corresponding authors.

E-mail addresses: yangyd25@mail.sysu.edu.cn (Yuedong Yang)

[#] These authors contributed equally to this work.

ABSTRACT

The single cell ATAC sequencing (scATAC-seq) technology provides insight into gene regulation and epigenetic heterogeneity at single-cell resolution, but cell annotation from scATAC-seq remains challenging due to high dimensionality and extreme sparsity within the data. Existing cell annotation methods mostly focused on cell peak matrix without fully utilizing the underlying genomic sequence. Here, we propose a method, SANGO, for accurate **s**ingle cell **an**notation by integrating **g**en**o**me sequences around the accessibility peaks within scATAC data. The genome sequences of peaks are encoded into low-dimensional embeddings, and then iteratively used to reconstruct the peak stats of cells through a fully-connected network. The learned weights are considered as regulatory modes to represent cells, and utilized to align the query cells and the annotated cells in the reference data through a graph transformer network for cell annotations. SANGO was demonstrated to consistently outperform competing methods on 55 paired scATAC-seq datasets across samples, platforms, and tissues. SANGO was also shown able to detect unknown tumor cells through attention edge weights learned by graph transformer. Moreover, according to the annotated cells, we found cell type-specific peaks that provide functional insights/ biological signals through expression enrichment analysis, cis-regulatory chromatin interactions analysis, and motif enrichment analysis.

The single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) technologies provides great opportunities for many biological applications, including detecting cell heterogeneity and regulatory elements [1], reconstructing differentiation trajectories [2], and identifying biological mechanisms for complex diseases[3]. One of the most fundamental problems in scATAC-seq data analysis is cell type identification, which is essential for comprehending the intricate composition of complex tissues and uncovering unknown cell types. Currently, a popular strategy is to cluster cells and then annotate the cell clusters through the peaks corresponding signature gene [4]. This process is cumbersome and complex involving professional experts. With the rapid increase of well-characterized public scATAC-seq datasets[5], it is promising to use the well labeled cells to automatically annotate newly generated datasets.

Since scATAC data involves inherently high dimensionality of accessible peaks and sparsity of sequencing reads per cell [6], numerous approaches [7] have been developed to transform scATAC-seq data into synthetic scRNA-seq data by estimating "gene activity matrix" through summing all counts within the gene body, overlapping scATAC-seq fragments with genes, or employing co-accessibility calculations [8]. This converted data is analogous to scRNAseq data and processed through scRNA-seq tools like Seurat[9], SingleR[10], SingleCellNet[11], and scNym[12]. Regarding the difference from scRNA-seq data, a few methods have been optimized specifically for the scATAC data through neural networks [13]. Unfortunately, these methods have simply summed the counts of surrounding peaks around a gene, and thus ignore the specificity of peaks.

To address this issue, a few methods annotate cell types directly using the peak-by-cell matrix data. For example, EpiAnno [14] keeps frequent peaks and inputs them into a nonlinear Bayesian neural network to capture the latent space. scATAnno [15] emphasizes on detecting unknown cell types not present in the reference data by estimating uncertainty scores. Albeit successful, these two methods only consider these peaks independently without considering their relative positions. More importantly, they didn't consider the genomic sequence information.

In fact, the peaks in scATAC-seq data could be discriminated by underlying genome sequences containing the accessibility of cell type-specific enhancers and the transcription factor binding motifs, which are informative of the developmental state and cell identity [16]. Genomic sequence information has been widely used for predicting gene expression[17], chromatin accessibility [18], extracting embeddings [19], and enhancer-promoter interaction prediction [20]. However, genomic information hasn't been used for cell annotation for scATAC data.

To this end, we propose SANGO, an accurate and scalable graph-based method for annotating cells within scATAC-seq data by integrating DNA sequence information. SANGO first learns the low-dimensional informative representations of scATAC-data from the underlying DNA sequence information of peaks through the channel attention convolutional neural network [21]. The learned low-dimensional representations of reference and query data are subsequently fed into a graph transformer to remove batch effects by propagating shared messages among similar cells. Ultimately, the graph transformer is fine-tuned through cell labels in the reference data, and used for predicting cell types for the query. SANGO was demonstrated to consistently outperform competing methods on 55 paired scATAC-seq datasets across samples, platforms, and tissues. SANGO was also shown able to detect unknown tumor cells. Moreover, according to the annotated cells, cell type-specific peaks could be used for downstream analyses to provide functional insights and biological signals.

Results

Overview of the SANGO.

As shown in Fig 1, SANGO is a deep learning-based method for annotating cells within scATAC-seq data. SANGO first extracts cell low-dimensional representations from DNA sequence information underlying accessibility peaks by predicting single cell chromatin accessibility (stage 1), and subsequently utilizes the learned cell representations to annotate cell types of the query dataset according to reference datasets (stage 2).

In stage 1, around each peak, L-base pair (bp) DNA sequence is extracted and one-hot encoded into an $L\times 4$ matrix. The matrix undergoes initial processing with convolutional filters before being inputted into a channel attention convolutional neural network (CA-CNN) to derive the d-dimensional embedding of the peak. The embedding is then used to predict binary accessibilities of the peak for all N_{cell} cells through a dense linear network transformation with a weight matrix W_c of $d \times N_{cell}$. The network is trained iteratively over all peaks through the binary cross-entropy loss. Finally, the learned weights in the dense network serve as a d-dimensional representation for the N_{cell} cells. Here, L and d are two hyperparameters.

In stage 2, the learned cell representations of the reference and query data are fed into the graph transformer for simultaneously removing batch effects and predicting cell types. Specifically, the graph transformer randomly initializes the edge similarity weights between cells and updates the edge similarity weights by the learned attentions. In parallel, the graph transformer propagates shared messages along the learned edge similarity weights between cells to remove batch effects. Finally, the graph transformer is optimized for the given cell labels in the reference data, and used to predict cell types for the query.

Performance of cell type annotation for intra-datasets

SANGO was first evaluated on 14 sets of paired intra data, each pair consisting of the annotated cell types as the reference data and the unannotated as the query. As shown in Fig 2a and Supplementary Figure 1, our method achieved the best performance with an average accuracy of 96.4%, which was 3.3% higher than the second-ranked method scJoint. The difference is mostly from sequence information since the removal of sequence (SANGO-noseq) caused 6.4% drop in the average accuracy. The 3nd best method is scNym achieving an average accuracy of 92.5%, partly because it was designed for scRNA-seq data and not specifically optimized for scATAC-seq data. Another two scRNA-seq analysis methods (SingleR and Seurat) achieved the lowest accuracies (61.9% and 81.6%), indicating that their used machine learning or linear correlation techniques are not so powerful compared to the deep learning used by scNym. We noticed that EpiAnno didn't perform well like because it wasn't designed to identify detailed cell types. On the dataset culled by EpiAnno to identify coarse cell types, EpiAnno was slightly worse than our method (Fig 2b) but much better than other 5 methods. Most methods achieved accuracy above 90% on this easy task.

To further illustrate the advantages of SANGO, we showed the river plots on the paired case of LargeIntestineB-LargeIntestineA. As shown in Fig 2c, for the most difficult discriminated B and T cells, our method achieved accuracies of 91% and 97.3%, respectively. Both scNym and EpiAnno achieved less than 50% accuracy for B cells and completely missed T cells. By comparison, scJoint achieved better performance for B cells while failed to predict T cells. Our method (Fig 2d) successfully separated T cells from B cells. For the rare Endothelial I cells, SANGO achieved higher prediction accuracy (91%) than SingleR (73%) and other methods (<10%). This is as expected because SANGO-noseq failed to separate these cells after removing genomic sequence information. These results demonstrated that SANGO provided embeddings for better intra-cell type compactness and inter-cell type separability, which were beneficial to further cell classification.

Performances across platform and tissue datasets

Since available reference datasets were primarily obtained from other platforms or tissues, it was essential to evaluate methods on cross-platform and tissue datasets. Here, we first compared methods on 19 paired datasets from different sequencing platforms (10x, snATAC-seq, and sciATAC-seq). As shown in Fig 3a and Supplementary Figure 2, SANGO consistently yielded the best performance with average accuracy of 77.6%, which was 10.1% higher than the second-best method Cellcano. The following methods, EpiAnno and scNym achieved average accuracies of 67% and 61.9%, respectively. Across the datasets, the competing methods showed divergent performances: SingleR and Seurat achieved lower accuracies on the paired dataset of MouseBrain(10x) and Cerebellum (28.8% and 43.2%), while EpiAnno, scNym, and Cellcano obtained lower accuracies on the paired dataset of MosP1 and Cerebellum (62%, 53%, and 59.4%). In contrast, our method consistently obtained the best performance on these two datasets (78.3% and 78.8%, respectively). As visualized in Fig 3b, for the MosP1_Cerebellum case data, Endothelial cells from Microglia cells that join together according to original data could be separated by integrating the genomic sequence information (SANGO-nograph, SANGO without graph transformer) while cells of the same types did not cluster well. The complete version, SANGO, demonstrated intra-cluster compactness and inter-cluster separation. In contrast, Seurat and scNym didn't separate Endothelial cells from Astrocyte cells. SingleR mixed Endothelial and Microglia cells.

For 22 paired cross-tissue datasets across seven tissues (Bone Marrow, Liver, Kidney, Lung, Heart, Intestine, and Mouse Brain), SANGO obtained superior performance with an average accuracy of 86.3% (Fig 3c and Supplementary Figure 3). This was significantly higher than SingleR (60%; P=5.7E-10), scNym (59.1%; P=9.7E-05), Seurat (55.9%; P=5E-06), scJoint (50.2%; P=7.6E-06), Cellcano (44.6%; P=7.8E-06), and EpiAnno (35.7%; P=9.3E-09) by T-test. Similarly, for the BoneMarrowB_Liver case data (Fig 3d), SANGO could solve the mixture problem of all cell types by the raw data. The competing methods did not cluster cells well and were difficult to separate Monocytes from others. All methods failed to isolate the T and Regulatory T cells, likely because they were the immune cells sharing similar gene expression.

We also included three clustering metrics for evaluating the batch effect on cross-platform and tissue datasets. As shown in Supplementary Figure 4, our method achieved the best performance in terms of average Adjusted Rand Index (ARI) on the cross-platform and tissue datasets, yielding 5.2% and 29.4% improvements compared to the second-ranked method scNym, respectively. Similar trend could be observed when measured by the Normalized Mutual Information (NMI) and Average Silhouette Width (ASW) metrics. Since part of competing methods didn't consider batch effect, we applied two recommended methods (Harmony and LIGER)[22] to remove batch effects of datasets. Similar to the previous observation [13], the batch removal didn't improve (indeed worsen) the classification (Supplementary Figure 5-6).

SANGO was shown robust to various parameter selection (Supplementary Figure 7). The employed CA-CNN module was better than Autoencoder because its substitution to Autoencoder caused 50.1%, 74.3%, and 50.4% drops in the average accuracy on the intra-, cross-platform, and the cross-tissue datasets, respectively (Supplementary Figure 8). We also tested using continuous scATAC values instead of binary status (peak or not), and the model accuracy has a sharp drop. This is likely because the scATAC matrix is highly sparse (about 97.2% zeros), and the continuous value model brought more complexity than signals.

Performance with different reference data sources

To evaluate our method on the multi-source or the atlas as reference data, we employed multi-source datasets from tissues mouse brain (consisting of four datasets) and Intestine (consisting of three datasets) as the reference. For each tissue, we iteratively utilized one dataset as the query data and the rest as the multi-source data, resulting in 7 paired multi-reference and query datasets. As shown in Fig 4a and Supplementary Figure 9, SANGO achieved the best performance with average accuracy of 93.2%, which was 6.4% and 7.4% higher than two next best methods (scNym and scJoint), respectively. Seurat and Cellcano achieved similar results while singleR obtained the lowest performance. We also tested the performance of our method when transferring labels from single reference dataset to the combined query datasets, and our method still performed the best (Supplementary Figure 10).

To investigate the performance on the real single-cell atlas data, we annotated cell types from the reference PBMC atlas to the query PBMC dataset from 10x Genomics. Since the query data did not have ground truth labels, we annotated the cell types by Seurat [23] as the reference labels. As shown in Fig 4c, most of the cell types predicted by SANGO were the same as Seurat's annotations, except that our method made opposite annotations on the Memory and Naïve B cells (Fig 4c, black circle). The SANGO annotations were confirmed through the peak signals around their marker genes (Fig 4d): the annotated Naïve B cells showed enriched peaks over the reported marker gene *TCL1A*[24], while the annotated Memory B cells showed enriched peaks over the marker gene *FCGR2B* [25] and the specifically expressed *TEX9* gene [26].

To investigate the performance on the challenging heterogeneity atlas data, we employed the large adult mouse atlas consisting of 13 tissues with about 80, 000 cells. The large tissue LungA was used as the query dataset with the rest as the reference because ~90% of the query cells have corresponding cell types occurring in at least one of other tissues (e.g. B cells in BoneMarrow, LargeIntestine, and Spleen tissues; Endothelial II cells in Heart and WholeBrain tissues). For better quantitative evaluations, we excluded query cells of "unknown" types that do not exist in the reference or of rare types that contain <1000 cells in the query. As shown in Fig. 4b, our method achieved the best performance. EpiAnno was not included because it reported errors in the dataset.

SANGO was also evaluated on two scenarios when the reference and query datasets do not contain the same cell populations. We selected the WholeBrain tissue as the query data and other tissues as the reference since the WholeBrain tissue has 9 cell types all included in the Cerebellum tissue. To mimic the first scenario where all query cell types appear in one specific reference dataset and not in any other datasets, we kept 9 cell types only in the Cerebellum, and removed the cell types from other tissues. For the second scenario where query cell types are included but not by any individual reference dataset, we removed cell types so that five cell types appeared only in the Cerebellum, and the other four cell types appeared only in other tissues. As shown in Supplementary Figure 11, SANGO demonstrated the best performance in both scenarios, outperforming the second-best method, scNym, by 4.8% and 1.1% in terms of accuracy, respectively. For the computational and memory costs(Supplementary Figure 12), our method demonstrated a running time and Graphics Processing Unit (GPU) memory comparable to EpiAnno (EpiAnno encountered an error for datasets containing over 40K cells), longer but acceptable when compared to other methods.

Revealing biological implications for normal tissues

To demonstrate the capability of our method in elucidating biological mechanisms, we applied the Prefrontal cortex as reference data to annotate the normal cortex data from the adult mouse brain. Since the query data didn't provide the ground truth labels, we first checked the chromatin accessibility over signature genes across predicted cell types through the coverage plot. As shown in Fig 5a, for peak signals over the genomic region $\pm$ 3kb of each cell type-specific gene, the epigenetic features within the scATAC-seq profile exhibited a distinct enrichment of peaks in the cell types predicted by SANGO. For example, the excitatory neuron cells displayed enriched peaks over the *Neurod6* gene, a canonical marker of excitatory neuron cells [27]. The Microglia showed enriched peaks aound the classical marker gene *Tmem119* [28]. Oligodendrocytes showed enriched signals over its signature gene *Mag*. Similar results could be observed in other cell types (Supplementary Figure 13). The enriched epigenetic features of signature genes support cell-type annotations predicted by SANGO.

To investigate the functional insights of the predicted cell populations, the cell type-specific peaks were analysed from three aspects. Firstly, by motif enrichment analysis through Signac [29], the results showed that most motifs were specific for the annotated cell types (Supplementary Figure 14a). Among these, the Excitatory neurons cell type achieved the top value 82% (41 cell type-specific motifs in the total top 50 motifs) while the Endothelial II cell type obtained the lowest value 52% in terms of cell type-specific motifs. The top 10 cell type-specific motifs were enriched in the corresponding cell types (Supplementary Figure 14b). As depicted in Fig 5b and Supplementary Figure 15, the binding motifs for each cell type were also supported by previous literature (Supplementary Note 1).

Secondly, the tissue-specific expression enrichment was calculated through the SNPsea analysis for single nucleotide polymorphisms (SNPs) in the set of cell type-specific peaks and the set of background peaks. Background peaks were obtained by omitting the union of these cell type-specific peaks from the complete peaks. The analysis quantified the enrichments of tissue-specific expression profiles across 79 tissues, revealing the top 30 significantly enriched tissues for the Ex. neurons as demonstrated in Fig 5c. There is significant enrichment of cell type-specific gene expression within the SANGO-identified peaks in brain-related tissues compared to the background peaks.

Finally, SANGO could reveal co-accessible sites specific to cell types. By predicting cis-regulatory chromatin interactions through Cicero (Fig 5d and Supplementary Figure S16-22), the cis-regulatory interactions were observed specific to each cell type. Notably, the cell type-specific peaks (cyan peaks) aligned well with the patterns of cell type-specific interactions, effectively reducing false-positive identifications in genomic regions lacking cell type-specific interactions. These results highlight the potential of these cell type-specific peaks in deciphering cis-regulatory grammar and cooperative interactions.

Identifying multi-level cell types in basal cell carcinoma data

To investigate the ability on multi-level cell type prediction, we evaluated SANGO on basal cell carcinoma (BCC) sample data composed of diverse immune subtype cells and tumor cells in the tumor microenvironment (TME). BCC sample was first annotated by referenced on a healthy adult human large atlas (HHLA) with merged immune cell types. As shown in Fig 6a-c, SANGO identified tumor cells as "unknown" with high unknown probability scores. For known cell types, most immune cells and Endothelial cells were correctly predicted, as also indicated by the river plot (Fig 6d). Differently, fibroblast cells were predicted as mural cells (commonly referred to as pericytes), likely due to the strong connection between pericytes and fibroblasts within the TME. These results demonstrated that our method could efficiently distinguish tumor cells from immune cells and identify the tumor cells as unknown types.

To test the ability for annotating subtypes, we annotated the merged immune cells by using the Tumor-Infiltrating Lymphocytes atlas from Basal Cell Carcinoma (BCC_TIL) [15] that contains diverse subtypes of immune cells. SANGO obtained 90% accuracy for identifying the immune

subtypes (Fig 6e,f). To further refine the predicted results, we followed previous studies [30, 31] to conduct clustering on the query data, and showed the clustering could refine the annotations (Supplementary Figure 23). Unfortunately, the clustering didn't always improve the annotations. On another dataset from the previous study [31], by using one pre-treatment sample (SU008_Immune_Pre) and one post-treatment sample (SU006_Total_Post) as query datasets and the rest as the reference, the clustering decreased the accuracy from 90 to 89.1% (Supplementary Figure 24). Thus, we kept the clustering as an optional step.

Discussion

Accurate cell type identification is essential for scATAC-seq data analysis. Here, We present SANGO, a scalable and accurate method to identify cell types by efficiently integrating DNA sequence information for solving the high dimensionality and sparsity of scATAC-seq data. SANGO uses a scalable graph transformer to remove batch effects between the annotated reference data and unlabeled query data. Extensive experiments demonstrated that SANGO achieved superior results on 55 paired scATAC-seq datasets across samples, platforms, and tissues. SANGO was also shown able to detect unknown tumor cells in TME. Moreover, the annotated cells were found to contain cell type-specific peaks, providing functional insights/biological signals.

Several methods including Cellcano annotate cells within scATAC-seq data using the gene activation matrix estimated from the peak-by-cell matrix. This process avoids high-dimensionality and sparsity of peak data while ignoring the specificity of peaks. To address this problem, several methods including EpiAnno directly use the peak-by-cell matrix data. However, these methods only independently consider peaks without keeping their relative positions. More importantly, they didn't consider the genomic sequence information. To this end, SANGO integrates the DNA sequence into the cell annotation task to learn low-dimensional informative representations. By this way, our methods displayed superior performance compared to competing methods.

While unsupervised clustering methods prevail, supervised cell type identification strategy has been extensively used in real data analyses [24, 32-34]. Current studies are focused on cell clustering mostly due to a lack of high-quality annotated datasets. As indicated in the recent review [35], the reference-based classification methods turn more and more popular with the increasing availability of high-quality atlas datasets. These datasets enable fast annotations of not only major cell types but also detailed subtypes. We utilized both the reference and query data in the training to solve the potential divergence. Such strategy was essential to improve the performance, as also indicated in previous studies[36, 37].

Relative to scRNA-seq, scATAC-seq provides another view to reveal regulatory dynamics and potential cell states that may not be apparent from gene expression alone. Particularly, scATAC-seq is valuable to discover subtle differences in regulatory mechanisms between cell types, which might be missed by scRNA-seq. Our algorithm aims to harness this potential, enhancing the analysis and understanding of complex cellular landscapes.

SANGO could be further improved by the following aspects. Firstly, SANGO can be extended to integrate multi-omics datasets. Secondly, with the rapid growth of omics data, the pretrained foundational models might provide a unified framework to elegantly integrate multi-source information including genome sequence information.Thirdly, the used deep learning models are not easily interpretable in terms of specific biological features such as motifs, requiring the development of interpretability for our models.

Methods

The Architecture of SANGO

SANGO is a scalable and accurate tool for annotating cells within scATAC-seq data by harnessing DNA sequence information. Fig 1 illustrates a two-stage workflow involving sequence information extraction and cell type prediction. In stage 1, SANGO employs the channel attention convolutional neural network to extract low-dimensional representations from the DNA sequence information underlying accessibility peaks for the reference and query data. These extracted representations are further fed into the graph transformer module in stage 2 to simultaneously eliminate batch effects and perform cell-type prediction.

Stage 1: Sequence information extraction

Channel attention block. The channel attention block is used to improve the performance of convolutional neural networks by learning channel attention for each convolutional block [38]. Concretely, for the input feature matrix $C \times F$, we first conduct the pooling operation for C channels and then perform local cross-channel interactions for each channel along with its k neighbors. Given the pooled channel feature $h \in \mathbb{R}^C$, we then compute the interacted channel feature h_i by regarding the interaction between h_i and its k neighbors through shared learning parameters as follows:

$$\hat{h}_i = \delta\left(\sum_{j=1}^{k} w^j h_i^j\right),\ h_i^j \in \Omega_i^k \quad (1)$$

where Ω_i^k indicates the set of k adjacent channels of h_i and δ is a Sigmoid function. j is the j-th neighboring channel to the i-th channel out of C channels. This process can be easily implemented by a fast 1D convolution with a kernel size of k: $\hat{h} = \delta(1D\_Conv_k(h))$. We follow the study [38] to adaptively select the k neighbors based on the channel dimension C as follows:

$$k = \left|\frac{log_2(C)}{o} + \frac{e}{o}\right|_{odd} \quad (2)$$

where $|m|_{odd}$ represents the nearest odd number of m. C is the number of channels. e and o are two hyperparameters that control the relationship between C channels and k neighbors. In this paper, we set o and e to 2 and 1 for all datasets, respectively. Finally, we conduct channel-wise multiplication between the input feature $C \times F$ matrix and the interacted channel feature $\hat{h}$ to maintain the shape of input data.

Channel attention convolutional neural network (CA-CNN). The channel attention block is seamlessly integrated into the convolutional neural network to facilitate cell representation extraction. As shown in Fig 1, stage 1 takes an L-bp (L = 1344 for all datasets) DNA sequence from the center of each peak as the input, which is then transformed into a 1344×4 matrix by one-hot encoding. The process begins with a 1D convolutional layer with 288 filters of size 17×4. This is followed by batch normalization, GELU activation, and then 3 max-pooling layers, resulting in a 448×288 output matrix. The output matrix is then processed by the channel attention convolutional neural network (CA-CNN) with a depth of 4 layers. Each layer has two convolutional blocks (followed by batch normalization, max pooling, and GELU activation function) and one channel attention block. For the convolutional block in the four layers of CA-CNN, we set the number of convolutional filters varying in [64, 128, 256, 512] and maintained a kernel width of 5. The output of CA-CNN is a 56×512 matrix, which is further processed through a 1D convolutional layer with 256 filters of width 1 to generate a 28×256 matrix. The matrix is further flattened into a 1×7168 vector.

Cell representation extraction. The learned 1×7168 vector is then fed into a d-unit (d=64) bottleneck layer to capture the low-dimensional embeddings of the peak. These embeddings are subsequently utilized to predict the peak accessibility of the N_{cell} cells through a dense layer. All learnable parameters in stage 1 are optimized iteratively by the binary cross-entropy loss between the predicted peak accessibility and the observed peak accessibility. Here, the learned weights in the dense network serve as a 64-dimensional representation for the N_{cell} cells. In addition, we provide an optional batch correction operation in stage 1. Specifically, we introduce a second parallel dense layer connected to the bottleneck layer, which is responsible for predicting batch-specific peak accessibility. This batch-specific peak accessibility is then multiplied by the batch-by-cell matrix to compute the batch-related peak accessibility for each cell. The resulting vector is then added to the existing peak accessibility of each cell. We apply this strategy to the cross-platform datasets since their reference genome versions are different.

Stage 2: Graph transformer for cell type prediction

The learned representations of the reference and query are concatenated to construct the paired reference-query data, which is then fed into the graph transformer for cell type prediction. In this process, the batch effects are mitigated by propagating shared information among similar cells through the attention mechanism used in the graph transformer. However, the attention mechanism entails an $O(N^2)$ computational complexity, which becomes impractical for larger datasets. Therefore, we employ an approximate scheme named **k**ernelized **G**umbel-**S**oftmax **o**perator (KGSO) [39] for message passing and similarity learning, which seamlessly synthesizes random feature map [40] and approximated sampling strategy [41], resulting in reducing the complexity from $O(N^2)$ to $O(N)$ via avoiding explicit computation of the all-pair cell similarities.

Kernelized Gumbel-Softmax operator. The kernelized Gumbel-Softmax operator consists of kernelized message passing and differentiable stochastic structures learning.

Kernelized message passing. We assume $z_u^{(0)}$ as the cell u initial representation. We define a full-graph attention network that estimates cell similarity and enables corresponding densely connected message propagating as follows:

$$\tilde{a}_{uv}^{(l)} = \frac{exp\left(\left(W_Q^{(l)} z_u^{(l)}\right)^T \left(W_K^{(l)} z_v^{(l)}\right)\right)}{\sum_{w=1}^{N} exp\left(\left(W_Q^{(l)} z_u^{(l)}\right)^T \left(W_K^{(l)} z_w^{(l)}\right)\right)}, \quad z_u^{(l+1)} = \sum_{v=1}^{N} \tilde{a}_{uv}^{(l)} \left(W_V^{(l)} z_v^{(l)}\right) \quad (3)$$

where $W_Q^{(l)}, W_K^{(l)}, and\ W_V^{(l)}$ are learnable parameters in the l-th layer. Symbol u represents a specific cell within the paired reference-query dataset, while symbol v represents the v-th cell among the cells in the dataset. To alleviate the challenging $O(N^2)$ complexity, we transform th e dot-then-exponentiate operator into a pairwise similarity function:

$$z_u^{(l+1)} = \sum_{v=1}^{N} \frac{ks\left(W_Q^{(l)} z_u^{(l)}, W_K^{(l)} z_v^{(l)}\right)}{\sum_{w=1}^{N} ks\left(W_Q^{(l)} z_u^{(l)}, W_K^{(l)} z_w^{(l)}\right)} \cdot \left(W_V^{(l)} z_v^{(l)}\right) \quad (4)$$

where $ks(.,.): \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a kernel function that measures the pairwise similarity. The kernel function can be approximated by random features (RF) [40], which serve as an unbiased estimation. Therefore, we can further convert the dot-then-exponentiate operation to an inner product in vector space as follows:

$$z_u^{(l+1)} = \sum_{v=1}^{N} \frac{\varphi(q_u)^T \varphi(k_v)}{\sum_{w=1}^{N} \varphi(q_u)^T \varphi(k_w)} \cdot V_v = \frac{\varphi(q_u)^T \sum_{v=1}^{N} \varphi(k_v) \cdot V_v^T}{\varphi(q_u)^T \sum_{w=1}^{N} \varphi(k_w)} \quad (5)$$

For simplicity, we assume $q_u = W_Q^{(l)} z_u^{(l)}, k_u = W_K^{(l)} z_u^{(l)}, and\ V_u = W_V^{(l)} z_u^{(l)}$, respectively. V_v^T is the transposition of the term $V_v = W_V^{(l)} z_v^{(l)}$. The term $\varphi(\cdot)$ is a low-dimensional feature map with random transformation. The two summations in Equation 5 are shared by each u, so that one only needs to compute them once. This property enables $O(N)$ computational complexity for full-graph message passing. Detailed proof can be found in the study [39].

Nevertheless, Equation 5 may cause the gradient to vanish since the normalization of the denominator tends to reduce the attention weights to zero. The core problem lies in the fact that message passing takes place within a densely connected graph with weighted edges. However, it's worth noting that only partial edges play a crucial role in the message-passing process. Consequently, our next step is to address this hurdle by extracting a sparse structure from the fully connected graph through a distillation process.

Differentiable stochastic structure learning. This section introduces a framework for differentiable optimization over discrete graph structures that arise from the fully connected graph. We use the reparameterization trick [42] to modify Equation 3 to allow differentiable learning in the following manner:

$$z_u^{(l+1)} = \sum_{v=1}^{N} \frac{exp\left(\left(q_u^T k_v + g_v\right)/\tau\right)}{\sum_{w=1}^{N} exp\left(\left(q_u^T k_w + g_w\right)/\tau\right)} \cdot V_v = \sum_{v=1}^{N} \frac{ks\left(\frac{q_u}{\sqrt{\tau}}, \frac{k_v}{\sqrt{\tau}}\right) e^{g_v/\tau}}{\sum_{w=1}^{N} ks\left(\frac{q_u}{\sqrt{\tau}}, \frac{k_w}{\sqrt{\tau}}\right) e^{g_w/\tau}} V_v \quad (6)$$

where g_v is sampled from a Gumbel distribution, while τ represents a temperature coefficient that governs the proximity to hard discrete samples [43]. Then, we can yield that

$$z_u^{(l+1)} \approx \sum_{v=1}^{N} \frac{\varphi\left(\frac{q_u}{\sqrt{\tau}}\right)^T \varphi\left(\frac{k_v}{\sqrt{\tau}}\right) e^{\frac{g_v}{\tau}}}{\sum_{w=1}^{N} \varphi\left(\frac{q_u}{\sqrt{\tau}}\right)^T \varphi\left(\frac{k_w}{\sqrt{\tau}}\right) e^{\frac{g_w}{\tau}}} \cdot V_v = \frac{\varphi\left(\frac{q_u}{\sqrt{\tau}}\right)^T \sum_{v=1}^{N} e^{\frac{g_v}{\tau}} \varphi\left(\frac{k_v}{\sqrt{\tau}}\right) V_v^T}{\varphi\left(\frac{q_u}{\sqrt{\tau}}\right)^T \sum_{w=1}^{N} e^{\frac{g_w}{\tau}} \varphi\left(\frac{k_w}{\sqrt{\tau}}\right)} \quad (7)$$

where Equation 7 facilitates message passing over a sampled latent graph while maintaining linear complexity. A full proof is available in the study [39].

Priori structures as relational bias. Priori cell edge relationships are beneficial for cell classification and clustering [30, 44, 45]. Here, we incorporate priori cell edge relationships as relational biases to adjust the attention weights between cells. Specifically, we modify the attention weight $\tilde{a}_{uv}^{(l)}$ by adding a term $\prod[a_{uv} = 1]\sigma\left(b^{(l)}\right)$, where $b^{(l)}$ represents a learnable scalar that acts as a relational bias for adjacent cell pairs (u, v), and σ represents the sigmoid activation function. Thus, the cell embeddings can be adjusted accordingly using the following step:

$$z_u^{(l+1)} \leftarrow z_u^{(l+1)} + \sum_{v, a_{uv}=1} \sigma\left(b^{(l)}\right) \cdot V_v \quad (8)$$

In this study, we use the k-nearest neighbors (KNN) algorithm to build the priori cell edge relationships within reference and query data.

Loss function. We apply the standard cross-entropy loss to minimize the cell classification error with known labels $Y = \{y_u\}_{u \in N_{r\_cell}}$, where N_{r_cell} represents the number of cells within the reference data.

$$L_{ce} = -\frac{1}{N_{r\_cell}} \sum_{v \in N_{r\_cell}} \sum_{t=1}^{T} \Pi[y_u = t] log \tilde{y}_{u,t} \quad (9)$$

where T is the number of cell types, $\Pi[\,.\,]$ is an indicator function. since graph topology learning increases the degrees of freedom, while the available number of training labels is not comparable to that. To mitigate this, we introduce edge-level regularization as follows:

$$L_e = -\frac{1}{NL}\sum_{l=1}^{L}\sum_{(u,v)\in\varepsilon}\frac{1}{d_u} log\pi_{uv}^{(l)} \quad (10)$$

where d_u represents the in-degree of cell u and $\pi_{uv}^{(l)}$ is the predicted probability for cell edge between cell u and cell v at the l-th layer. Equation 10 is a maximum likelihood estimation for edge in ε. We follow the study [39] to obtain $\pi_{uv}^{(l)}$ values as follows:

$$\pi_{uv}^{(l)} = \frac{\varphi\left(W_Q^{(l)} z_u^{(l)}\right)^T \varphi\left(W_K^{(l)} z_v^{(l)}\right)}{\varphi\left(W_Q^{(l)} z_u^{(l)}\right)^T \sum_{w=1}^{N} \varphi\left(W_K^{(l)} z_w^{(l)}\right)} \quad (11)$$

In summary, the total loss of our method is defined as follows:

$$L_{total} = L_{ce} + \lambda\, L_e \quad (12)$$

where λ is a hypermeter used for controlling the contribution of the edge-level regularization loss.

Unknown probability score

To identify cellular types that exist within the query data but are absent in the reference data, we introduce a strategy for evaluating the unknown probability score for each cell within the query data by considering its neighboring cells in the reference data. Concretely, for cell i in query data, we first identify its n closest cell neighbors (referred to as $N_{(i)}$) in reference data based on the similarity weights learned by the attention mechanism. The distance between the cell i and its neighbor j in the reference data is further quantified as $dist(i,j)$= 1 - similarity weight between cell i and cell j. Then, we establish the standard deviation of these nearest distances to capture the variations by the following formula: $sd_j = \sqrt{\sum_{j\in N_{(i)}} dist(i,j)^2/n}$. We further transform these distances into a similarity measurement: $S(i,j) = e^{-dist(i,j)/\left(2/sd_j\right)^2}$ through the Gaussian kernel function. Finally, the unknown probability score of the query cell i can be represented by $P_i = 1 - \frac{\sum_{j\in N_i} S(i,j) I(y_j^R = t_q)}{\sum_{j\in N_i} s(i,j)}$, where y_j^R is the cell label of the reference cell j and t_q is the cell type of the query cell i. In this study, the cell is identified as an unknown type if its P score surpasses the half of maximal score P_{max} in the query data.

Clustering-level label

After obtaining cell predictions using SANGO, we further acquired clustering tags for each cell through the Leiden algorithm by utilizing the embeddings generated by our method. Then, we labeled the cluster-level tags for each cluster by selecting the most frequent cell type predicted by our method within that cluster. We used clustering-level labels to annotate cells within tumor datasets.

Downstream analysis

SNPsea analysis. We conducted SNPsea analysis [46] utilizing the default configurations for SNPs within each group of cell type-specific peaks and the corresponding background peak set. Specifically, we evaluated the extent of tissue-specific expression enrichments in the profiles of 17,581 genes across various human tissues, using the Gene Atlas dataset [47].

Motif enrichment analysis. We performed a motif enrichment analysis utilizing the cell type-specific accessibility peaks through Signac [29]. Concretely, we calculated the GC content (the proportion of G and C nucleotides) for each differentially accessible peak and then selected a background set of 40,000 peaks in a manner that ensured the background set closely matched the overall GC content, accessibility, and peak width of the differential peaks. This process was accomplished by utilizing the FindMotifs function in Signac.

Cis-coaccessibility analysis. To predict cis-regulatory chromatin interactions relevant to each specific cell type, we used the Cicero tool [8]. Specifically, we separately preprocessed cell dataset objects for scATAC-seq data corresponding to each specific cell type. This preprocessing involved a series of operations, including the detect_genes, estimate_size_factors, preprocess_cds, and reduce_dimension functions, all of which were run with their default parameters. This extensive preparation culminated in the transformation of the data into a Cicero cell dataset object using the make_atac_cds function. Finally, cell type-specific chromatin interactions were acquired using the run_cicero function.

Coverage plot. The coverage plot is constructed through the Singac tool [29]. We first generate a fragment index file from the fragment data of the query dataset. We then seamlessly integrate the fragment file into Signac to visualize peak accessibility across genomic regions through the CoveragePlot function of Signac.

Datasets preprocessing

We followed studies [5, 48] converting the cell-by-peak count matrices of the reference and query data to the binary matrices, where a peak with value"1" indicates that one or more reads fall within that peak, and the value "0" indicates otherwise. We next performed feature selection following the studies [4, 49] to reserve the peaks that have at least one read count in at least 1% of cells. Cells that were not expressed in any peaks would be removed. Note that, the Healthy Adult Human Large Atlas (HHLA) was obtained from the previous study [15], where the data generated in ref [50]was annotated by GRCh38[51], and they have selected deep-sequenced 1000 cells per minor cell type and 890,130 adult-specific peaks, and labeled T cells, B cells and Myeloids as immune cells. On the other hand, to enable cell annotation between datasets from different experimental platforms, we needed to align the peaks of reference and query datasets. Specifically, for the situation of reference and query datasets with different genomes, we used the tool CrossMap [52] to map the genome of the target to that of reference data by converting genome coordinates between assemblies. Next, we used BEDTools [53] to obtain the overlaps between peaks from reference and query datasets. For fair quantification comparison on reference-query paired datasets from cross-samples, cross-platforms, and cross-tissues, we followed the literature [54] to remove cell types present in the query dataset but absent in the reference dataset.

Evaluation criteria

Cell annotation performance is evaluated through widely used metrics including overall accuracy (ACC), median F1 score (mF1), macro F1 scores, and Cohen's kappa. The accuracy measures the ratio of correctly assigned cells to the total cell count, gauging the overall assignment precision. The mF1 score considers cell type prediction as a binary classification task and calculates the median performance for each cell type. The Macro F1 metric treats all cell types equally, placing more emphasis on the accuracy of smaller clusters instead of other metrics. Cohen's kappa assesses the agreement between actual and predicted labels. Essentially, these metrics provide various perspectives and fairly capture differences in prediction performance across different evaluators.

Statistics and reproducibility

This study utilizes publicly accessible datasets. No statistical methods were used to pre-determine sample sizes but our sample sizes are the same

as those reported in previous publications [14]. After quality control, all data were incorporated into the analyses without any exclusions. The experiments were not randomized, and investigators were not blinded to allocation during both experimentation and outcome assessment. Data distribution was assumed to be normal but this was not formally tested (tested on a single dataset in Supplementary figure 25). All necessary data, code for replication of the analysis, are accessible at https://github.com/cquzys/SANGO.

Data availability

We downloaded the raw scATAC matrix data directly from the website as described in the following section and followed previous works [5, 48, 55] to binarize the matrix. (1) The datasets Bone Marrow A, Bone Marrow B, Lung A, Lung B, Kidney, Liver, Heart, Large Intestine A, Large Intestine B, Small Intestine, Whole brain A, Whole brain B, Cerebellum, and Prefrontal cortex are derived from the adult mouse atlas data [56], downloading from either the GEO access number GSE111586 or the website http://atlas.gs.washington.edu/mouse-atac/data/. These datasets are sequenced by the sci-ATAC-seq technology [57] and annotated through the mm9 reference genome. (2) The anterior datasets (Mos-A1, Mos-A2), middle datasets (Mos-M1, Mos-M2), and posterior datasets (Mos-P1, Mos-P2) are from the different sections of the secondary motor cortex in mouse brain [58], which can be accessed through GEO accession number GSE126724. These datasets are sequenced by snATAC-seq technology [59] and annotated through the GRCm38 reference genome. (3) The Mouse Brain (10x) dataset and the normal cortex dataset are sequenced by the 10x sequencing technology and annotated by the mm10 reference genome. These two datasets can be downloaded from https://support.10xgenomics.com/single-cell-atac/datasets/1.1.0/atac_v1_adult_brain_fresh_5k and https://www.10xgenomics.com/resources/datasets/fresh-cortex-from-adult-mouse-brain-p-50-1-standard-1-2-0, respectively. (4) The forebrain dataset can be downloaded through GEO accession number GSE100033, which is sequenced by the snATAC and annotated by the mm9 reference genome. (5) The PBMC atlas data, the Tumor-Infiltrating Lymphocytes atlas from Basal Cell Carcinoma (BCC_TIL), and the BCC sample data are obtained from the study [3, 15]. These datasets are annotated by the ENCODE hg19 reference genome and can be accessed through the GEO accession number GSE129785 or the download website https://www.synapse.org/#!Synapse:syn52559388/files/. (6) The PBMC (10x) data is obtained from the official 10x website: https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k, which is annotated by the GRCh38 reference genome. (7) The raw HHLA data can be obtained from the GEO accession number GSE184462 and the processed data can be downloaded from the website https://www.synapse.org/#!Synapse:syn52559388/files/. All of these datasets were preprocessed as described in the section Datasets preprocessing. Source data for Figures 2–6 is available with this manuscript.

Code availability

All source codes used in our experiments have been deposited at https://github.com/cquzys/SANGO. A Zenodo version is also available at https://zenodo.org/doi/10.5281/zenodo.10826453 (ref. 60).

Acknowledgments

This study has been supported by the National Key R&D Program of China (2022YFF1203100), National Natural Science Foundation of China (T2394502)， the Research and Development Project of Pazhou Lab (Huangpu) [2023K0606], and the Postdoctoral Fellowship Program of CPSF (GZC20233321).

Author Contributions

Y.Y. conceived and supervised the project. Y.Z., M.L, and N.S developed and implemented the SANGO algorithm. Y.Y, W.Y, and Y.Z. validated the methods and wrote the manuscript. P.S., J.F, and J.X conduct the biological analysis. Y.L and K.C discussed and performed the rebuttal experiments. All authors read and approved the final manuscript.

Competing interests

All authors have no competing interests to declare.

References

[1] J. D. Buenrostro, B. Wu, U. M. Litzenburger, et al., "Single-cell chromatin accessibility reveals principles of regulatory variation," *Nature,* vol. 523, no. 7561, pp. 486-490, 2015.

[2] H. Chen, L. Albergante, J. Y. Hsu, et al., "Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM," *Nature communications,* vol. 10, no. 1, p. 1903, 2019.

[3] A. T. Satpathy, J. M. Granja, K. E. Yost, et al., "Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion," *Nature biotechnology,* vol. 37, no. 8, pp. 925-936, 2019.

[4] L. Xiong, K. Xu, K. Tian, et al., "SCALE method for single-cell ATAC-seq analysis via latent feature extraction," *Nature communications,* vol. 10, no. 1, p. 4576, 2019.

[5] M. D. Luecken, M. Büttner, K. Chaichoompu, et al., "Benchmarking atlas-level data integration in single-cell genomics," *Nature methods,* vol. 19, no. 1, pp. 41-50, 2022.

[6] H. Chen, C. Lareau, T. Andreani, et al., "Assessment of computational methods for the analysis of single-cell ATAC-seq data," *Genome biology,* vol. 20, no. 1, pp. 1-25, 2019.

[7] J. M. Granja, M. R. Corces, S. E. Pierce, et al., "ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis," *Nature genetics,* vol. 53, no. 3, pp. 403-411, 2021.

[8] H. A. Pliner, J. S. Packer, J. L. McFaline-Figueroa, et al., "Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data," *Molecular cell,* vol. 71, no. 5, pp. 858-871. e8, 2018.

[9] R. Satija, J. A. Farrell, D. Gennert, et al., "Spatial reconstruction of single-cell gene expression data," *Nature biotechnology,* vol. 33, no. 5, pp. 495-502, 2015.

[10] D. Aran, A. P. Looney, L. Liu, et al., "Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage," *Nature immunology,* vol. 20, no. 2, pp. 163-172, 2019.

[11] Y. Tan and P. Cahan, "SingleCellNet: a computational tool to classify single cell RNA-Seq data across platforms and across species," *Cell systems,* vol. 9, no. 2, pp. 207-213. e2, 2019.

[12] J. C. Kimmel and D. R. Kelley, "scNym: Semi-supervised adversarial neural networks for single cell classification," *bioRxiv,* p. 2020.06. 04.132324, 2020.

[13] W. Ma, J. Luand H. Wu, "Cellcano: supervised cell type identification for single cell ATAC-seq data," *Nature Communications,* vol. 14, no. 1, p. 1864, 2023.

[14] X. Chen, S. Chen, S. Song, et al., "Cell type annotation of single-cell chromatin accessibility data via supervised Bayesian embedding," *Nature Machine Intelligence,* vol. 4, no. 2, pp. 116-126, 2022.

[15] Y. Jiang, Z. Hu, J. Jiang, et al., "scATAnno: Automated Cell Type Annotation for single-cell ATAC Sequencing Data," *bioRxiv,* 2023.

[16] D. Srivastava and S. Mahony, "Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms,* vol. 1863, no. 6, p. 194443, 2020.

[17] R. Schwessinger, J. Deasy, R. T. Woodruff, et al., "Single-cell gene expression prediction from DNA sequence at large contexts."

[18] H. Yuan and D. R. Kelley, "scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks," *Nature Methods,* vol. 19, no. 9, pp. 1088-1096, 2022.

[19] Z. Tayyebi, A. R. Pineand C. S. Leslie, "Scalable sequence-informed embedding of single-cell ATAC-seq data with CellSpace."

[20] K. Chen, H. Zhaoand Y. Yang, "Capturing large genomic contexts for accurately predicting enhancer-promoter interactions," *Briefings in Bioinformatics,* vol. 23, no. 2, p. bbab577, 2022.
[21] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458,* 2015.
[22] H. T. N. Tran, K. S. Ang, M. Chevrier, et al., "A benchmark of batch-effect correction methods for single-cell RNA sequencing data," *Genome biology,* vol. 21, pp. 1-32, 2020.
[23] T. Stuart, A. Butler, P. Hoffman, et al., "Comprehensive integration of single-cell data," *Cell,* vol. 177, no. 7, pp. 1888-1902. e21, 2019.
[24] C. Domínguez Conde, C. Xu, L. Jarvis, et al., "Cross-tissue immune cell analysis reveals tissue-specific features in humans," *Science,* vol. 376, no. 6594, p. eabl5197, 2022.
[25] M. Mackay, A. Stanevsky, T. Wang, et al., "Selective dysregulation of the FcγIIB receptor on memory B cells in SLE," *The Journal of experimental medicine,* vol. 203, no. 9, pp. 2157-2164, 2006.
[26] T. Sundell, K. Grimstad, A. Camponeschi, et al., "Single-cell RNA sequencing analyses: interference by the genes that encode the B-cell and T-cell receptors," *Briefings in Functional Genomics,* vol. 22, no. 3, pp. 263-273, 2023.
[27] L. Loo, J. M. Simon, L. Xing, et al., "Single-cell transcriptomic analysis of mouse neocortical development," *Nature communications,* vol. 10, no. 1, p. 134, 2019.
[28] C. Ruan and W. Elyaman, "A new understanding of TMEM119 as a marker of microglia," *Frontiers in Cellular Neuroscience,* vol. 16, p. 902372, 2022.
[29] T. Stuart, A. Srivastava, S. Madad, et al., "Single-cell chromatin state analysis with Signac," *Nature methods,* vol. 18, no. 11, pp. 1333-1341, 2021.
[30] J. Hu, X. Li, K. Coleman, et al., "SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network," *Nature methods,* vol. 18, no. 11, pp. 1342-1351, 2021.
[31] Y. Jiang, Z. Hu, J. Jiang, et al., "scATAnno: Automated Cell Type Annotation for single-cell ATAC-seq Data," *bioRxiv,* p. 2023.06. 01.543296, 2023.
[32] C. Xu, M. Prete, S. Webb, et al., "Automatic cell type harmonization and integration across Human Cell Atlas datasets," *bioRxiv,* p. 2023.05. 01.538994, 2023.
[33] C. V. Theodoris, L. Xiao, A. Chopra, et al., "Transfer learning enables predictions in network biology," *Nature,* pp. 1-9, 2023.
[34] Z.-Z. Hao, J.-R. Wei, D. Xiao, et al., "Single-cell transcriptomics of adult macaque hippocampus reveals neural precursor cell populations," *Nature neuroscience,* vol. 25, no. 6, pp. 805-817, 2022.
[35] L. Zappia and F. J. Theis, "Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape," *Genome biology,* vol. 22, pp. 1-18, 2021.
[36] C. Shengquan, Z. Boheng, C. Xiaoyang, et al., "stPlus: a reference-based method for the accurate enhancement of spatial transcriptomics," *Bioinformatics,* vol. 37, no. Supplement_1, pp. i299-i307, 2021.
[37] Q. Song, J. Suand W. Zhang, "scGCN is a graph convolutional networks algorithm for knowledge transfer in single cell omics," *Nature communications,* vol. 12, no. 1, p. 3826, 2021.
[38] Q. Wang, B. Wu, P. Zhu, et al., "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11534-11542.
[39] Q. Wu, W. Zhao, Z. Li, et al., "Nodeformer: A scalable graph structure learning transformer for node classification," *Advances in Neural Information Processing Systems,* vol. 35, pp. 27387-27401, 2022.
[40] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," *Advances in neural information processing systems,* vol. 20, 2007.
[41] E. Jang, S. Guand B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144,* 2016.
[42] D. P. Kingma, T. Salimansand M. Welling, "Variational dropout and the local reparameterization trick," *Advances in neural information processing systems,* vol. 28, 2015.
[43] C. J. Maddison, A. Mnihand Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," *arXiv preprint arXiv:1611.00712,* 2016.
[44] Y. Zeng, X. Zhou, J. Rao, et al., "Accurately clustering single-cell RNA-seq data by capturing structural relations between cells through graph convolutional network," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2020: IEEE, pp. 519-522.
[45] Y. Zeng, Z. Wei, Z. Pan, et al., "A robust and scalable graph neural network for accurate single-cell classification," *Briefings in Bioinformatics,* vol. 23, no. 2, p. bbab570, 2022.
[46] K. Slowikowski, X. Huand S. Raychaudhuri, "SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci," *Bioinformatics,* vol. 30, no. 17, pp. 2496-2497, 2014.
[47] A. I. Su, T. Wiltshire, S. Batalov, et al., "A gene atlas of the mouse and human protein-encoding transcriptomes," *Proceedings of the National Academy of Sciences,* vol. 101, no. 16, pp. 6062-6067, 2004.
[48] A. Ma, X. Wang, J. Li, et al., "Single-cell biological network inference using a heterogeneous graph transformer," *Nature Communications,* vol. 14, no. 1, p. 964, 2023.
[49] M. Zamanighomi, Z. Lin, T. Daley, et al., "Unsupervised clustering and epigenetic classification of single cells," *Nature communications,* vol. 9, no. 1, p. 2410, 2018.
[50] K. Zhang, J. D. Hocker, M. Miller, et al., "A single-cell atlas of chromatin accessibility in the human genome," *Cell,* vol. 184, no. 24, pp. 5985-6001. e19, 2021.
[51] V. A. Schneider, T. Graves-Lindsay, K. Howe, et al., "Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly," *Genome research,* vol. 27, no. 5, pp. 849-864, 2017.
[52] H. Zhao, Z. Sun, J. Wang, et al., "CrossMap: a versatile tool for coordinate conversion between genome assemblies," *Bioinformatics,* vol. 30, no. 7, pp. 1006-1007, 2014.
[53] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics,* vol. 26, no. 6, pp. 841-842, 2010.
[54] F. Yang, W. Wang, F. Wang, et al., "scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data," *Nature Machine Intelligence,* vol. 4, no. 10, pp. 852-866, 2022.
[55] K. E. Wu, K. E. Yost, H. Y. Chang, et al., "BABEL enables cross-modality translation between multiomic profiles at single-cell resolution," *Proceedings of the National Academy of Sciences,* vol. 118, no. 15, p. e2023070118, 2021.
[56] D. A. Cusanovich, A. J. Hill, D. Aghamirzaie, et al., "A single-cell atlas of in vivo mammalian chromatin accessibility," *Cell,* vol. 174, no. 5, pp. 1309-1324. e18, 2018.
[57] D. A. Cusanovich, R. Daza, A. Adey, et al., "Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing," *Science,* vol. 348, no. 6237, pp. 910-914, 2015.
[58] R. Fang, S. Preissl, Y. Li, et al., "Comprehensive analysis of single cell ATAC-seq data with SnapATAC," *Nature communications,* vol. 12, no. 1, p. 1337, 2021.
[59] S. Preissl, R. Fang, H. Huang, et al., "Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation," *Nature neuroscience,* vol. 21, no. 3, pp. 432-439, 2018.
[60] Zeng,Y., Luo, M., Shangguan, N., Shi, P., Feng, J., Xu, J., Chen, K., Lu, Y., Yu, W. and Yang, Y. Deciphering Cell Types by Integrating scATAC-seq Data with Genome Sequences. *Zenodo* https://zenodo.org/doi/10.5281/zenodo.10826453 (2024).

Figure legends

Fig. 1 | Schematic overview of SANGO framework for annotating cells within scATAC-seq data by integrating genome sequence. SANGO consists of two stages for sequence information extraction and cell type prediction, respectively. In stage 1, around the i-th peak, an input of L-bp length DNA sequence is extracted and one-hot encoded into an $L\times 4$ matrix. The matrix undergoes initial processing with C convolutional filters to generate the feature matrix with dimensions of $C \times F$. Subsequently, the matrix is inputted into a channel attention 1D convolutional neural network with the Sigmoid function δ and channel-wise multiplication $\otimes$. This is followed by a bottleneck layer to learn the d-dimensional embeddings of the peak. The embeddings are

subsequently used to predict binary accessibilities of the peak for all N_{cell} cells through a dense linear network transformation with a weight matrix W_c of size $d \times N_{cell}$. All learnable parameters in stage 1 are optimized iteratively by the binary cross-entropy loss over all peaks. Finally, the learned weights in the dense network serve as a d-dimensional representation for the N_{cell} cells. In stage 2, the learned representations of reference and query data are used to construct cell graph through similarity, and a graph transformer was employed to remove batch effects and predict cell labels $\hat{Y}$ according to the truth labels Y of N_{r_cell} cells in the reference data. Finally, the trained graph transformer is used to predict cell types of N_{q_cell} cells within the query data.

Fig. 2 | Performance of cell type annotation for intra-datasets. (a) Boxplots summarize the ACC scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. The hollow red dot within the boxplot represents the average values, while black dots denote outliers. This analysis includes n=14 biologically independent paired intra-datasets. The x-axis represents the various methods, while the y-axis denotes the measured values. **(b)** The radar plot shows the accuracy of each method on five datasets with merged cell types when using the Forebrain data as the reference, where the Brain dataset is the combination of the four mouse brain datasets. **(c)** River plots illustrate the predicted cell types and their relationships to the actual cell types on the query data LargeIntestineA when using the LargeIntestineB data as the reference. **(d)** UMAP visualization of representation generated by each method on the query data, with actual cell types represented by different colors.

Fig. 3 | Performance across platform or tissue datasets. (a) Boxplots summarize the ACC scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. The hollow red dot within the boxplot represents the average values, while black dots denote outliers. This analysis includes n=19 biologically independent paired cross-platform datasets. The x-axis represents the various methods, while the y-axis denotes the measured values. **(b)** UMAP visualization of the representation generated by each method on the query data Cerebellum when utilizing the MosP1 as the reference, with actual cell types represented by different colors. **(c)** Boxplots summarize the ACC scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. The hollow red dot within the boxplot represents the average values, while black dots denote outliers. This analysis includes n=22 biologically independent paired cross-tissue datasets. The x-axis represents the various methods, while the y-axis denotes the measured values. **(d)** UMAP visualization of the representation generated by each method on the query data Liver when utilizing the BoneMarrowB as the reference, with actual cell types represented by different colors.

Fig. 4 | Performance to utilize the multi-source data or the atlas data as the reference. (a) Comparative analysis of accuracy for SANGO and other competing methods when the reference data utilizes multi-source data from tissues mouse brain (consisting of four datasets) and Intestine (consisting of three datasets). For each tissue, we iteratively left one dataset as the query data and the rest as the multi-source data, resulting in 7 paired datasets. Boxplots summarize the ACC scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. The hollow red dot within the boxplot represents the average values, while black dots denote outliers. This analysis includes n=7 biologically independent paired datasets. The x-axis represents the various methods, while the y-axis denotes the measured values. **(b)** The bar plot depicts the accuracy of each method when using a real single-cell atlas consisting of 13 different tissues with about 80000 cells to annotate the lung tissue. **(c)** The UMAP visualization of the PBMC data, where cell labels were annotated by SANGO and Seurat, respectively. **(d)** Coverage plots of chromatin accessibility over signature or high expression genes across predicted Memory B and Naïve B cells: *TCL1A* for Naive cells, *FCGR2B* and *TEX9* for Memory B cells. The term "Region" in each subgraph represents a genomic region of the chromosome.

Fig. 5 | Revealing biological implications for normal tissues. (a) Coverage plots of chromatin accessibility for each predicted cell within the normal cortex data over cell type-specific signature genes: *Neurod6* for excitatory neuron cells, *TMEM119* for microglia cells, *Mag* for oligodendrocytes cells. The term "Region" in each subgraph represents a genomic region of the chromosome. **(b)** Overrepresented DNA motifs were identified by cell type-specific accessibility peaks in excitatory neurons, microglia, and oligodendrocytes, respectively. **(c)** During the SNPsea analysis, the top 30 tissues exhibiting substantial enrichment were identified, considering both the excitatory neuron-specific peaks identified by SANGO and the background peaks. To assess significance, vertical dashed and solid lines served as indicators, representing the one-sided P-value cutoff at the 0.05 level. This criterion evaluates whether all genes collectively display enrichment specific to a given annotation. The heatmaps illustrate Pearson correlation coefficients (PCC) for pairs of expression profiles, arranged using hierarchical clustering via the unweighted pair-group method with arithmetic means (UPGMA). **(d)** Cicero utilized scATAC-seq data from excitatory neuron cells, microglia cells, and oligodendrocyte cells to predict cis-regulatory chromatin interactions. Cell type-specific peaks identified by SANGO were highlighted in cyan.

Fig. 6 | Identifying multi-level cell types in basal cell carcinoma data. (a) UMAP visualization of the cells within basal cell carcinoma data, cells are colored by actual cell types. **(b)** UMAP visualization of unknown probability scores for each cell in the basal cell carcinoma data, representing the probability that the cell belongs to an unknown cell type, with higher scores indicating a higher probability. **(c)** The cell types of cells within basal cell carcinoma data are predicted by SANGO when using a healthy adult human large atlas (HHLA) with merged immune cell types as the reference, the cells with higher probability scores are recognized as unknown cell types. **(d)** River plot mapping coarse-grained cell types annotated by SANGO (left) to actual cell types (right). **(e)** The coarse-grained immune cells are further classified into fine-grained immune cells (UMAP visualization) by SANGO when using tumor-infiltrating lymphocytes atlas from basal cell carcinoma (BCC_TIL) containing diverse subtypes of immune cells as the reference. **(f)** River plot mapping cell subtypes annotated by SANGO (left) to actual cell labels (right).

Figures

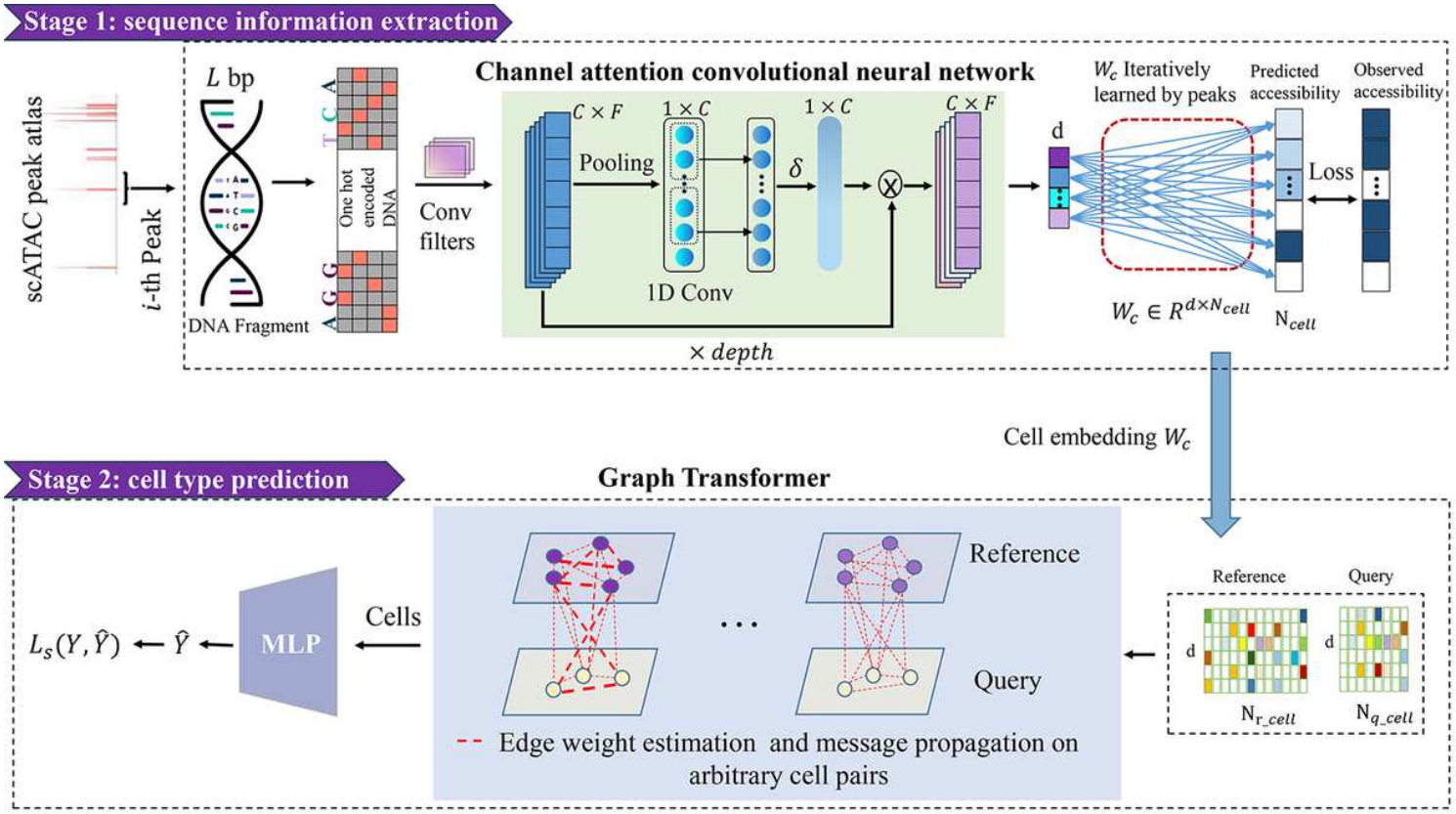

Fig. 1 | Schematic overview of SANGO framework for annotating cells within scATAC-seq data by integrating genome sequence. SANGO consists of two stages for sequence information extraction and cell type prediction, respectively. In stage 1, around the i-th peak, an input of L-bp length DNA sequence is extracted and one-hot encoded into an $L \times 4$ matrix. The matrix undergoes initial processing with C convolutional filters to generate the feature matrix with dimensions of $C \times F$. Subsequently, the matrix is inputted into a channel attention 1D convolutional neural network with the Sigmoid function δ and channel-wise multiplication $\otimes$. This is followed by a bottleneck layer to learn the d-dimensional embeddings of the peak. The embeddings are subsequently used to predict binary accessibilities of the peak for all N_{cell} cells through a dense linear network transformation with a weight matrix W_c of size $d \times N_{cell}$. All learnable parameters in stage 1 are optimized iteratively by the binary cross-entropy loss over all peaks. Finally, the learned weights in the dense network serve as a d-dimensional representation for the N_{cell} cells. In stage 2, the learned representations of reference and query data are used to construct cell graph through similarity, and a graph transformer was employed to remove batch effects and predict cell labels $\hat{Y}$ according to the truth labels Y of N_{r_cell} cells in the reference data. Finally, the trained graph transformer is used to predict cell types of N_{q_cell} cells within the query data.

Figure 1

See image above for figure legend.

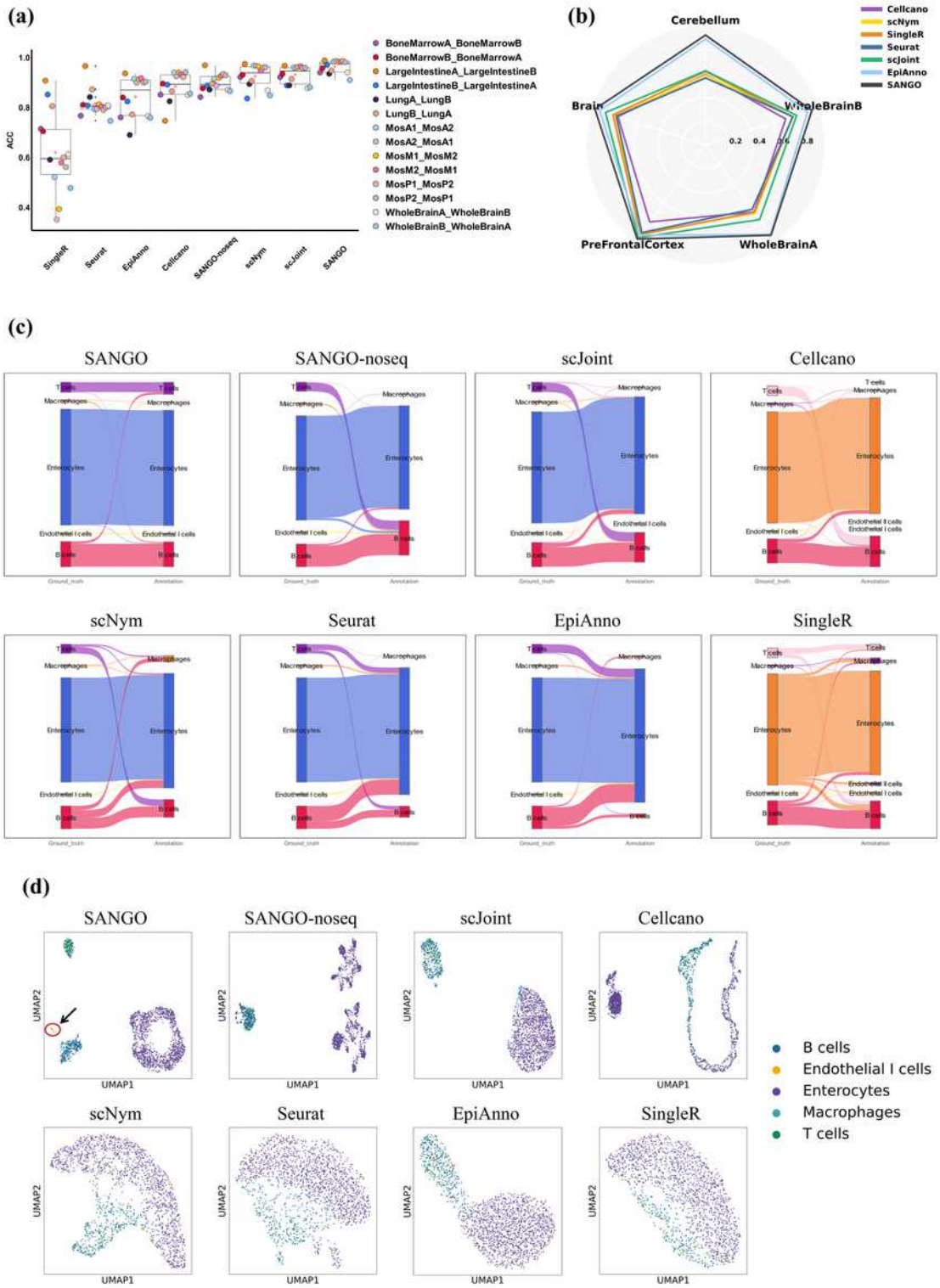

Figure 2

Performance of cell type annotation for intra-datasets. (a) Boxplots summarize the ACC scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. The hollow red dot within the boxplot represents the average values, while black dots denote outliers. This analysis includes n=14 biologically independent

paired intra-datasets. The x-axis represents the various methods, while the y-axis denotes the measured values. **(b)** The radar plot shows the accuracy of each method on five datasets with merged cell types when using the Forebrain data as the reference, where the Brain dataset is the combination of the four mouse brain datasets. **(c)**River plots illustrate the predicted cell types and their relationships to the actual cell types on the query data LargeIntestineA when using the LargeIntestineB data as the reference. **(d)** UMAP visualization of representation generated by each method on the query data, with actual cell types represented by different colors.

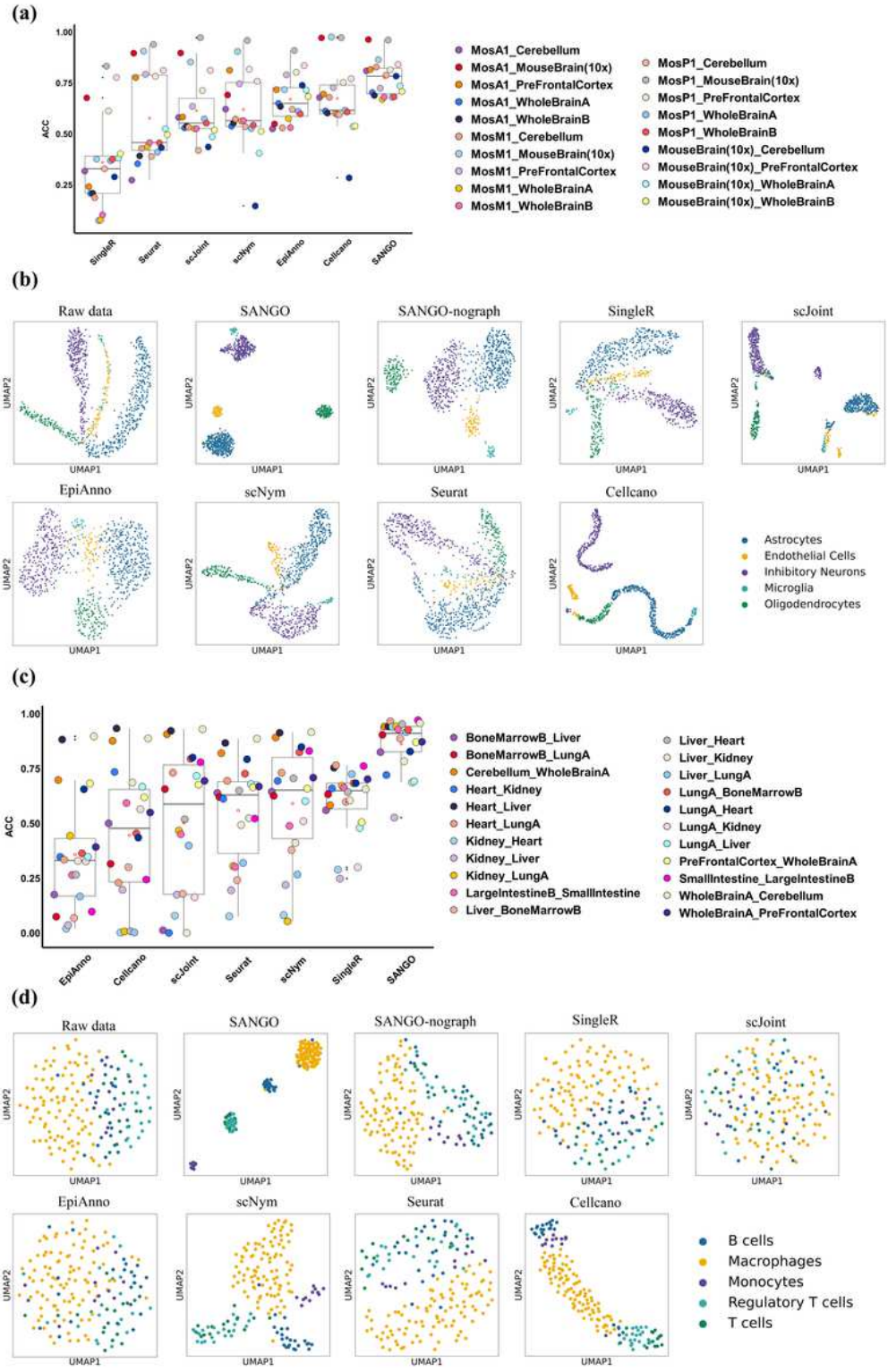

Figure 3

Performance across platform or tissue datasets. (a) Boxplots summarize the ACC scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. The hollow red dot within the boxplot represents the average values, while black dots denote outliers. This analysis includes n=19 biologically independent paired cross-platform datasets. The x-axis represents the various methods, while the y-axis denotes the measured values. **(b)**UMAP visualization of the representation generated by each method on the query data Cerebellum when utilizing the MosP1 as the reference, with actual cell types represented by different colors. **(c)** Boxplots summarize the ACC scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. The hollow red dot within the boxplot represents the average values, while black dots denote outliers. This analysis includes n=22 biologically independent paired cross-tissue datasets. The x-axis represents the various methods, while the y-axis denotes the measured values. **(d)**UMAP visualization of the representation generated by each method on the query data Liver when utilizing the BoneMarrowB as the reference, with actual cell types represented by different colors.

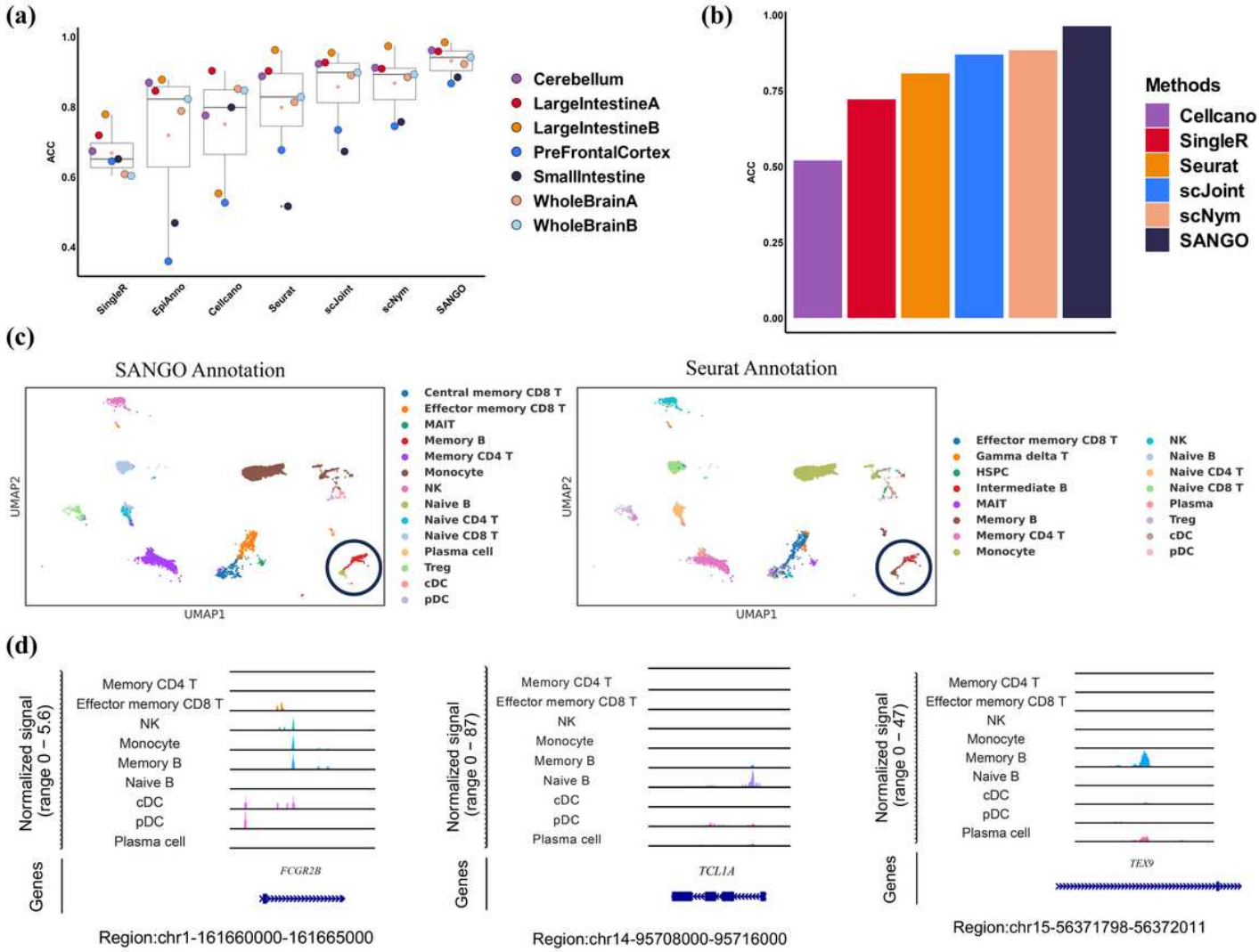

Figure 4

Performance to utilize the multi-source data or the atlas data as the reference. (a) Comparative analysis of accuracy for SANGO and other competing methods when the reference data utilizes multi-source data from tissues mouse brain (consisting of four datasets) and Intestine (consisting of three datasets). For each tissue, we iteratively left one dataset as the query data and the rest as the multi-source data, resulting in 7 paired datasets. Boxplots summarize the ACC scores for each method, which are defined by minima = 25th percentile – 1.5 × interquartile range (IQR), maxima = 75th percentile + 1.5 × IQR, interquartile range (hinges), and 1.5 times the interquartile range (whiskers), center = median and bounds of box = 25th and 75th percentile. The hollow red dot within the boxplot represents the average values, while black dots denote outliers. This analysis includes n=7 biologically independent paired datasets. The x-axis represents the various methods, while the y-axis denotes the measured values. **(b)** The bar plot depicts the accuracy of each method when using a real single-cell atlas consisting of 13 different tissues with about 80000 cells to annotate the lung tissue. **(c)** The UMAP visualization of the PBMC data, where cell labels were annotated by SANGO and Seurat, respectively. **(d)** Coverage plots of chromatin accessibility over signature or high expression genes across predicted Memory B and Naïve B cells:

TCL1A for Naive cells, *FCGR2B* and *TEX9* for Memory B cells. The term “Region” in each subgraph represents a genomic region of the chromosome.

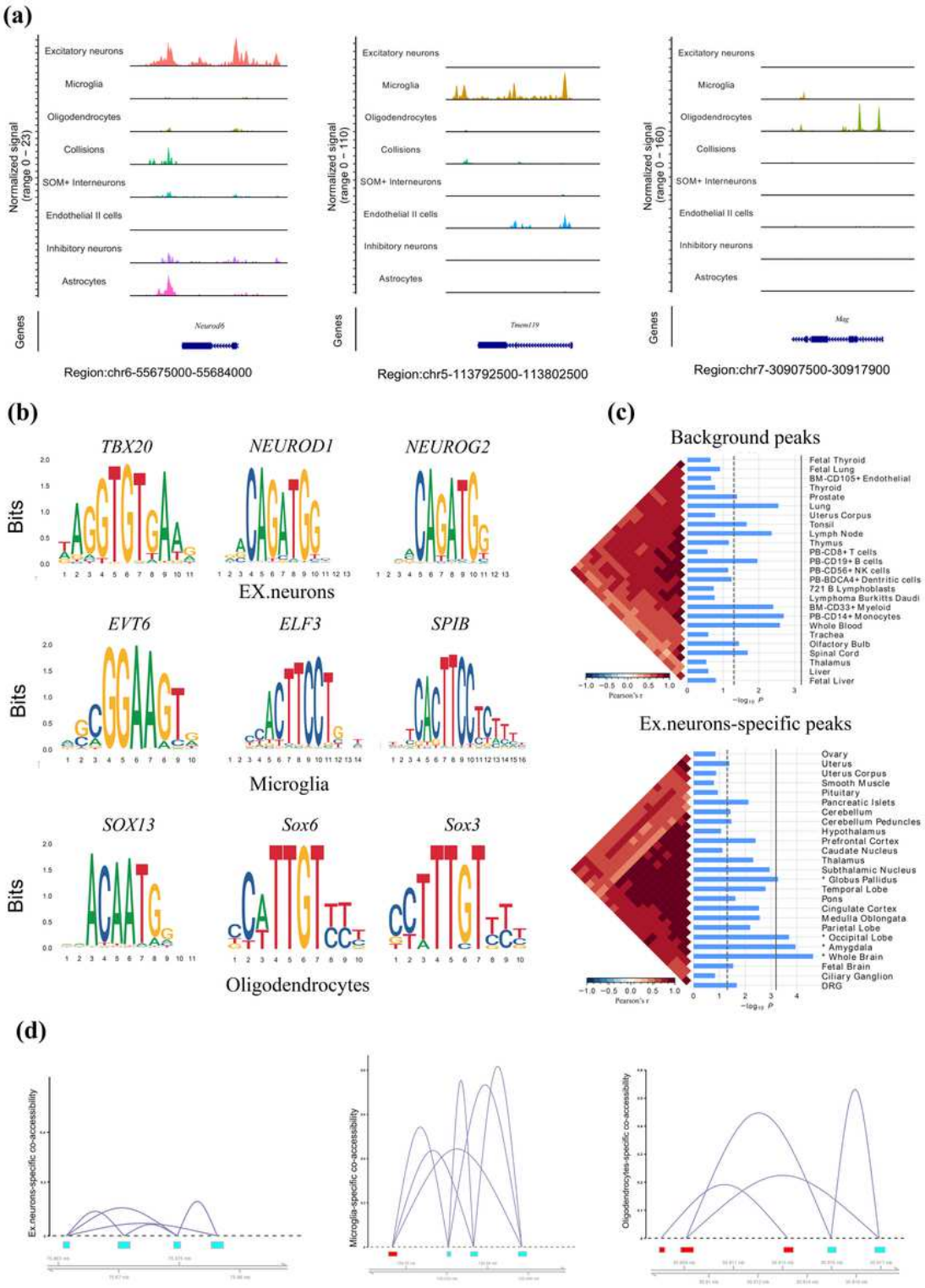

Figure 5

Revealing biological implications for normal tissues. (a) Coverage plots of chromatin accessibility for each predicted cell within the normal cortex data over cell type-specific signature genes: *Neurod6* for

excitatory neuron cells, *TMEM119* for microglia cells, *Mag* for oligodendrocytes cells. The term “Region” in each subgraph represents a genomic region of the chromosome. **(b)** Overrepresented DNA motifs were identified by cell type-specific accessibility peaks in excitatory neurons, microglia, and oligodendrocytes, respectively. **(c)** During the SNPsea analysis, the top 30 tissues exhibiting substantial enrichment were identified, considering both the excitatory neuron-specific peaks identified by SANGO and the background peaks. To assess significance, vertical dashed and solid lines served as indicators, representing the one-sided P-value cutoff at the 0.05 level. This criterion evaluates whether all genes collectively display enrichment specific to a given annotation. The heatmaps illustrate Pearson correlation coefficients (PCC) for pairs of expression profiles, arranged using hierarchical clustering via the unweighted pair-group method with arithmetic means (UPGMA). **(d)** Cicero utilized scATAC-seq data from excitatory neuron cells, microglia cells, and oligodendrocyte cells to predict cis-regulatory chromatin interactions. Cell type-specific peaks identified by SANGO were highlighted in cyan.

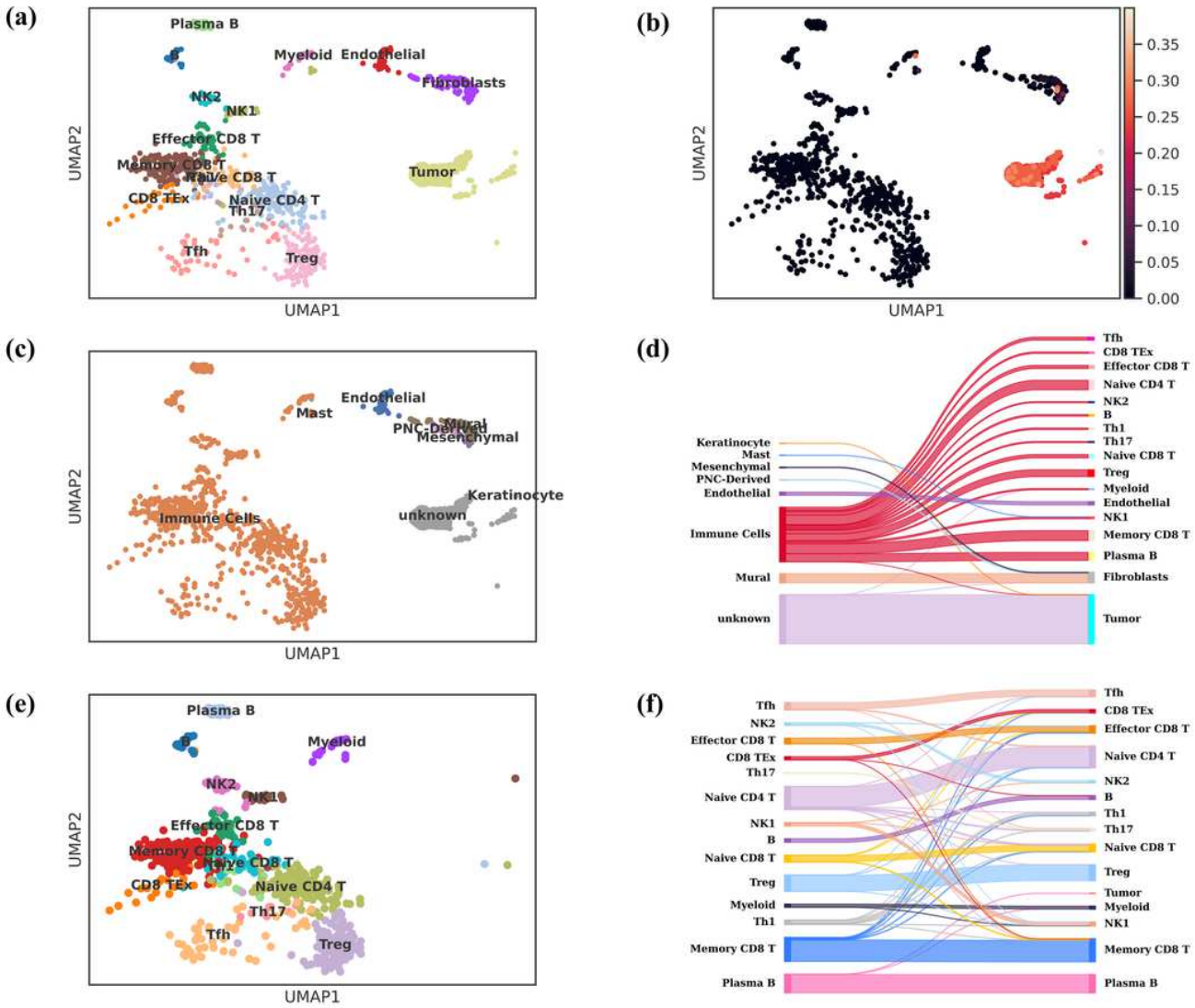

Figure 6

Identifying multi-level cell types in basal cell carcinoma data. (a) UMAP visualization of the cells within basal cell carcinoma data, cells are colored by actual cell types. **(b)** UMAP visualization of unknown probability scores for each cell in the basal cell carcinoma data, representing the prob-ability that the cell belongs to an unknown cell type, with higher scores indicating a higher probability. **(c)** The cell types of cells within basal cell carcinoma data are predicted by SANGO when using a healthy adult human large atlas (HHLA) with merged immune cell types as the reference, the cells with higher probability scores are recognized as unknown cell types. **(d)** River plot mapping coarse-grained cell types annotated by SANGO (left) to actual cell types (right). **(e)** The coarse-grained immune cells are further classified into fine-grained immune cells (UMAP visualization) by SANGO when using tumor-infiltrating lymphocytes atlas from basal cell carcinoma (BCC_TIL) containing diverse subtypes of immune cells as the reference. **(f)** River plot mapping cell subtypes annotated by SANGO (left) to actual cell labels (right).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- InventoryofSupportingInformationSANGO.docx
- ReportingSummary50471attach68119.pdf
- checklistYangAuthorGuidance17102581971.docx
- nreditorialpolicychecklist.pdf
- nrreportingsummary.pdf
- supplementary.pdf