

METHOD

Open Access



# MultiMAP: dimensionality reduction and integration of multimodal data

Mika Sarkin Jain<sup>1,2\*</sup>, Krzysztof Polanski<sup>2</sup>, Cecilia Dominguez Conde<sup>2</sup>, Xi Chen<sup>2,3</sup>, Jongeun Park<sup>2,4</sup>, Lira Mamanova<sup>2</sup>, Andrew Knights<sup>2</sup>, Rachel A. Botting<sup>5</sup>, Emily Stephenson<sup>5</sup>, Muzlifah Haniffa<sup>2,5</sup>, Austen Lamacraft<sup>1</sup>, Mirjana Efremova<sup>2,6\*</sup>  and Sarah A. Teichmann<sup>1,2\*</sup>

\* Correspondence: [mikasarkinjain@gmail.com](mailto:mikasarkinjain@gmail.com); [m.efremova@qmul.ac.uk](mailto:m.efremova@qmul.ac.uk); [st9@sanger.ac.uk](mailto:st9@sanger.ac.uk)

<sup>1</sup>Theory of Condensed Matter, Dept Physics, Cavendish Laboratory, University of Cambridge, JJ Thomson Ave, Cambridge CB3 0HE, UK

<sup>2</sup>Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK  
Full list of author information is available at the end of the article

## Abstract

Multimodal data is rapidly growing in many fields of science and engineering, including single-cell biology. We introduce MultiMAP, a novel algorithm for dimensionality reduction and integration. MultiMAP can integrate any number of datasets, leverages features not present in all datasets, is not restricted to a linear mapping, allows the user to specify the influence of each dataset, and is extremely scalable to large datasets. We apply MultiMAP to single-cell transcriptomics, chromatin accessibility, methylation, and spatial data and show that it outperforms current approaches. On a new thymus dataset, we use MultiMAP to integrate cells along a temporal trajectory. This enables quantitative comparison of transcription factor expression and binding site accessibility over the course of T cell differentiation, revealing patterns of expression versus binding site opening kinetics.

## Background

Multimodal data is rapidly growing in single-cell biology and many other fields of science and engineering. Emerging single-cell technologies are providing high-resolution measurements of different features of cellular identity, including single-cell assays for gene expression, protein abundance [1, 2], chromatin accessibility [3], DNA methylation [4], and spatial resolution [5]. Large-scale collaborations, including the Human Cell Atlas international consortium [6, 7], are generating an exponentially increasing amount of data, using these technologies. Each technology provides a unique view of cellular biology and has different strengths and weaknesses. Integrating these measurements to study a single biological system will open avenues for a more comprehensive view of cellular identity, cell-cell interactions, developmental dynamics, and tissue structure [8].

The integration of multi-omic data poses several challenges [9]. Different omics technologies measure distinct unmatched features with different underlying distributions and properties and hence produce data of different dimensionality. This makes it difficult to place data from different omics in the same feature space. Additionally, omics

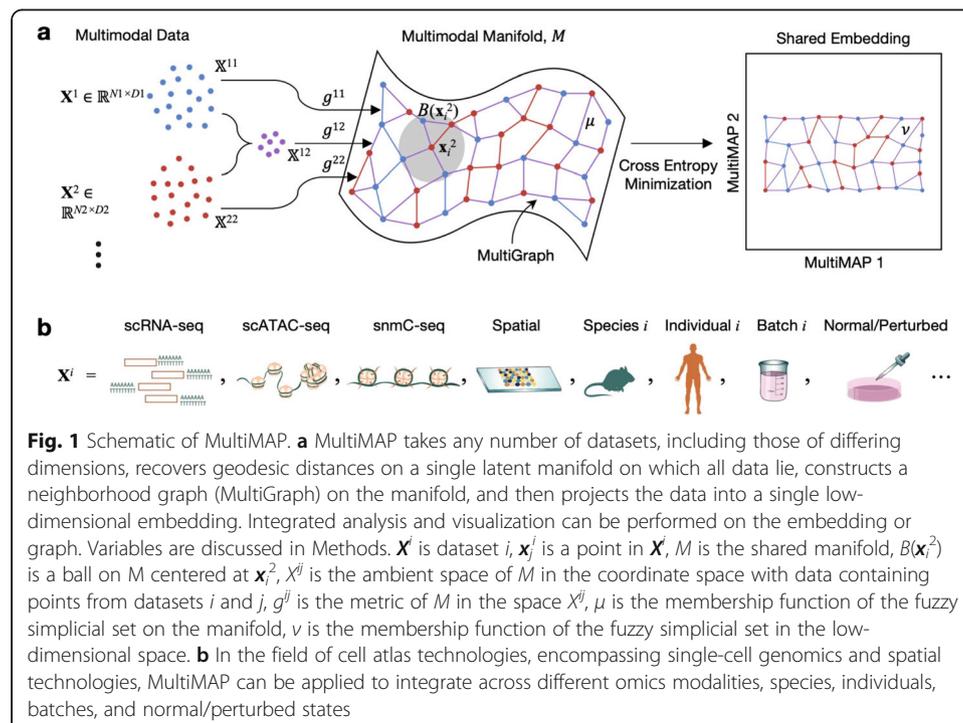


© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

technologies can also have different noise and batch characteristics which are challenging to identify and correct. Furthermore, as multi-omic data grows along two axes, the number of cells per omic and the number of omics per study, integration strategies need to be extremely scalable.

Most data integration methods project multiple measurements of information into a common low-dimensional representation to assemble multiple modalities into an integrated embedding space. Recently published methods employ different algorithms to project multiple datasets into an embedding space, including canonical correlation analysis (CCA) [10], nonnegative matrix factorization (NMF) [11, 12], or neural network models [13]. These methods have demonstrated utility, yet suffer from shortcomings, including challenges with scaling and being limited to consideration of features shared across data sets (e.g., the same genes). A further drawback is that methods that use linear models, such as CCA and NMF, are unable to capture nonlinear differences between datasets.

Here we introduce a new method that overcomes all these limitations: MultiMAP, an algorithm for the dimensionality reduction and integration of multiple datasets. MultiMAP integrates data by constructing a nonlinear manifold on which diverse high-dimensional data reside and then projecting the manifold and data into a shared low-dimensional embedding space (Fig. 1). MultiMAP generalizes the UMAP algorithm [14] to the setting of multiple datasets with different dimensions, while implementing novel techniques for estimating the manifold, constructing a joint graph on the manifold, and optimizing the low-dimensional embedding to account for the limitations and the challenges of multimodal and multi-omic data. In contrast to other integration strategies for single-cell data, MultiMAP can integrate any number of datasets, is not restricted to a linear mapping, leverages features that are not present in all datasets



(i.e., datasets can be of different dimensionalities), allows the user to specify the influence of each dataset on the embedding, and is effortlessly scalable to large datasets. The ability of MultiMAP to integrate datasets of different dimensionalities allows leveraging information that is not considered by methods that operate in a shared feature space. The power of MultiMAP's consideration of all features in each modality can be illustrated when integrating the 20,000-feature gene space of scRNAseq data with the 100,000-feature peak space of scATAC-seq data: by taking into account the full epigenetic landscape of cells, including distal enhancers, in addition to the transcriptome, cell states such as, e.g., memory T cells versus naive T cells, where it is known that memory is largely epigenetically encoded, can be more sharply defined in the manifold space.

We apply MultiMAP to challenging synthetic multimodal data, demonstrate its ability to integrate a wide range of single-cell omics datasets, and benchmark it against popular integration algorithms using a variety of performance metrics. Finally, we apply MultiMAP to the study of T cell development with new scATAC-seq data from fetal thymi. We show that MultiMAP can co-embed datasets across different technologies and modalities, while at the same time preserving the structure of the data, even with extensive biological and technical differences. The resulting embedding and shared neighborhood graph (MultiGraph) can be used for simultaneous visualization and integrative analysis of multiple datasets. With respect to single-cell genomics data, this allows for standard analysis on the integrated data, such as cluster label transfer, joint clustering, and trajectory analysis.

## Results

### The MultiMAP framework

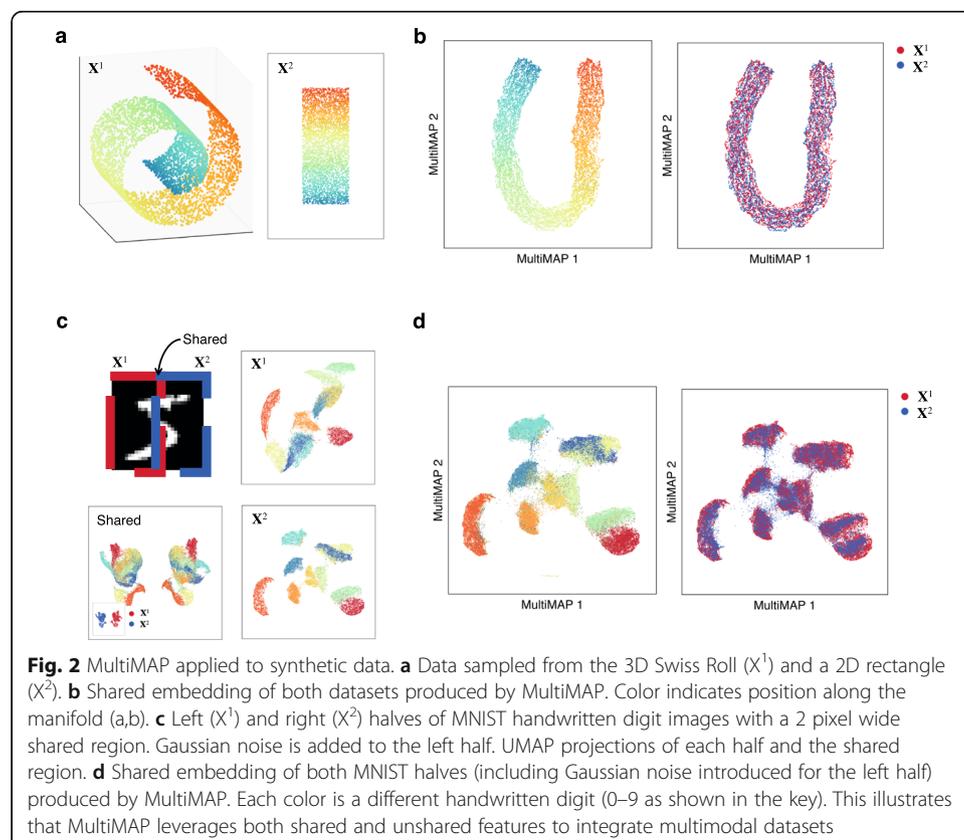
We introduce MultiMAP, an approach for integration and dimensionality reduction of multimodal data. MultiMAP is based on a framework of Riemannian geometry and algebraic topology and generalizes the UMAP framework to the setting of multiple datasets each with different dimensionality. MultiMAP takes as input any number of datasets of potentially differing dimensions and recovers geodesic distances on a single latent manifold on which all of the data is uniformly distributed. The distances are calculated between data points of the same dataset by normalizing distances with respect to a neighborhood distance specific to the dataset, and between data points of different datasets by normalizing distances between the data in a shared feature space with respect to a neighborhood parameter specific to the shared feature space. These distances are then used to construct a neighborhood graph (MultiGraph) on the manifold. Finally, the data and manifold space are projected into a low-dimensional embedding space by minimizing the cross entropy of the graph in the embedding space with respect to the graph in the manifold space. MultiMAP allows the user to modify the weight of each dataset in the cross entropy loss, and thus to modulate the contribution of each dataset to the layout. Integrated analysis can be performed on the embedding or the graph, and the embedding also provides an integrated visualization. The mathematical formulation of MultiMAP is elaborated in Additional file 2: Supplementary Methods [15–20].

In order to study MultiMAP in a controlled setting, we first applied it to two synthetic examples of multimodal data ("Methods"). The first synthetic data consists of

points sampled randomly from the canonical 3D “Swiss Roll” surface and the 2D rectangle (Fig. 2a). The dataset is considered multimodal data, because samples are drawn from different feature spaces but describe the same rectangular manifold. In addition, we are given the position along the manifold of 1% of the data. This synthetic setting illustrates that MultiMAP can integrate data in a nonlinear fashion and operate on datasets of different dimensionality, because data points along a similar position on the manifold are near each other in the embedding (Fig. 2b). The MultiMAP embedding properly unrolls the Swiss Roll dataset, indicating that the projection is nonlinear. The embedding also appears to preserve aspects of both datasets; the data is curved and at the same time unrolled.

To determine if MultiMAP can effectively leverage features unique to certain datasets, we used the MNIST database [21], where handwritten images were split horizontally with thin overlap (Fig. 2c; see “Methods” for details). The two datasets can be considered multimodal because they have different feature spaces but describe the same set of digit images. The thin overlapping region of the two halves is not enough information to create a good embedding of the data (Fig. 2c). Many distinct digits are similar in this thin central sliver, and hence they cluster together in the feature space of this sliver. Indeed, in a UMAP projection of the data in the shared feature space of this overlap, the clusters of different digits are not as well separated as in the UMAP projections of each half (Fig. 2c).

A multimodal integration strategy that effectively leverages all features would use the features unique to each half to separate different digits, and the shared space to bring



the same digits from each dataset close together (Fig. 2d). We show that with MultiMAP the different modalities are well mixed in the embedding space and the digits cluster separately, despite mostly different feature spaces and noise being added to only the second dataset. This indicates that MultiMAP is leveraging the features unique to each dataset and is also robust to datasets with different levels of noise.

Moreover, MultiMAP has weight parameters  $\omega^v$  which control the contribution of each dataset  $X^v$  to the final embedding, allowing the user to modulate which dataset has a greater influence on the MultiMAP embedding. When a dataset's weight is larger, its structure has a larger contribution to the MultiMAP embedding. Our results show that when integrating the MNIST data, for different choices of  $\omega^v$ , the datasets remain well integrated in the embedding space (Additional file 1: Fig. S1a,b). For our real biological datasets, we use fixed default values of the weighting parameter (0.8 for scRNA-seq and 0.2 for all other -omics), demonstrating that MultiMAP produces robust integration without the need to adjust the weighting.

Finally, to illustrate that our assumption of a shared manifold is robust to variable levels of overlap across datasets, we used MultiMAP to integrate datasets with varying numbers of shared clusters in the MNIST data (Additional file 1: Fig. S2). Our results show that MultiMAP is able to effectively integrate datasets that have only 1 out of 10 clusters shared between them. The transfer accuracy, silhouette score, and structure score of the MultiMAP integration remained largely constant as the number of overlapping clusters varied, demonstrating that MultiMAP is highly robust to differences in populations between datasets.

#### **MultiMAP integration of single-cell transcriptomics and chromatin accessibility**

Having shown that MultiMAP succeeds in integrating synthetic data, we applied the technique to real biological data. Epigenomic regulation underlies gene expression and cellular identity. Hence, integration of single-cell transcriptomics and epigenomics data provides an opportunity to investigate how epigenomic alterations regulate gene expression to determine and maintain cell identity. In addition, effective integration with transcriptomics data can improve the sensitivity and interpretability of the more sparse scATAC-seq data.

To assess MultiMAP's ability to integrate transcriptomic and epigenomic data, we applied it to integrate our previously generated high-coverage scATAC-seq data of mouse splenocytes [22] and generated corresponding single-cell transcriptomic profiles of the same tissue. The high coverage of the plate-based scATAC-seq data as well as the published cluster annotations of the subpopulations served as a good ground truth example to validate our method. Analysis of the transcriptomics data revealed similar subpopulations to the published scATAC-seq dataset, in addition to two RNA-specific clusters: a subpopulation of B cells with higher expression of Interferon-Induced (Ifit) genes and a subpopulation of proliferating cells (Additional file 1: Fig. S3a,b).

MultiMAP effectively integrated the two datasets, using both gene activity scores and the cell-type-specific epigenetic information outside of gene bodies. The different modalities are well mixed in the embedding space and cells annotated as the same type are close together, regardless of the modality for different choices of  $\omega^v$  (Fig. 3a, Additional file 1: Fig S1c,d). Next, we jointly clustered cells from both datasets using the

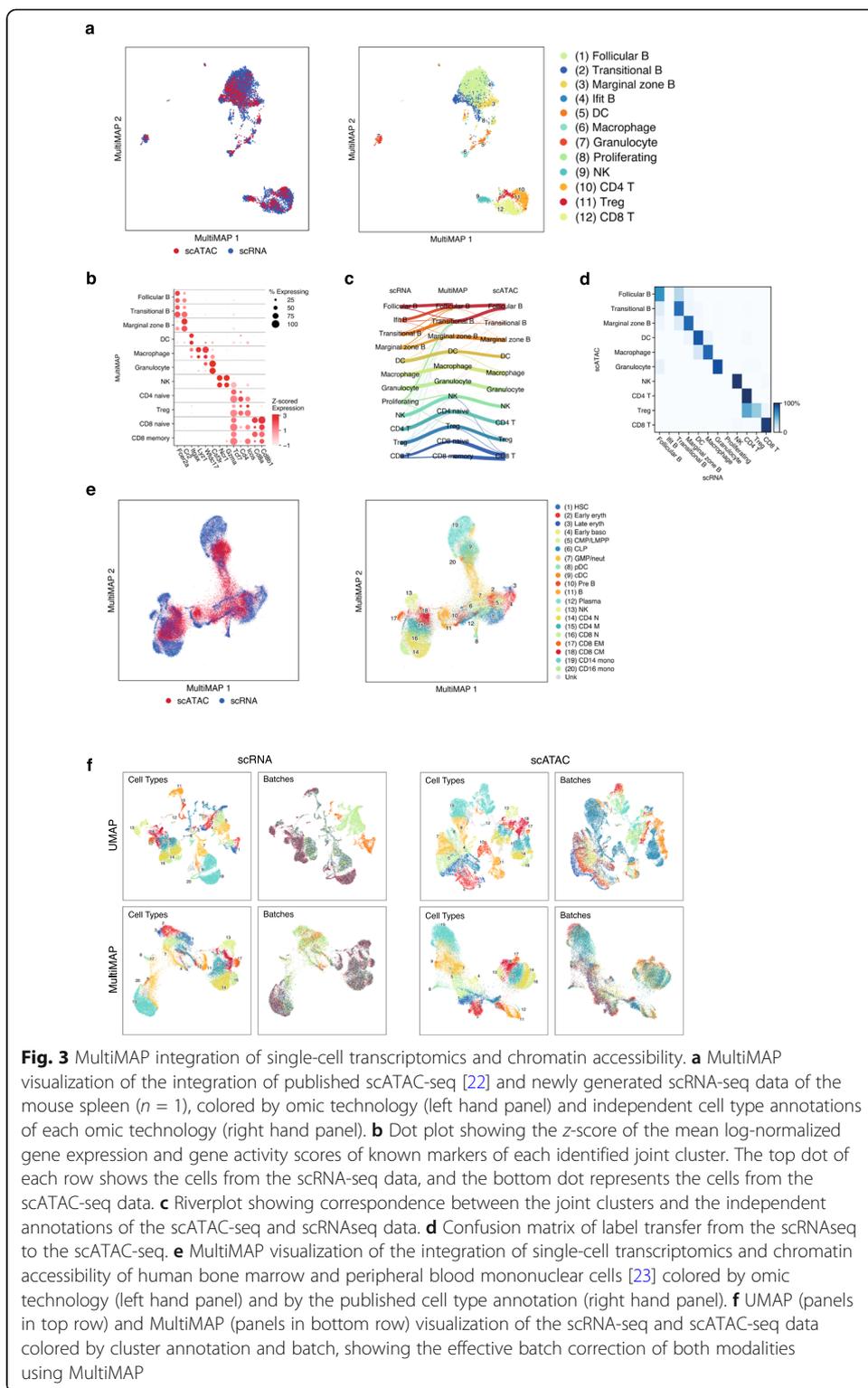
MultiGraph. This produced clusters with markers corresponding to known cell types [22] (Additional file 1: Fig S3c). The annotations produced by this joint clustering were generally consistent with independent annotations of each dataset (Fig. 3c). Two of the clusters determined to be proliferating cells and B cells with upregulated *Ifit* genes were found only in the scRNA-seq data, as expected (Fig. 3a, Additional file 1: Fig. S3b). In addition, the integration produced by MultiMAP is robust to different choices of the weight parameters (Additional file 1: Fig. S1c).

Further, we used the MultiGraph to directly predict the cell types of the scATAC-seq given the cell types of the scRNA-seq. Figure 3d shows the confusion matrix of the predictions, demonstrating that cells were generally annotated correctly. This illustrates the ability of MultiMAP to leverage annotation efforts of one omic technology to inform those of another. Interestingly, a small subset of cells from scRNA-seq previously annotated as T cells is now clearly separated on the MultiMAP plot, and clusters close to the B cells (Fig. 3a, Additional file 1: Fig S3). Doublet detection confirmed that this cluster is composed of doublet T/B cells. These doublets are spread throughout the UMAP plot of the scRNA-seq data, but are clearly distinct on the MultiMAP plot (Additional file 1: Fig. S3). This illustrates the power of MultiMAP both as a visualization tool, and to reveal new populations of cells.

Next, we applied MultiMAP to integration of multiple batches from each data modality, to assess the ability to account for batch effects. For this purpose, we used recently published scRNA-seq and scATAC-seq data of human bone marrow and peripheral blood mononuclear cells (PBMCs) [23]. This dataset consists of 16 experimental samples, representing different experimental batches. Another challenge is that cells are not in discrete clusters but rather on a continuum. MultiMAP is able to simultaneously correct batch effects and modality differences, integrating all 16 datasets into a consistent embedding (Fig. 3e). The different modalities are well mixed in the embedding and cells of the same type are close together, regardless of modality or batch. The cell-type annotations of all of the data were taken from the original publication [23], so they provide a good ground truth and independent validation of MultiMAP. Additionally, MultiMAP is able to correct batch effects present in different omics technologies. Applying MultiMAP to just the scRNA-seq data produces embedding that properly integrates cells of the same type regardless of batch, and the same is true when MultiMAP is applied to only the scATAC-seq data (Fig. 3f). It is also evident from this figure that clusters with cell types unique to a batch remain unmixed in the embedding. This indicates that MultiMAP is not forcing incompatible data to integrate and demonstrates that MultiMAP can integrate datasets even if they have extensive technical differences.

#### **MultiMAP integration of multiple modalities of mouse brain cells**

Recent advances in spatial sequencing technology enable the simultaneous measurement of gene expression and spatial locations of single cells, facilitating the study of tissue structure [5]. While these technologies provide spatial information, they often measure only a small fraction of the genes measured by scRNA-seq. Integration of spatial measurements and scRNA-seq has the potential to provide spatial context to



**Fig. 3** MultiMAP integration of single-cell transcriptomics and chromatin accessibility. **a** MultiMAP visualization of the integration of published scATAC-seq [22] and newly generated scRNA-seq data of the mouse spleen ( $n = 1$ ), colored by omic technology (left hand panel) and independent cell type annotations of each omic technology (right hand panel). **b** Dot plot showing the z-score of the mean log-normalized gene expression and gene activity scores of each identified joint cluster. The top dot of each row shows the cells from the scRNA-seq data, and the bottom dot represents the cells from the scATAC-seq data. **c** Riverplot showing correspondence between the joint clusters and the independent annotations of the scATAC-seq and scRNA-seq data. **d** Confusion matrix of label transfer from the scRNA-seq to the scATAC-seq. **e** MultiMAP visualization of the integration of single-cell transcriptomics and chromatin accessibility of human bone marrow and peripheral blood mononuclear cells [23] colored by omic technology (left hand panel) and by the published cell type annotation (right hand panel). **f** UMAP (panels in top row) and MultiMAP (panels in bottom row) visualization of the scRNA-seq and scATAC-seq data colored by cluster annotation and batch, showing the effective batch correction of both modalities using MultiMAP

scRNA-seq data, as well as to reveal finer grained biological differences in the spatial data by leveraging the greater number of cells and genes present in scRNA-seq data.

We applied MultiMAP to the integration of a Drop-seq scRNA-seq data of the mouse frontal cortex [24] and STARmap in situ gene expression dataset [25]. Despite the

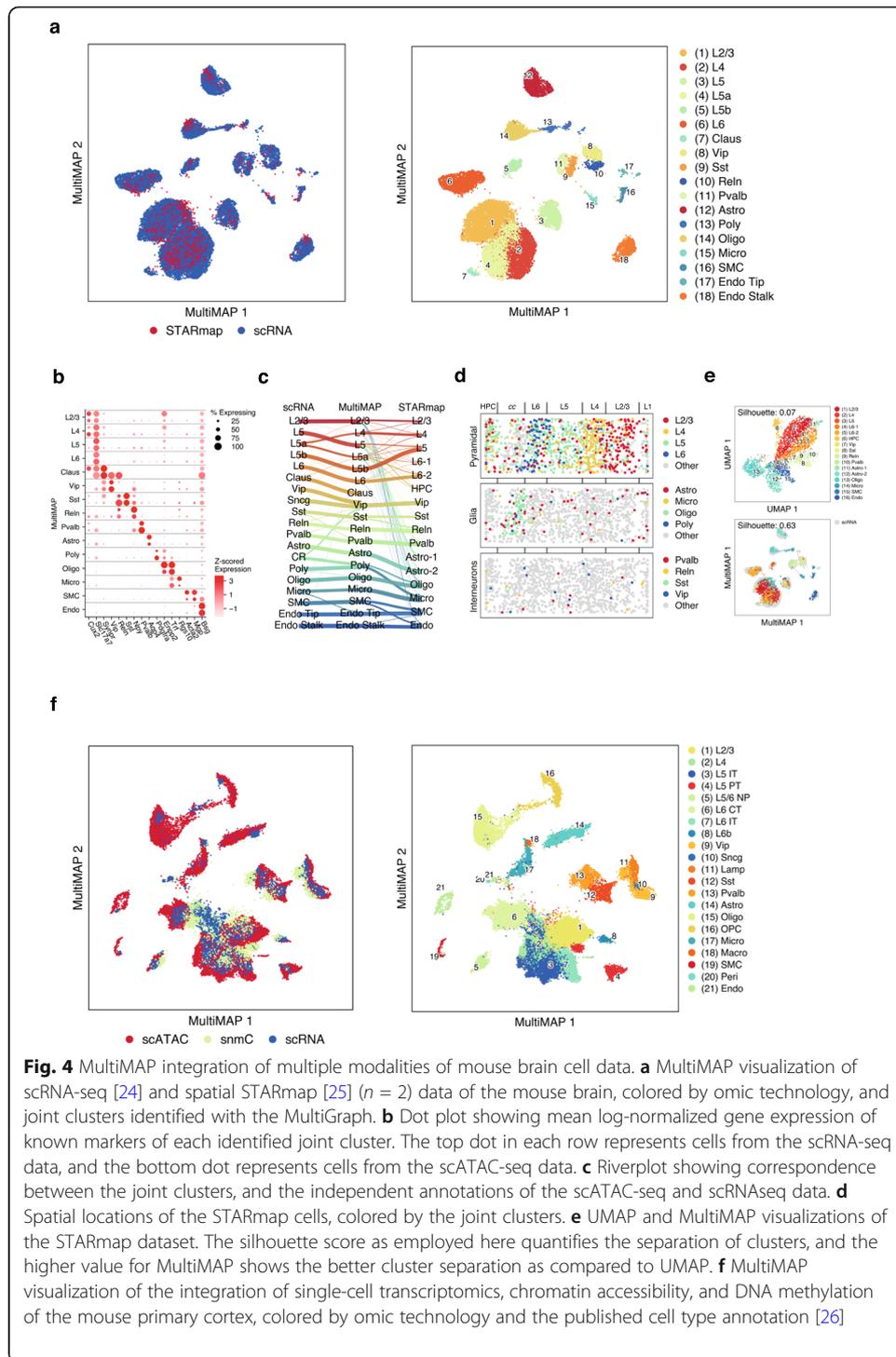
differences between the two datasets in the number of measured genes (only 1020 in STARmap) and the number of cells (71640 in Drop-seq versus 2137 in STARmap), our integrated analysis shows that MultiMAP successfully integrated the datasets. Clustering the integrated data using the MultiGraph produced clusters with markers corresponding to known cell types (Fig. 4a, b). One of the clusters, the claustrum, was found only in the scRNA-seq data, consistent with previous studies [11]. Integration with MultiMAP also resulted in improved cluster annotation for both datasets. The excitatory L4 neurons were previously only present in the STARmap data, as the motor cortex and prefrontal cortex that are part of the frontal cortex are considered to lack a layer 4 in mice [27]. However, after the integration, we also identified L4 cells in the scRNA-seq data previously annotated as L5 neurons (Fig. 4a, c, Additional file 1: Fig. S4). A similar population of pyramidal cells located between layers 3 and 5 were recently identified both with anatomical and single-cell studies [26, 28]. This was confirmed by expression of marker genes associated with L4, including *Cux2* and *Rorb* (Additional file 1: Fig. S4). This illustrates the power of MultiMAP to reveal new cell types.

MultiMAP also improves visualization of the STARmap data. Before integration with MultiMAP, many of the cell types of the spatial data did not cluster separately and were difficult to distinguish visually. In comparison, the MultiMAP embedding of the STARmap data exhibits tighter cell-type clusters and increased separation between cell types (Fig. 4e). This improvement was measured by the average Silhouette score in the embedding space, which is significantly larger for MultiMAP (Fig. 4e). We attribute the improved cluster separation seen in the MultiMAP plot to the fact that MultiMAP leverages all ~30 k genes, included genes that are present in only the scRNA-seq data, rather than only the ~1 k genes shared between the STARmap and scRNA-seq.

Integration with MultiMAP also enabled us to spatially locate all the joint cell types in the STARmap data, allowing study of the spatial structure of the tissue (Fig. 4d). The pyramidal neurons localize to layers 2–6 and oligodendrocytes localize to the layer below the cortex, whereas the interneurons do not appear to exhibit spatial organization. These observations are all consistent with the known spatial architecture of the mouse visual cortex [25].

To investigate the performance of MultiMAP on the integration of more than two modalities, we applied the approach to integrate recently published multi-omics datasets of the mouse primary motor cortex [26] consisting of 9 separate datasets, including 7 single-cell or single-nucleus transcriptomics datasets, one single-nucleus chromatin accessibility, and one single-nucleus DNA methylation (snmC-seq) dataset. MultiMAP successfully co-embedded more than 600,000 single-cell or single-nucleus samples assayed by six molecular modalities and identified the previously published cell subpopulations. The MultiMAP embedding displays good mixing of clusters from different modalities when the clusters correspond to the same cell type. Cell-type annotations were taken from the original publication of the data, so they provide a good ground truth and an independent validation of MultiMAP. We further see that cell types that exist in one modality, but not in the others, are not falsely aligned in the embedding space. This indicates that MultiMAP does not force incompatible data to integrate.

Finally, using the integration of scRNA-seq with the STARmap data, as well as the integration of the multi-omics spleen data, we assessed the impact of using only shared



**Fig. 4** MultiMAP integration of multiple modalities of mouse brain cell data. **a** MultiMAP visualization of scRNA-seq [24] and spatial STARmap [25] ( $n = 2$ ) data of the mouse brain, colored by omic technology, and joint clusters identified with the MultiGraph. **b** Dot plot showing mean log-normalized gene expression of known markers of each identified joint cluster. The top dot in each row represents cells from the scRNA-seq data, and the bottom dot represents cells from the scATAC-seq data. **c** Riverplot showing correspondence between the joint clusters, and the independent annotations of the scATAC-seq and scRNAseq data. **d** Spatial locations of the STARmap cells, colored by the joint clusters. **e** UMAP and MultiMAP visualizations of the STARmap dataset. The silhouette score as employed here quantifies the separation of clusters, and the higher value for MultiMAP shows the better cluster separation as compared to UMAP. **f** MultiMAP visualization of the integration of single-cell transcriptomics, chromatin accessibility, and DNA methylation of the mouse primary cortex, colored by omic technology and the published cell type annotation [26]

vs. all features. We find that using all features greatly improves the integration and results in embeddings that are visually and quantitatively superior, according to four performance metrics (Additional file 1: Fig. S5). This illustrates that non-shared features can be extremely helpful and demonstrates an advantage of MultiMAP over other methods which do not consider non-shared features.

### Benchmarking

We assessed and benchmarked the performance of MultiMAP against several popular approaches for integrating single-cell multi-omics, including Seurat [10], LIGER [11], iNMF [12], Conos [29], and GLUER [30].

These integration approaches differ in key regards, summarized in Fig. 5d. We used a diversity of performance metrics to comprehensively compare MultiMAP with other approaches, including transfer accuracy, silhouette score, alignment, preservation of the structure, and runtime. With these metrics, we quantified the separation of the joint clusters, how well mixed the datasets were after integration and how well they preserved the structure in the original datasets to investigate whether the methods integrate populations across datasets without blending distinct populations together. To measure transfer accuracy and silhouette score, we use cell-type annotations generated by other publications or different members of our lab group as ground truth, to ensure an independent validation. Out alignment and structure preservation metrics are based on the structural qualities of MultiMAP's output and do not rely on ground truth labels. This provides an orthogonal and complementary form of validation to using ground truth labels.

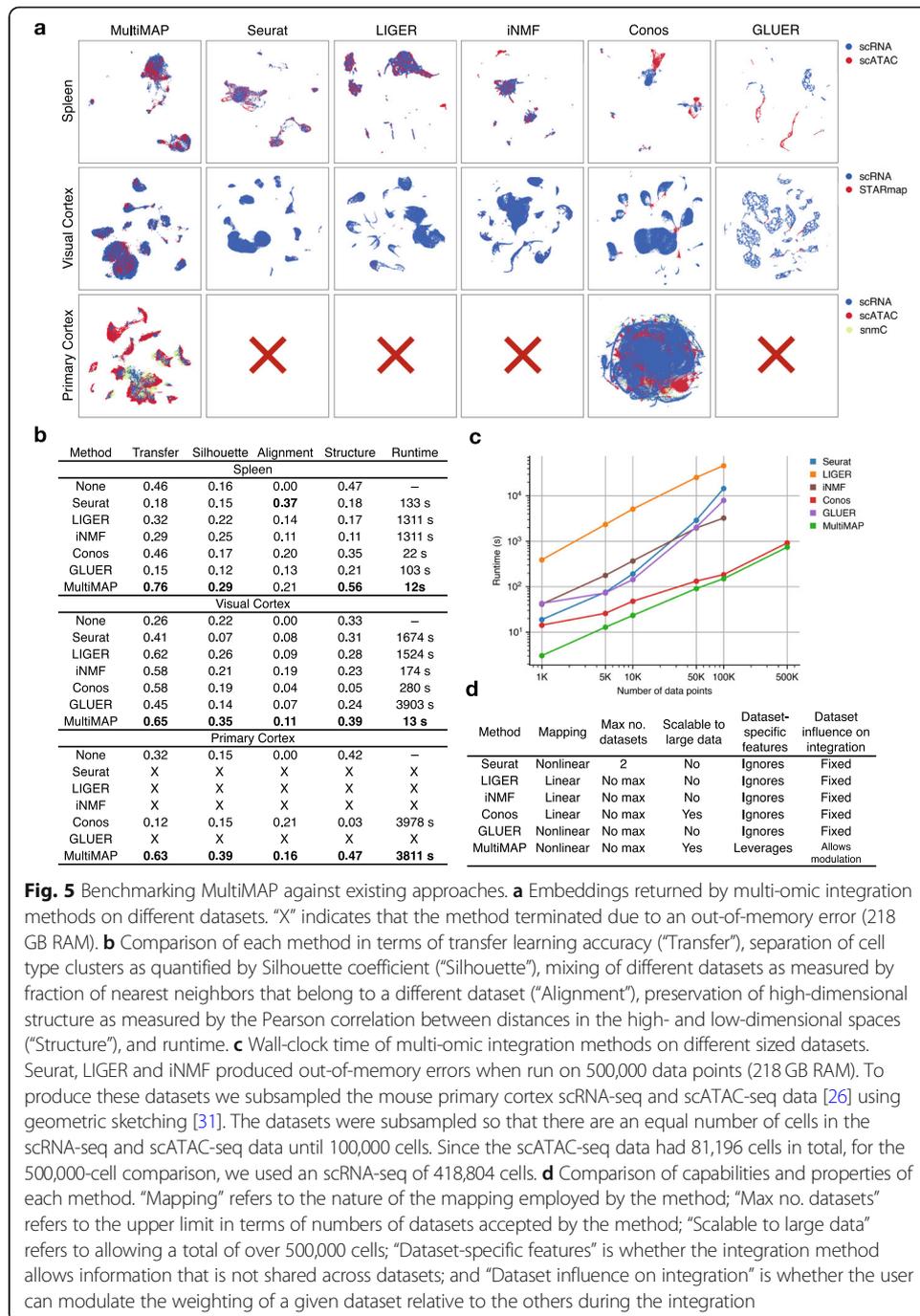
To this end, we generated single-nucleus data from human PBMCs using the Multiome ATAC + RNA kit. We obtained a PBMC atlas of 6344 nuclei of high-quality ATAC + RNA profiles. We analyzed and annotated the RNA and ATAC data separately, revealing all the known major PBMC types: CD14 and CD16 monocytes, cDCs and pDCs, naive and effector CD4 and CD8 T cells, Tregs, MAIT and gamma-delta T cells, NK and ILCs, naive and memory B cells, and plasmablasts (Additional file 1: Fig. S6a). Most cell types were well separated in both modalities with the exception of the NK and ILC clusters and the gamma-delta and the CD8 effector T cells that blended together in the ATAC data.

We used the PBMCs as a gold standard dataset to benchmark MultiMAP against the four other methods. As shown in the co-embedding and the metrics, MultiMAP successfully integrated the cell types across modalities and outperformed other methods (Additional file 1: Fig. S6b,c). The label transfer accuracy was particularly striking, with MultiMAP achieving a much higher score compared to other methods.

Furthermore, we also benchmarked MultiMAP using a variety of multi-omics data with published cell-type annotations, including the transcriptomics and chromatin accessibility spleen data, scRNA-seq and STARmap of the visual cortex, and the multi-omics data of the primary cortex. For all datasets, MultiMAP achieves top or near top performance on all metrics (Fig. 5). a, bThe embeddings produced by MultiMAP prove superior for transferring cell-type annotations between datasets, separating clusters of different cell populations, integrating datasets in a well-mixed manner, and capturing the high-dimensional structure of each dataset.

Critically, MultiMAP is significantly faster and more scalable than all other benchmarked methods, and significantly faster than LIGER and Seurat (Fig. 5c). Seurat, LIGER and iNMF were not able to scale to the primary cortex data of 600 k, producing out-of-memory errors despite access to 218 GB of RAM.

Finally, to assess the performance of MultiMAP for batch correction, we also applied it to three scRNA-seq studies of the human pancreas [32–34] that were recently used for comparison of eight batch correction methods [35]. Even though the main purpose of MultiMAP is the integration of several different omic technologies, MultiMAP



outperformed all other well-established batch correction methods in the field, demonstrating that MultiMAP can correct batches and integrate multiple omics data simultaneously (Additional file 1: Fig. S7).

### MultiMAP reveals patterns of T cell maturation along a multi-omic trajectory

Single-cell transcriptomics has enabled reconstruction of developmental trajectories and the study of dynamic processes such as differentiation and reprogramming. Bulk

RNA-seq and ATAC-seq data have further revealed regulatory events driving these processes [36]. However, joint analysis of single-cell expression and chromatin accessibility profiles along a time course trajectory would allow the study of dynamic chromatin regulation alongside gene expression, elucidating the epigenomic drivers of transcriptional change [37, 38].

In order to investigate the potential of integrating multi-omic data along a common differentiation trajectory, we focused on T cell development in the thymus. The thymus is an organ essential for the maturation and selection of T cells. Precursor cells migrate from the fetal liver and bone marrow to the thymus where they develop into different types of mature T cells [39]. We recently provided a comprehensive single-cell transcriptomics atlas of the human thymus during development, childhood, and adult life, and computationally predicted the trajectory of T cell development from early progenitors to mature T cells [39]. To expand on this and further investigate the gene regulatory mechanisms driving T cell development, we generated single-cell transcriptomics and chromatin accessibility data from a human fetal thymus sample at 10 weeks of gestation.

Clustering of the scRNA-seq data revealed cell types identified in our recently published transcriptomic thymus cell atlas [39], including several subpopulations of T cells across different stages of development, fibroblasts, endothelial cells, erythrocytes, thymic epithelial cells (TECs), NK and ILC3 cells, and macrophages and dendritic cells (Additional file 1: Fig. S8). The sparse scATAC-seq and the continuous nature of cell types along the maturation trajectory made it difficult to cluster the ATAC cells into different T cell subtypes (Additional file 1: Fig. S8). However, the integration with MultiMAP and the joint clusters obtained using the MultiGraph corresponded to the published thymus cell types [39] (Fig. 6a, b), allowing us to correctly annotate the cell types of the scATAC-seq data. Comparison with other integration methods shows that MultiMAP, by taking advantage of all features, outperforms other integration methods (Additional file 1: Fig S8f).

We then selected the T cell populations identified from the joint clustering and performed diffusion map pseudotime analysis using the alignment MultiMAP graph. The reconstructed development trajectory showed a continuous differentiation with the same trend as the published study, starting from early double negative (DN) CD4-CD8-, gradually progressing to double positive (DP) CD4+ CD8+ T cells, and then differentiating into single positive (SP) mature CD8+ or CD4+ T cells. Hallmark genes of T cell differentiation varied along the inferred pseudotime in a manner consistent with [39] (Fig. 6d), serving as validation of the trajectory inference and the integration produced by MultiMAP.

To identify transcription factors (TFs) that potentially regulate T cell development, we studied changes in TF expression and TF binding site accessibility along the differentiation trajectory. The top variable TFs/TF binding sites along the trajectory included many TFs that have been previously shown to be involved in T cell differentiation, including GATA3, SPI1, MEF2C, ERG, TCF3, TCF4, TFAP4, MYBL2, STAT1, NR4A2, and others [36, 39, 40] (Fig. 6e, Additional file 1: Fig. S9). The TFs that most varied along the trajectory showed changes in motif accessibility at the transition between the late DN and early DP stage of differentiation as shown before [40].

Moreover, our integrated trajectory allowed us to identify TFs where changes in motif accessibility and expression of the TF itself were closely coordinated, for example LEF1, IRF1, REL, FOS, and others, suggesting that these TFs actively regulate their target genes immediately and directly (Fig. 6e). In contrast, for TFs such as ETS1 and JUN, gene expression of the TF significantly precedes the accessibility of the corresponding TF binding sites, suggesting that additional regulatory mechanisms are potentially required for opening of the TF motifs.

## Discussion

Here we present a novel approach for dimensionality reduction and integration of multimodal data which considers full datasets, even when they have different feature spaces. MultiMAP embeds all datasets into a shared space to preserve both the manifold structure of each dataset independently, as well as in shared feature spaces. This enables both visualization and streamlined downstream analyses.

Existing integration methods require correspondence between all features profiled across omics technologies. In comparison, our method can incorporate different types of features, such as gene expression and open chromatin peaks or intergenic methylation, and thus takes advantage of the full power of multi-omics data. Ignoring the features unique to one dataset (as in most existing methods) may omit important information such as distinguishing features of certain subpopulations of cells, and thus yield an integrated embedding that does not distinctly cluster all subpopulations. Dataset-specific features often capture cell type or state heterogeneity not present in the shared features, for instance, cells may exhibit heterogeneity in chromatin structure outside of genes, or heterogeneity in genes other than the very small number of ~ 1000 genes shared by different technologies, as in the case of integrating scRNA with spatial data. Discarding this heterogeneity potentially stifles the discovery of new cell types or states. In addition, our integrated MultiGraph enables inference of joint developmental or differentiation trajectories that allow the study of dynamic chromatin regulation alongside gene expression. Using non-shared features that account for the full epigenetic landscape of cells, including distal enhancers, could help identify novel regulatory elements.

Another limitation is that linear integration approaches such as CCA and NMF are not able to correct for nonlinear distortions between datasets. In contrast, MultiMAP uses a nonlinear manifold learning approach and we demonstrate that it can effectively integrate in the presence of nonlinear differences between datasets. We find this to be a significant advantage of MultiMAP both for multi-omic integration and batch correction. Crucially, some methods such as LIGER, Seurat, and other CCA or NMF approaches are challenged by scaling to large datasets because they require matrix factorization. In contrast, MultiMAP readily scales to hundreds of thousands of cells due to its graph-based algorithm.

An additional feature of MultiMAP, not present in other existing strategies, is that the influence of each dataset on the shared embedding can be modulated. This is useful when integrating datasets of different qualities, or when aligning a query dataset to a reference dataset. Comparison with existing methods for integration shows that MultiMAP outperforms or has close to best performance in every aspect investigated.



MultiMAP is a robust and effective method for dimensionality reduction and integration of multimodal data and is extremely fast and scalable to massive datasets.

Using synthetic examples to illustrate the power of the method, we show that MultiMAP leverages the features unique to each dataset to effectively integrate and reduce the dimensionality of the data and is also robust to data with noise. Throughout our applications of MultiMAP to diverse single-cell multi-omic data, we demonstrate that our method can facilitate integration across transcriptomic, epigenomic, and spatially resolved datasets and derive biological insights jointly from multi-omic single-cell data. In addition, our method can align datasets across different technologies and modalities even with extensive biological and technical differences. Crucially, we show that MultiMAP is flexible enough to integrate datasets with different clusters and cell populations, illustrating that MultiMAP is applicable even when its central hypothesis is not strictly reflected by the data. The multimodal integration of three or more omics technologies opens many opportunities for the comprehensive study of tissues.

We note that our method is based on the hypothesis that multi-omics data are uniformly distributed on a latent manifold. A hypothesis of this sort, about the distribution of data in a latent space, is a central feature of many existing integration strategies. For example, CCA-based strategies (including Seurat and Conos) assume that the data reside in a maximally correlated manner in a latent space which is a linear projection of the original data. MultiMAP, in contrast, does not make as strong an assumption because we do not restrict the latent manifold to a linear projection of the data. While this kind of hypothesis is often realistic for data generated from the same tissue, there may be cases where this is not strictly the case. In practice, we find that MultiMAP can successfully accommodate datasets that depart from this central hypothesis, i.e., when clusters and cell populations are not shared across all datasets that are being integrated.

Perhaps the greatest potential lies in applying MultiMAP to datasets beyond those considered here. Integrative analysis with MultiMAP can be used to compare healthy and diseased states and identify pathologic features, or to uncover cell-type-specific responses to perturbations. Other examples include the integration of data across species to study the evolution of cell states and identify conserved cell types and regulatory programs. Along similar lines, the integration of *in vivo* with *in vitro* models such as organoids will reveal the quality or faithfulness of cells in a dish relative to their native counterparts. Finally, given the rapid development of joint multimodal single-cell genomics methods (e.g., CITEseq for protein and RNA, joint snRNA-, and ATACseq), it is relevant to point out that MultiMAP can be applied to multi-omic data acquired both from different cells and from the same cells.

## Conclusions

In this study, we introduce a novel algorithm for dimensionality reduction and integration of multiple datasets, which generalizes the UMAP algorithm to the setting of multiple datasets with different dimensions. MultiMAP is a nonlinear manifold learning algorithm that recovers a single manifold on which several datasets reside and then projects the data into a single low-dimensional space so as to preserve the manifold structure. It can be used for visualization of multimodal data, and as an integration approach that enables joint analyses.

We apply MultiMAP to the integration of a variety of single-cell transcriptomics, chromatin accessibility, methylation, and spatial data and show that it outperforms current approaches in preservation of high-dimensional structure, alignment of datasets, visual separation of clusters, transfer learning, and runtime. Furthermore, MultiMAP enables joint analysis of single-cell expression and chromatin accessibility profiles along a time course trajectory, allowing the study of dynamic chromatin regulation alongside gene expression.

In summary, given the broad appeal of dimensionality reduction methods (e.g., PCA, tSNE, UMAP), and the growth of multimodal data in many areas of science and engineering, we anticipate that MultiMAP will find wide and diverse use.

## Methods

### MultiMAP

MultiMAP (Fig. 1) is a new approach for the integration and dimensionality reduction of multimodal data based on a framework of Riemannian geometry and algebraic topology. MultiMAP takes as input any number of datasets of potentially differing dimensions. The datasets take the form  $X^i$ ,  $i = 1, 2, \dots$ , with  $x_j^i \in R^{D_i}$  being the  $j$ th point in dataset  $X^i$ . MultiMAP recovers geodesic distances on a single latent manifold  $M$  on which all of the data is uniformly distributed. The geodesic distances are calculated between data points of the same dataset by normalizing distances in each dataset's ambient space  $X^{ii}$  with respect to a neighborhood distance specific to the dataset, and between data points of different datasets by normalizing distances between the data in a shared ambient space  $X^{ij}$  with respect to a neighborhood distance specific to the shared feature space. We note that MultiMAP leverages both shared and unshared features. We do not wish for our approach to rely only on unmatched features because correspondence between shared features provides valuable information across omics. When integrating multi-omics data with MultiMAP, the ambient spaces are the PC components of each dataset's full feature space and of the shared feature space(s). These neighborhood distances are the radius of a constant-radius ball  $B$  on  $M$ . These distances are then used to construct a neighborhood graph (MultiGraph) on the manifold. Finally, the data and manifold space are projected into a low-dimensional embedding space by minimizing the cross entropy of the graph in the embedding space with respect to the graph in the manifold space. Specifically, this optimization minimizes cross entropy of a fuzzy set–representation  $(\nu, \{x_j^i\})$  of the graph in the embedding space with respect to a fuzzy set–representation  $(\mu, \{x_j^i\})$  of the graph in the manifold space. MultiMAP allows the user to modify the weight  $\omega^i$  of each dataset in the cross entropy loss, allowing the user to modulate the contribution of each dataset to the layout. Integrated analysis can be performed on the embedding or the graph, and the embedding also provides an integrated visualization. MultiMAP is novel in its graph construction on the shared manifold, the weights of the graph's edges, and the optimization of the low-dimensional embedding, each of which is motivated by manifold geometry. An extended description of MultiMAP, including mathematical background and motivation, is in the Supplementary information.

### Synthetic data

MultiMAP was applied to two synthetic examples of multimodal data, in order to study the technique in a controlled setting.

The first synthetic setting is schematized in Fig. 2a. This setting consists of one dataset ( $X^1$ ) of 10,000 points sampled randomly from the canonical 3D “Swiss roll” surface (generated with sklearn in Python), and a second dataset ( $X^2$ ) of 10,000 points sampled randomly from a 2D rectangle. The two datasets can be considered multimodal data because they have different feature spaces but describe a similar rectangular manifold. In addition, we are given the position along the manifold of 1% of the data. Distances between data in the different datasets are calculated for 1% of the data as the absolute differences between these positions. These distances are supplied to MultiMAP. The purpose of this setting is to determine if MultiMAP can integrate data in a nonlinear fashion and operate on datasets of different dimensionality.

The second synthetic setting is schematized in Fig. 2c. This setting consists of two datasets based on the MNIST database [41] which comprises 70,000  $28 \times 28$  pixel gray-scale images of handwritten digits 0–9. The first dataset ( $X^1$ ) consists of the  $28 \times 15$  pixel left half of each of images flattened into a 420 dimensional vector. The second dataset ( $X^2$ ) consists of the  $28 \times 15$  pixel right half of each of 70,000 digit images, also flattened into a 420 dimensional vector. Added to the first dataset is Gaussian noise with a mean of zero and a standard deviation equal to the maximum pixel value. The two halves overlap by a  $28 \times 2$  pixel region. Distances between data in the different datasets are calculated in this shared space and supplied to MultiMAP. The two datasets can be considered multimodal because they have different feature spaces but describe a similar population of digit images. The purpose of this setting is to determine if MultiMAP can effectively leverage features unique to certain datasets. The thin overlapping region of the two halves is not enough information to create a good embedding of the data. Many distinct digits are similar in this thin central sliver, and hence they should cluster together in the feature space of the two pixel overlap. Indeed, in a UMAP projection of the data in the shared feature space of this overlap, the clusters of different digits are not as well separated as in the UMAP projections of each half (Fig. 2c). A multimodal integration strategy that effectively leverages all features would use the features unique to each half to separate different digits, and the shared space to bring the same digits from each dataset close together.

### Acquisition and processing of human fetal thymic tissue

The developmental age was estimated from measurements of foot length and heel-to-knee length, and compared against a standard growth chart [42]. A piece of skin was collected from every sample for Quantitative Fluorescence-Polymerase Chain Reaction analysis using markers for the sex chromosomes and the following autosomes: 13, 15, 16, 18, 21, 22. The sample was of normal karyotype.

The tissue was processed immediately after isolation using enzymatic digestion. Tissue was transferred to a sterile  $10 \text{ mm}^2$  tissue culture dish and cut into  $< 1 \text{ mm}^3$  segments before being transferred to a 50-mL conical tube. Tissues were digested with 1.6 mg/mL collagenase type IV (Worthington) in RPMI (Sigma-Aldrich) supplemented with 10% (v/v) heat-inactivated fetal bovine serum (FBS; Gibco), 100 U/mL penicillin

(Sigma-Aldrich), 0.1 mg/mL streptomycin (Sigma-Aldrich), and 2 mM L-glutamine (Sigma-Aldrich) for 30 min at 37 °C with intermittent shaking. Digested tissue was passed through a 100- $\mu$ m filter, and cells collected by centrifugation (500g for 5 min at 4 °C). Cells were treated with 1 $\times$  red blood cells (RBC lysis buffer (eBioscience) for 5 min at room temperature and washed once with a flow buffer (PBS containing 5% (v/v) FBS and 2 mM EDTA) prior to cell counting. For scATAC-seq, cells were taken forward for nuclei isolation following 10X Genomics guidelines. Briefly, cells were centrifuged (300g for 5 min), added the lysis buffer (Tris-HCl (pH 7.4) 10 mM; NaCl 10 mM; MgCl<sub>2</sub> 3 mM; Tween-20 0.1%; NP-40 0.1%; Digitonin 0.01%; BSA 1%) and incubated on ice for 3 min (time optimized for thymus). Following the incubation, cells were washed (Tris-HCl (pH 7.4) 10 mM; NaCl 10 mM; MgCl<sub>2</sub> 3 mM; BSA 1%; Tween-20 0.1%) and centrifuged (300g for 5 min) and nuclei were resuspended in Diluted Nuclei Buffer (10X Genomics). Isolated nuclei were high-quality with well-resolved edges and no evidence of blebbing. The final nuclei concentration was determined prior to loading using a hemocytometer.

#### **Single-cell RNA and ATAC sequencing of human thymus**

scRNA-seq targeting 5000 cells per sample was performed using the Chromium Controller (10X Genomics). Single-cell cDNA synthesis, amplification, and sequencing libraries were generated using the Single Cell 5' Reagent Kit following the manufacturer's instructions. The libraries from up to eight loaded channels were multiplexed together and sequenced on an Illumina HiSeq 4000.

scATAC-seq targeting 5000 cells was performed using Chromium Single Cell ATAC Library and Gel Bead kit (10X Genomics). The libraries from up to eight loaded channels were multiplexed together and sequenced on an Illumina HiSeq 4000.

#### **Computational processing and analysis of the human fetal thymus single-cell genomics data**

scRNA-seq data were aligned and quantified using the Cell Ranger Single-Cell Software Suite (version 2.0, 10X Genomics) against the GRCh38 human reference genome provided by Cell Ranger. The scRNA-seq data was preprocessed using Seurat. Cells with fewer than 500 detected genes and more than 10% mitochondrial gene expression content were removed. Ribosomal genes, cell cycle genes [39], and genes associated with dissociation-induced effects [43] were removed. Clusters were identified using a community identification algorithm as implemented in the Seurat "FindClusters" function and annotated using canonical cell-type markers from [39].

The scATAC-seq data was aligned and preprocessed using CellRanger (10X Genomics). SnapATAC [44] was used for quality control, preprocessing, and generating cell-by-bin and log-normalized gene activity matrices. The binarized cell-by-bin matrix was used as input for term frequency-inverse document frequency (TF-IDF) weighting, using term frequency and smoothed inverse document frequency as the weighting scheme. Singular-value decomposition (SVD) was used for dimensionality reduction. Clustering and UMAP visualization were performed using Seurat. chromVar [45] was used to discover transcription factor dynamics and variation in their motif accessibility.

The 50 dimension reduced scATAC-seq and the 50 dimension reduced scRNA-seq data were supplied as input to MultiMAP. A shared feature space with both the scATAC-seq and scRNA-seq was constructed by removing genes from each dataset that were not present in the other, and then reducing the space to 50 dimensions using PCA. This shared space was supplied as input to MultiMAP, allowing the calculation of distances between cells from different datasets. The parameters of MultiMAP were all set to their default values, including the weight parameter for the scRNA-seq set to 0.8 and for ATAC-seq set to 0.2, on account of the higher-quality scRNA-seq.

The Leiden algorithm [46] was applied directly to the MultiGraph to jointly cluster all cells. The clusters were then annotated using canonical cell-type markers from [39]. Diffusion pseudotime (DPT) [47] was used for trajectory inference. The MultiGraph was supplied as input to the DPT function in SCANPY [48]. DPT was performed only on cells annotated as T cells. Cells were removed if they were positioned away from T cell clusters and close to fibroblasts and erythrocytes on the MultiMAP plot, as this likely indicated that they were incorrectly annotated. tradeSeq [49] was used to identify genes whose expression changes significantly along the trajectory.

#### **Acquisition and processing of human PBMCs**

PBMCs from two donors were acquired from a LeukoLab (Clinical division of AllCells). Frozen PBMC samples were thawed quickly at 37 °C in a water bath. Two pools made for technical duplicates with ~500,000 cells for each donor per pool (50/50). Nuclei isolation, transposition, ATAC-seq, and Gene Expression (GEX) sequencing libraries construction performed according to the manufacturer's demonstrated protocol (CG000365 Rev A; 10X Genomics) and Next GEM Single Cell Multiome ATAC and Gene Expression user guide (CG000338 Rev A; 10X Genomics). One lane per pool with a 3000 targeted nuclei recovery was loaded on a Chromium Next GEM Chip J. ATAC-seq and GEX indexed libraries were sequenced on a NovaSeq 6000 SP Flowcell according to the 10X Genomics recommendations, aiming for a minimum of 50,000 PE reads per cell for both types (ATAC-Seq and GEX) libraries.

#### **Computational processing and analysis of the human PBMCs Multiome ATAC+RNA data**

snRNA-seq and snATAC-seq data were aligned and quantified using the Cell Ranger ARC suite (10X Genomics) against the GRCh38 human reference genome provided by Cell Ranger. The snRNA-seq data was preprocessed using Seurat. Cells with fewer than 500 detected genes and more than 20% mitochondrial gene expression content were removed. Clusters were identified using a community identification algorithm as implemented in the Seurat "FindClusters" function and annotated using canonical cell-type markers.

SnapATAC [44] was used for quality control, preprocessing, and generating cell-by-bin and log-normalized gene activity matrices for the snATAC-seq data. The binarized cell-by-bin matrix was used as input for term frequency-inverse document frequency (TF-IDF) weighting, using term frequency and smoothed inverse document frequency as the weighting scheme. SVD was used for dimensionality reduction. Clustering and UMAP visualization were performed using Seurat.

The 50 dimension reduced snATAC-seq and the 50 dimension reduced snRNA-seq data were supplied as input to MultiMAP. A shared feature space with both the snATAC-seq and snRNA-seq was constructed by removing genes from each dataset that were not present in the other, and then reducing the space to 50 dimensions using PCA. This shared space was supplied as input to MultiMAP, allowing the calculation of distances between cells from different datasets. The parameters of MultiMAP were all set to their default values, including the weight parameter for the snRNA-seq set to 0.8 and for snATAC-seq set to 0.2, on account of the higher-quality snRNA-seq.

### Single-cell RNA sequencing of mouse spleen and data processing

The spleen from a 6-month-old C57BL/6Jax mouse was removed. The splenocytes were isolated by passing the spleen through a 70- $\mu$ m cell strainer (Fisher Scientific 10788201) into 30-ml ice-cold 1 $\times$  DPBS (Thermo Fisher 14190169) with 2 mM EDTA and 0.5% (w/v) BSA (Sigma A9418) using the plunger of a 2-ml syringe. Cells were spun down at 500g for 7 min at 4°. Then the supernatant was removed, and the cell pellet resuspended in 5 ml 1 $\times$  RBC lysis buffer (Thermo Fisher 00-4300-54). The cell suspension was vigorously vortexed for 5 s and left on the bench for 5 min to lyse the red blood cells. Then, 45 ml ice-cold 1 $\times$  DPBS was added, and cells were spun down at 500g for 7 min at 4°. The supernatant was removed, and 30 ml ice-cold 1 $\times$  DPBS with 0.1% BSA was used to resuspend the cell pellet. The cell suspension was passed through a Miltenyi 30  $\mu$ m Pre-Separation Filter (Miltenyi 130-041-407), and the cell number was determined using the C-chip counting chamber (VWR DHC-N01). The cells were spun down again, and the cell pellet resuspended in ice-cold 1 $\times$  DPBS with 0.1% BSA to reach a concentration of 1,000,000 cells per ml. The splenocytes were then loaded on the 10x Chromium Controller, aiming to recover ~ 5000 cells (Targeted Cell Recovery 5000 cells). cDNA and a sequencing library were made according to 10x Single Cell 3' Reagent Kits v2 manual. The library was sequenced on an Illumina HiSeq 4000 machine.

The resulting scRNA-seq data were preprocessed using Cell Ranger (10X Genomics) and downstream analysis were performed using the Seurat workflow. Cells with fewer than 200 detected genes and more than 10% mitochondrial gene expression content were filtered out. Downstream analyses such as normalization, clustering, and visualization were performed using Seurat. Clusters were identified using the community identification algorithm as implemented in the Seurat "FindClusters" function. Clusters were annotated using canonical cell-type markers from the original study [22]. Scrublet [50] was used for doublet detection.

### Acquisition and processing of previously published datasets

The mouse spleen scATAC-seq data was obtained from ArrayExpress (E-MTAB-6714) and preprocessed using the code provided by Chen et al. [22] ([https://github.com/dbrg77/plate\\_scATAC-seq](https://github.com/dbrg77/plate_scATAC-seq)). Briefly, reads from all cells were merged, and open chromatin regions were identified by peak calling with MACS2 [51]. Latent semantic indexing analysis was used for dimensionality reduction of the resulting cell-by-bin matrix. The binary cell-by-bin accessibility was used as input for TF-IDF weighting, using term frequency and smoothed inverse document frequency as the weighting scheme. SVD

was used for dimensionality reduction. SnapATAC [44] was used to generate gene activity count matrices, which were then log-normalized. The 50 dimension reduced accessibility of the scATAC-seq and the 50 dimension reduced gene expression of the scRNA-seq data were supplied as input to MultiMAP. A shared feature space with both the scATAC-seq and scRNA-seq was constructed by removing genes from each dataset that were not present in the other, and then reducing the space to 50 dimensions using PCA. This shared space was supplied as input to MultiMAP, allowing the calculation of distances between cells from different datasets. The parameters of MultiMAP were all set to their default values, including the weight parameter for the scRNA-seq set to 0.8 and for ATAC-seq set to 0.2 due to the higher-quality scRNA-seq. The Leiden algorithm was applied directly to the MultiGraph to jointly cluster all cells. Harmonic function-based node classification was performed directly on the MultiGraph to predict cell types of the scATAC-seq cells given the cell types of the scRNA-seq cells [52].

Human hematopoiesis scRNA-seq and scATAC-seq data were downloaded from <https://github.com/GreenleafLab/MPAL-Single-Cell-2019>. The scRNA-seq consists of 6 experimental batches, and the scATAC-seq consists of 10 experimental batches. Severe batch effects were observed, so this data was considered to consist of 16 separate datasets for the integration with MultiMAP. The scRNA-seq data was preprocessed using Seurat, and each batch was log-normalized and reduced to 50 dimensions with PCA. The cell-by-bin peak accessibility was used as provided by the authors. The binary cell-by-bin accessibility was used as input for TF-IDF weighting, using term frequency and smoothed inverse document frequency as the weighting scheme. Separately for each batch, the weighted data were reduced to 50 dimensions using SVD. Gene activities of the ATAC data were calculated using Cicero [53] and log-normalized. To integrate all of the data at once, all 16 datasets were provided as input to MultiMAP in the form of the 50 dimension reduced accessibility data of the scATAC-seq and the 50 dimension reduced gene expression of the scRNA-seq. Shared feature spaces containing two datasets were constructed by removing genes from each of the datasets that were not present in the other, and then reducing the space to 50 dimensions using PCA. These shared spaces were supplied as input to MultiMAP to calculate distances between cells from different datasets. The parameters of MultiMAP were all set to their default values, including the weight parameter for the scRNA-seq set to 0.8 and for ATAC-seq set to 0.2 due to the higher-quality scRNA-seq data.

scRNA-seq data of the mouse frontal cortex acquired with Drop-seq was obtained from [dropviz.org](https://dropviz.org). STARmap data of the mouse visual cortex was downloaded from <https://www.starmapresources.com/data/>. Each dataset was separately preprocessed with Seurat [10], log-normalized, and reduced to 50 dimensions with PCA. Both 50 dimensional reduced datasets were supplied as input to MultiMAP. A shared feature space with both the STARmap and scRNA-seq data was constructed by removing genes from each dataset that were not present in the other, and then reducing the space to 50 dimensions using PCA. This shared space was supplied as input to MultiMAP to calculate distances between cells from different datasets. The parameters of MultiMAP were all set to their default values, including the weight parameter for the scRNA-seq set to 0.8 and for Drop-seq set to 0.2, on account of higher-quality, tighter clusters generally observed in the scRNA-seq.

scRNA-seq, scATAC-seq, and snmC-seq data from the mouse primary cortex [26] was downloaded from the Neuroscience Multi-omics Archive (NeMO). The scRNA-seq was preprocessed using Seurat, log-normalized, and reduced to 50 dimensions with PCA. The binary cell-by-bin accessibility and gene activity count matrix of the scATAC-seq were obtained with SnapATAC [44]. The gene activity count data was log-normalized. Latent semantic indexing analysis was used for dimensionality reduction of the scATAC-seq accessibility. The binary cell-by-bin accessibility was used as input for TF-IDF weighting, using term frequency and smoothed inverse document frequency as weighting scheme. Weighted data were reduced to 50 dimensions using SVD. The DNA methylation data was preprocessed as described in [54], using the provided scripts. Briefly, after mapping, the methyl-cytosine counts and total cytosine counts were calculated in two sets of genome regions for each cell: the non-overlapping 100-kb bins tiling the mm10 genome, which was used for dimensionality reduction, and gene body regions  $\pm 2$  kb, which is used for the joint alignment. Posterior mCH and mCG rates were calculated based on beta-binomial distribution for the non-overlapping 100-kb bins matrix. The top 3000 highly variable features were taken and the data was reduced to 50 dimensions with PCA. Because gene body mCH proportions are negatively correlated with gene expression level, the direction of the methylation data was reversed by subtracting all values from the maximum methylation value [11]. The 50 dimensional reduced scRNA-seq, scATAC-seq, and snmC-seq were supplied as input to MultiMAP. Shared feature spaces containing each pair of two datasets and all three datasets together were constructed by removing genes from each of the datasets that were not present in the other, and then reducing the space to 50 dimensions using PCA. These shared spaces were supplied as input to MultiMAP, allowing the calculation of distances between cells from different datasets. The parameters of MultiMAP were all set to their default values. The weight parameter for the scRNA-seq set to 0.8 and for the other omics set to 0.2, on account of the higher-quality scRNA-seq data.

### Benchmarking

Benchmarking of MultiMAP, Seurat v3, LIGER, iNMF, Conos, and GLUER was performed using a variety of multi-omic data including the scRNA-seq and scATAC-seq data of the spleen, scRNA-seq, and STARmap of the visual cortex, and the scRNA-seq, scATAC-seq, and snmC-seq of the primary cortex. These datasets were chosen because they all have cell type annotations supplied in their original publications, which was used to independently validate the integration.

The scRNA-seq and STARmap data was log-normalized using Seurat and then used as an input for all integration methods, except GLUER where the raw data was used as an input and preprocessed using the SCANPY workflow. The scATAC-seq data was preprocessed as described above and the log-normalized gene activity matrix was used as an input for all integration methods. Seurat, LIGER, iNMF, Conos, and GLUER were executed as detailed in their tutorials, with all parameters set to their default values. Latent Semantic Indexing was used as the dimensionality reduction technique for the scATAC-seq data for weighting anchors in Seurat 3. CCA was used as the

dimensionality reduction technique for the scRNA-seq and STARmap data for weighting anchors in Seurat.

A diversity of performance metrics was used. After integration, label transfer of the cell type annotations from the scRNA-seq to each other omic was performed by setting the cell type of a query cell to the most frequent type among its 5 nearest labeled neighbors. The balanced accuracy of the label transfer (“Transfer”) was calculated using the annotations from the original publications as the ground truth. A high accuracy indicates that the same cell types from different modalities are near each other in the integrated embedding. After integration, the average Silhouette score [55] (“Silhouette”) across all cells was calculated using the cell type annotations from the original publications as the cluster labels. We note that the Silhouette score is not affected by the number of clusters as we use the same cell type labels, and hence number of clusters, for each integration method. A higher Silhouette score indicates the embedding is better separating distinct cell types. The degree of alignment (“Alignment”) of the different datasets in the integrated embedding was calculated as the proportion of each cell’s 5 nearest neighbors that originated in a different dataset, averaged over all cells. This metric was also used in [11]. A higher value of the alignment score indicates that the different datasets are more evenly mixed in the integrated embedding. The degree to which the embedding preserves the high-dimensional structure (“Structure”) of each dataset was calculated as the Pearson correlation between all pairwise distances in the high-dimensional spaces and the corresponding distances in the embedding. A higher correlation indicates that the embedding is more faithful to the high-dimensional structure. All of these performance metrics were also calculated in the shared feature space of the datasets to be integrated, to get baseline values of the metrics prior to the application of any integration strategy.

The wall-clock runtime of each method on each dataset was recorded. Additionally, to characterize the runtimes of the methods on a wide range of dataset sizes, the integration methods were run on datasets ranging from 1000 to 500,000 cells. To produce these datasets, we subsampled the mouse primary cortex scRNA-seq and scATAC-seq data [26] using geometric sketching [31]. The datasets were subsampled so that there were an equal number of cells in each of the scRNA-seq and scATAC-seq datasets, up to 100,000 cells. Since the scATAC-seq data had 81,196 cells in total, for the 500,000-cell comparison, we used an scRNA-seq of 418,804 cells. All methods were run with 3.1 GHz Intel i7 cores and 218 GB RAM.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-021-02565-y>.

**Additional file 1: Figures S1-S9**

**Additional file 2.** Supplementary Methods

**Additional file 3.** Review history

## Acknowledgements

We thank Jana Eliasova for the graphical illustrations. We are grateful to Emma Dann, Natsuhiko Kumasaka, and Zhihan Xu for critical feedback on the manuscript. We thank Ruben Chazarra-Gil and Vladimir Yu Kiselev for the comparison of different batch correction methods on the pancreas dataset. M.S.J. was supported by a Gates Cambridge Scholarship. J.-E.P. was supported by EMBO Long-Term and Advanced Fellowships. M.E. is funded by a Barts Charity Lectureship (MGU045). S.A.T. is funded by Wellcome (WT206194). The study was supported by Wellcome Human Cell Atlas Strategic Science Support (WT211276/Z/18/Z) and the Chan Zuckerberg Initiative (CZF2019-002445).

**Review history**

The review history is available as Additional file 3.

**Peer review information**

Barbara Cheifet was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

**Authors' contributions**

M.S.J., M.E., and S.A.T. conceived the study. M.S.J. conceived and developed MultiMAP. M.S.J. created the codebase with contributions from K.P. and M.E. C.D.C., J-E.P., R.A.B, E.S, L.M., A.E., and X.C. generated the single-cell data. M.E. analyzed the data and interpreted the results with contributions from M.S.J. and S.A.T. M.S.J., M.E., and S.A.T. wrote the manuscript with contributions from A.L., X.C., and C.D.C. All authors read and accepted the manuscript.

**Availability of data and materials**

scRNA-seq and scATAC-seq data generated for this publication were deposited in ArrayExpress: E-MTAB-9769 [56] for scRNA-seq of mouse splenocytes, E-MTAB-9840 [57] and E-MTAB-9828 [58] for scRNA-seq and scATAC-seq of the thymus, ArrayExpress. E-MTAB-11225 and E-MTAB-11226 [59] for the Multiome RNA + ATAC PBMC data. MultiMAP is publicly available at [github.com/Teichlab/MultiMAP](https://github.com/Teichlab/MultiMAP) [60] and zenodo [61].

**Declarations****Ethics approval and consent to participate**

The thymus tissue sample used for this study was obtained with written informed consent from the participant in accordance with the guidelines in The Declaration of Helsinki 2000. The human fetal tissue was obtained from the MRC/Wellcome Trust-funded Human Developmental Biology Resource (HDBR, <http://www.hdbbr.org>) with appropriate maternal written consent and approval from the Newcastle and North Tyneside NHS Health Authority Joint Ethics Committee (08/H0906/21 + 5). HDBR is regulated by the UK Human Tissue Authority (HTA; [www.hta.gov.uk](http://www.hta.gov.uk)) and operates in accordance with the relevant HTA Codes of Practice.

The mice for the splenocyte data were maintained under specific pathogen-free conditions at the Wellcome Trust Genome Campus Research Support Facility (Cambridge, UK). These animal facilities are approved by and registered with the UK Home Office. All procedures were in accordance with the Animals (Scientific Procedures) Act 1986. The protocols were approved by the Animal Welfare and Ethical Review Body of the Wellcome Trust Genome Campus.

**Competing interests**

S.A.T. has received remunerations for consulting and Scientific Advisory Board work from Genentech, Biogen, Roche, and GlaxoSmithKline as well as Foresite Labs over the past 3 years.

**Author details**

<sup>1</sup>Theory of Condensed Matter, Dept Physics, Cavendish Laboratory, University of Cambridge, JJ Thomson Ave, Cambridge CB3 0HE, UK. <sup>2</sup>Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK. <sup>3</sup>Southern University of Science and Technology, 1088 Xueyuan Ave, Nanshan, Shenzhen 518055, Guangdong Province, China. <sup>4</sup>KAIST, 291 Daehak-ro, Eoeun-dong, Yuseong-gu, Daejeon, South Korea. <sup>5</sup>Biosciences Institute, Newcastle University, Newcastle upon Tyne NE2 4HH, UK. <sup>6</sup>Barts Cancer Institute, Queen Mary University of London, London, UK.

Received: 26 August 2021 Accepted: 3 December 2021

Published online: 20 December 2021

**References**

1. Stoeckius M, Hafemeister C. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods*. 2017; 14(9):865–8. <https://doi.org/10.1038/nmeth.4380>.
2. Peterson VM, Zhang KX. Multiplexed quantification of proteins and transcripts in single cells. *Nat Biotechnol*. 2017; 35(10):936–9. <https://doi.org/10.1038/nbt.3973>.
3. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet*. 2019;20(4): 207–20. <https://doi.org/10.1038/s41576-018-0089-8>.
4. Karamaker ID, Vermeulen M. Single-cell DNA methylation profiling: technologies and biological applications. *Trends Biotechnol*. 2018;36(9):952–65. <https://doi.org/10.1016/j.tibtech.2018.04.002>.
5. Mayr U, Serra D, Liberali P. Exploring single cells in space and time during tissue development, homeostasis and regeneration. *Development*. 2019;146:12. <https://doi.org/10.1242/dev.176727>.
6. Regev A, Teichmann SA. The Human Cell Atlas. *Elife*. 2017;6. <https://doi.org/10.7554/eLife.27041>.
7. HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature*. 2019; 574(7777):187–92. <https://doi.org/10.1038/s41586-019-1629-x>.
8. Efremova M, Teichmann SA. Computational methods for single-cell omics across modalities. *Nat Methods*. 2020;17(1): 14–7. <https://doi.org/10.1038/s41592-019-0692-4>.
9. Lähnemann D, Köster J. Eleven grand challenges in single-cell data science. *Genome Biol*. 2020;21(1):31. <https://doi.org/10.1186/s13059-020-1926-6>.
10. Stuart T, et al. Comprehensive integration of single-cell data. *Cell*. 2019;177:1888–1902.e21.
11. Welch JD, et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*. 2019;177: 1873–1887.e17.
12. Gao C, Liu J. Iterative single-cell multi-omic integration using online learning. *Nat Biotechnol*. 2021;39(8):1000–7. <https://doi.org/10.1038/s41587-021-00867-x>.

13. Lopez R, et al. A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. arXiv [cs.LG]. 2019.
14. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. *J Open Source Software*. 2018;3(29):861. <https://doi.org/10.21105/joss.00861>.
15. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using umap. *Nat Biotechnol*. 2019;37(1):38–44. <https://doi.org/10.1038/nbt.4314>.
16. Spivak ID. Metric realization of fuzzy simplicial sets. Preprint; 2009.
17. Barr M. Fuzzy set theory and topos theory. *Can Math Bull*. 1986;29(4):501–8. <https://doi.org/10.4153/CMB-1986-079-9>.
18. Shang X-G, Jiang W-S. A note on fuzzy information measures. *Pattern Recogn Lett*. 1997;18:425–32. [https://doi.org/10.1016/S0167-8655\(97\)00028-7](https://doi.org/10.1016/S0167-8655(97)00028-7).
19. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* 3111–3119; 2013. <https://doi.org/10.5555/2999792.2999959>.
20. Tang J, Liu J, Zhang M, Mei Q. Visualizing large-scale and high-dimensional data. In: *Proceedings of the 25th International Conference on World Wide Web 287–297: International World Wide Web Conferences Steering Committee*; 2016. <https://doi.org/10.1145/2872427.2883041>.
21. Gradient-based learning applied to document recognition. *Intell Signal Process*. 2009. <https://doi.org/10.1109/9780470544976.ch9>.
22. Chen X, Miragaia RJ, Natarajan KN, Teichmann SA. A rapid and robust method for single cell chromatin accessibility profiling. *Nat Commun*. 2018;9(1):5345. <https://doi.org/10.1038/s41467-018-07771-0>.
23. Granja JM, Klemm S. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat Biotechnol*. 2019;37(12):1458–65. <https://doi.org/10.1038/s41587-019-0332-7>.
24. Saunders A, et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell*. 2018;174:1015–1030.e16.
25. Wang X, Allen WE. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*. 2018; 361(6400). <https://doi.org/10.1126/science.aat5691>.
26. Yao Z, et al. An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types. 2020.02.29.970558. 2020. <https://doi.org/10.1101/2020.02.29.970558>.
27. Brodmann K. *Brodmann's: Localisation in the Cerebral Cortex*; Springer; 2010. <https://doi.org/10.1007/b138298>.
28. Yamawaki N, Borges K, Suter BA, Harris KD, Shepherd GMG. A genuine layer 4 in motor cortex with prototypical synaptic circuit connectivity. *Elife*. 2014;3:e05422. <https://doi.org/10.7554/eLife.05422>.
29. Barkas N, Petukhov V. Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat Methods*. 2019;16(8): 695–8. <https://doi.org/10.1038/s41592-019-0466-z>.
30. Peng T, Chen GM, Tan K. GLUER: integrative analysis of single-cell omics and imaging data by deep neural network. <https://doi.org/10.1101/2021.01.25.427845>.
31. Hie B, Cho H, DeMeo B, Bryson B, Berger B. Geometric sketching compactly summarizes the single-cell transcriptomic landscape. *Cell Syst*. 2019;8:483–493.e7.
32. Muraro MJ, et al. A single-cell transcriptome atlas of the human pancreas. *Cell Syst*. 2016;3:385–394.e3.
33. Segerstolpe Å, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab*. 2016;24:593–607.
34. Baron M, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst*. 2016;3:346–360.e4.
35. Chazarra-Gil R, van Dongen S, Kiselev VY, Hemberg M. Flexible comparison of batch correction methods for single-cell RNA-seq using BatchBench. *Nucleic Acids Res*. 2021. <https://doi.org/10.1093/nar/gkab004>.
36. Roels J, Kuchmiy A. Distinct and temporary-restricted epigenetic mechanisms regulate human  $\alpha\beta$  and  $\gamma\delta$  T cell development. *Nat Immunol*. 2020;21(10):1280–92. <https://doi.org/10.1038/s41590-020-0747-9>.
37. Jia G, Preussner J. Single cell RNA-seq and ATAC-seq analysis of cardiac progenitor cell transition states and lineage settlement. *Nat Commun*. 2018;9(1):4877. <https://doi.org/10.1038/s41467-018-07307-6>.
38. Chen H, Albergante L. Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nat Commun*. 2019;10(1):1903. <https://doi.org/10.1038/s41467-019-09670-4>.
39. Park J-E, Botting RA. A cell atlas of human thymic development defines T cell repertoire formation. *Science*. 2020; 367(6480). <https://doi.org/10.1126/science.aay3224>.
40. Hosokawa H, Rothenberg EV. How transcription factors drive choice of the T cell fate. *Nat Rev Immunol*. 2020. <https://doi.org/10.1038/s41577-020-00426-6>.
41. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE*. 1998; 86(11):2278–324. <https://doi.org/10.1109/5.726791>.
42. Hern WM. Correlation of fetal age and measurements between 10 and 26 weeks of gestation. *Obstet Gynecol*. 1984; 63(1):26–32.
43. van den Brink SC, Sage F. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat Methods*. 2017;14(10):935–6. <https://doi.org/10.1038/nmeth.4437>.
44. Fang R, et al. Fast and accurate clustering of single cell epigenomes reveals Cis-regulatory elements in rare cell types. <https://doi.org/10.1101/615179>.
45. Schep AN, Wu B, Buenostro JD, Greenleaf WJ. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods*. 2017;14:975–8.
46. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mechanics*. 2008;2008:P10008.
47. Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods*. 2016;13(10):845–8. <https://doi.org/10.1038/nmeth.3971>.
48. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. 2018;19(1):15. <https://doi.org/10.1186/s13059-017-1382-0>.

49. Van den Berge K, et al. Trajectory-based differential expression analysis for single-cell sequencing data. *Nat Commun.* 2020;11(1):1201. <https://doi.org/10.1038/s41467-020-14766-3>.
50. Wolock SL, Lopez R, Klein AM. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* 2019;8:281–291.e9.
51. Grytten I, Rand KD. Graph Peak Caller: calling ChIP-seq peaks on graph-based reference genomes. *PLoS Comput Biol.* 2019;15(2):e1006731. <https://doi.org/10.1371/journal.pcbi.1006731>.
52. Zhu X, Ghahramani Z, Lafferty JD. Semi-supervised learning using Gaussian fields and harmonic functions. In: *Proceedings of the 20th International conference on Machine learning (ICML-03)*; 2003. p. 912–9.
53. Pliner HA, et al. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol Cell.* 2018;71:858–871.e8.
54. Kozareva V, et al. A transcriptomic atlas of the mouse cerebellum reveals regional specializations and novel cell types. <https://doi.org/10.1101/2020.03.04.976407>.
55. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math.* 1987;20:53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
56. Sarkin JM, Krzysztof P, Cecilia DC, Xi C, Jongeun P, Lira M, et al. scRNA-seq data of mouse splenocytes. *ArrayExpress.* E-MTAB-9769; 2021.
57. Sarkin JM, Krzysztof P, Cecilia DC, Xi C, Jongeun P, Lira M, et al. scRNA-seq data of human fetal thymus. *ArrayExpress.* E-MTAB-9840; 2021.
58. Sarkin JM, Krzysztof P, Cecilia DC, Xi C, Jongeun P, Lira M, et al. scATAC-seq data of human fetal thymus. *ArrayExpress.* E-MTAB-9828; 2021.
59. Sarkin JM, Krzysztof P, Cecilia DC, Xi C, Jongeun P, Lira M, et al. Multiome RNA + ATAC data of human PBMCs. *ArrayExpress.* E-MTAB-11225 and E-MTAB-11226; 2021.
60. Sarkin JM, Krzysztof P, Cecilia DC, Xi C, Jongeun P, Lira M, et al. Multiome RNA + ATAC data of human PBMCs. MultiMAP: dimensionality reduction and integration of multimodal data: Github. <https://github.com/Teichlab/MultiMAP>; 2021.
61. Sarkin JM, Krzysztof P, Cecilia DC, Xi C, Jongeun P, Lira M, et al. Multiome RNA + ATAC data of human PBMCs. MultiMAP: dimensionality reduction and integration of multimodal data: Zenodo; 2021. <https://doi.org/10.5281/zenodo.5747678>.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

