nature methods

Article

https://doi.org/10.1038/s41592-022-01651-8

Annotation of spatially resolved single-cell data with STELLAR

Received: 24 November 2021

Accepted: 14 September 2022

Published online: 24 October 2022

Check for updates

Maria Brbić[1,2,6], Kaidi Cao[1,6], John W. Hickey[3,6], Yuqi Tan[3], Michael P. Snyder[4], Garry P. Nolan[3,5] ✉ and Jure Leskovec[1] ✉

Accurate cell-type annotation from spatially resolved single cells is crucial to understand functional spatial biology that is the basis of tissue organization. However, current computational methods for annotating spatially resolved single-cell data are typically based on techniques established for dissociated single-cell technologies and thus do not take spatial organization into account. Here we present STELLAR, a geometric deep learning method for cell-type discovery and identification in spatially resolved single-cell datasets. STELLAR automatically assigns cells to cell types present in the annotated reference dataset and discovers novel cell types and cell states. STELLAR transfers annotations across different dissection regions, different tissues and different donors, and learns cell representations that capture higher-order tissue structures. We successfully applied STELLAR to CODEX multiplexed fluorescent microscopy data and multiplexed RNA imaging datasets. Within the Human BioMolecular Atlas Program, STELLAR has annotated 2.6 million spatially resolved single cells with dramatic time savings.

Development of spatial protein and RNA imaging technologies has opened new opportunities for understanding location-dependent properties of cells and molecules[1–4]. The power to capture spatial organization of cells within tissue plays an essential role in understanding cellular function and in studying complex intercellular mechanisms. To increase our knowledge of cells in healthy and diseased tissues, large tissue-mapping consortia efforts such as the Human BioMolecular Atlas Program (HuBMAP)[5], Human Tumor Atlas Network (HTAN)[6] and Human Cell Atlas (HCA)[7] have been generating comprehensive cell atlas datasets. These consortia efforts necessitate computational methods that can assist with robust characterizations of cells and guide our understanding of functional spatial biology. While computational tools for dissociated single-cell technologies have been used to characterize cells in spatial datasets[4,8–10], these tools ignore the spatial information crucial for more accurate annotation of spatial datasets.

Recently, methods that take into account spatial organization of cells have been developed for annotating spatial transcriptomics data[11–15]. However, these methods can not automatically assign cell-type labels to cells and thus require human reannotation once clusters are identified. While CELESTA tool has successfully avoided post hoc cluster reannotation for multiplexed in situ images[16], it is reliant on the previous human supervision provided in the form of the set of marker genes of expected cell types, which is often challenging to define for a new dataset. As new spatial datasets are being generated[17–21], there is a necessity for computational methods that simultaneously leverage molecular features and additional spatial context of cells while at the same time minimize manual human annotation effort.

Here, we present STELLAR (SpaTial cEll LeARning), a geometric deep learning tool for cell-type discovery and identification in spatially resolved single-cell datasets. Given annotated spatially resolved single-cell dataset with cells labeled according to their cell types (reference dataset), STELLAR learns spatial and molecular signatures that define cell types. Using the reference dataset, STELLAR then transfers the annotations to a completely unannotated spatially resolved

[1]Department of Computer Science, Stanford University, Stanford, CA, USA. [2]School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [3]Baxter Laboratories Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford University, Stanford, CA, USA. [4]Department of Genetics, Stanford University School of Medicine, Stanford University, Stanford, CA, USA. [5]Department of Pathology, Stanford University School of Medicine, Stanford University, Stanford, CA, USA. [6]These authors contributed equally: Maria Brbić, Kaidi Cao, John W. Hickey. ✉e-mail: gnolan@stanford.edu; jure@cs.stanford.edu

single-cell dataset with unknown cell types. The reference and unannotated datasets can belong to different dissection regions, different donors or different tissue types.

STELLAR has two unique properties. First, using graph convolutional neural networks[22,23], STELLAR learns latent low-dimensional cell representations that jointly capture spatial and molecular similarities of cells. In this way, cells that are spatially close together and that have similar levels of gene or protein expression are embedded close to each other. Indeed, we show that cell embeddings learned by STELLAR reveal higher-order tissue structures. Second, in a new completely unannotated dataset, STELLAR automatically assigns cells to cell types included in the reference set and also identifies cells with unique properties as belonging to a novel type that is not part of the reference set. Thus, STELLAR has an ability to assign cells to one of the cell types seen in the reference set, or discover a novel cell type for previously uncharacterized cell types. From a practical standpoint this ability is essential for leveraging reference sets generated in a disease-free state or in different tissues that may not have all the cell types represented in the target tissue but that may include most established cell types.

As a result of these unique properties, STELLAR solves major limitations with current annotation tools. STELLAR can be applied to both multiplexed protein and multiplexed RNA imaging datasets. In particular, we show the effectiveness of STELLAR on annotating CODEX multiplexed fluorescent microscopy datasets and multiplexed error-robust fluorescence in situ hybridization (MERFISH) spatial transcriptomics datasets. Encouraged by the results, we used STELLAR to annotate HuBMAP consortium data: so far STELLAR has annotated 2.6 million cells labeled with 54 protein markers from eight different regions of the human intestine data and across eight different donors[17]. On these datasets, STELLAR has saved hundreds of hours of manual work needed to assign cells to cell types.

Results

Overview of STELLAR

STELLAR takes as input (1) the reference dataset of annotated spatially resolved single-cell dataset with cells assigned to cell types, and (2) a completely unannotated spatially resolved single-cell dataset with cell types that are unknown (Fig. 1). STELLAR then assigns unannotated cells to cell types in the reference dataset, while for cells that do not fit into the reference cell types, it identifies individual novel cell types and assigns cells to them.

STELLAR learns low-dimensional cell representations (embeddings) that reflect molecular features of cells as well as their spatial organization and neighborhood. The cell embeddings are learned using graph convolutional neural networks (GCNs)[22,23]: neural networks that enable learning over arbitrary graph structures by encoding the structure of the local graph neighborhood into a dense vector embedding. In STELLAR, we construct a graph based on the spatial proximity of cells and use molecular features of cells as node features (Methods).

Given cell embeddings, STELLAR has two capabilities: (1) it assigns cells in the unannotated dataset to one of the cell types previously seen in the reference dataset, and (2), for cells that do not fit the existing known/labeled cell types STELLAR identifies novel cell types and assigns cells to them. STELLAR achieves this using an objective function that consists of two main components (Extended Data Fig. 1). In the first component, STELLAR gradually learns to separate cell types from the reference set by controlling intraclass variance of previously seen cell types using the adaptive margin mechanism[24]. In the second component, STELLAR discovers novel classes by generating auxiliary labels for the unannotated data that are used to guide the training. The auxiliary labels are generated based on the nearest neighbors of cells in the embedding space (Methods).

STELLAR identifies existing and discovers novel cell types

To demonstrate STELLAR's ability to assign cells to one of the cell types seen in the reference set and discover a novel cell type, we applied STELLAR to CODEX multiplexed imaging data[10]. We used data from human tonsil as the reference set (Fig. 2a and Extended Data Fig. 2) and tissue from a patient with Barrett's esophagus (BE) as the unannotated target tissue (Fig. 2b). Both imaging datasets underwent image processing and single-cell segmentation. Cell-type labels were manually curated and assigned by iterative unsupervised clustering of the single-cell data, analyzing expressions of protein markers and validating cluster accuracy with visualization in the tissue. Tonsils are often used for testing antibody panels as they contain high numbers of immune cell markers. The BE tissue also contains immune cells, but additionally contains differentiated epithelial cells that tonsils do not. Indeed, three subtypes of epithelial cells appear only in the BE dataset, while B cells appear only in tonsil data. Moreover, the two datasets have different distributions of cell types. For example, smooth muscle cells are major cluster in BE, but form the smallest group in the tonsil dataset (Extended Data Fig. 3).

Despite these differences, we found that STELLAR accurately assigned cell types to 93% of cells in the BE data, discovering also BE-specific subtypes of epithelial cells as a newly discovered cell type and correctly differentiating between glandular epithelial and secretory epithelial novel cell types (Fig. 2c). The novel cluster of paneth cells was correctly recognized as a novel class by STELLAR but assigned to glandular epithelial cells. The reason is that paneth cells are very rare in the BE dataset and form the smallest class of only 0.6% of the total cell number in the BE dataset. Mapping cell-type annotations back to spatial coordinates, STELLAR predictions also show agreement with ground-truth annotations and do not reflect issues recognizing cell types in certain areas of the tissue (Fig. 2d). The disagreement with ground-truth annotations mostly comes from mixing main cell types, in particular endothelial with stroma cells and endothelial with smooth muscle cells (Supplementary Fig. 1). These are all stromal cells and they likely share Vimentin, while lacking CD45 or Cytokeratin. They are also very often spatially located next to each other: endothelial cells will often be next to smooth muscle cells and vice versa. Furthermore, endothelial cells are elongated and the problem may be caused by current segmentation software tools that cosegment them with other cell-type markers causing STELLAR to assign them to other cell types.

One of the reasons we applied STELLAR to these two datasets was also to test the flexibility of the algorithm when cells may be found in different cellular environments. This is important for cross-tissue applications, but also in cases of applying STELLAR across disease states such as cancer where cells, normally restricted to one area, invade other areas of the tissue. Since we observed a high accuracy in cell-type label transfer, we asked whether labels that STELLAR generates result in unique cellular microenvironments or neighborhoods. This would mean that STELLAR is able to generalize even when the same cell types are present in a different microenvironment. To explore this, we used a method for cellular neighborhood analysis[25] with the ground-truth labels of the tonsil dataset and the STELLAR's predictions for the BE dataset. This method determines the composition of the nearest neighbors for each cell. Then these composition vectors for each cell are clustered into similarly composed cellular neighborhoods (Methods). This analysis revealed both shared and unique multicellular neighborhoods (Fig. 2e and Extended Data Fig. 4). For example, glandular epithelium and smooth muscle are exclusively found within BE tissue, while B cell zone, T cell zone and squamous epithelium/B cell neighborhoods are only found in the tonsil dataset. Furthermore, many neighborhoods shared between the datasets have little overlap in total percentage of cells indicating the structural distinctness of the two tissues. Thus, STELLAR can generalize to different spatial organizations while accurately transferring cell-type labels.

https://doi.org/10.1038/s41592-022-01651-8

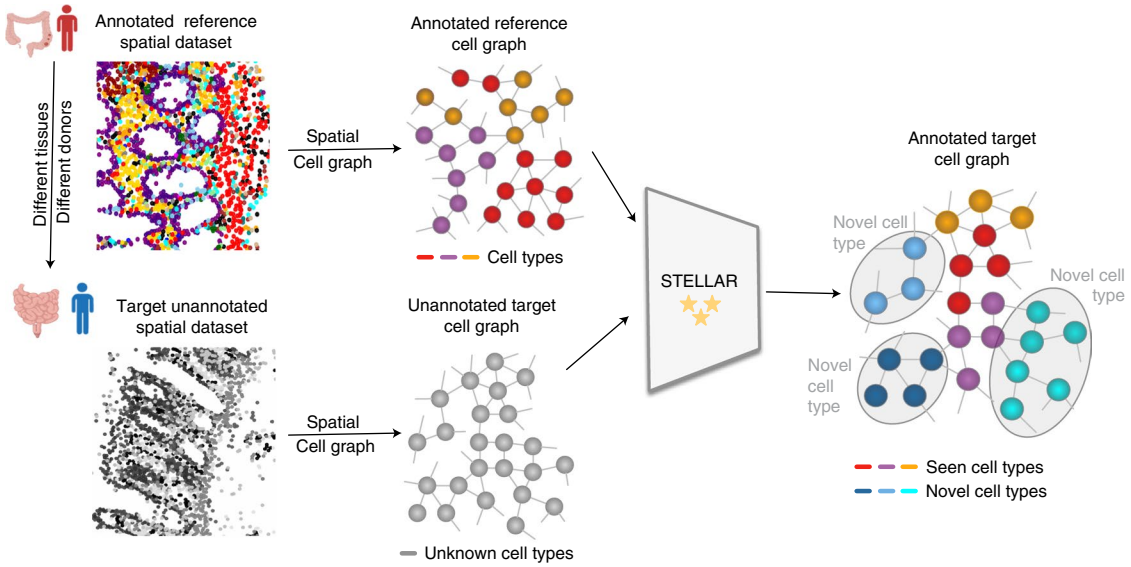


Fig. 1 | STELLAR is a geometric deep learning framework for annotating spatially resolved single-cell datasets. Given a reference spatially resolved single-cell dataset in which cells are annotated according to their cell types, STELLAR assigns cells in an unannotated spatial single-cell dataset to cell types included in the reference dataset or discovers novel cell types as a group of cells with unique properties not present in the reference dataset. STELLAR uses a graph convolutional encoder to learn low-dimensional cell embeddings that capture cell neighborhood as well as its molecular profile. Reference and unannotated datasets can originate from different sources as STELLAR transfers information across different tissues and different donors.

STELLAR outperforms other methods

On the BE dataset, we compared STELLAR to four other baselines including XGBoost[26], Support Vector Machine (SVM)[27], Random Forest[28], AdaBoost[29], as well as single-cell annotation baselines Seurat V4 (ref. [30]) and scNym[31]. We trained all baselines on the tonsil dataset, and evaluated them on the BE dataset. STELLAR substantially outperformed the best alternative method, achieving 61% higher accuracy (Fig. 2f and Supplementary Fig. 2). Single-cell baselines Seurat V4 and scNym achieved lower performance than other baselines so we additionally compared to another single-cell baseline scANVI[32]. However, the performance of scANVI was even lower, so we created a leaky setting for scANVI in which we used a fraction of BE annotations for model training. In such a scenario, the performance of scANVI substantially improved to 0.76 indicating that the performance drops are caused by differences between tonsil and BE datasets (Supplementary Methods and Extended Data Fig. 5a). Even when using BE annotations to annotate BE dataset, scANVI still did not outperform STELLAR. STELLAR was also the best performing method when evaluating performance using other evaluation metrics, such as F1-score, precision and recall (Extended Data Fig. 5b–d). To directly measure performance on cell types that are only present in the reference dataset, that is, shared between tonsil and BE datasets, we removed novel cell types. In this scenario, we still find that STELLAR outperformed other alternative methods by 6% (Supplementary Fig. 3).

We further systematically tested robustness of STELLAR to: (1) different data normalization strategies, (2) artificially introduced noisy annotations and (3) different number of withheld marker proteins. First, we find that STELLAR achieves best performance on *z* score normalized CODEX data, which agrees with previous studies comparing different combinations of normalization techniques and unsupervised clustering[33] (Extended Data Fig. 6a). STELLAR outperforms alternative methods even with different normalization strategies including when evaluated on unnormalized data, achieving 87% accuracy on unnormalized data. We additionally evaluated the effect of the noisy annotations on the performance. Specifically, we misannotated proportion of cells by randomly assigning wrong annotations to cells in the reference tonsil dataset. With 5% of wrong annotations, STELLAR achieves only 0.7% lower performance compared to performance in which all annotations are correct (Extended Data Fig. 6b). This means that STELLAR is robust even when a proportion of cells are misannotated. Finally, STELLAR is ran on the shared subsets of the genes and/or proteins between annotated reference dataset and unannotated target dataset. STELLAR is able to accurately transfer cell-type labels even when leaving out measured marker proteins (Extended Data Fig. 6c). While not as critical for transcriptomic datasets, this may be important for transferring cell-type labels in multiplexed imaging where there are limited numbers of antibody markers that can be chosen and antibody panels may not completely overlap.

Importance of capturing spatial organization of cells

Using graph convolutional neural networks, STELLAR leverages both spatial organization of cells as well as their molecular expressions. To directly measure the benefits of including spatial information, we designed a baseline in which we used STELLAR's objective function, but we replaced graph convolutional encoder with a fully connected neural (FCN) network layer disabling the usage of spatial information. When comparing FCN baseline to STELLAR, we find that capturing spatial organizations of the cells brings 11% improvement in cell-type annotation task compared to relying solely on the molecular information (Fig. 2g).

We next sought to analyze the optimal graph structure for the task. In STELLAR, we construct the graph by considering as neighbors all the cells whose distance is less than the preset threshold, which we initially set to 50 µm. We systematically changed the value of the threshold and analyzed the effect on the STELLAR performance on the CODEX tonsil/BE dataset. We found that the optimal distance is 30–70 µm, corresponding to approximately 5–30 average number of neighbors per cell. This optimum may indicate the single-cell local environment size that is most conserved, rather than broader structures within the tissue. Furthermore, even with a very low threshold of 10 µm where each node has only 0.02 neighbors per cell resulting in a graph with 98% of isolated nodes, or with a very high threshold of 90 µm where each cell has on average 50 neighbors, STELLAR performance does not degrade compared to using only molecular information. Overall, this analysis indicates that STELLAR is robust to input graph structure and that substantial performance gains can be expected by a meaningfully constructed spatial graph.

https://doi.org/10.1038/s41592-022-01651-8

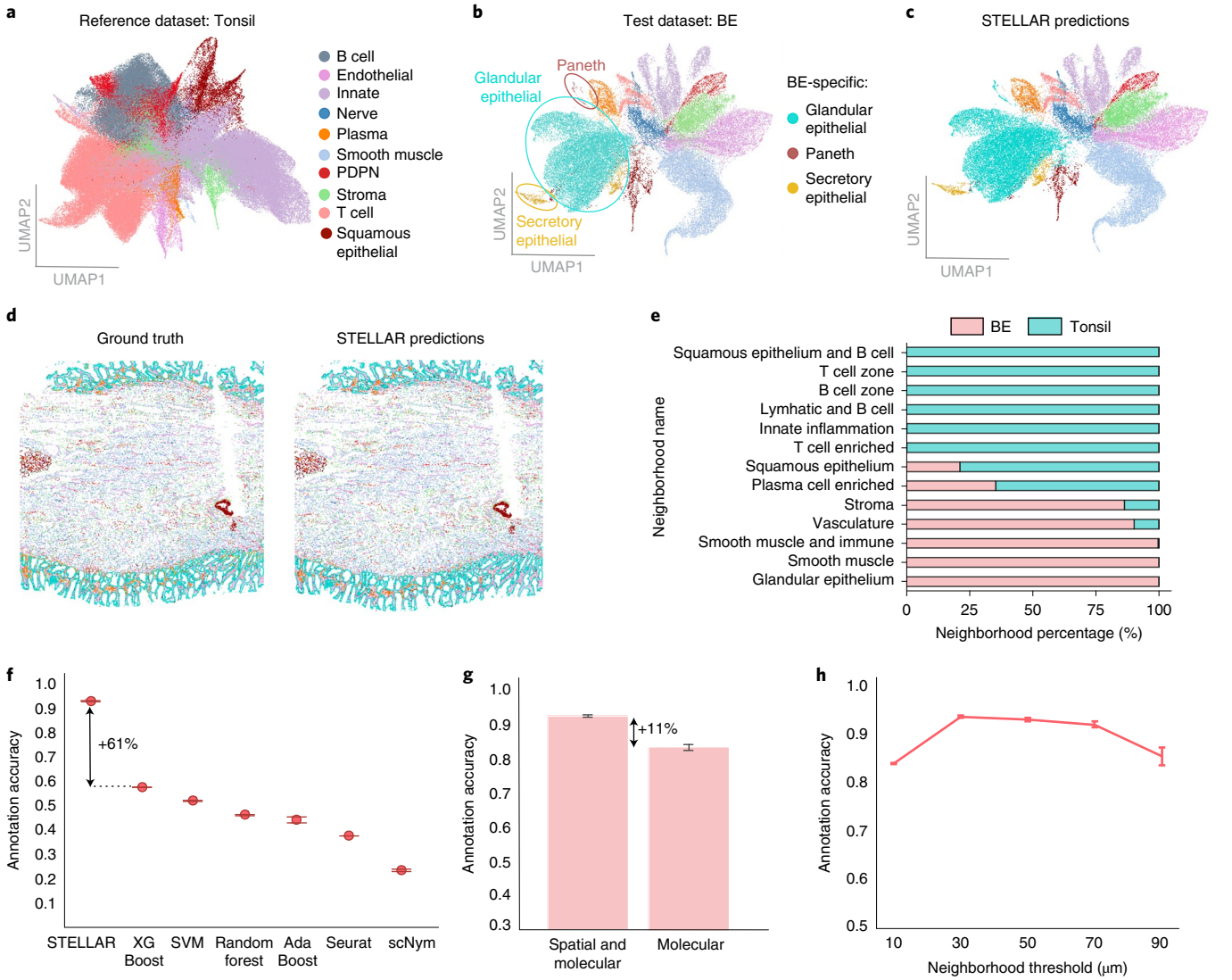


Fig. 2 | STELLAR accurately identifies cell types from the reference set and discovers novel cell types that have never been characterized in the reference set. a, Uniform manifold approximation and projection (UMAP) visualization of the healthy tonsil data used as the reference dataset. Colors denote ground-truth cell types annotations used to train STELLAR. PDPN stands for Podoplanin (PDPN) positive stromal cells. **b**,**c**, UMAP visualization of the BE data used as the test dataset. Three subtypes of epithelial cells are not found in the tonsil reference data. Colors denote ground-truth cell-type annotations (**b**) and STELLAR predictions (**c**). **d**, CODEX image of BE in spatial coordinates colored according to ground-truth annotations (left) and STELLAR predictions (right). **e**, Different neighborhood composition between reference dataset from tonsil tissue and target unnannotated dataset from BE. Neighborhoods are determined by cell composition enrichment in nearest neighbor vectors for each cell (Methods). **f**, Accuracy of cell-type assignments in the BE dataset by STELLAR compared to alternative approaches. Position of scatter plot points is computed as a mean accuracy score across five runs of each method. Error bars are from the standard deviation (s.d.). **g**, Accuracy of cell-type assignments in BE dataset by STELLAR and when modifying STELLAR to rely only on the molecular information in cell-type assignment. Height of each bar is computed as a mean accuracy score across five runs of each method. Error bars are from s.d. **h**, The effect of neighborhood threshold on the STELLAR's performance. All cells with distance less than the threshold are considered as neighbors. Position of scatter plot points is computed as a mean accuracy score across five runs of each method. Error bars are from s.d.

STELLAR is applicable to spatial transcriptomics data

While existing methods are focused on either spatial proteomics or transcriptomics data, in addition to CODEX datasets we evaluated STELLAR on a single-cell transcriptome-imaging dataset generated using MERFISH. In particular, we applied STELLAR to a large-scale spatially resolved cell atlas of the mouse primary motor cortex[8] consisting of 23 granular cell types from two mice. We used a dataset from one mouse as the reference annotated dataset and evaluated performance on the dataset from another mouse. We then systematically removed a number of cell types from the reference set and evaluated how STELLAR's performance was affected when gradually increasing the number of novel cell types. We measured accuracy separately on classes seen in the reference set and classes withheld from the reference set (novel classes). We find that STELLAR correctly assigned cell types to one of the classes in the reference set achieving 93% accuracy independently of the number of withheld cell types (Extended Data Fig. 7).

STELLAR successfully annotates HuBMAP data

A critical current bottleneck in analysis of the spatially resolved single cells for consortia efforts is an accurate assignment of granular cell-type labels across donors of the same tissue. For example, a typical CODEX dataset with a 48-marker panel used to analyze four tissues of the intestine requires at least 25 hours of work to cluster, merge, recluster, subcluster and assign cell types based on average

https://doi.org/10.1038/s41592-022-01651-8

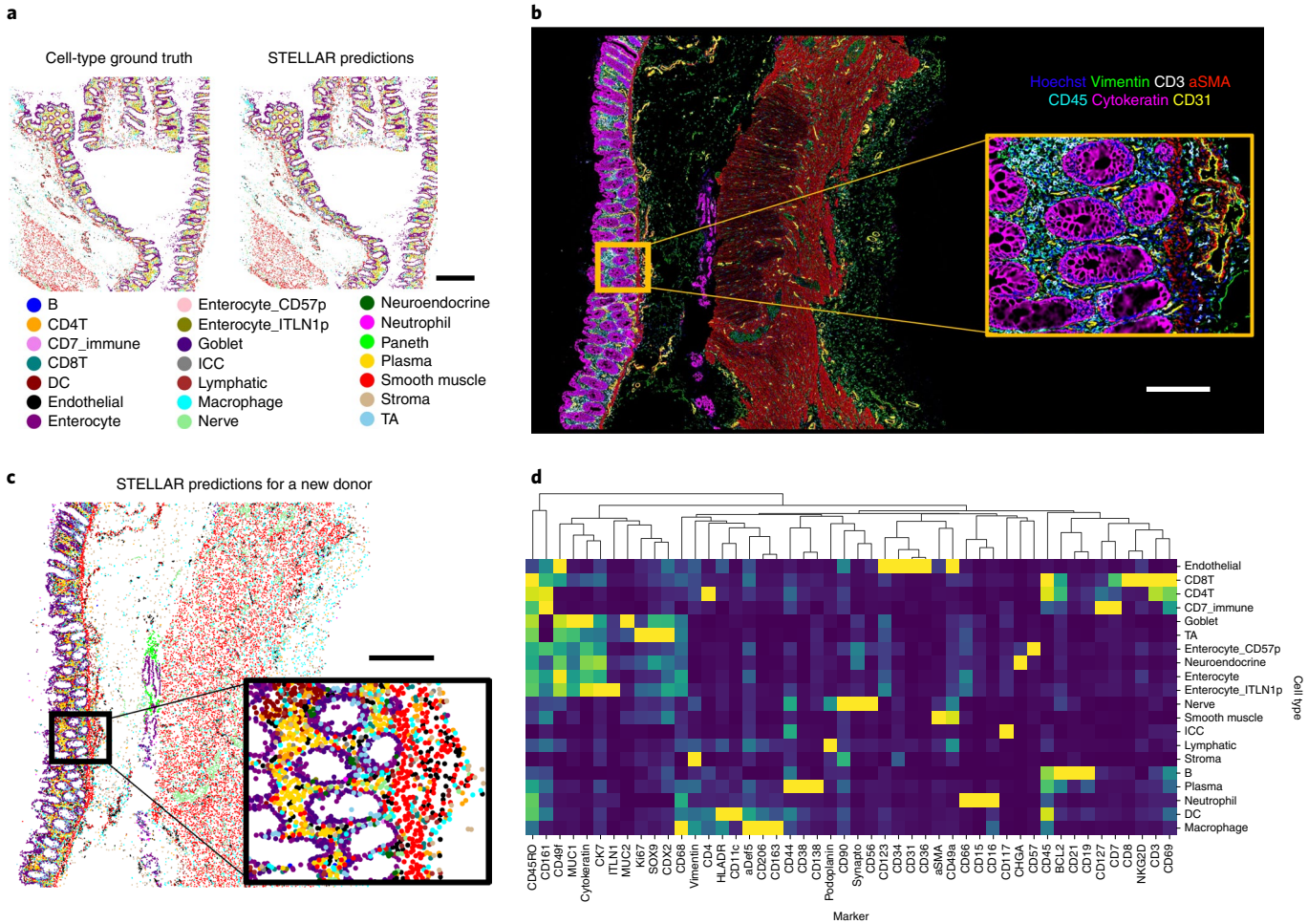


Fig. 3 | STELLAR transfers granular cell-type labels across tissue regions and donors from HuBMAP data and identifies main structures of healthy human intestine tissue. a, Cell-type maps with ground-truth and STELLAR predictions made after training on CODEX multiplexed imaging of three different intestine regions across the colon. This was tested five times across each of the three tissue sections with all four tissue regions being used as training datasets. Scale bar, 100 µm. **b**–**d**, We used data from one donor as the reference set, and applied STELLAR to annotate new, completely unannotated donor datasets. This is a representative image from a total of 16 tissues that we transferred cell-type labels to. **b**, Fluorescent images of 7 of 54 markers used in the CODEX multiplexed imaging of healthy intestine from a different donor. Hoechst, blue; Vimentin, green; CD3, white; aSMA, red; CD45, cyan; Cytokeratin, magenta and CD31, yellow. Scale bar of the zoomed-out image is 100 µm. **c**, STELLAR predictions of cell types after training on data from a different donor mapped to CODEX spatial coordinates. Scale bar, 100 µm. **d**, Average marker expression for cell types predicted by STELLAR for intestinal samples. DC, dendritic cell; ICC, interstitial cells of Cajal; TA, transit amplifying cell.

marker expressions and locations within CODEX images. As a part of the HuBMAP consortium, we have generated CODEX imaging data for tissues from eight donors for a total of 64 tissues from a healthy intestine. Combined with single-cell transcriptomic and epigenomic data, this molecularly detailed cell atlas contributes spatial organization of cell types as a reference map for understanding human intestinal biology[17].

We first evaluated the performance of STELLAR on this more granular data and used expert-annotated cell-type labels of images of different regions of a healthy colon generated from a single donor. We evaluated STELLAR in the leave-one-region-out setting where we trained on three regions and predicted annotations on the fourth region. STELLAR had high accuracy in cell-type label transfers (Fig. 3a) and had substantially higher performance (*F* score of 0.8) than tested unsupervised clustering methods without manual intervention (*F* score of 0.3 across different methods)[33]. This analysis showed STELLAR's ability to recognize fine-grained level cell types and encouraged us to use STELLAR to annotate 2.6 millions of cells generated across eight different donors.

STELLAR transfers information across donors

Encouraged by cross-region transfer results, we next used our expert-annotated samples from a single donor as training data and applied STELLAR to unannotated samples from two other donors. These datasets vary from each other based on tissue harvesting time, staining and imaging handler, and segmentation algorithms applied, which may alter marker distribution. Despite the differences, we find that the fluorescent CODEX data of new donors (Fig. 3b and Supplementary Fig. 4) agree well with the cell-type map predicted by STELLAR and also agree with the known distribution of main cell types in the intestine (Fig. 3c and Extended Data Fig. 8). We additionally confirmed the quality of STELLAR's predictions by looking at average marker expression profiles of predicted cell types for the samples from a different donor (Fig. 3d). Protein marker distributions match expert hand-annotated profiles in fine-grained cell types predicted by STELLAR. For example, CD4 is expressed only in cell types annotated as $CD4^+$ T cells, whereas CD8 is expressed exclusively in $CD8^+$ T cells. This additionally confirms that STELLAR predictions are reliable. We further applied STELLAR to six other donors, for a total of eight donors, across

https://doi.org/10.1038/s41592-022-01651-8

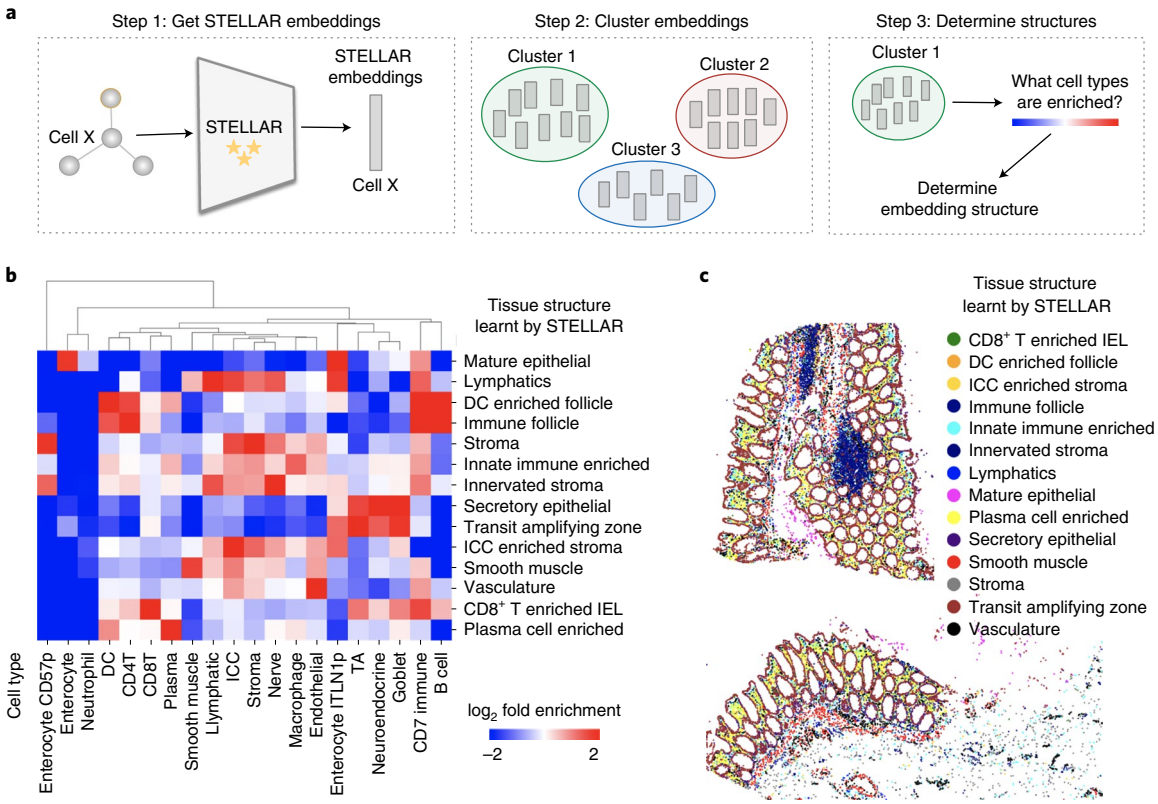


Fig. 4 | STELLAR's embeddings reveal higher-order tissue structures. a, To analyze embedding space learned by STELLAR, we clustered the cell embeddings using Louvain clustering. We then analyzed microenvironments of resultant clusters by checking which cell types are enriched in each cluster. The resultant clusters were manually annotated based on the cell-type enrichments showed in **b**. **b**, Resultant clusters in STELLAR's embedding space reveal that the embeddings capture main structures within the intestinal tissue. Heatmap shows enrichment of cell types assigned by STELLAR within the structural clusters as compared to tissue average percentages. **c**, Graphical display of the tissue structures learned by STELLAR found for one of the regions of the colon. Main structures such as immune follicles were identified by STELLAR. IEL stands for intraepithelial lymphocytes.

64 tissues and 2.6 million cells. STELLAR alleviated expert annotation effort tremendously: human annotations of these images would require approximately 320 hours of manual labor, while with STELLAR it took the expert only 4 hours to annotate the images.

We further validated STELLAR on datasets generated from different donors, which were prepared and stained at different times by different tissue handlers and were formulated into TMA (tissue microarray) or large tissue blocks, had different fixations conditions, used different image processing and cell segmentation algorithms, and used different antibody panels and clones for antibodies. This represents a particularly difficult case for cell-type label transfer. We used fresh frozen CODEX intestine HuBMAP dataset as a reference dataset and applied STELLAR to a CODEX formalin-fixed, paraffin-embedded colorectal cancer dataset[25]. We used the subset of antigens targeted that overlapped even if the clone was different. While only a subset of the marker set overlapped, we found that STELLAR outperformed other methods by at least 22% (Supplementary Fig. 5).

STELLAR embeddings capture higher-order tissue structures

The classification of cell–cell interactions enabled by multiplexed imaging of many cell types simultaneously unlocks the ability to identify and discover multicellular structures. Characterization of multicellular modules such as tertiary lymphoid structures is critical to understanding tissue function, its relationship with disease and how to design effective therapeutics[34]. We hypothesized that the latent cell embedding space learned by STELLAR could reveal biologically meaningful information about the tissue organization. To explore this, we obtained cell embeddings from STELLAR and then clustered these cells in the STELLAR's embedding space using the Louvain clustering[35] (Fig. 4a). We then evaluated the microenvironments represented by the resulting clusters by calculating the enrichment of cell types that neighbor these cells within these clusters as compared to tissue averages. This was also verified from looking at where the cell types with assigned cluster labels fell within the tissue using fluorescence staining and cell-type neighbors to confirm larger pathological structures. We found that the clusters represent main multicellular structural features of the intestine such as the immune follicle structure, which was enriched for $CD4^+$ T cells, B cells and dendritic cells, as well as secretory epithelial, which was enriched for goblet cells, neuroendocrine cells and transit amplifying cells (Fig. 4b,c). We also confirmed this works across tissue sections taken at different sites of the colon with similar architectures (Extended Data Fig. 9). Similarly, we find that high-level structural information was detected also for the MERFISH data (Extended Data Fig. 10). These analyses revealed that the embedding space of STELLAR provides a method for identifying multicellular structures in tissues.

Discussion

STELLAR is an effective framework for annotating spatially resolved protein and RNA single-cell data and extracting spatial cell-interaction information. Two properties make STELLAR a unique tool in the single-cell toolbox: (1) the ability to learn low-dimensional cell embeddings that leverage spatial and molecular information across different biological contexts, and (2) automatic identification of cell types from the reference set and discovery of novel cell types not present in the reference dataset. STELLAR learns a new cell embedding space that leverages spatial and molecular features using graph convolutional

neural networks. The cell embeddings learned by STELLAR can be used for any downstream analysis and provide a novel capability to identify main multicellular structures in the tissue. Our results support that incorporating spatial information directly leads to improved cell-type annotation performance.

STELLAR is intended to be used for transferring annotations across datasets, including datasets from different biological contexts, for example, across dissection regions, donors or related tissues. While doing so, STELLAR discovers expression patterns that define a novel cell type or cellular state. Our tool has a great value in transferring annotations across levels of granularity to a new biological context and discovering novel biological states that have not been characterized in previous experiments. The annotation transfer methodology in STELLAR distinguishes it from previous spatial annotation tools that rely on predefined marker genes that define cell types[16].

At present, segmentation tools used largely rely on nuclear segmentation, and those that incorporate membrane-based segmentation are far from perfect[36,37]. Once segmentation masks and measurements become more reliable in regards to shapes of cells, then morphological features of cells could be included in STELLAR as feature vectors in addition to the marker expression. By including morphological features of cells, further improvements in performance could be expected as well as identification of unique cell states.

Development of STELLAR was motivated by a growing need to leverage spatial and molecular information across different biological contexts[17]. Multiplexed imaging technologies will drive future efforts to understand both healthy and diseased tissue processes, enabled by large consortia efforts that generate comprehensive datasets and standardize computational methods[5–7]. Novel insights will come from the rich, but still unexplored and underused information about spatial cell organization. We anticipate that STELLAR will alleviate future annotation efforts of large-scale spatial cell atlases, identify main multicellular structures in tissues and reveal how cells cooperatively coordinate to enable tissues to function.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-022-01651-8.

References

1. Lewis, S. M. et al. Spatial omics and multiplexed imaging to explore cancer biology. *Nat. Methods* **18**, 997–1012 (2021).
2. Bodenmiller, B. Multiplexed epitope-based tissue imaging for discovery and healthcare applications. *Cell Systems* **2**, 225–238 (2016).
3. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
4. Hickey, J. W. et al. Spatial mapping of protein composition and tissue organization: a primer for multiplexed antibody-based imaging. *Nat. Methods* **19**, 284–295 (2021).
5. HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019).
6. Rozenblatt-Rosen, O. et al. The Human Tumor Atlas Network: charting tumor transitions across space and time at single-cell resolution. *Cell* **181**, 236–249 (2020).
7. Regev, A. et al. Science forum: the Human Cell Atlas. *eLife* **6**, e27041 (2017).
8. Zhang, M. et al. Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature* **598**, 137–143 (2021).
9. Black, S. et al. CODEX multiplexed tissue imaging with DNA-conjugated antibodies. *Nature Protocols* **16**, 3802–3802 (2021).
10. Goltsev, Y. et al. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell* **174**, 968–981 (2018).
11. Teng, H., Yuan, Y. & Bar-Joseph, Z. Clustering spatial transcriptomics data. *Bioinformatics* **38**, 997–1004 (2021).
12. Partel, G. & Wählby, C. Spage2vec: unsupervised representation of localized spatial gene expression signatures. *FEBS J* **288**, 1859–1870 (2021).
13. Zhao, E. et al. Spatial transcriptomics at subspot resolution with BayesSpace. *Nat. Biotech.* **39**, 1375–1384 (2021).
14. Hu, J. et al. SpaGCN: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* **18**, 1342–1351 (2021).
15. Zeng, Z., Li, Y., Li, Y. & Luo, Y. Statistical and machine learning methods for spatially resolved transcriptomics data analysis. *Genome Biol* **23**, 83 (2022).
16. Zhang, W. et al. Identification of cell types in multiplexed in situ images by combining protein expression and spatial information using CELESTA. *Nat. Methods* **19**, 759–769 (2022).
17. Hickey, J. W. et al. High resolution single cell maps reveals distinct cell organization and function across different regions of the human intestine. Preprint at *bioRxiv* (2021).
18. Greenbaum, S. et al. Spatio-temporal coordination at the maternal-fetal interface promotes trophoblast invasion and vascular remodeling in the first half of human pregnancy. Preprint at *bioRxiv* (2021).
19. Currlin, S. et al. 3D-mapping of human lymph node and spleen reveals integrated neuronal, vascular, and ductal cell networks. Preprint at *bioRxiv* (2021).
20. Neumann, E. K. et al. A multiscale atlas of the molecular and cellular architecture of the human kidney. Preprint at *bioRxiv* (2022).
21. Lake, B. B. et al. An atlas of healthy and injured cell states and niches in the human kidney. Preprint at *bioRxiv* (2021).
22. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks, in *Proc. International Conference on Learning Representations* (2016).
23. Hamilton, W., Ying, Z. & Leskovec, J. Inductive representation learning on large graphs. in *Proc. Adv. Neural Inform. Proc. Syst.* **30** (eds Guyon, I. et al.) (2017).
24. Cao, K., Brbic, M. & Leskovec, J. Open-world semi-supervised learning, in *Proc. International Conference on Learning Representations* (2022).
25. Schürch, C. M. et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell* **182**, 1341–1359 (2020).
26. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system, in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (eds Krishnapuram, B. et al.) (2016).
27. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
28. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
29. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997).
30. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
31. Kimmel, J. C. & Kelley, D. R. Semi-supervised adversarial neural networks for single-cell classification. *Genome Res.* **31**, 1781–1793 (2021).

32. Xu, C. et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).
33. Hickey, J. W., Tan, Y., Nolan, G. P. & Goltsev, Y. Strategies for accurate cell type identification in CODEX multiplexed imaging data. *Front. Immunol.* 3317 (2021).
34. Sautès-Fridman, C., Petitprez, F., Calderaro, J. & Fridman, W. H. Tertiary lymphoid structures in the era of cancer immunotherapy. *Nat. Rev. Cancer* **19**, 307–325 (2019).
35. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
36. Hollandi, R. et al. Nucleus segmentation: towards automated solutions. *Trends Cell Biol.* **32**, 295–310 (2022).
37. Van Buren, K. et al. Artificial intelligence and deep learning to map immune cell types in inflamed human tissue. *J. Immunol. Methods* **505**, 113233 (2022).

Methods

Overview of STELLAR

STELLAR learns spatial and molecular cell similarities that are transferable across different biological contexts, such as different dissections regions, donors or tissues. Across different contexts, STELLAR learns to automatically assign cells to cell types seen in the annotated reference set, or forms novel cell types if cells have unique properties that are not present in the reference dataset.

Specifically, STELLAR starts with an annotated reference cell graph $\mathcal{G}_a = \{\mathcal{V}_a, \mathcal{E}_a\}$ with molecular features for all nodes $\mathbf{x}_{v_i} \in \mathbb{R}^D, \forall v_i \in \mathcal{V}_a$, and an unannotated cell graph $\mathcal{G}_u = \{\mathcal{V}_u, \mathcal{E}_u\}$ with molecular features $\mathbf{x}_{v_i} \in \mathbb{R}^D, \forall v_i \in \mathcal{V}_u$. Here, $\mathcal{V}$ denotes the set of vertices and $\mathcal{E}$ denotes the set of edges in the graph $\mathcal{G}$. The nodes in each graph correspond to cells and cells are connected if they are spatially close. Node features correspond to gene or protein expressions of cells where D denotes the total number of measured genes or proteins. For the reference graph we assume we are given a vector of cell annotations $\mathbf{y} = \{y_i \in \{1, ..., K_a\}\}_{i=1}^{|\mathcal{V}_a|}$ that assigns each cell to one of the K_a cell types (or other annotations). STELLAR is ran on the shared subsets of the genes or proteins between annotated reference dataset and unannotated target dataset.

Given reference graph $\mathcal{G}_a$ and unannotated graph $\mathcal{G}_u$, STELLAR first applies the encoder function $f_\theta : \mathbb{R}^D \to \mathbb{R}^d$ that maps cells from both graphs into a joint embedding space that captures spatial and molecular similarities between the cells. The cell embedding encoder function f_θ is parameterized by learnable parameters θ of a graph convolutional neural network (GCN)[22]. The encoder function f_θ generates d-dimensional cell embeddings $\mathbf{z}_i \in \mathbb{R}^d, \forall v_i \in \mathcal{V}_a \cup \mathcal{V}_u$. On top of the encoder function, we add a single linear layer parameterized by a weight matrix $W : \mathbb{R}^D \to \mathbb{R}^{K_a+K_n}$, where K_n corresponds to expected number of novel cell types that need to be discovered. Next, a softmax layer is added that assigns each cell to one of the $K_a + K_n$ cell types.

Graph construction

Given spatial cell coordinates of each cell, STELLAR first calculates Euclidean distances $d_{i,j}$ for each pair of cells (v_i, v_j) from the same region, and edge (v_i, v_j) is added to the edge set $\mathcal{E}$ if $d_{i,j} < \tau$, where τ is a tunable threshold. We select the value of $\tau = 50$ μm and experiment with different threshold values in Fig. 2h. Graph construction step is independent of the subsequent method and can be changed as long as the constructed graph meaningfully reflects spatial similarities between cells.

STELLAR encoder

The STELLAR encoder contains one fully connected layer followed by the nonlinear activation function:

$$\mathbf{h}_i^{(1)} = \phi(W^{(0)}\mathbf{h}_i^{(0)} + \mathbf{b}^{(0)}), \tag{1}$$

where $\mathbf{h}_i^{(k)}$ is the hidden state of node v_i in kth layer of the neural network and $k = 0, 1$. W is a parameter matrix, $\mathbf{b}$ is bias vector and ϕ denotes nonlinear activation function. The hidden state $\mathbf{h}_i^{(0)}$ in layer 0 is set to node features $\mathbf{x}_{v_i}$ that is, a gene or protein expression vector. The rectified linear unit (ReLU) is used as the activation function ϕ: $\mathrm{ReLU}(\cdot) = \max(0, \cdot)$.

We then use a graph convolutional layer[22] to enable message passing among nearby cells:

$$\mathbf{h}_i^{(2)} = W_0^{(1)}\mathbf{h}_i^{(1)} + \sum_{j \in \mathcal{N}^i} W_1^{(1)}\mathbf{h}_j^{(1)}, \tag{2}$$

where $\mathcal{N}^i$ denotes neighborhood on node v_i. The final embedding of node v_i is $\mathbf{z}_i = \mathbf{h}_i^{(2)}$. In nonspatial STELLAR (Fig. 2f), we replace the graph convolutional layer with another fully connected layer.

STELLAR initialization

We initialize novel cell types with unlabeled data that lie out of distribution with respect to the labeled data. We first train a single-layer neural network g_ψ on labeled data with cross-entropy loss. We cluster the unlabeled data using the Louvain algorithm[35] with default parameters in scanpy[38] and then summarize the average entropy of each cluster with the trained g_ψ. We select K_n clusters with largest entropy and assign pseudo-labels corresponding to novel cell types as cells in the same cluster.

STELLAR objective function

The objective function in STELLAR assigns cells to cell types from the reference set or discovers novel cell types. Inspired by ref. [24], the objective function consists of two main components: (1) a component for discovering novel cell types, and (2) a component for learning to recognize cell types from the reference set (Supplementary Fig. 1).

In the component for novel cell-type discovery, we use an objective term that predicts pairwise similarities given cell embeddings obtained using STELLAR encoder, that is, we predict whether two cells are similar or not. For the reference graph, we use ground-truth annotations to learn to predict similarity between two cells, that is two cells are similar if they belong to the same cell type. For the unannotated graph, pseudo-labels are generated based on the distances between cells in the embedding space. In particular, for each cell within the mini-batch, we identify the most similar nearest neighbor cell and generate pseudo-labels for the given pair. In that way, pseudo-labels are generated only for the pairs in which there is the most confidence. To find nearest neighbors, cell embeddings $\mathbf{h}^{(1)}$ without ReLU are used to allow cells from the reference graph to be selected as the neighbors of cells from the unannotated graph. Formally, the component for discovering cell types (DCT) minimizes the following term:

$$\mathcal{L}_{\mathrm{DCT}} = \frac{1}{N_a + N_u} \sum_{\substack{\mathbf{z}_i, \mathbf{z}' \in \\ (\mathcal{Z}_a \cup \mathcal{Z}_u, \mathcal{Z}'_a \cup \mathcal{Z}'_u)}} -\log\langle\sigma(W^T\mathbf{z}_i), \sigma(W^T\mathbf{z}'_i)\rangle, \tag{3}$$

where W denotes a linear layer weight matrix, σ denotes a softmax function, N_a and N_u denote numbers of cells for annotated and unannotated graphs, respectively. $\mathcal{Z}_a$ and $\mathcal{Z}_u$ denote cell embeddings for annotated and unannotated graphs, and $\mathcal{Z}'$ and $\mathcal{Z}'$ denote the set of closest neighbors of $\mathcal{Z}_a$ and $\mathcal{Z}_u$ used for generating pseudo-labels.

The second component for recognizing cell types learns to distinguish cell types from the reference graph using ground-truth cell annotations y. Standard cross-entropy classification loss was enhanced with an adaptive margin mechanism that controls the learning speed of cell types from the reference set compared to novel cell types. Formally, STELLAR minimizes the following objective term to learn to recognize cell types (RCT) in the reference graph:

$$\mathcal{L}_{\mathrm{RCT}} = \frac{1}{N_a} \sum_{\mathbf{z}_i \in \mathcal{Z}_a} -\log \frac{e^{s(W_{y_i}^T\mathbf{z}_i + \bar{u})}}{e^{s(W_{y_i}^T\mathbf{z}_i + \bar{u})} + \sum_{k \neq y_i} e^{sW_k^T\mathbf{z}_i}}, \tag{4}$$

where s is temperature scaling parameter, $\bar{u}$ is uncertainty and W_{y_i} refers to the column vector that relates to class y_i. The uncertainty is estimated as the average confidence of unlabeled examples computed from the output of the softmax function:

$$\bar{u} = \frac{1}{N_u} \sum_{v_i \in \mathcal{V}_u} 1 - \max_k \Pr(y = k | v_i), \tag{5}$$

where k goes over all reference and novel cell types. At the start of the training uncertainty is large, leading to a larger margin and forcing larger intraclass variance[39]. As training proceeds, the margin becomes smaller and the objective boils down to standard cross-entropy.

Additionally, we use maximum entropy regularization term to avoid a trivial solution of assigning all cells to the same cell type. In particular, the regularization term is the following:

$$\mathcal{R} = \frac{1}{N_a + N_u} \sum_{z_i \in \mathcal{Z}_l \cup \mathcal{Z}_u} \sum_k \sigma_k(W^T z_i) \log \sigma_k(W^T z_i), \quad (6)$$

where k goes over all reference and novel cell types and σ_k denotes kth cell-type value of the softmax output.

Finally, the objective function in STELLAR combines reference cell-type recognition, novel cell-type discovery and regularization components:

$$\mathcal{L}_* = \mathcal{L}_{RCT} + \eta_1 \mathcal{L}_{DCT} + \eta_2 \mathcal{R}, \quad (7)$$

where η_1 and η_2 are regularization parameters.

Architecture and hyperparameters

The encoder network in STELLAR consists of one fully connected layer with ReLU activation and a graph convolutional layer with a hidden dimension of 128 in all layers. It uses the Adam optimizer with an initial learning rate of 10^{-3} and weight decay 0. The model is trained with a batch size of 512 for 20 epochs. A cluster sampler[40] first clusters input graphs into subgraphs and then assigns the subgraphs to mini-batches. The temperature scaling parameter s in equation (4) is set to 10. Regularization parameters in equation (7), η_1 add η_2, are set to 1 and 0.3, respectively. These hyperparameters were used across all experiments.

Number of novel cell types

STELLAR is initialized with the expected number of novel cell types as an input parameter. As the number of novel cell types is usually not known, STELLAR can be initialized with a large number of novel cell types and will automatically reduce the number by not assigning any cells to unneeded classification heads.

Neighborhood identification analysis

Neighborhood analysis was performed as described previously[25]. Briefly, a window size of ten nearest neighbors for each cell was taken across the tissue cell-type maps. These vectors were overclustered to 20 clusters using k-means clustering algorithm. The clusters were mapped back to the tissue and evaluated for cell-type enrichments to determine overall structure and merged down into final structures.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The CODEX datasets presented in this study can be found in the online repository Dryad at https://datadryad.org/stash/share/1OQtxew0Unh3iAdP-ELew-ctwuPTBz6Oy8uuyxqliZk. Specifically, the quantified single-cell data are provided (with cells in rows and protein expression, xy position and cell-type labels in columns). Additionally, we provide datasets used to transfer from the tonsil to BE tissue (BE_Tonsil_dryad.csv) and expert-annotated healthy human intestine (B004_training_dryad.csv), which was used to test the accuracy of STELLAR across the four regions of the colon regions of this dataset and also for training for transferring cell-type labels to unlabeled donors (B0056_unannotated_dryad.csv). MERFISH mouse cortex datasets are from Ref. [8].

Code availability

STELLAR was written in Python v.3.8 using the PyTorch library. The source code is available on Github at https://github.com/snap-stanford/stellar. The project website with links to data and code can be accessed at http://snap.stanford.edu/stellar/.

References

38. Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15 (2018).
39. Liu, B. et al. Negative margin matters: understanding margin in few-shot classification, in *Proc. European Conference on Computer Vision*, 438-455 (eds Vedaldi, A. et al) (2020).
40. Chiang, W.-L. et al. Cluster-GCN: an efficient algorithm for training deep and large graph convolutional networks, in *Proc. ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 257–266 (eds Teredesai, A. et al.) (2019).

Acknowledgements

This work was supported by the US National Institutes of Health (grant nos. 2U19AI057229-16, 5P01HL10879707, 5R01GM10983604, 5R33CA18365403, 5U01AI101984-07, 5UH2AR06767604, 5R01CA19665703, 5U54CA20997103, 5F99CA212231-02, 1F32CA233203-01, 5U01AI140498-02, 1U54HG010426-01, 5U19AI100627-07, 1R01HL120724-01A1, R33CA183692, R01HL128173-04, 5P01AI131374-02, 5UG3DK114937-02, 1U19AI135976-01, IDIQ17X149, 1U2CCA233238-01 and 1U2CCA233195-01); Cancer Research UK (grant no. C27165/A29073); and the Parker Institute for Cancer Immunotherapy. J.W.H. was supported by an NIH T32 Fellowship (grant no. T32CA196585) and an American Cancer Society: Roaring Fork Valley Postdoctoral Fellowship (grant no. PF-20-032-01-CSM). We also gratefully acknowledge the support of DARPA under grant nos. HR00112190039 (TAMI), N660011924033 (MCS); ARO under grant nos. W911NF-16-1-0342 (MURI), W911NF-16-1-0171 (DURIP); NSF under grant nos. OAC-1835598 (CINES), OAC-1934578 (HDR), CCF-1918940 (Expeditions), IIS-2030477 (RAPID), NIH under grant no. R56LM013365; Stanford Data Science Initiative, Wu Tsai Neurosciences Institute, Amazon, JPMorgan Chase, Docomo, Hitachi, Juniper Networks, Intel, KDDI and Toshiba.

Author contributions

M.B., K.C., J.W.H. and J.L. conceived the research. M.B., K.C., J.W.H. and Y.T. performed research and analyzed results. M.B., K.C. and J.L. contributed new analytical tools and created the algorithmic framework. J.W.H., M.P.S. and G.P.N. generated and analyzed the data. J.L., G.P.N. and M.P.S. supervised the research. All authors participated in interpretation and wrote the manuscript.

Competing interests

M.P.S. is cofounder and advisory board member of Personalis, Qbio, January AI, Mirvie, Filtricine, Fodsel, Protos. RTHM, Marble Therapeutics and Crosshair Therapeutics. G.P.N. has equity in and is a scientific advisory board member of Akoya Biosciences, Inc. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41592-022-01651-8.

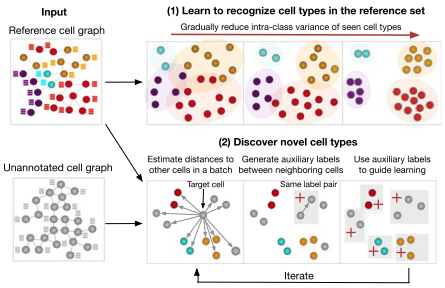
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-022-01651-8.

Correspondence and requests for materials should be addressed to Garry P. Nolan or Jure Leskovec.

Peer review information *Nature Methods* thanks Ellis Patrick, Darren Tyson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Rita Strack, in collaboration with the *Nature Methods* team.

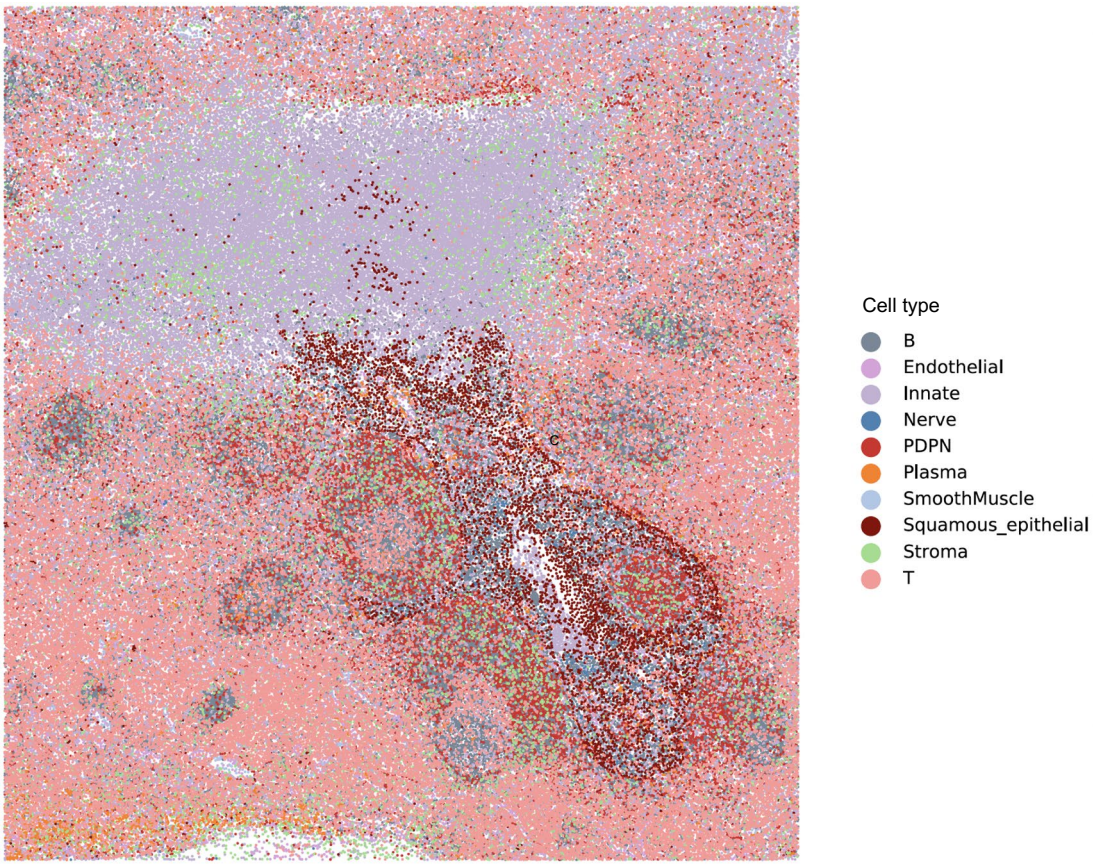
Reprints and permissions information is available at www.nature.com/reprints.

https://doi.org/10.1038/s41592-022-01651-8



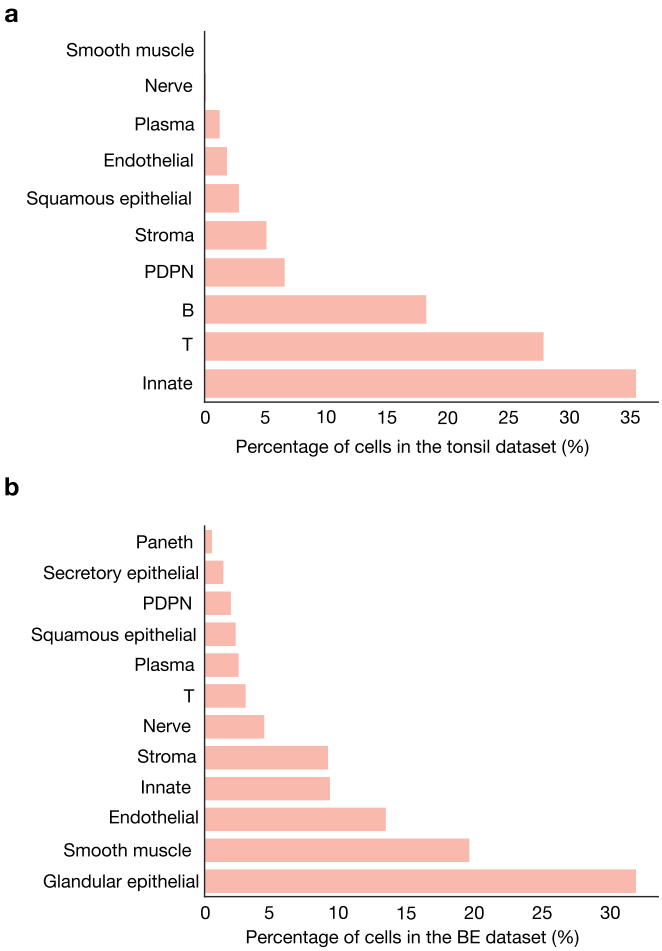
Extended Data Fig. 1 | STELLAR overview. STELLAR is a unique method in its ability to simultaneously recognize cell types seen in the reference set and discover novel cell types that have never been characterized in the reference set. This is made possible by an objective function that consists of two main components (Methods). First, STELLAR learns to gradually separate cell types from the reference set by controlling intra-class variance to allow the model to simultaneously learn to discover novel cell types. Simultaneously, STELLAR discovers novel classes by generating auxiliary labels (pseudo-labels) in the unannotated graph that are used to guide the training. The auxiliary labels are generated based on the nearest neighbors of each cell in the embedding space.

https://doi.org/10.1038/s41592-022-01651-8



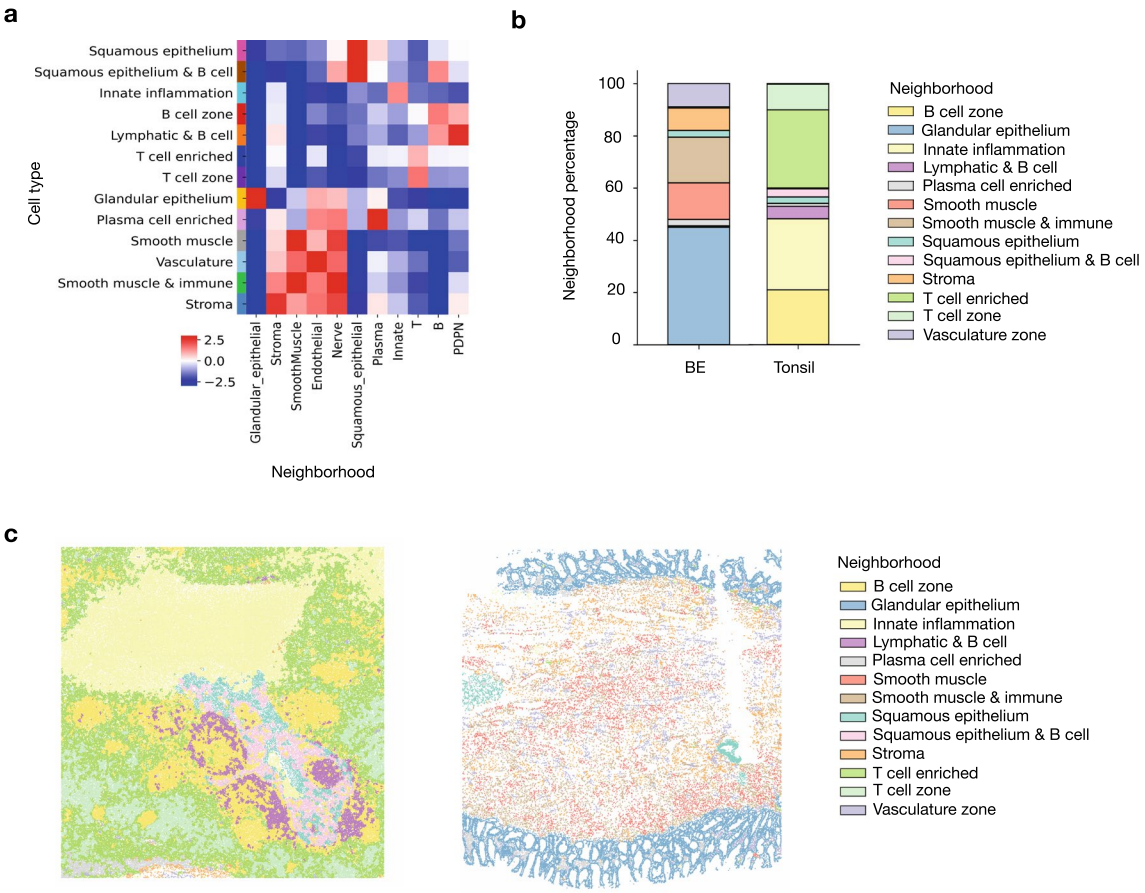
Extended Data Fig. 2 | CODEX image of reference dataset from human tonsil. Ground-truth labels of the tonsil CODEX multiplexed imaging dataset. Colors denote different cell types.

https://doi.org/10.1038/s41592-022-01651-8



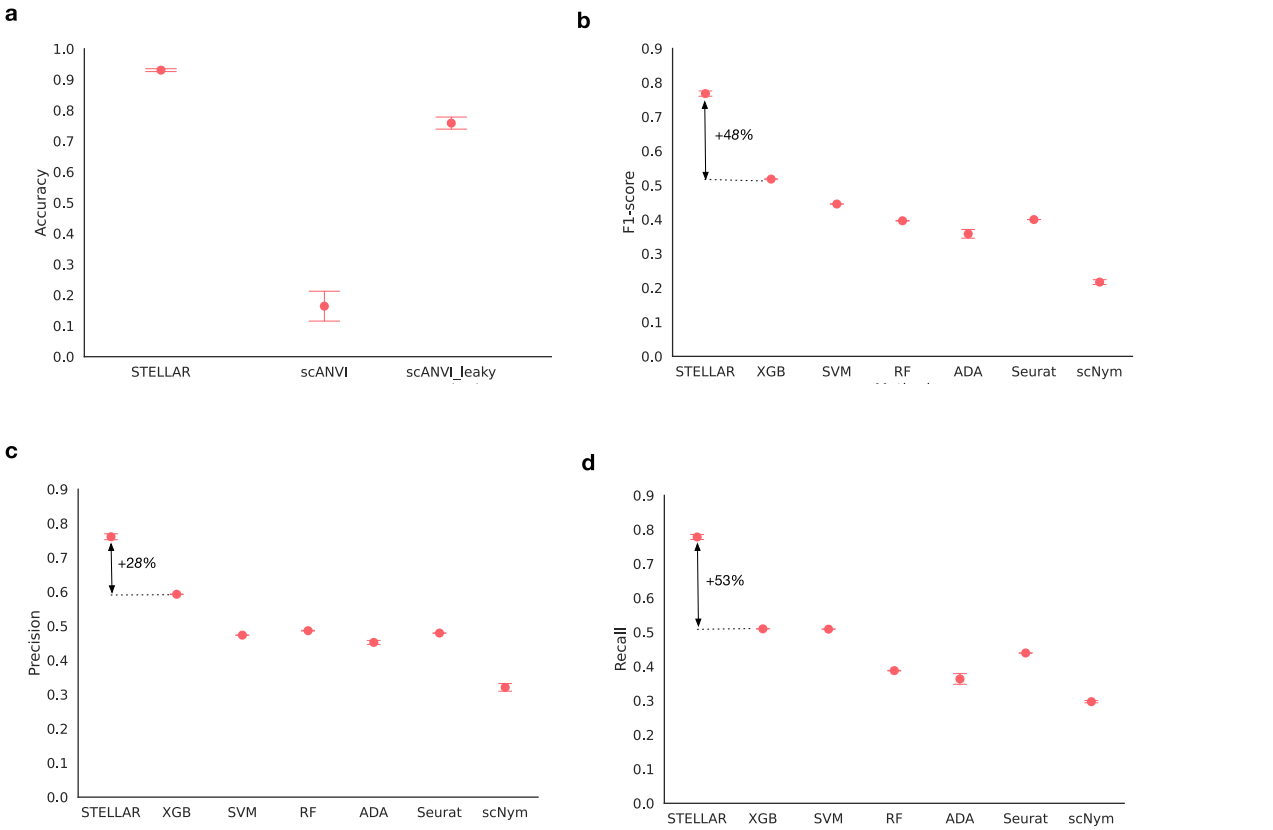
Extended Data Fig. 3 | Cell-type distributions on tonsil and BE datasets. Cell-type distributions of ground-truth labels on (a) tonsil reference dataset and (b) Barrett's esophagus dataset. PDPN stands for Podoplanin (PDPN) positive stromal cells.

https://doi.org/10.1038/s41592-022-01651-8



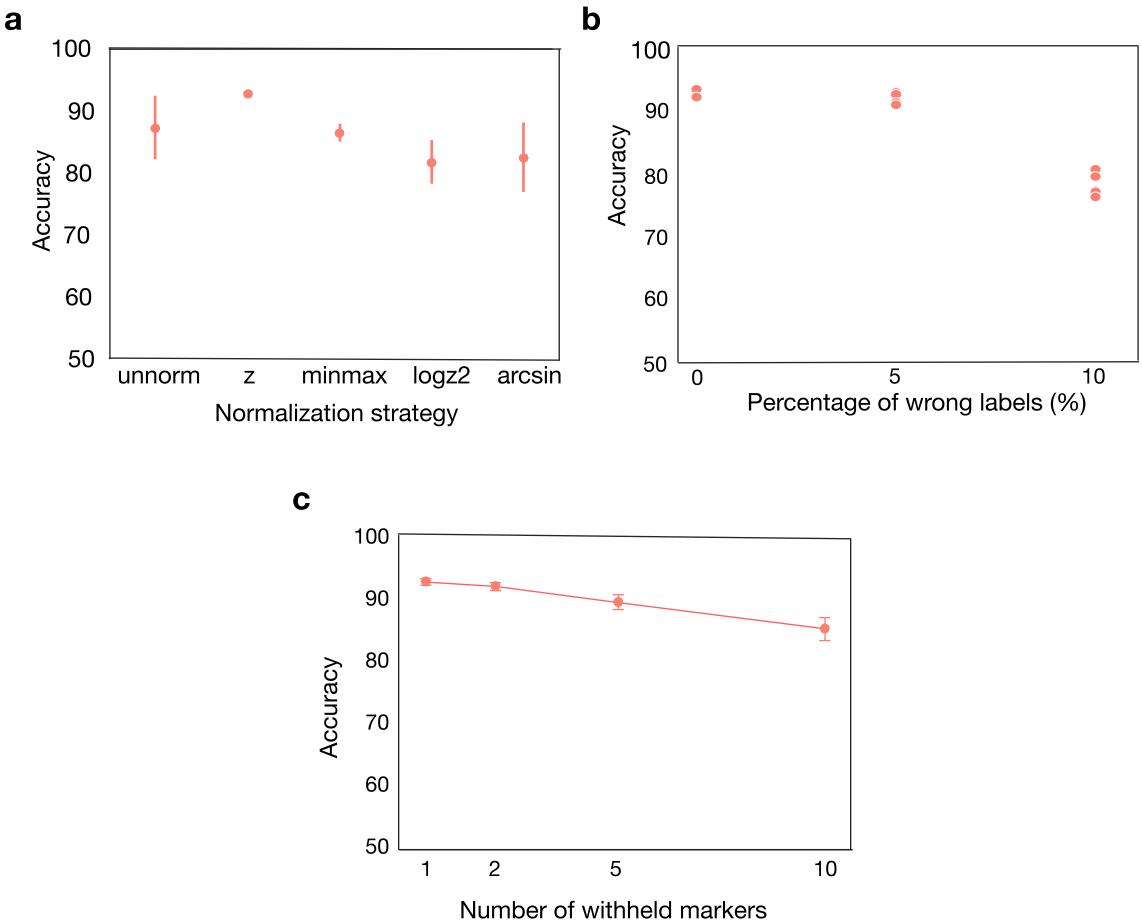
Extended Data Fig. 4 | Neighborhoods found in tonsil and Barrett's esophagus (BE) dataset. (a) Neighborhood heatmap showing the neighborhoods found across both tissues and cell types enriched compared to tissue averages. (b) Neighborhood composition between BE and tonsil tissues. (c) Neighborhood types mapped back to tissue coordinates.

https://doi.org/10.1038/s41592-022-01651-8



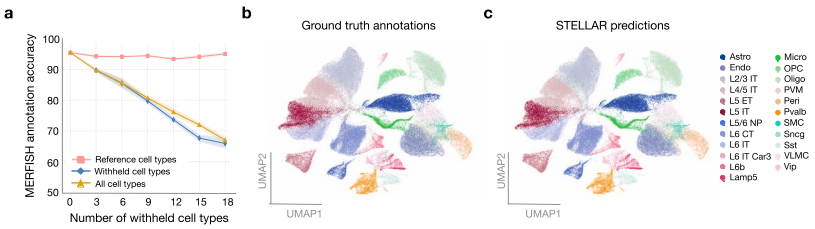
Extended Data Fig. 5 | Comparison of STELLAR to baseline methods on the Barrett's esophagus (BE) dataset. (a) Accuracy of STELLAR and scANVI on the BE dataset. Performance was evaluated as a mean score across n=5 runs of each method. Error bars are from standard deviation. scANVI stands for the setting evaluated in the same manner as STELLAR in which we train the model on tonsil dataset and evaluate on BE dataset. scANVI_leaky stands for the approach in which we use fraction of labels from BE dataset as the training data and use the rest of the BE dataset as the test set. Although the setting in which scANVI_leaky is evaluated does not present a fair comparison to STELLAR and other baselines, it indicates that the performance of drop of scANVI is caused by differences between tonsil and BE datasets. (b-d) Performance of STELLAR and alternative baselines on the BE dataset evaluated as (b) mean macro F1-score, (c) macro precision score, and (d) macro recall score across n=5 runs of each method. Error bars are from standard deviation. XGB stands for XGBoost, SVM for Support Vector Machine, RF for Random Forest, ADA for ADABoost, and Seurat for Seurat V4.

https://doi.org/10.1038/s41592-022-01651-8



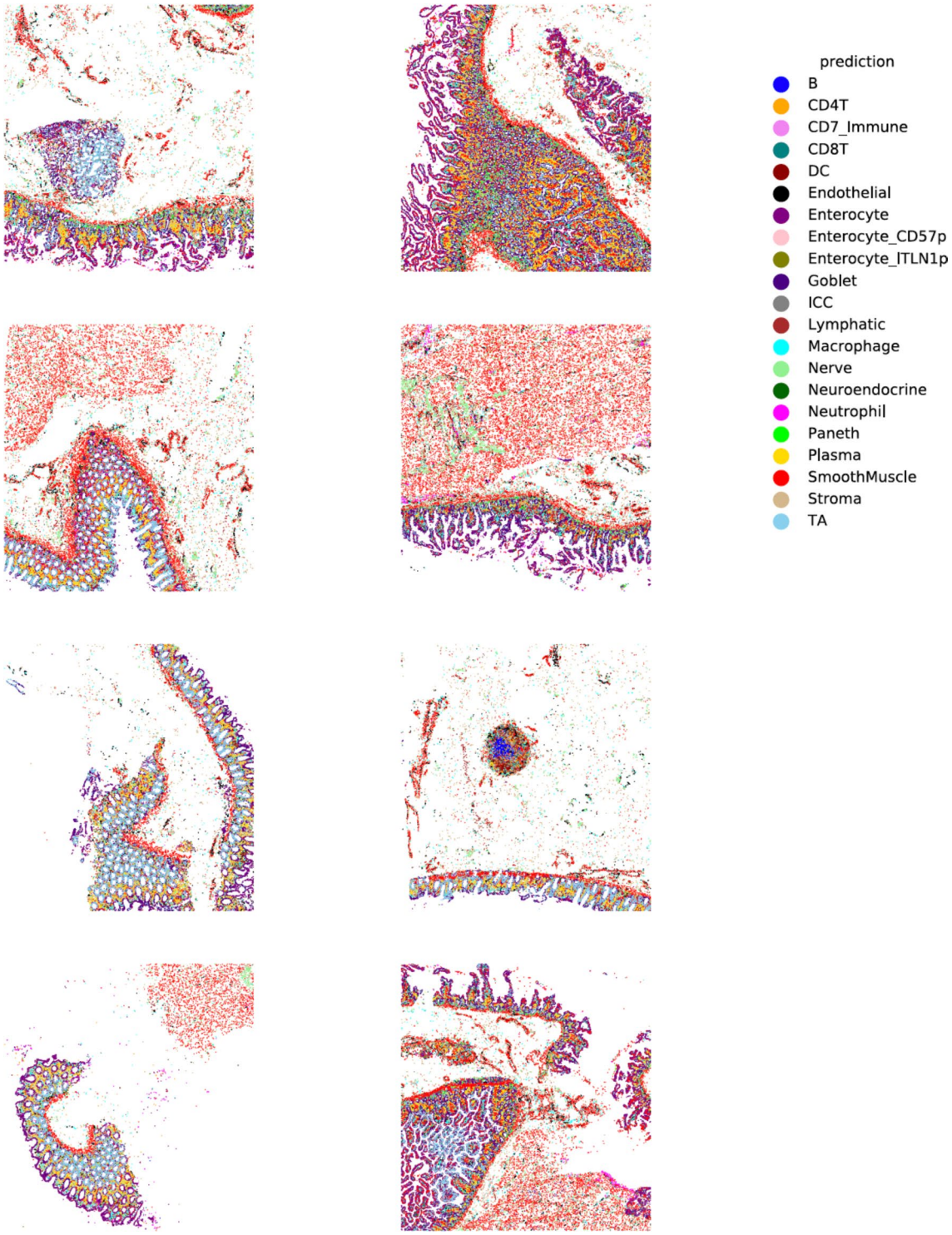
Extended Data Fig. 6 | Robustness of STELLAR evaluated on the Barrett's esophagus (BE) dataset. (a) Performance of STELLAR using different normalization strategies. 'Unnorm' stands for raw (unnormalized) data. Performance was evaluated as a mean accuracy score across n=5 runs of each normalization strategy. Error bars are from standard deviation. (b) Performance of STELLAR when misannotating proportion of randomly selected cells. In each run, cells were randomly selected and labels different than ground truth annotations were randomly assigned to cells in the annotated reference tonsil dataset. Performance was evaluated as an accuracy score across n=5 runs. Individual data points are shown. (c) Performance of STELLAR when removing different number of marker genes. In each run, different set of randomly selected marker genes was withheld from the reference tonsil dataset and BE datasets. Performance was evaluated as a mean accuracy score across n=5 runs. Error bars are from standard deviation.

https://doi.org/10.1038/s41592-022-01651-8



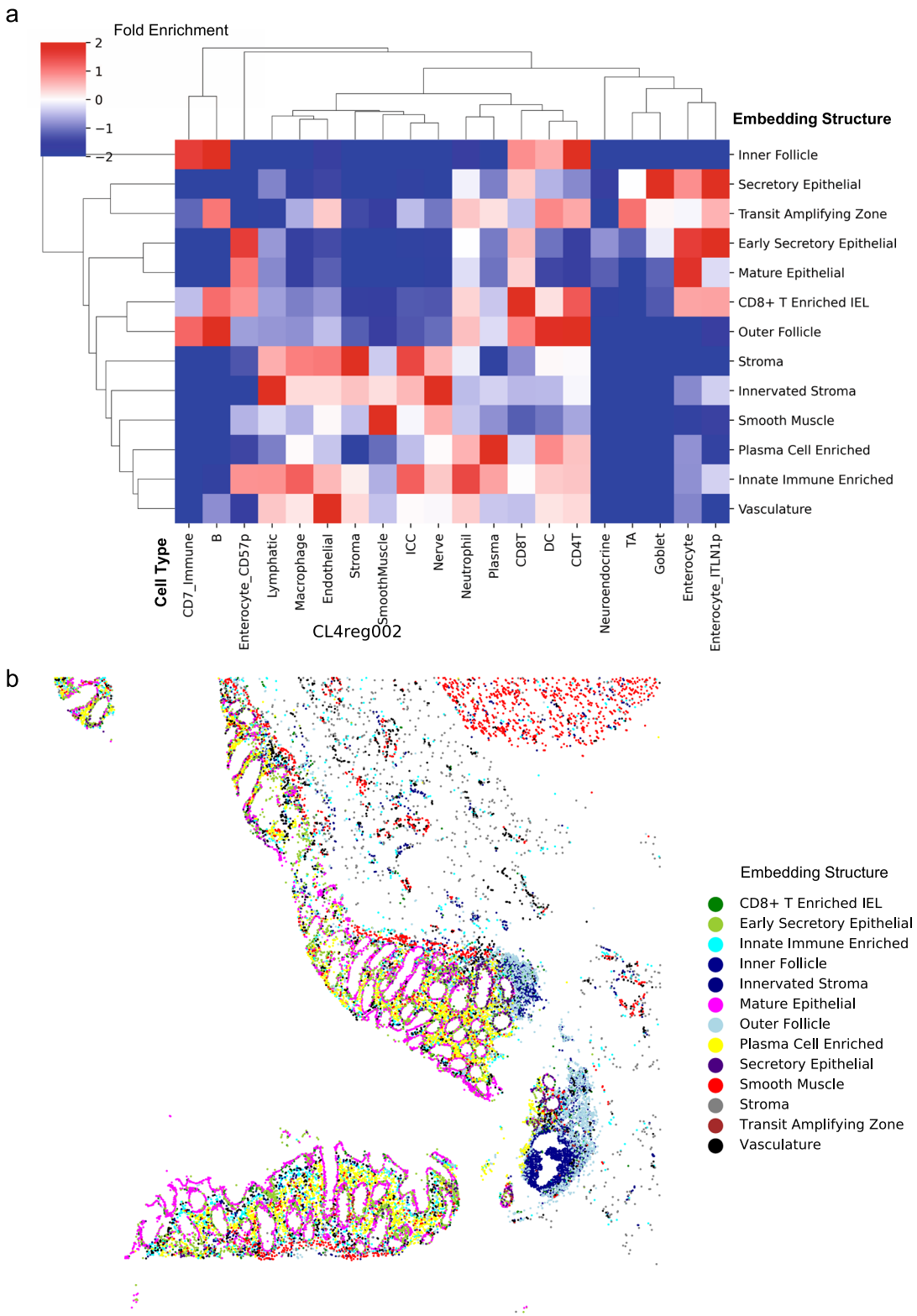
Extended Data Fig. 7 | Performance of STELLAR on the MERFISH dataset from mouse cortex. We applied STELLAR to a large-scale mouse primary motor cortex MERFISH dataset consisting of 23 granular cell types from two mice [8]. (a) Annotation accuracy of STELLAR on the MERFISH mouse cortex dataset with different numbers of withheld cell types. Position of scatter plot points is computed as a mean accuracy score across n=5 runs. Error bars are from standard deviation. We randomly removed a number of cell types from the reference set and evaluated STELLAR's performance by gradually increasing the number of removed cell types. We measured accuracy separately on classes seen in the reference set and classes withheld from the reference set. Performance is evaluated on the reference cell types, novel cell types withheld from the reference set during training, and jointly on all cell types. (b, c) UMAP visualization of MERFISH mouse cortex dataset from mouse used as the test set. Cells are colored according to (b) ground-truth annotations, and (c) STELLAR's predictions without any withheld cell types.

https://doi.org/10.1038/s41592-022-01651-8



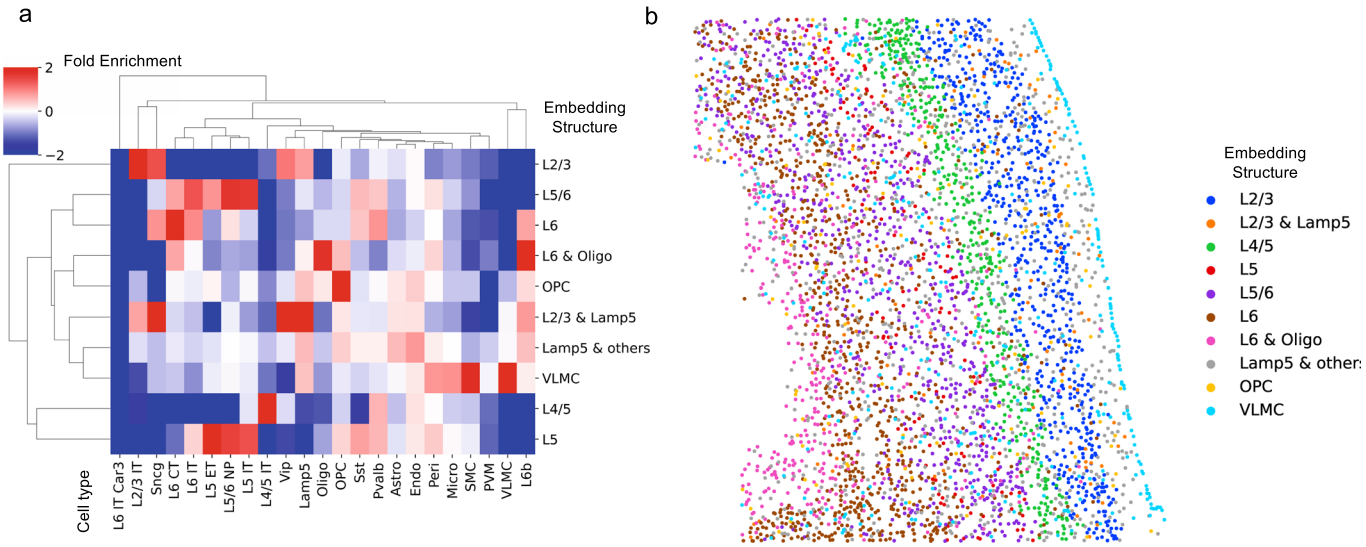
Extended Data Fig. 8 | STELLAR predictions on the dataset from healthy intestine. CODEX-imaged regions with cell types colored by prediction from STELLAR using data from the healthy intestine of a different donor as the reference set. Data from both small intestine and colon are shown. Colors denote different cell types. DC stands for dendritic cell, ICC stands for interstitial cells of Cajal, TA stands for transit amplifying cell.

https://doi.org/10.1038/s41592-022-01651-8



Extended Data Fig. 9 | Multicellular structures discovered bt STELLAR on CODEX healthy intestine data. Characterization of multicellular structures by clustering the embedding space from STELLAR on CODEX healthy intestine data. (a) Heatmap of average cell-type composition in clustered embeddings. (b) Representative tissue image colored by embedding structure. IEL stands for intraepithelial lymphocytes.

https://doi.org/10.1038/s41592-022-01651-8



Extended Data Fig. 10 | Multicellular structures discovered bt STELLAR on MERFISH mouse cortex data. Clusters in STELLAR's embedding space identify multicellular structures in tissues in MERFISH data from mouse cortex. (a) Heatmap of average cell-type composition in STELLAR clustered embeddings. (b) Representative tissue image colored by overall structure. L, lateral; OPC, oligodendrocyte precursor cell; PVM, perivascular macrophage; SMC, smooth muscle cell; VLMC, vascular leptomeningeal cell.

nature research

Corresponding author(s): Jure Leskovec

Last updated by author(s): 06/23/2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We used code for processing the CODEX multiplexed raw image files here https://github.com/nolanlab/CODEX as described in https://doi.org/10.1038/s41596-021-00556-8
Data analysis	~~Code available at https://github.com/snap-stanford/stellar~~

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets presented in this study can be found in the online repository Dryad. Specifically, the quantified single-cell data (cells-rows by protein expression-columns by x/y position-columns and cell type labels-columns) for all datasets are provided. Briefly expert annotated healthy human intestine (B004 training dryad.csv) was used for testing accuracy of STELLAR across the four regions of the colon regions of this dataset and also for training for transferring cell type labels to unlabeled donors (B0056 unnanotated dryad.csv). Additionally, the data used to transfer from the tonsil to BE tissue is there (BE Tonsil dryad.csv). For review purposes the data can be downloaded from https://datadryad.org/stash/share/1OQtxew0Unh3iAdP-ELew-ctwuPTBz6Oy8uuyxqliZk

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were set to maximize diversity across regions of the intestine from healthy donors and also depth of data acquisition.
Data exclusions	No data was excluded.
Replication	Each donor imaged was run on a different day.
Randomization	Randomization was not relevant for this study as there were not multiple groups requiring randomization.
Blinding	No blinding was performed in this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging