Inference of single cell profiles from histology stains with the Single-Cell omics from Histology Analysis Framework (SCHAF)

Charles Comiter[1,2,3,#], Eeshit Dhaval Vaishnav[5], Metamia Ciampricotti[6], Bo Li[1,13,14,17], Yiming Yang[1,13,17], Scott J. Rodig[7], Madison Turner[8,16], Kathleen L. Pfaff[8], Judit Jané-Valbuena[1], Michal Slyper[1], Julia Waldman[1], Sebastian Vigneau[1], Jingyi Wu[1], Timothy R. Blosser[9], Åsa Segerstolpe[1], Daniel Abravanel[1,15], Nikil Wagle[1,15], Xiaowei Zhuang[9,10], Charles M. Rudin[6], Johanna Klughammer[1,11], Orit Rozenblatt-Rosen[1,13], Koseki J. Kobayash-Kirschvink[1,12], Jian Shu[1,3], Aviv Regev[1,4,10,17,#]

[1]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[2]Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA 02142, USA
[3]Cutaneous Biology Research Center, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02129, USA
[4]Department of Biology, MIT, Cambridge, MA 02140, USA
[5]Sequome Inc., 10113 Berkshire Court, Cupertino, CA 95014, USA
[6]Department of Medicine, Thoracic Oncology Service, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[7]Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115, USA
[8]Center for Immuno-Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA
[9]Department of Chemistry and Chemical Biology, Department of Physics, Harvard University, Cambridge, MA 02138, USA
[10]Howard Hughes Medical Institute
[11]Gene Center and Department of Biochemistry, Ludwig-Maximilians-University, Munich, Germany
[12]Laser Biomedical Research Center, G. R. Harrison Spectroscopy Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[13]Center for Immunology and Inflammatory Diseases, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA.
[14]Department of Medicine, Harvard Medical School, Boston, MA, USA
[15]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA
[16]Department of Microbiology, Immunology, and Cancer Biology, University of Virginia, Charlottesville, VA 22908, USA
[17]Current address: Genentech, 1 DNA Way, South San Francisco, CA 94080, USA

#To whom correspondence should be addressed: ccomiter@broadinstitute.org (CC), aviv.regev.sc@gmail.com (AR)

Tissue biology involves an intricate balance between cell-intrinsic processes and interactions between cells organized in specific spatial patterns, which can be respectively captured by single-cell profiling methods, such as single-cell RNA-seq (scRNA-seq), and histology imaging data, such as Hematoxylin-and-Eosin (H&E) stains. While single-cell profiles provide rich molecular information, they can be challenging to collect routinely and do not have spatial resolution. Conversely, histological H&E assays have been a cornerstone of tissue pathology for decades, but do not directly report on molecular details, although the observed structure they capture arises from molecules and cells. Here, we leverage adversarial machine learning to develop SCHAF (Single-Cell omics from Histology Analysis Framework), to generate a tissue sample's spatially-resolved single-cell omics dataset from its H&E histology image. We demonstrate SCHAF on two types of human tumors—from lung and metastatic breast cancer—training with matched samples analyzed by both sc/snRNA-seq and by H&E staining. SCHAF generated appropriate single-cell profiles from histology images in test data, related them spatially, and compared well to ground-truth scRNA-Seq, expert pathologist annotations, or direct MERFISH measurements. SCHAF opens the way to next-generation H&E2.0 analyses and an integrated understanding of cell and tissue biology in health and disease.

INTRODUCTION

Advances in massively parallel, high-resolution molecular profiling now provide cellular and tissue level measurements at a genomic scale(Tang et al. 2019). These include methods for massively parallel single-cell or single-nucleus (sc/sn) profiling of RNA, chromatin, proteins, and their multi-modal combinations(Rood et al. 2022). Simultaneous advancements in data-driven analytics, largely in machine learning(Yan, Yoshua, and Geoffrey 2015), have allowed us to derive biological insights from such rich data(Regev et al. 2017; Rozenblatt-Rosen et al. 2017; Regev et al. 2018; Biancalani et al. 2020; Ding and Regev 2019), as well as from key modalities used in both research and routine clinical practice, especially H&E stains(van der Laak, Litjens, and Ciompi 2021).

However, substantial challenges remain in realizing the promise of these methods in the context of tissue biology, especially for histopathology, a cornerstone of medicine. While the costs and complexity of single-cell omics have reduced dramatically(Ziegenhain et al. 2017), they remain relatively expensive and time consuming and are not yet applied routinely in clinical settings. Experiments also remain prone to technical variations, leading to inter-sample discrepancies and batch effects(Lähnemann et al. 2020)[,(Luecken et al. 2020)]. Moreover, single-cell genomics does not directly capture spatial information nor is it directly related to the rich legacy of histology. While spatial transcriptomics methods, such as MERFISH(K. H. Chen et al. 2015), seqFISH+(Eng et al. 2019), Visium(Ståhl et al. 2016), ISS(Ke et al. 2013), Barista-seq(X. Chen et al. 2018), smFISH(Codeluppi, Borm, Zeisel, Manno, et al. 2018), osmFISH(Codeluppi, Borm, Zeisel, La Manno, et al. 2018), or Targeted ExSeq(Alon et al. 2020) measure spatially-resolved expression data, their throughput is still limited, and they involve high costs and complexities. Computational methods(Satija et al. 2015; Biancalani et al. 2021; Achim et al. 2015) have used

limited spatial signatures measured experimentally to project cell profiles to spatial positions, but they require some shared variables for such mapping (*i.e.*, genes measured by both modalities) and do not spatially resolve single-cell omics data based solely on the microscopic morphology in a histology image.

Recent advances in applied deep learning may open the way to address the challenge of mapping single-cell profiles to histology. In other domains, deep learning methods have successfully related data modalities of the same entities even when they do not have nominally shared variables (such as audio and video(H. Zhu et al. 2021; Kumar et al. 2022; Ngiam et al. 2011)). Moreover, recent studies showed that a model can be trained to generate a tissue's bulk RNA-seq profile from its histology image(Schmauch et al. 2020). Given these successes, we hypothesized that deep learning could also be applied to the more difficult modality transfer problem inferring single-cell expression profiles from histology.

Here, we present the Single-Cell omics from Histology Analysis Framework (SCHAF), an adversarial deep learning-based framework for transferring between histology images and single-cell omics data. SCHAF is based on the assumption that a histology image and a single-cell omics dataset from the same tissue sample can be explained by a single underlying latent distribution. Given a corpus of tissue data, where each sample has both a histology image and sc/snRNA-seq data, SCHAF discovers this common latent space from both modalities across different samples. SCHAF then leverages this latent space to construct an inference engine mapping a histology image to its corresponding (model-generated) single-cell profiles. This inferred dataset has genomic coverage, yielding expression information on all the genes in the training corpus, and is spatially resolved, with each predicted cellular expression profile mapped

to a radius of several dozen microns. Given the spatial information inherent in its predictions, SCHAF further produces a spatial portrait of a tissue's cell types with the help of an auxiliary annotator mapping from expression profiles to cell types. We demonstrate SCHAF's success in these tasks on data from two human tumor types, through multiple criteria, including validation by comparison to expert pathologist annotation and experimentally measured MERFISH data. SCHAF thus provides an important new tool for studies of tissue biology and disease.

RESULTS

SCHAF: Single-Cell omics from Histology Analysis Framework

SCHAF is trained to infer a tissue sample's single-cell omics dataset from its corresponding histological image by using a corpus of training samples with a matched histology image and a single-cell omics dataset from each training sample. The resulting predicted single-cell dataset is spatially resolved, such that it can be integrated with the input (test) histology image to create a single spatial, genomic scale, morphologically informative tissue data modality.

Briefly, SCHAF consists of the following steps (**Figure 1**): First, if necessary, it reduces domain discrepancies between histology images via normalization techniques (**Methods**). It then decomposes each histology image into many, partially overlapping, smaller square tiles, each ~950μm^2 (~30X30μm; smaller than the area of a spot in the Visium spatial transcriptomics method) and treated as representing one cell in its close neighborhood (without explicit cell segmentation). Next, it uses adversarial machine learning to build a model that translates one histology image tile into one cell's profile. Finally, it generates a tissue's full single-cell dataset by using this model to move each (partly overlapping) tile to a profile.

SCHAF infers single-cell profiles from tiles in a full image

The first challenge for SCHAF is that single-cell profiling lacks an explicit spatial structure and – even if some spatial structure were available – state-of-the-art deep learning algorithms for image translation (*e.g.*, Pix2Pix(Isola et al. 2017) or CycleGAN(J.-Y. Zhu et al. 2017)) would exhaust standard resources given the high-dimensionality of omics data and the potentially large size of a histology image.

To address these challenges, SCHAF reduces the problem of predicting a collection of single-cell profiles from a single macro-scalar image to the problem of predicting a collection of single-cell profiles from a collection of smaller image "tiles", each representing one cell and its close surroundings (**Figure 1**). SCHAF splits the full histology image into tiles, by sliding a small window over the whole image and including any tile not composed of mostly whitespace in its final collection (**Methods**). For a test image, SCHAF predicts one single-cell's expression profile from each tile (**Methods**). The tile size (roughly equal to half a dozen cells), is a hyperparameter in the model, which we select by optimization (**Methods**).

SCHAF learns histology-to-single-cell-dataset generation

SCHAF can effectively train an inference engine from a histology tile to a single-cell's expression profile with domain discrepancy-reduced data. Specifically, if necessary, SCHAF first uses off-the-shelf histology image normalization("{{ngMeta['og:Title']}}" n.d.) to move all histology images to a single domain (**Methods**). It splits each histology image into smaller tiles and trains one reconstructing autoencoder on all samples' single-cell expression profiles, and another convolutional adversarial autoencoder on all samples' image tiles to both reconstruct the

tiles and encode them to a latent space indistinguishable from that of the gene-expression reconstructing autoencoder. With these trained autoencoders, to predict a cell's expression profile from a tile, SCHAF encodes the tile with encoder of the histology image tile autoencoder and then decodes the encodings with the decoder of the gene expression autoencoder (**Figure 1, Methods**). Thus, every generated expression profile is also spatially positioned (to a tile) by definition.

Once SCHAF is trained, we construct an end-to-end process to generate a single-cell expression profiling dataset from a histology image (**Figure 1, Methods**). We normalize the histology image to the same domain as the images used in training, split the normalized histology image into tiles, and generate for each tile a single-cell's expression profile using the histology tile-to-expression-profile inference mechanism.

A framework to assess SCHAF's inference of single-cell profiles from histology in tumors

To demonstrate SCHAF, we applied it to three corpora of human tumor tissue, each consisting of matched single-cell or single-nucleus RNA-seq (sc/snRNA-Seq) and H&E histology we performed on adjacent portions of each specimen (**Figure 1**): a small cell lung cancer (SCLC) corpus with snRNA-seq data (sn-SCLC), a metastatic breast cancer (MBC) corpus with snRNA-seq data (sn-MBC), and an MBC corpus with scRNA-seq data (sc-MBC). The sn-SCLC corpus spanned 24 tumors (19 training, 5 for evaluation; MC, BL, YY, MS, ORR, AR and CR, unpublished data), the sn-MBC corpus had 8 tumors (6 for training, 2 for evaluation; DA, JJV, MS, JWu, SV, JWa, TRB, AS, DA, NW, XZ, JK, ORR, AR, NW), and the sc-MBC corpus had 7 tumors (5 for training, 2 for evaluation; DA, JJV, MS, JWu, SV, JWa, TRB, AS, DA, NW, XZ, JK, ORR, AR, NW) (**Supplementary Table 1**). The three corpora represent different tumor

types and profiling techniques. We demonstrated the results on two evaluation samples from sn-SCLC (MSKCC-16 and MSKCC-24), one from sn-MBC (HTAPP-6760), and one from sc-MBC (HTAPP-932). We analyzed 6,244 highly variable genes in the sn-SCLC corpus and 11,474 genes from the MBC corpora. For MBC, we also performed MERFISH on an immediately adjacent section of the same specimen to obtain spatial RNA data as well as collected expert pathologist annotations, allowing us to further assess the quality of spatial predictions, as we present below.

We evaluated the quality of SCHAF's inference of single-cell profiles for each dataset by four criteria, relative to a baseline. To establish a performance baseline, for each of the three corpora, we trained a variational autoencoder (VAE(Kingma and Welling 2013)) on the training sc/snRNA-seq dataset and compared each dataset inferred by SCHAF to an inferred "baseline" dataset generated from the corpus's VAE via random-sampling from a Gaussian latent space. We considered performance in correctly predicting the real test data based on: (**1**) pseudo-bulk expression profile proportions (the sum of all cells' expression profiles, normalized to unit sum) for the entire sample and per cell type; (**2**) gene-gene correlation matrices of the top 100 most highly variable genes across the entire dataset; (**3**) distribution of each gene's counts across the cells; and (**4**) annotated cell types and their organization in a low dimensional embedding.

SCHAF accurately predicts pseudo-bulk expression profiles

In all corpora, SCHAF's predicted pseudo-bulk expression proportions profiles (pseudo-bulk expression profiles normalized to unit sum; **Methods**) agreed closely with the corresponding real data (**Figure 2a**), with higher gene-correlations (**Figure 2b,** Pearson's $r > 0.67$ for both sn-SCLC samples, Pearson's $r > 0.87$ for both the sn-MBC and sc-MBC sample) than those of the

VAE-sampled baseline (Fisher-Steiger improvement-of-correlation test(Howell 2012) *p*-values of ~0 for all four samples, Pearson's r <0.55 for both sn-SCLC samples' baselines, Pearson's r <0.8 and <0.75 for the sc-MBC and sn-MBC samples' baselines, respectively), as well as far lower Jensen-Shannon divergence than those of the VAE-sampled baseline (**Figure 2c**). This is the case, despite the different cell numbers in the inferred, VAE-sampled baseline, and real datasets, and even given the relatively high correlation for the baseline. Note that due to the sparsity of sc/snRNA-seq, SCHAF tended to under-predict the expression of lowly expressed genes, and somewhat over-predict that of more highly-expressed genes (**Figure 2a**).

Gene-gene correlations in SCHAF-predicted and real data are well aligned

There was also good agreement in the gene-gene correlation across single cells between SCHAF-generated profiles and real scRNA-seq. First, the gene-gene correlation matrix for the top-100 highly variable genes in each sample shows similar patterns in real and SCHAF-generated data for each sample (albeit with lower signal for SCHAF), but not in the VAE-sampled baseline, which yielded negligible if any gene-gene correlations (**Figure 2d**). Notably, the SCHAF-inferred datasets generally exaggerated the stronger correlations, while capturing more subtle correlations less well (**Figure 2d**). (Note that when a gene did not have a predicted expression in any cell, it would also not yield a self-correlation, leading to a missing diagonal entry; **Figure 2d**). Moreover, when calculating the correlation coefficient between corresponding top-100 highly variable genes' entries of the two flattened gene-gene correlation matrices, SCHAF showed much higher agreement with real data (**Figure 2e**, Pearson's r=0.40-0.49 in sc-SCLC; 0.46-0.74 in sn-MBC and sc-MBC) than the VAE-sampled baseline (Pearson's r < 0.13 in sn-SCLC; r=0.18-0.28 in sn-MBC and sc-MBC), yielding a significant

improvement (*p*-values = ~0.0 Fisher-Steiger improvement-of-correlation test for all four samples).

Single-cell distributions of gene expression agree in SCHAF-predicted and real scRNA-seq

Next, we found that SCHAF's predicted datasets broadly preserved the distribution of genes' expression levels (count values) across single cells. For each gene, we calculated the Earth Mover's Distance between its distribution in SCHAF- or baseline-inferred data and its distribution in the original target dataset. The SCHAF-inferred data showed a higher similarity to the real data than the VAE-sampled baseline, with lower EMDs in three of the tested samples (*p*-values < 10^{-12}, Mann-Whitney U test), but not the fourth (p>0.08; sn-MBC HTAPP-6760). Furthermore, all EMDs between SCHAF-predicted gene expression distributions and real ones were less than 11, far below the maximum possible of 50.

SCHAF inference preserves cell type annotations and clusters

To compare how well SCHAF-generated profiles preserve cell types and their clusters, for each sample, we first used the original scRNA-seq data to train a neural network to classify a cell type from one sc/snRNA-Seq profile, then used this classifier to assign a cell type label to each real, SCHAF-inferred, or VAE-sampled baseline profile, and finally examined how those labels organize when we project the SCHAF-predicted, VAE-sampled baseline, or real single-cell profiles in low dimensions using UMAP(McInnes, Healy, and Melville 2018) and color each by its assigned cell type label (**Figure 3**).

The cell type assignment probabilities of profiles to an existing annotation by the classifier were significantly higher for SCHAF-inferred data than for the VAE-sampled baseline ($p << 10^{-24}$ in

each tested sample; Mann-Whitney U test, **Figure 3d**). Furthermore, for each sample, within each cell type, the pseudo-bulk expression proportions profile of cells assigned to a cell type agreed quite closely with that of real cell profiles of that cell type (Pearson's $r > 0.63$ in all cell types in both sn-SCLC samples, Pearson's $r > 0.76$ in all cell types in both MBC samples, **Figure 3c**), and exceeded that of the VAE-sampled baseline in nearly every case. Similarly, the 'meta gene-gene correlations' of SCHAF-inferred data per cell type, while relatively low, were substantially better than over the baseline in all cell types and samples (**Supplementary Figure 1**).

Strikingly, in all test samples across all corpora, the SCHAF-predicted datasets preserved cell type clusters (**Figure 3a**, middle) as more distinct and well-formed in a low dimensionality embedding than the VAE-sampled baseline (**Figure 3a**, bottom), and often in roughly comparable proportions to the real profiles (**Figure 3a**, middle) (Jensen-Shannon divergences < 0.125 for all samples). While these were not as well-separated as in the real datasets, they far exceeded the patterns in the VAE-sampled baseline (**Figure 3a**, bottom). To quantify the quality of these clusters, we compared the distributions of cells' cell type silhouette coefficients between each sample's SCHAF-inferred, VAE-sampled baseline, and original (measured) sc/snRNA-seq profiles. In all samples, the distributions of the SCHAF-inferred profiles were significantly higher than those of the corresponding VAE-sampled baselines and quite close to those in the real data, supporting our qualitative conclusions from the embeddings (**Figure 3b**, *p*-values close to 0, Mann-Whitney U tests comparing the inferred and random silhouette coefficient distributions).

SCHAF infers accurate spatial gene expression at the tile level and agrees with regional annotations

To evaluate SCHAF's performance in generating expression profiles at correct spatial positions (tiles), we focused on the sn-MBC and sc-MBC samples, where we performed multiplex FISH measurements of 212 genes on a consecutive section measured by MERFISH(K. H. Chen et al. 2015), as well as regional annotations on the H&E section from expert pathologists (SJR, MT, and KLP). We devised two measures to assess the quality of SCHAF's spatial inference: (**1**) the cross-tile (within sample) correlation between SCHAF-predicted expression of a given gene in each tile, and its MERFISH measured expression in the tile; and (**2**) the agreement in high-level cell type assignments (cancer, normal, fibrosis, immune, vasculature) between SCHAF-inferred and pathologist annotations.

To compare SCHAF-inferred spatial expression to MERFISH measurements, we "tiled" the aligned H&E (with SCHAF-inference) and MERFISH images (performed on a consecutive section) into multiple smaller tiles, each 1:1 between the two sections, and calculated, for each of the 212 genes measured by both scRNA-seq and MERFISH, the correlation coefficient between the mean SCHAF-expression and MERFISH-expression per tile across all the tiles. As null benchmarks, we randomly assigned spatial locations to either expression values from the VAE-sampled baseline (as above) or from the real scRNA-seq.

Overall, the tile-level spatial correlation to MERFISH measurements was higher for SCHAF-inferred data compared to the null baseline (**Figure 4b**). Out of 152 sn-MBC genes and 183 sc-MBC genes that were sufficiently expressed in at least 40 tiles and are in the 212-gene MERFISH panel, 33 and 34 genes, respectively, had substantial correlations between

SCHAF-inferred and direct MERFISH data (Pearson's $r > 0.3$, **Figure 4a,b, Supplementary Tables 2 and 3**), but there was virtually no correlation ($-0.1 < r < 0.1$) in the two baseline cases (p-value < 0.05, Fisher-Steiger improvement-of-correlation test in all cases, **Figure 4c**).

We also compared regional annotations by pathologists (tumor, normal, immune, fibrosis, vasculature; **Figure 4d-middle, e-top-right)**) to annotations derived by mapping the cell type assignments of single-cell profiles (as above) to spatial locations of corresponding samples' histology images (**Figure 4d-right, e-bottom, Supplementary Figs. 2 and 3**). We assigned all the pixels of each tile one color, according to the tile's assigned cell type category, creating an inferred spatial annotation for both the sc-MBC sample and the sn-MBC sample with either broad (**Figure 4d-right, e-bottom, Supplementary Figs. 2 and 3**) or refined (**Figure 4d-right, e-bottom, Supplementary Figs. 2 and 3**) cell type annotations.

We found good agreement between SCHAF-based and pathologist annotations distinguishing cancerous and non-cancer regions (sample HTAPP-932, **Figure 4e**) and for the multiple detailed annotations available for one sn-MBC sample (HTAPP-6760, **Figure 4d**). In this latter case, the pathology "Vasculature" annotation aligned with endothelial vascular and smooth muscle vascular portions in the SCHAF annotation maps, "Fibrosis" with SCHAF's fibroblasts, "Immune Cells" with T cells and macrophages, and "Tumor" with parts of the SCHAF mapping consisting almost entirely of malignant cells (**Figure 4d**). In contrast, the baseline annotations presented an uninformative picture: a spatially uniform distribution dominated by malignant cells with arbitrary "salt-and-pepper" scatter of other cell types (**Figure 4d**). Note, that in some cases SCHAF has been challenged with predicting a cell class, and hence its location. For example, SCHAF under-predicted fibroblasts in HTAPP-932 (**Figure 3a**), and correspondingly, the

fibrosis-annotated region in this tumor (**Figure 4d**). Nevertheless, overall, we show evidence for the success of SCHAF's spatial mapping.

DISCUSSION

Here, we presented SCHAF as a computational framework to predict spatial single-cell profiles from H&E images without the need for any molecular spatial data in the training. *In-silico* tissue-data modality transfer between histological imaging and scRNA-Seq domains would significantly increase the accessibility of spatially-resolved molecular profiles, not just in terms of time, effort, and expense, but also in the ability to predict molecular information for the vast number of clinical samples archived over decades, and to benefit from the insights provided by high-resolution methods in low-resource settings. To achieve this, SCHAF learns to use a tissue's histology image as a specification for inferring an associated single-cell expression dataset. More broadly, this can be seen as a biological application of a framework for inferring an entire dataset as opposed to a single output.

SCHAF is based on adversarial machine learning, with latent space representations of data from two modalities. This approach performed quite well even though SCHAF does not account for differences in cell sizes via segmentation, instead considering histology "tiles", each intended to encompass roughly one cell and its close surroundings, such that all parts of the image are captured in some tile. While it may be intuitive to segment cells first, and then match them to profiles, we reasoned that a cell's neighborhood holds important information on its intrinsic profile, as we have recently showed(Jerby-Arnon and Regev 2022). SCHAF's success suggests that *in-silico* data generation is not only a viable path to augment laboratory measurements, but also that biological tissue data can be successfully represented by a modality-agnostic,

lower-dimensional latent space, which can be exploited by adversarial machine learning techniques for a multitude of applications in data integration and modality transfer, as well as probed in the future for understanding the principles of tissue organization, including how molecular information leads to tissue structures and vice versa.

SCHAF was successful on corpora of matched H&E and sc/snRNA-seq data from two tumor types in three separate cases. As tumors are more variable and less canonical than healthy tissue, predicting a new tumor's profile, even of the same type, is a challenging task. Future studies are required to assess the ability to train a single model across a single corpus of multiple types of tissues, tumors, or technologies (*e.g.*, snRNA-seq and scRNA-seq or snRNA-Seq and scATAC-Seq). Doing so would require additional datasets with matched single-cell profiling and histology imaging, which are still surprisingly scarce, but are growing thanks to efforts such as the Human Tumor Atlas Network(Rozenblatt-Rosen et al. 2020) and the Human Cell Atlas(Regev et al. 2017),(Rood et al. 2022). These initiatives also generate additional spatial genomics data, which we did not use for training SCHAF, but leveraged to validate its inferred patterns to gain confidence in its performance and introduce further improvements, for example, to enhance predictions at the individual gene level.

SCHAF's initial success – and remaining limitations – should motivate several further research directions, including improving the initial model through more sophisticated machine learning or cell segmentation. Such efforts are important given the value of high-resolution single cell and spatial genomics data in understanding tumor biology and monitoring the impact of therapy, but the substantial costs to patients, caregivers and researchers required to collect such specimens and data. In addition, inference should also be possible in the reverse direction: constructing

histological images from single-cell expression data, which can help with understanding tissue biology. Given the high dimensionality of single-cell profiles and the many ways histology images could manifest, this could be seen as designing a mechanism for the generation of imaging data conditioned on a very high-dimensional specification. Constructing such a gene-expression-to-histology mechanism presents several challenges, including choosing an appropriate size of the output histology data, the need to infer many different potential histological configurations from the same input single-cell dataset, designing a generative model to be conditioned on a variable as high-dimensional as a transcriptome, departing from more traditional models (such as AC-GAN(Odena, Olah, and Shlens 2017)), and validating the predicted images, which, especially in tumors, would likely require matching of features rather than direct pixel-to-pixel alignment.

The underlying principles defining SCHAF can be extended to other cases in biology, both with similar data modalities (*e.g.*, cell profiles and microscopy images measured in cultured cells(Kobayashi-Kirschvink et al. 2022)), in different biological modalities (3D imaging(Wang et al. 2018) or temporal tracing(Schofield et al. 2018)), and non-biological settings. Finally, and more fundamentally, interpreting SCHAF's model could have implications for our understanding of tissue biology, by helping us understand which cellular and gene programs and configurations relate to which tissue features.

Figure Legends

Figure 1. SCHAF learns to predict a tissue's single-cell omics dataset from its histology image.

Schematic overview of SCHAF. From left: Training data, consisting of matched histology (H&E) stains and scRNA-seq are obtained from multiple patients with a particular type ("tumor samples"). Histology images are normalized (as needed) and tiled ("Input"), and then used to train a histology-image-tile autoencoder ("Model: SCHAF"; middle top) to reconstruct and encode to a latent space, which is indistinguishable from that of a gene-expression encoder adversarially trained against a latent discriminator ("Model: SCHAF", middle bottom). In inference, SCHAF takes an input histology image, normalizes it to the same domain as the images used in training, tiles it, and passes it through the image tile encoder followed by the gene-expression decoder to generate single cell profiles in spatial locations for the specific tumor-type on which it was trained ("Output").

Figure 2. SCHAF-inferred scRNA-seq profiles are similar to measured scRNA-seq and exceed a VAE-sampled baseline

a-c. SCHAF bulk gene-expression profiles agree better with measured scRNA-seq. **a.** Bulk gene-expression proportions (y axis) for each gene (dot) rank ordered by expression level (x axis) in real data (blue) from SCHAF-inferred (orange, top) or VAE-sampled baseline (orange, bottom) *vs*. measured (blue) scRNA-seq in each of the four test samples. **b-c.** Pearson's correlation coefficient (b, y axis) or Jensen-Shannon divergence (c, y axis) between the original (measured) scRNA-seq and SCHAF-inferred (orange) or VAE-sampled baseline (blue) expression profiles in each of the four test samples (x axis). **d-e.** SCHAF preserves real gene-gene correlations. **d.** Pearson's correlation coefficient (color bar) between each pair of the top 100 most highly-variable genes (rows, columns) in the original (measured) (top), SCHAF-inferred (middle), and VAE-sampled baseline (bottom), with genes ordered in all cases based on their clustering in the original data. **e.** Meta-correlations (y axis, Pearson's correlation coefficient between two correlation matrices' flattened entries corresponding to the 100 most highly-variable genes) between the original (measured) data and SCHAF-inferred (orange) or VAE-sampled baseline (blue).

Figure 3. SCHAF-inferred profiles preserve cell type subsets

a,b. SCHAF-inferred profiles preserve better-defined cell type subsets. **a.** UMAP embedding of cell profiles from the original scRNA-Seq (top), SCHAF-inferred (middle) or VAE-sampled baseline (bottom) for each test sample (column), colored by cell type labels from manual-annotation of real data (top) or corresponding classifier-assigned labels (middle and bottom) (color legends). **b.** Distributions of cell type silhouette coefficients (y axis) for the original scRNA-Seq, SCHAF-inferred, or VAE-sampled baseline (x axis). P-value of Mann-Whitney U test is noted on top. Light-orange line: median; black rectangle: 25^{th} to 75^{th} percentiles; whiskers: all non-outlier data; circles above and below whiskers: extreme outliers. **c.** SCHAF-inferred profiles better preserve cell type classes. Pearson's correlation coefficient (y axis) between the pseudobulk profiles in real scRNA-seq and in SCHAF-inferred (orange) or VAE-sampled baseline (blue) for profiles classified for each cell type (x axis). **d.** Distribution of the cell type classifier's assignment probabilities on SCHAF-inferred (orange) or VAE-sampled baseline (blue) profiles.

Figure 4. Accuracy of SCHAF's spatial expression inferences compared to MERFISH measurements and pathologist annotations.

a-c. SCHAF inferences are more consistent with MERFISH measurements. **a.** SCHAF-inferred (y axis) and MERFISH-measured (x axis) expression levels in each tile (dot) for each of six genes for sample HTAPP-932 (three left columns) or sample HTAPP-6760 (three right columns). The number of tiles considered for each gene is denoted. Dashed line: least-mean-square-error fit. **b.** Distribution of spatial-tile correlation coefficient (x axis) between the MERFISH measurements and either SCHAF-inferred (orange), a baseline of randomly positioned, VAE-generated profiles (blue) or a baseline of randomly positioned scRNA-Seq profiles (green) from each test MBC sample, for genes with both MERFISH and SCHAF prediction. **c.** Spatial tile correlation coefficient (y axis) for each of the genes in (a) (x axis) between the MERFISH measurements and either SCHAF-inferred (orange), a baseline of randomly positioned, VAE-generated profiles (blue) or a baseline of randomly-positioned scRNA-Seq profiles (green) from each test MBC sample. **d-e.** SCHAF inferences are more consistent with pathologist's annotations. H&E stain from samples HTAPP-6760 (d) and HTAPP-932 (e) colored by pathologist annotations (d-center, e-top-right) and by annotation derived based on the classified cell type of a SCHAF-inferred profile at each tile, for each annotation separately (d-left, e-top-left) and combined (d-right, e-bottom, top and middle: coarse and fine grained labels), or by classified cell types of randomly-positioned scRNA-seq profiles (d-right-bottom, e-left-bottom).

Supplementary Figure Legends

Supplementary Figure 1. Cell type specific gene-gene correlations

"Meta"-correlation coefficient (y axis; correlations between two correlation matrices' flattened entries corresponding to the gene-gene correlations between the 100 most highly-variable genes), between the original (measured) scRNA-seq and SCHAF-inferred (orange) or VAE-sampled -baseline (blue) for each cell type (x axis) in each of the four test samples (panels).

Supplementary Figure 2. Broad cell type spatial annotations of MBC sample HTAPP-932.

H&E stain of sample HTAPP-932 with tiles labeled by broad cell type annotations derived based on the classified cell type of a SCHAF-inferred profile at each tile (**a**), a baseline of randomly-positioned scRNA-Seq profiles (**b**), or a baseline of randomly positioned, VAE-generated profiles (**c**).

Supplementary Figure 3. Broad cell type spatial annotations of MBC sample HTAPP-6760.

H&E stain of sample HTAPP-6760 with tiles labeled by cell type annotations derived based on the classified cell type of a SCHAF-inferred profile at each tile (**a**), a baseline of randomly-positioned scRNA-Seq profiles (**b**), or a baseline of randomly positioned, VAE-generated profiles (**c**).

Supplementary Table Legends

Supplementary Table 1. Sample Datasets

Number of cells and genes in each patient sample's sc/sn-RNA-seq dataset.

Supplementary Table 2. SCHAF and MERFISH comparison for sample HTAPP-932

Number of tiles in and correlation between SCHAF and MERFISH expression levels for each gene occurring in both the MERFISH and scRNA-seq data for sample HTAPP-932.

Supplementary Table 3. SCHAF and MERFISH comparison for sample HTAPP-6760

Number of tiles in and correlation between SCHAF and MERFISH expression levels for each gene occurring in both the MERFISH and snRNA-seq data for sample HTAPP-6760.

METHODS

Human subjects

All MBC samples used in this study underwent IRB review and approval at Dana Farber Cancer Institute (DFCI) (protocol 05-246) as well as at the Broad Institute (protocol #15-370B). All SCLS samples were from patients undergoing a surgical resection or tissue biopsy at Memorial Sloan Kettering Cancer Center (MSKCC) identified and biospecimens collected prospectively from 2005 to 2013 (MSKCC protocol #06-107 A(13)). All patients from whom biospecimens were obtained provided informed consent through the corresponding Institutional Review Board-approved biospecimen collection and analysis protocol.

Matched scRNA-seq and H&E images data

Three corpora of tumors were used. A corpus of twenty-four (24) small cell lung cancer (SCLC) samples profiled by snRNA-seq and matching H&E stains was obtained from an unpublished study from Memorial Sloan Kettering Cancer Center (MSKCC). Two corpora of eight (8) and seven (7) metastatic breast cancer (MBC) tissue samples, with snRNA-seq and scRNA-seq data, respectively, were obtained from the Human Tumor Atlas Project Pilot (HTAPP). In the MSKCC corpus, the snRNA-seq data came from a portion of the same slice used in H&E staining. In the HTAPP corpora, the H&E and sc/sn-RNA-seq data of each patient sample came from different biopsies of the same metastasis. The histology images were of resolution 1 μm/pixel in the HTAPP corpora and 0.5 μm/pixel in the MSKCC corpus. Further information about these datasets' dimensionalities is in **Supplementary Table 1**. Cell types in the original datasets were manually annotated.

H&E data pre-processing

To prepare H&E stains for further analysis, all pixels that were sufficiently light (having a channel with pixel value > 200 of a maximum of 255), indicating non-tissue, were replaced with pure white pixels.

Sc/snRNA-seq data pre-processing

Cell profiles with < 200 detected genes were removed, followed by removing genes expressed in < 3 cells. Each cell's counts were normalized to a sum of 10,000, followed by *log1p* (*f(x) = log(x + 1)*) transformation of each normalized count value, *x*. Each cell's counts were then normalized to sum to 1, and each entry was divided by 10, a constant larger than any log-normalized count in any cell in any sample (over all corpora). Lastly, within each of the two tissue types (SCLC, MBC), all genes shared across all samples of the tissue type were identified. Then, the 1,024 most highly variable of these shared genes in each sample were found. Lastly, the union of all these genes (6,244 genes in sn-SCLC and 11,474 genes in both of the MBC corpora) was retained for further analysis. Quality-control steps were performed with Python's *scanpy* package(Wolf, Angerer, and Theis 2018) with default parameters except for the following: *pp.filter_cells* with *min_genes*=200*, pp.filter_genes* with *min_cells*=3, *pp.normalize_total* with *target_sum*=10,000, *sc.pp.log1p* and *pp.highly_variable_genes* with *n_top_genes*=1024.

Discrepancy correction in H&E stains

To resolve domain discrepancies between the different H&E stain samples, samples MSKCC-10 (for SCLC) and HTAPP-6760 (for MBC) were used as references for the SCLC and both MBC datasets, respectively. The *color_normalization.reinhard* function from the *histomicstk* package (https://digitalslidearchive.github.io/HistomicsTK/index.html) was used with default settings to normalize all histology images to occupy the same domain as the reference sample's image.

Histology image tiling

After resolving domain discrepancies, histology images were tiled into smaller, potentially-overlapping tiles. A square window of size *tile_size* X *tile_size* was slid over each histology image, starting at the top-left corner of each image, and moving the window *tile_move_dist* units at a time, first horizontally, then vertically once a row's tiles are exhausted, until each possible position of the window in each image is covered. *tile_size* = 31 μm and *tile_move_dist* = 16 μm were selected as those that yielded the best results (the strongest bulk correlations similar to those in **Figure 2b**) over a wide range of hyperparameters. All tiles where most pixels did not consist of whitespace were included in the image's set of tiles.

Learning to infer a single-cell-omics dataset from a histology image

All machine learning models were built and trained using the Python deep learning library *pytorch(Paszke et al. 2019)*.

Given a training dataset of tissue samples, each with a corresponding tiled H&E image and sc/snRNA-seq data, we trained SCHAF to infer an entire sc/snRNA-seq dataset from a histology image. If necessary, all training histology images were first integrated into a common discrepancy-free domain via image normalization as described above, and the histology images were tiled (as described above).

Next, a model was trained to learn a single-cell's expression profile from a single histology-image tile, using an adversarial autoencoder(Makhzani et al. 2015) based framework(Yang and Uhler 2019) for domain translation applied to image-tile and gene-expression domains. First, a standard autoencoder, *G*, was trained on all the expression profiles of all training samples, according to a mean-square-error (MSE) loss, optimized via stochastic gradient descent(Kingma and Ba, n.d.), for 50 epochs, with a batch size of 32, at a learning rate of 0.00005. Let *L* be the latent space of such an autoencoder. A convolutional adversarial autoencoder *T* was trained on all image-tiles of all training samples, to simultaneously reconstruct image tiles and encode to a latent space indistinguishable from *L* via an adversarial training regime of the tile network against an adversarial discriminator. The adversarial discriminator was trained according to a binary-cross-entropy loss with-logits (BCE), optimized via stochastic gradient descent(Kingma and Ba, n.d.) for 25 epochs, with a batch size of 32, at a learning rate of 0.004. *L* was trained according to a regularized mean-square-error loss, given by the loss function $f(i', i, z) = MSE(i', i) + beta*BCE(z, t)$, where *i* is the image input, *i'* is a reconstruction of *i* by *L*, *t* are labels signifying that a latent space comes from the gene-expression latent space, and *z is* the output of the adversarial discriminator an encoding of *i* by *L*. *L* was trained with *beta*=0.001, optimized via stochastic gradient descent(Kingma and Ba, n.d.), for 25 epochs, with a batch size of 32, at a learning rate of 0.001. The adversarial discriminator and *L* were trained in adversarial fashion with alternating gradient updates.

G and *T* both had an encoder-decoder architecture, with the encoder composed of four sequential [Linear, BatchNorm, ReLU] blocks, followed by one [Linear, BatchNorm] block, and the decoder composed of four sequential [Linear, BatchNorm, ReLU] blocks, followed by one

[Linear, ReLU] block. The dimensions of the five linear layers of the encoder were, in order, *number_of_genes*, 1024, 1024, 1024, and 128 neurons. The dimensions of the five linear layers of the decoder were, in order, 128, 1024, 1024, 1024, and *number_of_genes* neurons, where *number_of_genes* was the number of highly variable genes (HVG) in each corpus being retained for analysis. The adversarial discriminator had an architecture consisting of four [Linear, SpectralNorm, ReLU] blocks, followed by one [Linear, SpectralNorm] block. The dimensions of the five linear layers were, in order, 128, 64, 32, 32, and 2 neurons. The Linear, BatchNorm, ReLU, and SpectralNorm layers were implemented via pytorch's *nn.Linear, nn.BatchNorm1D, nn.ReLU,* and *nn.utils.spectral_norm* functions, respectively.

To translate from image-tile to a single-cell expression profile, after training *G* and *T*, the tile was encoded with *T*'s encoder and the encoding was decoded with *G*'s decoder.

For a full inference pipeline from H&E-stain to single-cell dataset, the input histology image was first normalized to the same domain as the integrated images used in training using the *histimoicstk* package, the normalized image was tiled with the same parameters used for training images, and a single-cell expression profile was inferred for each tile.

Test datasets for evaluation of non-spatial single-cell profiles predicted by SCHAF

Four evaluation samples were withheld for testing: two from the sn-SCLC corpus (MSKCC-16 and MSKCC-24) and two from the MBC corpus (HTAPP-6760 (sn-MBC) and HTAPP-932 (sc-MBC)). Each test sample was evaluated based on four criteria (below) in comparison to a random (VAE-sampled) baseline (as described below).

Generation of VAE-sampled baseline for evaluation

For each evaluation sample, the inferred dataset was compared to both the target (real) sc/snRNA-seq dataset as well as to a baseline dataset generated by sampling a variational autoencoder (VAE). First, a variational autoencoder (VAE) was trained on training sc/snRNA-Seq data from each corpus to both reconstruct expression profiles and encode to a Gaussian latent space. The VAE was trained according to a standard VAE loss $f(x', x, z) = MSE(x', x) + lambda*KL(z, z')$, where x is the data, x' its reconstruction by a VAE, z' is its encoding, z is a sample from a unit Gaussian, *MSE* is the mean-square-error, and *KL* is the KL divergence. The VAEs all had architectures identical to that of the autoencoder *G* above. After training, the VAE was used to generate a baseline dataset by sampling from the Gaussian latent space of the VAE for each cell in the original dataset and generating an expression profile. The VAEs were trained for 50 epochs, with a batch size of 32, optimized via stochastic gradient descent(Kingma and Ba, n.d.), at a learning rate of 0.00005, with *lambda* = 0.00000001.

Evaluation based on proportional pseudo-bulk expression profiles

To calculate pseudo-bulk profiles, in each of the original (real), SCHAF-inferred, or VAE-sampled datasets, all of the constituent cells' vectors were first summed to a single pseudo-bulk expression vector, which was then divided by the sum of counts over all genes to yield a (proportional) pseudo-bulk expression probability distribution for each dataset, and then ordered the genes in the vector by increasing proportion in the original (real) dataset (for visualization purposes). The *scipy.stats.pearsonr* and *scipy.spatial.distance.jensenshannon* functions were used to calculate the Pearson correlation coefficient and the Jensen-Shannon divergence between these vectors. To test the significance of improvement in the Pearson

correlation coefficient between the real data and the VAE sampled baseline *vs*. SCHAF, a Fisher-Steiger improvement-of-correlation test(Howell 2012) was used as implemented in https://github.com/psinger/CorrelationStats.

Evaluation based on gene-gene correlations

To compare the gene-gene relationships between the original (real), SCHAF-inferred, or VAE-sampled baseline, the Pearson correlation coefficient was computed between each pair of genes across all cells in each dataset, and correlation matrices of the 100 most highly-variable genes in the original (real) dataset were plotted for each dataset using the *numpy.corrcoef* function (**Figure 3a,b**). The "meta-correlations" between the SCHAF-inferred or VAE-sampled baseline and the original (real) datasets matrices were calculated by flattening the portion of each matrix corresponding to these 100 highly-variable genes to a one-dimensional vector and calculating the Pearson correlation coefficient using the *scipy.stats.pearsonr* function.

Evaluation by gene count distributions

To compare the distributions of each gene's expression across cells, for each gene, the distribution of counts over a dataset's cells was considered. For each dataset and each gene, the distribution was found by first making fifty (50) evenly-spaced bins of counts per cell, calculating the number of cells in each bin, and dividing each by the total number of cells, to create a fifty-part discrete probability distribution. For each gene, such a distribution was calculated in the SCHAF-inferred, VAE-sampled baseline, or original (real) datasets. The Earth Mover's Distance (EMD; intuitively, the minimum amount of work needed to make one distribution identical to the other) was calculated between the distribution in the original (real) dataset and the SCHAF-inferred or VAE sampled baseline datasets with the *pyemd* package's

emd function. In our setting, such a metric can take on a value anywhere in the interval [0, 50]. The distributions of these EMDs were compared using a Mann-Whitney test, with the *scipy* package's *stats.mannwhitneyu* function.

Evaluation by cell type annotations, composition, and relationship in low dimensions

To compare cell type composition and clusters in the original (real), SCHAF-inferred, or VAE-sampled baseline datasets, a cell type classifier was first trained on labeled (cell type annotated) cell profiles for each original (real) dataset, and then applied to the SCHAF or random baseline inferred profiles to predict a cell type label for each profile. The classifier followed a linear neural network architecture, composed of three sequential [Linear, BatchNorm, ReLU] blocks, followed by one [Linear, SoftMax] block. The dimensions of the four linear layers were, in order, *num_genes*, 1024, 256, 64, and *num_cell_types* neurons, where *num_genes* was the number of hyper variable genes (HVG) retained for analysis in each corpus, and *num_cell_types* was the number of different cell types in the dataset. The Linear, BatchNorm, ReLU, and SoftMax layers were implemented via pytorch's *nn.Linear, nn.BatchNorm1D, nn.ReLU,* and *nn.Softmax* functions, respectively. The model was trained according to a standard cross-entropy loss. Next, for each sample, the distribution of probabilities assigned to each cell's assignment by the classifier in the SCHAF-inferred and VAE-sampled baseline were compared (**Figure 3d**) and tested for significance using a Mann-Whitney U test.

To assess the relationship between cell profiles in low dimensions, Uniform Manifold Approximation and Projections (UMAPs(McInnes, Healy, and Melville 2018)) were generated separately for the original (real), SCHAF-inferred, and VAE-sampled baseline datasets with the *scanpy(Wolf, Angerer, and Theis 2018)* package with default parameters. First, Principal

Components Analysis (PCA) was performed with the *pp.pca* function, retaining the top 50 PCs. Then, a *k*-nearest neighbor (*k*-NN) graph was computed (*k*=15) with the *pp.neighbors* function, and a UMAP was generated with the *tl.umap* function, and plotted in two dimensions using the *pl.umap* function, coloring by cell type annotations (for the real datasets) or classifier prediction (for SCHAF-inferred and VAE-sampled baseline) (**Figure 3a**).

To quantify the quality of cell type groups, the distributions of cell type silhouette scores were compared between the original (real), SCHAF-inferred, and VAE-sampled baseline datasets of each evaluation sample (**Figure 3b**), using the *sklearn* library's *metrics.silhouette_samples* function on the PCs computed above.

To examine cell type-specific pseudo-bulk expression profiles (**Figure 3c**), for each evaluation sample, within each cell type, the pseudo-bulk gene-expression vector was calculated, as described above, for cells annotated by that cell type in the original dataset, or assigned this annotation by the cell type classifier in SCHAF-inferred or VAE-sampled datasets. The Pearson correlation coefficient was calculated for each cell type between the pseudo-bulk expression vectors of the original data and the SCHAF-inferred or VAE-sampled datasets random baseline, using the *scipy.stats.pearsonr* function.

Evaluation of spatial gene expression compared to MERFISH data

For samples HTAPP-932 and HTAPP-6760, segmented MERFISH data were converted into an image by first normalizing each cell's MERFISH counts to have sum 1,000, applying *log1p* (*f(x) = log(x + 1)*) to each entry, and dividing each entry by 10, and then, starting from an empty image, filling in the 5-pixel-radius circle corresponding to an expression point's coordinates with

the measured expression. This image was then resized using the *imutis* library's *resize* function and manually aligned with the sample's corresponding H&E image (and thus also with the SCHAF-inferred expression).

These MERFISH-SCHAF images were then tiled for each sample using the same tiling approach as above, but with a larger tile size (127μmX127μm=16,129 μm^2) and stride size of 63μm to compensate for potential shortcomings in image registration. In both samples, for the 212 genes that are in both SCHAF predictions and MERFISH measurements, the Pearson correlation coefficient between a tile's mean SCHAF expression value and mean MERFISH validation value was calculated, after discarding tiles without expression in both the gene's SCHAF and MERFISH channels. Two baselines were considered, based on the Pearson correlation between MERFISH and randomly-placed profiles from either the corresponding VAE-sampled baseline dataset or the original (real) sc/snRNA-seq data.

Evaluation by spatial regional annotation compared to expert pathologist annotations

For samples HTAPP-932 and HTAPP-6760 (**Figure 4d,e**), spatial annotations were compared between predicted cell types assigned to each inferred cell by their respective classifiers (as above) and expert pathologist annotations of the H&E section. Each tile was colored by the corresponding predicted cell's assigned cell type and compared to pathologist-assigned regional annotations.

Code availability

Code for comprehensive analysis as described here, and figure generation as shown here can be found at https://github.com/ccomiter/SCHAF.

Data availability

HTAPP data can be found as part of the HTAN-HTAPP data release on Synapse (syn20834712).

Links for accession of MSKCC datasets used in figures can be found at the below links:

MSKCC-24: https://github.com/ccomiter/SCHAF

MSKCC-16: https://github.com/ccomiter/SCHAF

MSKCC datasets used in training will be published in their respective studies (MC, BL, YY, ORR, AR, and CMR).

Author Contributions

CC, EDV, and AR conceived the study. CC designed and executed the study with guidance from AR. MC, BL, YY, and CMR mined and provided SCLC tissue data. JJ, MS, JWu, SV, JWa, TRB, AS, DA, NW, XZ, and JK mined and provided the MBC tissue data. SJR, MT, and KLP provided expert pathological annotations of MBC tissue data. EDV, KJKK, and JS provided methodological suggestions. AR provided feedback and guidance. KJKK, JS contributed to manuscript preparation and writing. CC and AR wrote the manuscript, with input from all authors.

Conflict of interest

AR is a co-founder and equity holder of Celsius Therapeutics, an equity holder in Immunitas, and until July 31, 2020 was an SAB member of ThermoFisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics, and Asimov. From August 1, 2020, AR is an employee of Genentech and

has equity in Roche. JS is a scientific advisor for Arcadia Science. CMR has consulted regarding oncology drug development with AbbVie, Amgen, AstraZeneca, D2G, Daiichi Sankyo, Epizyme, Genentech/Roche, Ipsen, Jazz, Kowa, Lilly, Merck, and Syros. He serves on the scientific advisory boards of Bridge Medicines, Earli, and Harpoon Therapeutics. EDV is the founder of Sequome, Inc. BL, YY, ORR, and NW are employees as of 9 August 2021, 7 June 2021,19 October 2020, and Feb 13, 2023, respectively and have equity in Roche. SJR receives research support from Bristol-Myers-Squibb and KITE/Gilead. SJR is a member of the SAB of Immunitas Therapeutics. NW is an equity holder in Relay Therapeutics and a consultant and equity holder in Flare Therapeutic. Prior to Jan 31 2023 he was an SAB member of Relay Therapeutics, an advisory board member for Eli Lilly, and received research support from Astra-Zeneca and Puma Biotechnologies. MC is an employee of Mission Bio as of 22 February 2022. A patent application has been filed related to this work through Broad Institute.

Acknowledgements

This paper is part of the Human Cell Atlas. We thank Leslie Gaffney for help with figure preparation, Ania Hupalowska for help with graphics, and Regev lab members for discussions. Work was supported by the Klarman Cell Observatory and Howard Hughes Medical Institute (AR). This project was also funded in part by federal funds from NCI, NIH Task Order HHSN261100039 under Contract HHSN261201500003. CC was supported by an MIT Department of Electrical Engineering and Computer Science Fellowship, Teaching Assistantship, and Research Assistantship. JS was supported by NIH New Innovator Award (DP2TR004354), NIH Common Fund (UG3CA275687), NICHD (R00HD096049), Massachusetts Life Science Center, Burroughs Wellcome Fund, Additional Ventures, and

Massachusetts General Hospital. KJKK was supported by the Japan Society for the Promotion of Science Postdoctoral Fellowship for Overseas Researchers, and the Naito Foundation Overseas Postdoctoral Fellowship. CMR was supported by NIH U24 CA213274 and R35 CA263816. EDV was supported by the MIT Presidential Fellowship.

References

Achim, Kaia, Jean-Baptiste Pettit, Luis R. Saraiva, Daria Gavriouchkina, Tomas Larsson, Detlev Arendt, and John C. Marioni. 2015. "High-Throughput Spatial Mapping of Single-Cell RNA-Seq Data to Tissue of Origin." *Nature Biotechnology* 33 (5): 503–9.

Alon, Shahar, Daniel R. Goodwin, Anubhav Sinha, Asmamaw T. Wassie, Fei Chen, Evan R. Daugharthy, Yosuke Bando, et al. 2020. "Expansion Sequencing: Spatially Precise In Situ Transcriptomics in Intact Biological Systems." *bioRxiv*. https://doi.org/10.1101/2020.05.13.094268.

Biancalani, Tommaso, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, et al. 2020. "Deep Learning and Alignment of Spatially-Resolved Whole Transcriptomes of Single Cells in the Mouse Brain with Tangram." *bioRxiv*. https://doi.org/10.1101/2020.08.29.272831.

———. 2021. "Deep Learning and Alignment of Spatially Resolved Single-Cell Transcriptomes with Tangram." *Nature Methods* 18 (11): 1352–62.

Chen, Kok Hao, Alistair N. Boettiger, Jeffrey R. Moffitt, Siyuan Wang, and Xiaowei Zhuang. 2015. "RNA Imaging. Spatially Resolved, Highly Multiplexed RNA Profiling in Single Cells." *Science* 348 (6233): aaa6090.

Chen, Xiaoyin, Yu-Chi Sun, George M. Church, Je Hyuk Lee, and Anthony M. Zador. 2018. "Efficient in Situ Barcode Sequencing Using Padlock Probe-Based BaristaSeq." *Nucleic Acids Research* 46 (4): e22.

Codeluppi, Simone, Lars E. Borm, Amit Zeisel, Gioele La Manno, Josina A. van Lunteren, Camilla I. Svensson, and Sten Linnarsson. 2018. "Spatial Organization of the Somatosensory Cortex Revealed by osmFISH." *Nature Methods* 15 (11): 932–35.

Codeluppi, Simone, Lars E. Borm, Amit Zeisel, Gioele La Manno, Josina A. van Lunteren, Camilla I. Svensson, and Sten Linnarsson. 2018. "Spatial Organization of the Somatosensory Cortex Revealed by Cyclic smFISH." *bioRxiv*. bioRxiv. https://doi.org/10.1101/276097.

Ding, Jiarui, and Aviv Regev. 2019. "Deep Generative Model Embedding of Single-Cell RNA-Seq Profiles on Hyperspheres and Hyperbolic Spaces." *bioRxiv*. bioRxiv. https://doi.org/10.1101/853457.

Eng, Chee-Huat Linus, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, et al. 2019. "Transcriptome-Scale Super-Resolved Imaging in Tissues by RNA seqFISH." *Nature* 568 (7751): 235–39.

Howell, David C. 2012. *Statistical Methods for Psychology*. Cengage Learning.

Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. "Image-to-Image Translation with Conditional Adversarial Networks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1125–34.

Jerby-Arnon, Livnat, and Aviv Regev. 2022. "DIALOGUE Maps Multicellular Programs in Tissue from Single-Cell or Spatial Transcriptomics Data." *Nature Biotechnology* 40 (10): 1467–77.

Ke, Rongqin, Marco Mignardi, Alexandra Pacureanu, Jessica Svedlund, Johan Botling, Carolina Wählby, and Mats Nilsson. 2013. "In Situ Sequencing for RNA Analysis in Preserved Tissue and Cells." *Nature Methods* 10 (9): 857–60.

Kingma, and Ba. n.d. "Adam: A Method for Stochastic Gradient Descent." *ICLR: International Conference on Learning*.

Kingma, Diederik P., and Max Welling. 2013. "Auto-Encoding Variational Bayes." *arXiv*

[stat.ML]. arXiv. http://arxiv.org/abs/1312.6114v10.

Kobayashi-Kirschvink, Koseki J., Shreya Gaddam, Taylor James-Sorenson, Emanuelle Grody, Johain R. Ounadjela, Baoliang Ge, Ke Zhang, et al. 2022. "Raman2RNA: Live-Cell Label-Free Prediction of Single-Cell RNA Expression Profiles by Raman Microscopy." *bioRxiv*. https://doi.org/10.1101/2021.11.30.470655.

Kumar, L. Ashok, D. Karthika Renuka, S. Lovelyn Rose, M. C. Shunmuga priya, and I. Made Wartana. 2022. "Deep Learning Based Assistive Technology on Audio Visual Speech Recognition for Hearing Impaired." *International Journal of Cognitive Computing in Engineering* 3 (June): 24–30.

Laak, Jeroen van der, Geert Litjens, and Francesco Ciompi. 2021. "Deep Learning in Histopathology: The Path to the Clinic." *Nature Medicine* 27 (5): 775–84.

Lähnemann, David, Johannes Köster, Ewa Szczurek, Davis J. McCarthy, Stephanie C. Hicks, Mark D. Robinson, Catalina A. Vallejos, et al. 2020. "Eleven Grand Challenges in Single-Cell Data Science." *Genome Biology* 21 (1): 31.

Luecken, M. D., M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, et al. 2020. "Benchmarking Atlas-Level Data Integration in Single-Cell Genomics." *bioRxiv*. https://doi.org/10.1101/2020.05.22.111161.

Makhzani, Alireza, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. 2015. "Adversarial Autoencoders." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1511.05644.

McInnes, Leland, John Healy, and James Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1802.03426.

Ngiam, Jiquan, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng.

2011. “Multimodal Deep Learning.” https://openreview.net/pdf?id=Hk4OO3W_bS.

“{{ngMeta[’og:Title']}}.” n.d. Accessed August 25, 2022. https://bio.tools/histomicstk.

Odena, Augustus, Christopher Olah, and Jonathon Shlens. 2017. “Conditional Image Synthesis with Auxiliary Classifier Gans.” In *International Conference on Machine Learning*, 2642–51. PMLR.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. 2019. “PyTorch: An Imperative Style, High-Performance Deep Learning Library.” In *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Regev, Aviv, Sarah A. Teichmann, Eric S. Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, et al. 2017. “The Human Cell Atlas.” *eLife* 6 (December). https://doi.org/10.7554/eLife.27041.

Regev, Aviv, Sarah Teichmann, Orit Rozenblatt-Rosen, Michael Stubbington, Kristin Ardlie, Ido Amit, Paola Arlotta, et al. 2018. “The Human Cell Atlas White Paper.” *arXiv [q-bio.TO]*. arXiv. http://arxiv.org/abs/1810.05192.

Rood, Jennifer E., Aidan Maartens, Anna Hupalowska, Sarah A. Teichmann, and Aviv Regev. 2022. “Impact of the Human Cell Atlas on Medicine.” *Nature Medicine* 28 (12): 2486–96.

Rozenblatt-Rosen, Orit, Aviv Regev, Philipp Oberdoerffer, Tal Nawy, Anna Hupalowska, Jennifer E. Rood, Orr Ashenberg, et al. 2020. “The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution.” *Cell* 181 (2): 236–49.

Rozenblatt-Rosen, Orit, Michael J. T. Stubbington, Aviv Regev, and Sarah A. Teichmann. 2017. "The Human Cell Atlas: From Vision to Reality." *Nature* 550 (7677): 451–53.

Satija, Rahul, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. 2015. "Spatial Reconstruction of Single-Cell Gene Expression Data." *Nature Biotechnology* 33 (5): 495–502.

Schmauch, Benoît, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, et al. 2020. "A Deep Learning Model to Predict RNA-Seq Expression of Tumours from Whole Slide Images." *Nature Communications* 11 (1): 3877.

Schofield, Jeremy A., Erin E. Duffy, Lea Kiefer, Meaghan C. Sullivan, and Matthew D. Simon. 2018. "TimeLapse-Seq: Adding a Temporal Dimension to RNA Sequencing through Nucleoside Recoding." *Nature Methods* 15 (3): 221–25.

Ståhl, Patrik L., Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, et al. 2016. "Visualization and Analysis of Gene Expression in Tissue Sections by Spatial Transcriptomics." *Science* 353 (6294): 78–82.

Tang, Xiaoning, Yongmei Huang, Jinli Lei, Hui Luo, and Xiao Zhu. 2019. "The Single-Cell Sequencing: New Developments and Medical Applications." *Cell & Bioscience* 9 (June): 53.

Wang, Xiao, William E. Allen, Matthew A. Wright, Emily L. Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, et al. 2018. "Three-Dimensional Intact-Tissue Sequencing of Single-Cell Transcriptional States." *Science* 361 (6400). https://doi.org/10.1126/science.aat5691.

Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis. 2018. "SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis." *Genome Biology* 19 (1): 15.

Yang, Karren D., and Caroline Uhler. 2019. "Multi-Domain Translation by Learning Uncoupled Autoencoders." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1902.03515.

Yan, Le Cun, B. Yoshua, and H. Geoffrey. 2015. "Deep Learning." *Nature* 521 (7553): 436–44.

Zhu, Hao, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. 2021. "Deep Audio-Visual Learning: A Survey." *International Journal of Automation and Computing* 18 (3): 351–76.

Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks." In *Proceedings of the IEEE International Conference on Computer Vision*, 2223–32.

Ziegenhain, Christoph, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. 2017. "Comparative Analysis of Single-Cell RNA Sequencing Methods." *Molecular Cell* 65 (4): 631–43.e4.

Figure 1

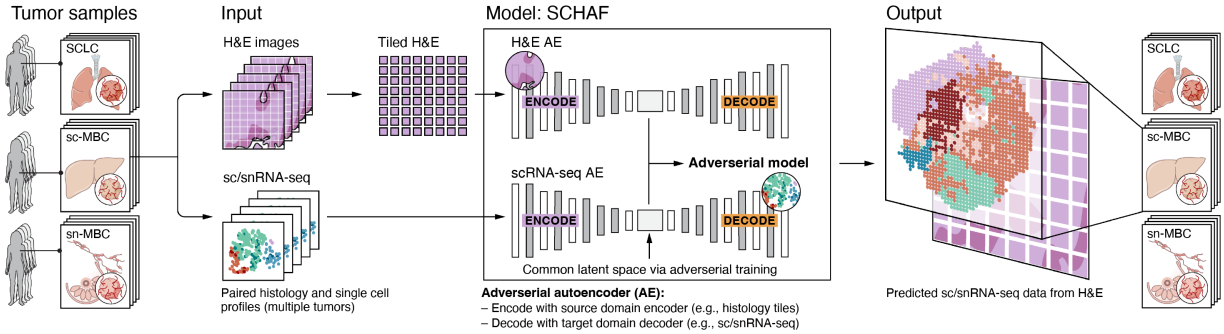


Figure 2

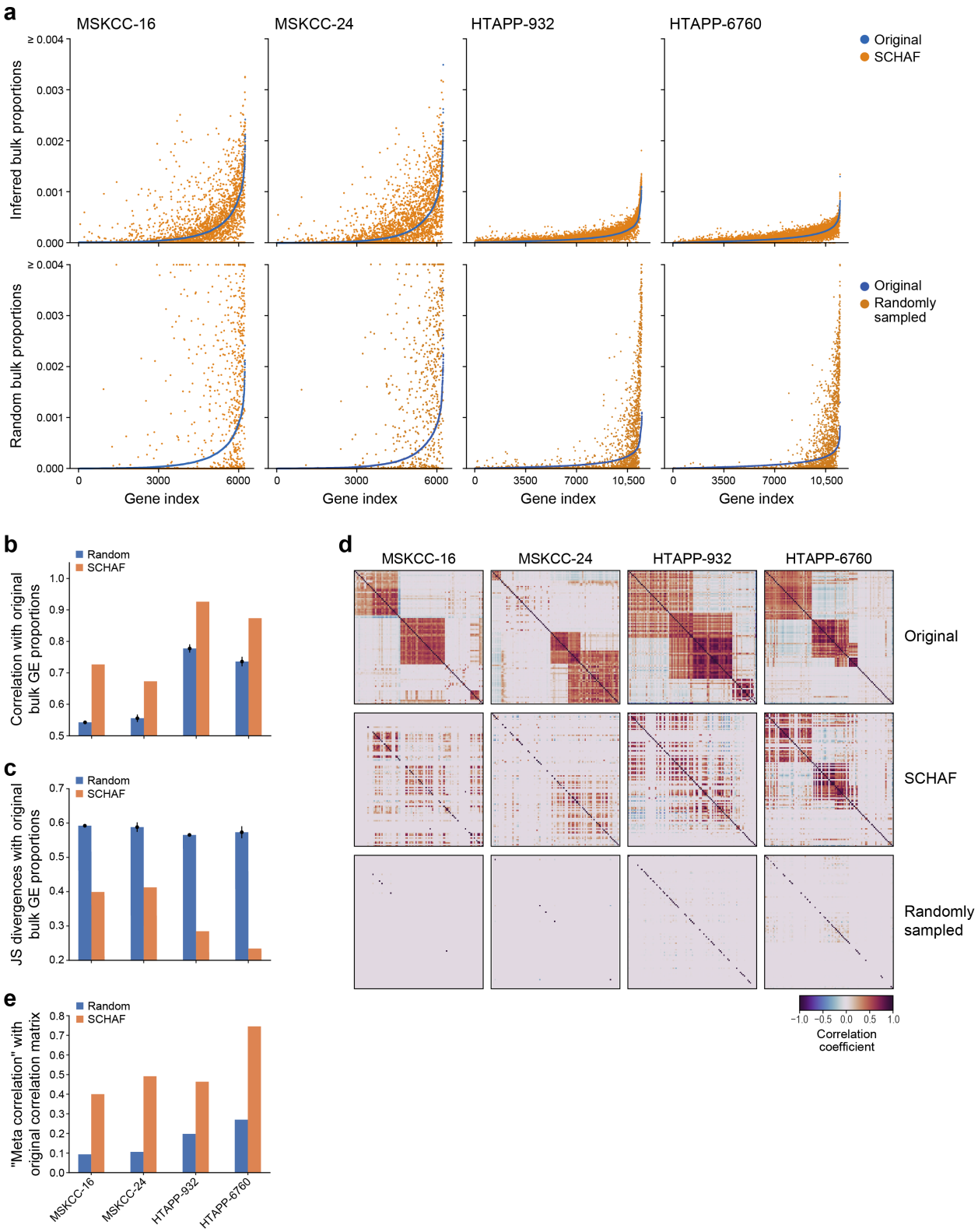


Figure 3

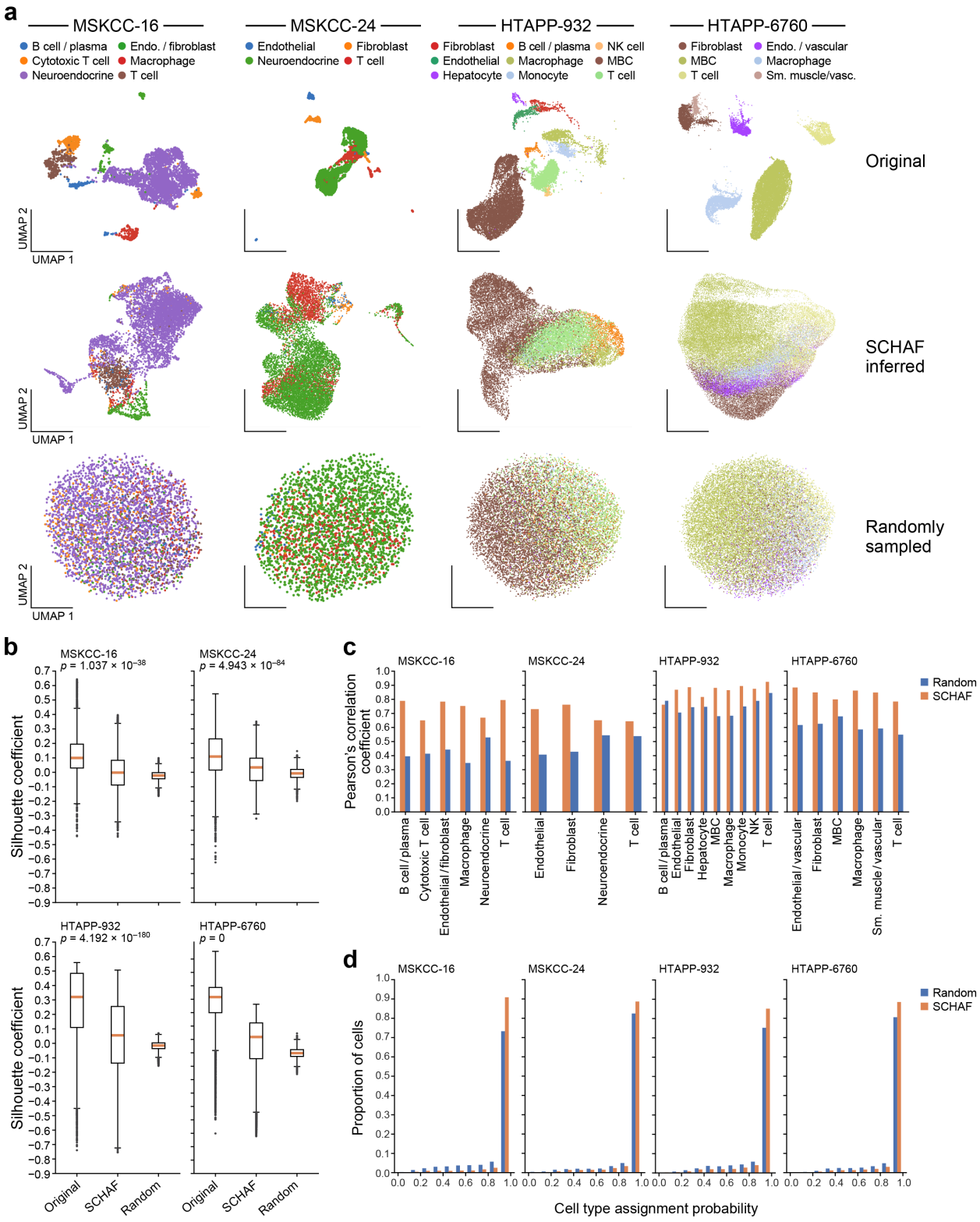
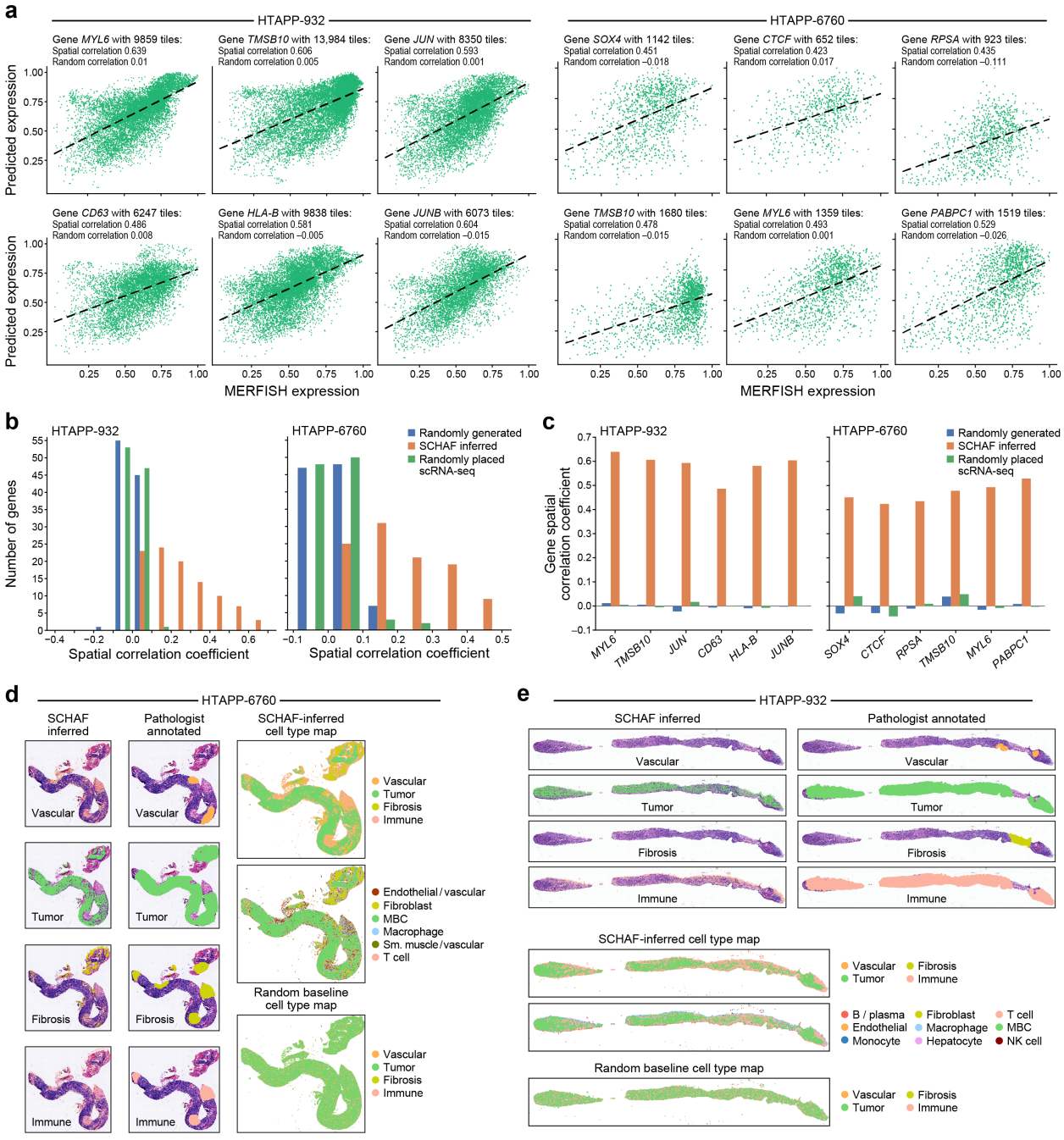
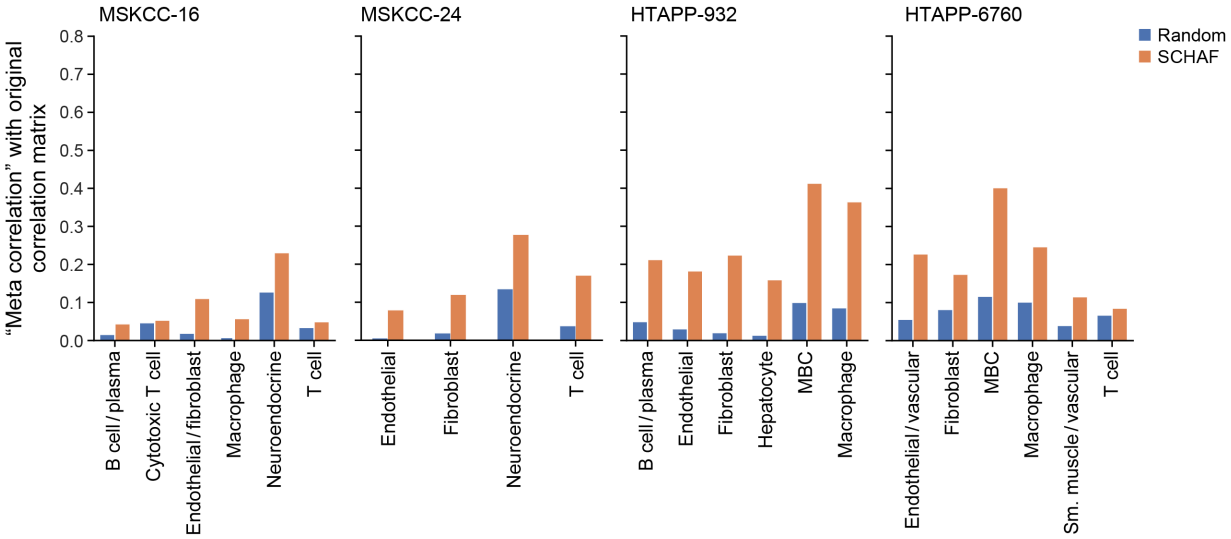


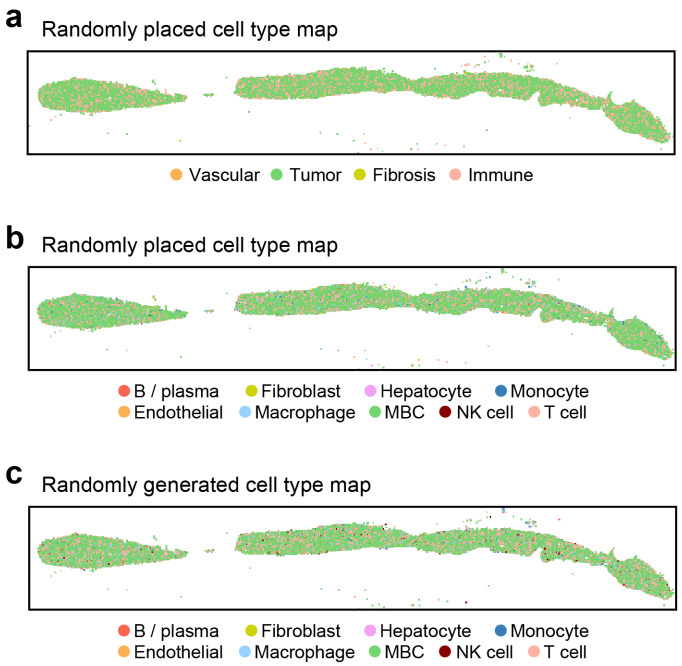
Figure 4



Supplementary Figure 1



Supplementary Figure 2



Supplementary Figure 3