nature methods

SUPPLEMENTARY INFORMATION

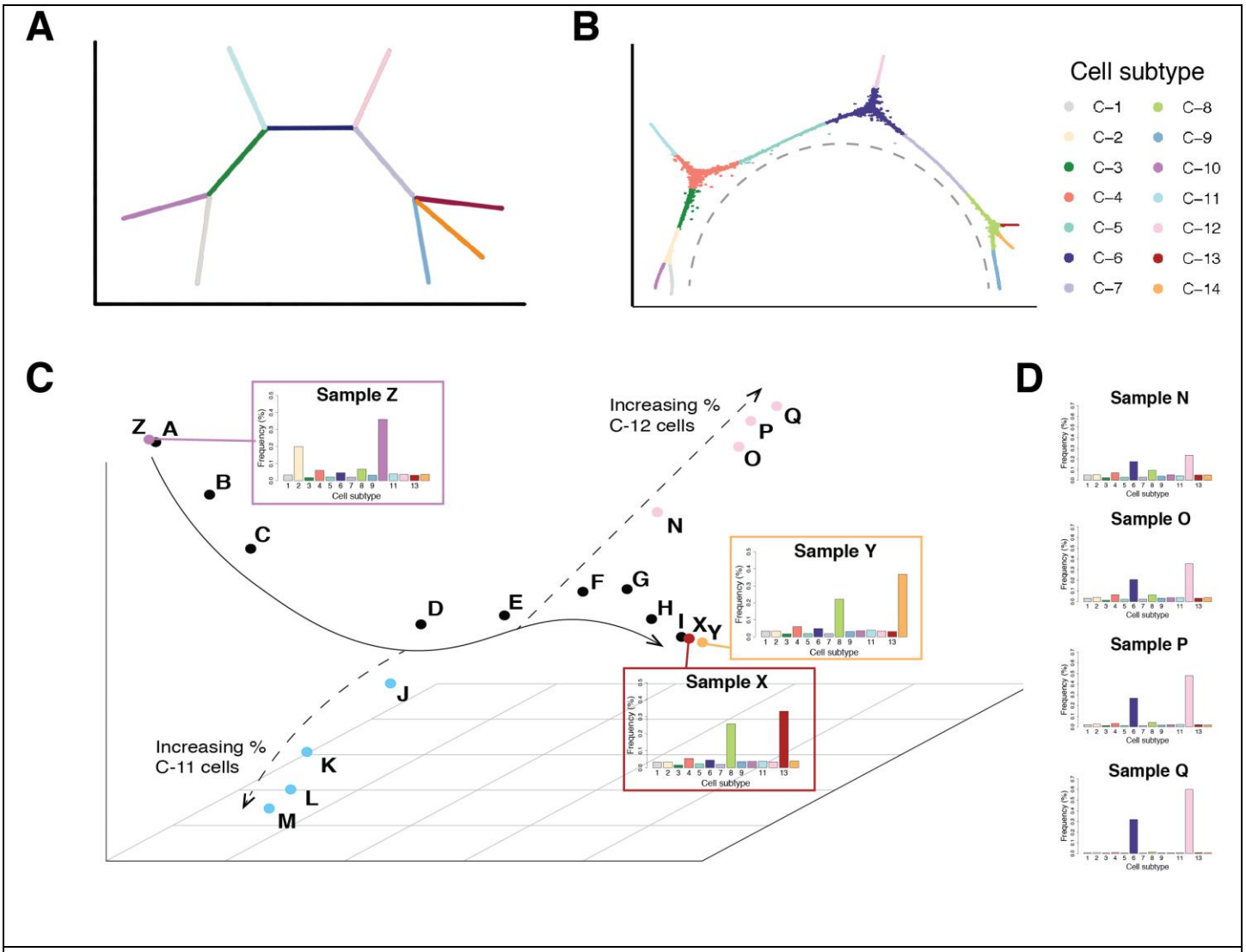
https://doi.org/10.1038/s41592-019-0689-z

In the format provided by the authors and unedited.

Uncovering axes of variation among single-cell cancer specimens

William S. Chen[1,5], Nevena Zivanovic[2,5], David van Dijk[1,3], Guy Wolf[4], Bernd Bodenmiller[2,6*] and Smita Krishnaswamy[1,3,6*]

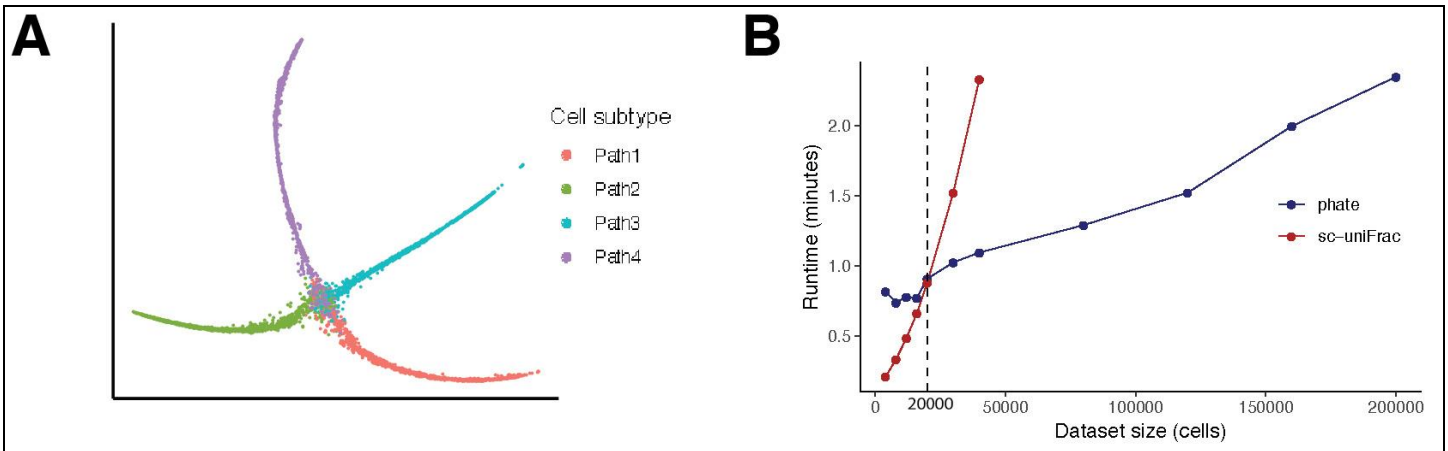
[1]Department of Genetics, Yale School of Medicine, New Haven, CT, USA. [2]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. [3]Department of Computer Science, Yale University, New Haven, CT, USA. [4]Department of Mathematics and Statistics, Université de Montréal, Montreal, Quebec, Canada. [5]These authors supervised this work: Bernd Bodenmiller, Smita Krishnaswamy. [6]These authors contributed equally: William S. Chen, Nevena Zivanovic. *e-mail: bernd.bodenmiller@imls.uzh.ch; smita.krishnaswamy@yale.edu



Supplementary Figure 1

PhEMD correctly recovers cell-state and biospecimen embeddings in a simulated single-cell dataset with known ground-truth structure.

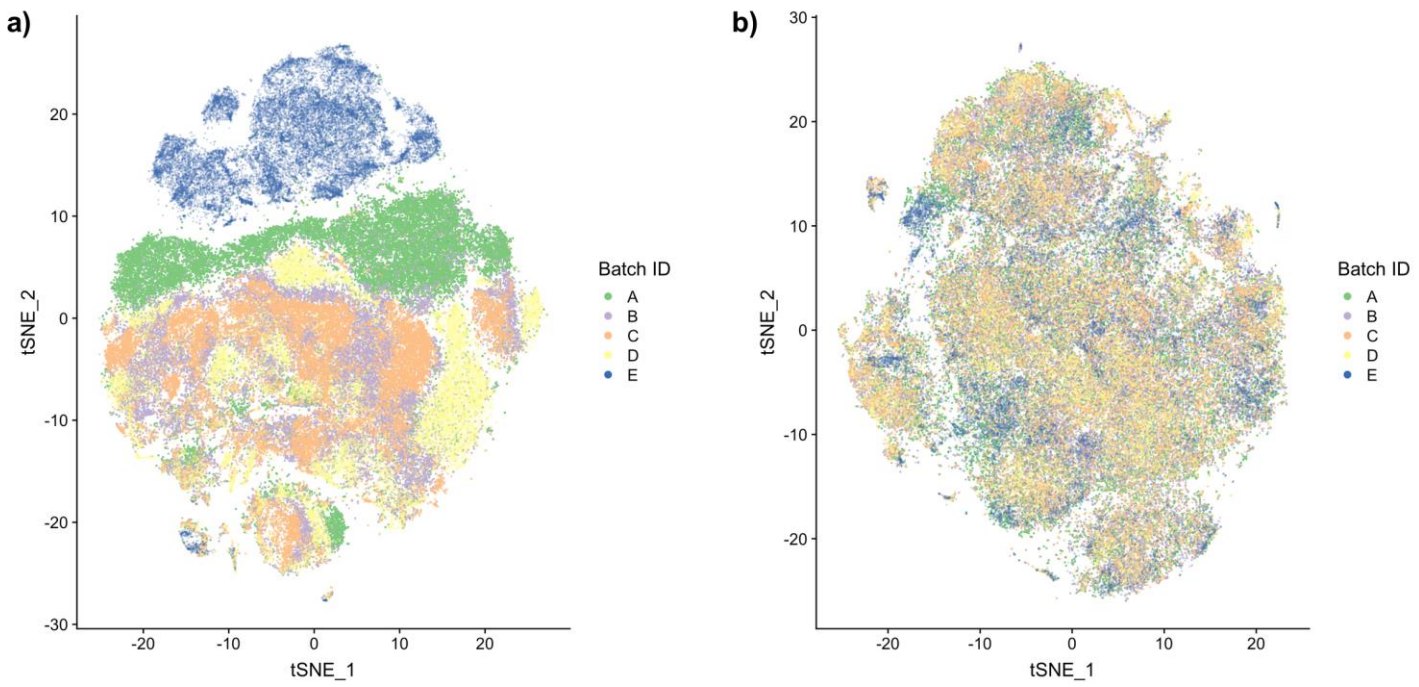
a) Ground-truth tree structure of the cell-state space of Synthetic Dataset A (see Online Methods for data parameters). b) PHATE embedding of the cell-state space of Synthetic Dataset A, colored by cell-subtypes identified by PHATE. Grey dotted line denotes major axis (comprised of cell subtypes C-1 through C-9) along which density is modulated for biospecimens A–I. c) Diffusion map embedding of biospecimens. Points colored black and labeled A–I represent samples that have density concentrated at various clusters along the trajectory from C-1 (“starting state”) and ending at C-9 (“terminal state”) highlighted in grey. The alphabetical ordering of samples from A–I correspond to increasing intra-sample relative proportions of starting state to terminal state points. Samples X and Y represent specimens with cells concentrated in clusters C-13 and C-14 respectively (i.e. highly similar cell subtypes), and Sample Z represents a specimen with cells concentrated in cluster C-11 (highly dissimilar to cell subtypes C-13 and C-14). d) Relative frequency histograms representing distribution of cells across different cell subtypes for selected samples forming a sub-trajectory in the biospecimen embedding.



Supplementary Figure 2

Runtime comparison between PHATE and sc-UniFrac for comparing single-cell specimens.

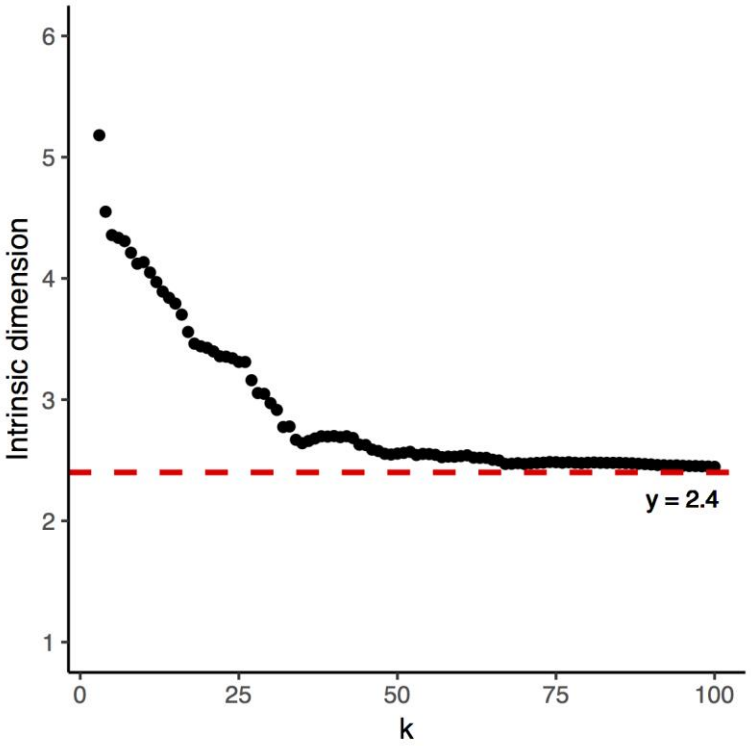
a) PHATE embedding of the cell-state space of Synthetic Dataset B colored by cell-subtypes identified by PHATE. b) Runtime comparison between PHATE and sc-UniFrac applied to datasets of increasingly larger sample sizes.



Supplementary Figure 3

Assessment of CCA batch correction results in multi-batch EMT drug screen experiment.

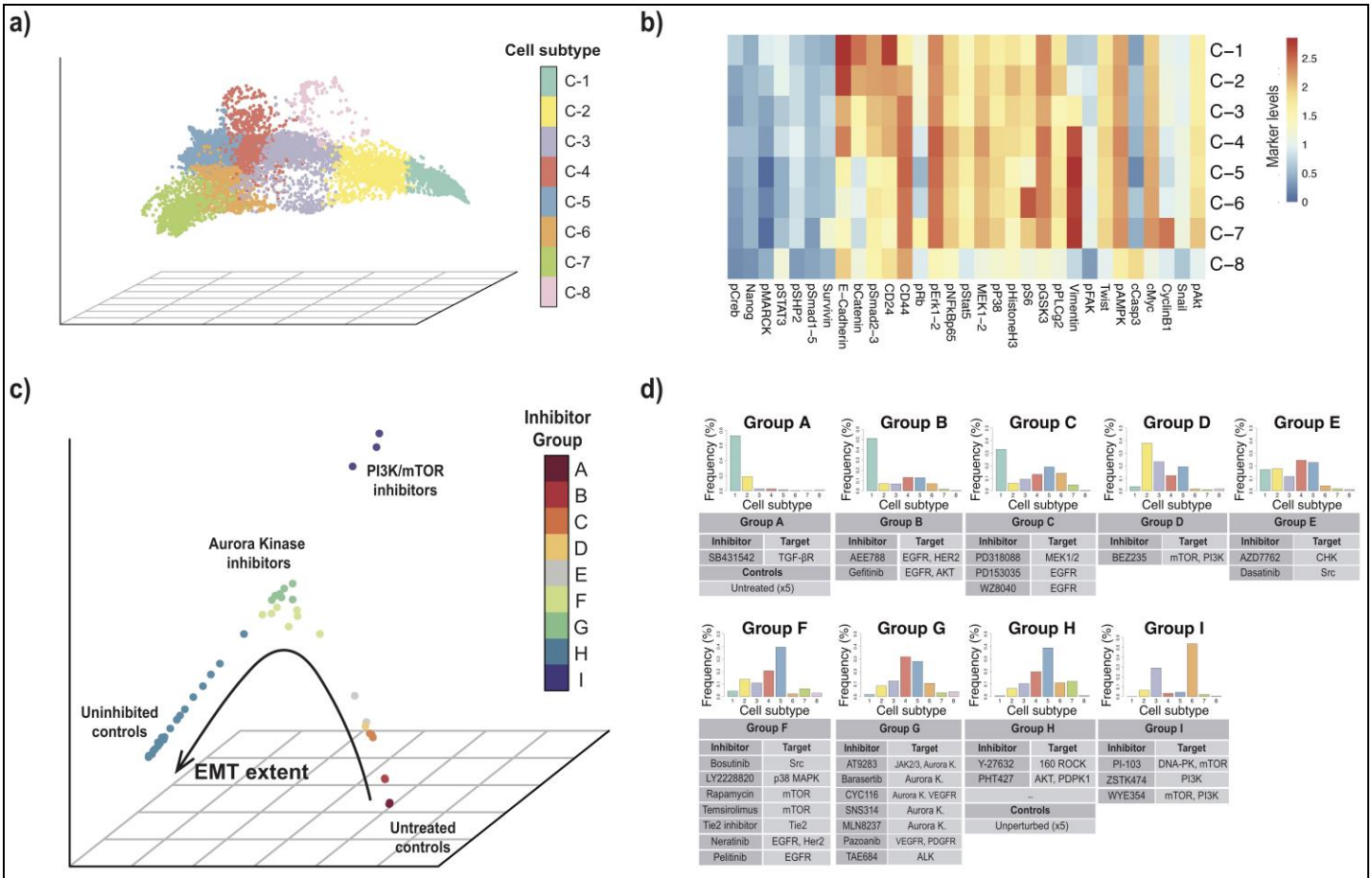
t-SNE embedding of cells from multiple CyTOF runs based on gene expression data a) pre- and b) post-CCA batch correction, with individual cells colored by experimental batch.



Supplementary Figure 4

Intrinsic dimensionality analysis of the PhEMD embedding comprised of 300-sample multi-batch EMT inhibition and control conditions.

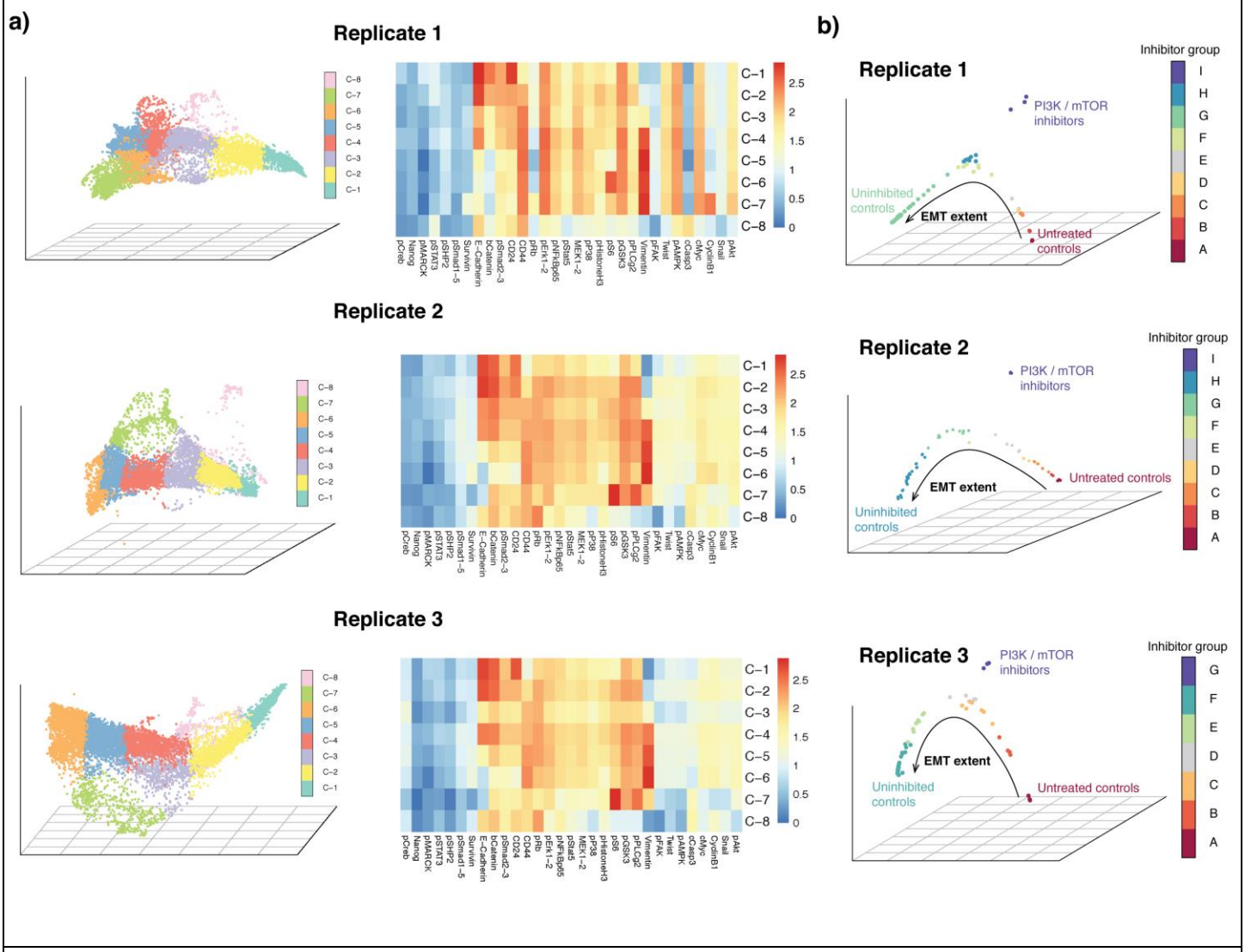
Intrinsic dimension computed using the maximum likelihood estimation (MLE) approach over a range of “k” (k-nearest-neighbors parameter) values.



Supplementary Figure 5

PhEMD analysis of 60 inhibition and control conditions measured in a single CyTOF run.

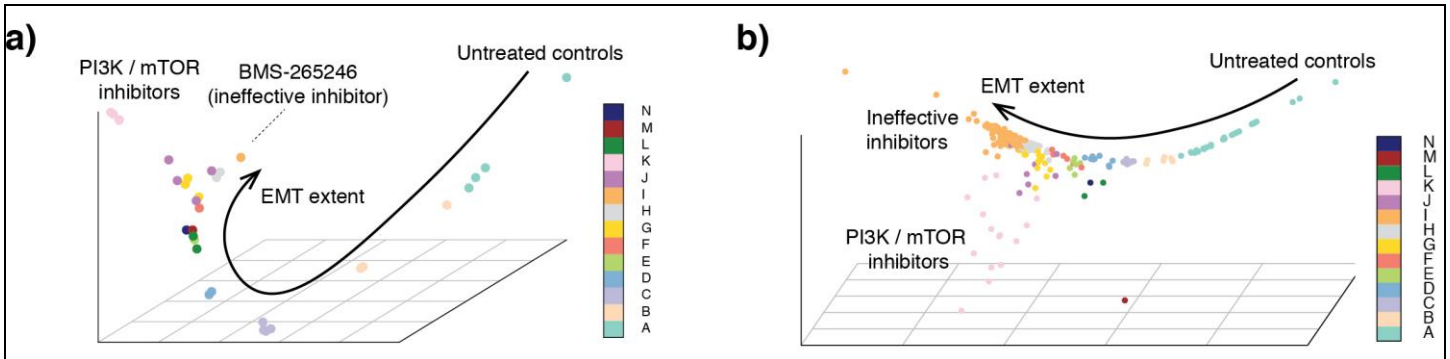
a) PHATE embedding of cells from all conditions of a single CyTOF run representing perturbed EMT cell state landscape, colored by cell subtype determined using spectral clustering. b) Heatmap of mean $\log_2$ protein expression levels for each subpopulation of cells representing a distinct cell subtype. c) Embedding of drug inhibitors, colored by clusters assigned by hierarchical clustering. d) Individual inhibitors assigned to each inhibitor group. Histograms represent bin-wise mean of relative frequency of each cell subtype for all inhibitors in a given group. The full list of inhibitors in each group can be found in Table S5.



Supplementary Figure 6

Reproducibility of results across three experimental replicates (independent cell culture experiments measured in distinct CyTOF runs) of EMT data.

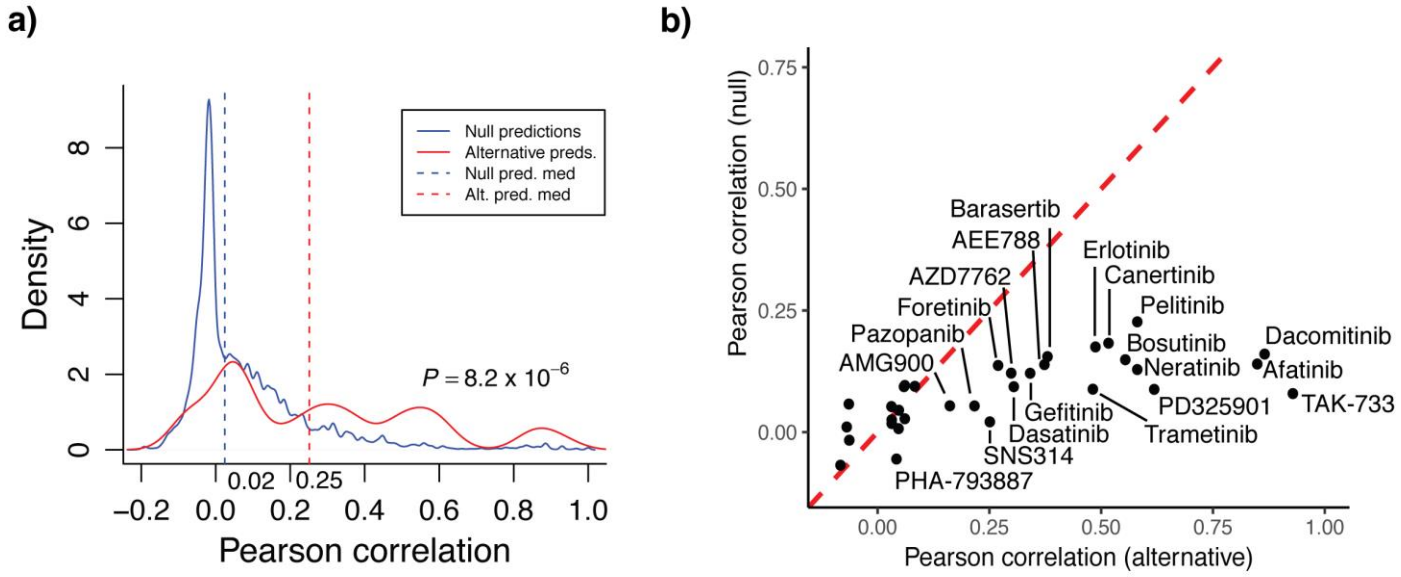
a) Cell subtype expression patterns and cell-state embeddings for three independent experimental replicates. b) PhEMD biospecimen embeddings and inhibitor clusters for three independent experimental replicates. The full list of inhibitors in each group can be found in Table S5.



Supplementary Figure 7

Identification of landmark points encompassing network geometry of full 300-specimen EMT experiment.

a) Diffusion map embedding of 300-specimen EMT experiment, plotting only the 34 landmark points identified using a previously published diffusion map sampling technique (see Online Methods). Points are colored based on cluster assignments as determined based on original clustering of all 300 samples (see Figure 3c). b) Reconstructed diffusion map embedding, generated by starting with the 34 landmark points and using a previously published out-of-sample extension technique to infer the embedding coordinates of all 300 samples relative to these 34 landmark points (see Online Methods).



Supplementary Figure 8

Predicting known drug-target binding specificity using PhEMD results based on single-cell profiling of EMT drug screen experiment.

a) Probability density functions representing distribution of Pearson correlations between predicted and known drug-target binding specificity profiles. The null (n=39,000 predictions from 1,000 independent permutations) vs. alternative (n=39 predictions) models demonstrated median correlation-based accuracy of 0.02 vs. 0.25, P=8.2*10^{-6}. Statistical testing was performed using a one-sided Mann-Whitney U-test. b) Pearson correlation-based prediction accuracy of null (n=1,000 permutations per inhibitor) vs. alternative (one prediction per inhibitor) models for predicting the drug-target binding specificity of each inhibitor. Given multiple null-model predictions for each inhibitor, the *y*-axis represents mean prediction accuracy of all predictions for a given inhibitor. See Online Methods for detailed properties of the null and alternative models.

10 Supplementary notes

Supplementary Note 1: Leveraging single-cell resolution to distinguish samples indistinguishable by bulk expression analysis

Traditional bulk sequencing and bulk expression analysis may reveal trends that inadequately reflect the true differences between biological samples. For example, a prior report studying pulsatile expression of p53 in cells before and after γ-irradiation treatment found that on bulk analysis, the average amplitude of pulses (i.e., magnitude of response to treatment) was greater with increasing dose of irradiation [1]. A natural conclusion from this observation may be that cells express increased p53 in response to irradiation-induced DNA damage. However, the group then performed the same experiment but obtained single-cell instead of bulk measurements. This experiment revealed that the pulse amplitude for a given cell was actually constant and independent of irradiation dose; the change in average pulse amplitude on bulk analysis was attributable to changing proportions (i.e., preferential survival and/or proliferation) of certain cell subpopulations rather than changes in individual cells themselves. Without single-cell resolution, this distinction could not be made.

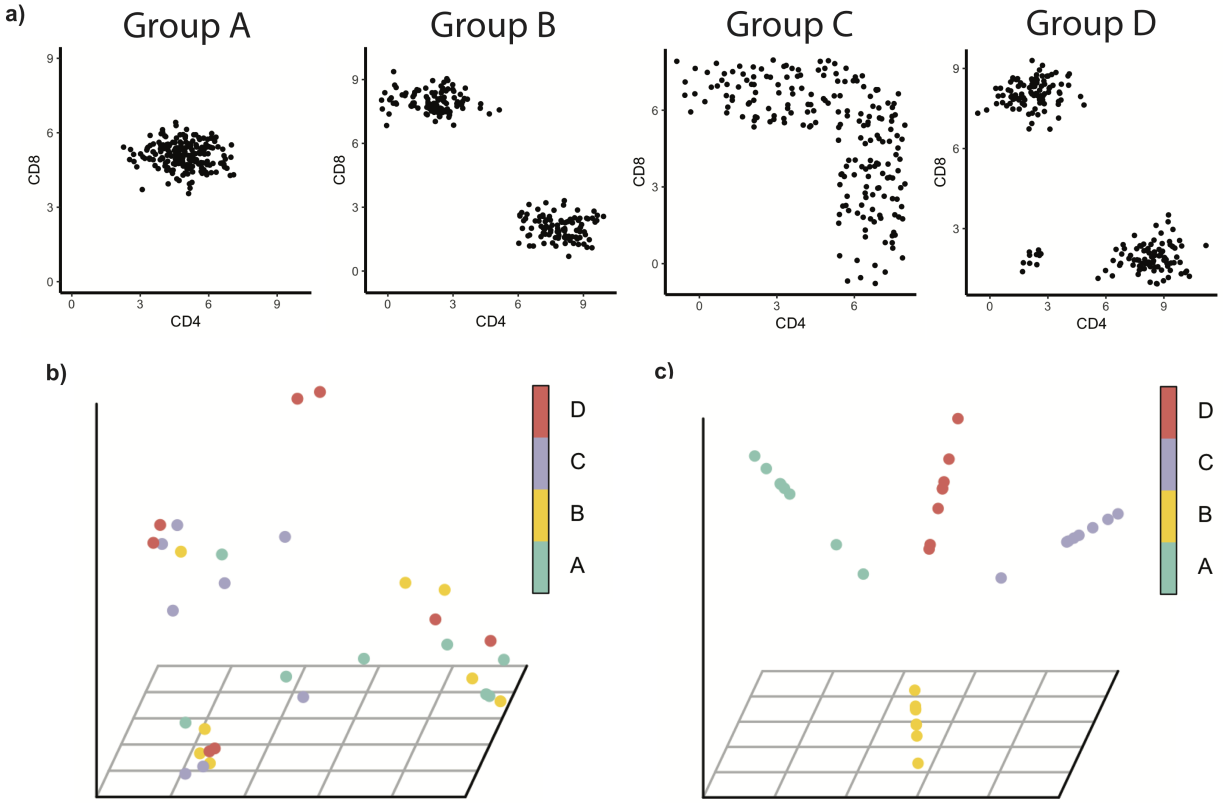
In addition to lacking the resolution to explain certain phenomena, bulk measurements may fail to detect true biological differences between experimental conditions altogether. To demonstrate this concept more concretely and highlight the utility of single-cell analytical approaches for distinguishing between biological samples, we computationally modeled a multi-sample dataset consisting of immune cells with collectively variable expression of CD4 and CD8 (Figure SN1A). Each sample was a cell population that fit one of four distribution patterns. Group A samples each consisted of a homogeneous immune cell population characterized by intermediate expression of both CD4 and CD8. Group B samples each consisted of two similarly-sized immune cell subpopulations: one $CD4^+$ and one $CD8^+$ subpopulation. Group C samples consisted of a mixture of $CD4^+$, $CD8^+$, and CD4/CD8 double-positive (DP) immune cells. Group D samples consisted of one $CD4^+$ and one $CD8^+$ subpopulation of roughly equal size and one additional rare subpopulation of CD4/CD8 double-negative (DN) immune cells. Note that these immune cell subtypes ($CD4^+$, $CD8^+$, DP, and DN) have been reported to exist in normal thymus as well as various disease states (e.g., breast and hematologic malignancies [2, 3]). Our simulated experiment consisted of 32 samples in total (8 of each of Groups A-D). By design, the bulk (average) expression of CD4 and CD8 for each sample was roughly the same for all samples, regardless of differences in cell subpopulation characteristics.

Our goal was to relate the 32 samples to one another in a biologically meaningful way. This could be done by generating a low-dimensional embedding that could be visualized to view the similarity of any two samples relative to the rest and identify groups of similar samples. We first attempted to do so using bulk measurements. We generated a sample-sample distance matrix by computing pairwise (Euclidean) distances between each pair of samples, with each sample represented as its average protein (i.e., CD4 and CD8) expression. We then embedded this distance matrix using a diffusion map. The result was an embedding that failed to differentiate samples based on biologically important differences. Specifically, samples of the same known, ground-truth subtype (i.e., Group A-D) failed to map to similar parts of the resulting embedding (Figure SN1B).

A better approach to comparing these samples was to compare the presence and abundance of all single-cell subpopulations in each sample. We formalize an approach ("PhEMD") in this

manuscript and demonstrate that it can be used to effectively distinguish single-cell samples from one another that cannot be distinguished based on bulk or average expression patterns. In this particular example, the PhEMD diffusion map embedding vastly outperformed the bulk approach described above and successfully differentiated samples based on biologically important differences in cell subpopulation characteristics and proportions (Figure SN1C).

Supplementary Figure SN1: Single-cell analysis can resolve differences between biological samples that are indistinguishable on bulk (average) expression analysis. a) Single-cell profiles of each biological sample in a computationally-generated immune cell dataset. Groups A-D each have 8 samples that fit the single-cell profile. By design, all samples have roughly the same bulk expression of CD4 and CD8. b) Diffusion map embedding generated by embedding a sample-to-sample distance matrix, where pairwise distances between samples were computed by taking the Euclidean distance between samples represented as bulk expression of CD4 and CD8. Bulk expression profiles do not adequately reflect the biological differences between samples in this dataset and cannot be used to distinguish samples in a biologically meaningful way. c) Diffusion map embedding generated by embedding a PhEMD distance matrix, which takes into account single-cell characteristics of each sample (see "Overview of PhEMD" in Results section). PhEMD successfully distinguishes samples with different single-cell profiles from one another.



Supplementary Note 2: Conceptualizing and comparing heterogeneous single-cell specimens

Single-cell samples are "high-dimensional" in that each sample is not only represented as many feature (i.e., gene) measurements but also consists of many datapoints. The analysis of high-dimensional data, especially in unsupervised or exploratory settings, often introduces various chal-

lenges that are collectively referred to as the curse of dimensionality [4, 5]. A popular approach to analyzing such data is to use manifold learning methods that assume the intrinsic geometry of the data can conceptually be modeled as a low dimensional manifold (i.e., a collection of smoothly-varying, locally low-dimensional data patches), which is immersed in the high dimensional ambient space of collected features [6]. Such methods often aim to uncover this intrinsic geometry by first capturing local neighborhoods, then using them to form a rigid structure of nonlinear relations in the data, and finally embedding this structure in a low-dimensional (e.g., 2D or 3D) space via a new set of features that preserve those relations (e.g., as distances).

At the core of most manifold learning methods is the assumption that there exists some natural distance metric that can be used to define local neighborhoods in the data. Indeed, popular manifold learning methods are often based on selection of nearest neighbors over simple Euclidean distance, even though this distance is only meaningful locally due to the curse of dimensionality. In multi-specimen data, on the other hand, specimens are no longer individual vectors, but rather form data clouds with varying numbers of datapoints (i.e., cells). Therefore, to construct an intrinsic data geometry between biospecimens, we have to first define (and compute) a notion of distance between specimens, which can then be used for further analysis.

To compare two specimens, we consider two notions of quantifying the difference between the distributions represented by them. First, two distributions can be compared by considering how distinguishable they are from each other. Indeed, if they are nearly indistinguishable from sampled data, then the two specimens should be considered very similar, while the easier it is to set their distributions apart, the more different the specimens are from each other. This notion is typically considered in machine learning for generative tasks, e.g., to produce artificial images that are indistinguishable from real ones [7]. Second, two distributions can be compared by quantifying how hard it is to transform one distribution into another. If only a small perturbation is required, then the specimens are close together, while drastic changes mean they should be far apart. Remarkably, these two notions are closely related via the Kantorovich-Rubinstein duality theorem [8, 9] and can be computed using Earth Mover's Distance (also known as the Wasserstein metric) as detailed below.

Motivation for "cell subtype" features: Individual cells can be thought of as points in an n-dimensional space, in which each dimension represents the expression level of a specific gene or protein marker. Thus, a single-cell sample containing thousands of cells can be thought of as a distribution of points in an n-dimensional space. Our goal is to perform pairwise comparisons of samples. Naturally, it follows that we aim to compare the n-dimensional distributions by computing a distance between two distributions. A naive geometric approach is to bin the distribution into a n-dimensional joint histogram with b bins in each dimension. However, this results in b^n bins, which implies extreme sparsity and computational intractability, as n is generally in the tens to thousands for single-cell genomic datasets.

In a joint histogram approach using equi-width bins, most bins are hardly populated. In other words, this representation of the data is highly inefficient. One way to address this issue is to perform an adaptive binning approach that partitions data recursively to generate *uniformly populated* bins rather than equi-width bins, as was done by Orlova et al. [10]. However, while this can potentially mitigate the issue of sparsity, the issue of binning granularity remains. Histograms that are too finely binned may be too sparse to reveal significant differences when compared. On the other hand, overly-coarse binning sacrifices resolving power. Hyper-rectangular histogram bin-

ning methods, whether equi-width or adaptive, tend to have limited success achieving a balance between data representation efficiency and resolving power [11].

An alternative representation of multidimensional distributions that optimizes efficiency and resolving power involves the use of "signatures" and "weights" rather than equi-width or adaptive histograms [11]. Signatures are defined as the main clusters (high-density regions) of a multidimensional distribution, and weights represent cluster size. In image-processing applications, this representation of images was found to yield the best results when comparing images to one another for the task of image retrieval. An analogy can be made to single-cell data modeling for the purpose of comparing single-cell specimens, as single-cell specimens are similarly multidimensional distributions. Using the signature-and-weights architecture, a "signature" can be thought of as a distinct cell subpopulation (or "cell subtype" e.g. memory B-cells or $CD8^+$ effector T-cells), and the corresponding "weight" represents the number of cells in the cell subtype. The advantages of signatures and weights with respect to data representation efficiency may be intuitively extended from computer vision literature to our application; biologically relevant cell subtypes are the ideal signatures, or "bins," for organizing single-cell data.

Comparing multidimensional distributions represented as signatures and weights: Our final representation of a biospecimen is a categorical frequency histogram representing the relative abundance ("weights") of all possible cell subtypes ("signatures") found in all specimens collectively. Since our ultimate goal is to compare the similarity of specimens, we need some metric to compare the similarity of these histograms. A major challenge is to identify a metric that captures the similarity of unique bins (i.e., "signatures" or "cell subtypes") in the final distance measure. As a simple example using the EMT model, for a specimen with 80% mesenchymal, 10% transitional, and 10% epithelial cells, we would expect a specimen with 50% mesenchymal, 40% transitional, and 10% epithelial cells to be more similar (closer in distance) than a specimen with 50% mesenchymal, 10% transitional, and 40% epithelial cells. This would be consistent with our intuitive sense of distance because 80-10-10 represents that most cells have fully transitioned from epithelial to mesenchymal states, 50-40-10 represents that most cells have partly or fully transitioned, and 50-10-40 represents that almost half of the cells have not transitioned at all. Earth Mover's Distance (EMD) is a distance metric that mathematically encodes this intuition and can be used as a robust measure of dissimilarity. EMD is the optimal transport distance for moving datapoints from the way they are proportioned between clusters in one specimen to the way they are proportioned in the other. The transport distance increases with not only the number of data points moved but also the distance each data point is moved in the embedded space, known as the *ground distance*

To more concretely explain our notion of "ground distance," we will use our experiment including control and inhibited EMT specimens as an example. What is the "ground distance" between the "bins" in this experiment? Recall that each bin represents a cell subtype (e.g. mesenchymal). Each cell subtype is associated with various different data points (individual cells assigned to that subtype), so it can be represented as the centroid of the cluster of cells that comprise it. Thus, we can quantify the ground distance between bins as the distance between their representative centroids.

To define a measure of distance between centroids, we first observe that, by design, all cells undergoing EMT across all specimens originated from the same homogeneous epithelial population. Thus, it seems most reasonable to represent these cells as lying on a continuous trajectory with the epithelial cell subtype defined as the "origin" cell state. While EMT may have a primary linear

progression from epithelial to mesenchymal state, we expect additional terminal cell states (e.g. apoptotic, senescent, proliferative) in our aggregate cell population. In other words, we expect the trajectory to be potentially branched. In fact, many single-cell experiments represent data that are modeled well by branched trajectories, such as models of cell differentiation, cellular reprogramming, and immune response. To represent and visualize our data as a branched trajectory, we use PHATE, a tool specifically designed for single-cell data that uses a diffusion process to learn the cell-state space [12].

With the PHATE embedding, we are able to visualize relationships between cell subtypes. Of note, the PHATE embedding has a useful property in that Euclidean distances between points in the low-dimensional PHATE space approximate the diffusion distance (i.e., manifold distance) derived from the native dimension space (akin to the properties of diffusion maps) [12]. Thus, a robust quantitative measure of dissimilarity between cell subtypes may be derived by computing the Euclidean distance between the centroids of two cell subtypes in the PHATE space. We used these Euclidean distances between the centroids of all cell subtypes as our measure of "ground-distance" between cell subtypes for our final computation of EMD-based specimen-to-specimen distances.

Supplementary Note 3: Evaluating accuracy of PhEMD in mapping multi-specimen, single-cell dataset with known ground-truth structure

The final PhEMD mapping of the synthetic multi-specimen dataset with known ground-truth structure was assessed as follows. First, we examined the single-cell specimens in which a large number of cells were concentrated in a single branch. We found that specimens with cellular density concentrated in branches close to one another on the cell-state manifold (e.g. Samples X and Y) tended to map to regions close to one another on the biological-specimen manifold compared to specimens with cellular density concentrated in branches far from one another on the cell-state manifold (e.g. Samples X and Z). Next, we examined Samples A–I: specimens in which cellular density was modulated so that Sample A had cells mostly in the arbitrary "starting" state of the manifold, Sample I had cells mostly in an arbitrary "terminal" state, and Specimens B through H had progressively fewer cells in the "starting" state and more cells in the "terminal" state. We found that in the final biospecimen embedding, Samples A–I appropriately formed a trajectory and were ordered based on their intra-specimen relative proportions of "starting state" to "terminal state" cells. Finally, we examined Samples J-Q: specimens in which point density was concentrated in intermediate branches diverging from the main trajectory of the cell-state manifold (i.e., cell subtypes C-11 and C-12). We found that PhEMD correctly mapped these specimens to distinct branches in the final single-cell specimen embedding and correctly ordered them in terms of increasing enrichment of the C-11 and C-12 cell types. Overall, this demonstrated that our approach accurately inferred both the cell-type frequencies in each specimen and the similarity between cell subtypes.

Supplementary Note 4: PhEMD incorporates a more scalable and flexible approach to comparing single-cell specimens compared to existing methods

cellAlign was designed to compare two experimental conditions (i.e., two heterogeneous cell populations) by first modeling each condition as an unbranched trajectory of cells, then assigning a pseudotime value to each cell based on its ordinal position in the trajectory, and finally computing a distance between the two experimental conditions as the "cost" of aligning the two pseudotemporal trajectories. By nature of its implementation, cellAlign could not be applied to cell populations

sampled from branched cell-state trajectories, as it assumed cells with the same pseudotime value had identical gene expression profiles (an assumption violated in the setting of branched cell-state trajectories). Our implementation of EMD did not make such an assumption and was thus more flexible for analyzing datasets with branched cell-state trajectories.

sc-UniFrac was a different method that was similarly designed to compare two single-cell experimental conditions but that faced scalability issues. Its memory requirements exceeded that of a standard laptop (2.5 GHz Intel Core i7 processor, 16 GB RAM) when attempting to compare experimental conditions containing collectively greater than 40,000 cells using default parameters. This prevented it from being useful for analyzing large multi-specimen datasets such as our drug-screen experiment spanning 300 experimental conditions and over 1.7 million cells. In contrast to sc-UniFrac, which was unable to be run on a laptop to analyze a set of 40,000 cells from two or more experimental conditions, PhEMD was successfully run on the same laptop to analyze a set of over 360,000 cells from 60 experimental conditions in under 10 minutes. In light of these memory-based limitations of sc-uniFrac, we compared the runtime of our implementation of EMD to sc-uniFrac using a smaller dataset consisting of 20 single-cell specimens each containing 500 cells sampled from a cell-state tree ("Synthetic Dataset B"). The cell-state tree was generated using the Splatter `R` package and was characterized by four branches sharing a single branch point. We found that our implementation of EMD correctly recovered the known cell-state space of the dataset (Supplementary Figure 3A) and had faster empiric runtime than when analyzing datasets including more than 21,000 cells in total (Supplementary Figure 3B).

In sum, unlike cellAlign, which could only be applied to datasets in which all cells across all specimens were mappable to a single unbranched trajectory (e.g., a simple differentiation process), our approach could be used to compare specimens comprised of cells sampled from an underlying cell-state manifold that was potentially branched. Compared to sc-UniFrac, our implementation of EMD was much more scalable (Figure S3), allowing for the efficient pairwise comparison of multiple specimens as was required to generate a final embedding containing many single-cell specimens.

Supplementary Note 5: Assessing batch effect in multi-run experiment

Batch effect is a well-known problem when comparing data from multiple single-cell RNA-sequencing [13, 14] or CyTOF [15] experiments. Because of this, single-cell specimens are ideally processed and measured in a single batch. However, comparing specimens across experimental runs is still of great interest. In some cases, the sheer number of specimens makes simultaneous processing impossible. In other cases, the experimental design (e.g. time-series analysis) precludes sample processing on the same plate or gene profiling of all specimens simultaneously. In order to enable these sorts of experiments, a number of methods have been recently published that correct for batch effect. We chose canonical correlation analysis (CCA), a new feature of the popular Seurat package, as our batch correction tool and demonstrated that PhEMD can leverage existing batch correction methods to compare hundreds of specimens from five experimental runs.

To assess the presence of batch effect in our multi-plate experiment prior to batch effect normalization, we performed t-SNE dimensionality reduction on an equal, random subsample of cells from each batch (Supplementary Figure 6). Since each batch used the same Py2T breast cancer cell line and contained a relatively similar mix of inhibition and control conditions, batches were expected to have more shared than non-shared cell subtypes. If true, this phenomenon would be appear as extensive inter-plate mixing in most regions of the t-SNE cell state space. This is be-

cause most sources of variation in the data were expected to be attributable not to the plate on which specimens were cultured or CyTOF run in which specimens were measured, but instead to specimen-specific biology. Visualizing the t-SNE embedding and coloring cells by their original batch (Supplementary Figure 6A), we noticed poor inter-plate mixing. This indicated that batch effect was present in the unnormalized data.

We then applied CCA to the expression measurements and ran t-SNE on the batch-corrected data (Supplementary Figure 6B). Reassuringly, we noticed that there was strong inter-plate mixing when coloring cells in the t-SNE embedding by their original plate. This suggests that CCA effectively corrected for the technical sources of variation that appeared to be dominating the initial t-SNE embedding based on un-normalized expression data (Supplementary Figure 6A). To assess whether batch effect correction not only removed technical sources of variation but also performed accurate data alignment, we examined the control conditions present on each plate. Two sets of identical control conditions were included on each plate: one set consisted of Py2T epithelial cells cultured with neither TGF-β nor drug inhibitor ("untreated controls"), and the other set consisted of Py2T cells stimulated with TGF-β and given no drug inhibitor ("uninhibited controls"). In our final clustering of specimens, we found that all of the untreated controls from all 5 plates clustered together and consisted almost entirely of the same epithelial cell population. Similarly, all of the uninhibited controls from all 5 plates clustered together and consisted predominantly of late-transitional and mesenchymal cells. Moreover, inhibitors targeting the same molecular target tended to group together, irrespective of batch (e.g. Clusters D, E, F). These findings suggest that CCA accurately aligned the expression data.

Supplementary Note 6: PHATE recovers known EMT cancer cell subtypes

C-1 was characterized by the following expression pattern: E-cadherin$^{(hi)}$ β-catenin$^{(hi)}$ CD24$^{(hi)}$ vimentin$^{(lo)}$ CD44$^{(lo)}$. C-5 and C-6 had roughly the opposite expression profile with respect to the markers described above (Figure 3B). E-cadherin is the hallmark cell adhesion marker of epithelial cells [16], and vimentin and CD44 are known mesenchymal markers involved in cell migration [16–19]. Moreover, recent studies found high CD44:CD24 expression to be indicative of breast cancer cell invasiveness and an as an EMT endpoint, suggestive of mesenchymal properties [20–22]. C-3 was characterized by low-intermediate expression of both E-cadherin and vimentin, and C-4 was characterized by cells with intermediate levels of E-cadherin and vimentin and increased expression of p-MEK1/2, p-ERK1/2, p-p38-MAPK, p-GSK-3β, and p-NFkB-p65. These subtypes were consistent with the "hybrid" cancer cells that co-express epithelial and mesenchymal markers (E+/M+) and simultaneously demonstrate both epithelial and mesenchymal properties [23–25]. Altogether, the subtypes identified by PHATE are consistent with known epithelial, mesenchymal, and "hybrid" EMT cell phenotypes, and the trajectory defined by subtypes C-1 through C-6 in our model represent the epithelial-to-mesenchymal transition process that one would expect to recover in our dataset.

In addition to modeling the main EMT trajectory, the PHATE cell-state embedding identified additional cell subtypes mapped to regions of the cell-state manifold off of the main EMT axis. C-7 and C-8 were mapped close to the C-6 mesenchymal subtype. C-7 was characterized by high expression of vimentin, CD44, cyclin B1, and pRb, and C-8 was characterized by high expression of vimentin, CD44, and phospho-S6. C-9 demonstrated high E-cadherin and cleaved caspase-3 expression and was consistent with an epithelial subpopulation undergoing apoptosis. By analyzing our single-cell data with PHATE, which applied no prior assumptions on the intrinsic geometry

of the cell-state embedding, we were able to uncover a more complex, continuous model of EMT than has been previously reported.

Supplementary Note 7: Analyzing EMT perturbations measured in a single CyTOF batch

To assess whether consistent results were obtained when applying PhEMD to batch-normalized and unnormalized expression data, we performed an analysis of a subset of 60 inhibition and control conditions that were measured in the same CyTOF run. We specifically assessed the consistency of the cell-state and higher-level biospecimen embeddings across experiments.

Cell subtype definition via manifold clustering: Our model of the cell-state space identified 8 unique cell subtypes across all unperturbed and perturbed EMT specimens (Figure S5A-B). These included the starting epithelial subtype (C-1), main mesenchymal subtype (C-5), and transitional subtypes on the major EMT-axis (C-2 through C-4). C-1 was characterized by the following expression pattern: E-cadherin$^{(hi)}$ β-catenin$^{(hi)}$ CD24$^{(hi)}$ vimentin$^{(lo)}$ CD44$^{(lo)}$. C-4 and C-5 had roughly the opposite expression profile with respect to the markers described above (Figure 2C). C-6 through C-8 had expression profiles consistent with C-7 through C-9 in our multi-batch experiment (Figure 3B, Figure S5B). Altogther, the cell subtypes recovered in the single-batch and batch-normalized experiments were consistent with one another and with known EMT cell subtypes.

Note that in order to construct the cell-state manifold more efficiently, it was beneficial to generate the reference cell-state embedding on a subsample of all cells across all single-cell samples (and then to map unembedded cells to cell subtypes using a nearest-neighbor approach). For the analysis of our EMT dataset, we chose to subsample 200 cells from each experimental condition. To assess whether this subsampling procedure had adverse effects on recovering accurate sample-to-sample distances, we first performed such a process on Synthetic Dataset B. We found that the sample-to-sample distances were accurate (Pearson rho$>$99% between computed and ground-truth distances) when subsampling 200 cells from each sample, even when the 200 cells comprised as little as 1% of all cells in each sample. We then assessed whether the subsampling procedure introduced variability into the sample-to-sample distances computed on our EMT dataset by comparing the correlation of results from 20 different random subsamples applied to the same EMT dataset. We found that the correlation between sample-to-sample distances across any two runs was $>$98%. Altogether, these results demonstrated that 200 cells were an adequate subsampling size to yield stable results and that PhEMD was robust to different cell subsamplings.

Constructing and clustering the EMD-based drug-inhibitor manifold: After modeling the EMT cell-state space with PHATE, we used PhEMD to map the experimental variable (i.e., single-cell specimen) state space as a low-dimensional embedding. Specifically, EMD was computed pairwise between specimens based on cell subpopulational differences among samples, and these specimen-to-specimen distances (i.e., measures of dissimlarity) were used to generate a final low-dimensional diffusion map in which specimens mapped closer to one another represented samples with more similar cell subtype relative abundances (Figure S5C). The embedding of drug inhibitors constructed as described above was then partitioned by applying hierarchical clustering to the network of inhibitors. Note that the hierarchical clustering was performed on the EMD-based sample-to-sample distance matrix prior to applying diffusion map dimensionality reduction. Hierarchical clustering revealed clusters of inhibitors with similar net effects on EMT; inhibitors assigned to the

same cluster were assumed to have similar effects on EMT. Moreover, by including "uninhibited" controls (samples in which TGF-β was applied to induce EMT in absence of any inhibitor) and "untreated" controls (samples in which neither TGF-β nor inhibitor was applied and no EMT was induced) in our experiment, we were able to identify inhibitors with notable effects on EMT. Those inhibition conditions that clustered with uninhibited controls likely had little to no effect on EMT, whereas those that clustered with untreated controls halted EMT strongly and likely at an early stage.

The final embedding of drug inhibitors revealed a manifold structure that highlighted the variable extent of EMT that had occurred in the different inhibition conditions (Figure S5C-D). Partitioning the embedding into nine clusters (Clusters A-I, Table S5), we found that Cluster A included the untreated controls and the TGF-β-receptor inhibitor condition, each of which consisted almost entirely of epithelial cells. These were the experimental conditions in which EMT was actually or effectively not induced. On the other hand, Cluster H included all five uninhibited control conditions and inhibitors ineffective at modulating EMT; inhibitors in this cluster were found to have mostly mesenchymal cells. Clusters B through G included inhibitors that had generally decreasing strength with respect to halting EMT (Figure S5C-D). The EGFR and MEK1/2 inhibitors in Clusters B and C strongly inhibited EMT, as indicated by a marked predominance of epithelial cells at time of CyTOF measurement. Cluster G mostly consisted of Aurora kinase inhibitors and was characterized by a mixture of epithelial, transitional, and mesenchymal cells with a relatively high proportion of C-4 cells (consistent with the E+/M+ "hybrid" EMT phenotype). The three inhibitors in Cluster I formed a small branch off the main EMT-extent trajectory in the inhibitor embedding (Figure 2C). These three inhibitors targeted PI3K and mTOR and each demonstrated a cell profile characterized by a relatively high proportion of C-6 cells. Examining these results alongside measurements of cell yield in each inhibition condition (Table S4), we attributed the relatively greater proportion of C-4 cells in the setting of Aurora kinase inhibition and of C-6 cells in the setting of PI3K/mTOR inhibition to preferential drug-induced death of other cell types. C-4 and C-6 cells were not uniquely generated by these inhibition conditions, as they were observed in other samples including the uninhibited EMT control conditions (Figure S5C), but appeared to have increased cell viability relative to other EMT cell types, especially in the setting of targeted kinase inhibition (Table S4). Note that these findings were consistent with those of the multi-batch experiment performed on batch-normalized data.

Supplementary Note 8: Learning a concise representation of network structure among a large set of single-cell specimens

We applied a previously published sampling technique to our PhEMD embedding [26]. The sampling technique used incompletely pivoted QR decomposition to identify "landmark points" (inhibition or control conditions) that approximately spanned the subspace of the single-cell sample embedding (Online Methods). Using this approach, we identified 34 landmark points that summarized our EMT perturbation state space (Figure S7A). The 34 landmark points included samples from all 14 of Clusters A-N, suggesting they spanned all classes of experimental conditions in our experiment. To more fully assess whether the landmark points adequately captured the perturbation landscape of our full 300-sample experiment, we applied an accompanying out-of-sample extension technique to infer the embedding coordinates of all 300 samples relative to these 34 landmark points (Online Methods). The resulting embedding had a similar geometry to that of our original 300-sample PhEMD embedding, suggesting that the 34 landmark points were suffi-

cient to capture the overall network structure of all 300 measured experimental conditions (Figure 3C, Figure S7B). Comparing the pairwise sample-to-sample distances of all 300 samples in the 34-dimensional landmark-point space to the experimentally computed EMD sample-to-sample distances, we found that there was strong correlation between these distances (r=0.92). These findings supported the notion that redundancies may exist in a drug screen experiment, and that one may not need to measure an exhaustive set of perturbation conditions in order to infer the effects of all perturbations.

Supplementary Note 9: Predicting the effects of three selected inhibitors on breast cancer EMT relatively to the effects of measured inhibitors based on known drug-target binding specificities

We sought to evaluate whether we could leverage known information on the mechanistic similarity between our inhibitors and additional inhibitors not measured in our experiment to predict the effects of these additional inhibitors on EMT. We selected saracatinib, ibrutinib, and dasatinib as three nonspecific Src inhibitors whose effects on EMT we wanted to predict. First, we generated a PhEMD embedding based on our CyTOF experimental results (not including the three selected inhibitors). Then, we obtained drug-target specificity data from a recently published inhibitor-profiling experiment [27] for inhibitors that overlapped between our experiment and the recently published one (including the 3 Src inhibitors of interest). We used the drug-target specificity data to compute pairwise cosine similarities between each of the 3 Src inhibitors and the samples in our initial PhEMD diffusion map embedding (that did not include the 3 inhibitors) (Online Methods). These pairwise similarities were used to perform Nystrom extension—a method of extending a diffusion map embedding to include new points based on partial affinity to existing points. In this way, we were able to predict the effects of the three Src inhibitors on breast cancer EMT relatively to inhibitors with known, measured effects (Online Methods).

To validate our extended embedding containing predicted Src inhibitor effects, we compared it to a "ground-truth" diffusion map embedding that used known (measured) CyTOF expression data for the 3 inhibitors and explicitly included the 3 inhibitors along with the rest in the initial embedding construction. Benchmarking our predictions against this ground-truth model, we found that our predictive model mapped the three inhibitors to the correct phenotypic space (Figure 4A-B). Specifically, saracatinib and ibrutinib were predicted to have an effect intermediate to those of specific MEK and EGFR inhibitors, and dasatinib was predicted to halt EMT less strongly than the other two Src inhibitors. These findings are consistent with ground-truth results based on direct CyTOF profiling and PhEMD-modeling of the three inhibitors (Figure 4B; Online Methods).

We also hypothesized that we could use drug-target information to not only relate unmeasured inhibitors to measured ones but also impute their single-cell compositions. To test this, we used the Nystrom-extended PhEMD embedding as input into a partial least squares regression model. We used this model to impute the cell subtype relative frequencies for the three unmeasured (imputed) Src inhibitors (Online Methods). As validation, we compared the predicted cell subtype relative frequencies to ground-truth CyTOF results (i.e., actual single-cell measurements) for the three inhibitors. PhEMD accurately predicted the cell subtype relative frequencies for the three inhibitors compared to the null model (P=0.01, P=0.01, P=0.03; Figure 4C).

Supplementary Note 10: Leveraging all cells using PHATE with nearest-node mapping

Unique cell subtypes are identified by running PHATE on an equal subsampling of cells from each experimental condition (i.e., single-cell specimen). Through this step, each of the subsampled

cells from each inhibition condition is assigned to a specific subtype. Note that cells that were not initially subsampled still lack a subtype assignment. To incorporate all cells into our analysis, we next iterate through the entire set of cells (including cells not initially subsampled) and map each to a subtype based on a nearest-neighbor approach. Note that each cell subtype detected by PHATE is defined as a cluster of cells with similar gene-expression profiles. To assign cell x, which is not initially included in the construction of the PHATE cell-state embedding, to a cell subtype, we first identify cell y in the initial embedding that was most similar to cell x, i.e. the cell with the lowest Euclidean distance from cell x. Cell x is then given the same cell subtype assignment as cell y. The end result is that each cell is assigned to a specific cell subtype.

References

[1] Lahav, G. *et al.* Dynamics of the p53-mdm2 feedback loop in individual cells. *Nature Genetics* **36**, 147–150 (2004).

[2] Young, K. J., Kay, L. S., Phillips, M. J. & Zhang, L. Antitumor activity mediated by double-negative t cells. *Cancer Research* **63**, 8014–8021 (2003). PMID: 14633734.

[3] Overgaard, N. H., Jung, J.-W., Steptoe, R. J. & Wells, J. W. Cd4 $^{+}$ /cd8 $^{+}$ double-positive t cells: more than just a developmental stage? *Journal of Leukocyte Biology* **97**, 31–38 (2015).

[4] Bellman, R. E. *Dynamic Programming* (Princeton University Press, Princeton, NJ, 1957).

[5] Bellman, R. E. & Dreyfus, S. E. *Applied Dynamic Programming* (Princeton University Press, Princeton, NJ, 1962).

[6] Moon, K. R. *et al.* Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Current Opinion in Systems Biology* **7**, 36–46 (2018).

[7] Arjovsky, M., Chintala, S. & Bottou, L. Wasserstein generative adversarial networks. In Precup, D. & Teh, Y. W. (eds.) *Proceedings of the 34th International Conference on Machine Learning*, vol. 70 of *Proceedings of Machine Learning Research*, 214–223 (PMLR, International Convention Centre, Sydney, Australia, 2017).

[8] Edwards, D. On the kantorovich–rubinstein theorem. *Expositiones Mathematicae* **29**, 387 – 398 (2011).

[9] Kantorovich, L. V. & Rubinstein, G. S. On a space of completely additive functions. *Vestnik Leningradskogo Universiteta* **13**, 52–59 (1958).

[10] Orlova, D. Y. *et al.* Earth Mover's Distance (EMD): A True Metric for Comparing Biomarker Expression Levels in Cell Populations. *PLOS ONE* **11**, e0151859 (2016).

[11] Rubner, Y., Tomasi, C. & Guibas, L. J. The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision* **40**, 99–121 (2000).

[12] Moon, K. R. *et al.* Visualizing transitions and structure for high dimensional data exploration. *bioRxiv* (2017). [Online; accessed 2018-06-22].

[13] Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology* **36**, 421–427 (2018).

[14] Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**, 411–420 (2018).

[15] Shaham, U. *et al.* Removal of batch effects using distribution-matching residual networks. *Bioinformatics* **33**, 2539–2546 (2017).

[16] Mani, S. A. *et al.* The Epithelial-Mesenchymal Transition Generates Cells with Properties of Stem Cells. *Cell* **133**, 704–715 (2008).

[17] Zhu, H. *et al.* The role of the hyaluronan receptor CD44 in mesenchymal stem cell migration in the extracellular matrix. *Stem Cells* **24**, 928–935 (2006).

[18] L Ramos, T. *et al.* MSC surface markers (CD44, CD73, and CD90) can identify human MSC-derived extracellular vesicles by conventional flow cytometry. *Cell communication and signaling* **14**, 2 (2016).

[19] Ivaska, J., Pallari, H.-M., Nevo, J. & Eriksson, J. E. Novel functions of vimentin in cell adhesion, migration, and signaling. *Experimental Cell Research* **313**, 2050–2062 (2007).

[20] Li, W. *et al.* Unraveling the roles of CD44/CD24 and ALDH1 as cancer stem cell markers in tumorigenesis and metastasis. *Scientific Reports* **7**, 13856 (2017).

[21] Ma, F. *et al.* Enriched CD44(+)/CD24(-) population drives the aggressive phenotypes presented in triple-negative breast cancer (TNBC). *Cancer Letters* **353**, 153–159 (2014).

[22] Ricardo, S. *et al.* Breast cancer stem cell markers CD44, CD24 and ALDH1: expression distribution within intrinsic molecular subtype. *Journal of Clinical Pathology* **64**, 937–946 (2011).

[23] Yu, M. *et al.* Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *Science* **339**, 580–584 (2013).

[24] Nieto, M., Huang, R.-J., Jackson, R. & Thiery, J. EMT: 2016. *Cell* **166**, 21–45 (2016).

[25] Jolly, M. K. *et al.* Implications of the Hybrid Epithelial/Mesenchymal Phenotype in Metastasis. *Frontiers in Oncology* **5**, 155 (2015).

[26] Salhov, M., Bermanis, A., Wolf, G. & Averbuch, A. Approximately-isometric diffusion maps. *Applied and Computational Harmonic Analysis* **38**, 399–419 (2015).

[27] Klaeger, S. *et al.* The target landscape of clinical kinase drugs. *Science* **358**, eaan4368 (2017).