January 2020

Uncovering Intratumoral And Intertumoral Heterogeneity Among Single-Cell Cancer Specimens

William Shelton Chen

Recommended Citation

Chen, William Shelton, "Uncovering Intratumoral And Intertumoral Heterogeneity Among Single-Cell Cancer Specimens" (2020). *Yale Medicine Thesis Digital Library*. 3890.
https://elischolar.library.yale.edu/ymtdl/3890

Uncovering Intratumoral and Intertumoral Heterogeneity Among Single-Cell Cancer Specimens

A Thesis Submitted to the Yale University School of Medicine
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Medicine

by
William S. Chen
2020

UNCOVERING INTRATUMORAL AND INTERTUMORAL HETEROGENEITY AMONG SINGLE-CELL CANCER SPECIMENS

William S Chen, Nevena Zivanovic, David van Dijk, Guy Wolf, Bernd Bodenmiller, and Smita Krishnaswamy. Department of Genetics, Yale University, School of Medicine, New Haven, CT.

ABSTRACT

While several tools have been developed to map axes of variation among individual cells, no analogous approaches exist for identifying axes of variation among multicellular biospecimens profiled at single-cell resolution. Developing such an approach is of great translational relevance and interest, as single-cell expression data are now often collected across numerous experimental conditions (e.g., representing different drug perturbation conditions, CRISPR knockdowns, or patients undergoing clinical trials) that need to be compared. In this work, “Phenotypic Earth Mover's Distance” (PhEMD) is presented as a solution to this problem. PhEMD is a general method for embedding a “manifold of manifolds,” in which each datapoint in the higher-level manifold (of biospecimens) represents a collection of points that span a lower-level manifold (of cells).

PhEMD is applied to a newly-generated, 300-biospecimen mass cytometry drug screen experiment to map small-molecule inhibitors based on their differing effects on breast cancer cells undergoing epithelial–mesenchymal transition (EMT). These experiments highlight EGFR and MEK1/2 inhibitors as strongly halting EMT at an early stage and PI3K/mTOR/Akt inhibitors as enriching for a drug-resistant mesenchymal cell subtype characterized by high expression of phospho-S6. More generally, these experiments reveal that the final mapping of perturbation conditions has low intrinsic dimension and that the network of drugs demonstrates manifold structure, providing insight into how these single-cell experiments should be computational modeled and

visualized. In the presented drug-screen experiment, the full spectrum of perturbation effects could be learned by profiling just a small fraction (11%) of drugs. Moreover, PhEMD could be integrated with complementary datasets to infer the phenotypes of biospecimens not directly profiled with single-cell profiling. Together, these findings have major implications for conducting future drug-screen experiments, as they suggest that large-scale drug screens can be conducted by measuring only a small fraction of the drugs using the most expensive high-throughput single-cell technologies—the effects of other drugs may be inferred by mapping and extending the perturbation space.

PhEMD is also applied to patient tumor biopsies to assess intertumoral heterogeneity. Applied to a melanoma dataset and a clear-cell renal cell carcinoma dataset (ccRCC), PhEMD maps tumors similarly to how it maps perturbation conditions as above in order to learn key axes along which tumors vary with respect to their tumor-infiltrating immune cells. In both of these datasets, PhEMD highlights a subset of tumors demonstrating a marked enrichment of exhausted CD8$^+$ T-cells. The wide variability in tumor-infiltrating immune cell abundance and particularly prominent exhausted CD8$^+$ T-cell subpopulation highlights the importance of careful patient stratification when assessing clinical response to T cell-directed immunotherapies.

Altogether, this work highlights PhEMD's potential to facilitate drug discovery and patient stratification efforts by uncovering the network geometry of a large collection of single-cell biospecimens. Our varied experiments demonstrate that PhEMD is highly scalable, compatible with leading batch effect correction techniques, and generalizable to multiple experimental designs, with clear applicability to modern precision oncology efforts.

Published in part:
Chen WS*, Zivanovic N*, van Dijk D, Wolf G, Bodenmiller B, Krishnaswamy S. Uncovering axes of variation among single-cell cancer specimens. *Nature Methods*, 2020.

Presented in part:
Chen WS, Zivanovic N, Pe'er D, Bodenmiller B, Krishnaswamy S. Phenotypic analysis of single-cell breast cancer inhibition data reveals insights into EMT. AACR Annual Meeting, Washington, DC, Apr 2017.

ACKNOWLEDGEMENTS:

I would like to acknowledge the Krishnaswamy and Bodenmiller laboratories for thought-provoking and productive discussions. I am especially appreciative of Prof. Smita Krishnaswamy for her incredible mentorship and support. I am also indebted to Nevena Zivanovic and Bernd Bodenmiller for their help generating and interpreting much of the single-cell data presented in this work.

This work was supported in part by the Chan–Zuckerberg Initiative Seed Networks for the Human Cell Atlas (S.K.), a Swiss National Science Foundation (SNSF) R'Equip grant (B.B), a SNSF Assistant Professorship grant PP00P3-144874 (B.B.), the SystemsX Transfer Project "Friends and Foes" (B.B.), the SystemX grants Metastasix and PhosphoNEtX (B.B.), the European Research Council (ERC) under the European Union's Seventh Framework Program (FP/2007-2013)/ERC Grant Agreement 336921 (B.B.), the CRUK IMAXT Grand Challenge (B.B.), and the following National Institutes of Health (NIH) grants: R01GM135929 (S.K, G.W.), UC4 DK108132 (B.B.), NIH–NIDDK T35DK104689 (W.C.).

Table of Contents

INTRODUCTION

Bulk vs. single-cell profiling

Next-generation sequencing (NGS) has revolutionized the way in which diseases can be studied. Bulk DNA sequencing (DNA-seq) of germline biospecimens can be leveraged to discover disease-specific polymorphisms and to investigate disease heritability at an unprecedented scope and level of detail (1–3). In the setting of cancer, bulk DNA-seq of liquid- or solid-tumor biopsies has been used to identify somatic alterations (e.g., mutations, copy number alterations, and structural variants) that can serve as biomarkers prognostic of clinical outcomes and predictive of response to therapies (4–9). Complementarily, bulk RNA-sequencing (RNA-seq) has been used to quantitate gene expression of protein-coding genes and long non-coding RNAs at the exon level of resolution. Paired with proteomic assays, NGS approaches have facilitated our understanding of cellular biology and genomic drivers of disease at all steps of the central dogma, from DNA to RNA to protein.

While instrumental in building our foundational understanding of cancer genomics, bulk tumor profiling faces the notable limitation of being unable to resolve intratumoral heterogeneity. By nature of the sample preparation procedure for bulk NGS, DNA or RNA fragments are isolated from all cells of a biospecimen in aggregate, and per-cell read counts cannot be determined. Thus, genomic variants identified via bulk DNA-seq can only be interpreted as being present in some fraction of profiled cells. Moreover, it is impossible to determine which of the variants co-occur in a given cancer cell. The readout of bulk RNA-seq is similarly coarse in that the reported expression of a given gene represents the average expression across all cells in the biospecimen without any consideration of cell-to-cell variation. In practice, when comparing expression values across biospecimens measured using bulk profiling or when performing association studies between specific DNA variants and clinical phenotypes, a simplifying assumption is often made that all (or at least a substantial-enough proportion of) cells in each biospecimen harbor the genomic variant or gene expression signature of interest. In reality, this assumption may not always be valid, and bulk measurements may fail to accurately reflect the expression profiles of individual cells. Bulk profiling may also fail to detect true biological differences between experimental conditions. The following example demonstrates these concepts more concretely and highlights the utility of single-cell analytical approaches for accurately characterizing and distinguishing between multicellular biospecimens.

Consider a multi-specimen dataset consisting of immune cells with collectively variable expression of CD4 and CD8. Each specimen is comprised of a cell population that fits one of four distribution patterns, as shown below (Figure 1A). Each Group A specimen consists of a homogeneous immune cell population characterized by intermediate expression of both CD4 and CD8. Each Group B specimen consists of two similarly-abundant immune cell subpopulations: one CD4$^+$ and one CD8$^+$ subpopulation. Group C specimens consist of a mixture of CD4$^+$, CD8$^+$, and CD4/CD8 double-positive (DP) immune cells. Group D specimens consist of one CD4$^+$ and one CD8$^+$ subpopulation of roughly equal abundance 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 (10, 11)). The simulated experiment consists of 32 specimens in total (eight of each of Groups A-D). By design, the bulk (average) expression of CD4 and CD8 for each biospecimen is roughly the same for all biospecimens, regardless of differences in cell subpopulation characteristics.

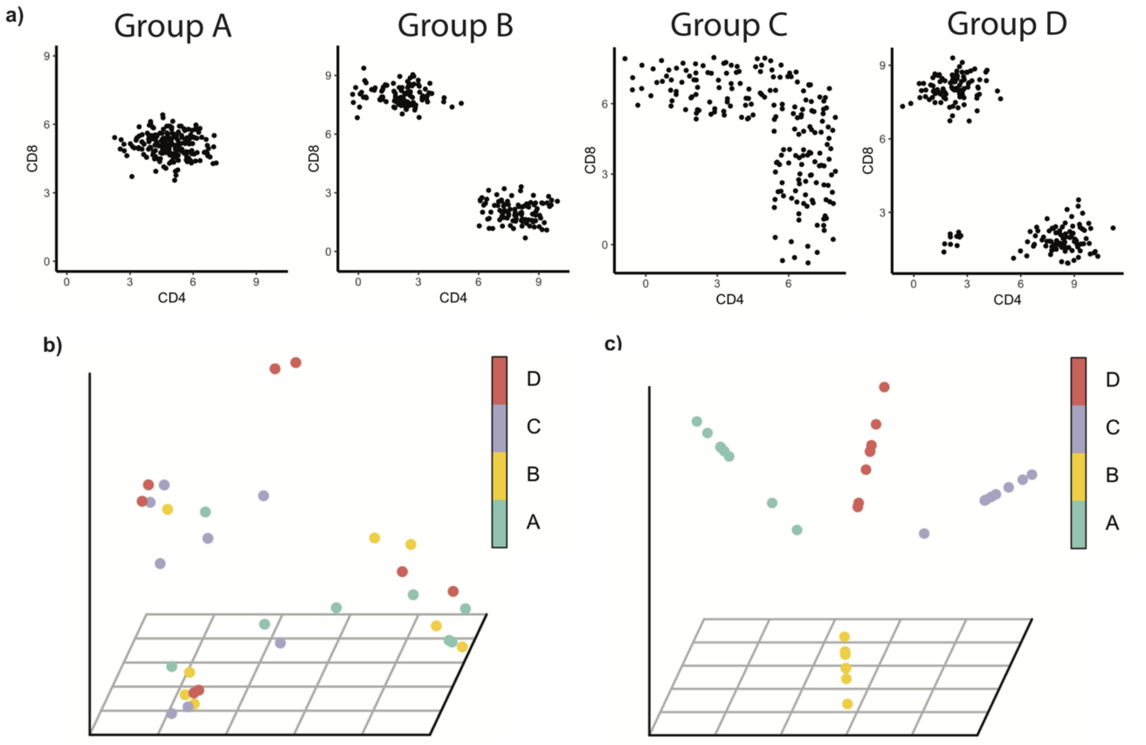


Figure 1. a) Single-cell profiles of each multicellular biospecimen in a computationally-generated immune cell dataset. Each point represents a single cell. Groups A-D each have 8 biospecimens that fit the single-cell profile (i.e., are comprised of some combination of the cell subpopulations depicted) for a total of 32 biospecimens. By design, all biospecimens have roughly the same bulk expression (mean across all cells) of CD4 and CD8. b) Diffusion map embedding generated by embedding a specimen-to-specimen distance matrix, where pairwise distances between specimens were computed by taking the Euclidean distance between specimens represented as bulk expression of CD4 and CD8. Bulk expression profiles do not adequately reflect the biological differences between specimens in this dataset and cannot be used to distinguish specimens in a biologically meaningful way. c) Diffusion map embedding generated by embedding a PhEMD distance matrix, which accounts for the single-cell characteristics of each specimen (see "Overview of PhEMD" in Results section). PhEMD successfully distinguishes specimens with different single-cell profiles from one another.

Next, consider the aim of relating the 32 specimens 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 specimens relative to the rest and to identify groups of similar biospecimens. First, consider an approach using bulk expression measurements. A biospecimen–biospecimen distance matrix can be generated by computing pairwise (Euclidean) distances between each pair of biospecimens, with each biospecimen represented as the average expression of each gene (i.e., CD4 and CD8) across all cells in the biospecimen. This distance matrix can then be embedded and visualized in two dimensions using the diffusion map nonlinear dimensionality reduction approach. The result is an embedding that fails to differentiate specimens based on biologically important differences. Specifically, specimens of the same known, ground-truth subtype (i.e., Group A-D) failed to map to similar parts of the resulting embedding (Figure 1B).

A better approach to comparing these specimens is to compare the presence and abundance of all single-cell subpopulations in each specimen. I aim to formalize such an approach in this thesis and demonstrate that it can be used to effectively distinguish single-cell specimens from one another that cannot be distinguished based on bulk or average expression patterns. In the above example, the approach yields a final low-dimensional map that vastly outperforms a bulk approach (Figure 1B) and successfully differentiates specimens based on biologically important differences in cell subpopulation characteristics and proportions (Figure 1C).

The exploration of cell-to-cell variation within a given biospecimen has been facilitated by the recent development of single-cell expression profiling (measurement of gene expression on a *per-cell* rather than *average-across-all-cells* level). Early studies leveraging these technologies have uncovered important insights not previously identified by bulk profiling. Several studies have highlighted the compositional heterogeneity of tumors as a mixture of specific cancer and non-malignant (e.g., immune and stromal) cell types and have revealed profound cellular heterogeneity among melanoma (12), clear-cell renal cell carcinoma (ccRCC) (13), and breast cancer cells (14), even within a single tumor biopsy. Additional studies have used single-cell profiling to better elucidate cell signaling, differentiation, and reprogramming in the context of cancer (15), aging (16), and other physiologic and disease processes (17–19). Among else, single-cell profiling is particularly useful for studying cancer, as cancer is understood to arise from the genomic mis-programming of a single cell and the downstream sequelae. While the analytical approaches presented in this work are generalizable to studying many biological phenomena at a single-cell level, the focus of this thesis will be on leveraging single-cell technologies to better understand cancer progression, cellular response to chemotherapies, and the tumor microenvironment.

Approaches to characterizing axes of variation among a collection of cells

As the readout of single-cell expression profiling is highly complex, new computational tools have been developed in parallel with single-cell profiling techniques in order to facilitate the extraction of biological insights. A particularly challenging property of single-cell expression data is its high dimensionality: each biospecimen is comprised of many cells, each of which is represented by tens to thousands of gene or protein measurements. The analysis of high-dimensional data, especially in unsupervised or exploratory settings, often introduces various challenges that are collectively referred to as the “curse of dimensionality” (20, 21). While the ambient dimension of single-cell data is often high (equal to the number of genes or proteins measured), the intrinsic dimension, or minimum number of variables needed to represent the data adequately, is often much lower. Mapping single-cell data from its ambient dimension to this lower-dimensional space is termed “dimensionality reduction,” which is often a critical first step for learning and visualizing the ways in which cells vary. It is also instrumental in identifying distinct, biologically meaningful cell subpopulations (e.g., by clustering cells in the lower-dimensional space). The following subsections provide an overview of several of the leading dimensionality reduction techniques for learning and visualizing axes of cell-to-cell variation among a set of cells measured using single-cell expression profiling. In the below subsections, each “dataset” refers to a heterogeneous cell population and each “point” represents a single cell, characterized by multiple measured features (i.e., gene expression values).

Principal Component Analysis (PCA)

Principal component analysis (PCA) is a dimensionality reduction technique that aims to find new, uncorrelated variables (“principal components”) that successively maximize variance while minimizing information loss from the original dataset (22). It does so by performing an orthogonal transformation of the original dataset such that the new variables are linear combinations of features in the original ambient-dimensional space. This transformation can be computed as the solution to an eigenvalue/eigenvector problem, in which principal components are defined as eigenvectors of the covariance matrix and corresponding eigenvalues represent the proportion of the data variance explained by the eigenvectors.

PCA is useful in many settings for learning a low-dimensional representation of the data, although it does make several key assumptions. Firstly, it assumes that the principal components are appropriately modeled as linear combinations of the original dimensions. Secondly, PCA assumes that principal components are orthogonal to one another. Thirdly, PCA assumes that the input data are scaled and normalized appropriately prior to application, as the approach is not scale invariant. In the event that any of these assumptions are violated, PCA may fail to recover optimal axes of variation in the data. Additionally, by design, PCA prioritizes preserving global structure (i.e., distances between faraway points) over local structure (i.e., distances between points within the same "neighborhood") when mapping from high- to low-dimensional space. Thus, the approach is especially sensitive to outliers and measurement noise.

t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a popular dimensionality reduction approach that preserves local relationships between points when mapping them from an ambient-dimensional to low-dimensional space (23). Put another way, points that are close to one another in the original representation of the data are mapped to be close to one another in the final low-dimensional t-SNE space. Consider two points *i* and *j* denoted by x_i and x_j respectively in the ambient-dimensional space and y_i and y_j respectively in the low-dimensional space. t-SNE first models each point in the ambient-dimensional space as a Gaussian probability distribution centered on the actual coordinates of the point (with a data-dependent variance proportional to a user-specified "perplexity" value), then computes pairwise similarity between points x_i and x_j as the conditional probability $P_{j|i}$ that x_i would select x_j as its neighbor if neighbors were selected in proportion to their probability density under a Gaussian centered at x_i (24). An analogous conditional probability $Q_{j|i}$ is computed between points y_i and y_j in the low-dimensional t-SNE space, wherein points yi and y_j are modeled as Student *t*-distribution with one degree of freedom (rather than as a Gaussian distribution). The final low-dimensional embedding is learned through gradient descent by minimizing the Kullback-Leibler (KL) divergence between P and Q. In the perfect case, the difference between $P_{j|i}$ and $Q_{j|i}$ is zero for all *i* and *j*, i.e., the pairwise relationships between points are perfectly preserved in the ambient and t-SNE dimensions.

Strengths of t-SNE include non-linearity, which renders it superior to linear approaches such as PCA when applied to curved manifolds, and preservation of local structure, which reveals subtle differences between similar yet distinct cell subpopulations. A limitation of t-SNE is its high computational resource demands. In its exact form, t-SNE has a quadratic time and space complexity, making applications to datasets larger than 10,000 points often computationally intractable. To mitigate this issue, various approximations and optimizations have been developed (25–27). Another limitation is the loss of global structure preservation in the final embedding. t-SNE effectively identifies neighborhoods of points but generally yields disjoint "clouds" of points. Thus, continuous (e.g. cellular differentiation) processes and trajectories are often fragmented in the t-SNE embedding, and relative distances between faraway points or clusters in the embedding are not preserved (28).

Uniform Manifold Approximation and Projection (UMAP)

Uniform Manifold Approximation and Projection (UMAP) is a dimensionality reduction approach that has been recently popularized due to its purported advantages over t-SNE in terms of improved scalability and preservation of both local and global data structure (29). Similarly to t-SNE, UMAP models points as probability distributions and performs gradient descent to iteratively "move" points more similar to one another in the ambient-dimensional space to be closer to one another in the low-dimensional embedding.

However, among other minor differences, UMAP omits the normalization of probabilities used in t-SNE (thus improving runtime) and uses binary cross-entropy instead of KL divergence as the cost function when comparing relationships between points in the ambient-dimensional and low-dimensional spaces. UMAP also employs a graph Laplacian approach to assigning the initial coordinates of the points in low-dimensional space (prior to the first iteration of gradient descent), in contrast to the random initialization employed by t-SNE. Early studies have claimed advantages of UMAP over t-SNE in terms of faster computational runtime, greater preservation of global data structure, and increased reproducibility of results across different iterations. However, there is ongoing debate as to whether global structure is truly better preserved using UMAP than t-SNE and if so, why exactly this may be (30).

Tree-based approaches

Several graph-based approaches have been developed to explicitly model single-cell expression datasets as an interconnected "web" or "tree" of cells. Particularly aimed at organizing and visualizing data with intrinsic trajectory structure (e.g., bifurcating differentiation processes), these approaches typically represent cells as nodes and relationships between similar cells as edges between nodes. Distances between cells can then be defined as the shortest path between representative nodes (i.e., minimum number of edges separating one cell from the other or, in the event of weighted edges, minimum sum of edge lengths in a path from one cell to the other). Several particular single-cell tools that employ such an approach include SPADE (31), Wishbone (32), and Monocle2 (33). These approaches are particularly useful for modeling continuous manifolds and for resolving local neighborhood structure. A key limitation of most tree-based approaches is poor scalability. In practice, when working with large datasets, these approaches often require cell subsampling or prior identification of “landmark points” which may then collectively comprise a relatively small number of graph nodes. Additionally, the number of branches recovered in the final tree can vary greatly and is often dependent on user-defined parameters, which may be challenging to tune if the expected number of branches is not known *a priori*.

Diffusion maps

Diffusion maps are another nonlinear dimensionality reduction technique based on the idea that a collection of points (e.g., cells) may be modeled such that a given point (e.g., cell) may “transition” to another point (e.g., similar cell state) with a probability proportional to the known similarity of the two points (34). Diffusion maps first model points as an interconnected graph, with connectivity between points generally based on their distance in the ambient-dimensional space (e.g., Gaussian kernel, which prioritizes preservation of local neighborhood structure). The point-to-point connectivity metric is then used to represent the probability of “transitioning” from one cell to another in one step of a random walk. A diffusion process is then performed over a diffusion time *t* (i.e., *t*-step random walk), wherein the local connectivity of the data is used to reveal the global geometric structure of the data. The end result is a set of *t*-step transition probabilities, which can be used to embed a low-dimensional map that captures both local and global structure in the data. Diffusion maps are particularly well-suited for modeling single-cell datasets with known trajectory structure and are modeled on underlying principles that reflect our intuitive understanding of cellular differentiation processes (e.g., “transition” from one cell state to the next). Diffusion maps are also attractive for their nonlinearity and inherent denoising properties. Limitations of diffusion maps include high computational runtime and sensitivity to scale parameter σ, which determines the scale at which the data are visualized (35). In the traditional implementation of diffusion maps, a fixed σ is used for all points in the dataset, often imposing a tradeoff between preserving global and local structure with a bias toward preserving global structure (34, 35). However, subsequent adaptations to the original implementation proposed by Coifman and Lafon have been developed to better preserve local structure (36).

PHATE

Potential of Heat-diffusion for Affinity-based Transition Embedding (PHATE) is a nonlinear dimensionality reduction technique that aims to preserve both local and global structure when mapping from high- to low-dimensional space (37). Similarly to diffusion maps, PHATE models cell-to-cell connectivity as one-step transition probabilities in a random walk model and then performs the diffusion process over diffusion time t to determine t-step transition probabilities. However, PHATE employs a distance metric ("potential distance") between points in the diffusion space distinct from the "diffusion distance" metric used in diffusion maps. In so doing, PHATE better preserves both local and global data geometry and yields a more stable embedding than diffusion maps (37).

Characterizing axes of variation among a collection of multicellular cancer specimens

The multitude of dimensionality reduction techniques described above have been adopted and adapted to elucidate clusters, patterns, and progressions from high-dimensional single-cell data. These techniques all rely on the ability to create a geometry from datapoints by comparing them on the basis of their features. More specifically, these techniques often compute a distance between the datapoints (i.e., cells) in order to organize cells into lower-dimensional embeddings, such as diffusion maps or t-SNE embeddings, in order to extract biologically meaningful clusters (i.e., cell subtypes) or trajectories (e.g., cell differentiation pathways) from the data.

However, single-cell experimental designs are becoming increasingly complex, with data now often collected across numerous experimental conditions to characterize libraries of drugs, pools of CRISPR knockdowns, or groups of patients undergoing clinical trials (12, 13, 38–42). The challenge in these experiments is to characterize the ways in which not only individual cells but also multicellular experimental conditions vary. Comparing single-cell experimental conditions (e.g., distinct perturbation conditions or patient samples) is challenging, as each condition is itself high-dimensional, comprised of a heterogeneous population of cells with each cell characterized by many gene measurements. Each "datapoint" in such settings should then be a patient or an experimental condition, which has a *collection* of measurements associated with it instead of a single measurement. In this setting, a datapoint is no longer a single set of features (i.e., a vector) but a collection of observations, each containing its own features (i.e., a two-dimensional matrix). Thus, existing techniques can no longer be directly applied for analysis in a straightforward manner.

While two prior studies presented approaches to comparing *two* single-cell biospecimens (43, 44), no existing methods address the problem of simultaneously relating *many* biospecimens and identifying biologically meaningful ways in which they differ. In this work, Phenotypic Earth Mover's Distance "*PhEMD*" is presented as a novel "manifold-of-manifolds" approach to mapping the key axes of variation among a large set of experimental conditions. PhEMD first leverages the observation that the structure of a single-cell experimental condition (i.e., multicellular biospecimen) can be well represented as a low-dimensional manifold (i.e., cell-state embedding) using techniques such as PHATE or diffusion maps. In this first-level manifold, individual datapoints represent cells, and distances between cells represent cell-to-cell dissimilarity. PhEMD models the cellular state space of each experimental condition as a "low-level" manifold and then models the experimental condition state space as a "higher-level" manifold. The ultimate goal of PhEMD is to generate this higher-level manifold, in which each datapoint represents a distinct experimental condition and distances between points represent biospecimen-to-biospecimen dissimilarity. In this work, the properties and potentially applications of this final higher-level manifold are explored in depth. Namely, the manifold can be visualized and clustered to reveal the key axes of variation among a large set of experimental conditions. Such embeddings can also be extended with additional data sources to impute experimental conditions not directly measured with single-cell technologies.

The accuracy of PhEMD is first validated on a synthetic dataset with known underlying data geometry. PhEMD is then applied to a newly-generated single-cell dataset to reveal insights into cancer progression and cancer drug-perturbation effects. Specifically, the dataset represents a large perturbation screen performed on breast cancer cells undergoing TGFβ_1-induced epithelial-to-mesenchymal transition (EMT), measured at single-cell resolution with mass cytometry. EMT is a process that is thought to play a role in cancer metastasis, whereby polarized epithelial cells within a local tumor undergo specific biochemical changes that result in cells with increased migratory capacity, invasiveness, and other characteristics consistent with the mesenchymal phenotype (45). In the drug-screen experiment, each perturbation condition consists of cells from the Py2T breast cancer cell line stimulated simultaneously with TGF-β_1 (to undergo EMT) and a unique kinase inhibitor, with the ultimate goal being to compare the effects of different inhibitors on our model EMT system. PhEMD is used to embed the space of the kinase inhibitors to reveal the main axes of variation among all inhibitors with respect to their effects on breast cancer cells undergoing EMT. Reproducibility of results is assessed through three biological replicates. Additionally, the drug-effect findings are further validated by showing that they are consistent with the drug-effect findings of a previously published study that profiled the drug-target binding specificities of several of the same drugs as those used in the present drug-screen experiment.

PhEMD is also applied to two distinct single-cell datasets to reveal insights into variation in the immune cell infiltrate of solid tumors profiled with single-cell resolution. The first dataset consists of a collection of 17 melanoma samples profiled using single-cell RNA-sequencing (scRNA-seq) and the second is comprised of 75 clear-cell renal cell carcinoma (ccRCC) samples profiled using mass cytometry. These experiments yield a low-dimensional map of patient tumors that highlights profound inter-tumoral heterogeneity with respect to tumor-infiltrating immune cells, demonstrating the potential utility of PhEMD for disease subtyping and patient stratification (e.g., for immunotherapy trials or clinical outcomes studies). Collectively, the analyses present a new generalizable analytical framework for organizing single-cell data and reveal new potential strategies for identifying effective cancer therapies.

Hypothesis

A “manifold-of-manifolds” approach to modeling multi-specimen single-cell data can accurately identify axes of variation among biospecimens and simultaneously reveal insights into both intra- and inter-specimen heterogeneity.

Specific Aims

Aim 1: *Develop a robust tool for uncovering axes of variation among single-cell biospecimens*

Aim 2: *Characterize the differing effects of 233 small-molecule inhibitors on breast cancer epithelial–mesenchymal transition (EMT)*

Aim 3: *Characterize the immune cell subpopulational variation among melanomas and among clear-cell renal cell carcinomas (ccRCCs)*

MATERIALS AND METHODS

The PhEMD analytical approach

In single-cell data, each cell is characterized by a set of features, such as protein or transcript expression levels of genes. The purpose of measuring these expression-based features for each cell (e.g., via single-cell RNA-seq or mass cytometry) is to answer biological questions especially related to the cell subpopulations present in a biospecimen. In particular, the features may be used for defining phenotypes of cells (38, 46), resolving cellular dynamics using transition-process modeling (32, 33, 47), and studying signaling networks (18, 48). In sum, the features are shared, quantitative characteristics of cells that may be used to organize a set of cells into a data geometry. An analogy can be made when attempting to compare single-cell specimens rather than individual cells. A biospecimen is a collection of cells. In order to compare single-cell biospecimens for the purpose of organizing a set of cell collections (e.g., different patient specimens or perturbation conditions), one must first determine useful features for a cell collection. Previous studies have shown that cell subtypes are highly useful features that are shared across all specimens and can be quantitatively measured (46, 49). Moreover, they can be used to represent single-cell specimens efficiently for downstream analyses. Just as transcript counts can be measured for selected genes in a single cell, so can cell counts be measured for selected cell subtypes in a cell collection.

In the present work, PHATE is used for the task of defining cell subtypes (37). PHATE is a diffusion-based single-cell dimensionality reduction technique that both identifies unique cell subpopulations and relates them to one another on a low-dimensional manifold that can be visualized. Of note, PHATE preserves an information theoretic distance between points (i.e., cells) in the diffusion space to derive a stable low dimensional embedding that reveals local, global, continual, and discrete non-linear structures in single-cell data. By applying PHATE to an aggregate of cells in a single-cell experiment, we can represent a biospecimen as the relative frequency of cells in each cell subtype. This representation of single-cell specimens is consistent with the "signatures-and-weights" representation of multidimensional distributions, first formalized by Rubner et al. (50), that was found to yield optimal data representation efficiency in other computer vision applications. In our case, a "signature" can be thought of as a distinct cell subtype (e.g., memory B-cells or CD8+ effector T-cells), and the corresponding "weight" represents the proportion of cells in a given specimen assigned to the cell subtype. However, comparing single-cell specimens represented as such is still a non-trivial task. Many studies represent single-cell specimens as their cell subtype composition and use known class labels (e.g., normal lung vs. lung adenocarcinoma) to group specimens and perform class-based comparisons (e.g., identifying cell subtypes enriched in a disease state) (39, 40). However, this approach is limited to comparing a few predefined classes of specimens and does not reveal insights into intra-class heterogeneity. Other studies organize a set of many single-cell specimens based on their relative frequency of one or a few important cell subtypes (41, 46, 51). However, this approach requires *a priori* knowledge of the most important cell subtypes and does not provide a complete view of specimen-to-specimen dissimilarity, especially in the context of high intra-specimen cellular heterogeneity.

The ideal metric for comparing specimens should take into account both the difference in weights of matching bins (e.g., number of epithelial cells) for all bins and the dissimilarity of the bins themselves (e.g., intrinsic dissimilarity between epithelial and mesenchymal cells). 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 to yield a final singular measure of distance, or dissimilarity, between two specimens (50). EMD can be conceptualized as the minimal amount of "effort" needed to move mass (e.g., cells) between bins of one histogram so that its shape matches that of the other histogram (i.e., all matching bins of two histograms have the same counts). Mathematically, EMD is defined by the following optimization problem:

$$\mathrm{EMD}(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}}$$

Such that $\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}$ is minimized subject to the following constraints:

1) $f_{ij} \geq 0 \qquad 1 \leq i \leq m, \quad 1 \leq j \leq n$
2) $\sum_{j=1}^{n} f_{ij} \leq w_{p_i} \qquad 1 \leq i \leq m$
3) $\sum_{i=1}^{m} f_{ij} = w_{q_j} \qquad 1 \leq j \leq n$

Definition 1. Earth Mover's Distance as an optimization problem.
$P = \{(p_1, w_{p_1}), \ldots, (p_m, w_{p_m})\}$, where p_i represents histogram bin i in the initial starting signature P and w_{p_i} represents the amount of "mass" present in bin i. Similarly, $Q = \{(q_1, w_{q_1}), \ldots, (q_n, w_{q_n})\}$, where q_j represents histogram bin j in the final signature Q and w_{q_j} represents the amount of "mass" present in bin j. f_{ij} represents the "flow" or amount of mass moved from bin p_i to bin q_j. d_{ij} represents the "ground distance" between bins p_i and q_j. Constraint 1 ensures that P and Q are the starting and final

signatures, respectively. Constraints 2 and 3 ensure that no more mass is moved from any bin p_i than is present initially.

EMD has been used in various applications including image retrieval (50, 52), visual tracking (53), and melodic similarity musical analysis (54): all tasks that require accurate comparison of multidimensional distributions (analogous to comparing single-cell specimens). Additionally, a prior study demonstrated proof-of-concept that Earth Mover's Distance can be used effectively to differentiate flow cytometry specimens of phenotypically distinct individuals (55). By design, EMD is a distance measure between probability distributions that is particularly invariant to small shifts in data (i.e., noise or technical variability) across specimens (50, 55). EMD also gives a "complete" measure of overall dissimilarity between two specimens, largely attributable to the fact that it takes into account both the difference in height of corresponding histogram bins between specimens (e.g., number of epithelial cells) and the concept that certain bins (e.g., cell subtypes) have a smaller "ground distance" (i.e., are more similar) than others. Including ground distance between bins in the EMD computation allows us to incorporate the idea that it requires more "effort" to move mass to a faraway bin than to a nearby bin (i.e., it requires more effort to convert cells to a more dissimilar cell signature than to a more similar cell signature). Recall that each cell subtype is associated with various different datapoints (individual cells assigned to that subtype), so it can be represented as the centroid of the cluster of cells that comprise it. In our application, we define the ground distance between two cell subtypes as the manifold distance between the cluster centroids of the two cell subpopulations representing the subtypes.

EMD is used to compare a pair of single-cell specimens, but the issue remains of how to relate a large set of samples simultaneously. For this task, we employ a manifold learning approach that assumes 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 derived from the high-dimensional ambient space of collected features (56). Such methods 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). Local neighborhood (and subsequently global network) structure is learned by first computing EMD between each pair of single-cell specimens; distances between "nearby" specimens are then preserved in the final learned manifold.

Leveraging the properties of EMD and manifold learning, we developed PhEMD as a novel "manifold-of-manifolds" approach to simultaneously relating a large set of single-cell specimens (Figure 2). PhEMD first aggregates cells from all biospecimens and applies a single-cell embedding technique (e.g., PHATE) to model the cell-state space (i.e., "first-level manifold"). PHATE simultaneously identifies all cell subtypes and relates them in a low-dimensional embedding. After constructing the PHATE cell-state manifold, PhEMD represents each biospecimen to be compared as a frequency histogram capturing relative abundance of each cell subtype (i.e., distribution of cells along a manifold). In the event that subsampling is performed when constructing the PHATE cell-state manifold, cells are assigned to a subtype using a nearest-neighbor approach[A]. PhEMD then uses EMD, incorporating manifold-distance as ground-distance between

[A] 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*.

bins, to compare two relative abundance histograms and derive a single value representing the dissimilarity between two single-cell specimens. PhEMD computes EMD pairwise for each pair of specimens to generate a distance matrix representing specimen–specimen dissimilarity. Finally, using this distance matrix, PhEMD generates a low-dimensional embedding of single-cell specimens (i.e., “higher-level manifold”) using diffusion maps to highlight specimen–specimen relationships in the context of overall network structure (57). Diffusion maps are useful in this case as they learn a nonlinear mapping of samples from high- to low-dimensional space, capture both local and global structure, and have intrinsic denoising properties. PhEMD identifies and visualizes clusters of similar samples based on the compositional similarity of their respective cell populations.

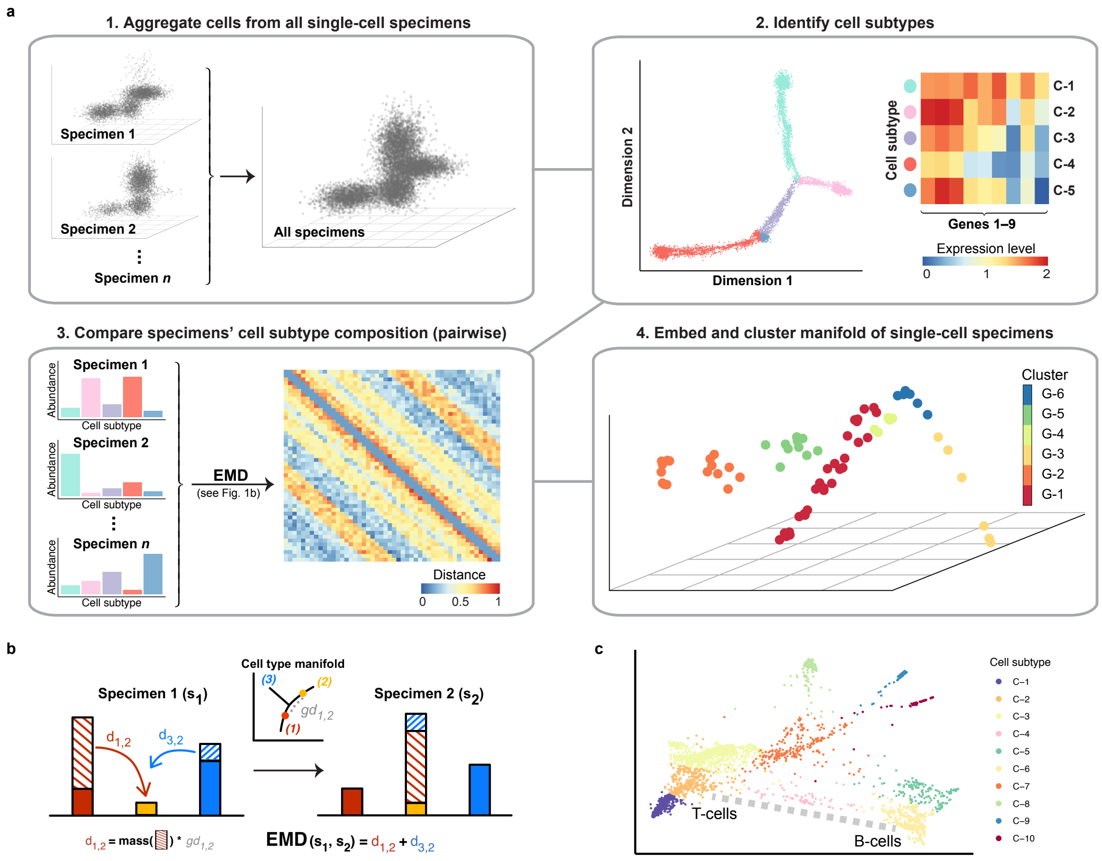


Figure 2. a) Flow diagram outlining the sequential steps performed in the PhEMD analysis pipeline. b) Schematic of the EMD computation, which accounts for both the differences in heights of matching bins and the intrinsic similarity of bins. c) Visual representation of "ground distance" (dissimilarity) between cell subtypes. The ground distance between subtypes C-2 and C-6 can be conceptualized as the length of the dotted path drawn in grey.

Pseudocode for the PhEMD algorithm is shown below in Algorithm 1.

Algorithm 1

```
procedure PhEMD(multispecimen.data)
                                        ⊳Map first-level manifold (e.g., cell-state embedding)
    data.all←aggregateData all specimens(multispecimen.data)
    first.level.embedding←embedDatapoints(data.all)
    first.level.clusts←clusterPoints(first.level.embedding)
    cluster.ground.dists←computeGroundDists(first.level.embedding; first.level.clusts)

                                        ⊳Map higher-level manifold (e.g., single-cell specimen embedding)
    specimen.clus.prop←GetClusterProportions(data.all; first.level.embedding; first.level.clusts)
    for each pair of specimens $s_i$; $s_j$ do
        Dists[i; j]←EMD(cluster.ground.dists; specimen.clus.prop[i]; specimen.clus. prop[j])
    specimen.embedding←DiffusionMap(Dists)
    specimen.clusters←ClusterSpecimens(Dists)
```

Data collection and processing

Py2T cell culture and stimulation

Py2T murine breast cancer cells were obtained from the laboratory of Gerhard Christofori, University of Basel, Switzerland (58). Cells were tested for mycoplasma contamination upon arrival and regularly during culturing and before being used for experiments. Cells were cultured at 37°C in DMEM (Sigma Aldrich), supplemented with 10% FBS, 2 mM L-glutamine, 100 U/mL penicillin, and 100 µg/mL streptomycin, at 5% CO_2. For cell passaging, cells were incubated with TrypLE Select 10X (Life Technologies) in PBS in a 1:5 ratio (v/v) for 10 minutes at 37°C.

Human recombinant TGF-β_1 was purchased from Cell Signaling Technologies as lyophilized powder and was reconstituted in PBS containing 0.1% carrier protein, according to the manufacturer's protocol to 400 ng/mL. The stock solution was kept at -

20°C until use. For daily treatment, TGF-β_1 stock was diluted into medium to 40 ng/mL working concentration. Following small-molecule inhibitor treatment, 10 µL TGF-β_1 was added to the cells for a final concentration of 4 ng/mL. As a control, PBS containing carrier protein diluted with growth medium was used.

Small-molecule inhibitors

A library of 234 small molecule kinase inhibitors was purchased from Selleckchem (Table S1). Small-molecule inhibitors were distributed within the 60 inner wells of five separate 96-well format deep well blocks with exception of wells within row E, which contained DMSO. Stock solutions of 2 mM small molecule inhibitor in DMSO were kept at -80°C until used. For daily treatment, the stock solution was equilibrated at room temperature for 1 h and then 5 µL of stock solution was added 995 µL of medium. Small-molecule inhibitor (or DMSO) was added to cells once per day, immediately after the cell growth media change and before application of TGF-β_1. Small-molecule inhibitor treatment was performed by adding 10 L of pre-diluted reagent to the cells in 80 µL cell growth medium; this resulted in a final concentration of 1 µM of small-molecule inhibitor and 0.1% DMSO.

Chronic kinase inhibition screen

For the chronic inhibition experiment, Py2T cells were seeded in 96-well plates (TPP, Techno Plastic Products AG) with a seeding density of 1800 cells per well in 80 µL of growth cell media. Only the 60 inner wells were used for analysis. In order to acquire sufficient sample size, five 96-well plates were used for single condition. After seeding, cells were allowed to recover for 36 hours to reach 50% confluence. Cells were treated simultaneously with TGF-β_1 or vehicle (PBS) and small-molecule inhibitor or vehicle

(DMSO) for 5 days, and medium was changed daily. All pipetting procedures were performed at room temperature using a Biomek FX Laboratory Automation Workstation (Beckman Coulter) supplied with 96-well pipetting pod.

In addition to experimental conditions treated with small-molecule inhibitors, at least five "uninhibited" control conditions and five "untreated" control conditions were included on each 96-well plate. Uninhibited control conditions were those in which TGF-β_1 was applied to induce EMT in absence of any inhibitor. Untreated control conditions were those in which neither TGF-β_1 nor inhibitor was applied and no EMT was induced.

Cell collection

The cell collection protocol was performed using a Biomek FX Laboratory Automation Workstation. The cell growth medium was removed using the multiple aspiration pipetting technique, and cells were washed twice with 37°C PBS. Dissociation reagent TrypLE Select 10X (Life Technologies) was diluted into PBS at a 1:5 ratio (v/v) was added to the cells and incubated for 10 min at 37°C. Cells were detached from plates. Five identically treated 96-well plates were combined into a single deep well block and were fixed for 10 min with PFA at the final concentration of 1.6% v/v. PFA was blocked with the addition of 600 μL 10% BSA in CSM. The cells were centrifuged for 5 min at 1040*g*, at 4°C. The supernatant was removed and the cells were resuspended in 300 μL of -20°C MeOH. Samples were then transferred onto dry ice and to -80°C storage.

Metal-labeled antibodies

Antibodies were obtained in carrier/protein free buffer and labeled with isotopically pure metals (Trace Sciences) using MaxPAR antibody conjugation kit (Fluidigm) according to the manufacturer's standard protocol. After determining the percent yield by measurement of absorbance at 280 nm, the metal-labeled antibodies were diluted in Candor PBS Antibody Stabilization solution (Candor Bioscience GmbH) for long-term storage at 4°C. Antibodies used in this study are listed in Table S2.

Mass-tag cellular barcoding and antibody staining

Cell samples in methanol were washed three times with Cell Staining Media (CSM, PBS with 0.5% BSA, 0.02% NaN_3) and once with PBS at 4°C. The cells were then resuspended at 1 million cells/mL in PBS containing barcoding reagents (^{102}Pd, ^{104}Pd, ^{105}Pd, ^{106}Pd, ^{108}Pd, and ^{110}Pd; Fluidigm) were conjugated to bromoacetamidobenzyl-EDTA (BABE, Dojindo) and two indium isotopes (^{113}In and ^{115}In, Fluidigm) were conjugated to 1,4,7,10-tetraazacy-clododecane-1,4,7-tris-acetic acid 10-maleimide ethylacetamide (mDOTA, Mycrocyclics) following standard procedures. Cells and barcoding reagent were incubated for 30 min at room temperature. Barcoded cells were then washed three times with CSM, pooled and stained with the metal-conjugated antibody mix (Table S2) at room temperature for 1 hour. Unbound antibodies were removed by washing cells three times with CSM and once with PBS. For cellular DNA staining, an iridium-containing intercalator (Fluidigm) was diluted to 250 nM in PBS containing 1.6% PFA, added to the cells at 4°C, and incubated overnight. Before measurement, the intercalator solution was removed and cells were washed with CSM, PBS, and doubly distilled H_2O. After the last wash step, cells were resuspended in MilliQ H_2O to 1 million cells/mL and filtered through a 40-µm strainer.

Mass cytometry data processing

EQ Four Element Calibration Beads (Fluidigm) were added to the cell suspension in a 1:10 ratio (v/v). Samples were measured on a CyTOF1 system (DVS Sciences). The manufacturer's standard operation procedures were used for acquisition at a cell rate of ~300 cells per second as described previously (59). After the acquisition, all FCS files from the same barcoded sample were concatenated using the Cytobank concatenation tool. Data were then normalized and bead events were removed (60). Cell doublet removal and de-barcoding of cells into their corresponding wells was done using a doublet-free filtering scheme and single-cell deconvolution algorithm (61). Subsequently, data were processed using Cytobank (http://www.cytobank.org/). Additional gating on the DNA channels (^{191}Ir and ^{193}Ir) was used to remove remaining doublets, debris, and contaminating particles. Final events of interest were exported as .csv files.

In-depth analysis of breast cancer EMT cell-state space and drug-inhibitor manifold from a single mass cytometry run

CyTOF measurements of cells undergoing unperturbed and perturbed EMT were generated and processed as described above. Data were then pooled from all experimental conditions, taking an equal random subsample from each condition to generate the cell-state embedding. Cell state definitions and relationships were modeled with PHATE. Subsequently, all cells from all experimental conditions were assigned a cell subtype using a nearest-neighbor approach.

Next, the cell subtype composition of each inhibition condition (i.e., relative frequencies of each cell subtype that sum to one for each sample) was determined. Using this cell subtype frequency-based representation of inhibition conditions, EMD was computed pairwise between single-cell samples. Euclidean distances between cluster centroids in the PHATE space (which approximate diffusion-based potential distances derived from the expression data native dimensions (37)) were used as a measure of intrinsic dissimilarity between cell subtypes for the EMD ground-distance matrix. EMD in this case represented the minimum "effort" required to transform one inhibition condition to another (conceptually equivalent to the total "effort" needed to move cells from relatively "overweight" parts of the branched, continuous, EMT cell-state manifold to relatively "underweight" parts). The EMD between every pair of inhibition conditions was computed to construct a network of drug inhibition conditions, represented as an EMD-based distance matrix. The resulting distance matrix was embedded using the diffusion map approach (as implemented in the 'destiny' Bioconductor `R` package (36)) and partitioned using hierarchical clustering (applied to the untransformed distance matrix) to highlight inhibitors with significant effects on EMT or similar effects to one another.

Integrating batch-effect correction to compare 300 EMT inhibition and control conditions measured in five experimental runs

CyTOF measurements of cells undergoing unperturbed and perturbed EMT were generated and processed as described in the above sections. Markers shared across all batches (n = 31) were used for downstream analyses. Data were pooled from all experimental conditions on a per-batch basis. Expression values were then linearly scaled for each gene to ensure all values were positive and in the same range across batches. After this initial normalization, an equal random subsample of cells from each batch (20,000 x 5) was used as the input for canonical correlation analysis (CCA) (62). CCA mapped expression data from each batch into an aligned, 8-dimensional space shared by all batches. The cell state manifold and cell subtype definitions were modeled by applying the PHATE dimensionality reduction and clustering method to the eight dimensions of the CCA-aligned space as input.

All cells from all experimental conditions were assigned a cell subtype using a nearest-neighbor approach. Next, the cell subtype composition of each inhibition condition (i.e., relative frequencies of each cell subtype that sum to one for each sample) was determined. Using this cell subtype-based representation of inhibition conditions, EMD was computed pairwise between single-cell samples. The ground distance (i.e. intrinsic dissimilarity) between cell subtypes was defined as the Euclidean distance between their respective centroids in the three-dimensional PHATE space. The resulting specimen-to-specimen distance matrix was embedded using the 'destiny' Bioconductor `R` package (36) and partitioned using hierarchical clustering (applied to the untransformed distance matrix) to identify 13 clusters of inhibitors with similar effects on EMT.

Intrinsic dimensionality analysis of the EMT perturbation state space

The bias-corrected maximum likelihood estimator approach was used to assess the intrinsic dimensionality of the EMT perturbation state space (63). The specimen-to-specimen distance matrix for the 300 samples was computed as described above and the intrinsic dimensionality of this embedding was estimated using the 'ider' `R` package (64). Intrinsic dimensionality was estimated over a range of values for knn parameter k from 1 through 100. The final value of intrinsic dimensionality was determined by examining the stable estimated value across a range of sufficiently large values for k (defined as >30).

Imputing the effects of inhibitions based on a small measured dictionary

A previously published sampling technique for identifying landmark points of an embedding was applied to assess whether the network geometry of all 300 inhibition and control conditions could be captured using a smaller subset of conditions (65). The technique, called "incompletely pivoted QR-based (ICPQR) dimensionality reduction," learns a concise embedding of a large collection of datapoints by identifying a subset of "landmark points" that collectively capture the geometry of the full collection of samples. The fundamental concept is that these N landmark points comprise an N-dimensional subspace and that all other existing and new points can be mapped in relation to these. ICPQR identifies the concise "landmark point" dictionary based on known pairwise distances between samples (e.g., our EMD-based distance matrix of sample-to-sample distances). The ICPQR procedure was applied as follows: first, the PhEMD distance matrix containing pairwise distances between our 300 experimental conditions was converted to an affinity matrix using a Gaussian kernel ($\sigma = 2$) and Markov-normalized to obtain probabilities. The (ICPQR) dimensionality reduction technique was then applied to this affinity matrix, using a μ distortion parameter of 0.01, to identify 34 landmark points. To assess whether the 34 landmark points adequately captured the geometry of the full collection of 300 samples, the landmark points identified were then used to impute the geometric coordinates of the remaining (non-landmark) points using the out-of-sample extension technique associated with ICPQR (65). The result was a 34-dimensional embedding of all 300 samples. We computed a 300x300 distance matrix based on the pairwise Euclidean distances between samples in this 34-dimensional space and then embedded it using the ‘destiny’ Bioconductor R package (36).

Incorporating drug-target binding specificity data to extend the PhEMD embedding and predict the effects of unmeasured inhibitors on TGFβ-induced EMT

We hypothesized that the influence of additional inhibitors on TGFβ-induced EMT could be predicted based on knowledge of inhibitor–inhibitor similarity from another data source. To test this, we obtained drug-target specificity data from a previously published experiment for a set of 39 inhibitors that overlapped between our experiment and theirs (66). Saracatinib, ibrutinib, and dasatinib were selected as three nonspecific Src inhibitors whose drug-target specificity data were known and whose effects on EMT we wanted to predict. Next, a PhEMD embedding was generated based on our CyTOF experimental results (not including the three selected inhibitors). To predict the effects of the three inhibitors on EMT relatively to other inhibitors in our experiment, we performed Nystrom extension on the diffusion map embedding. All 39 inhibitors that were found to have an effect on EMT in our experiment and that had known drug-target specificity profiles were included in the Nystrom extension. Pairwise distances between each “extended” point and each existing point in the original diffusion map were required for Nystrom extension. These distances were based on the similarity of drug-target specificity profiles between the two inhibitors, defined as $(1 - cosine\ similarity)^{20} * 4$ for all pairs of inhibitors with known drug-target specificity profiles. The remaining pairwise distances were imputed based on known PhEMD-based inhibitor–inhibitor dissimilarity and known pairwise drug target specificity-based dissimilarity using the MAGIC imputation algorithm (67).

A global shift in embedding coordinates between the original diffusion map (based on PhEMD distances) and the Nystrom extension points (based on normalized cosine similarity using drug-target specificity data) was observed. This was likely due to a difference in scale between PhEMD-based distances and cosine similarity-based distances. Nonetheless, we were able to use the Nystrom extension points alone to predict the effect of the three selected inhibitors on EMT. First, we visualized the Nystrom extension embedding to show the predicted relation of the three inhibitors to other inhibitors with known (measured) effects on EMT. Next, we used partial least squares regression (‘pls’ `R` package (68)) to predict the cell subtype relative frequencies that would result from applying the inhibitors to breast cancer cells undergoing TGFβ_1-induced EMT. Nystrom extension embedding coordinates were used as the input variables for the regression model. To validate our findings, we measured the three selected inhibitors directly using CyTOF and included them along with the rest of the inhibitors in the PhEMD analysis pipeline. We compared the actual to the predicted cell subtype relative frequencies and the actual to the predicted embedding coordinates relative to other similar, “nearby” inhibitors. To assess prediction accuracy, we compared our prediction error to the prediction error of the null hypothesis modeled by first randomizing the PhEMD-based and drug target specificity-based distance matrices and then generating a predictive model in the same way as in the alternative model. Prediction error was defined as the EMD between the predicted and actual (measured) cell subtype relative frequency distributions. The null hypothesis was modeled as a distribution of EMDs generated by randomizing the PhEMD-based and drug target specificity-based distance matrices 1,000 times and subsequently imputing cell subtype frequencies. *P*-values were computed by performing a permutation test (n=1,000) comparing our prediction error to that of the empirical null distribution and applying a one-sided significance test at a significance level of 0.05.

To more comprehensively assess PhEMD as a predictive tool, leave-one-out cross validation was performed on the 39 inhibitors with known (measured) cell subtype relative frequencies and drug-target specificity data. For each inhibitor, we constructed a PhEMD embedding based on known measurements of the 39 others and performed Nystrom extension to impute the relationship between the inhibitor and the measured ones. We then constructed a partial least squares regression model using the same input variables as above to predict the cell subtype relative frequencies of the inhibitor. Prediction error was defined the same as above (i.e. EMD between predicted and actual cell subtype relative frequency distributions). The null model was also defined in the same way as above by randomizing the PhEMD and distance matrices 100 times for the prediction of each inhibitor. To determine whether our alternative model was effective, we assessed whether the prediction errors in the alternative model (n=39) were lower than the EMDs in the null model (n=3,900) using a one-sided Mann-Whitney *U*-test.

Predicting drug-target binding specificities based on PhEMD results from EMT perturbation experiment

We hypothesized that if the PhEMD embedding were meaningful, it would have predictive power. In order to test this, we used the PhEMD embedding of inhibitors to predict the inhibitors' drug-target binding specificities. The drug-target binding specificity data were obtained from a previously published study that used a chemical proteomic approach to identify the protein targets of many clinical kinase inhibitors (66). We chose to predict the profiles of 39 inhibitors that were present in both the drug-target binding specificity experiment and ours, and that had at least 1 protein target identified by the binding specificity experiment. Next, we computed a 39-by-39 knn kernel (k=3) using the PhEMD inhibitor–inhibitor distances and then row-normalized the resulting matrix to 1 to turn it into a Markov operator. We then performed a leave-one-out cross validation, in which we set one of the inhibitor target values (i.e., drug-target binding specificity profiles) in the Klaeger et al. data to be unknown. Note that a drug-target binding specificity profile was represented as a vector of length 270, which represented the binding specificity between the drug and each of 270 potential protein targets. We predicted the drug-target binding specificity values using the MAGIC imputation method (67) with the PhEMD Markov operator as input and a diffusion parameter t of 2. We computed leave-one-out predictions for each of the 39 inhibitors. To quantify the performance of our predictive model, we computed Pearson correlation between the original ground-truth (experimentally measured) target values and the predicted values. To determine the accuracy of our predictions, we compared our results to a null model, in which we randomized the PhEMD matrix 1,000 times and each time ran the prediction using this randomized matrix. Prediction accuracy (Pearson correlations) of our alternative model (n=39 predictions, one per inhibitor) was compared to that of the null model (n=39,000 predictions, 1,000 per inhibitor) using a one-sided Mann-Whitney U-test.

Generation and analysis of dataset with known ground-truth branching structure

To evaluate the accuracy of the PhEMD analytical approach, high-dimensional single-cell data ('Synthetic Dataset B') were generated using Splatter, a previously published tool designed to simulate single-cell expression data (69). The basic tree structure represented in Figure 4A was generating using the following Splatter parameters: nGenes=100, de.prob=0.5, path.from=c(0,0,0,3,3,5,5,7,7,7). Each single-cell sample consisted of 2000 cells sampled from this cell-state manifold at varying degrees of cellular density spread across the cell-state space. For Samples A-I, cellular density was concentrated in cell subtypes C-1 through C-9 (constituting the main axis), with 55% of Sample A consisting of C-1 and C-2 cells and 55% of Sample I consisting of C-8 and C-9 cells. Samples B-H consisted of progressively fewer cells in the starting cell states (i.e., C-1 and C-2) and progressively more cells in the terminal cell states (i.e., C-8 and C-9). Samples X, Y, and Z were enriched for cells in C-10, C-13, and C-14 respectively. Samples J-M were comprised predominantly of C-11 cells and Samples N-Q were comprised predominantly of C-12 cells at increasing degrees of cell-type enrichment.

We applied PhEMD to the library-size normalized Splatter data as outlined in Figure 2. First, the tree structure was modeled by PHATE based on cells aggregated from all biological specimens. Then, the relative frequency of cells across different cell subtypes was computed for each specimen. EMD was computed pairwise for all cells using PHATE distances as a measure of ground-distance between cell subtypes. A final diffusion map embedding of biospecimens was generated using the 'destiny' Bioconductor `R` package.

Analysis of melanoma single-cell RNA-sequencing dataset

Data from a prior single-cell RNA-sequencing experiment were downloaded from the NCBI Gene Expression Omnibus website, accession number GSE72056 (12). These data contained read-count expression values that were log TPM-normalized values. 2 of the 19 samples were excluded from analysis due to low cell yield of immune cells. Initial feature selection was performed by selecting 44 features found in the initial publication characterization of this dataset to distinguish between key cell types (12). The PHATE model of the cell-state space was constructed using default parameters to identify ten cell subtypes. The remaining PhEMD analysis pipeline was completed as described in 'In-depth analysis of breast cancer EMT cell-state space and drug-inhibitor manifold from a single mass cytometry run'; a final embedding of biopsy samples was generated using the 'destiny' Bioconductor `R` package and partitioned using hierarchical clustering.

Analysis of clear cell renal cell carcinoma dataset

CyTOF data from a recent publication characterizing the immune landscape of clear cell renal cell carcinoma were downloaded from https://premium.cytobank.org/cytobank/projects/875 (13). Cell data were filtered and normalized using the method described in Methods section titled 'Mass cytometry data processing'. The PHATE model of the cell-state space was constructed with a diffusion parameter $t = 40$ to identify ten cell subtypes. The remaining PhEMD analysis pipeline was completed as described in 'In-depth analysis of breast cancer EMT cell-state space and drug-inhibitor manifold from a single CyTOF experiment'.

Statistical methods

Statistical tests were performed as detailed in the above subsections. Differences in group medians were assessed using a Mann-Whitney U-test. Benchmarking of prediction accuracy (point estimate) against a null distribution was performed using a permutation (i.e., randomization) test. All statistical comparisons were performed at a two-sided significance level of 0.05 unless otherwise stated.

Data availability

The mass cytometry data that support the findings of this study are available at https://community.cytobank.org/cytobank/projects/1296. Any additional data supporting the findings of this study are available upon request.

Code availability

PhEMD ("Phenotypic Earth Mover's Distance") takes as input a list of N matrices representing N single-cell specimens. An `R` implementation of PhEMD is publically available as a Bioconductor `R` package (package name: 'phemd') and can alternatively be downloaded from https://github.com/KrishnaswamyLab/phemd. Note that the cell-state space for all analyses presented in this manuscript was modeled using the PHATE method. However, alternative approaches are viable, and we have provided support for PHATE (37), Monocle2 (33), and Louvain community detection (as implemented in the Seurat software package (62)) for this purpose in the `R` package.

Author contributions

All *in vitro* EMT drug-screen experiments including CyTOF profiling of cells were performed by collaborator Nevena Zivanovic (NZ) at the University of Zurich in the laboratory of Prof. Bernd Bodenmiller. Data quality assurance of these data was performed jointly by the author of this thesis (WSC) and NZ. WSC obtained the data for all other single-cell datasets as described above, wrote the software implementation of PhEMD, and performed all computational analyses presented.

RESULTS

Overview of PhEMD

PhEMD is a method for embedding a “manifold of manifolds,” i.e., sets of datapoints in which each datapoint itself represents a collection of points that comprise a manifold. In the setting of analyzing single-cell data, each datapoint in the “manifold of manifolds” represents an experimental condition (i.e., single-cell specimen), which is itself comprised of a heterogeneous mixture of cells that span a cell-state manifold. PhEMD first embeds each biospecimen as a manifold and then derives a pairwise distance between the manifolds. Deriving an “higher-level” embedding then involves using these pairwise specimen-to-specimen distances to find a coordinate system (i.e., axes of variability) such that each point represents a specimen, and the distance between the points represents the dissimilarity between specimens. PhEMD derives such an embedding using the following general steps (Figure 2):

1. Compute a distance between each pair of datasets (i.e., experimental conditions) as follows:
   a) Embed points within each dataset using PHATE (37).
   b) Cluster datapoints using spectral clustering.
   c) Represent each dataset as a vector of relative cluster proportions.
   d) Compute the distance between two datasets using Earth Mover’s Distance (EMD).
2. Take the distance matrix derived from the previous step and compute a diffusion map embedding of the data (34).

When specifically applied to single-cell data, PhEMD leverages PHATE and spectral clustering to define cell subtypes, EMD to compute pairwise distances between biospecimens (based on their cell subtype relative abundances), and the diffusion map approach to generate a final low-dimensional embedding of biospecimens. Pseudocode and additional details on the PhEMD algorithm can be found in Methods.

Comparing specimens pairwise using Earth Mover's Distance (EMD)

A critical component of deriving the correct single-cell specimen embedding is computing accurate specimen-to-specimen distances. Two existing methods for doing so are cellAlign (43) and sc-UniFrac (44). However, both impose limiting assumptions or faced scalability issues that are addressed in our implementation of EMD.

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 cannot be applied to cell populations sampled from branched cell-state trajectories, as it assumes cells with the same pseudotime value have identical gene expression profiles (an assumption violated in the setting of branched cell-state trajectories). Our implementation of EMD does not make such an assumption and is thus more flexible for analyzing datasets with branched cell-state trajectories.

sc-UniFrac is a different method that was similarly designed to compare two single-cell experimental conditions but that faces scalability issues. Its memory requirements exceed 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 prevents 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 is unable to be run on a laptop to analyze a set of 40,000 cells from two or more experimental conditions, PhEMD can be 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 A"). The cell-state tree was generated using the Splatter R package and was characterized by four branches sharing a single branch point. Our implementation of EMD correctly recovered the known cell-state space of the dataset (Figure 3A) and had faster empiric runtime than when analyzing datasets including more than 21,000 cells in total (Figure 3B).

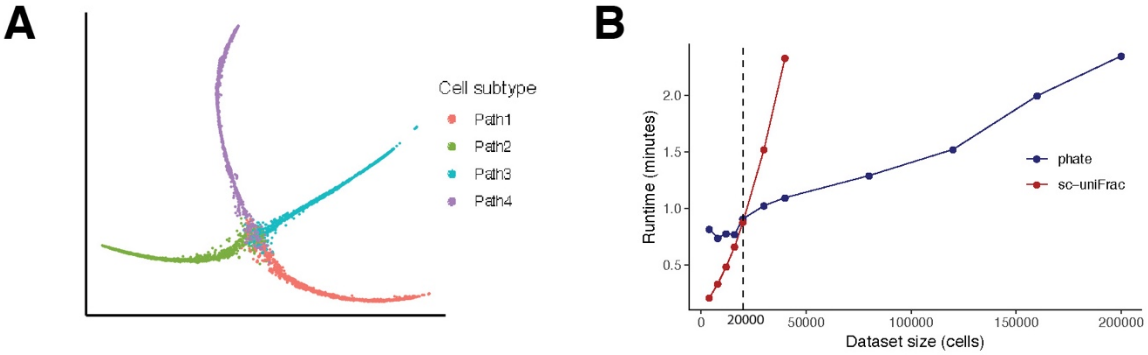


Figure 3. a) PHATE embedding of the cell-state space of Synthetic Dataset A colored by cell-subtypes identified by PHATE. b) Runtime comparison between PHATE and sc-UniFrac applied to datasets of increasingly larger sample sizes.

In sum, unlike cellAlign, which can 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 can be used to compare specimens comprised of cells sampled from an underlying cell-state manifold that is potentially branched. Compared to sc-UniFrac, our implementation of EMD is much more scalable, allowing for the efficient pairwise comparison of multiple specimens as is required to generate a final embedding containing many single-cell specimens.

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

We first applied PhEMD to simulated single-cell data with known ground-truth structure to determine whether PhEMD could accurately model both the cellular heterogeneity within each specimen and the specimen-to-specimen heterogeneity based on cell subtype relative abundances. The simulated cells lay on a continuous branched trajectory, wherein progression along a branch represented concurrent changes in gene expression in select differentially expressed genes (69). The distribution of cell density across branches was varied between specimens to simulate a heterogeneous multi-specimen dataset. PhEMD correctly recovered the branched cell-state manifold structure using PHATE (Figure 4A-B). The specimen-to-specimen EMD-based comparison and resulting PhEMD embedding were also found to be accurate (Figure 4C).

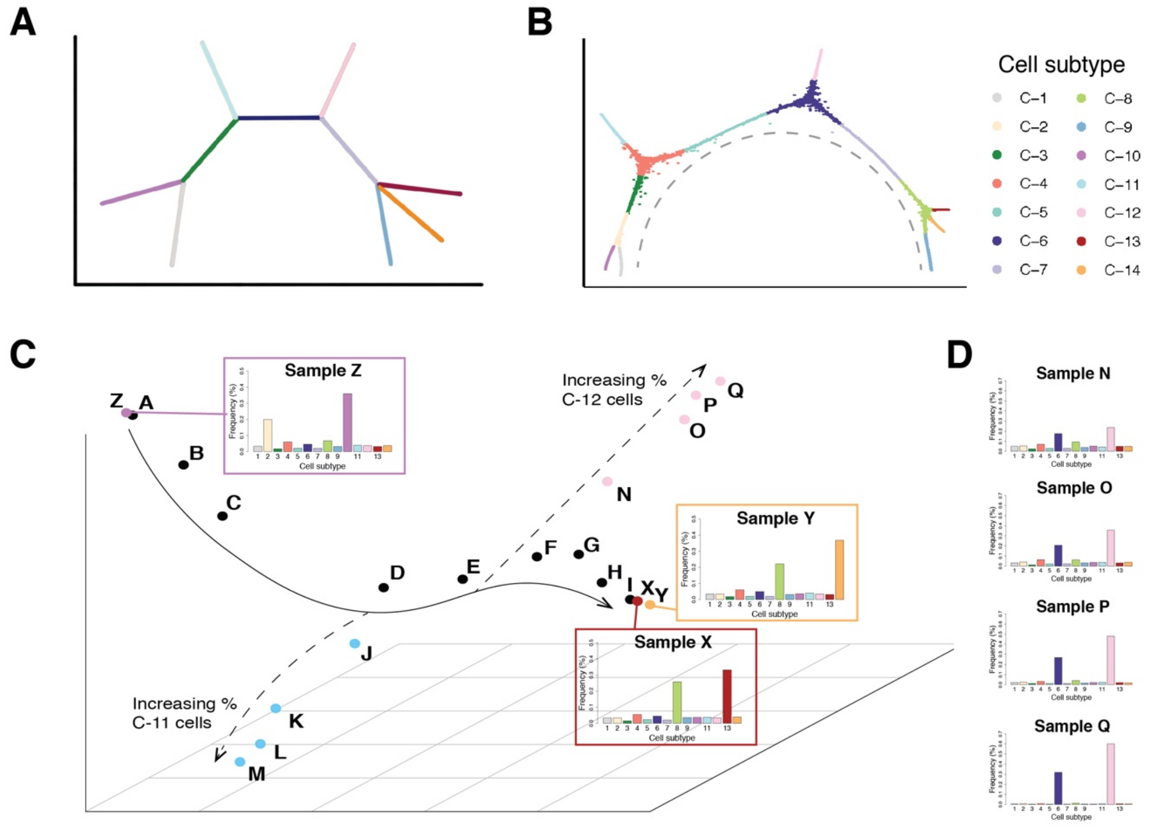
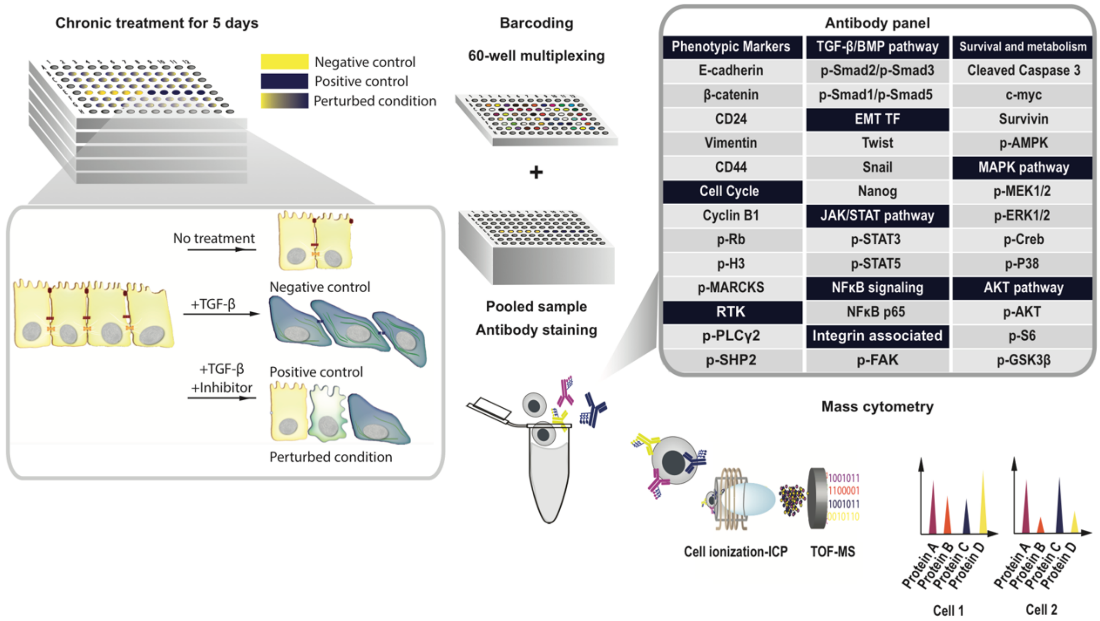


Figure 4. a) Ground-truth tree structure of the cell-state space of Synthetic Dataset B (see Methods for data parameters). b) PHATE embedding of the cell-state space of Synthetic Dataset B, colored by cell-subtypes identified by PHATE. Grey dotted line denotes major axis (comprised of cell subtypes C-1 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.

The accuracy of the final PhEMD biospecimen map was then 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.

Assessing the differing effects of selected drug perturbations on EMT in breast cancer

To study key regulators of epithelial-to-mesenchymal transition (EMT) in breast cancer, we performed a drug screen consisting of 300 inhibition and control conditions, collectively inhibiting over 100 unique protein targets in murine breast cancer cells undergoing TGFβ_1-induced EMT (Figure 5, Table S1). These specimens collectively contained over 1.7 million cells measured in a total of five mass cytometry runs. Time-of-flight mass cytometry (CyTOF) was used on day 5 of cell culture to measure the concurrent expression of 31 protein markers in each cell (Table S2), and PhEMD was used to model both the cell-state transition process and the perturbation-effect manifold. Batch correction was performed using canonical correlation analysis (CCA) prior to modeling the cell-state and single-cell specimen embeddings in order to analyze all experimental conditions across all plates simultaneously.



Antibody panel

Phenotypic Markers	TGF-β/BMP pathway	Survival and metabolism
E-cadherin	p-Smad2/p-Smad3	Cleaved Caspase 3
β-catenin	p-Smad1/p-Smad5	c-myc
CD24	EMT TF	Survivin
Vimentin	Twist	p-AMPK
CD44	Snail	MAPK pathway
Cell Cycle	Nanog	p-MEK1/2
Cyclin B1	JAK/STAT pathway	p-ERK1/2
p-Rb	p-STAT3	p-Creb
p-H3	p-STAT5	p-P38
p-MARCKS	NFκB signaling	AKT pathway
RTK	NFκB p65	p-AKT
p-PLCγ2	Integrin associated	p-S6
p-SHP2	p-FAK	p-GSK3β

Figure 5. Experimental design for measuring perturbation effects of small molecule inhibitors on TGFβ1-induced EMT.

Batch effect correction in multi-run EMT experiment

Batch effect is a well-known problem when comparing data from multiple single-cell RNA-sequencing (62, 70) or CyTOF (71, 72) 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 (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 appear as extensive inter-plate mixing in most regions of the t-SNE cell state space. This is because 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 (Figure 6A), we noticed poor inter-plate mixing. This indicated that batch effect was present in the unnormalized data.

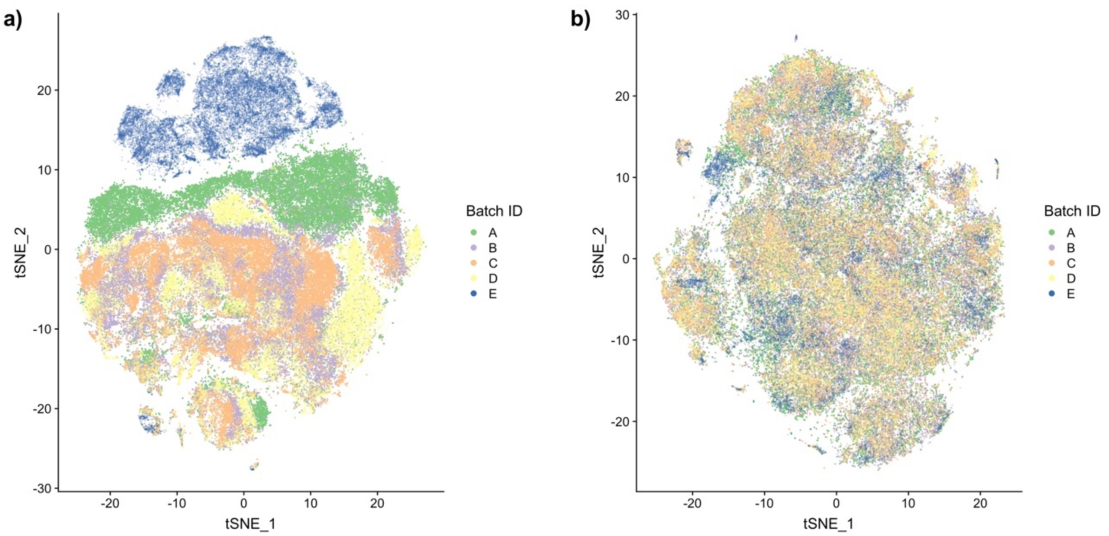
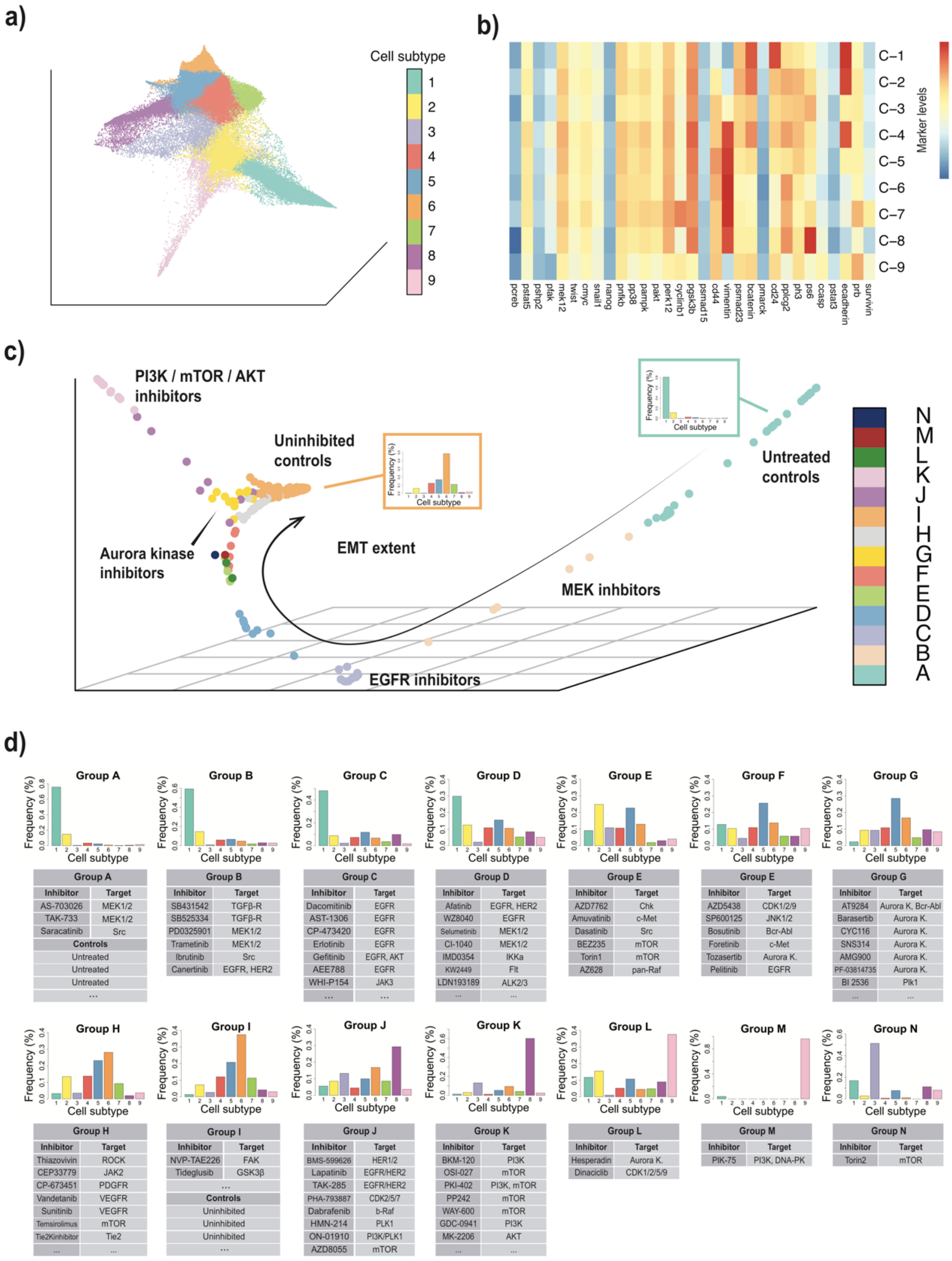


Figure 6. t-SNE embedding of cells from multiple CyTOF runs based on gene expression data a) pre- and b) post-CCA batch correction, with individual cells colored by experimental batch.

We then applied CCA to the expression measurements and ran t-SNE on the batch-corrected data (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 (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_1 nor drug inhibitor (“untreated controls”), and the other set consisted of Py2T cells stimulated with TGF-β_1 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.

Cell-subtype definition via manifold clustering

By design, all cells undergoing EMT were derived from the same homogeneous epithelial cell population. Thus, a continuous manifold with potentially branched structure (as modeled by PHATE) was ideal to model the cell-state space. Applied to the batch-corrected expression data, and PHATE identified nine cell subtypes across all unperturbed and perturbed EMT conditions (Figure 7A-B).



Group A	
Inhibitor	**Target**
AS-703026	MEK1/2
TAK-733	MEK1/2
Saracatinib	Src
Controls	
Untreated	
Untreated	
Untreated	
...	

Group B	
Inhibitor	**Target**
SB431542	TGFβ-R
SB525334	TGFβ-R
PD0325901	MEK1/2
Trametinib	MEK1/2
Ibrutinib	Src
Canertinib	EGFR, HER2

Group C	
Inhibitor	**Target**
Dacomitinib	EGFR
AST-1306	EGFR
CP-473420	EGFR
Erlotinib	EGFR
Gefitinib	EGFR, AKT
AEE788	EGFR
WHI-P154	JAK3
...	...

Group D	
Inhibitor	**Target**
Afatinib	EGFR, HER2
WZ8040	EGFR
Selumetinib	MEK1/2
CI-1040	MEK1/2
IMD0354	IKKa
KW2449	Flt
LDN193189	ALK2/3
...	...

Group E	
Inhibitor	**Target**
AZD7762	Chk
Amuvatinib	c-Met
Dasatinib	Src
BEZ235	mTOR
Torin1	mTOR
AZ628	pan-Raf

Group E	
Inhibitor	**Target**
AZD5438	CDK1/2/9
SP600125	JNK1/2
Bosutinib	Bcr-Abl
Foretinib	c-Met
Tozasertib	Aurora K.
Pelitinib	EGFR

Group G	
Inhibitor	**Target**
AT9284	Aurora K, Bcr-Abl
Barasertib	Aurora K.
CYC116	Aurora K.
SNS314	Aurora K.
AMG900	Aurora K.
PF-03814735	Aurora K.
BI 2536	Plk1
...	...

Group H	
Inhibitor	**Target**
Thiazovivin	ROCK
CEP33779	JAK2
CP-673451	PDGFR
Vandetanib	VEGFR
Sunitinib	VEGFR
Temsirolimus	mTOR
Tie2Kinhibitor	Tie2
...	...

Group I	
Inhibitor	**Target**
NVP-TAE226	FAK
Tideglusib	GSK3β
...	
Controls	
Uninhibited	
Uninhibited	
Uninhibited	
...	

Group J	
Inhibitor	**Target**
BMS-599626	HER1/2
Lapatinib	EGFR/HER2
TAK-285	EGFR/HER2
PHA-793887	CDK2/5/7
Dabrafenib	b-Raf
HMN-214	PLK1
ON-01910	PI3K/PLK1
AZD8055	mTOR

Group K	
Inhibitor	**Target**
BKM-120	PI3K
OSI-027	mTOR
PKI-402	PI3K, mTOR
PP242	mTOR
WAY-600	mTOR
GDC-0941	PI3K
MK-2206	AKT
...	...

Group L	
Inhibitor	**Target**
Hesperadin	Aurora K.
Dinaciclib	CDK1/2/5/9

Group M	
Inhibitor	**Target**
PIK-75	PI3K, DNA-PK

Group N	
Inhibitor	**Target**
Torin2	mTOR

Figure 7. a) PHATE embedding of cells from all 300 experimental conditions, colored by cell subtype. b) Heatmap representing $\log_2$ protein expression levels for each cell subpopulation representing its respective cell subtype. c) Diffusion map embedding of control and drug-inhibited conditions, colored by clusters determined 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 S3.

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 7B). E-cadherin is the hallmark cell adhesion marker of epithelial cells (73), and vimentin and CD44 are known mesenchymal markers involved in cell migration (73–76). 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 (77–79). 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-3, 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 (80–82). 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.

Constructing and clustering the EMD-based drug-inhibitor manifold

After modeling the EMT cell-state space with PHATE, PhEMD mapped the experimental variable (i.e., multicellular biospecimen) state space as a low-dimensional embedding (Figure 7C). Hierarchical clustering revealed clusters of inhibitors with similar net effects on EMT. Moreover, “uninhibited” controls (TGF-β_1 applied in absence of any inhibitor) and “untreated” controls (neither TGF-β_1 nor inhibitor applied) were included to distinguish inhibitors with notable effects on EMT.

The final embedding of drug inhibitors highlighted the variable extent of EMT that had occurred in the different inhibition conditions (Figure 7C-D). This diffusion map embedding was low-dimensional with an intrinsic dimensionality of 2.4 (Figure 8), implying relatively few axes of variation that could be appropriately visualized in three dimensions.

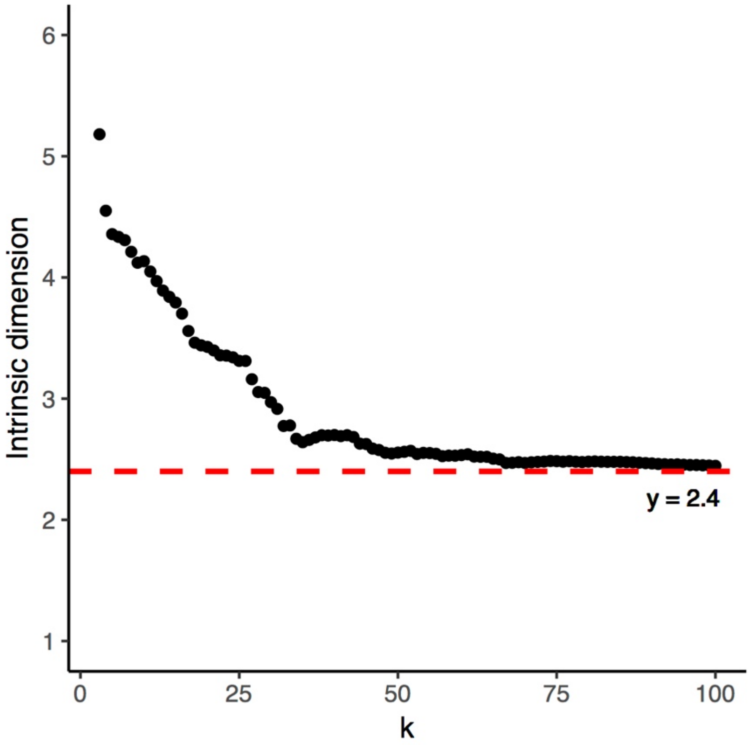


Figure 8. Intrinsic dimension of the PhEMD embedding comprised of 300-sample multi-batch EMT inhibition and control conditions, computed using the maximum likelihood estimation (MLE) approach over a range of "k" (k-nearest-neighbors parameter) values.

Fourteen inhibitor clusters (Clusters A-N) were identified (Table S3). Cluster A included the untreated controls and the TGFβ1-receptor inhibitor condition, each of which consisted almost entirely of epithelial cells (C-1). These were experimental conditions in which EMT was effectively not induced. On the other hand, Cluster I included all uninhibited control conditions and inhibitors ineffective at modulating EMT; inhibitors in this cluster were found to have mostly mesenchymal (C-6) cells. Clusters B through H included inhibitors that had generally decreasing strength with respect to halting EMT (Figure 7C-D). The inhibitors in Clusters J and K formed a prominent trajectory off the main EMT-extent trajectory in the inhibitor embedding (Figure 7C). Clusters J and K were enriched in cell subtype C-8, with Cluster K inhibitors inducing cell populations that were almost entirely comprised of C-8 cells.

All of the Cluster K inhibitors targeted PI3K, Akt, or mTOR protein kinases – three members of a well-characterized pathway. Compared to the predominant mesenchymal subtype observed in the uninhibited controls (C-6), C-8 was comprised of cells with similarly high expression of vimentin and CD44 and markedly higher expression of phospho-S6 (Figure 7). This expression profile was consistent with an alternative-mesenchymal EMT subtype. Examining the cell yield of these inhibitors compared to the respective uninhibited control conditions in their respective batches, we found that the cell yield of the Cluster K inhibitors was on average 60% lower than the TGFβ_1-only controls (Table S4). Based on these findings and a prior report that high expression of phospho-S6 was associated with resistance to PI3K inhibitors (83), the C-8 subtype is likely a mesenchymal cell population relatively resistant to inhibition of the PI3K-Akt-mTOR axis.

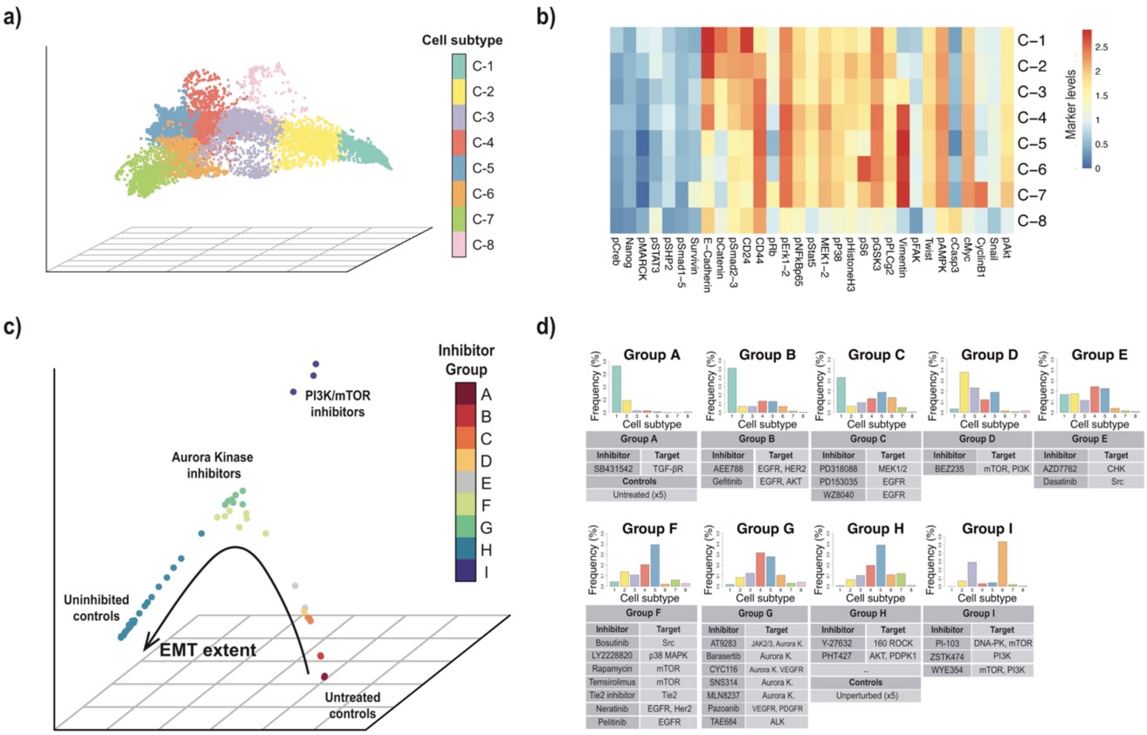
In general, small molecule inhibitors that had the same molecular target tended to cluster together, consistent with the intuitive notion that drugs with similar mechanisms of action likely have similar net effects on a given cell population (e.g. Cluster C, Cluster G). However, several inhibitors with the same reported primary target generated different resulting single-cell profiles and were clustered into different inhibitor clusters. This phenomenon may be due to differences in inhibitor potency and differences in off-target effects.

Analyzing EMT perturbations measured in a single CyTOF run

An analysis of a subset of 60 inhibition and control conditions measured in the same mass cytometry run (and hence not requiring batch normalization) was performed to assess whether applying PhEMD to batch-normalized and single-batch expression data would yield consistent results. Three replicates involving independent cell culture experiments measured in distinct mass cytometry runs were analyzed to demonstrate reproducibility of results.

Cell subtype definition via manifold clustering

Our model of the cell-state space identified eight unique cell subtypes across all unperturbed and perturbed EMT specimens (Figure 9A-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 9). C-6 through C-8 had expression profiles consistent with C-7 through C-9 in our multi-batch experiment (Figure 7B, Figure 9B). Altogether, the cell subtypes recovered in the single-batch and batch-normalized experiments were consistent with one another and with known EMT cell subtypes.



Group A	
Inhibitor	**Target**
SB431542	TGF-βR
Controls	
Untreated (x5)	

Group B	
Inhibitor	**Target**
AEE788	EGFR, HER2
Gefitinib	EGFR, AKT

Group C	
Inhibitor	**Target**
PD318088	MEK1/2
PD153035	EGFR
WZ8040	EGFR

Group D	
Inhibitor	**Target**
BEZ235	mTOR, PI3K

Group E	
Inhibitor	**Target**
AZD7762	CHK
Dasatinib	Src

Group F	
Inhibitor	**Target**
Bosutinib	Src
LY2228820	p38 MAPK
Rapamycin	mTOR
Temsirolimus	mTOR
Tie2 inhibitor	Tie2
Neratinib	EGFR, Her2
Pelitinib	EGFR

Group G	
Inhibitor	**Target**
AT9283	JAK2/3, Aurora K.
Barasertib	Aurora K.
CYC116	Aurora K. VEGFR
SNS314	Aurora K.
MLN8237	Aurora K.
Pazoanib	VEGFR, PDGFR
TAE684	ALK

Group H	
Inhibitor	**Target**
Y-27632	160 ROCK
PHT427	AKT, PDPK1
...	
Controls	
Unperturbed (x5)	

Group I	
Inhibitor	**Target**
PI-103	DNA-PK, mTOR
ZSTK474	PI3K
WYE354	mTOR, PI3K

Figure 9. 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.

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 A. We found that the sample-to-sample distances were accurate (Pearson ρ > 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 greater than 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 dissimilarity) 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 9C). 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_1 was applied to induce EMT in absence of any inhibitor) and “untreated” controls (samples in which neither TGF-β_1 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 9C-D). Partitioning the embedding into nine clusters (Clusters A-I, Table S5), we found that Cluster A included the untreated controls and the TGFβ_1-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 9C-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 9C). 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 9C), 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. Altogether, consistent results were observed across all single-batch and multi-batch analyses with respect to the resulting cell-state and higher-level biospecimen embeddings, demonstrating PhEMD's reproducibility and robustness to batch-normalized data (Figure 7, Figure 9–10).

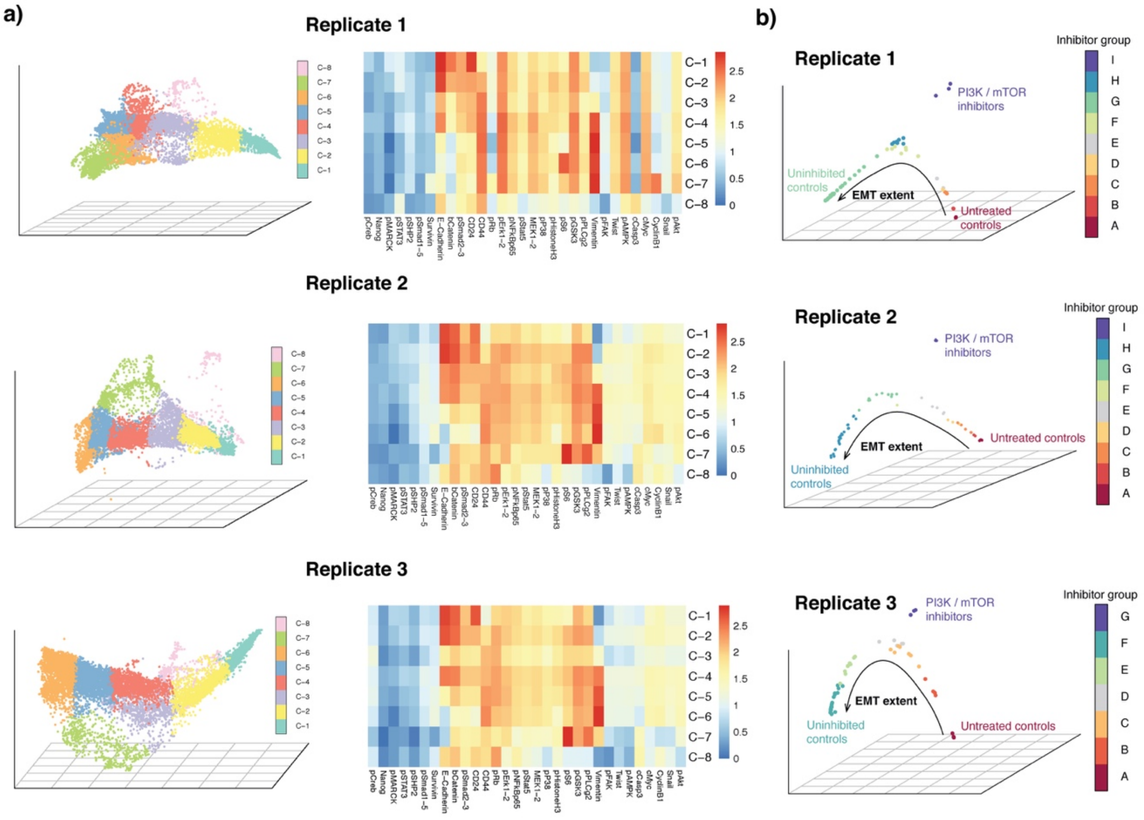


Figure 10. a) Cell subtype expression patterns and cell-state embeddings for three independent experimental replicates. b) PhEMD biospecimen embeddings and inhibitor clusters for three independent experimental replicates. The full list of inhibitors in each group can be found in Table S5.

Imputing the effects of inhibitors based on a small measured dictionary

In our model breast cancer system, we were able to use PhEMD to assess the effects of a large panel of inhibitors on TGFβ_1-induced EMT. We found that these inhibitors could be grouped into clusters based on the similarity of their effects and embedded in low dimension (with an intrinsic dimensionality of 2.4) to highlight complex, non-linear relationships between samples. Visualizing this embedding of inhibition conditions in 3D, we found that samples were distributed with varying density along a branched, continuous manifold. For example, the embedding space containing Cluster H inhibitors was characterized by high point density, while the embedding space containing Cluster B points was more sparsely populated (Figure 7C). We also noted that clusters often contained multiple inhibitors that targeted the same protein kinases. These findings suggested that we may have been able to capture the geometry of the drug-inhibition state space without measuring every single inhibition condition. If true, this finding would have implications for potentially reducing the cost of conducting single-cell drug-screen experiments, as it would suggest that only a small fraction of all inhibitors may need to be experimentally tested using expensive single-cell profiling techniques to assess the efficacy of a drug.

To test this hypothesis, we applied a previously published sampling technique to our PhEMD embedding (65). 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. Using this approach, we identified 34 landmark points that summarized our EMT perturbation state space (Figure 11A). 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. The resulting embedding had a similar geometry to that of our original 300-sample PhEMD embedding, suggesting that the 34 landmark points were sufficient to capture the overall network structure of all 300 measured experimental conditions (Figure 7C, Figure 11B). Comparing the pairwise sample–sample distances of all 300 samples in the 34-dimensional landmark-point space to the experimentally computed EMD sample–sample distances, we found that there was strong correlation between these distances (ρ=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. This highlights a potential opportunity for reducing the cost and improving the feasibility of future single-cell drug-screen experiments. Based on our findings, only a small fraction (11%) of all inhibitors may need to be experimentally measured using expensive single-cell profiling techniques to learn the full spectrum of perturbation effects.

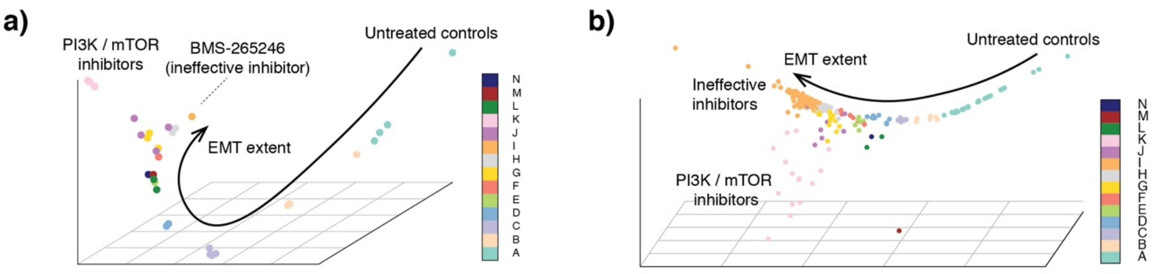


Figure 11. a) Diffusion map embedding of 300-specimen EMT experiment, plotting only the 34 landmark points identified using a previously published diffusion map sampling technique (see Methods). Points are colored based on cluster assignments as determined based on original clustering of all 300 samples (see Figure 7C). 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 Methods).

Validating the PhEMD embedding using external information on similarities between small-molecule inhibitors

We sought to validate our PhEMD drug-screen embedding by comparing the drug-drug similarities learned from our experiment (in the context of effects on EMT) to drug-drug similarities based on known drug-target binding specificities from a prior experiment (66). Since the prior experiment and ours measured an overlapping set of inhibitors, they could be conceptualized as two complementary "views" of the same shared inhibitors. We hypothesized that for the inhibitors shared between the two experiments, one view of the data might inform the other. Intuitively, this would support the notion that drugs with more similar protein targets action may tend to have more similar effects on EMT (and vice versa). Our approach to assessing this hypothesis was twofold: 1) We used a measure of inhibitor–inhibitor similarity, derived from the drug-target specificity data, to extend our PhEMD embedding and predict the effects of unmeasured inhibitors on our model EMT system, and 2) We used our PhEMD embedding to predict the drug-target specificity of inhibitors shared between the two drug-screen experiments.

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

For the first task, 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 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). 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 (84–86). 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 (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 12A-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 12B; Methods).

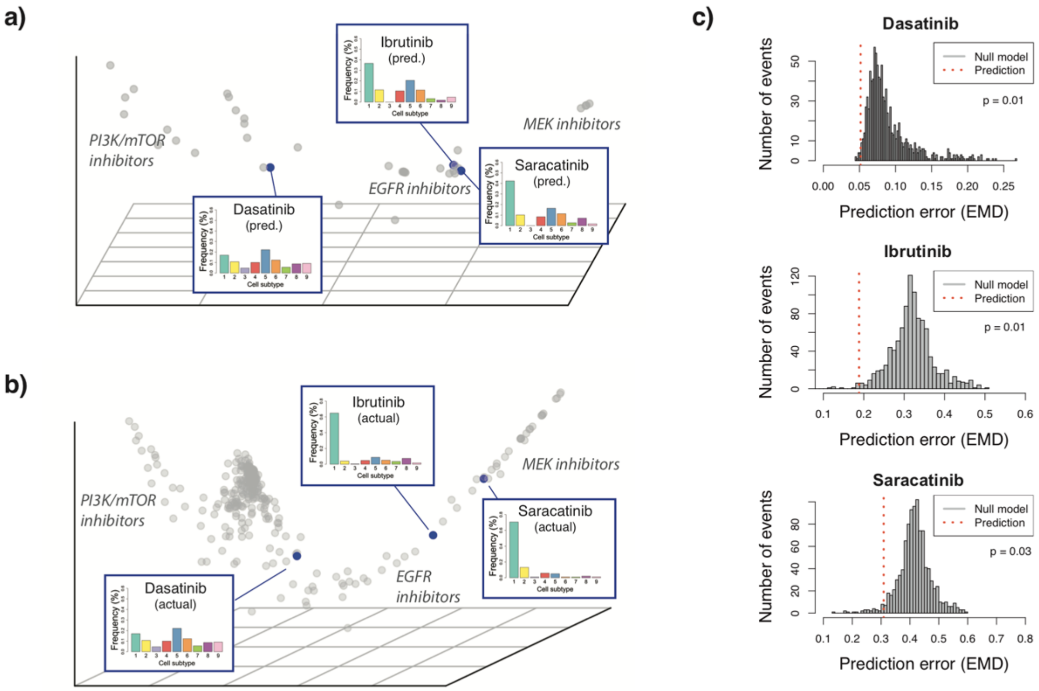


Figure 12. a) Nystrom extension embedding showing predicted effect of 3 selected inhibitors (dasatinib, ibrutinib, saracatinib) on EMT relatively to other measured inhibitors. b) PhEMD diffusion map embedding showing measured effects of 3 selected inhibitors on EMT. c) Histogram showing distribution of prediction error for null model (n=1000 independent permutations). Dotted red line represents prediction error for actual prediction (i.e., alternative model). *P*-values were computed using a one-sided permutation test.

Imputing the single-cell phenotypes of three unmeasured inhibitors based on drug-target similarity to measured inhibitors

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 (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 12C).

To assess more generally whether PhEMD could be integrated with complementary data to accurately predict perturbation effects, we performed leave-out-out cross validation on all 39 inhibitors in our CyTOF experiment with known drug-target specificity data (Methods). We found that single-cell profile predictions leveraging our imputed PhEMD embedding were significantly more accurate than a null model (*P*=0.005). Altogether, these findings suggested that PhEMD offered information that could be integrated with additional data sources and data types to support not only comparison of biospecimens directly measured but also prediction of single-cell phenotypes for additional, unmeasured specimens.

Predicting drug-target binding specificities based on PhEMD results from EMT perturbation experiment

We found that knowledge of drug-target binding specificity could be used to predict inhibitor effects in our model EMT system. We then sought to assess whether the reverse was true – whether the learned relationships between inhibitors from our EMT perturbation experiment could be used to predict drug-target binding specificities. For this prediction task, we used the 39 inhibitors that were present in both the drug-target profiling experiment and ours, and that had at least 1 protein target identified by their experiment. We then computed leave-one-out predictions using the MAGIC imputation algorithm (67) and results from our EMT perturbation screen experiment to predict the drug-target binding specificities of each inhibitor. Prediction accuracy was defined as the correlation between predicted and measured drug-target binding specificities for a given drug. Our predictive model that incorporated PhEMD results into the prediction was significantly more accurate than the null model (*P*=6.57x10^{-5}; Figure 13). This suggested that while the two experiments measured two distinct sets of inhibitor features, the inhibitor–inhibitor relationships learned from both experiments were consistent.

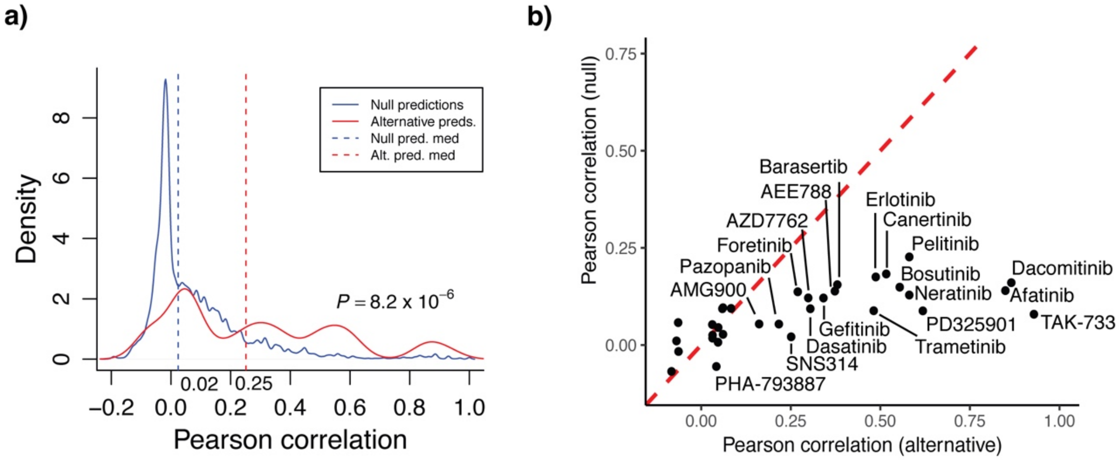


Figure 13. a) Probability density functions representing distribution of Pearson correlations between predicted and known drug-target binding specificity profiles. The null (n=39,000 predictions from 1,000 independent permutations) vs. alternative (n=39 predictions) models demonstrated median correlation-based 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 Methods for detailed properties of the null and alternative models.

PhEMD highlights manifold structure of tumor specimens measured using CyTOF and single-cell RNA-sequencing

To demonstrate an additional application of the PhEMD analytical approach, we used PhEMD to characterize the specimen-to-specimen heterogeneity in immune cell profiles of multiple tumor specimens. We first applied PhEMD to a single-cell RNA-sequencing dataset consisting of the "healthy" (non-malignant) cells of 17 melanoma biopsies. The cell-state embedding identified a total of 10 cell subtypes with gene expression profiles consistent with previously reported subpopulations of B cells, T cells, endothelial cells, epithelial cells, NK cells, and monocytes (Figure 14A-B) (12). Cell subtypes C-1 and C-2 both represented CD8$^+$ T cells. C-1 demonstrated high expression of TIGIT, CTLA4, and LAG3 and was consistent with a T-cell exhaustion profile (87). C-3 was comprised of CD4$^+$ T cells. C-6 and C-7 represented CD19$^+$ BLK$^+$ B cells with differences in the expression of SELL and CCR7. C-8 represented CD14$^+$ monocytes (88), C-9 represented PECAM1$^+$ vWF$^+$ CDH5$^+$ endothelial cells, and C-10 represented epithelial cells with high collagen expression.

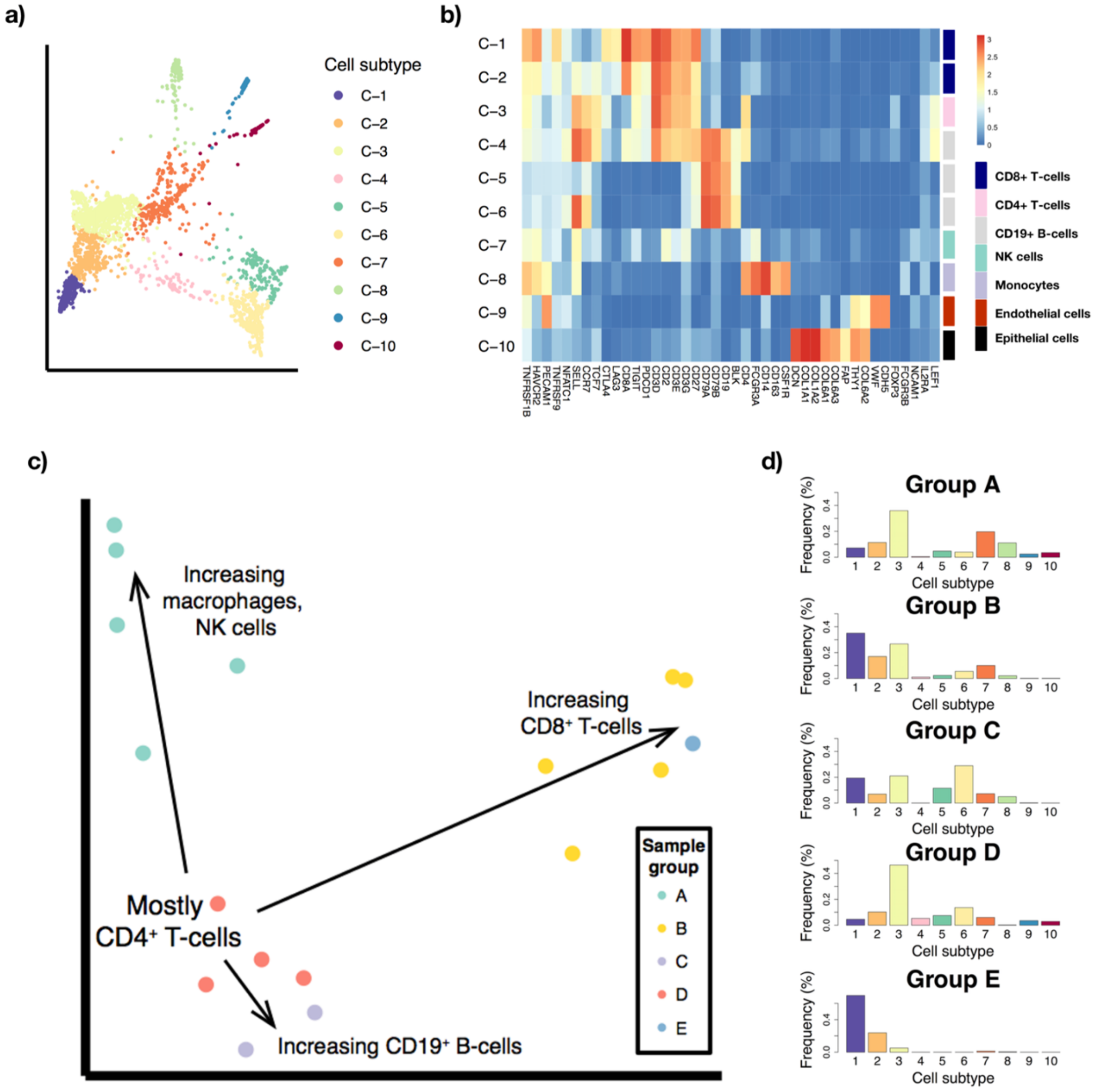


Figure 14. PhEMD applied to scRNA-sequencing data of 17 melanoma samples (non-tumor cells only) highlights heterogeneous immune response amongst different patients. a) PHATE cell state embedding colored by cell subtype. b) Heatmap showing mean RNA expression values of each cluster, colored by a $\log_2$ scale. c) Diffusion map embedding of samples (colored by group assignment) revealing multiple trajectories that represent increasing relative frequency of selected cell populations. d) Summary histograms, each representing the bin-wise mean relative frequency of cell subtypes for all samples assigned to a given group. The sample IDs (as assigned in the original dataset published by Tirosh et al. (12)) of all samples in each inhibitor group can be found in Table S6.

When comparing and mapping patient specimens, PhEMD identified the specimen 'Mel75' as having a unique immune cell profile characterized by the greatest proportion of exhausted CD8^{+} T-cells. These cell-state and tumor-comparison findings corroborated previously published results on the immune cell subtypes and inter-specimen heterogeneity present in this cohort (12). In addition to confirming prior findings, this analysis yielded an embedding that revealed the manifold structure of the single-cell specimen state space. With respect to a reference group of biospecimens (Cluster D) that were comprised mostly of $CD4^+$ T-cells and were mapped to one part of the manifold, three axes of variation emerged that corresponded to increasing relative proportions of B-cells (C-5, C-6), macrophages (C-7), and exhausted $CD8^+$ T-cells (C-1) (Figure 14C-D, Table S6). While it was well-understood that a set of individual cells, such as those undergoing differentiation, may demonstrate manifold structure (56, 89), our PhEMD embedding suggested that a set of patients with a shared phenotype (e.g., melanoma) may also lie on a continuous manifold (90).

To further explore this concept, we applied PhEMD to a mass cytometry dataset containing the T-cell infiltrates of 75 clear cell renal cell carcinoma (ccRCC) specimens (13). At the cellular level, our analysis recapitulated previous findings of important T-cell subpopulations present (13). Cell subtype C-1 represented cells with absent or low expression of both CD4 and CD8. C-2 through C-4 represented CD4+ T-cells with increasing expression of CD4, CD7, CCR7 and FOXP3, consistent with a regulatory T-cell profile. C-5 represented CD4+ T-cells with high Ki-67, a well-known proliferative marker. C-8 represented CD8+ cells with high expression of CD11b and CD45RA.The trajectory from C-6 to C-7 to C-9 to C-10 represented CD8+ T-cells with increasing expression of CD8, CD38, CD86, Ki-67, Tim-3, and PD-1. C-9 and C-10 cells demonstrated the highest expression of the above markers, consistent with a T-cell exhaustion profile (Figure 15A-B) (87).

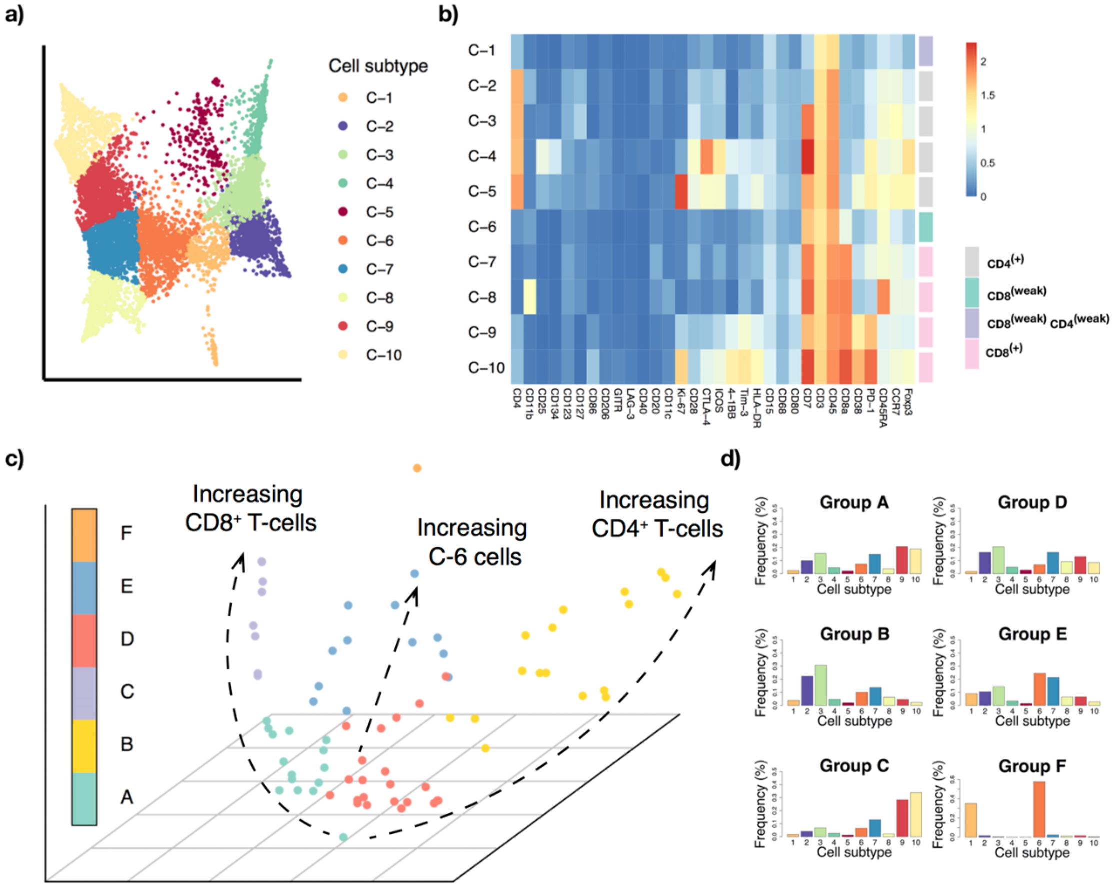


Figure 15. PhEMD applied to mass cytometry data of 75 ccRCC samples gated for T-cells. a) PHATE embedding of T-cell manifold colored by cell subtype. b) Heatmap showing mean protein expression values of each cell subtype cluster, colored by a $\log_2$ scale. c) Diffusion map embedding of all tumors colored by tumor subgroup, defined by hierarchical clustering. The main axes of inter-sample variability are highlighted as dotted-black trajectories. d) Summary histograms, each representing the bin-wise mean relative frequency of cell subtypes for all samples assigned to a given group. The sample IDs (as assigned in the original publication of these data (13)) of all samples in each inhibitor group can be found in Table S7.

We then modeled the diversity in immune cell signatures as a tumor-specimen embedding that could be used to characterize specimen-to-specimen variation (Figure 15C). A group of tumor specimens (Cluster B) mapping to one end of the PhEMD embedding was characterized by a marked predominance of CD4^{+} T-cells (C-2, C-3), and progression toward the other end of the tumor-space manifold represented a relative decrease in CD4+ T-cells and marked relative increase in CD8+ PD1+ exhausted T-cells (C-9, C-10) (Figure 15C, Table S7). This finding was supported by the initial report of substantial inter-patient variability in T-cell profiles especially related to CD8$^+$ cells (13). The detection of a subset of patients with exhausted T-cell enrichment may be of particular clinical interest, as immunotherapy agents that combat T-cell exhaustion have become a mainstay of advanced-stage ccRCC treatment, but patients continue to have highly variable treatment responses (91, 92). Future single-cell tumor-profiling experiments assessing treatment response may be able to use PhEMD as a tool to identify subgroups of patients that might especially benefit from PD-1 or PD-L1 inhibitor immunotherapy.

DISCUSSION

Here, we have demonstrated the successful mapping of single-cell experimental conditions using our proposed PhEMD embedding technique. We extensively studied the Py2T murine breast cancer cell line treated with TGF-β_1 and perturbed with over 200 kinase inhibitors, measured using mass cytometry. In this experiment, PhEMD revealed the structure of the kinase inhibitor space based on each drug's effect on the Py2T cell populations undergoing EMT. The final embedding of inhibitors was found to have low-dimensional structure, with drugs mapping to one of three main axes. We have shown that the embedding produced by PhEMD is useful in several ways:

1. Visualizing the experimental variable (i.e., single-cell specimen) state space.
2. Identifying clusters of similar experimental variable settings (e.g., similar drugs with respect to their measured effects on a given cell population).
3. Characterizing axes of variability among specimens in terms of biologically-

interpretable differences in the types and abundances of cell subpopulations present.

4. Extending the experimental variable state space through inference of unmeasured experimental settings based on similarity to existing (measured) settings.

PhEMD can enable a new paradigm of searching for effective therapeutic agents by identifying a small subset drugs that collectively capture the network geometry of a larger drug set. We demonstrated this application by computing a dictionary of 34 experimental conditions and showing that these experimental conditions were sufficient to capture the network geometry of the 300-specimen state space. This finding has the potential to reduce experimental burden in future drug discovery efforts. For example, one can first apply PhEMD to measurements obtained using one profiling technique (e.g., mass cytometry) to identify a small set of dictionary specimens from a large set of candidates and then investigate this smaller set further using complementary technologies that may be more limited in scale (e.g., single-cell RNA sequencing).

The PhEMD embedding can be integrated with additional data sources and data types for even larger and richer analyses. By using drug-target specificity data from a complementary inhibitor profiling experiment along with data imputation approaches, we were able to accurately predict the effects of inhibitors not directly measured in our experiment on TGFβ_1-induced breast cancer EMT. This approach is useful for analyzing drug-screen experiments, as it enables an initial mapping of a modest set of drugs (e.g., dictionary points) measured with single-cell resolution to be extended to include additional drugs. This application is not limited to perturbation screen data and can be useful for imputing the phenotypes of specimens (of any type) that are not directly measured using single-cell profiling. For example, examining a cohort of patients in which only some patients were biopsied and genomically profiled, one could potentially incorporate a non-genomic based measure of patient–to-patient similarity (e.g., based on clinicopathologic features) to predict the single cell-based phenotypes of all patients in the cohort.

We explored the applicability of PhEMD to other experimental designs besides drug screens by applying it to single-cell data from two clinical tumor-biopsy cohorts. These analyses revealed that PhEMD can uncover manifold structure in the tumor-specimen space that is biologically meaningful based on the observed proportions of the specimens' cell subpopulations. When applied to the melanoma and ccRCC datasets, PhEMD revealed "trajectories" of patients, with the most notable axis in both datasets consisting of patients with an increasing proportion of exhausted CD8$^+$ T-cells. It is possible that the abundance of tumor-infiltrating, exhausted T-cells may predict response to immunotherapy, although additional studies are needed to assess this. The PhEMD method may be useful for developing personalized cancer treatment regimens involving immunotherapy.

This study is not without limitations. Our approach specifically compares cell subtype *relative* abundances among biospecimens, which entails normalizing each biospecimen by its total cell count. In this setting, since relative abundances by definition sum to one for each biospecimen, the Earth Mover's Distance is a true metric and is robust across all pairwise comparisons of biospecimens. Comparing cell subtype relative abundances rather than absolute abundances is also often preferable from a biological perspective, as biospecimens (e.g., biopsy samples) may demonstrate variation in cell yield that is a technical artifact of little biological interest. Nevertheless, there exist experimental scenarios in which cell yield is of biological importance. In future work, we aim to incorporate cell yield into specimen-to-specimen comparisons and into the final biospecimen embedding. Another area of active investigation is exploring alternative methods of embedding the cell-subtype and biospecimen-state space. In the presented experiments, PHATE was used to model the cell-subtype space and diffusion maps were used to generate the biospecimen-state space. Future work may assess the utility of other methods that are potentially applicable for these tasks.

In the present study, PhEMD was used to characterize mass cytometry and single-cell RNA-sequencing data, though PhEMD may be applied to data generated by other single-cell profiling platforms as well. Many experimental designs may benefit from PhEMD—for example, comparisons of specimens pre- and post-treatment (or receiving different treatments), time-series analyses of cells undergoing transition processes, and organization of heterogeneous-yet-related specimens for the purpose of disease subtyping. Additionally, applying PhEMD to large-scale functional genomics (e.g., single-cell CRISPR) screens may yield embeddings that reveal complex relationships between genes. We have demonstrated in our analysis of over 1.7 million cells across 300 specimens and five mass cytometry runs that PhEMD is highly scalable and robust to batch effect. PhEMD offers the efficiency, flexibility, and model interpretability necessary to analyze single-cell experiments of increasingly large scale and complexity.

REFERENCES

1. Lappalainen T, Scott AJ, Brandt M, Hall IM. Genomic Analysis in the Age of Human Genome Sequencing. *Cell* 2019;177(1):70–84.

2. Zhang MY, Churpek JE, Keel SB, Walsh T, Lee MK, et al. Germline ETV6 mutations in familial thrombocytopenia and hematologic malignancy. *Nat. Genet.* 2015;47(2):180–185.

3. Chen WS, Feng EL, Aggarwal R, Foye A, Beer TM, et al. Germline polymorphisms associated with impaired survival outcomes and somatic tumor alterations in advanced prostate cancer [Internet]. *Prostate Cancer Prostatic Dis.* [published online ahead of print: November 19, 2019]; doi:10.1038/s41391-019-0188-4

4. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;490(7418):61–70.

5. Tutt A, Robson M, Garber JE, Domchek SM, Audeh MW, et al. Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and advanced breast cancer: a proof-of-concept trial. *The Lancet* 2010;376(9737):235–244.

6. Fong PC, Boss DS, Yap TA, Tutt A, Wu P, et al. Inhibition of Poly(ADP-Ribose) Polymerase in Tumors from *BRCA* Mutation Carriers. *N. Engl. J. Med.* 2009;361(2):123–134.

7. Jänne PA, Yang JC-H, Kim D-W, Planchard D, Ohe Y, et al. AZD9291 in EGFR Inhibitor–Resistant Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* 2015;372(18):1689–1699.

8. Chen WS, Bindra RS, Mo A, Hayman T, Husain Z, et al. CDKN2A Copy Number Loss Is an Independent Prognostic Factor in HPV-Negative Head and Neck Squamous Cell Carcinoma. *Front. Oncol.* 2018;8:95.

9. Chen WS, Aggarwal R, Zhang L, Zhao SG, Thomas GV, et al. Genomic Drivers of Poor Prognosis and Enzalutamide Resistance in Metastatic Castration-resistant Prostate Cancer. *Eur. Urol.* 2019;76(5):562–571.

10. Young KJ, Kay LS, Phillips MJ, Zhang L. Antitumor activity mediated by double-negative T cells. *Cancer Res.* 2003;63(22):8014–8021.

11. Overgaard NH, Jung J-W, Steptoe RJ, Wells JW. CD4 $^{+}$ /CD8 $^{+}$ double-positive T cells: more than just a developmental stage?. *J. Leukoc. Biol.* 2015;97(1):31–38.

12. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 2016;352(6282):189–196.

13. Chevrier S, Levine JH, Zanotelli VRT, Silina K, Schulz D, et al. An Immune Atlas of Clear Cell Renal Cell Carcinoma. *Cell* 2017;169(4):736-749.e18.

14. Wagner J, Rapsomaniki MA, Chevrier S, Anzeneder T, Langwieder C, et al. A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer. *Cell* 2019;177(5):1330-1345.e18.

15. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, et al. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell* 2018;174(5):1293-1308.e36.

16. Kowalczyk MS, Tirosh I, Heckl D, Rao TN, Dixit A, et al. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.* 2015;25(12):1860–1872.

17. Buganim Y, Faddah DA, Cheng AW, Itskovich E, Markoulaki S, et al. Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase. *Cell* 2012;150(6):1209–1222.

18. Krishnaswamy S, Spitzer MH, Mingueneau M, Bendall SC, Litvin O, et al. Conditional density-based analysis of T cell signaling in single-cell data. *Science* 2014;346(6213):1250689–1250689.

19. Cacchiarelli D, Qiu X, Srivatsan S, Manfredi A, Ziller M, et al. Aligning Single-Cell Developmental and Reprogramming Trajectories Identifies Molecular Determinants of Myogenic Reprogramming Outcome. *Cell Syst.* 2018;7(3):258-268.e3.

20. Bellman R. *Dynamic programming*. Princeton, NJ: Princeton Univ. Pr; 1984:

21. Bellman RE, Dreyfus SE. *Applied dynamic programming*. Princeton, N.J: Princeton Univ. Press; 1971:

22. Jolliffe IT. Principal Component Analysis and Factor Analysis [Internet]. In: *Principal Component Analysis*. New York, NY: Springer New York; 1986:115–128

23. van der Maaten L, Hinton G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* 2008;9:2579–2605.

24. Abraham A, Dutta P, Mandal JK, Bhattacharya A, Dutta S. *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 2 [Internet]*. 2019:

25. Amir ED, Davis KL, Tadmor MD, Simonds EF, Levine JH, et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* 2013;31(6):545–552.

26. van der Maaten L. Barnes-Hut-SNE [Internet]. *ArXiv13013342 Cs Stat* [published online ahead of print: March 8, 2013];http://arxiv.org/abs/1301.3342. cited February 1, 2020

27. Linderman GC, Rachh M, Hoskins JG, Steinerberger S, Kluger Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat. Methods* 2019;16(3):243–245.

28. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* 2019;10(1):5416.

29. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction [Internet]. *ArXiv180203426 Cs Stat* [published online ahead of print: December 6, 2018];http://arxiv.org/abs/1802.03426. cited February 1, 2020

30. Kobak D, Linderman GC. *UMAP does not preserve global structure any better than t-SNE when using the same initialization [Internet]*. Bioinformatics; 2019:

31. Qiu P, Simonds EF, Bendall SC, Gibbs KD, Bruggner RV, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.* 2011;29(10):886–891.

32. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* 2016;34(6):637–645.

33. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* 2017;14(10):979–982.

34. Coifman RR, Lafon S. Diffusion maps. *Appl. Comput. Harmon. Anal.* 2006;21(1):5–30.

35. Haghverdi L, Buettner F, Theis FJ. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* 2015;31(18):2989–2998.

36. Angerer P, Haghverdi L, Büttner M, Theis FJ, Marr C, et al. destiny: diffusion maps for large-scale single-cell data in R. *Bioinforma. Oxf. Engl.* 2016;32(8):1241–1243.

37. Moon KR, van Dijk D, Wang Z, Gigante S, Burkhardt DB, et al. Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* 2019;37(12):1482–1492.

38. Bodenmiller B, Zunder ER, Finck R, Chen TJ, Savig ES, et al. Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. *Nat. Biotechnol.* 2012;30(9):858–867.

39. Lavin Y, Kobayashi S, Leader A, Amir ED, Elefant N, et al. Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses. *Cell* 2017;169(4):750-765.e17.

40. Ribas A, Shin DS, Zaretsky J, Frederiksen J, Cornish A, et al. PD-1 Blockade Expands Intratumoral Memory T Cells. *Cancer Immunol. Res.* 2016;4(3):194–203.

41. Behbehani GK, Samusik N, Bjornson ZB, Fantl WJ, Medeiros BC, et al. Mass Cytometric Functional Profiling of Acute Myeloid Leukemia Defines Cell-Cycle and Immunophenotypic Properties That Correlate with Known Responses to Therapy. *Cancer Discov.* 2015;5(9):988–1003.

42. Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, et al. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* 2019;176(1–2):377-390.e19.

43. Alpert A, Moore LS, Dubovik T, Shen-Orr SS. Alignment of single-cell trajectories to compare cellular expression dynamics. *Nat. Methods* 2018;15(4):267–270.

44. Liu Q, Herring CA, Sheng Q, Ping J, Simmons AJ, et al. Quantitative assessment of cell population diversity in single-cell landscapes. *PLOS Biol.* 2018;16(10):e2006687.

45. Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. *J. Clin. Invest.* 2009;119(6):1420–1428.

46. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir ED, et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* 2015;162(1):184–197.

47. Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* 2016;13(10):845–848.

48. Sachs K. Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science* 2005;308(5721):523–529.

49. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* 2017;14(5):483–486.

50. Rubner Y, Tomasi C, Guibas LJ. The Earth Mover’s Distance as a Metric for Image Retrieval. *Int J Comput Vis* 2000;40(2):99–121.

51. Liu LL, Landskron J, Ask EH, Enqvist M, Sohlberg E, et al. Critical Role of CD2 Co-stimulation in Adaptive Natural Killer Cell Responses Revealed in NKG2C-Deficient Humans. *Cell Rep.* 2016;15(5):1088–1099.

52. Wang F, Guibas LJ. Supervised Earth Mover’s Distance Learning and Its Computer Vision Applications [Internet]. In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C eds. *Computer Vision – ECCV 2012*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012:442–455

53. Qi Zhao, Zhi Yang, Hai Tao. Differential Earth Mover's Distance with Its Applications to Visual Tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* 2010;32(2):274–287.

54. Typke R, Wiering F, Veltkamp RC. Transportation distances and human perception of melodic similarity. *Music. Sci.* 2007;11(1_suppl):153–181.

55. Orlova DY, Zimmerman N, Meehan S, Meehan C, Waters J, et al. Earth Mover's Distance (EMD): A True Metric for Comparing Biomarker Expression Levels in Cell Populations. *PLOS ONE* 2016;11(3):e0151859.

56. Moon KR, Stanley JS, Burkhardt D, van Dijk D, Wolf G, et al. Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Curr. Opin. Syst. Biol.* 2018;7:36–46.

57. Courty N, Flamary R, Ducoffe M. Learning Wasserstein Embeddings [Internet]. *ArXiv171007457 Cs Stat* [published online ahead of print: October 20, 2017];http://arxiv.org/abs/1710.07457. cited June 21, 2018

58. Waldmeier L, Meyer-Schaller N, Diepenbruck M, Christofori G. Py2T Murine Breast Cancer Cells, a Versatile Model of TGFβ-Induced EMT In Vitro and In Vivo. *PLoS ONE* 2012;7(11):e48651.

59. Ornatsky O, Bandura D, Baranov V, Nitz M, Winnik MA, et al. Highly multiparametric analysis by mass cytometry. *J. Immunol. Methods* 2010;361(1–2):1–20.

60. Finck R, Simonds EF, Jager A, Krishnaswamy S, Sachs K, et al. Normalization of mass cytometry data with bead standards. *Cytometry A* 2013;83A(5):483–494.

61. Zunder ER, Finck R, Behbehani GK, Amir ED, Krishnaswamy S, et al. Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell deconvolution algorithm. *Nat. Protoc.* 2015;10(2):316–333.

62. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 2018;36(5):411–420.

63. Levina E, Bickel P. Maximum likelihood estimation of intrinsic dimension. In: Saul L, Weiss Y, Bottou L eds. *Advances in Neural Information Processing Systems*. MIT Press; 2005:

64. Hino H. *ider: Various Methods for Estimating Intrinsic Dimension [Internet]*. 2017:

65. Salhov M, Bermanis A, Wolf G, Averbuch A. Approximately-isometric diffusion maps. *Appl. Comput. Harmon. Anal.* 2015;38(3):399–419.

66. Klaeger S, Heinzlmeir S, Wilhelm M, Polzer H, Vick B, et al. The target landscape of clinical kinase drugs. *Science* 2017;358(6367):eaan4368.

67. van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* 2018;174(3):716-729.e27.

68. Mevik B-H, Wehrens R, Liland KH. *pls: Partial Least Squares and Principal Component Regression [Internet]*. 2019:

69. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* 2017;18(1):174.

70. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* 2018;36(5):421–427.

71. Shaham U, Stanton KP, Zhao J, Li H, Raddassi K, et al. Removal of batch effects using distribution-matching residual networks. *Bioinformatics* 2017;33(16):2539–2546.

72. Amodio M, van Dijk D, Srinivasan K, Chen WS, Mohsen H, et al. Exploring single-cell data with deep multitasking neural networks. *Nat. Methods* 2019;16(11):1139–1145.

73. Mani SA, Guo W, Liao M-J, Eaton ENg, Ayyanan A, et al. The Epithelial-Mesenchymal Transition Generates Cells with Properties of Stem Cells. *Cell* 2008;133(4):704–715.

74. Zhu H, Mitsuhashi N, Klein A, Barsky LW, Weinberg K, et al. The role of the hyaluronan receptor CD44 in mesenchymal stem cell migration in the extracellular matrix. *Stem Cells Dayt. Ohio* 2006;24(4):928–935.

75. L Ramos T, Sánchez-Abarca LI, Muntión S, Preciado S, Puig N, et al. MSC surface markers (CD44, CD73, and CD90) can identify human MSC-derived extracellular vesicles by conventional flow cytometry. *Cell Commun. Signal. CCS* 2016;14:2.

76. Ivaska J, Pallari H-M, Nevo J, Eriksson JE. Novel functions of vimentin in cell adhesion, migration, and signaling. *Exp. Cell Res.* 2007;313(10):2050–2062.

77. Li W, Ma H, Zhang J, Zhu L, Wang C, et al. Unraveling the roles of CD44/CD24 and ALDH1 as cancer stem cell markers in tumorigenesis and metastasis. *Sci. Rep.* 2017;7(1):13856.

78. Ma F, Li H, Wang H, Shi X, Fan Y, et al. Enriched CD44(+)/CD24(-) population drives the aggressive phenotypes presented in triple-negative breast cancer (TNBC). *Cancer Lett.* 2014;353(2):153–159.

79. Ricardo S, Vieira AF, Gerhard R, Leitão D, Pinto R, et al. Breast cancer stem cell markers CD44, CD24 and ALDH1: expression distribution within intrinsic molecular subtype. *J. Clin. Pathol.* 2011;64(11):937–946.

80. Yu M, Bardia A, Wittner BS, Stott SL, Smas ME, et al. Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *Science* 2013;339(6119):580–584.

81. Nieto MA, Huang RY-J, Jackson RA, Thiery JP. EMT: 2016. *Cell* 2016;166(1):21–45.

82. Jolly MK, Boareto M, Huang B, Jia D, Lu M, et al. Implications of the Hybrid Epithelial/Mesenchymal Phenotype in Metastasis. *Front. Oncol.* 2015;5:155.

83. Elkabets M, Vora S, Juric D, Morse N, Mino-Kenudson M, et al. mTORC1 inhibition is required for sensitivity to PI3K p110α inhibitors in PIK3CA-mutant breast cancer. *Sci. Transl. Med.* 2013;5(196):196ra99.

84. Bengio Y, Paiement J-F, Vincent P, Delalleau O, Roux NL, et al. Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering [Internet]. In: *NIPS*. 2003:177–184

85. Fowlkes C, Belongie S, Chung F, Malik J. Spectral grouping using the Nyström method. *IEEE Trans. Pattern Anal. Mach. Intell.* 2004;26(2):214–225.

86. Williams CKI, Seeger M. Using the Nyström Method to Speed Up Kernel Machines [Internet]. In: Leen TK, Dietterich TG, Tresp V eds. *Advances in Neural Information Processing Systems 13*. MIT Press; 2001:682–688

87. Anderson AC, Joller N, Kuchroo VK. Lag-3, Tim-3, and TIGIT: Co-inhibitory Receptors with Specialized Functions in Immune Regulation. *Immunity* 2016;44(5):989–1004.

88. Etzerodt A, Moestrup SK. CD163 and Inflammation: Biological, Diagnostic, and Therapeutic Aspects. *Antioxid. Redox Signal.* 2013;18(17):2352–2363.

89. Bendall SC, Davis KL, Amir ED, Tadmor MD, Simonds EF, et al. Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development. *Cell* 2014;157(3):714–725.

90. Damond N, Engler S, Zanotelli VRT, Schapiro D, Wasserfall CH, et al. A Map of Human Type 1 Diabetes Progression by Imaging Mass Cytometry. *Cell Metab.* 2019;29(3):755-768.e5.

91. Hammers HJ, Plimack ER, Infante JR, Rini BI, McDermott DF, et al. Safety and Efficacy of Nivolumab in Combination With Ipilimumab in Metastatic Renal Cell Carcinoma: The CheckMate 016 Study. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* 2017;35(34):3851–3858.

92. Motzer RJ, Tannir NM, McDermott DF, Arén Frontera O, Melichar B, et al. Nivolumab plus Ipilimumab versus Sunitinib in Advanced Renal-Cell Carcinoma. *N. Engl. J. Med.* 2018;378(14):1277–1290.

SUPPLEMENTARY TABLES

Table S1. List of inhibitors included in EMT drug-screen experiment.

Compound in DMSO	Cas Number	Reported Target	Catalog Number	Plate / Well
Axitinib	319460-85-0	VEGFR, PDGFR, c-Kit	S1005	A / B2
Dovitinib (TKI-258)	405169-16-6, 804551-71-1	FLT3	S1018	A / B3
Lapatinib Ditosylate (Tykerb)	388082-77-7, 231277-92-2 (free base), 1187538-35-7 (4-methylbenzenesulfonate)	EGFR, HER2	S1028	A / B4
Sorafenib (Nexavar)	475207-59-1, 284461-73-0 (free base)	VEGFR, PDGFR, Raf	S1040	A / B5
BMS-599626 (AC480)	714971-09-2, 873837-23-1 (HCl), 873837-22-0 (H_2O)	EGFR, HER2	S1056	A / B6
SB 203580	152121-47-6, 224047-03-4, 869185-85-3 (HCl)	p38 MAPK	S1076	A / B7
KU-55933	587871-26-9	ATM	S1092	A / B8
LY294002	154447-36-6, 934389-88-5 (HCl)	PI3K	S1105	A / B9
JNJ-38877605	943540-75-8, 1093204-17-1 (X methanesulfonate), 1093204-20-6 (XHCl)	c-Met	S1114	A / B10
Brivanib alaninate (BMS-582664)	649735-63-7	VEGFR	S1138	A / B11
Saracatinib (AZD0530)	379231-04-6, 893428-72-3 (Fumaric acid), 893428-71-2 ($3H_2O$)	Src, Bcr-Abl	S1006	A / C2
CI-1033 (Canertinib)	267243-28-7, 289499-45-2 (2HCl)	EGFR, HER2	S1019	A / C3
Motesanib Diphosphate (AMG-706)	857876-30-3, 453562-69-1 (free base), 945716-97-2 (H_2O)	VEGFR, PDGFR, c-Kit	S1032	A / C4
Sunitinib Malate (Sutent)	341031-54-7, 557795-19-4 (free base), 1126641-10-8 (Maleic acid)	VEGFR, PDGFR, c-Kit, Flt	S1042	A / C5
Masitinib (AB1010)	790299-79-5, 1048007-93-7 (methanesulfonate)	c-Kit, PDGFR, FGFR, FAK	S1064	A / C6
SB 202190	152121-30-7, 350228-36-3 (HCl)	p38 MAPK	S1077	A / C7
GSK1904529A	1089283-49-7	IGF-1R	S1093	A / C8
OSU-03012	742112-33-0	PDK-1	S1106	A / C9
PD 0332991 (Palbociclib) HCl	827022-32-2, 571190-30-2 (free base)	CDK	S1116	A / C10
AG-490	133550-30-8	JAK, EGFR	S1143	A / C11

AZD6244 (Selumetinib)	606143-52-6, 942275-12-9 (4-methylbenzenesulfonate)	MEK	S1008	A / D2
CI-1040 (PD184352)	212631-79-3	MEK	S1020	A / D3
Nilotinib (AMN-107)	641571-10-0	Bcr-Abl	S1033	A / D4
Tandutinib (MLN518)	387867-13-2, 1227636-16-9 (3H2O)	Flt	S1043	A / D5
GDC-0941	957054-30-7, 957054-33-0 (dimethanesulfonate), 957054-54-5 (xTFA)	PI3K	S1065	A / D6
MK-2206 2HCl	1032350-13-2, 1032349-93-1 (free base), 1032349-77-1 (HCl)	Akt	S1078	A / D7
PF-04217903	956905-27-4, 956906-93-7 (methanesulfonate), 1159490-81-9 (HCl)	c-Met	S1094	A / D8
Danusertib (PHA-739358)	827318-97-8	Aurora Kinase, FGFR, Bcr-Abl, c-RET, Src	S1107	A / D9
Triciribine (Triciribine phosphate)	35943-35-2	Akt	S1117	A / D10
SNS-032 (BMS-387032)	345627-80-7, 345627-90-9 (HCl)	CDK	S1145	A / D11
BIBF1120 (Vargatef)	656247-17-5, 790241-30-4 (methanesulfonate), 959761-73-0 (HCl)	VEGFR, PDGFR, FGFR	S1010	A / F2
Deforolimus (Ridaforolimus)	572924-54-0, 697252-87-2	mTOR	S1022	A / F3
PD0325901	391210-10-9, 870474-62-7	MEK	S1036	A / F4
Vandetanib (Zactima)	443913-73-3, 338992-53-3 (TFA), 524722-52-9 (HCl)	VEGFR	S1046	A / F5
Crizotinib (PF-02341066)	877399-52-5, 877399-53-6 (acetate)	c-Met, ALK	S1068	A / F6
SU11274	658084-23-2	c-Met	S1080	A / F7
Vatalanib 2HCl (PTK787)	212141-51-0, 212141-54-3 (free base)	VEGFR, c-Kit, Flt	S1101	A / F8
BI 2536	755038-02-9, 876126-71-5 (H2O)	PLK	S1109	A / F9
XL-184 (Cabozantinib)	849217-68-1,1140909-48-3 (L-(-)-Apple Acid)	VEGFR, c-Met, Flt, Tie-2, c-Kit	S1119	A / F10
PLX-4720	918505-84-7	Raf	S1152	A / F11
Afatinib (BIBW2992)	439081-18-2, 936631-70-8 (Maleic acid), 1254955-21-9 (XHCl)	EGFR, HER2	S1011	A / G2

Erlotinib HCl	183319-69-9, 183321-74-6 (free base), 248594-19-6 (methanesulfonate)	EGFR	S1023	A / G3
VX-680 (MK-0457, Tozasertib)	639089-54-6, 639090-58-7 (sulfate)	Aurora Kinase	S1048	A / G5
PHA-665752	477575-56-7, 1262750-60-6(HCl)	c-Met	S1070	A / G6
Brivanib (BMS-540215)	649735-46-6	VEGFR	S1084	A / G7
U0126-EtOH	1173097-76-1, 109511-58-2 (FREE BASE)	MEK	S1102	A / G8
Foretinib (GSK1363089, XL880)	849217-64-7, 1332889-22-1 (H2O)	c-Met, VEGFR	S1111	A / G9
Everolimus (RAD001)	159351-69-6, 1245613-55-1	mTOR	S1120	A / G10
Roscovitine (Seliciclib, CYC202)	186692-46-6	CDK	S1153	A / G11
CP-724714	537705-08-1	EGFR, HER2	S1167	B / B2
ENMD-2076	1291074-87-7	Flt, Aurora Kinase, VEGFR	S1181	B / B3
Amuvatinib (MP-470)	850879-09-3	c-Met, c-Kit, PDGFR, Flt, c-RET	S1244	B / B4
AMG-208	1002304-34-8	c-Met	S1316	B / B5
AS-605240	648450-29-7	PI3K	S1410	B / B6
AS703026 (pimasertib)	1236699-92-5, 1236361-78-6 (HCl)	MEK	S1475	B / B7
CCT129202	942947-93-5	Aurora Kinase	S1519	B / B8
R406 (free base)	841290-80-0	Syk	S1533	B / B9
KU-60019	925701-49-1	ATM	S1570	B / B10
KW 2449	1000669-72-6	Flt, Bcr-Abl, Aurora Kinase	S2158	B / B11
TGX-221	663619-89-4	PI3K	S1169	B / C2
PIK-90	677338-12-4	677338-12-4	S1187	B / C3
JNJ-7706621	443797-96-4	CDK, Aurora Kinase	S1249	B / C4
TG100-115	677297-51-7, 677297-55-1 (2HCL)	PI3K	S1352	B / C5
Staurosporine	62996-74-1	PKC	S1421	B / C6
SB 525334	356559-20-1	TGF-beta/Smad	S1476	B / C7

XL765	1349796-36-6, 1123889-87-1	PI3K, mTOR	S1523	B / C8
CP 673451	343787-29-1, 343787-32-6 (4-methylbenzenesulfonate)	PDGFR	S1536	B / C9
BS-181 HCl	1092443-52-1 (free base)	CDK	S1572	B / C10
WZ3146	1214265-56-1	EGFR	S1170	B / D2
PIK-75	372196-77-5, 372196-67-3 (free base)	PI3K, DNA-PK	S1205	B / D3
PD173074	219580-11-7	FGFR	S1264	B / D4
GSK1059615	958852-01-2, 1356195-42-0 (H2O . Na)	PI3K, mTOR	S1360	B / D5
Aurora A Inhibitor I	1158838-45-9	Aurora Kinase	S1451	B / D6
HMN-214	173529-46-9	PLK	S1485	B / D7
AT7519	844442-38-2, 902135-91-5 (HCl), 902135-89-1 (methanesulfonate)	CDK	S1524	B / D8
AZD8055	1009298-09-2, 1201799-04-3 (D(-)-Tartaric Acid), 1201799-05-4 (Fumaric acid)	mTOR	S1555	B / D9
BIRB 796 (Doramapimod)	285983-48-4, 1283526-53-3 (HCl)	p38 MAPK	S1574	B / D10
LY2784544	1229236-86-5, 1229236-87-6 (HCl)	JAK	S2179	B / D11
WZ4002	1213269-23-8	EGFR	S1173	B / F2
YM201636	371942-69-7, 371933-96-9 (2HCl)	PI3K	S1219	B / F3
Vemurafenib (PLX4032)	918504-65-1	Raf	S1267	B / F4
ON-01910	1225497-78-8	PLK	S1362	B / F5
Thiazovivin	1226056-71-8, 1228446-06-7 (TFA)	ROCK	S1459	B / F6
PHA-793887	718630-59-2, 718630-60-5 (HCl)	CDK	S1487	B / F7
Hesperadin	422513-13-1	Aurora Kinase	S1529	B / F8
KRN 633	286370-15-8	VEGFR, PDGFR	S1557	B / F9
TWS119	601514-19-6	GSK-3	S1590	B / F10
AST-1306	1050500-29-2, 897383-62-9 (free base)	EGFR	S2185	B / F11
PD98059	167869-21-8	MEK	S1177	B / G2

OSI-930	728033-96-3	c-Kit, VEGFR	S1220	B / G3
IC-87114	371242-69-2	PI3K	S1268	B / G4
Ki8751	228559-41-9	VEGFR, c-Kit, PDGFR	S1363	B / G5
SP600125	129-56-6, 67072-00-8 (potassium salt)	JNK	S1460	B / G6
PIK-93	593960-11-3	PI3K, VEGFR	S1489	B / G7
BIX 02188	1094614-84-2	MEK	S1530	B / G8
AT7867	857531-00-1	Akt, S6 kinase	S1558	B / G9
BMS-265246	582315-72-8	CDK	S2014	B / G10
AZD8931	848942-61-0, 1196531-39-1 (diFumaric acid)	EGFR, HER2	S2192	B / G11
Raf265 derivative	927880-90-8	VEGFR, Raf	S2200	C / B2
PP242	1092351-67-1, 1173019-76-5 (H2O)	mTOR	S2218	C / B3
Palomid 529	914913-88-5	PI3K	S2238	C / B4
TAK-733	1035555-63-5	MEK	S2617	C / B5
DCC-2036 (Rebastinib)	1020172-07-9, 1020172-08-0 (2HCl), 1033893-29-6 (4-methylbenzenesulfonate)	Bcr-Abl	S2634	C / B6
AS-252424	900515-16-4	PI3K	S2671	C / B7
NVP-BSK805	1092499-93-8 (free base)	JAK	S2686	C / B8
AMG 900	945595-80-2	Aurora Kinase	S2719	C / B9
AZ628	878739-06-1	Raf	S2746	C / B11
BMS 794833	1174046-72-0, 1174161-83-1 (HCl)	c-Met, VEGFR	S2201	C / C2
Cyt387	1056634-68-4, 1056636-08-8 (XHCl)	JAK	S2219	C / C3
WP1130	856243-80-6	DUB, Bcr-Abl	S2243	C / C4
LDN193189	1062368-24-4, 1062368-62-0 (HCl)	TGF-beta/Smad	S2618	C / C5
CCT128930	885499-61-6	Akt	S2635	C / C6
PF-00562271	939791-38-5, 717907-75-0 (free base), 939791-39-6 (methanesulfonate)	FAK	S2672	C / C7
WAY-600	1062159-35-6	mTOR	S2689	C / C8
ZM 336372	208260-29-1	Raf	S2720	C / C9

TG101348 (SAR302503)	936091-26-8, 1374744-69-0 (2ClH.H2O)	JAK	S2736	C / C10
AMG458	913376-83-7	c-Met	S2747	C / C11
NVP-BHG712	940310-85-0	VEGFR, Src, Raf, Bcr-Abl	S2202	C / D2
SB590885	405554-55-4	Raf	S2220	C / D3
BKM120 (NVP-BKM120)	944396-07-0, 1312445-63-8 (HCl), 1370351-44-2 (0.5H2O)	PI3K	S2247	C / D4
AZD5438	602306-29-6	CDK	S2621	C / D5
A66	1166227-08-2	PI3K	S2636	C / D6
GSK1120212 (Trametinib)	871700-17-3, 871702-06-6 (sodium salt)	MEK	S2673	C / D7
TG101209	936091-14-4	Flt, JAK, c-RET	S2692	C / D8
PF-03814735	942487-16-3	Aurora Kinase	S2725	C / D9
PKI-402	1173204-81-3, 1173204-82-4 (XHCl)	PI3K	S2739	C / D10
NVP-BGT226	1245537-68-1, 915020-55-2 (free base)	PI3K	S2749	C / D11
R935788 (Fostamatinib disodium, R788 disodium)	1025687-58-4, 901119-35-5 (free base),1180490-89-4 (acetate)	Syk	S2206	C / F2
CAL-101 (GS-1101)	870281-82-6	PI3K	S2226	C / F3
Indirubin	479-41-4	GSK-3	S2386	C / F4
OSI-027	936890-98-1, 1187559-66-5 (sodium salt)	mTOR	S2624	C / F5
GSK2126458	1086062-66-9	PI3K, mTOR	S2658	C / F6
PCI-32765 (Ibrutinib)	936563-96-1	Src	S2680	C / F7
A-769662	844499-71-4	AMPK	S2697	C / F8
Dacomitinib (PF299804,PF-00299804)	1110813-31-4, 1042385-75-0 (H2O)	EGFR	S2727	C / F9
PHA-767491	845714-00-3, 942425-68-5 (HCl)	CDK	S2742	C / F10
Arry-380	937265-83-3	HER2	S2752	C / F11
PIK-293	900185-01-5	PI3K	S2207	C / G2

PIK-294	900185-02-6	PI3K	S2227	C / G3
Quercetin (Sophoretin)	117-39-5	PI3K, PKC, Src, Sirtuin	S2391	C / G4
R788 (Fostamatinib)	901119-35-5	Syk	S2625	C / G5
WYE-125132	1144068-46-1	mTOR	S2661	C / G6
AS-604850	648449-76-7	PI3K	S2681	C / G7
KX2-391	897016-82-9, 1038395-65-1 (2HCl), 1080645-95-9 (methanesulfonate), 1201926-60-4 (Maleic acid)	Src	S2700	C / G8
AG-1478 (Tyrphostin AG-1478)	153436-53-4, 170449-18-0 (HCl)	EGFR	S2728	C / G9
PF-04691502	1013101-36-4	mTOR, PI3K, Akt	S2743	C / G10
ARQ 197 (Tivantinib)	905854-02-6, 1000873-98-2, 1228508-24-4	c-Met	S2753	C / G11
NVP-BVU972	1185763-69-2	c-Met	S2761	D / B2
TAK-285	871026-44-7, 871027-78-0 (methanesulfonate)	EGFR	S2784	D / B3
GDC-0068	1001264-89-6	Akt	S2808	D / B4
Desmethyl Erlotinib (CP-473420)	183321-86-0, 183320-51-6 (HCl)	EGFR	S2826	D / B5
TG 100713	925705-73-3	PI3K	S2870	D / B6
Wortmannin	19545-26-7, 1405-03-4	PI3K	S2758	D / B7
AZD2014	1009298-59-2	mTOR	S2783	D / B8
Dabrafenib (GSK2118436)	1195765-45-7, 1195768-06-9 (methanesulfonic acid)	Raf	S2807	D / B9
TPCA-1	507475-17-4	IKK	S2824	D / B10
WHI-P154	211555-04-3, 296234-84-9 (HCl)	JAK	S2867	D / B11
CH5424802	1256580-46-7, 1256589-74-8 (HCl)	ALK	S2762	D / C2
INCB28060	1029712-80-8, 1029714-89-3 (XHCl), 1197376-85-4 (2HCl)	c-Met	S2788	D / C3
INK 128 (MLN0128)	1224844-38-5	mTOR	S2811	D / C4
Torin 1	1222998-36-8	mTOR	S2827	D / C5
Piceatannol	10083-24-6, 21100-92-5	Syk	S3026	D / C6

Dinaciclib (SCH727965)	779353-01-4	CDK	S2768	D / C7
Sotrastaurin (AEB071)	425637-18-9, 1058706-32-3 (HCl), 1058706-35-6 (Maleic acid)	PKC	S2791	D / C8
Tyrphostin AG 879 (AG 879)	148741-30-4	HER2	S2816	D / C9
Semaxanib (SU5416)	194413-58-6	VEGFR	S2845	D / C10
VX-702	745833-23-2, 479543-46-9	p38 MAPK	S6005	D / C11
3-Methyladenine	5142-23-4, 80681-18-1(HCl)	PI3K	S2767	D / D2
Tofacitinib (CP-690550, Tasocitinib)	477600-75-2, 540737-29-9 (citrate)	JAK	S2789	D / D3
BYL719	1217486-61-7	PI3K	S2814	D / D4
SAR131675	1433953-83-3	VEGFR	S2842	D / D5
Tofacitinib citrate (CP-690550 citrate)	540737-29-9, 477600-75-2 (free base)	JAK	S5001	D / D6
MK-2461	917879-39-1, 1196681-15-8, 1170702-87-0 (sodium salt)	c-Met	S2774	D / D7
CEP33779	1257704-57-6	JAK	S2806	D / D8
Tideglusib	865854-05-3	GSK-3	S2823	D / D9
IMD 0354	978-62-1, 634914-41-3 (sodium salt)	IKK	S2864	D / D10
Dovitinib Dilactic acid (TKI258 Dilactic acid)	852433-84-2, 405169-16-6 (free base)	FLT3	S2769	D / F2
WP1066	857064-38-1	JAK	S2796	D / F3
Torin 2	1223001-51-1	mTOR	S2817	D / F4
Baricitinib (LY3009104,incb28050)	1187594-09-7, 1187594-10-0 (TFA)	JAK	S2851	D / F5
MK-5108 (VX-689)	1010085-13-8	Aurora Kinase	S2770	D / G2
AZD4547	1035270-39-3	FGFR	S2801	D / G3
NVP-TAE226	761437-28-9	FAK	S2820	D / G4
Golvatinib (E7050)	928037-13-2 , 1007601-96-8 (L(+)-Tartaric Acid), 1007601-91-3 (Fumaric acid)	c-Met	S2859	D / G5
Linifanib (ABT-869)	796967-16-3	PDGFR, VEGFR	S1003	E / B2

Cediranib (AZD2171)	288383-20-0	VEGFR, Flt	S1017	E / B3
Imatinib Mesylate	220127-57-1	PDGFR, c-Kit, Bcr-Abl	S1026	E / B4
Rapamycin (Sirolimus)	53123-88-9	mTOR	S1039	E / B5
Enzastaurin (LY317615)	170364-57-5	PKC	S1055	E / B6
SB 216763	280744-09-4	GSK-3	S1075	E / B7
Linsitinib (OSI-906)	867160-71-2	IGF-1R	S1091	E / B8
GDC-0879	905281-76-7	Raf	S1104	E / B9
GSK690693	937174-76-0	Akt	S1113	E / B10
AT9283	896466-04-9	Bcr-Abl, JAK, Aurora Kinase	S1134	E / B11
BEZ235 (NVP-BEZ235)	915019-65-7	mTOR, PI3K	S1009	E / C2
Dasatinib (BMS-354825)	302962-49-8	Src, Bcr-Abl, c-Kit	S1021	E / C3
Pazopanib HCl	635702-64-6	VEGFR, PDGFR, c-Kit	S1035	E / C4
Temsirolimus (Torisel)	162635-04-3	mTOR	S1044	E / C5
SB 431542	301836-41-9	TGF-beta/Smad	S1067	E / C6
PD153035 HCl	183322-45-4	EGFR	S1079	E / C7
MLN8054	869363-13-3	Aurora Kinase	S1100	E / C8
TAE684 (NVP-TAE684)	761439-42-3	ALK	S1108	E / C9
XL147	956958-53-5	PI3K	S1118	E / C10
Barasertib (AZD1152-HQPA)	722544-51-6	Aurora Kinase	S1147	E / C11
Bosutinib (SKI-606)	380843-75-4	Src	S1014	E / D2
Gefitinib (Iressa)	184475-35-2	EGFR	S1025	E / D3
PI-103	371935-74-9	DNA-PK, PI3K, mTOR	S1038	E / D4
Y-27632 2HCl	129830-38-2	ROCK	S1049	E / D5
ZSTK474	475110-96-4	PI3K	S1072	E / D6
NVP-ADW742	475488-23-4	IGF-1R	S1088	E / D7

ZM-447439	331771-20-1	Aurora Kinase	S1103	E / D8
SGX-523	1022150-57-7	c-Met	S1112	E / D9
MLN8237 (Alisertib)	1028486-01-2	Aurora Kinase	S1133	E / D10
SNS-314	1146618-41-8	Aurora Kinase	S1154	E / D11
E7080 (Lenvatinib)	417716-92-8	VEGFR	S1164	E / F2
WZ8040	1214265-57-2	EGFR	S1179	E / F3
AG-1024	65678-07-1	IGF-1R	S1234	E / F4
BX-912	702674-56-4	PDK-1	S1275	E / F5
Pelitinib (EKB-569)	257933-82-7	EGFR	S1392	E / F6
TSU-68	252916-29-3	VEGFR, PDGFR , FGFR	S1470	E / F7
LY2228820	862507-23-1	p38 MAPK	S1494	E / F8
AZD7762	860352-01-8	Chk	S1532	E / F9
PD318088	391210-00-7	MEK	S1568	E / F10
Neratinib (HKI-272)	698387-09-6	HER2, EGFR	S2150	E / F11
CYC116	693228-63-6	Aurora Kinase, VEGFR	S1171	E / G2
Tivozanib (AV-951)	475108-18-0	VEGFR, c-Kit, PDGFR	S1207	E / G3
WYE-354	1062169-56-5	mTOR	S1266	E / G4
MGCD-265	875337-44-3	c-Met, VEGFR, Tie-2	S1361	E / G5
PHA-680632	398493-79-3	Aurora Kinase	S1454	E / G6
AEE788 (NVP-AEE788)	497839-62-0	EGFR, Flt, VEGFR, HER2	S1486	E / G7
Quizartinib (AC220)	950769-58-1	Flt	S1526	E / G8
PHT-427	1191951-57-1	Akt	S1556	E / G9
Tie2 kinase inhibitor	948557-43-5	Tie-2	S1577	E / G10
BGJ398 (NVP-BGJ398)	872511-34-7	FGFR	S2183	E / G11

Table S2. List of antibodies included in EMT drug-screen experiment.

Isotope	Target	Clone	Clone Reactivity	Clone Applications	Manufacturer	Lot	Description and Clone Validation (Manufacturer)	Staining Concentration [µg/ml]	User Clone Validation (Py2T by Mass Cytometry)
La139	Purified Mouse Anti-CREB (pS133) / ATF-1 (pS63)	J151-21	Human, Mouse, Rat (predicted)	WB, **FC**	BD	558359	https://www.bdbiosciences.com/us/reagents/research/antibodies-buffers/cell-biology-reagents/cell-biology-antibodies/purified-mouse-anti-creb-ps133-atf-1-ps63-j151-21/p/558359	2	Py2T, MEK1/2 signaling perturbation
Pr141	pStat5 (pTyr694)	47	Mouse; Human	WB, **FC**	BD	2150654	https://www.bdbiosciences.com/us/applications/research/stem-cell-research/stem-cell-signaling/human/purified-mouse-anti-human-stat5-py694-47stat5py694/p/611965	4.9	30 min vanadate treatment, 125µM vs 30 min Untreated
Nd142	pSHP2 (pTyr580)	D66F10	Human; Mouse; Rat	WB, IP, **FC**	CST	2	https://www.cellsignal.com/products/primary-antibodies/phospho-shp-2-tyr580-d66f10-rabbit-mab/5431	4	18 h TPA vs Untreated
Nd143	pFAK (pTyr397)	poly7	Human; Mouse	WB	CST	5	https://www.cellsignal.com/products/primary-antibodies/phospho-fak-tyr397-antibody/3283?site-search-type=Products	2.5	5 Days 4 ng/mL TGFb vs 5 Days Untreated
Nd144	MEK1/2 (pSer221)	166F8	Human; Mouse	WB, IHC, **FC**	CST	13	https://www.cellsignal.com/products/primary-antibodies/phospho-mek1-2-ser221-166f8-rabbit-mab/2338?site-search-type=Products	4	30 min 4 ng/mL TGFb + Dabrafenib, AZ628 (bRaf, cRaf inhibitors) vs 30 min 4 ng/mL TGFb
Nd145	Twist	poly ABD29	Mouse; Human	IH(P), ICC	Millipore	ABD29	http://www.merckmillipore.com/NL/en/product/Anti-Twist1-Twist-related-protein-1-Antibody,MM_NF-ABD29?ReferrerURL=https%3A%2F%2Fwww.google.com%2F&bd=1	4	3 Days 4 ng/mL TGFb vs 3 Days Untreated
Nd147	c-myc	D84C12	Human; Mouse	WB, IF, **FC**	CST	7	https://www.cellsignal.com/products/primary-antibodies/c-myc-d84c12-rabbit-mab/5605?site-search-type=Products	6	18 h TPA vs Untreated
Nd148	Snail	ab180714	Human; Mouse	IHC-Fr, WB, ICC/IF, IHC-P	Abcam	AF3639	https://www.abcam.com/snail-slug-antibody-ab180714.html	5	3 Days 4 ng/mL TGFb vs 3 Days Untreated
Nd149	Nanog	D2A3	Mouse	WB, IP, FC, IF, ChiP	BD	2	https://www.cellsignal.com/products/primary-antibodies/nanog-d2a3-xp-rabbit-mab-mouse-specific/8822	3	Not validated by user
Nd150	NFkB (p65)	Polyclonal	Human, Mouse	FC, ChIP, ICC, ChIP/Chip, EMSA, IP, IHC-P, WB,	Abcam	NA	https://www.abcam.com/NF-kB-p65-antibody-ChIP-Grade-ab7970.html	3	Not validated by user

				IHC-Fr, ICC/IF					
Eu151	pP38 (pThr180/pTyr182)	36/p38	Human; Mouse	**FC**, WB	BD	2150665	https://www.bdbiosciences.com/eu/applications/research/b-cell-research/intracellular-antigens/human/pe-mouse-anti-p38-mapk-pt180py182-36p38-pt180py182/p/612565	4	30 min vanadate treatment, 125µM vs 30 min Untreated
Sm152	pAMPK (pThr172)	40H9	Human; Mouse; Rat....	WB, IP, ICH	CST	18	https://www.cellsignal.com/products/primary-antibodies/phospho-ampka-thr172-40h9-rabbit-mab/2535	4	30 min vanadate treatment, 125µM vs 30 min Untreated
Eu153	pAkt (pSer473)	D9E	Human; Mouse; Rat....	WB, IP, IHC, IF, **FC**	CST	20	https://www.cellsignal.com/products/primary-antibodies/phospho-akt-ser473-d9e-xp-rabbit-mab/4060	5	30 min 4 ng/mL TGFb + GDC09411 (PI3K inhibitor) vs 30 min 4 ng/mL TGFb
Sm154	pErk1/2 (pThr202/pTyr204)	20A	Human; Mouse	**FC**, WB	BD	2153932	https://www.bdbiosciences.com/us/applications/research/intracellular-flow/intracellular-antibodies-and-isotype-controls/anti-rat-antibodies/pe-mouse-anti-erk12-pt202py204-20a/p/561991	2	30 min 4 ng/mL TGFb + PD325901 (MEK1/2 inhibitor) vs 30 min 4 ng/mL TGFb
Gd156	CyclinB1	GNS-11	Human; Mouse	WB, IP, **FC**	BD	4241979	https://www.bdbiosciences.com/us/applications/research/apoptosis/purified-antibodies/purified-mouse-anti-cyclin-b1-gns-11/p/554179	8	Cyclin B1 vs IdU
Gd158	pGSK3 (pSer9)	D85E12	Human; Mouse	WB, IP, IF, **FC**	CST	5558BF	https://www.cellsignal.com/products/primary-antibodies/phospho-gsk-3b-ser9-d85e12-xp-rabbit-mab/5558	1	30 min 4 ng/mL TGFb + MK2260 (AKT inhibitor) vs 30 min 4 ng/mL TGFb
Tb159	pSmad1/5 (pSer463/Ser465)	41D10	Human; Mouse	WB, IF, **FC**	CST	9516BF	https://www.cellsignal.com/products/primary-antibodies/phospho-smad1-5-ser463-465-41d10-rabbit-mab/9516	6	Not validated by user
Gd160	CD44	IM7	Human; Mouse	**FC**	BD	550538	https://www.bdbiosciences.com/eu/applications/research/t-cell-immunology/t-follicular-helper-tfh-cells/surface-markers/mouse/purified-rat-anti-mouse-cd44-im7/p/550538	0.085	3 Days 4 ng/mL TGFb vs 3 Days Untreated
Dy162	Vimentin	D21H3	Human; Mouse	WB, ICH, IF, **FC**	CST	5741BF	https://www.cellsignal.com/products/primary-antibodies/vimentin-d21h3-xp-rabbit-mab/5741	1	3 Days 4 ng/mL TGFb vs 3 Days Untreated
Dy164	pSmad2/3 (pSmad2(pSer465/Ser467)/pSmad3(pSer423/Ser425)	D27F4	Human; Mouse	WB	CST	5	https://www.cellsignal.com/products/primary-antibodies/phospho-smad2-ser465-467-smad3-ser423-425-d27f4-rabbit-mab/8828	2	30 min 4 ng/mL TGFb + SB431542 (TGFßR inhibitor) vs 30 min 4 ng/mL TGFb
Ho165	ß-Catenin	D13A1	Human; Mouse	WB, IP, IHC, IF, **FC,** ChIP	CST	8814BF	https://www.cellsignal.com/products/primary-antibodies/non-phospho-active-b-catenin-ser33-37-thr41-d13a1-rabbit-mab/8814	2	3 Days 4 ng/mL TGFb vs 3 Days Untreated

Er167	pMARCK (pSer167/Ser170)	D13E4	Human; Mouse; Rat	WB, IF, **FC**	CST	3	https://www.cellsignal.com/products/primary-antibodies/phospho-marcks-ser167-170-d13e4-xp-rabbit-mab/8722	7	Not validated by user
Er168	CD24	30-F1	Mouse	**FC**	Biolegend	138502	https://www.biolegend.com/en-us/products/purified-anti-mouse-cd24-antibody-6616	3	3 Days 4 ng/mL TGFb vs 3 Days Untreated
Tm169	pPLC gamma-2 (pTyr759)	K86-689.37	Human; Mouse	**FC**	BD	2150657	https://www.bdbiosciences.com/us/applications/research/intracellular-flow/intracellular-antibodies-and-isotype-controls/anti-human-antibodies/pe-mouse-anti-plc-2-py759-k86-68937/p/558490	5	30 min 4 ng/mL TGFb + PP121 (PDGFR inhibitor) vs 30 min 4ng/mL TGFß
Er170	pHistone H3 (pSer28)	HTA28	Human; Mouse	WB, **CyTOF**, ICC, IP, ICFC	Biolegend	641002	https://www.biolegend.com/de-at/products/purified-anti-histone-h3-phosphorylated-ser28-antibody-5169	1.5	Untreated Py2T IdU vs pH3
Yb171	pS6 p(pSer235/Ser236)	N7-548	Human; Mouse	**FC**	BD	2150655	https://www.bdbiosciences.com/eu/applications/research/intracellular-flow/intracellular-antibodies-and-isotype-controls/anti-human-antibodies/pe-mouse-anti-s6-ps235ps236-n7-548/p/560433	2	30 min 4ng/ mL TGFb + 1 µM PD325901 vs 30 min 4ng/ mL TGFb
Yb172	Cleaved Caspase 3	C92-605	Human; Mouse	**FC**, WB, IP	CST	559565	https://www.bdbiosciences.com/us/applications/research/intracellular-flow/intracellular-antibodies-and-isotype-controls/anti-human-antibodies/purified-rabbit-anti--active-caspase-3-c92-605/p/559565	5	5 Days Dinaciclib (1µM) + 4ng/mL TGFß vs 5 Days 4ng/mL TGFß
Yb173	pSTAT3 (pThr727)	49/pSTAT3	Human; Mouse	**FC**, IF, WB	BD	2150654	https://www.bdbiosciences.com/eu/applications/research/t-cell-immunology/th17-cells/intracellular-markers/cell-signalling-and-transcription-factors/human/purified-mouse-anti-stat3-ps727-49p-stat3/p/612542 , https://www.bdbiosciences.com/eu/applications/re	6	Not validated by user
Yb174	E-Cadherin	36/E-Cadh	Human; Mouse	WB, IP, IF, IHC	BD	610182	https://www.bdbiosciences.com/eu/applications/research/stem-cell-research/cancer-research/human/purified-mouse-anti-e-cadherin-36e-cadherin/p/610181	1	11 Days Untreated Py2T vs 11 Days 4ng/mL TGF-ß
Lu175	pRb (pSer807/811)	D20B12	Human; Mouse; Rat....	WB, IP, IF, IHC, **FC**	CST	5	https://www.cellsignal.com/products/primary-antibodies/phospho-rb-ser807-811-d20b12-xp-rabbit-mab/8516	4.5	Untreated Py2T IdU vs CyclinB
Yb176	Survivin	71G4B7	Human; Mouse; Rat	WB, IP, IHC, IF, **FC**	CST	14	https://www.cellsignal.com/products/primary-antibodies/survivin-71g4b7-rabbit-mab/2808?site-search-type=Products	4	BIRC5 overexpression

Table S3. Clusters of inhibitors with similar effects in multiple-batch EMT drug-screen experiment.

Cluster A	Cluster B	Cluster C	Cluster D	Cluster E	Cluster F	Cluster G	Cluster H
TAK-733 (MEK12)	Ibrutinib (Src)	Dacomitinib (EGFR)	LDN193189 (TGF-beta:Smad)	AZ628 (Raf)	AZD5438 (CDK)	AMG 900 (Aurora Kinase)	Cyt387 (JAK12)
Untreated control	Trametinib (MEK12)	GSK2126458 (PI3K)	Afatinib (EGFR)	Amuvatinib (cMet)	Foretinib (c-Met)	NVP-BGT226 (PI3K)	Indirubin (GSK-3b)
Untreated control	Canertinib (EGFR)	Erlotinib (EGFR)	CI-1040 (MEK1:2)	Torin1 (mTOR)	Tozasertib (Aurora Kinase)	PF-00562271 (FAK)	NVP-BHG712 (VEGFR)
Untreated control	PD0325901 (MEK1:2)	Vargatef (VEGFR)	Selumetinib (MEK1:2)	AZD7762 (Chk)	SP600125 (JNK12)	PF-03814735 (Aurora KinaseAB)	R935788 (Syk)
Untreated control	SB525334 (TGFbR1)	AST-1306 (EGFR)	KW 2449 (Flt)	BEZ235 (mTOR)	Bosutinib (Bcr-Abl)	TG101209 (Flt)	Rebastinib (Bcr-Abl)
Untreated control	SB431542 (TGFR)	AZD8931 (EGFR)	WZ3146 (EGFR)	Dasatinib (Src)	Pelitinib (EGFR)	BI 2536 (PLK1)	Deforolimus (mTOR)
Saracatinib (Src)		CP-473420 (EGFR)	IMD 0354 (IKKa)			AT9283 (AuroraK)	Sunitinib (VEGFR)
Untreated control		WHI-P154 (JAK3)	PD153035 (EGFR)			Barasertib (AuroraK)	Vandetanib (VEGFR)
Untreated control		AEE788 (EGFR)	PD318088 (MEK12)			CYC116 (AuroraK)	AT7867 (Akt)
Untreated control		Gefitinib (EGFR)	WZ8040 (EGFR)			MLN8054 (AuroraK-A)	CP 673451 (PDGFRb)
Untreated control						MLN8237 (AuroraK-A)	Ki8751 (VEGFR)
Untreated control						Neratinib (HER2)	R406 (Syk)
AS703026 (MEK12)						Pazopanib (VEGFR1)	Thiazovivin (ROCK)
Untreated control						SNS314 (AuroraK-A)	XL765 (PI3K)
Untreated control						TAE684 (ALK)	YM201636 (PI3K)
Untreated control							CEP33779 (JAK2)
Untreated control							CH5424802 (ALK)
Untreated control							Dovitinib (FLT3)
Untreated control							Semaxanib (VEGFR)
Untreated control							Linifanib (PDGFRb)
Untreated control							LY2228820 (p38MAPKa)
Untreated control							Rapamycin (mTOR)
Untreated control							Temsirolimus (mTOR)
Untreated control							Tie2Kinhibitor (Tie2)

Cluster I
A-769662 (AMPK)
A66 (PI3K)
AMG458 (c-Met)
Arry-380 (HER2)
AS-252424 (PI3K)
AS-604850 (PI3K)
TGFb-only control
BMS 794833 (c-Met)
CAL-101 (PI3K)
CCT128930 (Akt)
TGFb-only control
TGFb-only control
TGFb-only control
TGFb-only control
TGFb-only control
Fostamatinib (Syk)
KX2-391 (Src)
NVP-BSK805 (JAK12)
Palomid 529 (PI3K)
PF-04691502 (mTOR)
PHA-767491 (Cdc7:CDK9)
PIK-293 (PI3K)
PIK-294 (PI3K)
Quercetin (PI3K)
Raf265 (VEGFR)
SB590885 (bRaf)
TG101348 (JAK2)
Tivantinib (c-Met)
Tyrphostin (EGFR)
WP1130 (DUB)
WYE-125132 (mTOR)
ZM 336372 (cRaf)
AG-490 (JAK)
Axitinib (VEGFR)
…

Cluster J
BMS-599626 (EGFR)
Lapatinib (EGFR)
AZD8055 (mTOR)
HMN-214 (PLK1)
ON-01910 (PLK1)
PHA-793887 (CDK)
Dabrafenib (b-Raf)
TAK-285 (EGFR)

Cluster K
BKM120 (PI3K)
OSI-027 (mTOR)
PKI-402 (PI3K)
PP242 (mTOR)
WAY-600 (mTOR)
GDC-0941 (PI3K)
MK-2206 (Akt123)
GSK1059615 (PI3K)
PIK-90 (PI3K)
AZD2014 (mTOR)
BYL719 (PI3K)
INK128 (mTOR)
PI-103 (PI3K, DNA-PK)
WYE354 (mTOR)
ZSTK474 (PI3K)

Cluster L
Hesperadin (AuroraKinaseB)
Dinaciclib (CDK2)

Cluster M
PIK-75 (PI3K)

Cluster N
Torin2 (mTOR)

Table S4. Cell yield of each experimental condition in EMT drug-screen experiment

Dinaciclib (CDK2)	42
PIK-75 (PI3K)	54
SP600125 (JNK12)	82
Tozasertib (Aurora Kinase)	98
BI 2536 (PLK1)	129
Hesperadin (AuroraKinaseB)	145
HMN-214 (PLK1)	237
IMD 0354 (IKKa)	257
ON-01910 (PLK1)	288
Foretinib (c-Met)	294
Torin2 (mTOR)	298
WZ3146 (EGFR)	384
KW 2449 (Flt)	494
AZD8055 (mTOR)	581
Canertinib (EGFR)	617
CYC116 (AuroraK)	630
PIK-90 (PI3K)	631
Afatinib (EGFR)	743
GSK2126458 (PI3K)	755
PD0325901 (MEK1:2)	803
GSK1059615 (PI3K)	808
Torin1 (mTOR)	826
SNS314 (AuroraK-A)	853
GDC-0941 (PI3K)	909
Triciribine (Akt)	914
PHA-793887 (CDK)	946
Amuvatinib (cMet)	977
AT9283 (Bcr-Abl)	1000
MLN8237 (AuroraK-A)	1010
Erlotinib (EGFR)	1025
Everolimus (mTOR)	1028
BEZ235 (mTOR)	1074
AS703026 (MEK12)	1086
MK-2206 (Akt123)	1087
Barasertib (AuroraK)	1133
Axitinib (VEGFR)	1155
AMG 900 (Aurora Kinase)	1199
ENMD-2076 (Flt)	1210
AuroraA (inhibitor)	1263
AZD2014 (mTOR)	1277
AST-1306 (EGFR)	1318
INK128 (mTOR)	1328
AZD8931 (EGFR)	1339
BMS-265246 (CDK1:cyclinB)	1347

Saracatinib (Src)	1409
Deforolimus (mTOR)	1417
PF-03814735 (Aurora KinaseAB)	1438
BS-181 (CDK)	1445
Untreated control	1457
AZD7762 (Chk)	1592
Vandetanib (VEGFR)	1598
SB525334 (TGFbR1)	1650
JNJ-7706621 (CDK1:CyclinB)	1711
XL765 (PI3K)	1742
CCT129202 (AuroraKinaseABC)	1751
PD173074 (FGFR1)	1789
Untreated control	1834
Untreated control	1907
Pelitinib (EGFR)	1908
Vargatef (VEGFR)	1926
CP-473420 (EGFR)	1939
Crizotinib (c-Met)	1947
AT7867 (Akt)	1985
BYL719 (PI3K)	2011
Untreated control	2032
AMG-208 (c-Met)	2047
NVP-BGT226 (PI3K)	2059
LY2784544 (JAK2)	2080
WZ4002 (EGFR)	2117
CI-1040 (MEK1:2)	2141
Untreated control	2152
AT7519 (CDK1:cyclinB)	2156
Untreated control	2158
TGX-221 (PI3K)	2180
Selumetinib (MEK1:2)	2205
Lapatinib (EGFR)	2210
BMS-599626 (EGFR)	2211
Untreated control	2219
Neratinib (HER2)	2250
TWS119 (GSK3b)	2259
KU-60019 (ATM)	2330
KRN 633 (VEGFR)	2335
CP-724714 (EGFR)	2359
PD98059 (MEK12)	2363
OSI-930 (cKit)	2376
U0126 (MEK1:2)	2409
Ki8751 (VEGFR)	2422
IC-87114 (PI3K)	2445
Palbociclib (CDK4:6)	2484

SNS-032 (CDK2:7:9)	2503
BIRB 796 (p38MAPK)	2509
Dasatinib (Src)	2510
TG100-115 (PI3K)	2521
YM201636 (PI3K)	2543
PIK-93 (PI3K)	2559
Motesanib (VEGFR)	2592
Staurosporine (PKC)	2612
BIX 02188 (MEK5)	2622
Untreated control	2625
Danusertib (Aurora Kinase)	2644
PHA-665752 (c-Met)	2695
KU-55933 (ATM)	2725
Untreated control	2731
SU11274 (c-Met)	2761
LY294002 (PI3K)	2771
Sunitinib (VEGFR)	2785
TGFb-only control	2788
Untreated control	2845
Dovitinib (FLT3)	2866
AZD5438 (CDK)	2888
TGFb-only control	2909
AS-605240 (PI3K)	2930
TGFb-only control	2930
R406 (Syk)	2942
SB 203580 (p38 MAPK)	2972
Masitinib (c-Kit)	2977
Vemurafenib (bRAF)	2996
TGFb-only control	3001
TGFb-only control	3006
Brivanib (VEGFR)	3014
CP 673451 (PDGFRb)	3036
OSU-03012 (PDK-1)	3053
Vatalanib (VEGFR)	3055
Sorafenib (VEGFR)	3075
Roscovitine (CDK)	3080
Tandutinib (Flt3)	3080
TGFb-only control	3082
PD153035 (EGFR)	3085
ZSTK474 (PI3K)	3104
Trametinib (MEK12)	3126
Untreated control	3146
Cabozantinib (VEGFR2)	3187
Nilotinib (Bcr-Abl)	3223
Untreated control	3269

TGFb-only control	3282
MLN8054 (AuroraK-A)	3290
AEE788 (EGFR)	3297
WHI-P154 (JAK3)	3330
PLX-4720 (bRAF)	3343
SB 202190 (p38 MAPK)	3361
PI-103 (DNA-PK)	3383
Untreated control	3425
Thiazovivin (ROCK)	3441
AG-490 (JAK)	3465
PP242 (5Days)	3477
PF-04217903 (c-Met)	3505
Brivanib (VEGFR)	3563
JNJ-38877605 (c-Met)	3622
Untreated control	3656
TGFb-only control	3668
WYE354 (mTOR)	3672
Temsirolimus (mTOR)	3698
Pazopanib (VEGFR1)	3763
PD318088 (MEK12)	3812
TAE684 (ALK)	3875
Wortmannin (PI3K)	3902
GSK1904529A (IGF-1R)	3906
TAK-285 (EGFR)	3918
TGFb-only control	3948
SB431542 (TGFR)	4021
TG101209 (Flt)	4046
TGFb-only control	4052
NVP-BVU972 (c-Met)	4096
CH5424802 (ALK)	4126
TGFb-only control	4272
TAK-733 (MEK12)	4280
Untreated control	4287
TGFb-only control	4321
Rapamycin (mTOR)	4379
TPCA-1 (IKK2)	4464
TG100713 (PI3K)	4561
GDC-0068 (Akt123)	4635
SAR131675 (VEGFR)	4940
WP1066 (JAK2)	4973
PKI-402 (PI3K)	5139
BKM120 (PI3K)	5148
Dabrafenib (b)	5279
WZ8040 (EGFR)	5283
Untreated control	5294

LY2228820 (p38MAPKa)	5340
NVP-ADW742 (IGF-1R)	5371
Dovitinib (FLT3)	5387
Gefitinib (EGFR)	5393
TGFb-only control	5456
Semaxanib (VEGFR)	5487
TGFb-only control	5595
TGFb-only control	5596
MK-5108 (Aurora KinaseA)	5609
TGFb-only control	5624
Untreated control	5711
AZD4547 (FGFR)	5713
Tyrphostin (HER2)	5714
Piceatannol (Syk)	5827
TGFb-only control	5832
NVP-TAE226 (FAK)	5860
Untreated control	5874
TGFb-only control	5902
3-Methyladenine (PI3K)	5947
Golvatinib (c-Met)	5966
TGFb-only control	6028
TGFb-only control	6101
MK-2461 (c-Met)	6113
Untreated control	6174
Bosutinib (Src)	6247
TGFb-only control	6262
Tofacitinib (citrate)	6392
INCB28060 (c-Met)	6412
TGFb-only control	6445
TGFb-only control	6454
TGFb-only control	6515
Tofacitinib (JAK3)	6555
Dacomitinib (EGFR)	6561
TGFb-only control	6633
Tideglusib (GSK-3)	6677
Sotrastaurin (PKC)	6685
TGFb-only control	6688
Untreated control	6695
TGFb-only control	6789
TGFb-only control	6793
Baricitinib (JAK1)	6864
Quizartinib (Flt3)	6942
Ibrutinib (Src)	7074
VX-702 (p38 MAPK)	7140
MGCD-265 (c-MET)	7228

TGFb-only control	7335
Cediranib (VEGFR)	7378
Linsitinib (IGF-1R)	7397
TGFb-only control	7443
TGFb-only control	7610
TGFb-only control	7667
E7080 (VEGFR2)	7722
TGFb-only control	8071
OSI-027 (F5.csv)	8209
CEP33779 (JAK2)	8214
PHA680632 (AuroraK)	8257
TGFb-only control	8275
Tivozanib (VEGFR1)	8297
Linifanib (PDGFRb)	8455
BX912 (PDK-1)	8462
GSK690693 (Akt1)	8986
GDC0879 (B-Raf)	9309
WAY-600 (mTOR)	9364
ZM-447439 (AuroraK-A)	9415
TGFb-only control	9670
Enzastaurin (PKC)	9894
PF-00562271 (FAK)	9920
AG1024 (IGF-1R)	10343
PHT427 (Akt)	10497
Imatinib (PDGFR)	10597
BGJ398 (FGFR1)	10677
TSU68 (VEGFR1)	10677
XL147 (PI3K)	11193
Y-27632 (p160ROCK)	11615
Tie2Kinhibitor (Tie2)	11676
SGX523 (HGFR)	12691
AS-604850 (PI3K)	12762
A66 (PI3K)	12871
PIK-293 (PI3K)	14011
AZ628 (Raf)	14111
SB216763 (GSK-3a)	14224
TGFb-only control	14558
PIK-294 (PI3K)	14641
CAL-101 (PI3K)	14686
Palomid 529 (PI3K)	14718
R935788 (Syk)	14770
WYE-125132 (mTOR)	14920
LDN193189 (TGF-beta:Smad)	15493
TGFb-only control	15640
Tyrphostin (EGFR)	15705

PF-04691502 (mTOR)	15787
AMG458 (c-Met)	15862
Untreated control	15898
Fostamatinib (Syk)	15905
Tivantinib (c-Met)	15933
KX2-391 (Src)	16034
Arry-380 (HER2)	16075
BMS 794833 (c-Met)	16319
A-769662 (AMPK)	16487
TGFb-only control	16537
Untreated control	16582
Quercetin (PI3K)	16632
ZM 336372 (cRaf)	16714
TGFb-only control	17125
TG101348 (JAK2)	17243
NVP-BSK805 (JAK12)	17281
Untreated control	17555
Indirubin (GSK-3b)	17809
Untreated control	18074
Untreated control	18117
WP1130 (DUB)	18293
Raf265 (VEGFR)	18399
SB590885 (bRaf)	18431
TGFb-only control	18846
TGFb-only control	18909
NVP-BHG712 (VEGFR)	19703
AS-252424 (PI3K)	21606
CCT128930 (Akt)	22384
Cyt387 (JAK12)	22753
PHA-767491 (Cdc7:CDK9)	24855
Rebastinib (Bcr-Abl)	25245

Table S5. Clusters of inhibitors with similar effects in single-batch EMT drug-screen experiment.

Replicate 1

Cluster A	Cluster B	Cluster C	Cluster D	Cluster E	Cluster F	Cluster G	Cluster H	Cluster I
Untreated control 1	AEE788 (EGFR)	PD153035 (EGFR)	BEZ235 (mTOR)	AZD7762 (Chk)	Bosutinib (Bcr-Abl)	AT9283 (AuroraK)	TGFb-only control 1	PI-103 (PI3K)
Untreated control 2	Gefitinib (EGFR)	PD318088 (MEK12)		Dasatinib (Src)	LY2228820 (p38MAPKa)	Barasertib (AuroraK)	TGFb-only control 2	WYE354 (mTOR)
Untreated control 3		WZ8040 (EGFR)			Neratinib (HER2)	CYC116 (AuroraK)	TGFb-only control 3	ZSTK474 (PI3K)
Untreated control 4					Pelitinib (EGFR)	MLN8237 (AuroraK-A)	TGFb-only control 4	
Untreated control 5					Rapamycin (mTOR)	Pazopanib (VEGFR1)	TGFb-only control 5	
SB431542 (TGFR)					Temsirolimus (mTOR)	SNS314 (AuroraK-A)	AG1024 (IGF-1R)	
					Tie2Kinhibitor (Tie2)	TAE684 (ALK)	BGJ398 (FGFR1)	
							BX912 (PDK-1)	
							Cediranib (VEGFR)	
							E7080 (VEGFR2)	
							Enzastaurin (PKC)	
							GDC0879 (B-Raf)	
							GSK690693 (Akt1)	
							Imatinib (PDGFR)	
							Linifanib (PDGFRb)	
							Linsitinib (IGF-1R)	
							MGCD-265 (c-MET)	
							MLN8054 (AuroraK-A)	
							NVP-ADW742 (IGF-1R)	
							PHA680632 (AuroraK)	
							PHT427 (Akt)	
							Quizartinib (Flt3)	
							SB216763 (GSK-3a)	
							SGX523 (HGFR)	
							TSU68 (VEGFR1)	
							Tivozanib (VEGFR1)	
							XL147 (PI3K)	
							Y-27632 (p160ROCK)	
							ZM-447439 (AuroraK-A)	

Replicate 2

Cluster A	Cluster B	Cluster C	Cluster D	Cluster E	Cluster F	Cluster G	Cluster H	Cluster I
Untreated control 1	AZD7762 (Chk)	Gefitinib (EGFR)	Neratinib (HER2)	AT9283 (AuroraK)	WZ8040 (EGFR)	Cediranib (VEGFR)	AG1024 (IGF-1R)	PI-103 (PI3K)
Untreated control 2	PD318088 (MEK12)	PD153035 (EGFR)	Pelitinib (EGFR)	Barasertib (AuroraK)		CYC116 (AuroraK)	BGJ398 (FGFR1)	WYE354 (mTOR)
Untreated control 3	Tie2Kinhibitor (Tie2)	Bosutinib (Bcr-Abl)	AEE788 (EGFR)	MLN8237 (AuroraK-A)		Linsitinib (IGF-1R)	BX912 (PDK-1)	ZSTK474 (PI3K)
Untreated control 4		Dasatinib (Src)		SNS314 (AuroraK-A)		E7080 (VEGFR2)	TGFb-only control 1	
Untreated control 5				BEZ235 (mTOR)		MGCD-265 (c-MET)	TGFb-only control 2	
SB431542 (TGFR)						MLN8054 (AuroraK-A)	TGFb-only control 3	
						Rapamycin (mTOR)	TGFb-only control 4	
						Temsirolimus (mTOR)	TGFb-only control 5	
						Tivozanib (VEGFR1)	Enzastaurin (PKC)	
						TSU68 (VEGFR1)	GDC0879 (B-Raf)	
						XL147 (PI3K)	GSK690693 (Akt1)	
							Imatinib (PDGFR)	
							Linifanib (PDGFRb)	
							LY2228820 (p38MAPKa)	
							NVP-ADW742 (IGF-1R)	
							Pazopanib (VEGFR1)	
							PHA680632 (AuroraK)	
							PHT427 (Akt)	
							Quizartinib (Flt3)	
							SB216763 (GSK-3a)	
							SGX523 (HGFR)	
							TAE684 (ALK)	
							Y-27632 (p160ROCK)	
							ZM-447439 (AuroraK-A)	

Replicate 3

Cluster A

Untreated control 1
Untreated control 2
Untreated control 3
Untreated control 4
Untreated control 5
SB431542 (TGFR)

Cluster B

Gefitinib (EGFR)
AZD7762 (Chk)
PD153035 (EGFR)
PD318088 (MEK12)
Tie2Kinhibitor (Tie2)

Cluster C

AEE788 (EGFR)
WZ8040 (EGFR)
Pelitinib (EGFR)
Neratinib (EGFR/HER2)
Bosutinib (Bcr-Abl)
Cediranib (VEGFR)
Dasatinib (Src)

Cluster D

AT9283 (AuroraK)
Barasertib (AuroraK)
CYC116 (AuroraK)
MLN8237 (AuroraK-A)
BEZ235 (mTOR)

Cluster E

Enzastaurin (PKC)
Linsitinib (IGF-1R)
LY2228820 (p38MAPKa)
MGCD-265 (c-MET)
MLN8054 (AuroraK-A)
SNS314 (AuroraK-A)
Tivozanib (VEGFR1)
TSU68 (VEGFR1)
XL147 (PI3K)

Cluster F

TGFb-only control 1
TGFb-only control 2
TGFb-only control 3
TGFb-only control 4
TGFb-only control 5
AG1024 (IGF-1R)
BGJ398 (FGFR1)
BX912 (PDK-1)
E7080 (VEGFR2)
GDC0879 (B-Raf)
GSK690693 (Akt1)
Imatinib (PDGFR)
Linifanib (PDGFRb)
NVP-ADW742 (IGF-1R)
Pazopanib (VEGFR1)
PHA680632 (AuroraK)
PHT427 (Akt)
Quizartinib (Flt3)
Rapamycin (mTOR)
SB216763 (GSK-3a)
SGX523 (HGFR)
TAE684 (ALK)
Temsirolimus (mTOR)
Y-27632 (p160ROCK)
ZM-447439 (AuroraK-A)

Cluster G

PI-103 (PI3K)
WYE354 (mTOR)
ZSTK474 (PI3K)

Table S6. Clusters of biospecimens with similar single-cell profiles in melanoma scRNA-seq experiment

Cluster A	Cluster B	Cluster C	Cluster D	Cluster E
Mel53	Mel58	Mel60	Mel67	Mel75
Mel81	Mel65	Mel89	Mel72	
Mel82	Mel71		Mel80	
Mel84	Mel74		Mel94	
Mel88	Mel79			

Table S7. Clusters of biospecimens with similar single-cell profiles in ccRCC mass cytometry experiment

Cluster A	Cluster B	Cluster C	Cluster D	Cluster E	Cluster F	Cluster G	Cluster H
rcc11	rcc12	rcc13	rcc15	rcc16	rcc18	rcc2	rcc55
rcc14	rcc19	rcc26	rcc34	rcc33	rcc20	rcc27	
rcc17	rcc24	rcc32	rcc40	rcc35	rcc21	rcc28	
rcc36	rcc31	rcc37	rcc41	rcc46	rcc22	rcc29	
rcc42	rcc39	rcc4	rcc43		rcc23	rcc30	
rcc45	rcc5	rcc59	rcc44		rcc3	rcc38	
rcc56	rcc51	rcc64	rcc48		rcc50	rcc75	
rcc57	rcc76	rcc65	rcc53		rcc52	rcc81	
rcc58	rcc9		rcc54		rcc72		
rcc6			rcc60		rcc74		
rcc68			rcc62		rcc77		
rcc69			rcc63		rcc8		
rcc71			rcc67				
rcc73			rcc7				
rcc80			rcc70				
			rcc78				
			rcc79				
			rcc82				