Dissecting heterogeneous cell populations across drug and disease conditions with PopAlign

Sisi Chen[1,4,6], Jong H. Park[1,4], Tiffany Tsou[1,4], Paul Rivaud[1,4], Emeric Charles[2], John Haliburton[3], Flavia Pichiorri[5], and Matt Thomson[1,4,6]

[1]Division of Biology and Biological Engineering, California Institute of Technology. Pasadena, California, 91125, USA.

[2]Department of Molecular and Cell Biology, University of California - Berkeley, Berkeley. CA 94720, USA.

[3]Augmenta Bioworks Inc. 3475 Edison Way, Suite K, Menlo Park. CA 94025

[4]Beckman Center for Single-cell Profiling and Engineering. Pasadena, California, 91125, USA.

[5]Department of Hematologic Malignancies Translational Science, City of Hope. Monrovia, California, 91016, USA.

[6]correspondence to: sisichen@caltech.edu, mthomson@caltech.edu

Abstract

Single-cell measurement techniques can now probe gene expression in heterogeneous cell populations from the human body across a range of environmental and physiological conditions. However, new mathematical and computational methods are required to represent and analyze gene expression changes that occur in complex mixtures of single cells as they respond to signals, drugs, or disease states. Here, we introduce a mathematical modeling platform, PopAlign, that automatically identifies subpopulations of cells within a heterogeneous mixture, and tracks gene expression and cell abundance changes across subpopulations by constructing and comparing probabilistic models. Probabilistic models provide a low-error, compressed representation of single cell data that enables efficient large-scale computations. We apply PopAlign to analyze the impact of 40 different immunomodulatory compounds on a heterogeneous population of donor-derived human immune cells as well as patient-specific disease signatures in multiple myeloma. PopAlign scales to comparisons involving tens to hundreds of samples, enabling large-scale studies of natural and engineered cell populations as they respond to drugs, signals or physiological change.

Introduction

All physiological processes in the body are driven by heterogeneous populations of single cells [1, 2, 3]. Single-cell measurement technologies can now profile gene expression in thousands of cells from heterogeneous cell populations across different tissues, physiological conditions, and disease states. However, converting single cell data into models that provide a population-level understanding of processes like an immune response to infection or cancer progression remains a fundamental challenge. All human tissues contain many different subpopulations of cells, and each subpopulation can undergo distinct changes in gene expression and cellular abundance in response to signals, drugs, or environmental conditions. New conceptual and mathematical frameworks are required to model and track the changes that occur within distinct subpopulations of cells within a heterogeneous tissue as they respond to perturbations or succumb to disease.

In this paper, we introduce a computational framework, PopAlign, that identifies, aligns, and tracks subpopulations of single cells within a heterogeneous cell population profiled by single cell mRNA-seq [2, 4, 5, 6]. Mathematically, PopAlign constructs a probabilistic model of each cell population across a series of samples. PopAlign (a) automatically identifies and models subpopulations of cells (b) aligns cellular subpopulations across experimental conditions (signaling, disease) and (c) quantifies changes in cell abundance and gene expression for all aligned subpopulations of cells.

The key conceptual advance underlying PopAlign is representational: we model the distribution of gene expression states within a heterogeneous cell population using a probabilistic mixture model that we infer from single cell data. PopAlign identifies and represents subpopulations of cells as independent Gaussian densities within a reduced gene expression space. PopAlign, then, makes quantitative statistical alignments between subpopulations across samples, and thus enables targeted and quantitative comparisons in gene expression state and cellular abundance. Probabilistic modeling is enabled by a novel low dimensional representation of cell-state in terms of a set of gene expression features learned from data [7, 8, 9]. Unlike PopAlign, geometric methods based on global cell clustering [10, 11] do not provide a natural language for mathematically representing a subpopulation of cells or statistical metrics for quantifying shifts in population structure across experimental samples.

Critically, PopAlign fulfills a fundamental need for comparative analysis methods that can scale to hundreds of experimental samples. Fundamentally, PopAlign runtime scales linearly with the number of samples because computations are performed on probabilistic models rather than on raw single cell data. Probabilistic models provide a reduced representation of single cell data, reducing the memory footprint of a typical 10,000-cell experimental sample by $50 - 100$x. Further, downstream computations including population alignment are performed on the models themselves, often reducing the number of computations by an order of magnitude. By contrast methods based on extraction of geometric features (clusters) from single cell data either by clustering (Louvain) or tSNE rely on pairwise computations between individual cells, which is compute-intensive, and

requires storing of many raw single cell data sets in memory.

We assess the accuracy and generality of PopAlign using twelve datasets from a mouse tissue survey (Tabula Muris) [12] as well as new experiments on human peripheral blood cells, including a screen of immunomodulatory drugs and a comparison of healthy patients to disease (multiple myeloma). We show that PopAlign can identify and track cell-states across a diverse range of tissues, drug perturbation experiments, and human disease states. The probabilistic models have high representational accuracy and identify biologically meaningful cell-states from data. We performed an experimental screen of 40 immunomodulatory compounds applied to primary human immune cells, and used PopAlign to discover the biggest hits at a population-level and also for specific cell types within the mixture. Finally, we used PopAlign to extract general and treatment-specific signatures of disease progression from multiple myeloma patient samples. Moving forward, PopAlign sets the stage for the analysis of large-scale experimental screens of drugs and genetic perturbations on heterogeneous cell populations extracted from primary human tissue samples.

Key Contribution

- Probabilistic modeling of cell populations is key conceptual and practical advance that enables multi-scale analysis of single cell datasets across samples, subpopulations, and individual cells.
- Application of method to data sets from mouse tissues and primary human cells demonstrates accuracy of models and ability to track cell-state specific gene expression changes in response to drugs and disease states.

Results

PopAlign represents heterogeneous cell populations with probabilistic mixture models

We develop a mathematical and computational framework (PopAlign) that (i) identifies and aligns cell-states across paired populations of single cells (a reference population and a test population), and then (ii) quantifies shifts in cell-state abundance and gene expression between aligned populations (Fig. 1). The method has three steps: probabilistic mixture model construction, model alignment, and parameter analysis. PopAlign can be applied to analyze gene expression and population structure changes in heterogeneous populations of cells as they respond to signals, drugs, and disease conditions.

We consider two populations of cells, a test and a reference population, ($\boldsymbol{D}^{\text{test}}$ and $\boldsymbol{D}^{\text{ref}}$), that are

profiled with single cell mRNA-seq (Fig. 1a). Profiling of each population generates a set of gene expression vectors, e.g. $D^{\text{Test}} = \{\boldsymbol{g_i}\}_{i=1}^{k}$ where $\boldsymbol{g} = (g_1, g_2, \cdots, g_n)$, is an n dimensional gene expression vector that quantifies the abundance of each mRNA species in single cell $\boldsymbol{g}$ and k is the number of profiled single cells.

To compare the reference and test cell populations, we first, construct a probabilistic model of the gene expression distribution for each set of cells (Fig. 1b). The high dimensional nature of gene expression ($n \sim 20,000$) space makes the inference and interpretation of probabilistic models challenging. Therefore, we represent each cell, not as a vector of genes, but as a vector of gene expression programs or gene expression features that are extracted from the data, so that each single cell is represented as a vector $\boldsymbol{c} = (c_1, c_2 \cdots c_m)$ of m feature coefficients, c_i, which weight the magnitude of gene expression programs in a given cell(See Methods - Extraction of gene feature vectors using matrix factorization). We extract these gene features using a particular matrix factorization method called orthogonal non-negative matrix factorization (oNMF) that produces a useful set of features because all vectors are positive and composed of largely non-overlapping genes (See SI Fig. 1b and 1g). This allows us to naturally think of a cell's transcriptional state as a linear sum of different positive gene expression programs.

Following dimensionality reduction, for a given cell population, we think of cell states as being sampled from an underlying joint probability distribution over this feature space, $P(\boldsymbol{c})$, that specifies the probability of observing a specific combination of gene expression features/programs, $\boldsymbol{c}$, in the cell population. We estimate a probabilistic model, $P^{\text{test}}(\boldsymbol{c})$ and $P^{\text{ref}}(\boldsymbol{c})$, for the reference and test cell populations that intrinsically factors each population into a set of distinct subpopulations each represented by a Gaussian probability density (density depicted as individual 'clouds' in Fig. 1b):

$$P^{\text{test}}(\boldsymbol{c}) = \sum_{i=1}^{l} w_i \, \phi_i^{\text{test}}(\boldsymbol{c}) \qquad (1)$$
$$\text{where} \;\; \phi_i^{\text{test}}(\boldsymbol{c}) = \mathcal{N}(\boldsymbol{c}; \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$$

where $\mathcal{N}(\boldsymbol{c}; \mu_i, \Sigma_i)$ are multivariate normal distributions with weight w_i; centroids $\boldsymbol{\mu_i}$ and covariance matrices $\boldsymbol{\Sigma_i}$. The distributions $\phi_i^{\text{test}}(\boldsymbol{c}) = \mathcal{N}(\boldsymbol{c}; \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$, mixture components, represent individual subpopulations of cells; l is the number of Gaussian densities in the model. We estimate the parameters of the mixture model $(\{\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}, w_i\})$ from single cell data using the expectation-maximization algorithm [13, 6] with an additional step to merge redundant mixture components to compensate for fitting instabilities (See Methods - Merging of redundant mixture components).

The parameters associated with each Gaussian density ,$(\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}, w_i)$, have a natural correspondence to the biological structure and semantics of a cellular subpopulation. The relative abundance of each subpopulation corresponds to the weight $w_i \in [0, 1]$; the average cell gene expression state of each subpopulation corresponds to the (m dimensional) Gaussian centroid vector $\boldsymbol{\mu_i}$, and the shape or spread of the subpopulation is captured by the covariance matrix $\boldsymbol{\Sigma_i}$. Intuitively, the local Gaussian densities provide a natural 'language' for comparisons between samples. Each Gaussian

is a region of high density in gene feature space, and we compare cell populations by asking how the density of cells shifts across experimental conditions.

Statistical alignment of cellular subpopulations between samples

To compare the test and reference models, we 'align' each mixture component in the test population model, $\phi_i^{\text{test}}(\boldsymbol{c}) \in \{\phi_i^{\text{test}}(\boldsymbol{c})\}$, to a mixture component, $\{\phi_i^{\text{ref}}(\boldsymbol{c})\}$, in the reference population model (Fig. 1c). Alignment is performed by finding the 'closest' reference mixture component in gene feature space. Mathematically, to define closeness, we use Jeffrey's divergence, a statistical metric of similarity on probability distributions. We chose Jeffrey's divergence over other metrics because it is symmetric while also having a convenient parametric form (see Methods)

Specifically, for each $\phi_i^{\text{test}} \in \{\phi_i^{\text{test}}(\boldsymbol{c})\}$, we find an $\phi_j^{\text{ref}} \in \{\phi_j(c)\}^{\text{ref}}$, the closest mixture in the reference set:

$$\underset{\phi_j^{\text{ref}}(\boldsymbol{c}) \in \{\phi_j^{\text{ref}}(\boldsymbol{c})\}}{\arg\min} \; D_{\text{JD}}\left(\phi_i^{\text{test}}(\boldsymbol{c}) \parallel \phi_j^{\text{ref}}(\boldsymbol{c})\right), \tag{2}$$

where the minimization is performed over each $\{\phi_i^{\text{ref}}(\boldsymbol{c})\}$ in the set of reference mixtures, and D_{JD} is the Jeffrey's divergence (14). Intuitively, for each test mixture, we find the reference mixture ϕ_j that is closest in terms of position and shape in feature space. For each alignment, we can calculate an explicit p-value from an empirical null distribution $P(D_{JD})$ that estimates the probability of observing a given value of D_{JD} in an empirical set of all subpopulation pairs within a single cell tissue database (See Methods - Scoring alignments).

Tracking cell-state shifts through mixture model parameters

Following mixture alignment, we analyze quantitative differences in mixture parameters between the reference and test models to track shifts in gene expression state, gene expression covariance, and cellular abundances across the identified subpopulations of cells(Fig. 1d). Mathematically, for each aligned mixture pair, $(\phi_i^{\text{test}}, \phi_j^{\text{ref}})$ with parameters $\{\boldsymbol{\mu_i}^{\text{ref}}, \boldsymbol{\Sigma_i}^{\text{ref}}, w_i^{\text{ref}}\}$ and $\{\boldsymbol{\mu_j}^{\text{test}}, \boldsymbol{\Sigma_j}^{\text{test}}, w_j^{\text{test}}\}$, we calculate:

$$\Delta\mu_i = ||\boldsymbol{\mu_i}^{\text{ref}} - \boldsymbol{\mu_j}^{\text{test}}||_2 \tag{3}$$

$$\Delta\Sigma_i = D_{\text{C}}(\boldsymbol{\Sigma_i}^{\text{ref}}, \boldsymbol{\Sigma_j}^{\text{test}}) \tag{4}$$

$$\Delta w_i = |w_i^{\text{ref}} - w_j^{\text{test}}| \tag{5}$$

where $\Delta\mu_i$ measures shifts in mean gene expression; Δw_i quantifies shifts in cell-state abundance; $\Delta\Sigma_i$ quantifies shifts in the shape of each mixture including rotations and changes in gene expres-

sion variance (see Methods) [14]. We calculate these shifts in parameters for all mixture pairs to assess the impact of drug perturbations or environmental changes on the underlying cell population.

PopAlign identifies and aligns cell-states across disparate mouse tissues

To test the accuracy and generality of PopAlign, we first constructed and aligned probabilistic models across a wide range of mouse tissues from a recent public study (Tabula Muris) [12]. The Tabula Muris study contains single cell data collected from 12 different tissue samples with $\sim 40,000$ cells total.

For all tissues analyzed, the probabilistic mixture models produce an accurate and interpretable decomposition of the underlying cell states (SI Fig. 3). Accuracy of the models can be assessed by comparing the synthetic (model generated) data to raw experimental data held out from model training. PopAlign models generate synthetic data that replicates the geometric structures and statistical variations found in the tissue data in tSNE or PCA plots with quantitative error of $\sim 12\%$ (See methods; Fig. 2; 3a,b; SI Tissues; see methods).

In addition to providing an accurate representation, the mixture models decompose the cell populations into a biologically interpretable set of cellular subpopulations represented by individual $\phi_i(\boldsymbol{c})$, the mixture components (Fig. 3c,d). The PopAlign mixture components, $\{\phi_i(\boldsymbol{c})\}$ commonly contain cells of a single cell 'type' as defined by labels supplied by the Tabula Muris project. In example tissues, PopAlign extracts known tissue resident cell-types including (Fig. 3c,d) basal cells, luminal cells, macrophages, and T-cells (in mammary gland) and skeletal muscle cells, mesenchymal stem cells, endothelial cells, and macrophages (in limb muscle). Broadly, across all tissue models, 70% of the mixture components classified for a single cell-type provided by Tabula Muris (SI Fig. 3).

Through alignment of model components across tissues, PopAlign enables high-level comparisons of tissue composition. By aligning Mammary Gland to Limb Muscle (Fig. 3e), we identified 'common' cell-types between the two tissues including B-cells (p=0.0006), T-cells (p=0.001), endothelial cells (p=0.0013), and macrophages (p=0.004, 0.0076) (SI Fig. 4), and also revealed tissue scale differences in relative abundance. T-cells are highly prevalent ($w = .3$ in the mammary gland but rare in the limb muscle $w = .05$) (Fig. 3g); endothelial cells are highly abundant in the limb muscle ($w = .32$), but rare in the mammary gland ($w = .06$) (Fig. 3g). Between shared cell types, such as macrophages, we reveal common programs such as FC-receptor Signaling and Lysosome, as well as tissue-specific gene expression programs such as TGF-Beta, Phagocytosis, and Leukocyte Chemotaxis(Fig. 3h). PopAlign can, thus, give insight into the underlying composition of a tissue, shedding light onto principles of tissue organization with respect to tissue function.

PopAlign can perform global comparisons of cell state across tens to hundreds of samples

We tested the ability of PopAlign to compare large numbers of samples, using synthetic collections of samples bootstrapped from Tabula Muris data survey. We found that PopAlign runtime scales linearly with sample number and can analyze 100 samples in approximately 100 minutes on a typical workstation with 8 cores and 64GB RAM (Fig. 4a). By first building models, PopAlign front-loads the computation to produce a low-error (Fig. 2) representation of the data that achieves a 50-100x reduction in the memory footprint. Memory efficiency speeds up downstream tasks, such as the calculation of pairwise divergences between subpopulations (Fig. 4b) necessary for aligning them across samples.

Applying PopAlign to compare all 12 tissues of Tabula Muris shows the method is general across many types of experiments, including comparisons of disparate tissues that do not contain overlapping populations. PopAlign achieves generality because it aligns subpopulations by performing a local computation for each test subpopulation (i.e. the minimization of Jeffrey's Divergence relative to reference subpopulations), that can be accepted or rejected by a hypothesis test. Other methods for comparing samples across experiments essentially perform batch correction to align multiple datasets, before pooling data and jointly identifying clusters [10, 11]. These alignment methods discover a global transformation to bring together cells that are known or inferred to be transcriptionally similar. Not only are these alignment methods computationally slow, scaling exponentially with cell/sample number (Fig. 4a), but they also require overlapping populations or force them to overlap, thus limiting the generality of the approach.

In the 12-sample Tabula Muris comparison, PopAlign uncovered meaningful signatures of cell distributions and gene expression patterns that reflect and expand upon known biology. For example, we found that T cells (Fig. 4c) and B cells (Fig. 4d) are most abundant in organs where they are known to mature developmentally (the thymus [15] and spleen[16] respectively), endothelial cells (Fig. 4e) are most prevalent in highly vascularized tissues (Kidney and Limb Muscle), and macrophages (Fig. 4f) are highly prevalent in the Lung, which accumulates debris and bacteria that must be engulfed and destroyed. The analysis also highlights surprising results, such as the observation that T cells are very abundant in the mammary gland (Fig. 4c). We also found distinct patterns of gene program activation (e.g. Lung macrophages are highly phagocytic) in macrophage populations across tissues (Fig. 4g), consistent with previous reports of functional diversity among macrophages [17]. These results demonstrate that PopAlign is an efficient computational framework for extracting meaningful shifts in abundance and gene expression that scales to large numbers of samples, and is not constrained by requirements for overlapping cell populations between samples.

PopAlign identifies universal and cell-type specific impacts of drugs

A key application of PopAlign is to study heterogeneous cell populations from the human body as they respond to environmental change, drug treatments, and disease. The human immune system is an important application domain for PopAlign as an extremely heterogeneous physiological system that is central for disease and cell engineering applications [2, 18, 19, 20, 21]. Being able to screen the effects of different drugs on complex immune cell populations, and understand how they affect cell function, is fundamentally important to our ability to design drug therapies for disease treatment. Thus, we performed an analysis of commercially available immunological compounds on human immune cells and used PopAlign to discover how these compounds alter specific cellular subtypes.PopAlign allows us to explore the data hierarchically, first by using quantitative statistical metrics to rank samples and identify interesting drug hits at the population-level, and then by dissecting the impact of these hits on subpopulation composition and gene expression programs.

We performed our screen using 40 drugs (Fig. 5a) from a commercially available compound library (Selleck Chem) on peripheral blood mononuclear cells (PBMCs) from a healthy 22-year old male donor. PBMCs normally contain a mixture of different immune cell types, but our model revealed that blood samples from this particular donor were dominated by monocytes (18%) and T cells (82%) (Fig. 1a).

We first identified hits at a high-level by ranking drugs based on how similar the drug-exposed populations are to the unperturbed control populations (6 independent replicates). Statistically, we could define hits as drugs which have a negative log likelihood ratio metric (See Methods - Ranking populations) that lies below the control range (gray box). Within this group, high ranking drugs include a group of glucocorticoids (compounds labeled in orange - Fig. 5b), as well as mTOR inhibitors (pink), alprostadil (a prostaglandin) (purple).

Many immune-regulating drugs are known to be broadly suppressive or activating, but their cell-type specific effects are not very well understood. By quantifying and ranking cell type specific shifts, we found that 26 drugs exert significant gene expression shifts (Δ) on monocytes (Fig 5c) while 14 drugs exerted significant effects on T cells (FDR-corrected p-values < 0.05) (Fig 5d). Of these drugs, 8 drugs impacted both cell types (Fig 5c,5d, all drugs highlighted in color). Most drugs either did not affect abundances ($\Delta w \approx 0$) or increased monocyte abundance up to 5% at the expense of T cell abundance (SI Fig 6a,b).

The ability to find the transcriptional impacts of genes that are universal across cell types can reveal important insights into a drugs fundamental mechanisms. In our screen, we discovered that although drug-responsive genes were mostly cell type specific (Fig 5e, Fig), for some drugs, up to 15% of impacted genes were shared between cell types (See Supplementary File 1, which supplies differentially expressed genes for all drugs/cell types). For example, budesonide, up-regulated 11 genes and downregulated 14 genes in both T cells and monocytes (Fig 5f). The overlapping down-regulated genes include many genes associated with actin-based motility - such as actin genes (ACTB, ACTG1), an anti-adhesion peptide (CD52), a myosin interacting protein (CD74)

[22] and an actin-sequestering protein (TSMB10) [23]. This result is consistent with earlier observations that glucocorticoids impede T cell polarization and motility [24] and monocyte migratory behavior[25], and suggest that broad leukocyte motility deficits may be partly responsible for the general immunosuppressive effects of glucocorticoids.

Our analyses also allowed us to discover a highly T-cell specific drug, dexrazoxane, which exerted the largest changes on T cell state (mean $\Delta\mu$ = 2.64, p-val = 2.54e-5, Fig. 5g), but no changes in monocytes (mean $\Delta\mu$ = 0.29, p-val = 1, Fig 5h) . Dexraxozane did not generate any differentially expressed genes in monocytes (Fig 5e). We found that in T cells, dexrazoxane upregulates many cell survival genes including antioxidant enzymes (GPX4, PRDX1) and CORO1A, which is essential for T cell survival [26] (Fig 5i). Dexrazoxane is normally used as a chemoprotectant agent to reduce toxic side effects of chemotherapy on cardiac tissue [27]. Our finding that dexrazoxane specifically impacts T cells by up-regulating genes that reduce oxidative stress has not been previously reported and could potentially be useful in modulating T cell behavior for other diseases.

PopAlign to allows us to rapidly identify cell-type specific effects of drugs. Identification of the most impactful drugs would be difficult using common visualization approaches like t-SNE (SI Fig. 7) or UMAP, which show qualitative changes (see highlighted conditions), but cannot be readily interpreted because the nonlinear embedding means that changes are not quantifiable. Here, using a small screen of 40 drugs from an immunomodulatory compound library, we were able to use PopAlign to discover universal and cell-type specific mechanisms of drugs, including the observation that glucocorticoids broadly down regulate motility genes and dexrazoxane specifically impacts T cells by upregulating pro-survival genes. Understanding the cell type specific impacts of drugs, which have so far been obscured, will be integral for designing precision therapeutics that have targeted effects within a heterogeneous tissue.

PopAlign finds general and treatment-specific signatures of multiple myeloma

Given the success of the PopAlign framework in extracting cell-type specific responses in the immune drug response data, we applied the method to study underlying changes in cell state due to a disease process. As a model system, we applied PopAlign to compare human PBMC samples from healthy donors to patients being treated for multiple myeloma (MM). Multiple myeloma is an incurable malignancy of blood plasma cells in the bone marrow. Both the disease and associated treatments result in broad disruptions in cell function across the immune system [28, 29, 30, 31] further contributing to disease progression and treatment relapse. In MM patients, immune cells with disrupted phenotypes can be detected in the peripheral blood[32, 30, 33]. An ability to monitor disease progression and treatment in the peripheral blood could therefore provide a powerful new strategy for making clinical decisions.

We obtained samples of frozen PBMCs from two healthy and four multiple myeloma patients undergoing various stages of treatment (SI Table 1). We profiled $> 5,000$ cells from each patient,

and constructed and aligned probabilistic models to one reference healthy population (Fig. 6a-f).

PopAlign identified several common global signatures in the MM samples at the level of cell-type abundance and gene expression. Across all samples, we find previously known signatures of multiple myeloma including a deficiency in B cells [30, 34, 35], and an expansion of monocyte/myeloid derived cells [32], and critically, new impairments in T-cell functions.

Plotting Δw across all patients, we find high-level changes in subpopulation abundances, which are known to be prognostic of disease progression [33]. We find that all MM patients experience a contraction in B cell numbers (Fig. 4b), and 2 out of 4 see a dramatic expansion ($\Delta w >> .10$) of monocytes (Fig. 4c). Changes in T cell levels, however, can be highly variable, with outlier patient MM4 experiencing a large increase in effector T cell ($\Delta w = .2$), and a complete elimination of resting T cells ($\Delta w = .2$). For this patient, who was receiving a thalidomide-derived drug therapy, these deviations are consistent with thalidomide's known stimulatory effects on T-cells [36].

Especially in patients with apparently normal abundances (i.e. Δw are small), uncovering subpopulation - specific changes in transcription can point to specific modes of immune dysfunction. We use PopAlign to find that monocyte subpopulations in patients acquire immunosuppressive phenotypes, evidenced by upregulated expression of CD11b and CD33. Both genes are specific markers of myeloid derived suppressor cells[37] which are negative regulators of immune function associated with cancer. By plotting the monocyte-specific mean gene expression values for both CD11b and CD33, we see that all patients except patient MM3 score highly for both MSDC markers. (Fig. 6k). Patients with high MDSC populations typically have a poor prognosis, underscoring the need to monitor MDSC populations in patients.

Importantly, we also find that naive and effector T cells across all multiple myeloma patients have transcriptional defects in pathways essential for T cell function. By plotting $\Delta\mu$, we show that both populations of T cells experience large mean transcriptional shifts, compared to T cells from our second healthy donor, healthy2 (Fig. 6m). By examining the $\mu's$ in terms of gene expression vectors (Fig. 6n), we find that in multiple myeloma, T cells reduce their expression of two key features - Leukocyte Motility, and Cytotoxic Lymphocyte Killing. Surprisingly, the impact on motility is apparent even on the expression of beta-actin (ACTB) (Fig. 6o), a core subunit of the actin cytoskeleton, and which was the top hit in the Leukocyte Motility feature. We find similar declines in the distribution of Perforin 1 (PFN1), a pore-forming cytolytic protein that was found as a top hit in the Cytotoxic Lymphocyte program (Fig. 6p).

Our analysis establishes that we can extract consistent and also patient-specific transcriptional signatures of human disease and treatment response from PBMCs. Interpreting these signatures in the context of disease progression or drug response can provide insight into treatment efficacy and can form the basis of a personalized medicine approach. Our framework enables new applications by providing a highly scalable way of extracting, aligning, and comparing these disease signatures, across many patients at one time.

Discussion

In this paper, we introduce PopAlign, a computational and mathematical framework for tracking changes in gene expression state and cell abundance in a heterogeneous cell populations across experimental conditions. The central advance in the method is a probabilistic modeling framework that represents a cell population as a mixture of Gaussian probability densities within a low dimensional space of gene expression features. Models are aligned and compared across experimental samples, and by analyzing shifts in model parameters, we can pin-point gene expression and cell abundance changes in individual cell populations.

PopAlign constitutes a conceptual advance over existing single cell analytical methods. PopAlign is explicitly designed to track changes within complex cell populations. Since human diseases like cancer and neurodegeneration arise due to interactions between a wide variety of cell-types within a tissue, population level models will be essential for building a single cell picture of human disease and for understanding how disease interventions like drug treatments impact the wide range of cell-types within a tissue.

Mathematically, existing single cell analysis methods rely on heuristic cluster based analysis to extract subpopulations of cells. Fundamentally, such approaches lack well defined statistical metrics for making comparisons across samples. By conceptualizing a single-cell population as a probability distribution in gene expression space, we define a discrete mathematical object whose parameters can be interpreted, and which can be used to explicitly calculate quantitative statistical metrics for subpopulation alignment. Our probabilistic representation allows us to quickly and scalably learn drug responses even on a complex mixture of cells, in 'one shot'. This scalability allowed us to analyze data from large-scale drug screen on resting human immune cells, and identify both universal and cell-type specific mechanisms of drugs

In the future, we hope that PopAlign can be used as a part of a work-bench for single cell analysis and treatment of human disease. By applying PopAlign to data sets from the human immune system, we highlight the potential power of PopAlign for identifying drug/signal targets and for deconstructing single cell disease states. PopAlign identified cell-type specific signatures of disease treatment in multiple myeloma patients exposing a potential defect in T-cell activation and motility in three patient samples. This result points to a potential use of PopAlign for guiding treatment interventions by exposing the spectrum of transcriptional states within a diseased tissue and revealing the impact of drug treatments on diseased cell-states as well as the cellular microenvironment and immune cell-types. Such insights could lead to single cell targeting of drug combinations to treat human disease as an essentially population level phenomena.

Methods

Mathematical framework

We consider two populations of cells, a reference population, ($\boldsymbol{D}^{\text{test}}$ and $\boldsymbol{D}^{\text{ref}}$), and a test population. Following profiling by single cell mRNA-seq, each population of cells is a set of gene expression vectors, $\boldsymbol{D} = \{\boldsymbol{g_i}\}_{i=1}^{k}$ where k is the number of cells in the population, and $\boldsymbol{g} = (g_1, g_2, \cdots, g_n)$, is an n dimensional vector that quantifies the abundance of each mRNA species. While raw mRNA-seq measurements generate integer valued gene count data, due to measurement noise and data normalization, we consider $\boldsymbol{g}$ to be embedded in an n dimensional Euclidean vector space, gene expression space, $\boldsymbol{g} \in \mathbb{R}^n$. The high dimensional nature of gene expression space poses the key challenge for construction and interpretation of statistical models.

We think of the gene expression vectors, $\{\boldsymbol{g_k}\}$ as being distributed according to an underlying probability density function, $P(\boldsymbol{g})$, that quantifies the probability of observing a particular joint gene expression state, $\boldsymbol{g} = (g_1, g_2, \cdots, g_n)$, in a given cell population. Our broad goal is to estimate a statistical model of $P(\boldsymbol{g})$,based upon single cell measurements. The model provides a parametric representation of the gene expression density in each condition. Then, we seek to use this representation to track changes in the structure of the cell population across conditions.

In general, the mathematical challenge we face is model estimation in the high dimensional nature of gene expression space. For human cells $n > 20,000$, and single cell profiling experiments can routinely probe 10,000 cells per sample. The number of parameters in our probabilistic models scales quadratically with n, and mixture model learning has data requirements that are exponential in n [8]. Therefore, we first reduce the dimensionality of the problem by building models in a common low dimensional space defined by gene expression programs discovered from pooled data across all samples.

Data normalization

Single cell gene expression data must be normalized to 1) to account for the variation in the number of transcripts captured per cell and 2) to balance the wide disparity in the scale of values across different genes due to measurement noise and gene drop-out.

The total number of transcripts captured for each single cell can vary from 1000 to 100,000 unique transcripts per cell. Technical variability in reagents and library prep steps can have a large impact on the number of transcripts retrieved per cell. To scale out these differences, we divide each gene expression value g_i by the total number of transcripts and then multiply by a scaling factor β.

Additionally, across genes, mean transcript values can span 5 orders of magnitude. Transforming the data using the logarithm brings values across all genes close in scale, while also reducing the skew in the data distributions. The equation for transforming a raw gene expression value g_i (for a

single gene) into a normalized gene expression value, g_i' is:

$$g_i' = \log(\beta \frac{g_i}{\sum_i^n g_i} + 1) \tag{6}$$

where n is the total number of genes, and β is a scaling factor, and we add a 1 pseudo-count to each gene expression value. We found that by setting $\beta = 1000$ to be roughly the median number of total transcript counts in a cell (1000 transcripts), we achieve a smooth transition in the distribution of transformed g_i' when raw g_i values step from 0 to 1. Gene expression values are thus denoted in units of $log(\hat{g} + 1)$ where $\hat{g}$ is cell-normalized and rescaled.

Extraction of gene feature vectors with matrix factorization

We circumvent the curse of dimensionality ([7]) by building models in a common low-dimensional space defined by gene expression features or programs. Mathematically, we represent the transcriptional state of each single cell, $\boldsymbol{g}$, as a linear combination of gene expression feature vectors, $\{\boldsymbol{f_i}\}$:

$$\boldsymbol{g} = \sum_{i=1}^{m} c_i \, \boldsymbol{f_i} \tag{7}$$

where $\mathbf{f_i} \in \mathbb{R}^{\mathbf{n}}$ specifies a gene expression feature, and c_i is a coefficient that encodes the weighting of vector f_i in $\boldsymbol{g}$, the gene expression state of a single cell. The key result in [7] is that a cell's gene expression state, $\boldsymbol{g}$ can be represented as a linear combination of m gene expression module vectors, $\boldsymbol{f_i}$ where $m << n$. This insight allows us to construct a low dimensional representation of a cell population and, then, to estimate statistical models within the low dimensional space. [7, 38, 39].

The gene features, $\{\boldsymbol{f_i}\}$ can be extracted using a wide range of matrix factorization and machine learning technique including Singular Value Decomposition, and its matrix factorization relatives like sparse PCA as well as methods like layered neural networks [7]. We use a technique called orthogonal non-negative matrix factorization [40] (oNMF) to define a space of orthogonal gene expression features vectors. Like other linear dimensionality reduction techniques, such as PCA, oNMF factors the original data matrix, $\boldsymbol{D}^{\text{train}}$ (SI Fig. 1a) into two matrices $\boldsymbol{D} \approx \boldsymbol{F}\ \boldsymbol{C}$ (SI Fig. 1b,c). Factorization occurs through minimization of an objective function with positivity and orthogonality constraints:

$$\begin{aligned} &\arg\min_{\boldsymbol{F},\boldsymbol{C}} ||\boldsymbol{D}^{\text{train}} - \boldsymbol{F}\boldsymbol{C}||_2 \\ &\text{subject to } \ \boldsymbol{F}^T\boldsymbol{F} = I, C_{ij} \geq 0, F_{ij} \geq 0, \end{aligned} \tag{8}$$

The optimization minimizes the (Frobenius) norm of the difference between the training data, $\boldsymbol{D}^{\text{train}}$, and its factored representation $\boldsymbol{F}\boldsymbol{C}$. The columns of $\boldsymbol{F}$ contain gene features, $\boldsymbol{f_i}$. The matrix $\boldsymbol{C}$ is m by k, where k is the number of single cells in $\boldsymbol{D}^{\text{train}}$. Each column of $\boldsymbol{C}$ encodes the weighting of the m gene features across a given single cell. The entries of $\boldsymbol{F}$ and $\boldsymbol{C}$; are constrained

to be positive, and the columns of $\boldsymbol{F}$ (the gene features) are constrained to be orthogonal. $\boldsymbol{F}$ is an n by m matrix (genes by features). Each column contains n weights where each weight corresponds to the weight of a given gene in that feature, $\boldsymbol{f_i}$.

Standard non-negative matrix factorization has been shown to provide a useful set of features for gene expression analysis because feature vectors have positive entries, and so we can naturally think about the gene expression state of a cell, $\boldsymbol{g}$, as being assembled as a linear sum of positive gene expression programs.

In PopAlign, we incorporate orthogonality in $\boldsymbol{F}$ as a secondary constraint to aid interpretation. Empirically, we found that orthogonality aids in interpretation of the features as well as in model construction because the orthogonal gene expression features are interpretable as non-overlapping sets of genes (SI Fig. 1b) that can individually be analyzed by gene set enrichment analysis (SI Fig. 1f)(see Methods - Gene Set Enrichment Analysis). Second, orthogonality tended to force individual cell states to be represented by more than one feature which aided stability during model parameter estimation.

To perform oNMF, we select m, the number of features to be extracted through an optimization that balances accuracy and dimensionality explicitly. In oNMF (as opposed to PCA and SVD), m is a parameter given to the optimization. Choosing m involves balancing the tension between the 'expressiveness' in the feature set and its dimensionality. Higher m reduces the error in the representation while also breaking up blocks of genes into smaller modules that represent independent gene expression pathways with finer granularity. However, as m increases, the typical computational and sampling challenges associated with high dimensionality emerge.

Practically, we balance this tension in PopAlign by constructing a loss function with a penalty that increases with m:

$$\arg\min_{m} \; f(m) = \; ||\boldsymbol{D}^{\text{train}} - \boldsymbol{F_m C_m}||_2 + m^{\alpha}. \tag{9}$$

For each value of m, we perform oNMF on $\boldsymbol{D}^{\text{train}}$ yielding $\boldsymbol{F_m}$ and $\boldsymbol{C_m}$, and thus an error $||\boldsymbol{D}^{\text{train}} - \boldsymbol{F_m C_m}||^2$. This error is, then, incremented by the term m^{α} which penalizes higher values of m and hence the dimensionality of the feature set. We set $\alpha = .7$ based upon numerical experimentation on model data sets (SI Fig. 2). For any choice of m, we can estimate the accuracy of the representation by plotting reconstructed data FC (SI Fig. 1d) against normalized data D (SI Fig. 1a). The SI shows such plots and the PopAlign software package outputs these plots by default.

Practically, given data sampled a set of cell populations,(eg $D^{\text{test}_1}, D^{\text{test}_2}, D^{\text{ref}}$) we pool data from all cell populations into a training data set, $\boldsymbol{D}^{\text{train}}$, and perform oNMF. If we are analyzing a large number of data sets or sets with many single cells, we generate $\boldsymbol{D}^{\text{train}}$ by sampling 500-1,000 cells uniformly at random from the reference and test cell populations and selecting a m via ((9)).

Because the feature vectors are not always purely orthogonal, we recast the complete dataset into the feature space using a non-negative least squares. Specifically, for each gene expression profile, $\boldsymbol{g}$, we find, $\boldsymbol{c}$ via:

$$\begin{aligned} &\arg\min_{\boldsymbol{c}} ||\boldsymbol{g} - \boldsymbol{F}\ \boldsymbol{c}||_2 \\ &\boldsymbol{c_i} \geq 0 \end{aligned} \tag{10}$$

where $\boldsymbol{F}$ is a fixed feature set learned from oNMF on $\boldsymbol{D}^{\text{train}}$. The gene expression vector g is thus compressed into a k-dimensional vector c_i that provides a high-level programmatic representation of cell state in terms of gene expression 'features'. Finally, we interpret the biological meaning of the feature vectors in terms of annotated gene expression programs using gene set enrichment analysis (see Methods - Gene Set Enrichment Analysis). Using matrix factorization, we map a cell population, $D = \{\boldsymbol{g}\}$, from an $n \sim 20,000$ dimensional gene expression space into a gene feature space that is often of order $10 - 20$ dimensions [7].

$$D = \{\boldsymbol{g_i}\} \rightarrow \{\boldsymbol{c_i}\},$$

where $\{\boldsymbol{c_i}\}$ are $m \times 1$ dimensional vectors that now represent the cell population in the reduced gene feature space.

Gene set enrichment analysis

To interpret the gene features in terms of annotated gene sets, we perform geneset enrichment analysis using the cumulative hypergeometric distribution. We define each feature vector by the collection of genes that have weightings greater than 4 times the standard deviation ($> 4\sigma$). Using this collection of genes, we then calculate the null probability of drawing k genes ($P(X > k)$) from a specific annotated gene set using the hypergeometric cumulative probability distribution:

$$P(X > k) = \sum_{i=1}^{k} \frac{\binom{Y}{i}\binom{N-Y}{Z-i}}{\binom{N}{Z}}$$

, where N is the total number of genes, Z is the number of genes in the feature that are $> 4\sigma$, Y is the number of genes in each annotated gene set, and k is the number of genes that overlap with annotated gene set.

Gene sets are sorted by their associated null probability; the 10 gene sets with the lowest null probabilities are reported for each feature. The gene sets in our dictionary are pulled from GO, KEGG, and REACTOME, and are supplied with our code. SI Fig. 1f shows an example of gene set enrichment results for two features.

oNMF error analysis

The error associated with each feature set F_m was assessed by comparing data entries between a cross validation dataset $\boldsymbol{D_x}$ and its reconstructed matrix $\boldsymbol{F_m C_x}$. We binned the data in $\boldsymbol{D_x}$ into bins of equal width ~ 0.1 (in units of log(TPT+1)). We retrieved data values from each bin, $(\boldsymbol{D_x})_i$, and then plotted their means against the means of corresponding data values in the reconstructed matrix, $(\boldsymbol{F_m C_x})_i$ (SI Fig. 1d). The standard error in each bin is calculated as the mean squared deviation of the reconstructed data from the original data:

$$\sigma_i = \sqrt{\frac{\sum_i^n ((\boldsymbol{F_m}\,\boldsymbol{C_x})_i - (\boldsymbol{D_x})_i)^2}{n}},$$
$$\sigma = \sum_i \sigma_i.$$

To quantify the amount of dispersion relative to the mean, we also calculate the coefficient of variation for each bin:

$$CV_i = \frac{\sigma_i}{(\boldsymbol{D_x})_i},$$

We find empirically that the average CV is $\sim 30 - 35\%$ across all bins for most feature sets.

Representing a cell population as a Gaussian Mixture Model in gene feature space

Following the feature based representation, we construct a statistical model of each given cell population within the reduced gene feature space. Mathematically, we have exchanged a probability distribution in gene expression space for a probability distribution in gene feature space:

$$P(\boldsymbol{g}) \rightarrow P(\boldsymbol{c}), \tag{11}$$

where $\boldsymbol{g}$ is $n \times 1$ and $\boldsymbol{c}$ is $m \times 1$, and $m << n$. We can now estimate a statistical model of $P(\boldsymbol{c})$.

To account for the heterogeneity of cell-states within a tissue, we model cell-populations using Gaussian mixture models. Gaussian densities provide a natural representation of a transcriptional state in gene expression space which is consistent with measured gene expression distributions as well as empirical models of transcription [41, 42, 43]. Theoretical models of stochastic transcription commonly yield univarite gene expression distributions where mRNA counts are Poisson or Gamma distributed. Normal distributions provide a reasonable approximation to these distributions with a computationally tractable inference procedure.

We represent the cell population as a mixture of Gaussian densities, so that for a given cell population D, we construct:

$$P(\boldsymbol{c}) = \sum_{i=1}^{l} w_i \, \phi_i(\boldsymbol{c}) \tag{12}$$
$$\phi_i(\boldsymbol{c}) = \mathcal{N}(\boldsymbol{c}; \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$$

where $P(\boldsymbol{c})$ is a mixture of Gaussian densities, $\mathcal{N}(\boldsymbol{c}; \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$, with centroid, μ_i; covariance matrix, ,Σ_i; and scalar weighting w_i. $\boldsymbol{\mu_i}$ is a vector in the m dimensional feature space, and $\boldsymbol{\Sigma_i}$ is a symmetric $m \times m$ matrix. l is the number of Gaussian mixtures or components in the statistical model. The Gaussian mixture model (GMM) represents a cell populations as a mixture of individual Gaussian densities. Biologically, we think of each density as parameterizing a subpopulation of cells.

We can estimate the parameters $\boldsymbol{\mu_i}$, $\boldsymbol{\Sigma_i}$, and w_i based upon training data, using maximum likelihood estimation with likelihood function:

$$\mathcal{L}(\{\mu_i, \Sigma_i, \pi_i\}|\boldsymbol{D'}) = \sum_{j=1}^{k} \log P(\boldsymbol{c_j}; \{\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}, w_i\}), \tag{13}$$

where $\boldsymbol{c_j}$ are single cell profiles drawn from a cell population $\boldsymbol{D'}$ and cast into feature space; k is the number of single cells in the cell population $\boldsymbol{D'}$. For a given experimental data set, $\mathcal{L}$ defines a function over the space of model parameters. To select model parameters given data, we can attempt to maximize the value of $\mathcal{L}$. In general for Gaussian Mixture models, likelihood maximum is complicated by the geometry of $\mathcal{L}$ which is not concave and can have multiple local and global maxima [44]. $\mathcal{L}$ can be maximized approximately using expectation maximization.

Expectation-maximization is a heuristic algorithm that finds (local) maximum likelihood parameters. Although it is known to have fundamental problems - including weak performance guarantees and a propensity to overfit data, new methods [8] place constraints that are invalid for our application (such as shared covariance matrices). We find empirically that the EM algorithm performs well, learning low-error representations of $\boldsymbol{c}$ (SI Fig. S2), and that we can overcome fitting instabilities by algorithmically merging components. Practically, we perform expectation maximization using sci-kit learn. We regularize the variance of individual mixtures to constrain variance to be non-zero to avoid fitting instabilities. We determine mixtures number through the Bayesian Information Criterion (BIC) which optimizes a trade-off between model complexity and accuracy on training data.

Merging of redundant mixture components

One drawback of the EM algorithm is its propensity to fit 'redundant' mixture components to the same local density with significant overlap. For our application, this redundancy complicates interpretation and comparisons across samples. We overcome this problem by taking advantage of

mixture model properties to algorithmically merge redundant mixtures using the Jeffrey's divergence.

For multivariate Gaussian distributions, the Jeffrey's divergence has a closed analytic form.

$$D_{\mathrm{JD}}(\phi_i \| \phi_j) = \frac{1}{2}(D_{\mathrm{KL}}(\phi_i \| \phi_j) + D_{\mathrm{KL}}(\phi_j \| \phi_i)) \tag{14}$$

where ϕ_0 and ϕ_1 are two independent components from the same mixture model (12), and μ and Σ are their associated parameters.

D_{KL} is the Kullback Leibler divergence and has a convenient parametric form for Gaussian distributions:

$$D_{\mathrm{KL}}(\mathcal{N}_i(\boldsymbol{c}; \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) \| \mathcal{N}_j(\boldsymbol{c}; \boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})) = \frac{1}{2}\left(\mathrm{tr}\left(\Sigma_i^{-1}\Sigma_j\right) + (\mu_i - \mu_j)^{\mathrm{T}}\Sigma_i^{-1}(\mu_i - \mu_j) - k + \ln\left(\frac{\det \Sigma_i}{\det \Sigma_j}\right)\right)$$

For each mixture model, we iteratively attempt to merge component pairs with the lowest Jeffrey's divergence and accept mergers that increase the BIC of the model given the data. With each merge step, model parameters for candidate pair are recalculated, and the updated model is accepted or rejected based on the new BIC. Mergers are performed until the first rejection. This procedure removes redundant mixture components from the model.

Sampling data from Gaussian mixture models

Given parameters, $\{(\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}, w_i)\}_{i=1}^{l}$ for a given mixture model. We can 'generate' synthetic data from the model through a simple sampling procedure. A given model has, l mixtures, and we first select a mixture from the set of l mixtures with probabilities weighted by w_i. Following selection of a mixture, j, we use standard methods to draw a 'point' in the m dimensional feature space from $\mathcal{N}_|(\boldsymbol{c}, \boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})$.

Analysis of model error

Model error was assessed by comparing the empirical distributions of model-generated data with experimental data. We performed error analysis within 2D projections of the data, because fully binning a k-dimensional space can be highly memory intensive. Briefly, each 2D projection was binned into 25 bins (5x5). For each bin, the deviation between the model generated data and empirical data was calculated in terms of percent error in each bin. Total error within each projection is calculated as a weighted average of this percent error over all bins in a projection:

$$\text{error} = \sum_i \frac{N_i}{N_{\text{T}}} \frac{|N_i - \widehat{N_i}|}{N_i},$$

where N_i is the number of experimental data points in bin i, and $\widehat{N_i}$ is the number of model generated data points in bin i. N_{T} is the total number of experimental data points. This metric weights the per-bin fractional error by the probability density of each binned region in the projection.

Alignment of models across reference and test populations

To compare the test and reference models, we 'align' each mixture component in the test population model, $\phi_i^{\text{test}}(\boldsymbol{c}) \in \{\phi_i^{\text{test}}(\boldsymbol{c})\}$, to a mixture component, $\{\phi_i^{\text{ref}}(\boldsymbol{c})\}$, in the reference population model (Fig. 1c). Alignment is performed by finding the 'closest' reference mixture in gene feature space. Mathematically, to define closeness, we use, Jeffrey's divergence, a statistical metric of similarity on probability distributions. Specifically, for each $\phi_i^{\text{test}} \in \{\phi_i^{\text{test}}(\boldsymbol{c})\}$, we find an $\phi_j^{\text{ref}} \in \{\phi_j(c)\}^{\text{ref}}$, the closest mixture in the reference set:

$$\underset{\phi_j^{\text{ref}}(\boldsymbol{c}) \in \{\phi_j^{\text{ref}}(\boldsymbol{c})\}}{\arg\min} \; D_{\text{JD}} \left(\phi_i^{\text{test}}(\boldsymbol{c}) \parallel \phi_j^{\text{ref}}(\boldsymbol{c})\right), \tag{15}$$

where the minimization is performed over each $\{\phi_i^{\text{ref}}(\boldsymbol{c})\}$ in the set of reference mixtures, and D_{JD} is the Jeffrey's divergence (14). Intuitively, for each test mixture, we find the reference mixture ϕ_j that is closest in terms of position and shape in feature space.

Aligning subpopulations using Jeffrey's divergence incorporates information about the shape and position of the probability distributions (i.e. covariance), while disregarding the relative abundance of each subpopulation.

Scoring alignments For each alignment, we can calculate an explicit p-value from an empirical null distribution $P(D_{JD})$ that estimates the probability of observing a given value of D_{JD} in an empirical data set of all subpopulation pairs within a single cell tissue database. To assign a p-value to alignments, we calculate the probability of observing two cell-states with a given Jeffrey's divergence by chance using an empirical null distribution generated from the tissue data set. Specifically, we constructed a mixture model for all tissue pairs in Tabula Muris. Then, we calculate all pair-wise Jeffrey's divergence scores for all the underlying mixtures. This calculation gives us a global distribution over D_{JD} for cell-states in the mouse. This distribution provides a null distribution for typical statistical closeness between cell-states in feature space.

Differential gene expression between aligned subpopulations using L1-norm error metric We can discover up- and down-regulated genes between any two subpopulations by finding genes which have significant shifts in their distributions. We quantify the extent of this shift using a signed L1 distance metric over a discretized domain. For each gene, we calculate two empirical

distributions $P_1(a)$, $P_2(a)$, one for each subpopulation. These distributions are discretized over identical histogram binnings, a_i, over the gene's entire range of expression, $a_i \in \mathcal{A}$. Then we define the signed L1 distance as:

$$d = \text{sgn}(\bar{g_2} - \bar{g_1}) \sum_{a \in \mathcal{A}} ||P_1(a) - P_2(a)||_1$$

where $\bar{g_1}$ and $\bar{g_2}$ denote the respective means of subpopulations 1 and 2. The sign allows us to distinguish whether a gene is down-regulated in population 2 (negative) or upregulated in population 2 (positive). The signed L1 distance metric ranges from -2 to +2, both of which correspond to completely non-overlapping distributions (which we have not yet seen in gene expression data). In experiments with control samples, we use comparisons between control samples to calculate an empirical cutoff for interpreting an L1-distance as significant. We determine the cutoff as the point at which the fraction of control genes that exceed this value is less than 0.001 (or some user-selected p-value). For the drug screen, this L1 threshold is calculated to be around 0.5. Qualitatively, we see that genes with L1-distances of ¿ 0.5 have distributions which are visually in obvious agreement with the labeled directionality, which holds true across a range of different experimental samples.

Model interpretation through parameter analysis

Following mixture alignment, we analyze quantitative differences in mixture parameters between the reference and test sample to track shifts in gene expression state, gene expression covariance, and cellular abundances across the identified cell-states in the cell population. Specifically, for each aligned mixture pair, $(\phi_i^{\text{test}}, \phi_j^{\text{ref}})$ with parameters $\{\mu_i^{\text{ref}}, \Sigma_i^{\text{ref}}, w_i^{\text{ref}}\}$ and $\{\mu_j^{\text{test}}, \Sigma_j^{\text{test}}, w_j^{\text{test}}\}$ we calculate:

$$\begin{aligned}
\Delta\mu_i &= ||\boldsymbol{\mu_i^{\text{ref}}} - \boldsymbol{\mu_j^{\text{test}}}||_2 \\
\Delta\Sigma_i &= D_{\text{C}}(\boldsymbol{\Sigma_i^{\text{ref}}}, \boldsymbol{\Sigma_j^{\text{test}}}) \\
\Delta w_i &= |w_i^{\text{ref}} - w_j^{\text{test}}|
\end{aligned}$$

where $\Delta\mu_i$ measures shifts in mean gene expression; Δw_i quantifies shifts in cell-state abundance; $\Delta\Sigma_i$ quantifies shifts in the shape of each mixture including rotations and changes in gene expression variance. For $\Delta\Sigma_i$, we use the following distance metric on covariance matrices [14]:

$$D_{\text{C}}(\Sigma_i, \Sigma_j) = \sqrt{\sum_{z=1}^{m} \ln^2 \lambda_z(\Sigma_i, \Sigma_j)}$$

where λ_z is a generalized eigenvalue of $\boldsymbol{\Sigma_i}$ and $\boldsymbol{\Sigma_j}$ or a solution to $\boldsymbol{\Sigma_i}\, v = \lambda_z \boldsymbol{\Sigma_j} v$, and m is again the number of gene features.

We calculate these shift parameters for all mixture pairs, and then analyze the shifts to assess the impact of signaling conditions or environmental changes on the underlying cell population.

Ranking populations against a control using the log-likelihood ratio metric The log-likelihood ratio serves as a quantitative measure of how similar perturbed populations are to a control population. To calculate the log-likelihood ratio, we use both the control model (non-perturbed) as well as the perturbed model to compute probability scores for cells sampled from the perturbed sample, and then take the ratio of those probability scores. The LLR is the mean logged ratio between the two probability values, and will be near 0 if the two probabilities are very close (i.e. ratio is near 1) but will be negative if the perturbed sample differs substantially from the control.

$$\mathcal{LLR} = \frac{1}{n}\sum_{i=1}^{n} log\frac{\mathcal{L}(d_i|\theta_{ctrl})}{\mathcal{L}(d_i|\theta_{drug})}$$

In our drug screen, we observed that the LLR is sensitive to transcriptional changes but this sensitivity is decreased if the affected subpopulation is rare. All of the drugs which have significant LLR (p-value adjusted ¡ 0.05) results in a gene expression shift (delta mu) in either T cells or monocytes. However, drugs that induce cell type-specific changes in very rare cells often do not generate LLRs that are significantly different than the control. For example, Dexrazoxane has an LLR value that barely makes the p-value cut off (Fig 5b), but induces significant gene expression changes in T cells (Fig 5d). There are only 5% T cells in the Dexrazoxane sample (about 40 T cells total), which likely diminishes the T cell impact on the global LLR score.

Cell type classification of mixtures using marker genes For the Tabula muris data set, cells and mixtures were classified using cell-type annotations provided by the study. Mixtures were classified according to the cell-type with the maximum abundance within a given mixture (Fig. 2).

For drug screen experiments, we classified the independent mixtures as T cells (CD3D+), Monocytes (LYZ+), or B cells (CD20+).

For immune cell experiments with healthy and multiple myeloma patients, we classified the independent mixture components as effector T cells (CD3+ / CD57+), naive T-cells(CD3+ / CD28+), erythrocytes (HBB+), canonical monocytes (CD14+ / CD16low), and nonclassical monocytes (CD14low / CD16++) [45] (Fig. 4e). We aligned populations in our test samples - GM-CSF (Fig. 4b) and IFNG (Fig. 3c) - to the reference populations (control) by finding pairs of components that minimize the Jeffrey's divergence (Fig. 4d).

Software implementation

PopAlign has been implemented as a Python3 software package. The package requires common scientific computing libraries (numpy, matplotlib, pandas, seaborn, tables, MulticoreTSNE, ad-

justText) that can be easily installed with pip and our requirements file. A guide on how to get set up and install dependencies is provided on the packages Github page. The software runs on local machines, as well as on Amazon Web Services headless EC2 instances, providing a powerful setting for large-scale analyses. The architecture of this package is built around classes that store experimental samples as objects and provide a set of specific methods to perform tasks such as normalization, dimensionality reduction, model construction, model alignment, and parameter comparison.

Experimental methods

Single-cell RNA-sequencing. Cryopreserved PBMCs (Hemacare) from healthy and multiple myeloma patients were thawed in a 37C waterbath for 2 minutes after which the cells were transferred to a 15mL conical tube. Prewarmed RPMI 1640 was then added to the 15mL conical to a final volume of 10mL and centrifuged for 2 minutes at 300RCF to pellet the cells. Supernatant was removed and cells were resuspended to 1 million cells/mL in RPMI1640 supplemented with 10% FBS and 17,400 cells were loaded into each TENX lane.

Sample multiplexing using Multi-Seq Cryopreserved PBMCs sourced from Hemacare ($\sim$ 50 million cells) were thawed in a 37C waterbath for 2 minutes after which the cells were transferred to a 15mL conical tube. Prewarmed RPMI1640 was then added to the 15mL conical to a final volume of 10mL and centrifuged for 2 minutes at 300RCF to pellet the cells. Supernatant was removed and cells were resuspended in 10mL of RPMI1640 supplemented with 10% FBS and 1% pen/strep. The cell suspension was then plated onto a 100mm low attachment plate and rested in a CO2 incubator at 37C for 16 hours.

After resting, 200,000 cells were loaded into each well of a 96-well plate and exposed to 1 uM of drug in RPMI1640 plus serum. Drugs used were drawn from the Immunology and Inflammation-related library sold by SelleckChem. After 24 hours of exposure, cells were enzymatically dissociated into a single-cell suspension using TrypLE and multiplexed using Multi-seq lipid-modified oligos. [46].

References

[1] K. J. Pienta, N. McGregor, R. Axelrod, and D. E. Axelrod, "Ecological Therapy for Cancer: Defining Tumors Using an Ecosystem Paradigm Suggests New Opportunities for Novel Cancer Treatments," *Translational Oncology*, vol. 1, p. 158, Dec. 2008.

[2] M. J. Stubbington, O. Rozenblatt-Rosen, A. Regev, and S. A. Teichmann, "Single-cell transcriptomics to explore the immune system in health and disease," *Science*, vol. 358, no. 6359, pp. 58–63, 2017.

[3] S. J. Horning, "A new cancer ecosystem," *Science*, vol. 355, pp. 1103–1103, Mar. 2017.

[4] G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas, "Massively parallel digital transcriptional profiling of single cells," *Nature Communications*, vol. 8, p. ncomms14049, Jan. 2017.

[5] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll, "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets," *Cell*, vol. 161, pp. 1202–1214, May 2015.

[6] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Sept. 2003. Google-Books-ID: AKuMj4PN_EMC.

[7] G. Heimberg, R. Bhatnagar, H. El-Samad, and M. Thomson, "Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing," *Cell Systems*, vol. 2, pp. 239–250, Apr. 2016.

[8] S. Dasgupta, "Learning mixtures of gaussians," in *Foundations of computer science, 1999. 40th annual symposium on*, pp. 634–644, IEEE, 1999.

[9] E. Candes, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?: Recovering low-rank matrices from sparse errors," in *2010 IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 201–204, Oct. 2010.

[10] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, "Integrating single-cell transcriptomic data across different conditions, technologies, and species," *Nature biotechnology*, vol. 36, no. 5, p. 411, 2018.

[11] L. Haghverdi, A. T. Lun, M. D. Morgan, and J. C. Marioni, "Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors," *Nature biotechnology*, vol. 36, no. 5, p. 421, 2018.

[12] S. R. Quake, T. Wyss-Coray, S. Darmanis, T. M. Consortium, *et al.*, "Single-cell transcriptomic characterization of 20 organs and tissues from individual mice creates a tabula muris," *bioRxiv*, p. 237446, 2018.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[14] W. Förstner and B. Moonen, "A metric for covariance matrices," in *Geodesy-The Challenge of the 3rd Millennium* (E. W. Grafarend, F. W. Krumm, and V. S. Schwarze, eds.), pp. 299–309, Berlin, Heidelberg: Springer Berlin Heidelberg, 2003.

[15] H. Takaba and H. Takayanagi, "The mechanisms of t cell selection in the thymus," *Trends in immunology*, 2017.

[16] A. Brown, "Immunological functions of splenic b-lymphocytes.," *Critical reviews in immunology*, vol. 11, no. 6, pp. 395–417, 1992.

[17] E. L. Gautier, T. Shay, J. Miller, M. Greter, C. Jakubzick, S. Ivanov, J. Helft, A. Chow, K. G. Elpek, S. Gordonov, *et al.*, "Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages," *Nature immunology*, vol. 13, no. 11, p. 1118, 2012.

[18] K. Murphy and C. Weaver, *Janeway's Immunobiology, 9th edition*. CRC Press, 2016.

[19] W. A. Lim and C. H. June, "The principles of engineering immune cells to treat cancer," *Cell*, vol. 168, no. 4, pp. 724 – 740, 2017.

[20] K. Newton and V. M. Dixit, "Signaling in innate immunity and inflammation," *Cold Spring Harbor Perspectives in Biology*, vol. 4, p. a006049, 03 2012.

[21] A.-C. Villani, S. Sarkizova, and N. Hacohen, "Systems immunology: Learning the rules of the immune system," *Annual Review of Immunology*, vol. 36, pp. 813–842, 2018/09/09 2018.

[22] G. Faure-André, P. Vargas, M.-I. Yuseff, M. Heuzé, J. Diaz, D. Lankar, V. Steri, J. Manry, S. Hugues, F. Vascotto, J. Boulanger, G. Raposo, M.-R. Bono, M. Rosemblatt, M. Piel, and A.-M. Lennon-Duménil, "Regulation of dendritic cell migration by CD74, the MHC class II-associated invariant chain," *Science*, vol. 322, pp. 1705–1710, Dec. 2008.

[23] X. Zhang, D. Ren, L. Guo, L. Wang, S. Wu, C. Lin, L. Ye, J. Zhu, J. Li, L. Song, H. Lin, and Z. He, "Thymosin beta 10 is a key regulator of tumorigenesis and metastasis and a novel serum marker in breast cancer," *Breast Cancer Res.*, vol. 19, p. 15, Feb. 2017.

[24] N. Müller, H. J. Fischer, D. Tischner, J. van den Brandt, and H. M. Reichardt, "Glucocorticoids induce effector T cell depolarization via ERM proteins, thereby impeding migration and APC conjugation," *J. Immunol.*, vol. 190, pp. 4360–4370, Apr. 2013.

[25] J. Ehrchen, L. Steinmüller, K. Barczyk, K. Tenbrock, W. Nacken, M. Eisenacher, U. Nordhues, C. Sorg, C. Sunderkötter, and J. Roth, "Glucocorticoids induce differentiation of a specifically activated, anti-inflammatory subtype of human monocytes," *Blood*, vol. 109, pp. 1265–1274, Feb. 2007.

[26] P. Mueller, J. Massner, R. Jayachandran, B. Combaluzier, I. Albrecht, J. Gatfield, C. Blum, R. Ceredig, H.-R. Rodewald, A. G. Rolink, and J. Pieters, "Regulation of T cell survival through coronin-1-mediated generation of inositol-1,4,5-trisphosphate and calcium mobilization after T cell receptor triggering," *Nat. Immunol.*, vol. 9, pp. 424–431, Apr. 2008.

[27] S. E. Lipshultz, N. Rifai, V. M. Dalton, D. E. Levy, L. B. Silverman, S. R. Lipsitz, S. D. Colan, B. L. Asselin, R. D. Barr, L. A. Clavell, C. A. Hurwitz, A. Moghrabi, Y. Samson, M. A. Schorin, R. D. Gelber, and S. E. Sallan, "The effect of dexrazoxane on myocardial injury in doxorubicin-treated children with acute lymphoblastic leukemia," *N. Engl. J. Med.*, vol. 351, pp. 145–153, July 2004.

[28] S. K. Kumar, V. Rajkumar, R. A. Kyle, M. van Duin, P. Sonneveld, M.-V. Mateos, F. Gay, and K. C. Anderson, "Multiple myeloma," *Nature Reviews Disease Primers*, vol. 3, pp. 17046 EP –, 07 2017.

[29] E. Malek, M. de Lima, J. J. Letterio, B.-G. Kim, J. H. Finke, J. J. Driscoll, and S. A. Giralt, "Myeloid-derived suppressor cells: The green light for myeloma immune escape," *Blood Reviews*, vol. 30, no. 5, pp. 341 – 348, 2016.

[30] L. M. Pilarski, E. Joy Andrews, M. J. Mant, and B. A. Ruether, "Humoral immune deficiency in multiple myeloma patients due to compromised b-cell function," *Journal of Clinical Immunology*, vol. 6, pp. 491–501, Nov 1986.

[31] M. Bolzoni, D. Ronchetti, P. Storti, G. Donofrio, V. Marchica, F. Costa, L. Agnelli, D. Toscani, R. Vescovini, K. Todoerti, S. Bonomini, G. Sammarelli, A. Vecchi, D. Guasco, F. Accardi, B. D. Palma, B. Gamberi, C. Ferrari, A. Neri, F. Aversa, and N. Giuliani, "Il21r expressing cd14+cd16+ monocytes expand in multiple myeloma patients leading to increased osteoclasts," *Haematologica*, vol. 102, no. 4, pp. 773–784, 2017.

[32] C. Botta, A. Gullà, P. Correale, P. Tagliaferri, and P. Tassone, "Myeloid-derived suppressor cells in multiple myeloma: Pre-clinical research and translational opportunities," *Frontiers in Oncology*, vol. 4, p. 348, 2014.

[33] T. Dosani, F. Covut, R. Beck, J. J. Driscoll, M. de Lima, and E. Malek, "Significance of the absolute lymphocyte/monocyte ratio as a prognostic immune biomarker in newly diagnosed multiple myeloma," *Blood Cancer Journal*, vol. 7, pp. e579 EP –, 06 2017.

[34] A. C. Rawstron, F. E. Davies, R. G. Owen, A. English, G. Pratt, J. A. Child, A. S. Jack, and G. J. Morgan, "B-lymphocyte suppression in multiple myeloma is a reversible phenomenon specific to normal b-cell progenitors and plasma cell precursors," *British journal of haematology*, vol. 100, no. 1, pp. 176–183, 1998.

[35] R. J. Pessoa-Magalhaes, M.-B. Vidriales, B. Paiva, C. F. Gimenez, R. Garcia-Sanz, M. V. Mateos, N. Gutierrez, Q. Lecrevisse, J. F. Blanco, J. Hernandez, *et al.*, "Analysis of the immune system of multiple myeloma patients achieving long-term disease control, by multidimensional flow cytometry," *Haematologica*, pp. haematol–2012, 2012.

[36] M. Winqvist, F. Mozaffari, M. Palma, S. Eketorp Sylvan, H. Mellstedt, A. Osterborg, and J. Lundin, "In vivo effects of lenalidomide on t cell proliferation and immune checkpoint molecules in patients with advanced stage cll: Results from a phase ii study," *Blood*, vol. 126, no. 23, pp. 4164–4164, 2015.

[37] V. Bronte, S. Brandau, S.-H. Chen, M. P. Colombo, A. B. Frey, T. F. Greten, S. Mandruzzato, P. J. Murray, A. Ochoa, S. Ostrand-Rosenberg, P. C. Rodriguez, A. Sica, V. Umansky, R. H. Vonderheide, and D. I. Gabrilovich, "Recommendations for myeloid-derived suppressor cell nomenclature and characterization standards," *Nature Communications*, vol. 7, pp. 12150 EP –, 07 2016.

[38] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, pp. 1289–1306, Apr. 2006.

[39] D. Mumford and A. Desolneux, *Pattern theory: the stochastic analysis of real-world signals*. AK Peters/CRC Press, 2010.

[40] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix tri-factorizations for clustering," *KDD 06*, 2006.

[41] Z. S. Singer, J. Yong, J. Tischler, J. A. Hackett, A. Altinok, M. A. Surani, L. Cai, and M. B. Elowitz, "Dynamic heterogeneity and dna methylation in embryonic stem cells," *Molecular cell*, vol. 55, no. 2, pp. 319–331, 2014.

[42] R. M. Kumar, P. Cahan, A. K. Shalek, R. Satija, A. J. DaleyKeyser, H. Li, J. Zhang, K. Pardee, D. Gennert, J. J. Trombetta, *et al.*, "Deconstructing transcriptional heterogeneity in pluripotent stem cells," *Nature*, vol. 516, no. 7529, p. 56, 2014.

[43] A. Sanchez, S. Choubey, and J. Kondev, "Stochastic models of transcription: From single molecules to single cells," *Methods*, vol. 62, no. 1, pp. 13–25, 2013.

[44] S. Watanabe, *Algebraic geometry and statistical learning theory*, vol. 25. Cambridge University Press, 2009.

[45] S. Gordon and P. R. Taylor, "Monocyte and macrophage heterogeneity," *Nature Reviews Immunology*, vol. 5, no. 12, p. 953964, 2005.

[46] C. S. McGinnis, D. M. Patterson, J. Winkler, M. Y. Hein, V. Srivastava, D. N. Conrad, L. M. Murrow, J. S. Weissman, Z. Werb, E. D. Chow, *et al.*, "Multi-seq: Scalable sample multiplexing for single-cell rna sequencing using lipid-tagged indices," *bioRxiv*, p. 387241, 2018.

Acknowledgements Eric Chow, Chris McGinnis, David Patterson, Allan Pool Hermann, Jase Gehring, Members of the Thomson Lab, Beckman Institute Single-cell Profiling and Engineering Center (SPEC).

Figure Captions

Figure 1. Summary of PopAlign framework PopAlign provides a scalable method for deconstructing quantitative changes in population structure including cell-state abundance and gene expression across many single cell experimental samples. (a) Users input PopAlign single-cell gene expression data from a 'Reference' sample, and at least one 'Test' sample, which are each a collection of n-dimensional gene expression vectors $\boldsymbol{g}$, shown as single dots. (b) For each sample, PopAlign estimates a low-dimensional probabilistic model that represents the distribution of gene expression states as a mixture of local Gaussian densities ϕ_i with parameters encoding subpopulation abundance (w_i), mean gene expression state (μ_i), and population spread (Σ_i). PopAlign reduces the dimensionality of the input data by representing each gene expression vector as set of m gene expression features ($m = 10-20$), thus representing each cell as an m-dimensional vector of coefficients $\boldsymbol{c}$. (c) Each ϕ_i^{test} in the test population is aligned to the closest ϕ_i^{ref} in the Reference sample by minimizing Jeffrey's divergence. (d) Following alignment, the parameters of aligned subpopulation pairs are compared to identify subpopulation-specific shifts in cellular abundance Δw, shifts in mean gene expression state $\Delta\mu$ and shifts in subpopulation shape $\Delta\Sigma$.

Figure 2. PopAlign models represent experimental data with high qualitative and quantitative accuracy (a) Experimental data for ~ 3600 bone marrow cells projected into an $m = 15$ dimensional gene feature space. 2D plots show single cells projected along gene feature pairs (c_i, c_j), and a single selected 3D projection (inset) is shown. Blue axis denote shared axis between 2D and 3D plot. (b) Model generated data for the same 2D and 3D feature space projections shown in (a). In the 3D projection (inset), each maroon circle denotes the centroid (μ) of a Gaussian mixture component where the circle radius is proportional to w_i, the mixture weight. In all cases, the model generated data replicates the qualitative geometric structures in the experimental data. (c) Model error across 2D projections from (a) and (b) quantified by analyzing percent deviation in point density for experimental vs model generated data. Quantification of error is performed using a numerical error metric based on binning the 2D projections (See Methods - Analysis of model error). The quantitative error in model-generated is on average 11.3% (red line) compared with 8.5% error when comparing random sub-samples of experimental data (blue line).

Figure 3. Probabilistic models identify, align, and dissect cellular subpopulations across disparate tissues Experimental single-cell data (black) for two tissues, mammary gland (a) and limb muscle (b), are plotted together with PopAlign model-generated data (teal) using a 2D t-SNE transformation. Both experimental datasets contain ~ 3600 cells. For each tissue, mixture model centroids (μ) are indicated as numbered disks. (c-d) For models from both tissues, mixture components (x-axis) are scored using cell type annotations supplied by Tabula Muris (y-axis). Each cell of the heatmap represents the percentage of cells associated with each mixture component that have a specific cell type label. Columns (but not rows) sum to 1. (e) Alignments between mixture component centroids (μ) from the reference population (Mammary Gland) and the test population (Limb Muscle) are shown as connecting lines. All m-dimensional mu vectors are transformed using principal components analysis (PCA) and plotted using the first 3 PCS. Width of each line

is inversely proportional to the p-value associated with the alignment (see Legend). (f) Null distribution of Jeffrey's divergence used to calculate p-values. Jeffrey's divergence was calculated for all possible pairs of mixture components from models of all tissues from Tabula Muris. (g) We rank aligned subpopulations in terms of maximum Δw and show top two pairs. These subpopulations are identified as T cells and endothelial cells, and are highlighted using a blue dotted line in (e). We find that T cells are highly abundant in Mammary gland while endothelial cells are highly abundant in muscle. (h) Comparing subpopulation centroids (μ) for macrophages in terms of gene expression features. Macrophages in Mammary gland and Limb Muscle share common features (black font), but also have tissue-specific features (red font). Corresponding alignments are highlighted in (e) with a gray dotted line.

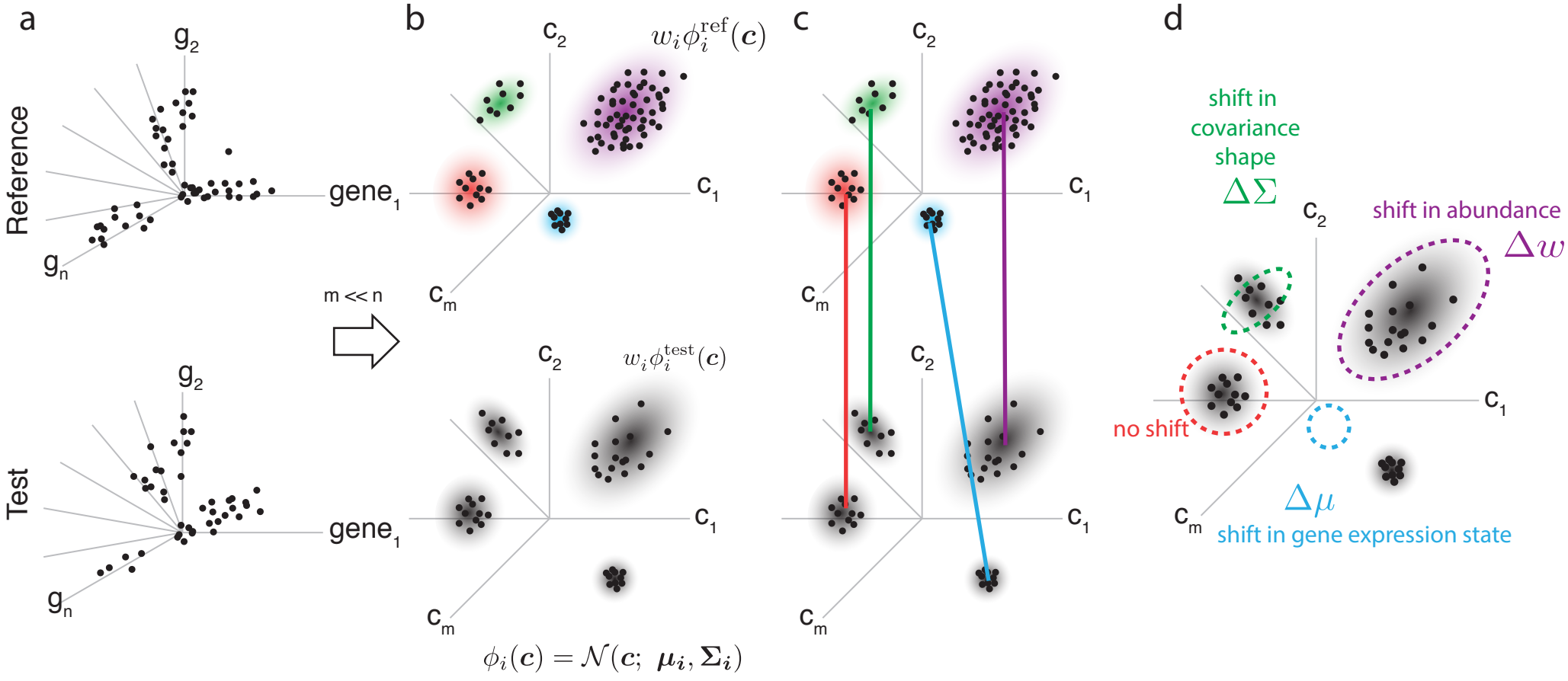
Figure 4. PopAlign can perform global comparisons of cell states across dozens to hundreds of experimental samples (a) Computational runtime versus number of samples for PopAlign (blue) vs Seurat's CCA-based alignment method (red). PopAlign scales linearly with the number of samples, while CCA scales exponentially and encounters an out-of-memory error when applied to ¿ 8 samples. Samples are bootstrapped from all 12 samples of the mouse tissue survey Tabula Muris. Benchmarking tests performed on typical workstation (8 cores, 64GB RAM). (b) Heatmap of a pairwise similarity metric between subpopulations from all 12 tissues demonstrates PopAlign can identify cogent cell-type specific clusters even when applied on very disparate tissue types. The similarity metric is defined as exp(-JD) where JD is the Jeffrey's Divergence between two subpopulations. Inset highlights subpopulations clustered as macrophages, displaying tissue and cell type labels extracted from Tabula Muris annotations. (c-f) Models for all tissues are aligned to a reference model (Mammary Gland) and corresponding abundances (w) are plotted for selected subpopulations classified as (c) T cells (d) B cells (e) endothelial cells (f) macrophages. (g) Mean gene expression state () for macrophage across all tissues show variation in key immune pathways (highlighted in red).

Figure 5. PopAlign identifies universal and cell-type specific impacts of immunomodulatory drugs (a) Rendering of GMM model for the control sample 1 projected onto the first 2 principal components. Abundance weights (w) are represented by the size of the circle, and supplied as a text label. (b) Ranking of all drugs based on population-level similarity to control population using the log-likelihood ratio metric (LLR). P-values are calculated using an FDR-corrected one-sample t-test of the 6 control replicates against the drug's mean LLR. Dashed line: p-value = 0.05. (c)-(d) Gene expression shifts ($\Delta\mu$) for drug-exposed (c) monocyte and (d) T cell subpopulations with respect to their aligned subpopulation in control sample 1. Each small black dot represents a separate bootstrapped model built from a randomly chosen subsample (80%) of the same data. The large dot indicates the mean $\Delta\mu$, and is colored by - log(p-value). P-values are calculated using an FDR-corrected one-sample t-test testing the 6 control replicates against the drug's mean Δmu. Gray box: The 95% confidence interval of the control mean. Dashed line: p-value = 0.05. (e) Number of T-cell-specific, monocyte-specific, and overlapping genes across drugs with an LLR p-value < 0.05. The maximum percentage of shared genes is 16%. (f) L1-distance metrics are shown for Budesonide, a glucocorticoid that impacts both monocytes and T cells. Differentially expressed genes that overlap between both cell types are denoted in yellow. Up-regulated genes:

blue. Down-regulated genes: red. Non-significant L1-distance values: white. (g) T-cell gene expression shifts ($\Delta\mu$) for dexrazoxane relative to controls. (h) Monocyte gene expression shifts ($\Delta\mu$) for dexrazoxane relative to controls. (i) Gene expression distributions showing up-regulation of GPX4, CORO1A, and PRDX1 in Dexrazoxane-exposed T cells.

Figure 6. Discovering signatures of disease and treatment in PBMCs from multiple myeloma patients (a-f) Experimental single cell mRNA-seq data from two healthy donors and four multiple myeloma patients (MM1-4) are projected into 16-dimensional gene feature space. 3D plots show single cells in a subset of three gene features that highlight separation between different immune cell types. Mixture model centroids (μ) are indicated as numbered disks. Subpopulations in test samples are aligned to the reference (a) and changes in abundance (Δw) are plotted for (g) B cells, (h) monocytes, i) naive T cells, and j) effector T cells, showing general and patient specific changes. (k) Mean gene expression levels for two markers of myeloid derived suppressor cells (MDSC), CD33 and CD11b, are plotted for all monocyte subpopulations. Error bars denote confidence interval of the mean. (m) $|\Delta\mu|$ for naive T cell and effector T cell populations relative to healthy1. (n) Heatmap of mixture component μ vectors in terms of feature coefficients c_i for aligned naive and effector T cells across samples. MM subpopulations exhibit reduced expression of two features (red font): Leukocyte motility and Cytotoxic Lymphocyte Killing. (o) Distribution of beta-actin (ACTB) expression for all effector T cell subpopulations across samples. Violin shows distribution, and mean is denoted by white circle. (p) Distribution of perforin 1 (PFN1) expression for all effector T cell subpopulations across samples. For single gene plots (k), (o), (p), units are in terms of normalized and transformed gene expression $(\log(\mathrm{g}+1))$.

Figure 1



Single-cell Gene Expression Data

Low-dimensional Probabilistic Model

Subpopulation Alignment

Compare parameters between alignments

Figure 2

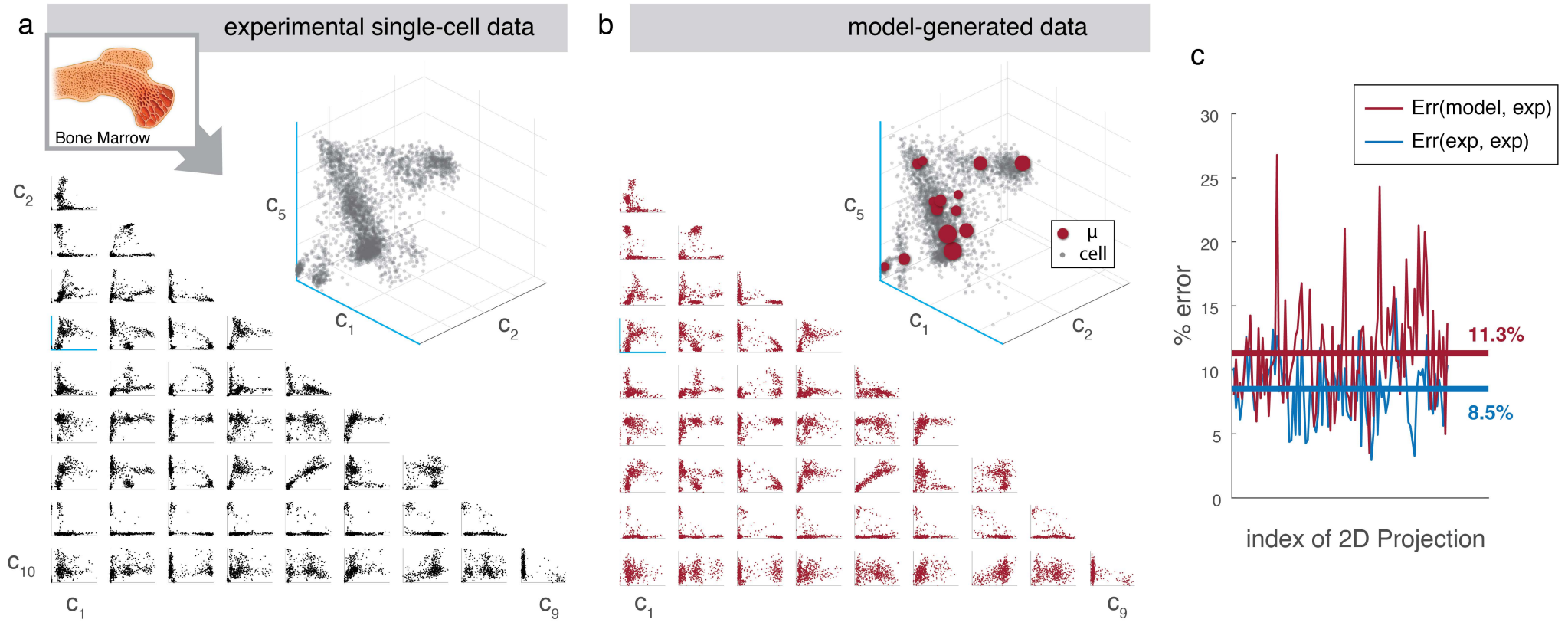
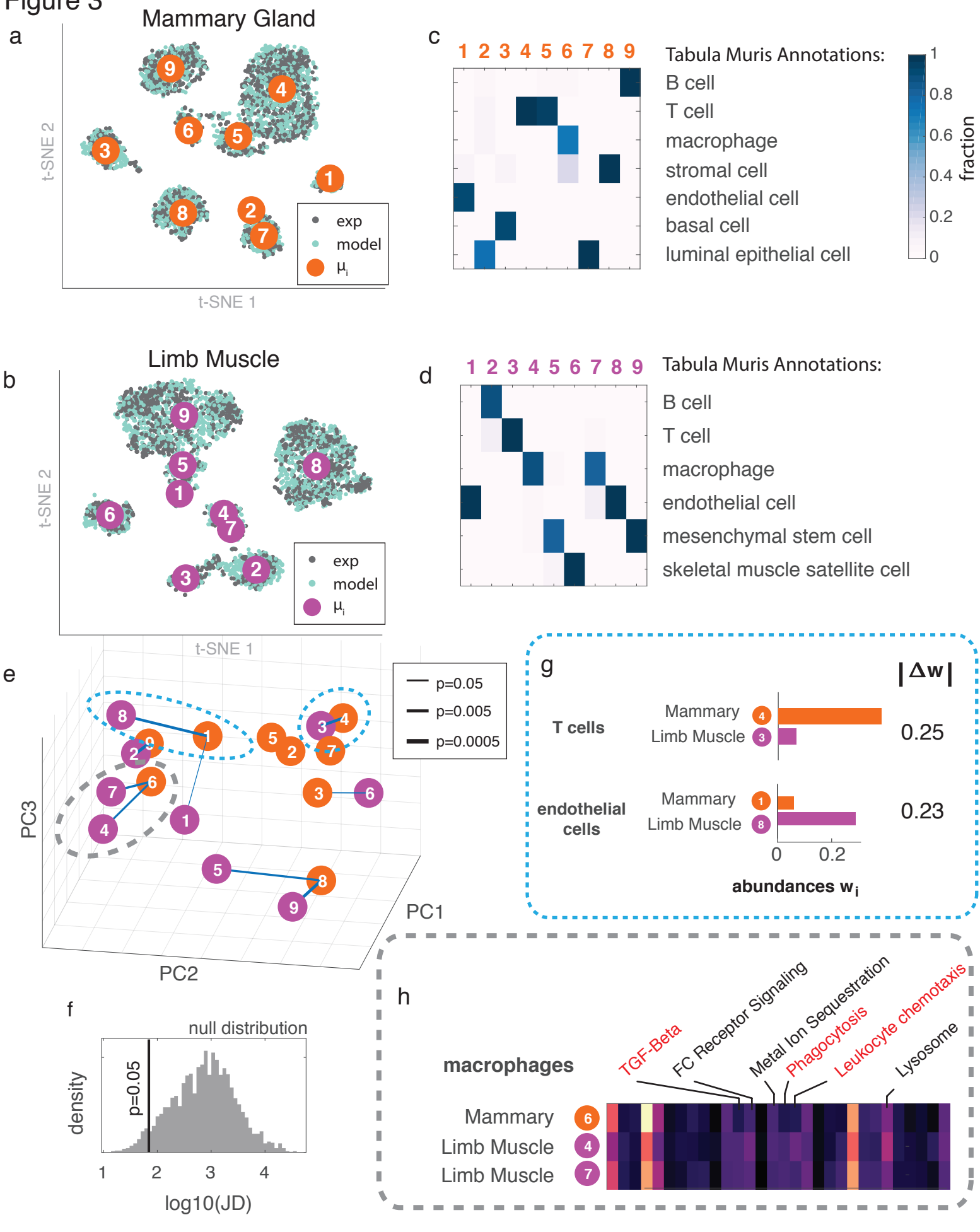

a
experimental single-cell data
Bone Marrow
C2
C5
C1
C2
C10
C1
C9
b
model-generated data
C5
μ
cell
C1
C2
C1
C9
c
Err(model, exp)
Err(exp, exp)
30
25
20
15
10
5
0
% error
11.3%
8.5%
index of 2D Projection

Figure 3


a
Mammary Gland
t-SNE 2
t-SNE 1
exp
model
μi
b
Limb Muscle
t-SNE 2
t-SNE 1
exp
model
μi
c
1 2 3 4 5 6 7 8 9
Tabula Muris Annotations:
B cell
T cell
macrophage
stromal cell
endothelial cell
basal cell
luminal epithelial cell
fraction
d
1 2 3 4 5 6 7 8 9
Tabula Muris Annotations:
B cell
T cell
macrophage
endothelial cell
mesenchymal stem cell
skeletal muscle satellite cell
e
PC3
PC2
PC1
p=0.05
p=0.005
p=0.0005
f
null distribution
density
p=0.05
log10(JD)
g
|Δw|
T cells
Mammary
Limb Muscle
0.25
endothelial cells
Mammary
Limb Muscle
0.23
abundances wi
h
macrophages
TGF-Beta
FC Receptor Signaling
Metal Ion Sequestration
Phagocytosis
Leukocyte chemotaxis
Lysosome
Mammary
Limb Muscle
Limb Muscle

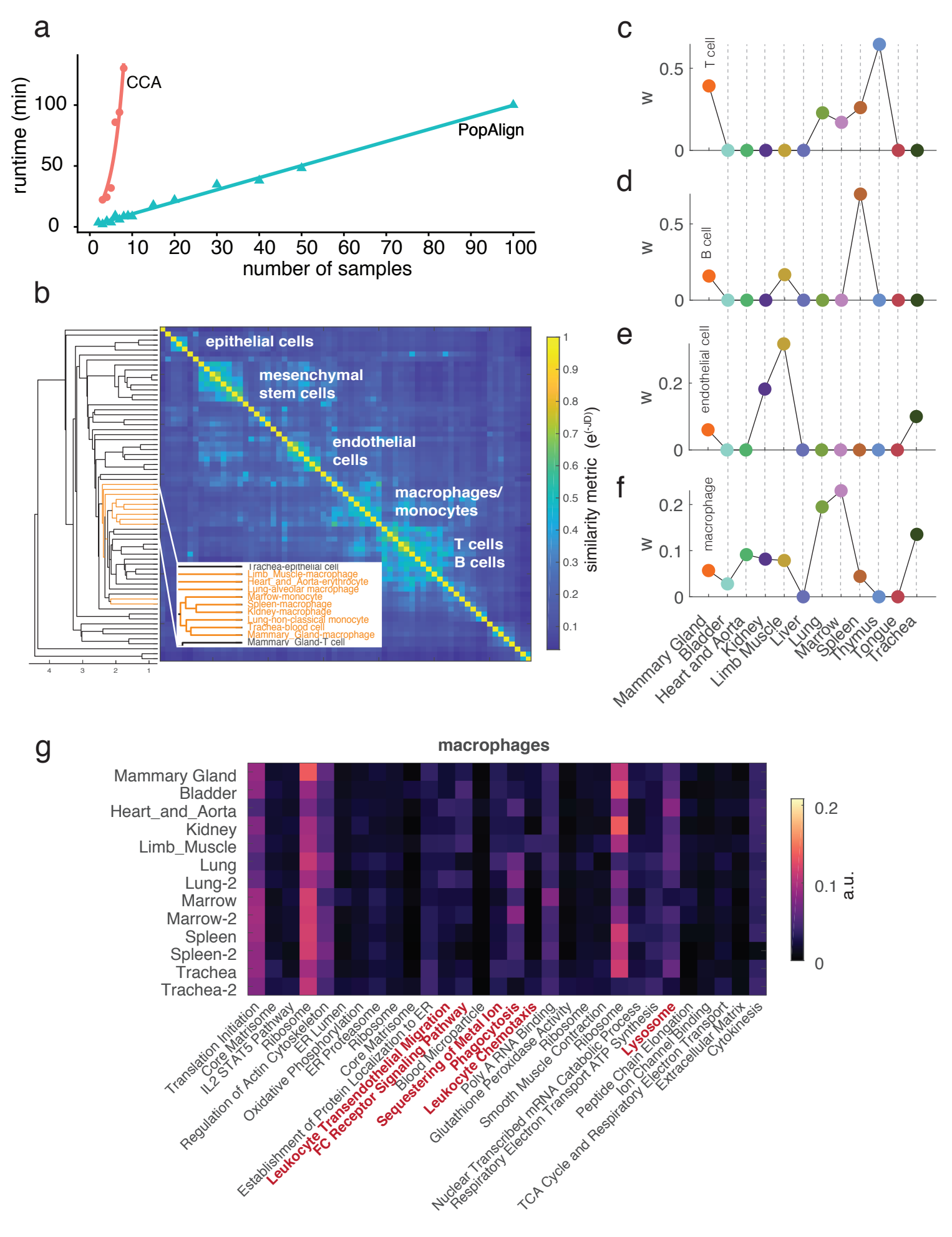

a
runtime (min)
100
50
0
CCA
PopAlign
0 10 20 30 40 50 60 70 80 90 100
number of samples
b
epithelial cells
mesenchymal stem cells
endothelial cells
macrophages/ monocytes
T cells
B cells
Trachea-epithelial cell
Limb_Muscle-macrophage
Heart_and_Aorta-erythrocyte
Lung-alveolar macrophage
Marrow-monocyte
Spleen-macrophage
Kidney-macrophage
Lung-non-classical monocyte
Trachea-blood cell
Mammary_Gland-macrophage
Mammary_Gland-T cell
similarity metric (e(-JD))
1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1
4 3 2 1
c
T cell
w
0.5
0
d
B cell
w
0.5
0
e
endothelial cell
w
0.2
0
f
macrophage
w
0.2
0.1
0
Mammary Gland
Bladder
Heart and Aorta
Kidney
Limb Muscle
Liver
Lung
Marrow
Spleen
Thymus
Tongue
Trachea
g
macrophages
Mammary Gland
Bladder
Heart_and_Aorta
Kidney
Limb_Muscle
Lung
Lung-2
Marrow
Marrow-2
Spleen
Spleen-2
Trachea
Trachea-2
0.2
0.1
0
a.u.
Translation Initiation
Core Matrisome
IL2 STAT5 Pathway
Ribosome
Regulation of Actin Cytoskeleton
Oxidative Phosphorylation
ER Lumen
ER Proteasome
Ribosome
Core Matrisome
Establishment of Protein Localization to ER
Blood Microparticle
Leukocyte Transendothelial Migration
FC Receptor Signaling Pathway
Sequestering of Metal Ion
Phagocytosis
Leukocyte Chemotaxis
Poly A RNA Binding
Glutathione Peroxidase Activity
Ribosome
Smooth Muscle Contraction
Ribosome
Nuclear Transcribed mRNA Catabolic Process
Respiratory Electron Transport ATP Synthesis
Lysosome
Peptide Chain Elongation
TCA Cycle and Respiratory Electron Transport
Ion Channel Binding
Extracellular Matrix
Cytokinesis

Figure 5: PopAlign identifies universal and cell-type specific impacts of drugs

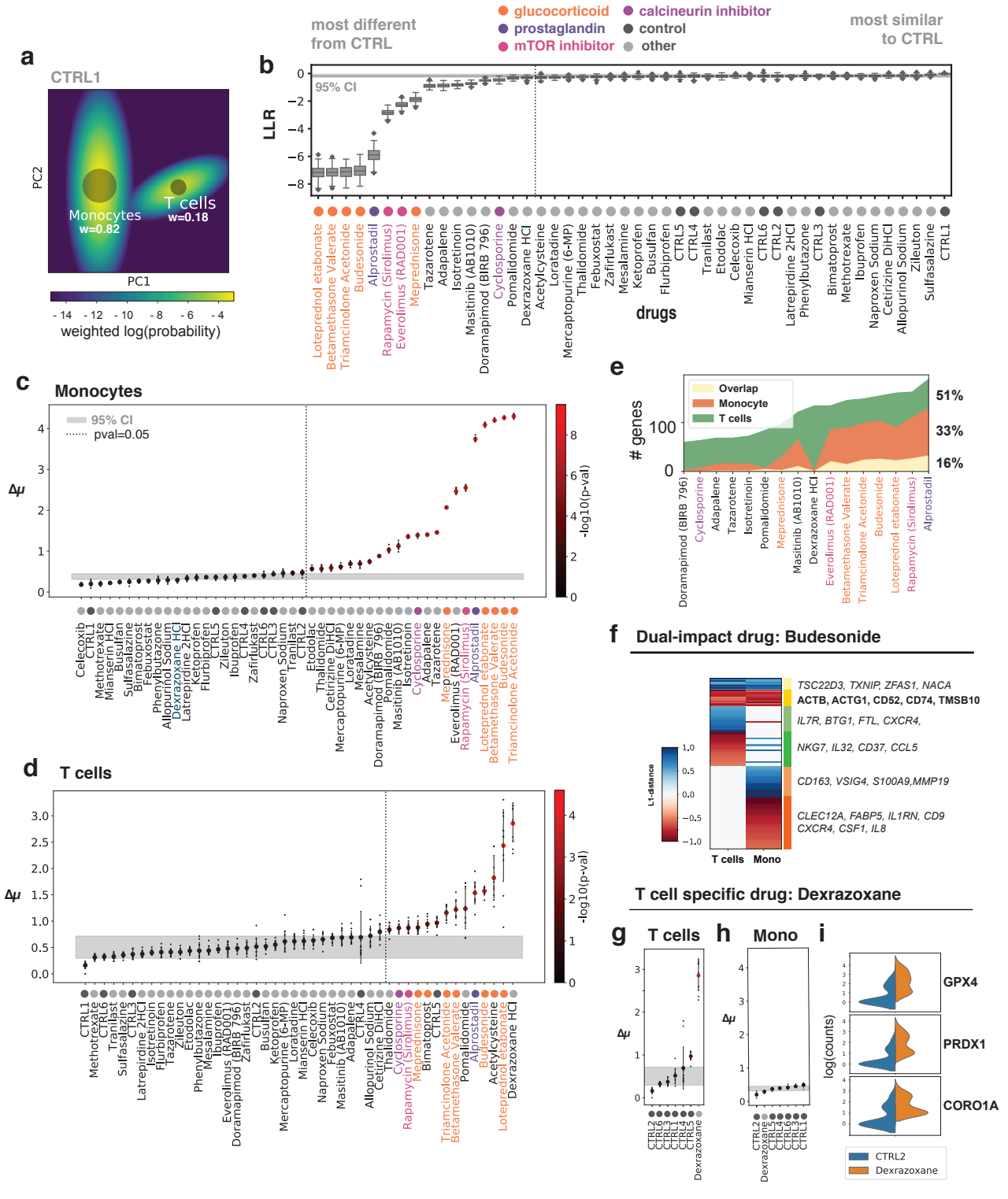
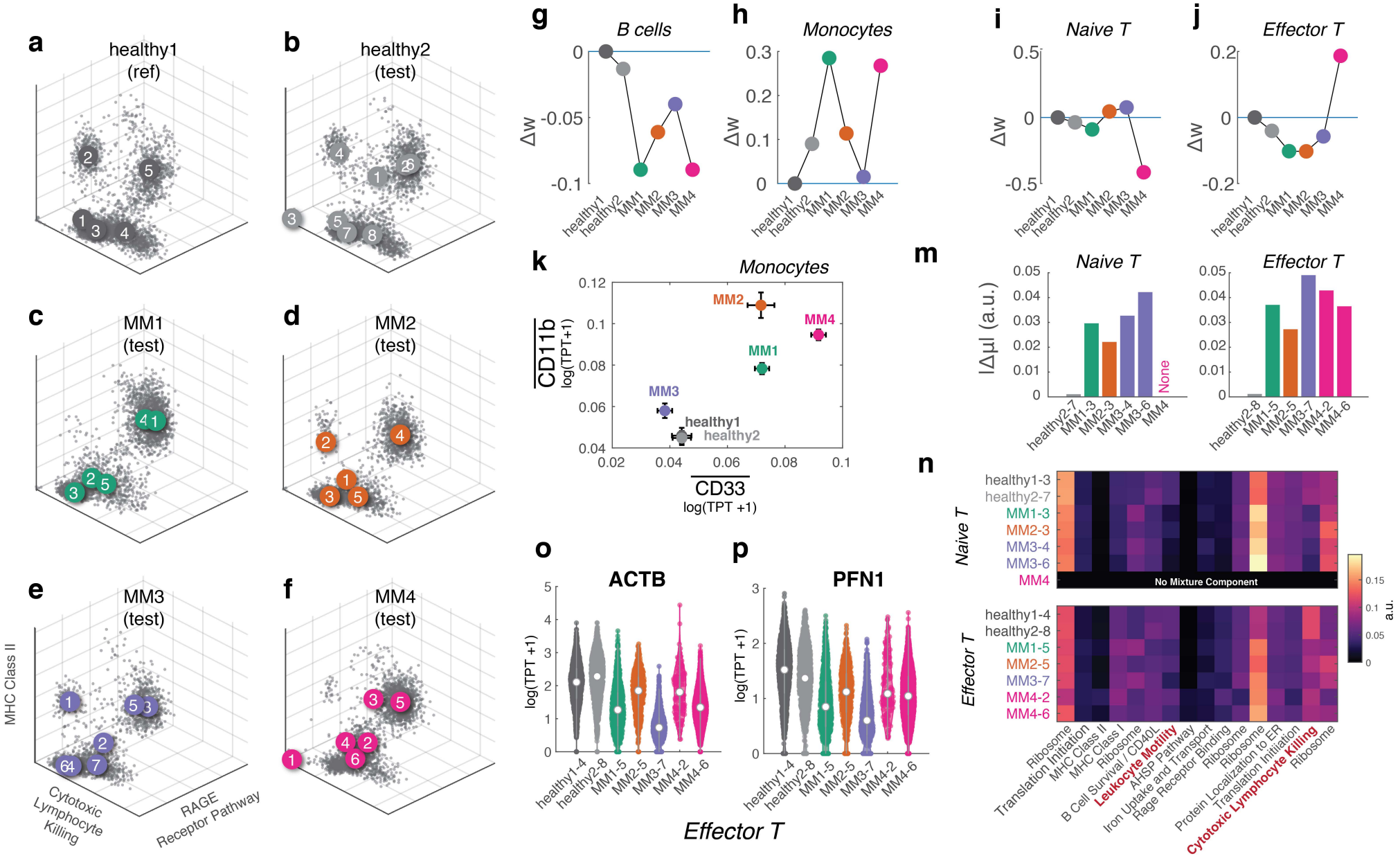


Figure 6


a
healthy1
(ref)
b
healthy2
(test)
c
MM1
(test)
d
MM2
(test)
e
MM3
(test)
f
MM4
(test)
MHC Class II
Cytotoxic
Lymphocyte
Killing
RAGE
Receptor Pathway
g
B cells
Δw
healthy1
healthy2
MM1
MM2
MM3
MM4
h
Monocytes
i
Naive T
j
Effector T
k
Monocytes
CD11b
log(TPT+1)
CD33
log(TPT +1)
m
Naive T
Effector T
|Δμl (a.u.)
None
n
Naive T
Effector T
No Mixture Component
a.u.
Ribosome
Translation Initiation
MHC Class II
MHC Class I
Ribosome
B Cell Survival / CD40L
Leukocyte Motility
AHSP Pathway
Iron Uptake and Transport
Rage Receptor Binding
Ribosome
Protein Localization to ER
Translation Initiation
Cytotoxic Lymphocyte Killing
Ribosome
o
ACTB
p
PFN1
log(TPT +1)
Effector T