Dissecting heterogeneous cell populations across drug and disease conditions with PopAlign

Sisi Chen[a,b,1], Paul Rivaud[a,b], Jong H. Park[a,b], Tiffany Tsou[a,b], Emeric Charles[c], John R. Haliburton[d], Flavia Pichiorri[e], and Matt Thomson[a,b,1]

[a]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125; [b]Beckman Center for Single-Cell Profiling and Engineering, California Institute of Technology, Pasadena, CA 91125; [c]Department of Molecular and Cell Biology, University of California Berkeley, Berkeley, CA 94720; [d]Augmenta Bioworks Inc., Menlo Park, CA 94025; and [e]Department of Hematologic Malignancies Translational Science, City of Hope, Monrovia, CA 91016

Edited by Jonathan S. Weissman, University of California, San Francisco, CA, and approved September 25, 2020 (received for review March 31, 2020)

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.

single-cell genomics | probabilistic models | single cell mRNA-seq

Introduction

All physiological processes in the body are driven by heterogeneous populations of single cells (1–3). 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.

Many single-cell analysis tools have been designed to provide in-depth characterization of cell states and developmental trajectories within a single-cell population. However, once clusters or trajectories are identified, existing methods do not provide a formal way to compare different populations or samples with each other. In this paper, we introduce a computational framework, PopAlign, that was designed to provide an integrated representation of the cell population within a sample, so that samples can be compared at different scales of representation, ranging from gene-expression programs, to cell states, to the structure of the entire cell population. PopAlign identifies, aligns, and tracks subpopulations of single cells within a heterogeneous cell population profiled by single-cell RNA sequencing (scRNA-seq) (2, 4, 5). Mathematically, PopAlign constructs a low-dimensional probabilistic model of each cell population across a series of samples. PopAlign 1) automatically identifies and models subpopulations of cells; 2) aligns cellular subpopulations across experimental conditions (signaling, disease); and 3) 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 identified by orthogonal nonnegative matrix factorization (oNMF). 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 low-dimensional representation of cell state in terms of a set of gene-expression features learned from data (6–8).

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- to 100-fold. Further, downstream computations, including population alignment, are performed on

Significance

Many physiological processes are driven by changes across heterogeneous populations of cells. However, we currently lack a conceptual framework for comparing single-cell transcriptional data collected from populations as they respond to perturbations. Here, we develop a framework, called PopAlign, that models a cell population as a probability distribution in gene-expression space. Individual subpopulations are represented as local densities that are statistically aligned across samples. The models present an integrated representation that allows comparisons at multiple scales, from gene-expression programs to cell state to population structure. The models are memory-efficient and can be scaled across hundreds of samples, which we demonstrate using public data and data from human immune cells responding to drugs and disease.

Author contributions: S.C., P.R., F.P., and M.T. designed research; S.C., P.R., J.H.P., T.T., E.C., J.R.H., and M.T. performed research; S.C., P.R., and M.T. contributed new reagents/analytic tools; S.C., P.R., and M.T. analyzed data; S.C. and M.T. wrote the paper; and F.P. provided clinical guidance and context for data interpretation.

Competing interest statement: S.C., M.T., and P.R. have filed a US and Patent Cooperation Treaty patent for the PopAlign computational framework.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: mthomson@caltech.edu or sisi.chen1@gmail.com.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2005990117/-/DCSupplemental.

First published October 30, 2020.

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 t-distributed stochastic neighbor embedding (tSNE) rely on pairwise computations between individual cells, which is compute-intensive and requires storing of many raw single-cell datasets in memory.

PopAlign is particularly well-suited to analyzing large-scale datasets with many samples, such as those generated for surveys of diseased tissues (9, 10) or data generated by using new sample multiplexing technologies (11–13). Multiplexing technologies specifically allow convenient interrogation of cell populations across hundreds of different experimental conditions. However, PopAlign also offers unique capabilities for analyzing small numbers of samples because it provides a formalized representation of cell-state and quantitative statistical metrics for analyzing subpopulation and whole-population changes.

We assessed the accuracy and generality of PopAlign using 12 datasets from a mouse-tissue survey (Tabula Muris) (14) as well as experiments on human peripheral blood cells, including a screen of immunomodulatory drugs and a comparison of healthy patients to disease (multiple myeloma [MM]). 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 MM 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.

Results

PopAlign Represents Heterogeneous Cell Populations with Probabilistic Mixture Models. We develop a mathematical and computational framework (PopAlign) that 1) identifies and aligns cell states across paired populations of single cells (a reference population and a test population), and then 2) 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 reference population ($\boldsymbol{D}^{\text{ref}}$) and a test population ($\boldsymbol{D}^{\text{test}}$), that are profiled with single-cell mRNA-seq (Fig. 1*A*). Profiling of each population generates a set of gene-expression vectors, e.g., $\boldsymbol{D}^{\text{test}} = \{\boldsymbol{g}_i\}_{i=1}^{k}$, where $\boldsymbol{g} = (g_1, g_2, \ldots, g_n)$ 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. We normalized and scaled $\boldsymbol{g}$ to account for technical variability in transcript capture and then log-transformed for downstream computations (*SI Appendix*, *Data Normalization*).

To compare the reference and test-cell populations, we first constructed a probabilistic model of the gene-expression distribution for each set of cells (Fig. 1*B*). The high-dimensional nature of gene-expression ($n \sim 20,000$) space makes the

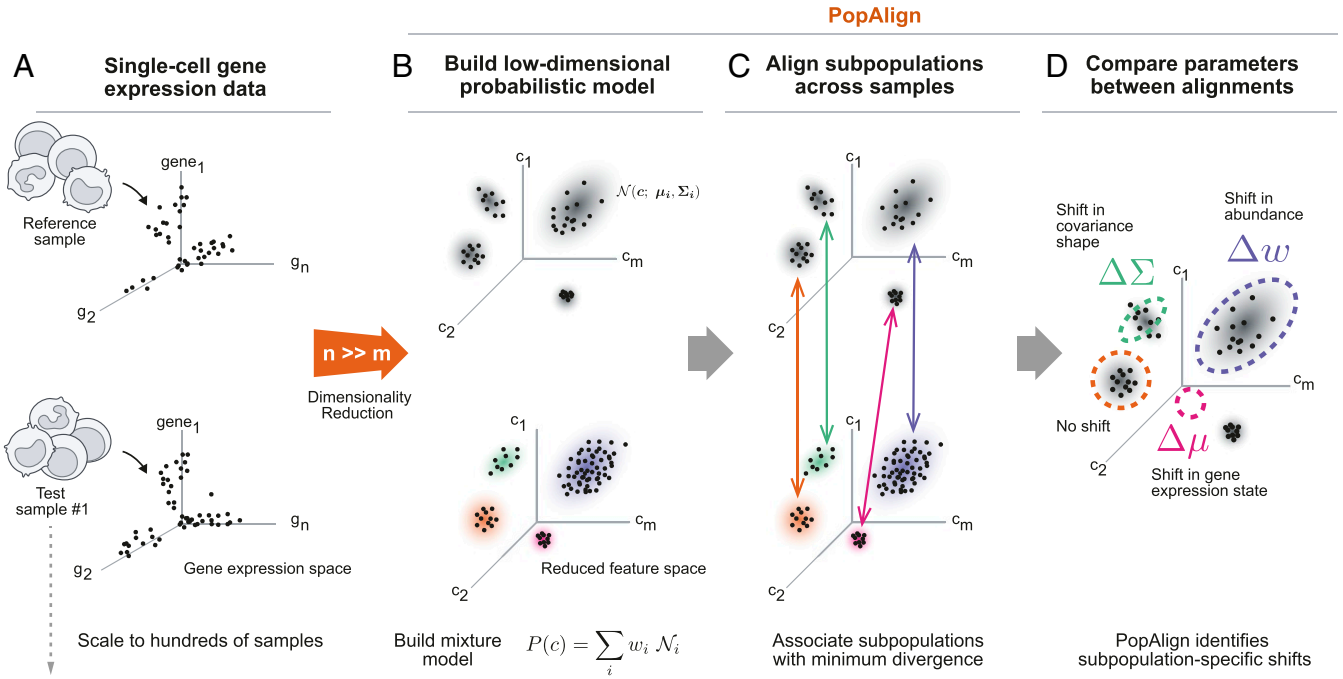

Fig. 1. Summary of PopAlign framework. PopAlign provides a scalable method for deconstructing quantitative changes in population structure, including cell-state abundance and gene expression, across many single-cell experimental samples. (*A*) Users input single-cell gene-expression data from a "Reference" sample and at least one "Test" sample, which each are a collection of n-dimensional gene-expression vectors $\boldsymbol{g}$, shown as single dots. PopAlign reduces the dimensionality of the input data by representing each gene-expression vector as a set of m gene-expression features ($m = 10 - 20$), thus representing each cell as an m-dimensional vector of coefficients $\boldsymbol{c}$. (*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 $\mathcal{N}_i$ with parameters encoding subpopulation abundance (w_i), mean gene-expression state (μ_i), and population spread (Σ_i). (*C*) Each $\mathcal{N}_i^{\text{test}}$ in the Test population is aligned to the closest $\mathcal{N}_i^{\text{ref}}$ in the Reference sample by minimizing Jeffreys 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 $\Delta\Sigma$.

inference and interpretation of probabilistic models challenging. Therefore, we represented each cell, not as a vector of genes, but as a vector of gene-expression programs or gene-expression features that were extracted from the data, so that each single cell was represented as a vector $\boldsymbol{c} = (c_1, c_2 \cdots c_m)$ of m feature coefficients, c_i, which weighted the magnitude of gene-expression programs in a given cell (*SI Appendix*, *Extraction of Gene Feature Vectors Using Matrix Factorization*).

We extracted these gene features using a particular matrix-factorization method called orthogonal nonnegative matrix factorization (oNMF) (15) that produces a useful set of features because all vectors are positive and composed of largely nonoverlapping genes (*SI Appendix*, Figs. S1*B*, S2, and S3). This allowed us to naturally think of a cell's transcriptional state as a linear sum of different positive gene-expression programs (16). Other methods like principal components analysis (PCA) can be more difficult to interpret because they produce vectors which contain contributions from overlapping sets of genes. If two features, f_1 and f_2, contain similar sets of genes, then representations which use these features can hide underlying cell-state similarity. oNMF forces shared genes to be separated into their own program, so, in the example above, shared genes become a third program f_3, and the two cell types can be represented as $f_1 + f_3$ and $f_2 + f_3$.

Choosing the number of features to use involves balancing a tradeoff between accuracy and dimensionality. To provide a principled way to choose the number of features, we constructed a loss function $f(m)$ that places a penalty on high values of m (*SI Appendix*, Figs. S4 and S5 and *Extraction of Gene Feature Vectors*) and uses the function to identify a local minimum. The PopAlign package allows users to modulate the exact choice of m by tuning the loss function, thus constructing a coarse- or fine-grained representation that is appropriate for the exact use case.

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 estimated a probabilistic model, $P^{\text{test}}(\boldsymbol{c})$ and $P^{\text{ref}}(\boldsymbol{c})$, for the reference and test-cell populations that intrinsically factored each population into a set of distinct subpopulations, each represented by a Gaussian probability density (density depicted as individual "clouds" in Fig. 1*B*):

$$P^{\text{test}}(\boldsymbol{c}) = \sum_{i=1}^{l} w_i \, \phi_i^{\text{test}}(\boldsymbol{c}) \qquad [1]$$

$$\text{where } \phi_i^{\text{test}}(\boldsymbol{c}) = \mathcal{N}(\boldsymbol{c}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i),$$

where $\mathcal{N}(\boldsymbol{c}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ 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 estimated the parameters of the mixture model $(\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, w_i\})$ from single-cell data using the expectation-maximization algorithm (17, 18) with an additional step to merge redundant mixture components to compensate for fitting instabilities (*SI Appendix*, *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 "aligned" 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. 1*C*). Alignment was performed by finding the "closest" reference mixture component in gene-feature space (*SI Appendix*, *Alignment of Models*). Mathematically, to define closeness, we used the Jeffreys divergence, a statistical metric of similarity on probability distributions. We chose the Jeffreys divergence over other metrics because it is symmetric while also having a convenient parametric form (*SI Appendix*, *Model Interpretation through Parameter Analysis*).

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

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

where the minimization is performed over each $\{\phi_i^{\text{ref}}(\boldsymbol{c})\}$ in the set of reference mixture components, and D_{JD} is the Jeffreys divergence (19). Intuitively, for each test mixture component, we find the reference mixture component 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 (*SI Appendix*, *Scoring Alignments*).

Alignments identify subpopulations with maximal transcriptional-state similarity across samples. Since transcriptional-state similarity can arise even in the absence of a direct lineage or identity relationship, alignments do not guarantee cell-type identity, but, rather, highlight predicted relationships that should be interpreted in the context of prior knowledge and can be explored through further downstream analysis.

The directionality of alignment can impact the results and highlight different classes of phenomena. For instance, suppose a cell state in the reference population splits into two progeny branches in a test sample. If we align to the reference, both branches in the test sample will align to the same reference cell state. However, reversing the directionality will give only one alignment—the original cell state will align only with its closest progeny. Both procedures are useful because they address different questions; the first allows us to identify all cell states that align to a particular cell state, and the second allows us to identify only the most similar cell state. For this reason, we offer the option to align in both directions, from test samples to reference, and the reverse.

Additionally, we can perform alignments two-way, which is useful for identifying these branching events or even missing populations. For instance, if a cell state is present in the test sample, but not in the reference sample, the alignment results will change depending on the directionality. An alignment will be found going one way, but not the other way. In this case, we run the alignment procedure in both directions and only retain alignments in which the aligned pair are each other's best match. Alignments which are only found one way are flagged to indicate a missing population or branching event. Thus, two-way alignments are useful for providing a stringent assessment of similar cell states across samples and also for flagging potentially interesting orphaned populations and branching events. These settings offer a suite of different approaches that have their uses in different contexts (*SI Appendix*, *Directionality of Alignment*).

Tracking Cell-State Shifts through Mixture Model Parameters. Following mixture alignment, we analyzed 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. 1*D*). Mathematically, for each aligned mixture pair, $(\boldsymbol{\phi}_i^{\text{test}}, \boldsymbol{\phi}_j^{\text{ref}})$ 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:

$$\Delta\boldsymbol{\mu}_i = \|\boldsymbol{\mu}_i^{\text{ref}} - \boldsymbol{\mu}_j^{\text{test}}\|_2, \quad [3]$$

$$\Delta\boldsymbol{\Sigma}_i = D_{\text{C}}(\boldsymbol{\Sigma}_i^{\text{ref}}, \boldsymbol{\Sigma}_j^{\text{test}}), \quad [4]$$

$$\Delta w_i = |w_i^{\text{ref}} - w_j^{\text{test}}|, \quad [5]$$

where $\Delta\boldsymbol{\mu}_i$ measures shifts in mean gene expression; $\Delta\boldsymbol{\Sigma}_i$ quantifies shifts in the shape of each mixture component, including rotations and changes in gene-expression variance using the Forstner metric, D_{C}, on aligned mixture component covariance matrices (*SI Appendix*, *Model Interpretation through Parameter Analysis*) (20); and Δw_i quantifies shifts in cell-state abundance. We calculated 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) (14, 21). 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 Appendix*, Fig. S9). Accuracy of the models can be assessed by comparing the synthetic (model-generated) data to raw experimental data held out from model training (Fig. 2). PopAlign models generate synthetic data that replicates the geometric structures and statistical variations found in the tissue data in tSNE or two-dimensional (2D) feature projections with quantitative error of $\sim 18\%$ (Figs. 2 and 3 *A* and *B* and *SI Appendix*, *Analysis of Model Error*). The quantitative error across 2D projections is roughly equal to the error observed at 50% subsampling (*SI Appendix*, Fig. S8).

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. 3 *C* and *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. 3 *C* and *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 (*SI Appendix*, Fig. S9), 70% of the mixture components classified for a single-cell type provided by Tabula Muris (*SI Appendix*, Fig. S10).

Through alignment of model components across tissues, PopAlign enables high-level comparisons of tissue composition. By aligning mammary gland to limb muscle (Fig. 3*E*), 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 Appendix*, Fig. S4), and also revealed tissue scale differences in relative abundance. T cells are highly prevalent ($w = 0.3$ in the mammary gland, but rare in the limb muscle $w = 0.05$) (Fig. 3*G*); endothelial cells are highly abundant in the limb muscle ($w = 0.32$), but rare in the mammary gland ($w = 0.06$) (Fig. 3*G*). Between shared cell types, such as macrophages, we reveal common programs such as ER-Trafficking and Metal Ion Sequestration, as well as tissue-specific gene-expression programs found specifically in limb-muscle macrophages (Fig. 3*H*). PopAlign can, thus, give insight into the underlying composition of a tissue, shedding

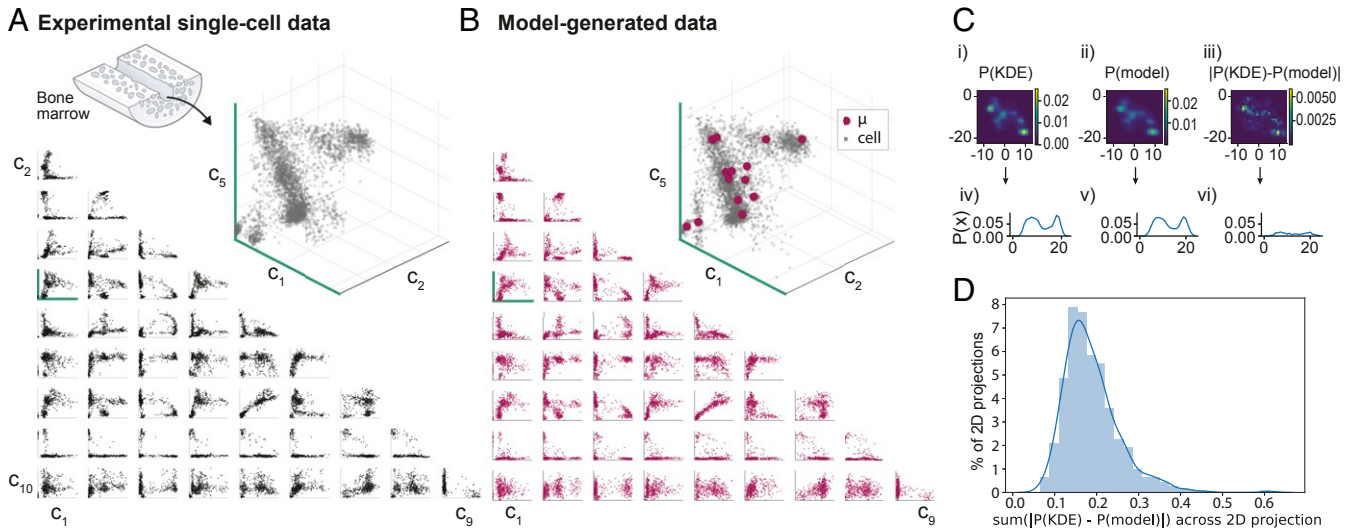

Fig. 2. PopAlign models represent experimental data with high qualitative and quantitative accuracy. (*A*) Experimental data for 3,267 bone marrow cells projected into an $m = 10$-dimensional oNMF feature space. The 2D plots show single cells projected along oNMF feature pairs (c_i, c_j), and a single selected three-dimensional (3D) projection (*A*, *Inset*) is shown. The blue axis denotes a shared axis between 2D and 3D plots. (*B*) Model-generated data for the same 2D and 3D feature-space projections shown in *A*. (*B*, *Inset*) In the 3D projection, each maroon circle denotes the centroid (μ) of a Gaussian mixture component. In each projection, the model-generated data replicate the qualitative geometric structures in the experimental data. (*C*) Random 2D and one-dimensional (1D) projection plots for quantifying error between a kernel-density estimate (KDE) of the data (*i* and *iv*) to the model (*ii* and *v*). The L1 error between the model and KDE is computed and displayed on the right (*iii* and *vi*). The marginal 1D distributions (*iv*, *v*, and *vi*) are summed along the *x* axis of each 2D projection. The kernel-density estimate is built by using all data points with a bandwidth of 0.75. (*D*) Distribution of summed L1 errors across 2D projections for 500 random projections. Mean error is 0.188 ± 0.066.

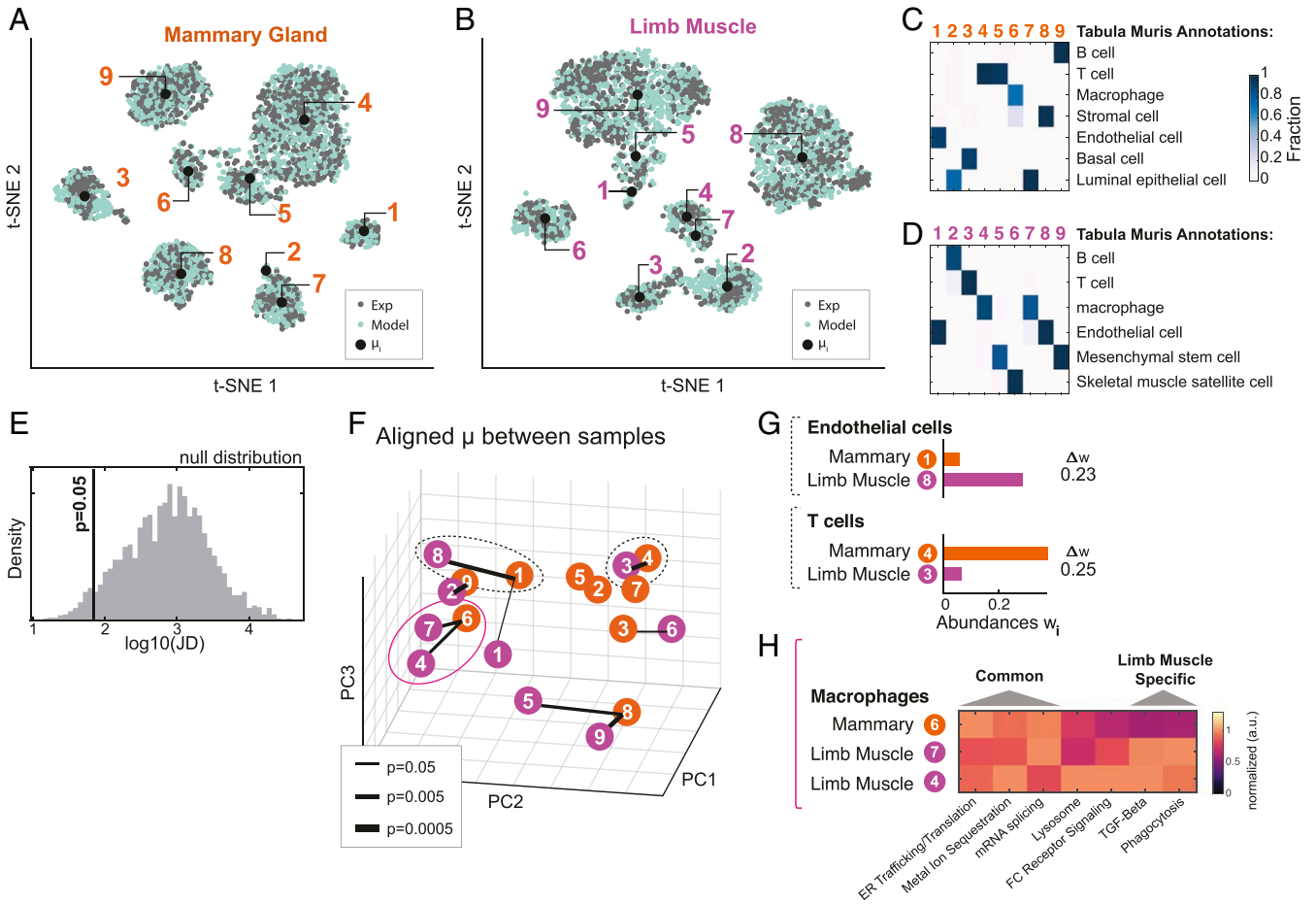

Fig. 3. Probabilistic models identify, align, and dissect cellular subpopulations across disparate tissues. (*A* and *B*) For two tissues, mammary gland (*A*) and limb muscle (*B*), experimental single-cell data (black) are plotted together with PopAlign model-generated data (teal) using a 2D tSNE transformation. For each tissue, mixture model centroids (μ) are indicated by a numbered black dot. (*C* and *D*) Heatmaps for mammary gland (*C*) and limb muscle (*D*) showing the percentage of cells associated with each mixture component that have a specific cell annotation label supplied by Tabula Muris. Columns (but not rows) sum to one. (*E*) Null distribution of Jeffreys divergence (JD) using all possible pairs of mixture components within each model. Threshold for $P < 0.05$ is indicated by a vertical line. (*F*) 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 μ vectors are transformed by using PCA and plotted by using the first three principal components. The width of each line is inversely proportional to the P value associated with the alignment (see key). (*G*) We ranked aligned subpopulations in terms of maximum Δw and show the top two pairs (T cells and endothelial cells) that are highlighted in *F* with a gray dotted line. T cells are highly abundant in mammary gland, while endothelial cells are highly abundant in limb muscle. (*H*) Comparing subpopulation centroids (μ) for macrophages in terms of annotated oNMF features. Macrophages in mammary gland and limb muscle share common features (*Left*), but two features (*Right*) are expressed at higher levels specifically in limb-muscle macrophages. Coefficient values for each feature have been normalized by the maximum value across the column. Corresponding alignments are highlighted in *F* with a red ellipse.

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 min on a typical workstation with eight cores and 64 GB RAM (Fig. 4*A*). By first building models, PopAlign front-loads the computation to produce a low-error (Fig. 2) representation of the data that achieves a 50- to 100-fold reduction in the memory footprint. Memory efficiency speeds up downstream tasks, such as the calculation of pairwise divergences between subpopulations (Fig. 4*B*) necessary for aligning them across samples.

Applying PopAlign to compare all 12 tissues of Tabula Muris shows that 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 Jeffreys divergence relative to reference subpopulations), that can be accepted or rejected using 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 (22, 23). For example, canonical correlation analysis (CCA) finds a global linear transformation of the data that minimizes transcriptional differences between samples. Not only are many batch-correction methods computationally expensive (Fig. 4*A*), they can also require overlapping subpopulations for interpretation, 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.

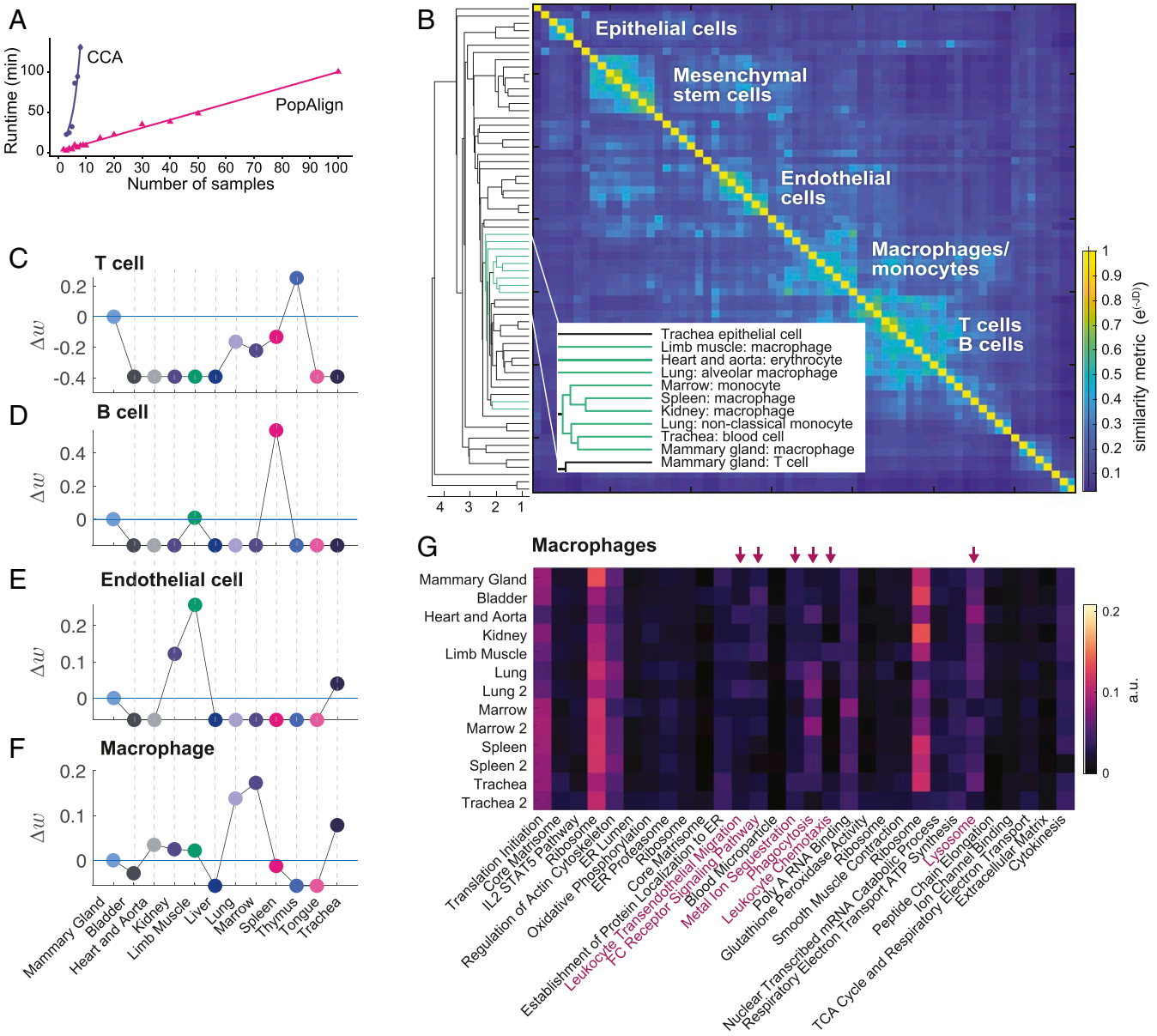

Fig. 4. PopAlign can perform global comparisons of cell states across dozens to hundreds of experimental samples. (*A*) Computational runtime versus number of samples for PopAlign (blue) vs. CCA-based alignment method (red). PopAlign scales linearly with the number of samples, while CCA scales with polynomial time 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 were performed on typical workstations (eight cores, 64 GB RAM). (*B*) Heatmap of a pairwise similarity metric between subpopulations from all 12 tissues demonstrates that 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 Jeffreys divergence between two subpopulations. *B*, *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 T cells (*C*), B cells (*D*), endothelial cells (*E*), and macrophages (*F*). (*G*) Mean gene-expression state (μ) for macrophages across all tissues show variation in features associated with key immune pathways (highlighted with red font and arrows).

For example, we found that T cells (Fig. 4*C*) and B cells (Fig. 4*D*) are most abundant in organs where they are known to mature developmentally [the thymus (24) and spleen (25), respectively], endothelial cells (Fig. 4*E*) are most prevalent in highly vascularized tissues (kidney and limb muscle), and macrophages (Fig. 4*F*) 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. 4*C*). We also found distinct patterns of gene-program activation (e.g., lung macrophages are highly phagocytic) in macrophage populations across tissues (Fig. 4*G*), consistent with previous reports of functional diversity among macrophages (26). 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, 27–30). Being able to analyze 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.

We performed our screen using 40 drugs (Fig. 5*A*) from a commercially available compound library (Selleck Chem) on peripheral blood mononuclear cells (PBMCs) from a healthy 22-y-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. 5*B*).

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 (six independent replicates). Statistically, we could define "hits" as drugs which have a negative log-likelihood ratio metric (*SI Appendix*, *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. 5*B*), as well as mTOR inhibitors (pink) and 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 these shifts across specific cell types, we found that 26 drugs exerted significant gene-expression shifts (Δw) on monocytes (Fig 5*C*), while 14 drugs exerted significant effects on T cells (false discovery rate [FDR]-corrected P values < 0.05) (Fig 5*D*). Of these drugs, eight drugs (highlighted in color) impacted both cell types (Fig. 5 *C* and *D*). Most drugs either did not affect abundances ($\Delta w \approx 0$) or increased monocyte abundance up to 5% (*SI Appendix*, Fig. S11 *A* and *B*).

The ability to find the transcriptional impacts of genes that are universal across cell types can reveal important insights into a drug's fundamental mechanisms. In our screen, we discovered that, although drug-responsive genes were mostly cell-type-specific (Fig 5*E*), for some drugs, up to 15% of impacted genes were shared between cell types (*SI Appendix*, Supplementary File 1, which supplies differentially expressed genes for all drugs/cell types). For example, budesonide up-regulated 11 genes and down-regulated 14 genes in both T cells and monocytes (Fig 5*F*). The overlapping down-regulated genes included many genes associated with actin-based motility—such as actin genes (beta-actin [ACTB] and ACTG1), an antiadhesion peptide (CD52), a myosin-interacting protein (CD74) (31), and an actin-sequestering protein (TSMB10) (32). This result is consistent with earlier observations that glucocorticoids impede T cell polarization and motility (33) and monocyte migratory behavior (34) and suggests 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 $\Delta\boldsymbol{\mu} = 2.64$, $P = 2.54\text{e-}5$; Fig. 5*G*), but no changes in monocytes (mean $\Delta\boldsymbol{\mu} = 0.29$, $P = 1$; Fig 5*H*). Dexrazoxane did not generate any differentially expressed genes in monocytes (Fig 5*E*). We found that in T cells, dexrazoxane up-regulated many cell-survival genes, including antioxidant enzymes (GPX4 and PRDX1) and CORO1A, which is essential for T cell survival (35) (Fig 5*I*). Dexrazoxane is normally used as a chemoprotectant agent to reduce toxic side effects of chemotherapy on cardiac tissue (36). Our finding that dexrazoxane specifically impacts T cells by up-regulating genes that reduce oxidative stress could potentially be useful in modulating T cell behavior for other diseases.

PopAlign 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 tSNE (*SI Appendix*, Fig. S12) or uniform manifold approximation and projection (UMAP) (37), which show qualitative changes (see highlighted conditions), but are not quantifiable due to the nonlinear embedding. 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 up-regulating prosurvival 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 MM. 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 MM. MM 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 (38–41), further contributing to disease progression and treatment relapse. In MM patients, immune cells with disrupted phenotypes can be detected in the peripheral blood (40, 42, 43). 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 MM patients undergoing various stages of treatment (*SI Appendix*, Table S1). We profiled $> 5,000$ cells from each patient and constructed and aligned probabilistic models to one reference healthy population (Fig. 6*A*).

PopAlign identified several common global signatures in the MM samples at the level of cell-type abundance and gene expression. Across all samples, we found previously known signatures of MM, including a deficiency in B cells (40, 44, 45), an expansion of monocyte/myeloid derived cells (42), and, critically, new impairments in T cell functions.

Plotting Δw across all patients, we found high-level changes in subpopulation abundances, which are known to be prognostic of disease progression (43). We found that all MM patients experienced a contraction in B cell numbers (Fig. 6*B*), and two out of four saw a dramatic expansion ($\Delta w \gg 0.2$) of monocytes (Fig. 6*C*). Changes in T cell levels, however, can be highly variable. Most patients saw a reduction in effector T cells (Fig. 6*D*) and no change in resting T cells (Fig. 6*E*). However, outlier patient MM4 had a large increase in effector T cells (Fig. 6*D*; $\Delta w = 0.2$) and a complete elimination of resting T cells (Fig. 6*E*; $\Delta w = 0.2$). For this patient, who was receiving a thalidomide-derived drug therapy, these deviations are consistent with thalidomide's known stimulatory effects on T cells (46).

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 used 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 (47), which are negative regulators of immune function associated with cancer.

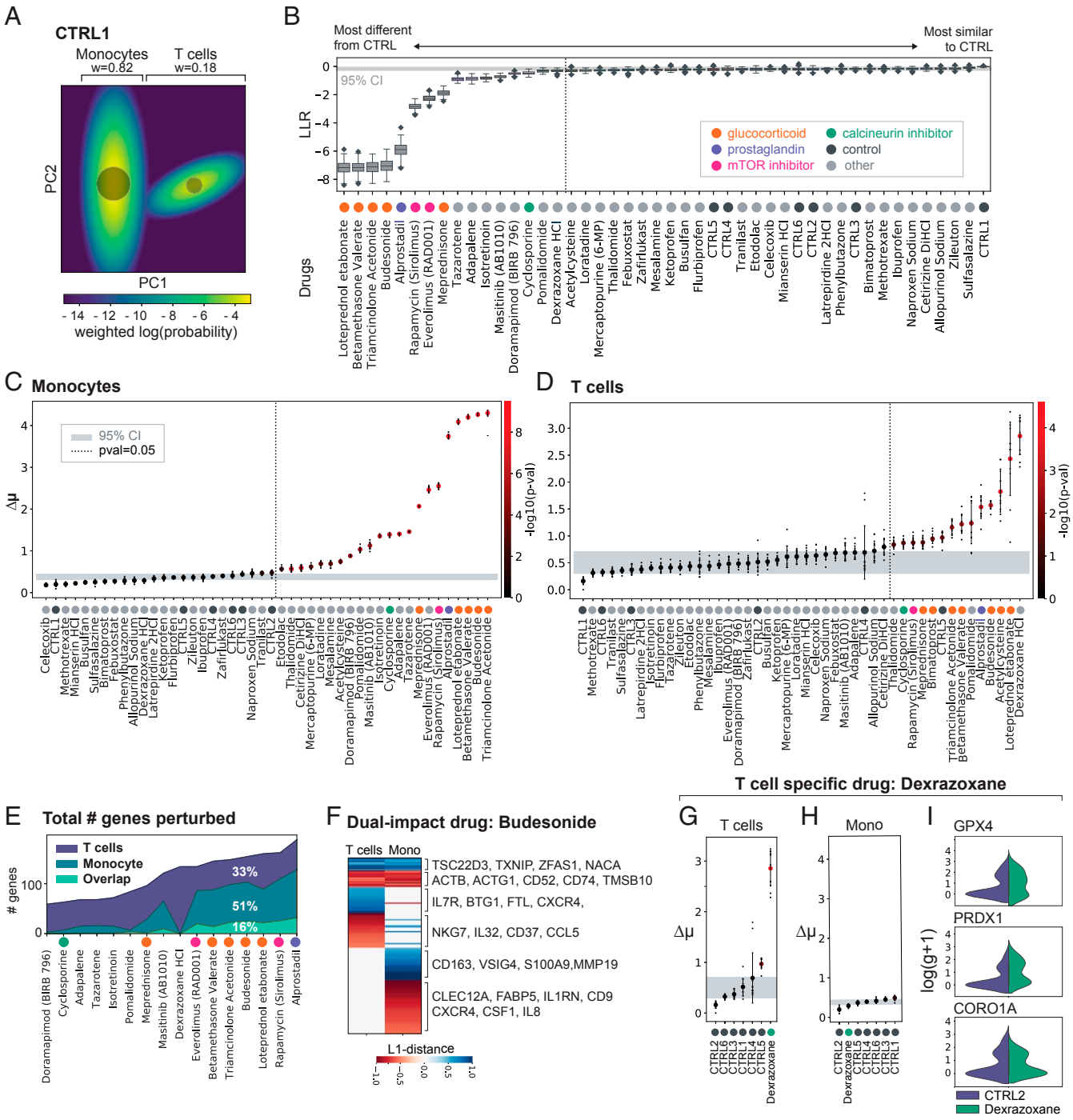

Fig. 5. PopAlign identifies universal and cell-type-specific impacts of immunomodulatory drugs. (*A*) Rendering of Gaussian mixture model for CTRL1 projected onto the first two principal components. Abundance weights (w) are represented by the size of the circle. (*B*) Drug ranking based on population-level similarity to control population using the log-likelihood ratio metric (LLR). FDR-corrected *P* values using one-sample *t*-test of the six control replicates against each drug are indicated. Dashed line, $P = 0.05$. (*C* and *D*) Gene-expression shifts ($\Delta\mu$) for drug-exposed monocyte (*C*) and T cell (*D*) subpopulations with respect to their aligned subpopulation in CTRL1. 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). FDR-corrected *P* values using one-sample *t* test of the six control replicates against each drug are indicated. Gray box: The 95% CI of the control mean. Dashed line: $P = 0.05$. (*E*) Number of T cell-specific, monocyte-specific, and overlapping genes that are perturbed in response to top-ranking drugs. (*F*) Per-gene L1-distance metrics are shown for budesonide, a dual-impact drug. Differentially expressed genes that impact both T cells and monocytes are shown at the top. (*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.

By plotting the monocyte-specific mean gene-expression values for both CD11b and CD33, we saw that all patients except patient MM3 scored highly for both myeloid-derived suppressor cell (MDSC) markers (Fig. 6*F*). Patients with high MDSC populations typically have a poor prognosis, underscoring the need to monitor MDSC populations in patients.

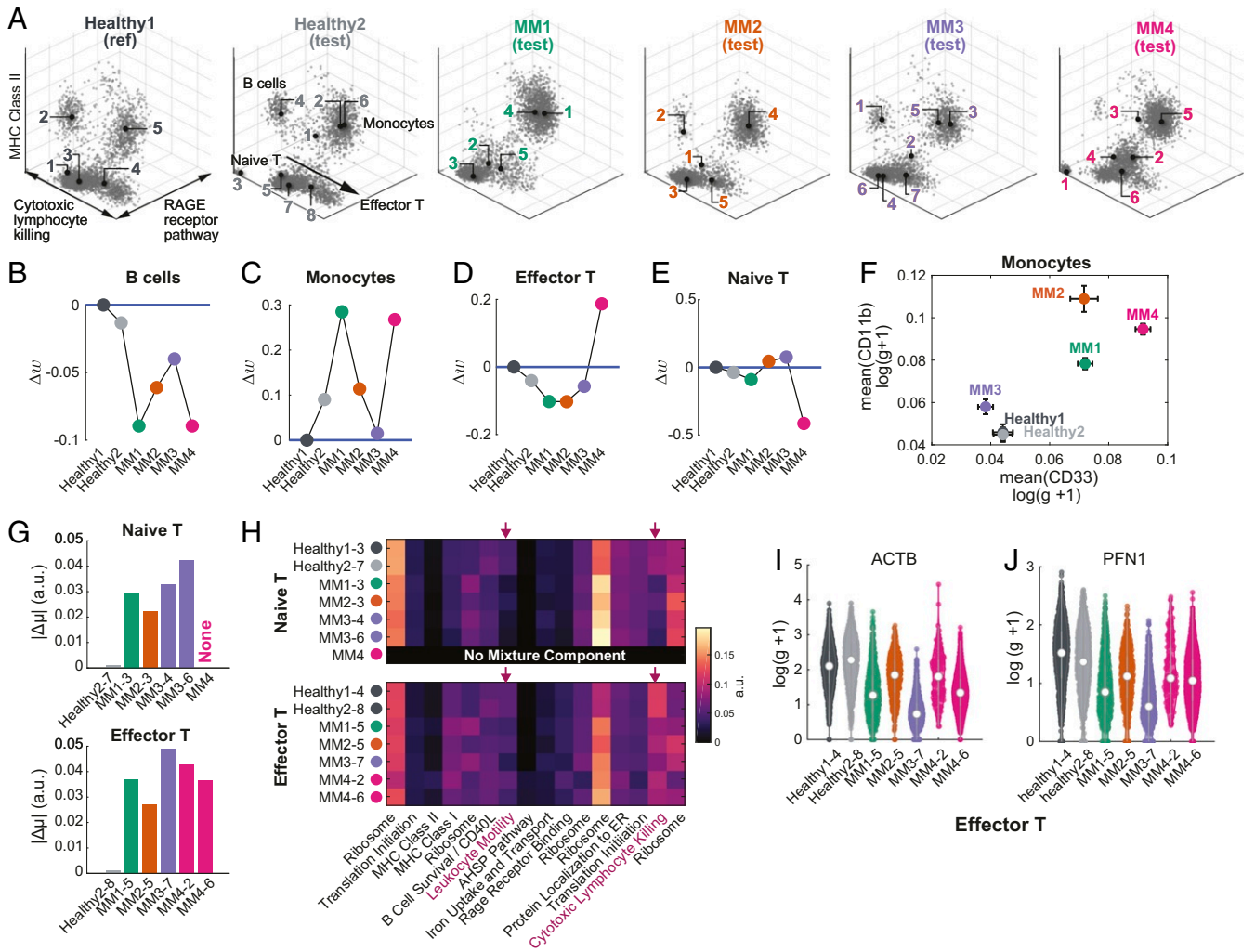

Fig. 6. Discovering signatures of disease and treatment in PBMCs from MM patients. (*A*) Experimental single-cell mRNA-seq data from two healthy donors and four MM patients (MM1 to 4-) are projected into 16-dimensional gene-feature space. The 3D plots show single cells in a subset of three gene features that highlight separation between different immune-cell types. Mixture model centroids (μ) are indicated by a numbered black dot. (*B–E*) Subpopulations in test samples are aligned to the reference (healthy1), and changes in abundance (Δw) are plotted for B cells (*B*), monocytes (*C*), effector T cells (*D*), and naive T cells (*E*). (*F*) Mean gene-expression levels for two markers of MDSCs, CD33 and CD11b, are plotted for all monocyte subpopulations. Error bars denote CI of the mean. (*G*) $|\Delta\mu|$ for naive T cell and effector T cell populations relative to healthy1. A.u., arbitrary units. (*H*) 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. (*I*) Distribution of ACTB expression for all effector T cell subpopulations across samples. Violin shows distribution, and mean is denoted by white circle. (*J*) Distribution of PFN1 expression for all effector T cell subpopulations across samples. For single gene plots—*F*, *I*, and *J*—units are in terms of normalized and log-transformed gene expression ($\log(\hat{g}+1)$).

Importantly, we also found that naive and effector T cells across all MM patients had 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. 6*G*). By examining the $\mu's$ in terms of gene-expression features (Fig. 6*H*), we found that in MM, 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 ACTB (Fig. 6*I*), a core subunit of the actin cytoskeleton, 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. 6*J*).

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 provides 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 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 pinpoint 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 screens 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 workbench for single-cell analysis and treatment of human disease. By applying PopAlign to datasets 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 MM 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. Discussion of the mathematical framework, including data normalization, selection of feature number, analysis of model error, alignment of models, and model interpretation through parameter analysis, is provided in *SI Appendix*.

Single-Cell RNA-Seq for MM and Healthy Donor Samples. All human cell samples cryopreserved PBMCs were thawed in warm Roswell Park Memorial Institute Medium (RPMI-1640) at 37 °C and pelleted at 300 × *g* for 2 min. Cells were resuspended to 1e6 cells/mL in RPMI-1640. For each sample, 17,400 cells were loaded into a 10X Genomics lane using single-cell 3′ v2 reagents.

Multiplexed Single-Cell RNA-Seq Using Multiseq. Cryopreserved PBMCs sourced from Hemacare were thawed and rested in RPMI-1640 in CO_2 incubator at 37 °C for 16 h before drug exposure. After resting, 200,000 cells were seeded into each well of a 96-well plate and exposed to 40 drugs selected from the immunology- and inflammation-related small-molecule compound library sold by SelleckChem. Drugs were used at 1 μM concentrations in RPMI-1640 plus 10% fetal bovine serum. After 24 h of exposure, cells were dissociated into a single-cell suspension by using TrypLE and multiplexed by using Multiseq lipid-modified oligos (11) before running on two 10X Genomics lanes using single-cell 3′ v3 reagents.

Study Approval. All studies were performed on PBMCs obtained commercially from Hemacare. The California Institute of Technology Institutional Review board (IRB) has determined that this work is exempt from the requirement for IRB review and approval (Reference #17-0727), and informed consent was not required.

Data Availability. Single-cell gene-expression data have been deposited in Figshare (https://doi.org/10.6084/m9.figshare.11837097) (48). The software package, implemented in Python 3, can be found at GitHub, https://github.com/thomsonlab/popalign.

ACKNOWLEDGMENTS. We thank Justin Bois, Eric Chow, Allan-Hermann Pool, Jase Gehring, Tami Khazaei, and members of the M.T. laboratory for helpful feedback and discussions; Chris McGinnis and David Patterson for experimental guidance; and Inna-Marie Strazhnik for figure editing and illustrations. This work was performed at the Beckman Institute Single-Cell Profiling and Engineering Center. M.T. was supported by the Shurl and Kay Curci Foundation and the Heritage Medical Research Institute.

1. K. J. Pienta, N. McGregor, R. Axelrod, D. E. Axelrod, Ecological therapy for cancer: Defining tumors using an ecosystem paradigm suggests new opportunities for novel cancer treatments. *Transl. Oncol.* **1**, 158–164 (2008).
2. M. J. Stubbington, O. Rozenblatt-Rosen, A. Regev, S. A. Teichmann, Single-cell transcriptomics to explore the immune system in health and disease. *Science* **358**, 58–63 (2017).
3. S. J. Horning, A new cancer ecosystem. *Science* **355**, 1103 (2017).
4. G. X. Y. Zheng *et al.*, Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
5. E. Z. Macosko *et al.*, Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
6. G. Heimberg, R. Bhatnagar, H. El-Samad, M. Thomson, Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Syst.* **2**, 239–250 (2016).
7. S. Dasgupta, "Learning mixtures of Gaussians" in *40th Annual Symposium on Foundations of Computer Science* (IEEE, Piscataway, NJ, 1999), pp. 634–644.
8. E. Candes, X. Li, Y. Ma, J. Wright, "Robust principal component analysis?: Recovering low-rank matrices from sparse errors" in *2010 IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)* (IEEE, Piscataway, NJ, 2010), pp. 201–204.
9. H. Mathys *et al.*, Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).
10. P. van Galen *et al.*, Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity. *Cell* **176**, 1265–1281.e24 (2019).
11. C. S. McGinnis *et al.*, MULTI-seq: Scalable sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. bioRxiv https://doi.org/10.1101/387241 (8 August 2018).
12. J. Gehring, J. Hwee Park, S. Chen, M. Thomson, L. Pachter, Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins. *Nat. Biotechnol.* **38**, 35–38 (2020).
13. H. M. Kang *et al.*, Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
14. S. R. Quake *et al.*, Single-cell transcriptomic characterization of 20 organs and tissues from individual mice creates a Tabula Muris. bioRxiv https://doi.org/10.1101/237446 (29 March 2018).
15. C. Ding, L. Tao, P. Wei, H. Park. "Orthogonal nonnegative matrix t-factorizations for clustering" in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 126–135 (2006).
16. D. Lee, H. Seung, Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
17. A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, 1–38 (1977).
18. D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge, UK, 2003).
19. H. Jeffreys, *The Theory of Probability* (Oxford University Press, Oxford, UK, 1998).
20. W. Förstner, B. Moonen, "A metric for covariance matrices" in *Geodesy—The Challenge of the 3rd Millennium*, E. W. Grafarend, F. W. Krumm, V. S. Schwarze, Eds. (Springer, Berlin, Germany, 2003), pp. 299–309.
21. O. Botvinnik, J. Webber, J. Batson, A. Pisco, Data from "Single-cell RNA-seq data from microfluidic emulsion (v2)." Figshare. Dataset. https://doi.org/10.6084/m9.figshare.5968960.v3 . Accessed 1 May 2019.
22. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
23. L. Haghverdi, A. T. Lun, M. D. Morgan, J. C. Marioni, Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
24. H. Takaba, H. Takayanagi, The mechanisms of T cell selection in the thymus. *Trends Immunol.* **38**, P805–P816 (2017).
25. A. Brown, Immunological functions of splenic B-lymphocytes. *Crit. Rev. Immunol.* **11**, 395–417 (1992).
26. E. L. Gautier *et al.*, Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages. *Nat. Immunol.* **13**, 1118–1128 (2012).
27. K. Murphy, C. Weaver, *Janeway's Immunobiology* (CRC Press, Boca Raton, FL, ed. 9, 2016).
28. W. A. Lim, C. H. June, The principles of engineering immune cells to treat cancer. *Cell* **168**, 724–740 (2017).

29. K. Newton, V. M. Dixit, Signaling in innate immunity and inflammation. *Cold Spring Harb. Perspect. Biol.* **4**, a006049 (2012).
30. A. C. Villani, S. Sarkizova, N. Hacohen, Systems immunology: Learning the rules of the immune system. *Annu. Rev. Immunol.* **36**, 813–842 (2018).
31. G. Faure-André *et al.*, Regulation of dendritic cell migration by CD74, the MHC class II-associated invariant chain. *Science* **322**, 1705–1710 (2008).
32. X. Zhang *et al.*, Thymosin beta 10 is a key regulator of tumorigenesis and metastasis and a novel serum marker in breast cancer. *Breast Cancer Res.* **19**, 15 (2017).
33. N. Müller, H. J. Fischer, D. Tischner, J. van den Brandt, H. M. Reichardt, Glucocorticoids induce effector T cell depolarization via ERM proteins, thereby impeding migration and APC conjugation. *J. Immunol.* **190**, 4360–4370 (2013).
34. J. Ehrchen *et al.*, Glucocorticoids induce differentiation of a specifically activated, anti-inflammatory subtype of human monocytes. *Blood* **109**, 1265–1274 (2007).
35. P. Mueller *et al.*, 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.* **9**, 424–431 (2008).
36. S. E. Lipshultz *et al.*, The effect of dexrazoxane on myocardial injury in doxorubicin-treated children with acute lymphoblastic leukemia. *N. Engl. J. Med.* **351**, 145–153 (2004).
37. L. McInnes, J. Healy, UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426 (9 February 2018).
38. S. K. Kumar *et al.*, Multiple myeloma. *Nat. Rev. Dis. Primers* **3**, 17046 (2017).
39. E. Malek *et al.*, Myeloid-derived suppressor cells: The green light for myeloma immune escape. *Blood Rev.* **30**, 341–348 (2016).
40. L. M. Pilarski, E. Joy Andrews, M. J. Mant, B. A. Ruether, Humoral immune deficiency in multiple myeloma patients due to compromised B-cell function. *J. Clin. Immunol.* **6**, 491–501 (1986).
41. M. Bolzoni *et al.*, *IL21R* expressing $CD14^{+}CD16^{+}$ monocytes expand in multiple myeloma patients leading to increased osteoclasts. *Haematologica* **102**, 773–784 (2017).
42. C. Botta, A. Gullà, P. Correale, P. Tagliaferri, P. Tassone, Myeloid-derived suppressor cells in multiple myeloma: Pre-clinical research and translational opportunities. *Front. Oncol.* **4**, 348 (2014).
43. T. Dosani *et al.*, Significance of the absolute lymphocyte/monocyte ratio as a prognostic immune biomarker in newly diagnosed multiple myeloma. *Blood Canc. J.* **7**, e579 (2017).
44. A. C. Rawstron *et al.*, B-lymphocyte suppression in multiple myeloma is a reversible phenomenon specific to normal B-cell progenitors and plasma cell precursors. *Br. J. Haematol.* **100**, 176–183 (1998).
45. R. J. Pessoa-Magalhaes *et al.*, Analysis of the immune system of multiple myeloma patients achieving long-term disease control, by multidimensional flow cytometry. *Haematologica* **98**, 79–86 (2013).
46. M. Winqvist *et al.*, 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* **126**, 4164 (2015).
47. V. Bronte *et al.*, Recommendations for myeloid-derived suppressor cell nomenclature and characterization standards. *Nat. Commun.* **7**, 12150 (2016).
48. S. Chen, PopAlign_Data. Figshare. Dataset. https://doi.org/10.6084/m9.figshare.11837097.v3. Deposited 9 November 2020.