Take ACTION to characterize the functional identity of single cells

Shahin Mohammadi[1], Vikram Ravindra[1], David F. Gleich[1] & Ananth Grama[1]

[1]*Department of Computer Sciences, Purdue University, West Lafayette, IN 47904, USA*

Single-cell transcriptomic data has the potential to radically redefine our view of cell type identity. Cells that were previously believed to be homogeneous are now clearly distinguishable in terms of their expression phenotype. Methods for automatically characterizing the functional identity of cells, and their associated properties, can be used to uncover processes involved in lineage differentiation as well as sub-typing cancer cells. They can also be used to suggest personalized therapies based on molecular signatures associated with pathology. We develop a new method, called *ACTION*, to infer the functional identity of cells from their transcriptional profile, classify them based on their principal functions, and reconstruct regulatory networks that are responsible for mediating their identity. Results from using *ACTION* to sub-type cancer cells in Melanoma patients reveal novel biomarkers along with their underlying regulatory networks.

Complex tissues typically consist of heterogeneous populations of interacting cells that are specialized to perform different functions. The functional identity of a cell reflects the degree to which it performs each function. This can be characterized by viewing a cell as a point within a space that consists of all possible functions, which we call the *functional space* of cells. Recent advances in single cell technologies have greatly expanded our view of the functional identity of cells. Cells that were previously believed to constitute a homogeneous group are now recognized as an ecosystem of cell types [1]. Within the tumor microenvironment, for example, the exact composition of these cells, as well as their molecular makeup, have a significant impact on diagnosis, prognosis, and treatment of cancer patients [2].

The functional identity of each cell is closely associated with its underlying type [3] and a number of methods have been proposed to directly identify cell types from the transcriptional profiles of single cells [4–9]. A majority of these methods rely on classic measures of distance between transcriptional profiles to establish cell types and their relationships. However, these measures fail to capture weakly expressed but highly cell type-specific genes [10]. They often require user-specified parameters, such as the underlying number of cell types, which critically determine their performance. Finally, once the identity of a cell has been established using these methods, it is often unclear what distinguishes one cell type from others.

To address these issues, we propose a new method, called *Archetypal-analysis for cell type identification (ACTION)*, for identifying cell types, characterizing their functional identity, and uncovering underlying regulatory factors from single-cell expression datasets. At the core of our method is a biologically-inspired metric that captures the similarity of cells. This metric accounts for the specificity of marker genes and defines a signature for each cell that is robust to noise. At the same time, it is sensitive to weak cell type-specific signals. This metric provides a view on the functional space, and we use it to identify a small set of *principal functions* that are performed by the population of cells. A combination of these principal functions then approximates the functional identity of each cell. Finally, we develop a statistical framework to identify key marker genes for each cell type, as well as transcription factors that are responsible for mediating the observed expression of these markers. We use these regulatory elements to construct cell type-specific transcriptional regulatory networks.

We show that the *ACTION* metric effectively represents known functional relationships between cells. Furthermore, if we use the dominant principal function of each cell to estimate its putative cell type, *ACTION* outperforms state-of-the-art methods for identifying cell types. Since we establish its superior performance in identifying existing cell types, we report on a case study of cells collected from a tumor microenvironment

of 19 melanoma patients [11]. We identify two novel subclasses of *MITF*-high patients, one of which exhibits significantly worse prognosis. We construct the transcriptional regulatory network associated with this subclass and pinpoint regulatory factors that mediate its function. These factors provide novel biomarkers, as well as potential therapeutic targets for future development.

Results

The *ACTION* framework consists of three major components, shown in Figure 1: (i) a robust, yet sensitive measure of cell-to-cell similarity, (ii) a geometric approach for identification of principal functions, and (iii) a statistical framework for constructing cell type-specific transcriptional regulatory networks (TRNs). Our cell-to-cell similarity metric is rooted in the notion that the transcriptional profile of a cell is dominated by housekeeping genes, whereas its functional identity is determined by a combination of weak but preferentially expressed genes. To define a robust yet sensitive metric, we suppress shared housekeeping expression and enhance cell type-specific signal. The next component of our method is a geometric approach for identifying principal functions of cells. Each of these principal functions is represented by a corner of the convex hull of points corresponding to the functional space of cells. We refer to these corners as *archetypes*. Finally, *ACTION* uses a novel method to orthogonalize archetypes and assess the significance of each transcriptional factor in mediating the transcriptional phenotype associated with each archetype. This method is then used to construct the characteristic transcriptional regulatory network (TRN) of each cell type. In what follows, we describe, validate, and discuss each component in detail.

The *ACTION* metric outperforms other metrics in representing functional relationships between single cells

A fundamental component of many methods for identifying cell types is a measure to quantify the similarity between individual cells. Most prior methods rely on traditional measures, such as linear correlation that are not specifically targeted towards transcriptomic profiles. In contrast, we define a similarity metric, or formally a kernel, specifically designed for measuring the similarity between single cells [10]. We first project raw transcriptional profiles of cells to the orthogonal subspace spanned by housekeeping genes. This removes the dominant non-discriminatory housekeeping contribution from the transcriptional profiles. We then boost the contribution of cell type-specific genes using an information theoretic approach. Finally, we combine the adjusted transcriptional profile of cells with the estimated expression-specificity vector of genes to define a robust measure of cell-to-cell similarity. Our approach is illustrated in Figure 2a and the mathematical models underlying the metric are described in the Methods section, Component 1.

To establish the superiority of our metric, we compare it against an alternate measure specifically designed for single cell analysis, *SIMLR* [12]. SIMLR combines a number of distance metrics to learn a joint similarity score that maximizes the block diagonal structure of the resulting matrix. In addition, we also compare *ACTION* with the normalized dot product defined over the reduced subspace constructed by two nonlinear dimension-reduction techniques: *multidimensional scaling (MDS)* and *Isomap*. While *ACTION* is a non-parametric method, the other methods have one or more parameters that need to be specified by the user. For SIMLR, we need to specify the true number of cell types. For all methods besides *ACTION*, we must specify the dimension of the low-dimensional subspace. To give them the best chance at competing with *ACTION*, we evaluate ten different values for the dimension of projected subspace (from 5 to 50 with

increments of 5) and report the best results obtained over all configurations.

To assess the quality of computed similarities between cells, we pair each metric with the kernel k-means algorithm in order to assess their ability to identify discrete cell types. We apply this technique to four different datasets (see Methods, Datasets). These datasets arise from a variety of different single cell technologies, have hundreds to thousands of cells, and span a wide range of normal and cancerous cells. We compare the computed cell types against the annotated cell types in the original dataset using two scores which quantify complementary aspects of clustering (*Normalized Mutual Information, NMI*, and *Adjusted Rand Index, ARI*). Both of these scores are close to 0 for a random assignment and approach 1 when the computed clustering exactly matches the annotations. In each case, we perform 100 independent trials with random initialization for k-means and report the mean of NMI and ARI scores.

Figures 2b-c present the performance of the cell type identification technique when operating with different similarity measures. In summary, our results demonstrate that in all cases the *ACTION* metric either outperforms or is jointly the best among competing metrics. For the *Immune* dataset, for instance, there is a tie between *ACTION, MDS,* and *SIMLR* with respect to best *NMI* score, whereas the tie is only between *ACTION* and *MDS* for best ARI scores. Another tie occurs on the *Melanoma* dataset with *SIMLR* for NMI. In all other cases, the *ACTION* metric significantly outperforms the other methods (*t*-test; *p*-val ≤ 0.05, computed within the population of random initializations). These results establish the *ACTION* metric as a *fast*, *nonparametric*, and *accurate* method for computing similarity among single cells. We use this measure throughout the rest of our study.

The *ACTION* method successfully uncovers functional identity of single cells

Using the *ACTION* metric as a measure of similarity between cells, we develop a new method for characterizing the *functional identity* of cells in a given experiment. Our method is based on a geometric interpretation of cellular functions. In this view, each cell corresponds to a data-point in a high-dimensional space. Our method identifies *"extreme"* corners, or *archetypes* in this space, each of which represents a *principal function*. The functional identity of each cell is subsequently characterized as a convex combination of these principal functions. The choice of the number of principal functions or archetypes is based on a novel non-parametric statistical procedure.

Discrete view of cell types Cells can be classified into discrete types using their dominant principal function. To validate cell types identified using *ACTION*, we compare our method to four recently proposed methods: SCUBA [5], SNNCliq [7], single-cell ParTI [8, 13], and TSCAN [9] (see Supplementary Text 1 for a brief description of these methods) to predict annotated cell types on the same four datasets (see Methods, Component 2). All methods (including ACTION) have no user-configurable parameters – except for *SNNCliq*. We used the default options for *SNNCliq* ($k = 3$ and $r = 0.7$). For the *Melanoma* dataset, *SNNCliq* did not terminate after *72h*, after which we stopped the experiment.

We present a comprehensive analysis of the results for all combinations of datasets and methods in predicting cell types in terms of their NMI and ARI scores (Figure 3). In all cases, the *ACTION* method is observed to perform best in identifying known cell types. For some datasets, such as the *Pollen* dataset, the ACTION cell types have higher quality across all methods. This is attributed to the underlying separation among these cell types. The *SNNCliq* method is the runner-up in two out of four cases for NMI, but only

one out of four when measured via ARI. This happens because this method considerably over-estimates the number of underlying cell types, which is reflected in low ARI values. The *SCUBA* method, on the other hand, is the runner-up in two out of four datasets with respect to ARI, but only for one case in terms of *NMI*. This suggests that, while *ACTION* performs the best in all datasets, there is not a consensus as to which other method performs well across datasets.

Continuous view of cell states While the functional identities of cells can be discretized to define cell types, they can also be explored in the continuous space of all principal functions. To illustrate this *continuous view*, we perform a case study on the *Melanoma* dataset (Figure 4). Each point corresponds to a cell. Given the functional profile of cells, defined in a k-dimensional space with k being the number of archetypes, we map cells to a two-dimensional plane using the Stochastic Neighbor Embedding (SNE) method with a deterministic initialization (see Supplemental Text 4). Our non-parametric method selected 8 archetypes for the Melanoma data, each is marked with a text label (A1, . . . , A8) and assigned a unique color. We interpolate the color of each cell using its distance to all archetypes to highlight the continuous nature of the data. We use markers from *LM22* dataset [14] to distinguish different subtypes of T-cells. For the tumor cells, we perform further analysis of active transcription factors, as described in the next section and the methods section, to identify key driving regulators that distinguish each archetype.

Figure 4 demonstrates the ability of our method to identify both isolated cell-types with specialized principal functions, as well as the ones with a mixed combination of functions. As an example, T-cells constitute a continuous spectrum across functional space of cells, which is consistent with previous studies [15]. Subclasses of melanoma cells, on the other hand, exhibit distinct separation that, to a high degree, is associated with the *functional activity* of the *MITF* gene. Functional activity of a specific transcription factor

is a statistical measure, directly inferred from the transcriptional activity of its target genes, that assesses the importance of that factor in the regulation of an observed transcriptional phenotype (see Methods, Component 3). Subclasses *B* and *C*, represented by archetypes *A4* and *A5*, respectively, are functionally active for MITF but they have distinct phenotypic behaviors and survival rates. In what follows, we construct the transcriptional regulatory network responsible for mediating underlying characteristics of these subclasses.

The *ACTION* method constructs accurate models of the regulatory networks that drive functional identity of cells

We propose a new method to construct regulatory pathways responsible for mediating the phenotypes associated with each archetype. We first perform an *archetype orthogonalization* to compute a residual expression and identify marker genes that are unique to each archetype. We then assess the role of each transcription factor (TF) in controlling these marker genes. Significant TFs, together with their top-ranked target genes (TGs), constitute the underlying transcriptional regulatory network (TRN) that is responsible for mediating a given principal function, and consequently, the phenotype associated with cells dominantly associated with that function (see Methods, Component 3, and Figure 5a for additional details).

To evaluate the effectiveness of our approach, we perform a case study to identify regulatory pathways of MITF-associated tumor cells (those in subclasses *B* and *C* from Figure 4). We set a p-value threshold of 0.05 on our functional activity measure in each case to identify significant transcription factors (TFs), see Supplemental Text 7 for the full list. We first reiterate that both *subclass B* and *subclass C* have *MITF* as a common functionally active transcription factor. In addition, both have *SOX10* as active too. These two factors are canonical markers for melanoma cells in the "proliferative" state [16]. Both of these subclasses

are also highly significant (Fisher's test p-value of *subclass B* = 7.9×10^{-11}, and that for *subclass C* = 9.3×10^{-14}) for known markers of the proliferative state [16]. Further analysis of these factors, however, reveals that while both of these subclasses are highly MITF-associated, the degree of association is higher for *subclass C*. Examining downstream targets of MITF that are activated in each subclass (see Supplementary Text 8), we identified that *GPNMB, MLANA, PMEL,* and *TYR* are shared between two subclasses, whereas *ACP5, CDK2, CTSK, DCT, KIT, OCA2* and TRPM1/P1 are unique to *subclass C*. Besides *MITF* and *SOX10*, there are 38 other factors identified for subclass *B* and 17 other factors for subclass *C*. In particular, well-known oncogenes *TP53* and *MYC* are differentially activated in *subclass B* and *subclass C*, respectively. Thus, we use both *TP53* and *MITF* as characteristic biomarkers for subclasses *B* and both *MYC* and *MITF* for subclass *C*.

We then construct subclass-specific transcriptional regulatory networks (TRN) for these two subclasses. (The complete TRN for each of the 8 archetypes are available for download, see Supplemental Text 9). These factors have a total of 51 and 91 distinct target genes in their corresponding network that are functionally active. In order to investigate how the difference among these genes contribute to the overall survival of patients, we assessed the association between identified genes in each network and survival rate of patients, measured via multivariate Cox regressions [17]. We note that genes in *subclass C* significantly deviate from the null distribution of Cox coefficients for all genes (Kolmogorov-Smirnov test; p-val = 5.4×10^{-10}), whereas *subclass B* does not (p-value = 0.31), which translates into worse prognosis for *subclass C* (Figure 4).

To further study the underlying regulatory mechanisms that drive this poor-outcome phenotype, we focus on only the most significant transcription factors (those with functional activity p-values $\leq 10^{-3}$ rather than ≤ 0.05 above) in *subclass C* and construct their associated regulatory network. Figure 5a shows the

interaction network among highly significant TFs and their major targets in *subclass C*.

While some of these factors, and their target genes, were previously directly or implicitly associated with Melanoma, this network provides a novel system view of these interactions and highlights new regulatory interactions. For instance, amplification of the *MYC* oncogene has been long associated with poor outcome in Melanoma patients [18]. Also *E2F1* is a critical transcription factor that is involved in cell cycle transition from G1 to S phase, and its overexpression is commonly associated with poor patient survival in high-grade tumors [19]. The *LEF1* factor has a dual role. On one hand, it acts as a downstream effector of the Wnt signaling pathway and is associated with phenotype switching in Melanoma cells between proliferative and invasive states [20]. On the other hand, it has been suggested that *LEF1* has a distinct, Wnt-independent, role in activating *E2F1* [21]. Finally, we note that *LEF1* regulates both *MITF* and *MYC*. Collectively, we hypothesize that *LEF1* is a key TF that regulates phenotype switching from proliferative to invasive state in *subclass C*, by controlling other transcription factors, including *MITF*, *MYC,* and *E2F1*.

To revisit the problem of survival analysis, and to recover genes that affect this prognostic change, we project individual Cox coefficients for each gene onto the TRN of *subclass C* (Figure 5b). Two of the most significantly associated genes, *KIT* and *OCA2*, are among *MITF* targets that are unique to *subclass C* but not *subclass B*. The Kaplan-Meier plots for these two genes are visualized alongside the TRN. In addition, there are multiple targets of *MYC, LEF1,* and *E2F1* that are also associated with poor outcome for melanoma patients. These genes can provide further avenues for therapeutic interventions.

Methods

Datasets

Single cell gene expression datasets For all our studies, we rely on the following datasets collected from publicly available sources:

***Immune* (from the Supplementary Material of original paper)** : Comprehensive qPCR based assay of 1522 immune cells. This dataset spans 30 different types of stem, progenitor, and fully differentiated cells [22].

***Melanoma* (GEO: GSE72056)** : This dataset measures the expression profile of 4,645 malignant, immune, and stromal cells isolated from 19 freshly procured human melanoma tumors. These cells are classified into 7 major types [11].

***MouseBrain* (GEO: GSE60361)** : This dataset contains the expression profile of 3005 cells from the mouse cortex and hippocampus. These cells classify into seven major types, including *astrocytes-ependymal, endothelial-mural, interneurons, microglia, oligodendrocytes, pyramidal CA1*, and *pyramidal SS* [23].

***Pollen* (SRA: SRP041736)** : This is a small, but commonly used dataset that contains different cell types in developing cerebral cortex. It consists of 301 cells that classify into 11 distinct cell types [24].

Transcriptional Regulatory Network (TRN) We collect transcription factor (TF) – target gene (TG) interactions from the TRRUST database [25]. This dataset contains a total of $6,314$ regulatory interactions between 651 TFs and $2,102$ TGs.

Component 1: Computing a biologically-inspired metric to represent functional relationships between cells

The transcriptome of each cell consists of genes that are expressed at different levels and have different specificity with respect to the underlying cell types. *Housekeeping genes* correspond to the subset of genes responsible for mediating core cellular functions. These functions are needed by all cells to function properly, which result in ubiquitous expression of these genes across all cell types [26]. While fundamental to cellular function, these genes are not informative with respect to the identity of cells. That is, the mere fact that a housekeeping gene is expressed at a high-level in a given cell does not provide any information regarding its cell type. On the other hand, cell type-specific genes are preferentially expressed in one or a few selected cell types to perform cell type-specific functions. Unlike housekeeping genes, cell type-specific genes are, typically, weakly expressed, but are highly relevant for grouping cells according to their common functions. Our goal here is to define a similarity measure between cells that suppresses the noise contributed by universally expressed genes and enhances the signal contained in cell type-specific genes.

Suppressing universal but highly expressed genes To suppress the ubiquitously high expression of housekeeping genes, we adopt a method that we developed recently for bulk tissue measurements and extend it to single cell analysis [10]. This method projects a standardized representation of expression profiles of cells onto the orthogonal subspace of housekeeping genes. Let us denote the *raw expression profile* of cells using matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, where each row corresponds to a gene and each column represents a cell. We use $\boldsymbol{x}_i$ to denote the expression profile of ith cell. In addition, let us denote the signature vector of housekeeping genes by $\boldsymbol{v}$. As a first order estimate, a housekeeping signature is computed by taking the average expression over all cells: $\boldsymbol{v} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$. This choice is optimal in a least-square sense when the

chance of observing a gene is uniform across all cells. After estimating this baseline expression, we z-score normalize the profile of each cell: $\boldsymbol{z}_i = \frac{\boldsymbol{x}_i - \mu_i}{\sigma_i}$, where μ_i and σ_i are the mean and sample standard deviation of the entries in the ith cell profile. Similarly, we z-score normalize the signature vector of housekeeping genes, $\boldsymbol{v}$, to create a new vector $\boldsymbol{z}_v$. Finally, we project out the impact of the housekeeping gene expressions on each cell's profile as follows:

$$\boldsymbol{z}_i^{\perp} \quad = \quad \Big(\mathbf{I} - \frac{\boldsymbol{z}_v \boldsymbol{z}_v^T}{\|\boldsymbol{z}_v\|_2^2}\Big)\boldsymbol{z}_i. \tag{1}$$

This operation projects z_i to the orthogonal complement of the space spanned by the housekeeping genes. We then concatenate the column vectors $\boldsymbol{z}_i^{\perp}$ to create a *housekeeping (HK) adjusted cell signature* matrix $\mathcal{Z}^{\perp}$.

Enhancing signal from cell type-specific genes Next, to enhance the signal contributed by preferentially expressed genes, we propose an information theoretic approach [27]. The main idea is to use Shannon's entropy to measure the informativeness of gene expressions. If a gene is uniformly expressed across cells, it contains less information as opposed to the case in which it is selectively expressed in a few cells. To this end, we start with the positive projection of HK adjusted cell signatures, $\mathcal{P}^{(+)}(\mathcal{Z}^{\perp})$. Then, we normalize this matrix to construct a stochastic matrix $\mathbf{P}$, in which every row has sum one. Let $\boldsymbol{p}_i$ be the row vector associated with the ith gene. We compute the uniformity, or normalized entropy, of $\boldsymbol{p}_i$ as: $u(i) \; = \; -\sum_j p_{ij} \log(p_{ij}) / \log(n)$, where p_{ij} is an entry in the matrix $\mathbf{P}$ and n is the number of genes. Finally, we use these uniformity values as a basis to boost contributions from the most informative genes. To this end, we compute a scaling factor that is greater than one for cell type-specific and less than one for universally expressed genes, respectively. To do so, we first partition genes as either informative or noninformative using a *participation ratio* [28] measure applied to the uniformity values. A participation ratio measure continuously estimates the effective number of entries E in a given real-valued vector, which we

turn into a non-parametric cut-off value u^* as detailed in Supplemental Text 3. Then for each gene i, we define a scaling factor as $w_i = u^*/u(i)$. Finally, we compute the kernel matrix as follows:

$$\mathbf{K} = (\mathcal{Z}^{\perp})^T \mathbf{diag}(\boldsymbol{w}^2)\mathcal{Z}^{\perp} \tag{2}$$

In this formulation, if we denote $\mathbf{Y} = \mathbf{diag}(\boldsymbol{w})\mathcal{Z}^{\perp}$, then $\mathbf{K}$ is a dot-product kernel defined as $\mathbf{Y}^T\mathbf{Y}$. We will refer to $\mathbf{Y}$ as the *adjusted transcriptional profile* of cells, and $\mathbf{K}$ as the *cell similarity kernel*, or *ACTION metric*.

Component 2: Fitting a geometric construct to characterize functional identity of cells

The *functional identity* of cells that perform multiple functions can be expressed as a combination of a suitably constrained set of *principal functions*. The functional space of cells, thus, can be represented by a low-dimensional geometric construct, such as a polytope. The convex hull of a given set of points is the minimum volume polytope that encloses all points. This can be envisioned as a rubber band fitting to the outermost points. Constructing the convex hull in high-dimensional space is computationally expensive and susceptible to noise and overfitting. As an alternative, we seek a limited number of points on the convex hull that enclose as many points as possible, while being resilient to noise and outliers. Each point here represents a cell and each corner, or archetype, of this polytope is a candidate cell that best represents a unique *principal function*. To find these candidate cells, we use a modified version of the *successive projection algorithm (SPA)* combined with a novel model selection technique to identify an optimal number, k, of candidate cells on the approximate convex hull that best represent distinct pure cells with specialized principal functions. Finally, we use the *principal convex hull algorithm (PCHA)* to relax these corners to allow others cells to contribute to the identity of each archetype/corner.

Identifying the best k representative cells Formally, given a matrix $\mathbf{Y}$ representing the *adjusted transcriptional profile* of cells, we aim to construct an optimal set $\mathcal{S}$ of k columns such that each selected column is an ideal representative of the cells that perform a given principal function. Let us assume that matrix $\mathbf{Y}$ can be decomposed as $\mathbf{Y} = \mathbf{Y}(:, \mathcal{S})\mathbf{H} + \mathbf{N}$, where $\mathcal{S}$ is the selected column subspace of matrix $\mathbf{Y}$, $\mathbf{H}$ is nonnegative with column-sums equal to one, and $\mathbf{N}$ represents bounded noise, where $\|\mathbf{N}(:, j)\|_2 \leq \varepsilon$. That is, we can select $|\mathcal{S}| = k$ columns from matrix $\mathbf{Y}$ to represent rest of the columns, with consideration for noise. A matrix satisfying this condition is called *near-separable* and is known as the *near-separable nonnegative matrix factorization (NMF)* when $\mathbf{Y}$ is nonnegative. For a matrix satisfying near-separability, there is an efficient algorithm, with provable performance guarantees, that can identify columns in $\mathcal{S}$. Furthermore, premultiplying matrix $\mathbf{Y}$ with a nonsingular matrix $\mathbf{Q}$ preserves its separability, but if chosen carefully, can enhance the conditioning of the problem and accuracy of results. To find the optimal preconditioner matrix $\mathbf{Q}$, we use a theoretically-grounded method based on identifying a minimum volume ellipsoid at the origin that contains all columns of $\mathbf{Y}$ (Supplementary Text 5).

Estimating the optimal number of cells to represent principal functions Given that *SPA* selects k columns of $\mathbf{Y}$, *given* k, the next issue is how to find the optimal value of k that captures most variation in data without overfitting. We devised a novel monitoring technique that assesses the current k-polytope to see if there is any evidence of oversampling cell-space. If so, it stops the algorithm. Otherwise, it continues by adding new archetypes. Informally, oversampling happens when we start adding new archetypes to regions in the space that are already well-covered by other archetypes, in which case the newly added archetype would be significantly close to one or more other archetypes, compared to the rest of the archetypes. Given that each archetype is a candidate cell, we can measure relationship between them using the *ACTION* metric. The distribution of similarities resembles a normal distribution; however, as we start to oversample, the right

tail of the distribution starts getting heavier. To distinguish the pairs of archetypes in this heavy-tailed region, we z-score normalize pairwise similarities between archetypes and select all pairs whose z-transformed similarity scores are above 1.96, which corresponds to 95% confidence level under Gaussian assumption for the underlying distribution. Then, we build an *archetype similarity graph* using these pairs of close archetypes. In this graph, oversampling can be identified by the emergence of dense local regions. We use the Erdős-Rényi (ER) random graph model as a background to assess density of each sub-region, or connected component, in the archetype similarity graph [29, 30]. If we find at least one of the connected components that is significantly dense, which is a sign of oversampling, then we terminate the algorithm and choose the last value of k before oversampling happens.

Optimizing archetypes by relaxing the pure cell assumption After estimating k ideal candidate cells, or pure cells, we use archetypal-analysis (AA) [31], which can be viewed as a generalization of near-separability to relax corners by locally adjusting them to have contributions from multiple cells. Formally, we can formulate *AA* as follows:

$$\begin{aligned} \underset{\mathbf{C},\mathbf{H},\boldsymbol{\alpha}}{\text{minimize}} \quad & \|\mathbf{Y} - \mathbf{YCH}\| \\ \text{subject to} \quad & \|\mathbf{C}(:,i)\|_1 = 1. \\ & \|\mathbf{H}(:,i)\|_1 = 1. \\ & 0 \leq \mathbf{C}, 0 \leq \mathbf{H} \end{aligned} \tag{3}$$

Near-separable non-negative matrix factorization is a special case of AA in which $\mathbf{Y}$ is non-negative, $\mathbf{C}$ has exactly k nonzeros, and none of the columns have more than one element. We use an efficient algorithm, called *Principal Convex Hull Analysis (PCHA)*, to solve the above problem to a local optima.

The matrix $\mathbf{A} = \mathbf{YC}$ then stores the *archetypes*. Column stochasticity of $\mathbf{C}$ indicates that archetypes are convex combinations of data points, and column stochasticity of $\mathbf{H}$ indicates each data point can be

represented as convex combination of archetypes.

A complete pseudo-code fitting all these components together is provided in Supplementary Text 6.

Component 3: Constructing the driving transcriptional regulatory network for each archetype

In order to understand what control mechanisms are responsible for mediating the transcriptional phenotype of each archetype, we first have to identify key marker genes that distinguish a given archetype from the rest of archetypes (see Figure 5a for an illustrative guide to this section). To this end, we first orthogonalize each archetype with respect to all other archetypes. In this formulation, what remains, referred to as the *residual expression* of genes, ranks genes according to their importance in a given archetype. Let matrix $\mathbf{A} = \mathbf{YC}$ represent the identified archetypes. Let $\mathbf{A}^{(+)} = \mathcal{P}^{(+)}(\mathbf{A})$ be the projection to positive entries and let $\boldsymbol{a}_i^{(+)}$ stand for the column i of $\mathbf{A}^{(+)}$. Moreover, let $\mathbf{A}_{-i}^{(+)}$ denote the matrix without the ith column. Our goal is to project $\boldsymbol{a}_i^{(+)}$ into the subspace orthogonal to the columns spanned by $\mathbf{A}_{-i}^{(+)}$. Then, the orthogonalization step can be written as:

$$\boldsymbol{a}_i^{\perp} = \left(\mathbf{I} - \mathbf{A}_{-i}^{(+)}(\mathbf{A}_{-i}^{(+)^T}\mathbf{A}_{-i}^{(+)})^{-1}\mathbf{A}_{-i}^{(+)^T}\right)\boldsymbol{a}_i^{(+)} \tag{4}$$

Finally, we construct matrix $\mathbf{A}_{+}^{\perp}$ where each column is $\boldsymbol{a}_i^{\perp}$. Genes with high residual expression in each archetype are controlled through a complex regulatory network. To uncover these relationships, we aim to identify transcription factors that are significantly associated with the expression of marker genes, which we will refer to as *functionally active* TFs. Functional activity of TFs is inferred directly from the expression of their target genes; thus, these TF activities can be controlled at different stages, ranging from transcriptional to post-translation regulations. To infer these activities, we first need to classify their target genes as either active or inactive in a given context (archetype). To this end, we partition genes according to their residual expression and declare top-ranked genes as active. We use the minimum hypergeometric (mHG)

method [32] to find the optimal partition of genes and assign a p-value to it. The main step of this algorithm is similar to classic enrichment analysis: for a fixed size l, we use the hypergeometric p-value to assess the over-representation of target genes for a given TF among top-l markers for an archetype. Then, we compute the same statistic for all $1 \leq l \leq m$, where m is the total number of genes. The minimum hypergeometric tail that is obtained, referred to as the *mHG score*, specifies the best cut, $l^{(best)}$, and all target genes that are ranked higher than $l^{(best)}$ among marker genes are selected as regulated targets for that TF. Finally, we use the obtained *mHG score* to assess the significance of the TF itself. This can be accomplished using a dynamic programming algorithm that assesses the probability of observing the same or more significant mHG score within the population of all binary vectors of size m with exactly r nonzeros, where r is the number of targets for the current TF. The set of all significant transcription factors (TFs), together with their target genes (TGs) that fall above the cut that results in the mHG score, are used to construct the final transcriptional regulatory network (TRN).

Data Availability

All datasets used and all the Matlab codes used are available for download from `https://github.com/shmohammadi86/ACTION`.

References

1. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Research* **25,** 1491–1498 (Oct. 2015).

2. Jamal-Hanjani, M. *et al.* Translational Implications of Tumor Heterogeneity. *Clinical Cancer Research* **21,** 1258–1266 (Mar. 2015).

3. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology* **34,** 1145–1160 (Nov. 2016).

4. Qiu, P. *et al.* Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nature Biotechnology* **29,** 886–891 (Oct. 2011).

5. Marco, E. *et al.* Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences* **111,** E5643–E5650 (Dec. 2014).

6. Grün, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525,** 251–5 (Sept. 2015).

7. Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31,** 1974–1980 (June 2015).

8. Korem, Y. *et al.* Geometry of the Gene Expression Space of Individual Cells. *PLOS Computational Biology* **11** (ed Iakoucheva, L. M.) e1004224 (July 2015).

9. Ji, Z. & Ji, H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Research* **44,** e117–e117 (July 2016).

10. Mohammadi, S. & Grama, A. *De novo identification of cell type hierarchy with application to compound marker detection* in *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB '16* (ACM Press, New York, New York, USA, 2016), 251–260.

11. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352,** 189–196 (Apr. 2016).

12. Wang, B. *et al. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning* tech. rep. (May 2016).

13. Hart, Y. *et al.* Inferring biological tasks using Pareto analysis of high-dimensional data. *Nature Methods* **12,** 233–235 (Jan. 2015).

14. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nature methods* **12,** 1–10 (2015).

15. Antebi, Y. E. *et al.* Mapping Differentiation under Mixed Culture Conditions Reveals a Tunable Continuum of T Cell Fates. *PLoS Biology* **11** (ed Bhandoola, A.) e1001616 (July 2013).

16. Verfaillie, A. *et al.* Decoding the regulatory landscape of melanoma reveals TEADS as regulators of the invasive cell state. *Nature communications* **6,** 6683 (2015).

17. Anaya, J. OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. *PeerJ Computer Science* **2,** e67 (June 2016).

18. Kraehn, G. M. *et al.* Extra c-myc oncogene copies in high risk cutaneous malignant melanoma and melanoma metastases. *British journal of cancer* **84,** 72–9 (Jan. 2001).

19. Alla, V. *et al.* E2F1 in melanoma progression and metastasis. *Journal of the National Cancer Institute* **102,** 127–33 (Jan. 2010).

20. Eichhoff, O. M. *et al.* Differential LEF1 and TCF4 expression is involved in melanoma cell phenotype switching. *Pigment cell & melanoma research* **24,** 631–42 (Aug. 2011).

21. Zhou, F. *et al.* LEF-1 activates the transcription of E2F1. *Biochemical and biophysical research communications* **365,** 149–53 (Jan. 2008).

22. Guo, G. *et al.* Mapping Cellular Hierarchy by Single-Cell Analysis of the Cell Surface Repertoire. *Cell Stem Cell* **13,** 492–505 (Oct. 2013).

23. Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347,** 1138–42 (Mar. 2015).

24. Pollen, A. A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature Biotechnology* **32,** 1053–1058 (Aug. 2014).

25. Han, H. *et al.* TRRUST: a reference database of human transcriptional regulatory interactions. *Scientific Reports* **5,** 11432 (June 2015).

26. Eisenberg, E. & Levanon, E. Y. *Human housekeeping genes, revisited* 2013.

27. Schug, J. *et al.* Promoter features related to tissue specificity as measured by Shannon entropy. *Genome biology* **6,** R33 (2005).

28. Farkas, I. J. *et al.* Spectra of real-world graphs: Beyond the semicircle law. *Physical Review E* **64,** 026704 (July 2001).

29. Magner, A., Mohammadi, S. & Grama, A. Combining Density and Overlap (CoDO): A New Method for Assessing the Significance of Overlap Among Subgraphs. *CoRR* **abs/1605.06167** (2016).

30. Koyutürk, M., Szpankowski, W. & Grama, A. Assessing significance of connectivity and conservation in protein interaction networks. *Journal of computational biology : a journal of computational molecular cell biology* **14,** 747–64.

31. Cutler, A. & Breiman, L. Archetypal Analysis. *Technometrics* **36,** 338 (Nov. 1994).

32. Eden, E. *et al. Discovering motifs in ranked lists of DNA sequences.* PhD thesis (Technion - Israel Institute of Technology, Mar. 2007), 77.

33. Jiang, L. *et al.* GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biology* **17,** 144 (Dec. 2016).

34. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31,** 2989–2998 (Sept. 2015).

35. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods* **11,** 333–337 (Jan. 2014).

36. Gillis, N. & Vavasis, S. A. Semidefinite Programming Based Preconditioning for More Robust Near-Separable Nonnegative Matrix Factorization. *SIAM Journal on Optimization* **25,** 677–698 (Jan. 2015).

Acknowledgements This work is supported by the NSF Center for Science of Information STC (CCF-093937), NSF Grant BIO 1124962, NSF Big data IIS-1546488, NSF CAREER award (CCF-1149756), the DARPA SIMPLEX problem, and the Sloan Foundation.

Author Contributions SM developed the method, implemented the method, ran the computational experiments, analyzed the results, and drafted the manuscript. VR performed initial experiments that led to the final method. DG and AG helped design the method, analyzed results, and assisted with the writing. All authors read and approved the final manuscript.

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to S.M. (email: mohammadi@purdue.edu).

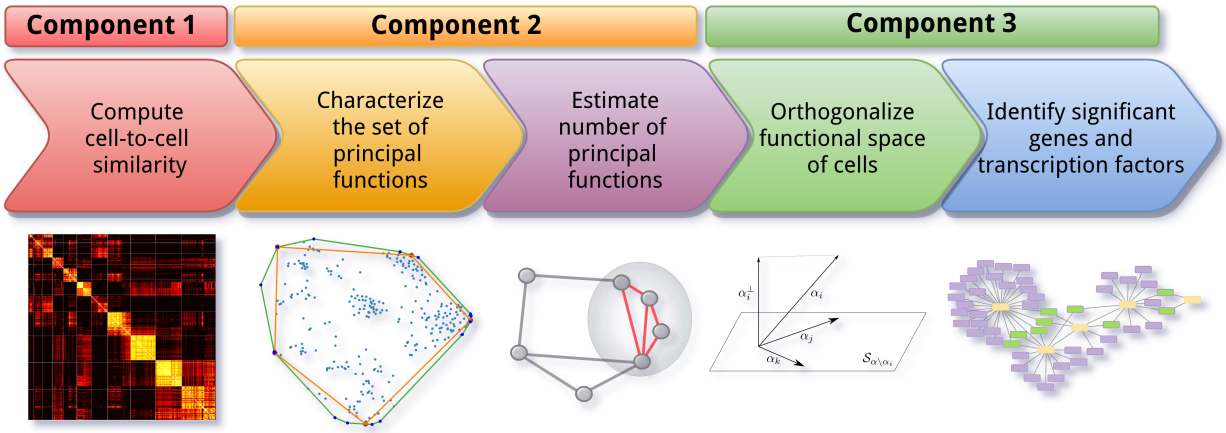

Figure 1: **Overview of *ACTION*.** *ACTION* consists of five main steps: (**i**) A biologically-inspired metric to capture similarity among cells. (**ii**) A geometric approach for identifying the set of principal functions. (**iii**) An automated mechanism for identifying the number of principal functions needed to represent all cells. (**iv**) An orthogonalization procedure for identifying key markers for each principal function. (**v**) A statistical approach for identifying key regulatory elements in the transcriptional regulatory network. These steps are grouped into three main components in the *ACTION* method that are each discussed in the methods section.

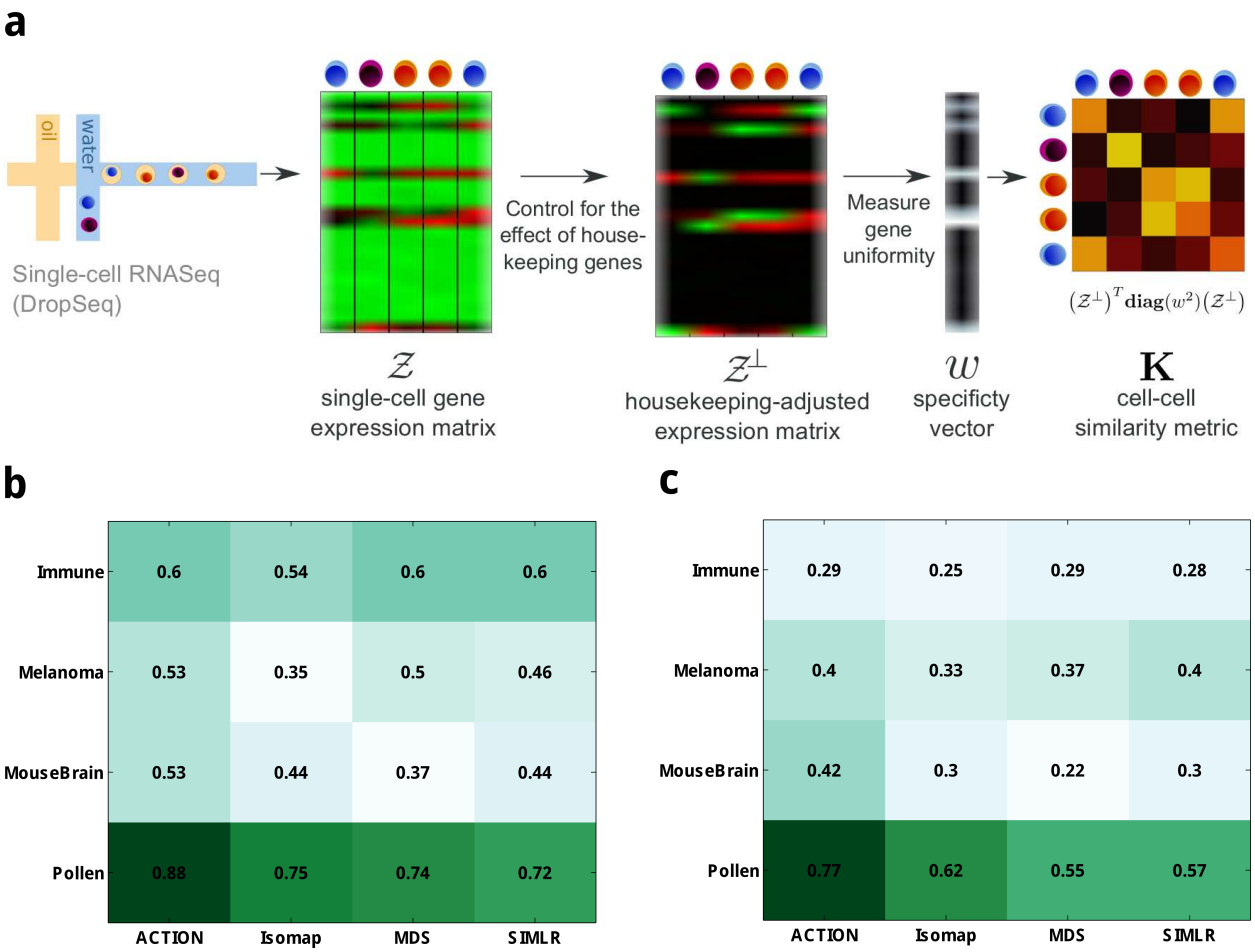

Figure 2: ***ACTION* Similarity Metric.** (**a**) Workflow of *ACTION* cell-to-cell similarity metric. (**b**) Performance of *ACTION* in terms of Normalized Mutual Information (NMI). (**c**) Performance of *ACTION* with respect to Adjusted Rand Index (ARI) measure. Both NMI and ARI values in panels (**b**) and (**c**) are between zero and one with larger values indicating better results. Performance of similarity measures is evaluated in the context of a kernel k-means clustering technique applied to each of the computed metrics. The intensity of green cells show cases where a method performs better than others for a given dataset. the *ACTION* metric performs equally good or significantly better than other methods in identifying similarity between cells.

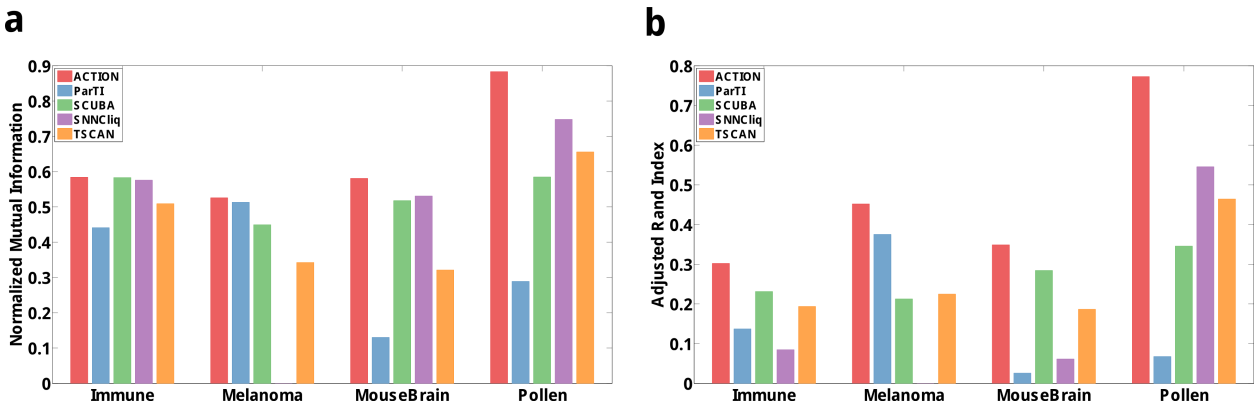

Figure 3: **Performance of *ACTION* in identifying discrete cell types.** Cell types are identified by classifying cells according to their dominant principal function. (**a**) Normalized Mutual Information (NMI) of cell type identification. (**b**) Adjusted Rand Index (ARI) performance measure. In both panels, performance of *SNNCliq* for the *Melanoma* dataset is left blank, as it did not finish in the given time. The NMI and ARI measures are computed against the cell types provided in the original sample annotations. With respect to both measures, the *ACTION* method outperforms all other methods in identifying cell types.

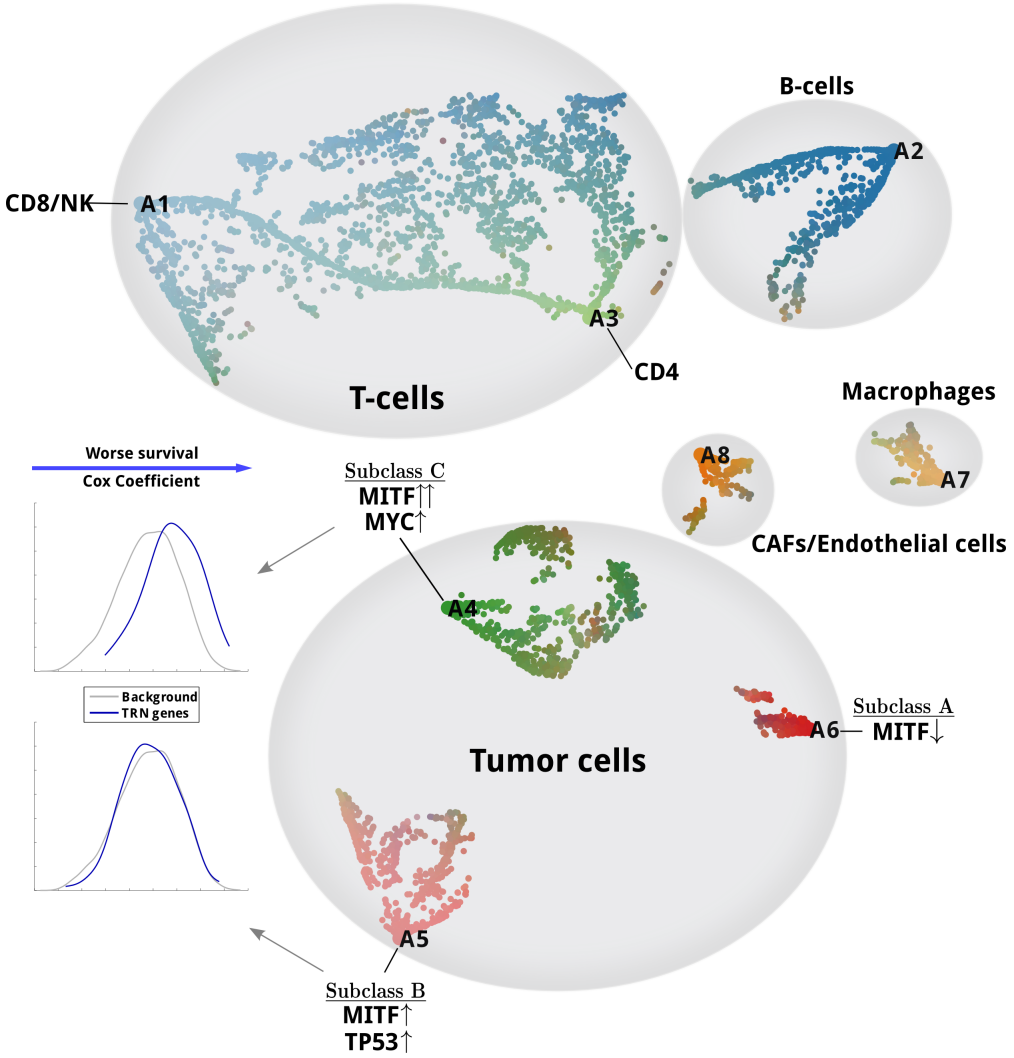

Figure 4: **A continuous view on the space of principal functions in the Melanoma dataset**

Figure 4 *(previous page)*: Each archetype, representing a principal function, is illustrated using a textual label (A1-A8). Each small dot represents a cell. Cells are color coded based on their proximity to archetypes. All data points are projected onto a 2D plane using a carefully initialized Stochastic Neighbor Embedding method (SNE, see Supplemental Text 4). The functional space of cells exhibit a mix of cell state continuum, such as in the case of T-cells, as well as discrete cell types. Three subclasses of melanoma tumor cells are marked accordingly in the map. Subclasses *B* and *C* are both MITF-associated. Among them, genes that participate in the transcriptional regulatory network (TRN) for *subclass B* do not show any significant shift in Cox coefficient, compared to the background of all genes, whereas in *subclass C* they do. In this sense, high-expression of genes in the TRN of *subclass C* is significantly associated with worse outcome in the melanoma patients.

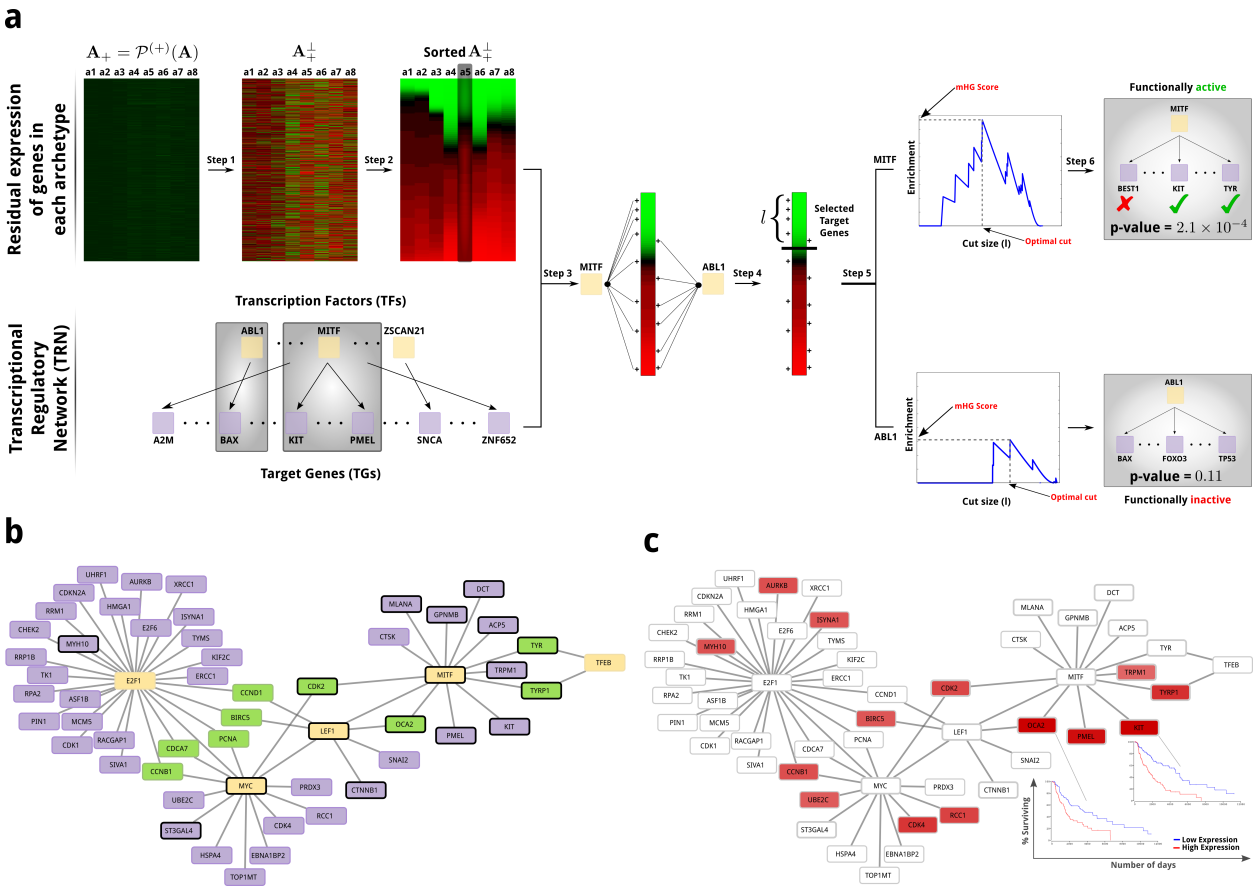

Figure 5: **The transcriptional regulatory network (TRN) for MITF-associated Melanoma patients highlights a number of genes that have not previously been associated with Melanoma – along with some known markers**.

Figure 5 *(previous page)*: (**a**) Main steps involved in the construction of archetype-specific TRNs: (1) Orthogonalize archetypes with respect to each other, (2) Sort genes based on their residual expression, (3) Map gene targets for TFs to the sorted list genes, (4) Enrichment analysis for fixed cut size l, (5) Find optimal cut size and compute minimum HyperGeometric (mHG) score, and (6) Assess significance of the mHG score using Dynamic Programming (DP). (**b**) A subset of the TRN of *subclass A* induced by using only the most significant TFs. The yellow nodes are transcription factors (TF), the purple nodes are target genes (TG), and green nodes are target genes that bridge different TFs. Genes marked with black border are known to be involved in the proliferative subclass of Melanoma. (**c**) The TRN of *subclass A* with genes color-coded according to their Cox coefficient. Red genes are the ones whose high expression is associated with worse outcome, and brightness of the color relates to the severity of the outcome. Kaplan-Meier plots for two of the targets of MITF that are unique to *subclass A* but not *subclass C* are shown on the plot.

Supplementary Material

1. Overview of prior methods for cell-type identification

Various methods have been developed for cell type identification. **SNNCliq** [7] computes a similarity graph among cells, referred to as *shared nearest neighbor (SNN)*. It then uses a graph-based clustering algorithm to identify dense subgraphs. **TSCAN** [9] starts by grouping genes with similar expression patterns into "modules" and represents all cells in this reduced space. It then performs principal component analysis (PCA) over the module space to further reduce dimensions. Finally, cells are clustered by fitting a mixture of multivariate normal distributions to the data, with the number of components estimated using the Bayesian Information Criterion (BIC). **SCUBA** [5] first uses k-means with gap statistic to cluster data along an initial binary tree by analyzing bifurcation events for time-course data. Then,it refines the tree using a maximum likelihood scheme. BackSPIN [23] is based on the SPIN algorithm, which permutes correlation matrix of cell types to extract its underlying structure. BackSPIN then couples it with a divisive splitting procedure to identify clusters from the ordered similarity matrix. Two methods are specifically designed to identify rare cell types. RaceID [6] uses k-means to first cluster cells, with the number of clusters identified using gap statistic. Then, it identifies rare cell types as outliers that are not explained by an appropriate noise model, accounting for both biological and technical variations. GiniClust [33] aims to identify marker genes that are specific to rare cell types using the concept of Gini index. Then, it computes distances between cell types in this reduced subspace and uses DBSCAN clustering algorithm to identify cell types. In addition to these methods, there are approaches that visualize cell types on a continuous spectrum in a given space. Haghverdi *et al.* [34] use diffusion maps to model the continuous spectrum of cells. In another direction, Korem *et al.* [8], adopted a previously developed method, called **Pareto task inference (ParTI)** [13], and

applied it to single cell datasets. While ParTI uses a similar notion of archetypal analysis to what we do, our method provides key regulatory pathways underlying these archetypes. Our method is further founded on a biologically-inspired, kernel-based approach, has a novel, deterministic initialization procedure, and an stable algorithm to identify the effective number of archetypes.

2. Comparison between kernel k-means and ACTION in identifying cell types.

A comparison of using kernel k-means versus archetypal analysis (AA) are presented in Table 1. Initially, we observe that using AA instead of kernel k-means to identify discrete cell types enhances Adjusted Rand Index (ARI) scores while preserving Normalized Mutual Information (NMI) scores. However, in the most cases, these differences are marginal, which suggests that the observed superior performance of *ACTION* is dominantly due its kernel. This is not surprising since the main difference in the objective function of kernel k-means versus AA is that in the former assignments are continuous, whereas in the latter they are binary. While both binarizing AA and solving kernel k-means are heuristics to solve the NP-hard cell type assignment problem for each cell, interestingly, the former surpasses in performance.

On the other hand, there are major benefits for using AA to identify the functional identity of cells. First, it allows us to define a continuous *functional space* to represent each cell. Secondly, AA is less sensitive to the sampling density of cells [13]. Thus, we will use AA throughout our paper.

3. Participation ratio threshold for informative and non-informative genes

Given the uniformity values $u(i)$ for each gene. We first shift these values such that the minimum value is zero: $\hat{u}(i) = u(i) - \min_i u(i)$. Then we compute the participation ratio of $\hat{u}(i)$: $E = (\sum_i \hat{u}(i)^2)^2 / (\sum_i \hat{u}(i)^4)$.

	ARI			
	Immune	Melanoma	MouseBrain	Pollen
ACTION	0.30	0.45	0.35	0.77
k-means	0.28	0.38	0.34	0.77
	NMI			
	Immune	Melanoma	MouseBrain	Pollen
ACTION	0.58	0.53	0.58	0.88
k-means	0.58	0.54	0.58	0.88

Table 1: Comparison between performance of kernel k-means versus AA with the ACTION kernel

This value E tells us that there are approximately E non-zero values in $\hat{u}(i)$ over all i, which we round to the nearest integer. The value $\hat{u}^*$ is simply the Eth largest entry in the sorted vector. The value u^* used in the weighting is then reshifted by the minimum value: $u^* = \hat{u}^* + \min_i u(i)$.

4. Deterministic initialization of SNE embedding

To ensure that our low-dimensional visualizations are replicable, we describe the exact, deterministic process we use to create them. The first step is to compute a Fiedler embedding of a similarity matrix derived from the cell coordinates returned by PCHA.

1. Take $\mathbf{H}$ from the PCHA as input.

2. Set $\tilde{\mathbf{H}} = [\mathbf{H} \quad \mathbf{I}]$

3. Let $\tilde{h}_i$ be the ith column of $\tilde{\mathbf{H}}$, and compute entries of the matrix $D_{ij} = \|\tilde{h}_i - \tilde{h}_j\|_2$ (that is, Euclidean distance between vectors $\tilde{h}_i$ and $\tilde{h}_j$).

4. Convert Distances to Similarity following Network Similarity Fusion [35] affinity matrix construction

   (a) Let d_i^* be the average distance from i^{th} cell to its top $k = \mathbf{round}(n/10)$ closest neighbors, with n being the total number of cells. (If you sort columns of the matrix $\mathbf{D}$, this is just the top k entries.)

   (b) Set $\Sigma_{i,j} = (d_i^* + d_j^* + 2\varepsilon + D_{ij})/3$, where ε is 2^{-52}.

   (c) Set $\tilde{\Sigma}_{i,j} = \begin{cases} \Sigma_{i,j} + \varepsilon & \Sigma_{i,j} \geq \varepsilon \\ \varepsilon & \text{Otherwise.} \end{cases}$

   (d) Set $W_{i,j}$ to be the probability that a normally distributed random variable with mean 0 and standard deviation $\tilde{\Sigma}_{i,j}$ has value $D_{i,j}$.

5. Set $\mathbf{G} = (\mathbf{W} + \mathbf{W}^T)/2$ be the weighted graph between cells.

6. Set $\mathbf{L} = \text{diag}(\mathbf{G} \cdot \text{ones}(n, 1)) - \mathbf{G}$ (that is, $\mathbf{L}$ is the combinatorial Laplacian of $\mathbf{G}$).

7. Compute the three smallest eigenvalues and eigenvectors of $\mathbf{L}$, $(\boldsymbol{v}_1, \lambda_1), (\boldsymbol{v}_2, \lambda_2), (\boldsymbol{v}_3, \lambda_3)$. Note that λ_1 is zero because of the Laplacian structure.

8. Set $\boldsymbol{x} = \boldsymbol{v}_2/\sqrt{\lambda_2}$

9. Set $\boldsymbol{y} = \boldsymbol{v}_y/\sqrt{\lambda_3}$

10. Let $\boldsymbol{x}, \boldsymbol{y}$ be the initial coordinates of each cell for the tSNE procedure.

5. Performance of SPA with preconditioner

Let $\mathbf{Y} = \mathbf{WH}$, where matrix $\mathbf{W}$ is defined as $\mathbf{Y}(:, \mathcal{S})$, with $\mathcal{S}$ being the selected column subspace of matrix $\mathbf{Y}$, and $\mathbf{H}$ is a non-negative matrix with column-sums equal to one. Moreover, let matrix $\tilde{\mathbf{Y}} = \mathbf{Y} + \mathbf{N}$, where the noise is bounded: $\|\mathbf{N}(:, j)\|_2 \leq \varepsilon$. Then, the performance of the SPA algorithm has the following upper bound guarantee:

$$\max_{1 \leq j \leq k} \min_{s \in \mathcal{S}} \| \tilde{\mathbf{Y}}(:, s) - \mathbf{W}(:, j) \| \leq \mathcal{O}\Big(\epsilon \kappa^2(\mathbf{W})\Big) \tag{5}$$

More recently, other techniques have been developed to enhance the robustness of *SPA* to noise [36]. These methods are based on the fact that premultiplying matrix $\mathbf{Y}$ by a nonsingular matrix $\mathbf{Q}$ preserves its separability. In this case, the upper bound limit changes to: $\mathcal{O}\Big(\epsilon \kappa(\mathbf{W}) \kappa^3(\mathbf{QW})\Big)$. Thus, by carefully choosing matrix $\mathbf{Q}$, we can enhance the conditioning of the problem. Ideally, if $\mathbf{Q} = \mathbf{W}^{-1}$, then $\kappa^3(\mathbf{QW}) = 1$ and we reduced the upper bound from quadratic to linear. While $\mathbf{W}^{-1}$ is not accessible, it turns out that we can approximate $\mathbf{W}^{-1}$ using a *minimum volume ellipsoid* centered around the origin that contains all columns of the original matrix $\mathbf{X}$. Formally, this can be solved using the following SDP to identify matrix $\mathbf{A}^*$:

$$\mathbf{A}^{(*)} = \underset{\mathbf{A} \in \mathbb{S}_+^k}{\operatorname{argmax}} \, \mathbf{det}(\mathbf{A})$$

$$\text{s.t.: } \mathbf{Y}(:, j)^T \mathbf{A} \mathbf{Y}(:, j) \leq 1; \forall j$$

Since $\mathbf{A^T}$ is symmetric positive definite, we then compute $\mathbf{A^T} = \mathbf{Q}^T \mathbf{Q}$ using Cholesky factorization and use it as a preconditioner.

6. Pseudo-code for fitting a geometric construct over single cells

See Algorithm 1.

Algorithm 1 SPA algorithm with prewhitening

Input: $\mathbf{Y} \in \mathbb{R}^{m \times n}$: adjusted expression profile of cells

Output: $\mathbf{A} \in \mathbb{R}^{m \times k}$: principal functions, $\mathbf{H} \in \mathbb{R}_{+}^{k \times n}$: functional identity of cells

Solve **minimum volume ellipsoid** problem to identify preconditioner $\mathbf{Q}$.

$\mathbf{K} = \mathbf{Y}^T\mathbf{Y}, \mathbf{R} = \mathbf{Q}\mathbf{Y}, \mathcal{S} = \{\}$

for $i = \{1, \cdots, max_k\}$ **do**

  $\alpha = \text{argmax}_j \|\boldsymbol{r}_j\|_2$ $\{\boldsymbol{r}_j$ is the jth column$\}$

  $\boldsymbol{\beta} = \mathbf{R}(:, \alpha)$

  $\mathbf{R} \leftarrow (\mathbf{I} - \frac{\boldsymbol{\beta}\boldsymbol{\beta}^T}{\boldsymbol{\beta}^T\boldsymbol{\beta}})\mathbf{R}$ {Orthogonal Projection}

  $\mathcal{S} \leftarrow \mathcal{S} \cup \{\boldsymbol{\beta}\}$

  Construct archetype similarity graph from $\mathbf{G} = \mathbf{K}(\mathcal{S}, \mathcal{S})$

  if $subgraph_{density}(\mathbf{G})$ is significant **then**

    break

  end if

end for

Initialize $\mathbf{C}_0$ using selected columns in $\mathcal{S}$, and run kernel **PCHA** with $\mathbf{K}$ to estimate matrices $\mathbf{C}$ and $\mathbf{H}$

$\mathbf{A} = \mathbf{Y}\mathbf{C}$

7. List of functionally active transcription factors

The following table lists all transcription factors (TFs) that are either significant in *subclass B, archetype 5*, *subclass C, archetype 4*, or both. Second and third columns of the table are the p-value of the functional activity of TFs. These factors are sorted according to the relative importance in these subclasses *B* and *C*. Green rows are the ones that are significant in both. *TP53* and *MYC*, marked in red, are used in conjunction with *MITF* to distinguish these two classes.

TF	*Subclass B*	*Subclass C*
MITF	0.00E+00	2.14E-04
E2F1	5.18E-06	7.62E-01
LEF1	3.19E-04	3.83E-01
MYC	8.39E-04	3.03E-01
TFEB	2.30E-04	2.63E-02
E2F4	1.32E-02	5.11E-01
TCF7	3.32E-02	1.00E+00
CNBP	4.05E-02	1.00E+00
FUBP1	4.05E-02	1.00E+00
RBL1	4.05E-02	1.00E+00
SRSF1	4.05E-02	1.00E+00
TBL1X	4.05E-02	1.00E+00
SNIP1	1.30E-02	2.90E-01
MTA1	4.05E-02	6.36E-01
HOXA1	1.30E-02	1.96E-01

OTX2	3.32E-02	1.71E-01
ONECUT2	4.05E-02	1.31E-01
SOX9	2.84E-03	4.57E-03
NFIB	3.60E-01	3.14E-02
TBP	4.62E-01	2.52E-02
USF1	4.62E-01	2.49E-02
KLF5	7.85E-01	3.97E-02
POU4F2	3.60E-01	1.73E-02
ATF3	1.00E+00	4.54E-02
PAX3	3.59E-01	1.62E-02
ESR1	1.07E-01	4.57E-03
MYF6	1.00E+00	4.02E-02
HDAC7	1.00E+00	3.97E-02
MAML1	1.00E+00	3.97E-02
NKX2-3	1.00E+00	3.97E-02
PPARG	1.00E+00	3.87E-02
LCOR	6.33E-01	2.45E-02
SMAD4	7.85E-01	2.63E-02
GTF2I	1.00E+00	3.14E-02
TAF1	1.00E+00	3.00E-02
SOX10	1.03E-02	3.00E-04
ETS2	9.57E-01	2.63E-02
ETS1	1.00E+00	2.53E-02

MYF5	1.00E+00	2.49E-02
HOXA7	8.15E-01	1.50E-02
EGR2	8.24E-01	1.39E-02
STAT3	1.00E+00	1.47E-02
JUND	9.48E-01	1.39E-02
KLF4	1.00E+00	1.24E-02
TWIST2	6.52E-01	7.91E-03
KLF6	8.10E-01	8.97E-03
TP53	3.11E-01	3.06E-03
TFCP2	8.15E-01	7.30E-03
ESR2	1.00E+00	7.30E-03
PARP1	1.00E+00	7.30E-03
ETV4	1.00E+00	4.57E-03
PPARD	1.00E+00	4.57E-03
CTNNB1	3.50E-01	1.58E-03
AR	1.00E+00	1.58E-03
BRCA1	1.00E+00	2.14E-04

8. Regulated downstream targets of MITF factor in Subclasses *B* and *C*

The following table lists the full set of significant downstream targets of MITF in both subclasses *B& C*. Genes *GPNMB, MLANA, PMEL* and *TYR* are shared between two subclasses, whereas the rest of targets are unique to one of them. For genes that have significant effect on the survival rate, their Cox coefficient

is presented in the table. A positive Cox coefficient indicates that high expression of the given genes is associated with poor survival.

Target	Subclass B	Subclass C	Cox Coefficient
ACP5	✓		-
CDK2	✓		0.218
CTSK	✓		-
DCT	✓		-
KIT	✓		0.3214
OCA2	✓		0.3038
TRPM1	✓		0.188
TYRP1	✓		0.2422
GPNMB	✓	✓	-
MLANA	✓	✓	-
PMEL	✓	✓	0.2765
TYR	✓	✓	-
BEST1		✓	-
BIRC7		✓	-
FOS		✓	-
MET		✓	-

9. Regulatory networks

We provided transcriptional regulatory networks (TRNs) for all eight archetypes in the Melanoma dataset. There are two files per network, one edge list file and one node annotation file, both in txt format. The former contains edges in form of ”TF_name tab TG_name,” whereas in the second file there is a row for each node (either TF or TG) that provides additional information about it. This information includes: (i) type (TF or TG), (ii) -log10 of the functional activity p-value, if node is a TF, or zero otherwise, (iii) residual expression of genes (either TF or TG) after orthogonalization of archetype, (iv) heuristic importance of node (for visualization purposes only), and (v) cox survival coefficient.