Take ACTION to identify high-resolution cell types and associated transcriptional pathways

Shahin Mohammadi[*], Vikram Ravindra, David Gleich, and Ananth Grama

Department of Computer Sciences, Purdue University

Abstract

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 that automatically identify cell types and their properties based on expression profiles can be used to uncover processes involved in lineage differentiation as well as sub-typing. They can also be used to suggest personalized therapies based on molecular signatures associated with pathology. We develop a new method, called ACTION, for projecting cells onto the state space of functional profiles, classifying them according to their principal functions, and reconstructing cell type-specific regulatory networks. Results on sub-typing cancer cells in Melanoma patients reveal novel biomarkers along with their regulatory networks.

Background

Complex tissues consist of heterogeneous populations of interacting cells that are specialized to perform different functions. With rapid growth in single cell transcriptomic technologies, the observed diversity of known cell types has greatly expanded. What were once believed to be homogeneous groups of cells can now viewed as ecosystems of varying cell types [1]. In tumor microenvironments, for example, immune, stromal, and cancerous cells coexist, cooperate, and compete for resources. The exact composition of these cells, as well as their molecular makeup, have significant impact on diagnosis, prognosis, and treatment of cancer patients [2]. Single cell technologies have already been proven useful for dissecting this complex microenvironment [3]. Using the rapidly growing datasets of single cell gene expression profiles, a key challenge is to identify *de novo* cell types directly from genome-wide transcriptomic phenotypes [4]. An important problem in cell type identification is the existence of rare but key cell types, such as circulating tumor cells [5]. Beyond identifying cell types, it is also import to identify factors that distinguish them from other cell types.

We propose a new method, called *Archetypal-analysis for cell type identificaTION (ACTION)*, to identify cell types from single cell expression datasets. Our method is robust to biological noise, identifies a wide range of cell types with varying relative populations, and provides a novel mechanism for constructing transcriptional regulatory networks (TRN) that mediate characteristic behaviors of each cell type. At the core of our method is a biologically-inspired metric for similarity of cells, as characterized by their transcriptional profiles. This metric accounts for specificity of marker genes and defines a signature for each cell that is robust to noise. At the same time, it is sensitive enough to capture weak cell type-specific signals. This metric helps us construct a geometric representation for the space of principal functions, which are groups of distinguishing

[*] Corresponding author: mohammadi@purdue.edu

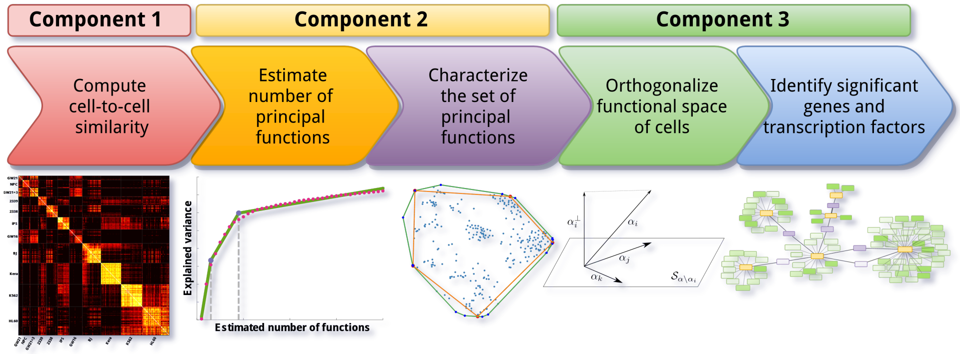

Component 1
Component 2
Component 3
Compute cell-to-cell similarity
Estimate number of principal functions
Characterize the set of principal functions
Orthogonalize functional space of cells
Identify significant genes and transcription factors
Explained variance
Estimated number of functions
$\alpha_i^{\perp}$
α_i
α_j
α_k
$S_{\alpha\backslash\alpha_i}$

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

functions that are uniquely performed by specialized cells. In this space, assigning cells to their closest principal function accurately identifies cell types. Finally, we develop a statistical framework to identify key marker genes, 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.

Our method provides a flexible approach for directly mapping characteristic transcriptional regulatory networks of cells from the raw transcriptomic data. We apply our method to the problem of subtyping Melanoma patients and identify a coherent subclass, which closely resembles noninvasive tumors [6]. For this subclass, we characterized key marker genes, as well as their underlying pathways. This analysis highlights a MITF-associated regulatory network and suggests a potential mechanism for distinguishing invasive and proliferative types of melanoma.

Significance. A few methods have been proposed for the problem of cell type identification [7]–[13]. A common theme underlying these methods is to cluster coherent cells as putative cell types [4]. At the core of these clustering methods is a similarity measure that defines relationships among cells. A majority of prior methods rely on classical measures such as correlation or Euclidean distance to define such relationships. However, this approach is confounded by ubiquitously and highly expressed levels of housekeeping genes. Cell type-specific markers, on the other hand, have a weaker signal in comparison. This, in turn, causes a majority of traditional techniques to be driven by biological noise contributed by housekeeping genes [14]. To overcome this, methods – such as *ACTION* – that are robust to biological noise but are sensitive enough to identify cell type-specific signals are critically needed. Once the identity of a cell has been established, it is unclear what distinguishes it from other cell types. Transcriptional regulatory networks (TRNs) are important aspects of this differentiation process. Understanding cell type-specific TRNs has the potential to explain distinguishing mechanisms underlying observed transcriptional phenotypes. *ACTION* is among the first set of methods to directly infer cell type-specific networks from single cell expression datasets.

Results and discussion

The *ACTION* framework consists of three major components, shown in Figure 1: (i) A robust 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 functional roles of a cell form an embedded hierarchy, with successively refined set of tissue-specific functions. When used with a classic clustering algorithm such as *k*-means, *ACTION* metric surpasses all other measures of cell similarity in identifying cell types. The next component of our method is a geometric approach for identifying principal functions of cells, each represented by an archetype (corner) of the convex hull in the functional space of cells. Finally, *ACTION* uses a novel method that utilizes the geometric view of cell functions to construct the transcriptional regulatory network (TRN) that mediates characteristic behavior of each cell type. In what follows, we describe, validate, and discuss each component in detail.

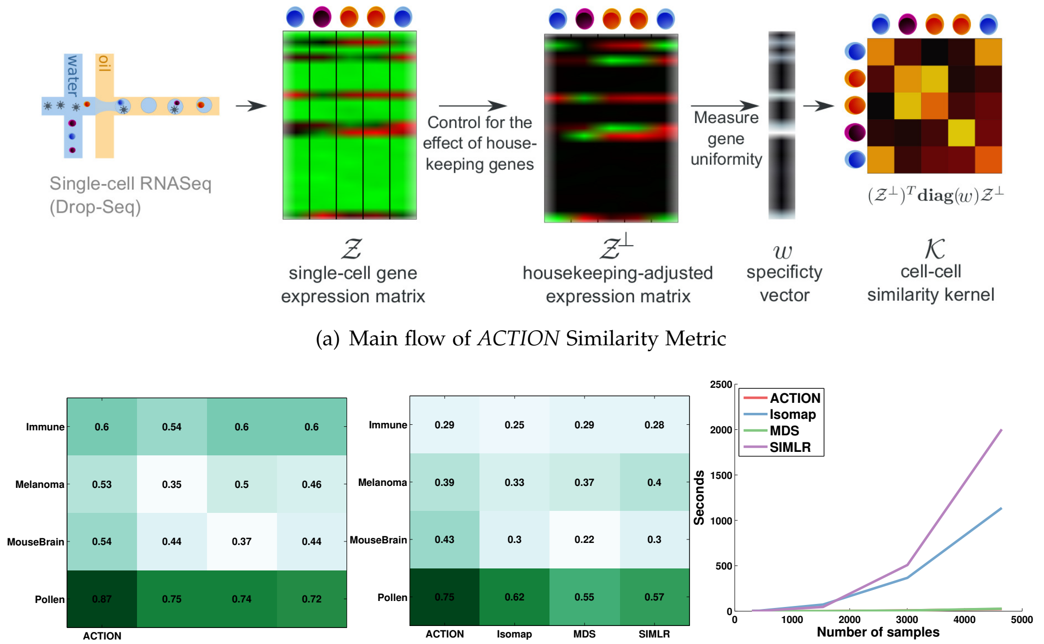
Component 1: Measuring cell-to-cell similarity

An essential component of any method for identifying cell types is the ability to quantify similarity between individual cells. Most prior methods rely on traditional measures, such as Euclidean distance, that are not specifically targeted towards transcriptomic profiles. In contrast, we define a similarity metric, or formally a kernel, specifically designed for measuring similarity between cells [14]. Our approach is based on the observation that housekeeping genes, while not informative of cell type identity, significantly impact traditional measures of cell similarity due to their ubiquitous and high expression levels. Suppressing these genes significantly enhances the signal-to-noise ratio (SNR) in expression profiles, allowing us to extract a stronger cell type-specific signal.

Novel methodology Our method starts by projecting transcriptional signatures to the orthogonal subspace spanned by housekeeping genes. We then boost the contribution of cell type-specific genes using an information theoretic approach. Finally, we combine these two measures to define a robust measure of cell-to-cell similarity. This approach is illustrated in Figure 2. The mathematical models underlying the metric are described in the Methods section.

Validation To establish the superiority of our metric, we compare it against one measure specifically designed for single cell analysis, *SIMLR*, and two general measures: *multidimensional scaling (MDS)*, and Isomap. SIMLR [15], combines a number of distance metrics to learn a joint similarity score that maximizes the block diagonal structure of the resulting matrix. Both *MultiDimensional Scaling (MDS)* and *Isomap* are nonlinear dimension reduction techniques. The former method projects points into a low-dimensional space, such that distances between samples are preserved to the extent possible. The latter method first computes the nearest neighborhood graph of data points. It then uses shortest path between vertices as a measure of distance between them. Finally it uses MDS to embed these distances in a low-dimensional space. After projecting the data to a lower dimension space in either MDS or Isomap, one can use linear correlation in the transformed subspace to measure similarity between cells. While *ACTION* is a non-parametric method, other methods need additional input. For SIMLR, we need to provide the true number of cell types. In order to give the other methods the best chance at competing with *ACTION*, we evaluate them using ten different values for 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 use each of the four measures to cluster cells and identify cell types. Each cluster is assumed to represent a unique cell



(a) Main flow of *ACTION* Similarity Metric

(b) Normalized Mutual Information

(c) Adjusted Rand Index

(d) Timing (s)

Figure 2: ***Evaluation of ACTION Similarity Metric.*** *(**a**) Workflow of ACTION metric. (**b**) Performance of ACTION in terms of Normalized Mutual Information (NMI). (**c**) Performance of ACTION with respect to Adjusted Rand Index (ARI) measure. (**d**) Overall running time of different methods, in log-scale. 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 kernels. Green cells show cases where a method performs better than others for a given dataset.*

type, and the clusters are determined using the commonly used kernel *k*-means algorithm. We compare the computed cell types with the true (known) cell types in terms of *Normalized Mutual Information (NMI)* and *Adjusted Rand Index (ARI)*. Normalized Mutual Information is an information theoretic measure that is zero for random clustering (when the identified clustering contains no information about true cell types), and one for a clustering that perfectly matches a given gold standard. The ARI measure is also between zero and one; however, it evaluates the cases in which a given pair of cells are either co-clustered in both true and identified, or classified separately in both.

In each case, we perform 100 independent clusterings with random initialization and report the average of NMI and ARI scores as quality measures (relative ordering of results is robust with respect to other aggregating functions, such as median or max). These experiments are independently performed for each dataset. Figures 2b-d present the performance of the cell type identification technique operating with different similarity measures, both in terms of their clustering quality (NMI and ARI) and total running time.

Discussion of results on similarity metric

To evaluate performance of each similarity metric, we analyzed four different datasets, which are listed in Section . These datasets have different number of cells, ranging from hundreds to thousands, span a wide range of normal and cancerous cells, and are measured using different single cell technologies.

For both *MouseBrain* and *Pollen* datasets, *ACTION* metric significantly outperforms other metrics in terms of both NMI and ARI measures. For the *Melanoma* dataset, *ACTION* has significantly better NMI, but there is a tie between *ACTION*, MDS, and SIMLR with respect to the ARI measure. Finally, for the *Immune* dataset, there is a tie between *ACTION*, MDS, and SIMLR for both measures. In all studies, t-test with p-val $\leq 10^{-2}$ has been used to assess significance of difference between observed NMI/ARI values. In summary, our results demonstrate that in all cases *ACTION* metric is either significantly better or at least as good as any other methods. Thus establishes 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. We note however, that our overall framework is flexible with respect to choice of other similarity metrics.

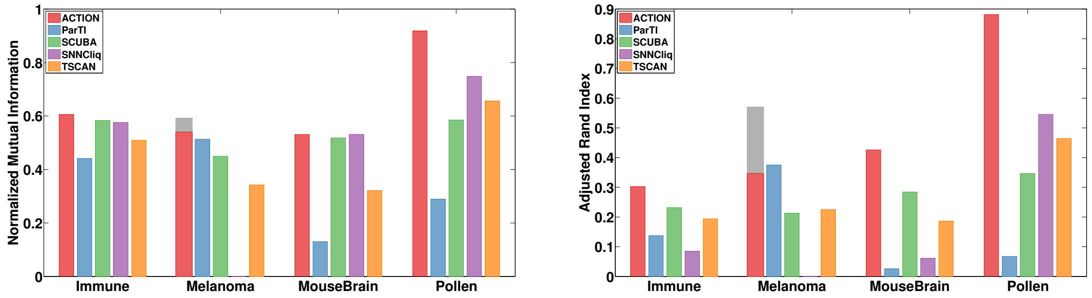
Component 2: A geometric view to identify discrete cell types

Novel methodology Using the *ACTION* metric as a measure of similarity between cells, we develop a new method for identifying *de novo* cell types in a given experiment. Our method is based on a geometric interpretation of cellular functions. Each cell is a data-point in a high-dimensional space. Our method identifies *"extreme"* corners in this space, and each cell is characterized by its distance to every corner. The corners identified by *ACTION* represent "pure" cells that are specialized to perform a principal function. This is in contrast to methods such as unsupervised clustering (e.g., *k*-medoids) that identify *the most common* centers. Our focus on identifying the extreme points (and thus, principal functions), allows us to better identify rare cell types.

Validation. Each corner or archetype represents a principal function. We first validate these by considering each archetype as a *characteristic cell type*. We then identify the type of each cell by determining the closest archetype and assigning this type. We compare our method to four recently proposed methods: SCUBA [7], SNNCliq [10], single-cell ParTI [11], [12], and TSCAN [13]. Details of these methods are given in the methods section. Clique size and density of quasi cliques of *SNN_Cliq* are left as default parameters ($k = 3$ and $r = 0.7$). Increasing clique size k did not improve performance, but significantly increased the running time. With these parameters, *SNNCliq* did not terminate in *72h* for the largest dataset (Melanoma), after which we stopped the experiment. We present a comprehensive analysis of the results for all other combinations of datasets/methods.

Discussion of results on cell-type identification. Figure 3 shows comparative performance of different methods in predicting cell types in various datasets. In all cases, except ARI for the *Melanoma* dataset, *ACTION* yields superior results compared to the state-of-the-art methods for cell-type identification. In general, NMI measure exhibits lower range of variation across methods, whereas ARI has a higher range of variability. To further investigate the difference between *ParTI* and *ACTION* on the Melanoma dataset, we manually evaluated each archetype identified in these methods. Our results indicate that the source of difference is that *ACTION* identifies more refined subtypes of T-cells and subclasses of tumor cells, whereas *ParTI* combines these subtypes/classes. These subgroup details are missing from the annotations provided for the dataset by authors. Combining cell types that are classified as different subtypes of T-cells or subclasses of tumor cells significantly enhances the computed performance measures of *ACTION* in this dataset. This is shown using gray boxes in the corresponding figure.

Analysis and validation of the principal functions. While cells can be classified based on their closest archetype, they can also be viewed on a continuum [12]. To illustrate this *continuous view*, we use the distance from each archetype as a low-dimensional embedding of



(a) NMI

(b) ARI

Figure 3: ***Performance of ACTION in identifying cell types.*** *Each cell type is identified by assigning it to its dominant function, represented by its closest principal function. (**a**) Normalized Mutual Information (NMI) of cell type identification. (**b**) Adjusted Rand Index (ARI) performance measure. In both (sub)figures, performance of SNNCliq for the Melanoma dataset is left blank, as it did not finish in the given time. NMI and ARI measures are computed against the true cell types from sample annotations. Gray bars for the Melanoma dataset show the difference in performance of ACTION after aggregating T-cell subtypes and tumor subclasses into a joint archetype.*

the cells. We use the Fielder embedding, followed by adjustment using Stochastic Neighbor Embedding (SNE) method to visualize this low-dimensional embedding in Figure 4. Each archetype is marked with a text labeled (A1, . . . , A11) point and assigned a unique color. Each point corresponds to a cell. We interpolate its color using its distance to all archetypes to highlight the continuous nature of the data. The labels for the groups are based on three sources. First, we perform enrichment analysis on the cells assigned to each archetype. Then, we use markers provided in the original datasets to identify the cell type-specific expression in each archetype. Finally, we use markers from *LM22* dataset [2] to classify subtypes of immune cells.

Figure 4 illustrates the ability of our method to identify both isolated cell-types with specialized principal functions, as well as cells with a combination of functions. As an example, different subclasses of T-cell constitute a spectrum with the corners (or archetypes) representing specialized functions that are performed by a pure T-cell subtype. In addition to given cell types, we also find an additional archetype, *A6*, which links between T-cells and B-cells and we hypothesize to be a lymphocyte progenitor.

In terms of tumor cells, many of the patients form their own archetypes. The two exceptions to this rule, A5 and A10, define a "*MITF axis*", which is shown in the subfigure (MITF is one of the transcription factors known be related to various types of Melanoma [6], [16]). Archetype *A5* is enriched in five patients with varying degrees of expression for MITF from mid to high. We collectively refer to patients in Archetype A5 as *MITF-associated* patients. Archetype *A10*, on the other hand, contains patients 81 and 82, both of who have low levels of MITF. In what follows, we construct the transcriptional regulatory network responsible for mediating observed phenotype of MITF-associated patients in A5.

Component 3: Constructing subclass-specific transcription regulatory network of MITF-associated patients

Novel Methodology We propose a new method to construct regulatory pathways responsible for mediating phenotypes associated with each archetype. To this end, we first perform an *archetype orthogonalization* (details described in Section), to compute residual expression and identify marker genes that are unique

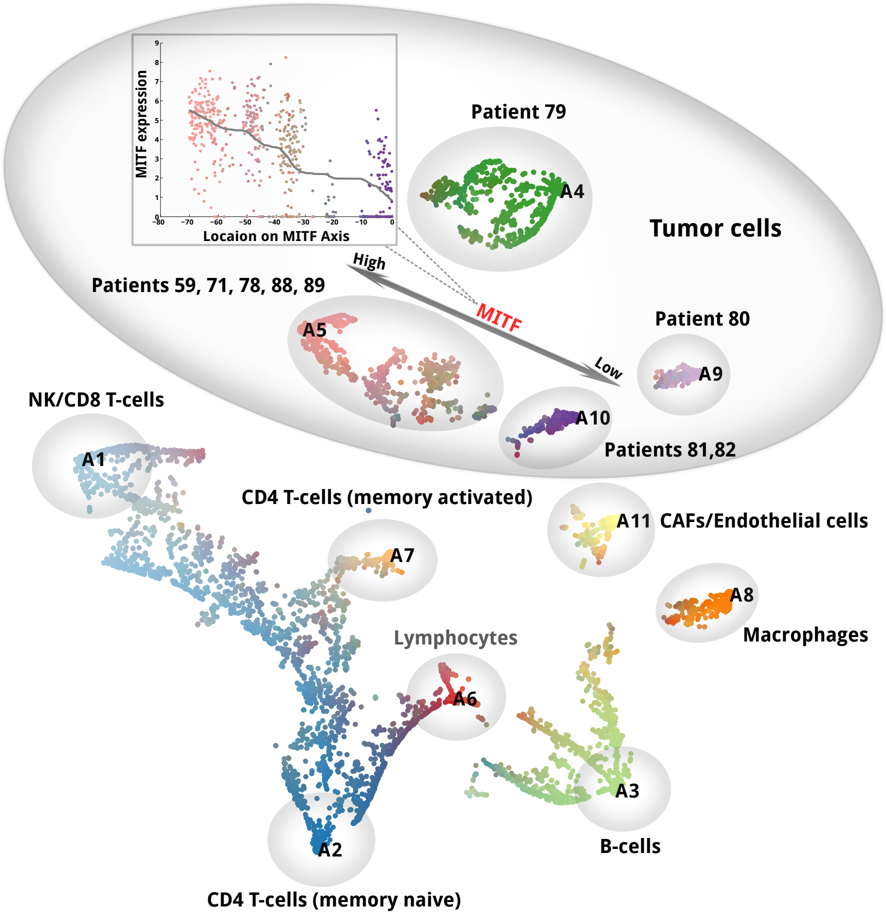


Figure 4: ***A continuous view of cell types in the Melanoma dataset identifies subclasses of immune cells and highlights a MITF-related "axis"*** *Each archetype, representing a principal function, is illustrated using a textual label. Each small dot represents a cell. Cells are color coded based on their proximity to archetypes. All data points are projected on to a 2D plane using Fielder embedding followed by Stochastic Neighbor Embedding (SNE). Each archetype is analyzed and annotated using three separate sources.*

to the archetype. Then, we rank all genes according to their *residual expression*. Finally, we project these scores to the transcriptional regulatory network (TRN) to find key transcription factors (TFs) responsible for mediating the observed transcriptional phenotype. For each TF, we assess the over-representation of its targets among top-ranked genes (according to the residual expression score). We use a dynamic programming algorithm [17] to assign exact p-values to each TF. For each TF, its "top ranked" target genes, according to the cut that yields the minimum hypergeometric score, are also selected as part of the regulatory network.

We apply this technique to identify regulatory pathways of MITF-associated samples. A p-value threshold of 0.05 is used to identify significant TFs. The final constructed network is presented in Figure 5. This network consists of six key transcription factors (in yellow), 85 target genes (in green/purple). Purple nodes are target genes that are jointly regulated by two TFs. We marked enriched functions of each group in the figure, accordingly, and highlighted elements that are already known to be associated with Melanoma.

Validation *MITF* is one of the best-characterized markers for Melanoma, and is also used in the original paper to classify patients [16]. It is notable here that our method identified MITF directly using data from the activity of its targets. Furthermore, since these transcription factors are identified based on the activity of their target, they are "related" to the subclasses, however, the mechanism of their control can be diverse.

Among other factors, *BHLHE40* has the highest number of activated targets. This factor, among other functions, regulates *M-MITF*, a melanocyte-restricted isoform of *MITF*, and potently reduces expression of *MITF* under hypotoxic conditions [18]. Angiogenesis, or growth of blood vessels, is a hallmark of cancer. *MEOX2* plays multiple roles in this process. At low levels, it activates nuclear factor-κB (NK-κB), a proangiogenic signaling pathway, whereas in high doses, it has an inhibitory role [19]. Similarly, *TSG101* plays different roles depending on the context. In fibroblasts, it acts as a tumor suppressor gene, whereas it has a tumor-enhancing role in some epithelial tumors. This bidirectional regulation is postulated to be through expression of *MMP-9*in different cell types [20]. The role of other factors is less-studied.

Experimental evidence To further validate our results, we use the transcriptome of 10 patients with *invasive* and *proliferative* melanoma subtypes from Verfaillie *et al.* [6]. *Proliferative* subtype is characterized by high levels of *MITF*, as well as *SOX10* and *PAX3*. In contract, *invasive* subtype is known to have low levels of *MITF* and high levels of epithelial-to-mesenchymal (EMT) transcription factor ZEB1, and is associated with metastatic dissemination. Nodes in our *MITF*-associtied TRN resemble the *proliferative* subtype. Thus, we use marker genes for this class to validate our results. There are a total of 770 marker genes for the *proliferative* subtype and among 91 total genes in our network, 8 genes coincide with them (p-value = 0.01). These genes include *DCT, MITF, PAX3, PPFIBP2, PRKCZ, TP53, TYR,* and *TYRP1*, all of which have high residual expression compared to all other nodes. Beside the *MITF* subnetwork, *TP53, PRKCZ,* and *PPFIBP2* are also enriched in this set. Interestingly, a key factor involved in the *invasive* subtype, *MEOX2*, is also identified as a node in our network. As mentioned earlier, depending on the level of its expression, this gene can play different roles for proliferative versus invasive subclasses.

Collectively, these results illustrate the effectiveness of the *ACTION* in identifying novel cancer subtypes, their underlying regulatory network, and characteristic markers. This, in turn, presents new avenues for diagnosis and prognosis of melanoma patients, as well as new therapeutic targets for further investigation.

Materials and methods

Datasets

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

***Immune* (from Supplementary Material)** : Comprehensive qPCR based assay of 1522 immune cells. This dataset spans 30 different types of stem, progenitor, and fully differentiated cells [21].

***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 [16].

***MouseBrain* (GEO: GSE60361)** : This dataset contains the expression profile of 3005 cells from the mouse cortex and hippocampus. These cells classify into 7 major types, including *astrocytes-ependymal, endothelial-mural, interneurons, microglia, oligodendrocytes, pyramidal CA1*, and *pyramidal SS* [8].

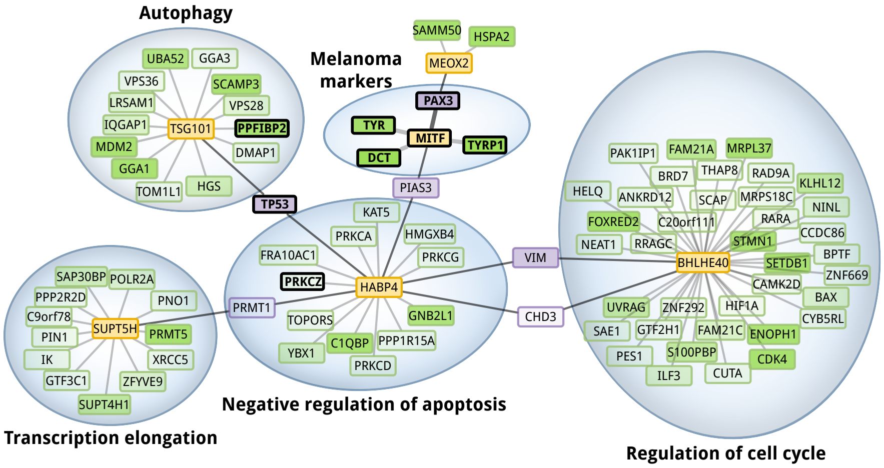


Figure 5: ***The transcriptional regulatory network (TRN) for MITF-associated Melanoma patients highlights a number of genes that have not previously been associated with Melanoma – along with some known markers.*** *The yellow nodes are Transcription Factors (TF), green nodes are Target Genes (TG), and purple nodes are target genes that bridge different TFs. Genes marked with black border are known to be involved in the proliferative subclass of Melanoma. Targets of each TF are used to annotate its dominant function, which are visualized using blue circles.*

***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 [22].

Immune subtype markers We collected immune cell markers for 22 subclasses from a recent paper [2]. This dataset contains a total of 547 markers, spanning 7 different T-cell subtypes, B-cells, NK cells, and myeloid derived subclasses. This dataset is collected and heavily curated from publicly available databases.

Transcriptional Regulatory Network (TRN) We collect transcription factor (TF) – target gene (TG) interactions from the RegNetwork database [23], which aggregates data from 25 different databases. This dataset contains a total of $151,214$ regulatory interactions between $1,408$ TFs and $20,230$ TGs.

Overview of prior methods for cell-type identification

Various methods have been developed to tackle the problem of cell type identification. **SNN-Cliq** [10] 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** [13] 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** [7] 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 [8] is based on 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 [9] 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 [24] 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.* [25] proposed to use diffusion maps to model the continuous spectrum of cells. On the other hand, Korem *et al.* [11], adopted a previously developed method, called **Pareto task inference (ParTI)** method [12], and applied it to single cell datasets.

Overview and justification for *ACTION*'s components

In the following sections, we describe various components of *ACTION*, as shown in Figure 1. We first explain exactly how the metric, illustrated in Figure 2(a), is computed from a matrix of raw cell expression profile data (Step 1 in the overview). Next, we explain how *ACTION* identifies the principal functions of a set of cells, assuming it knows the number of principal functions (Step 3 in the overview). We use an elbow method based on the quality of the principal functions to choose the actual number of principal functions (Step 2 in the overview). Finally, we explain how to estimate the transcriptional regulatory network for a specific principal function (Step 5) by orthogonalizing the functional space of cells (Step 4).

Step 1: A biologically-inspired metric for similarity of cells

Justification 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* are the subset of genes responsible for mediating core cellular functions, such as translation, transcription, and DNA repair. 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 fact that a housekeeping gene is expressed in a 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 group of cell types to perform cell type-specific functions. Unlike housekeeping genes, cell type-specific genes 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 housekeeping genes and enhances the signal contained in cell type-specific genes.

Suppressing housekeeping 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 [14]. The core of this method is to project a standardized representation of expression profiles of cells onto the orthogonal subspace of housekeeping genes. Let us denote given expression profiles 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 the shorthand $\boldsymbol{x}_i$ to denote the expression profile of i^{th} cell. In addition, let us denote the signature vector of housekeeping genes by $\boldsymbol{v}$. As a first order estimate, 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. Then, 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:

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

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 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, which in essence is similar to the one used previously for marker detection [27]. The 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 first shift all entries of $\mathcal{Z}^{\perp}$ by its minimum value to ensure positivity. Then, we normalize this shifted matrix to construct a new matrix $\mathbf{P}$, in which every row has sum one. Let $\boldsymbol{p}_j$ be the row vector associated with the jth gene. Then, we compute the entropy of $\boldsymbol{p}_j$ as: $H(j) = -\sum_j p_{ji} \log(p_{ji})$, where p_{ji} is an entry in the matrix $\mathbf{P}$. Finally, we use these entropy values as a basis to boost contributions from the most informative genes. To this end, we compute a scaling factor for each gene as follows. First, we partition genes as either informative or noninformative by finding the location of the most rapid shift in uniformity values, which resembles a L-shaped curve. Let us denote the entropy of the gene on the edge of this partition by H^*. Then for each gene j, we define a scaling factor as $s_j = H^*/H(j)$. Finally, we compute the kernel matrix as follows:

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

where function $\mathbf{diag}()$ creates a diagonal matrix from elements of a given vector, and each entry $w_i = s_j^2$. In this formulation, if we denote $\mathbf{Q} = \mathbf{diag}(\boldsymbol{c})\boldsymbol{Z}^{\perp}$, then $\mathbf{K} = \mathbf{Q}^T\mathbf{Q}$ defines a dot-product kernel.

Steps 2 and 3: A geometric approach to identify principal functions (representing pure cell types)

Transcriptional profiles of cells that perform multiple functions can be represented using a limited repertoire of principal functions. The functional space of cells, thus, can be represented by a low-dimensional geometric construct.

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. The functional space of cells that perform multiple functions can be represented using a limited repertoire of principal functions, which has recently been shown to be embedded within a reduced convex hull [12]. The corners, or archetypes, of this space represent principal functions, associated with specialized groups of cells. Identifying the enclosing 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. To this end, we first use the *successive projection algorithm (SPA)* to identify k transcriptional profiles as initial corners for the covering convex hull, each of which corresponds to a pure cell that is specialized to perform a set of unique principal functions. Then, we use principal convex hull algorithm (PCHA) combined with our distance kernel to adjust these corners by allowing others cells to contribute to the identity of each archetype/corner. This is combined with a standard model selection technique to estimate the number of principal functions.

A quick sketch of our procedure is as follows. We expand on this description in subsequent sections. For each $k = 1, \dots, K_{\max}$, (i) ***identify potential "pure" cells***: use SPA on the raw expression data $\mathbf{X}$ to find k pure cells that are near extreme points of the functional space; and (ii) ***adjust the corners***: initialize PCHA using the profiles of those k cells and iterate using the kernel $\mathcal{K}$. Then let $V(k)$ be the PCHA objective function with k archetypes. Finally after all models have been adjusted, (iii) ***estimate the number of cell types*** from $V(k)$ such that it balances the number of cells and the total explained variance.

Estimating "pure" cells as extreme corners of the functional subspace of cells

Given a raw expression matrix $\mathbf{X}$, we aim to identify an "optimal" set $\mathcal{S}$ of k "pure cells." These cells can be viewed as extreme corners of the convex hull of the functional space of cells, and all other samples can be written as convex combinations of these basis vectors. Under a strict assumption, known as *separability*, we seek to identify k columns such that $\mathbf{X} = \mathbf{X}(:, \mathcal{S})\mathbf{H}$, where $\mathcal{S}$ is the selected column subspace of matrix $\mathbf{X}$ and $\mathbf{H}$ is non-negative. This means that every column of $\mathbf{X}$ is a non-negative linear combination of a subset $\mathcal{S}$ of all columns. In terms of cells, this means that every cell's expression profile is a combination of a few cells. However, this is a very strong assumption that rarely holds in real data. A relaxation of this assumption, referred to as *near-separability*, seeks to estimate $\mathbf{X} \approx \mathbf{X}(:, \mathcal{S})\mathbf{H} + \mathbf{N}$, where the noise is bounded: $\|\mathbf{N}(:, j)\|_2 \le \varepsilon$. This decomposition is known as *near-separable Nonnegative Matrix Factorization (NMF)*. The *Successive Projection Algorithm (SPA)* is an efficient algorithm for solving near-separable NMF with provable performance guarantees [28]. If ε satisfies the technical condition $\varepsilon \le \mathcal{O}\left(\frac{\sigma_{min}(\mathbf{W})}{\sqrt{k}\kappa^2(\mathbf{W})}\right)$, then:

$$\min_{0 \le \mathbf{H}} \|\mathbf{X} - \mathbf{X}(:, \mathcal{S})\mathbf{H}\| \le \mathcal{O}\Big(\epsilon \kappa^2(\mathbf{W})\Big) \qquad (3)$$

More recently, other techniques have been developed to enhance the robustness of *SPA* to noise [29]. These methods are based on the fact that premultiplying matrix $\mathbf{X}$ by an orthogonal matrix $\mathbf{Q}$ preserves its separability. Thus, by carefully choosing matrix $\mathbf{Q}$, we can enhance the conditioning of the problem. Here, we use the *prewhitening technique*, which uses SVD decomposition of matrix $\mathbf{X}$ to estimate a noise-reduced approximation matrix. Algorithm 1 presents the SPA algorithm combined with prewhitening technique that we use to estimate a set of k cells.

Algorithm 1 SPA algorithm with prewhitening

```
Input: $\mathbf{X} \in \mathbb{R}^{m \times n}$: expression profile of cells
Output: $\mathcal{S}$: selected subset of columns in matrix $\mathbf{X}$
$[\mathbf{U}_k, \mathbf{\Sigma}_k, \mathbf{V}_k] = \mathbf{SVD}(\mathbf{X}, k)$
$\widetilde{\mathbf{X}} = \underbrace{\mathbf{\Sigma}_k^{-1}\mathbf{U}_k^T}_{\mathbf{Q}} \mathbf{X} = \mathbf{V}_k^T$ {Prewhitening}
$\mathcal{S} = \{\}, \mathbf{R} = \widetilde{\mathbf{X}} \Rightarrow$ Initialize
for $i = \{1, \cdots, k\}$ do
    $\alpha = \text{argmax}_j \|\boldsymbol{r}_j\|_2$ {$\boldsymbol{r}_j$ is the $j$th 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}
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{\boldsymbol{\beta}\}$
end for
```

Adjusting selected corners to allow contributions from all cells

Archetypal-analysis (AA) [30] can be viewed as a generalization of near-separable NMF. While in near-separable NMF all columns are represented using k columns in $\mathbf{X}$, in *AA* this constraint is relaxed to be a convex combination of all columns in $\mathbf{X}$. Formally, we can formulate *AA* as follows:

$$\begin{aligned} \underset{\mathbf{C}, \mathbf{H}, \boldsymbol{\alpha}}{\text{minimize}} \quad & \|\mathbf{X} - \mathbf{XCH}\| \\ \text{subject to} \quad & \|\mathbf{C}(:, i)\|_1 = 1. \\ & \|\mathbf{H}(:, i)\|_1 = 1. \\ & 0 \le \mathbf{C}, 0 \le \mathbf{H} \end{aligned} \qquad (4)$$

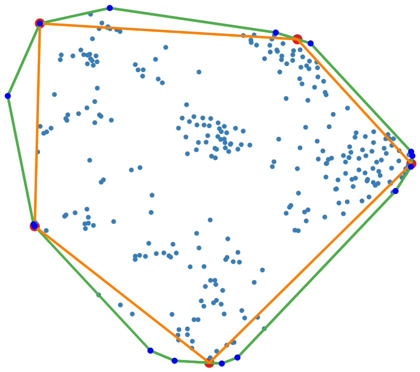


Figure 6: *Example of running PCHA algorithm. Light blue points are the data points, dark blue points are corners of the convex hull of data, the green lines represent the convex hull, and red points are the positions of archetypes selected by PCHA for k=5.*

Near-separable NMF is a special case of AA in which $\mathbf{C}$ has exactly k nonzeros and none of the columns have more than one element. The matrix $\mathbf{W} = \mathbf{XC}$ here stores the *archetypes*. Joint column stochasticity of $\mathbf{C}$ and $\mathbf{H}$ indicates that archetypes are convex combinations of data points, and each data point can be represented as convex combination of archetypes.

There is an algorithm, called *Principal Convex Hull Analysis (PCHA)*, to solve the above problem. The intuition behind PCHA is to fit a polytope to the data points, which approximates the optimal polytope containing as many data points as possible. Figure 6 illustrates this phenomena.

We use a kernelized version of PCHA algorithm that minimizes the objective:

$$\operatorname{trace}(-\mathbf{X}^T\mathbf{X}\mathbf{C}\mathbf{H} - \mathbf{H}^T\mathbf{C}^T\mathbf{X}^T\mathbf{X} + \mathbf{H}^T\mathbf{C}^T\mathbf{X}^T\mathbf{X}\mathbf{C}\mathbf{H}) \tag{5}$$

in which we directly provide the *ACTION* kernel $\mathcal{K}$ as $\mathbf{X}^T\mathbf{X}$ and initialize $\mathbf{C}$ based on the solution to SPA.

Estimating the total number of archetypes needed to represent all cell types

A key challenge in all parametric methods is to identify the optimal configuration for associated parameters. In our formulation, the total number of archetypes (corner points) must be provided by the user or directly estimated from the data. To automatically identify this number, one can use various measures of *"goodness"* to assess overall performance as we increase the number of archetypes. A balance between the number of archetypes and the goodness of solution provides an optimal compromise. We use variance explained by the fit as a measure to find the optimal number of archetypes. For each archetype count (up to a max value), we fit a convex hull to the data and compute explained variance.

The explained variance has an elbow-shape, meaning that it starts increasing rapidly, then it plateaus. The corner of this L-curve is an optimal choice for the number of archetypes. To find this point automatically, we fit a piecewise linear model to the data with two split points. This allows us to distinguish both rapid and more gradual shift patterns in the L-curve. Formally:

$$f(c) = \begin{cases} m_1c + b_1, & \text{for } 0 \leq c < c_i \\ m_2c + b_2, & \text{for } c_i \leq c < c_j \\ m_3c + b_3, & \text{for } c < c_j \leq c_{max} \end{cases} \tag{6}$$

where c is the archetype count and c_i and c_j are two free parameters. We evaluate every pair of $(c_i, c_j); 1 \leq c_i < c_j \leq c_{max}$ and fit a minimum least squares fit to each piece. The configuration with minimum overall error is selected as c_i^{best} and c_j^{best}. For this specific configuration, let m_2 and m_3 represent the slope of the second and the third linear fits. Then, if $\frac{m_2}{m_3}$ is less that or equal to a user-defined parameter threshold$_{\text{min}}$, then we select the first split point (c_i^{best}). Otherwise, we have a rapidly shifting curve and the slopes of second and third segments are very close. Thus, we select the second split point as the choice of k. Figure 7 illustrates an example of fitting process. The pink dots represent the explained variance for archetypal fits with increasing number of archetypes. Green lines show the piecewise linear fit to the data. The optimal number of archetypes is selected according to $best_j$ in this case, which is nine.

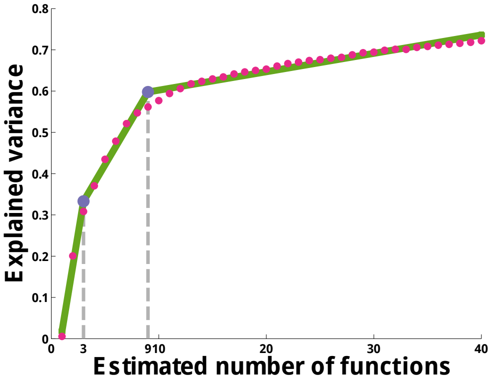


Figure 7: *Illustration of identification of total number of functions for the Pollen dataset*

Steps 4 and 5: Constructing the transcriptional regulatory network corresponding to each archetype

Each archetype represents a principal function performed by a group of cells. However, what makes these functions unique and the functional specializations they represent is not clear from the archetype signatures. To identify marker genes in each archetype, and to shed light on the underlying network regulating the observed transcriptional phenotype, we developed a novel approach based on orthogonalizing the space of principal functions.

Archetype orthogonalization to identify cell type-specific markers

A key factor in analyzing principal functions represented by each archetype is to identify what distinguishes one archetype from others. To identify shared and unique aspects represented by each archetype, we present a new method, called *arechetype orthogonalization*. The idea is to remove effects that are shared with any other archetypes before analyzing a given archetype.

Recall the result of PCHA is $\mathbf{C}$ and $\mathbf{H}$. The result $\mathbf{XC}$ represents the archetypes in the space of gene expression profiles. Let us denote the vector representation of archetype i by $\boldsymbol{a}_i$ and let $\mathbf{A}$ be the matrix of all archetypes. Let $\mathbf{A}_{-i}$ denote the matrix of archetypes without the ith column. Then our goal is to project $\boldsymbol{a}_i$ into the subspace orthogonal to the columns spanned by $\mathbf{A}_{-i}$. This can be computed as:

$$\boldsymbol{a}_i^{\perp} = \left(\mathbf{I} - \mathbf{A}_{-i}(\mathbf{A}_{-i}^T\mathbf{A}_{-i})^{-1}\mathbf{A}_{-i}^T\right)\boldsymbol{a}_i \qquad (7)$$

For each archetype, we can sort all genes according to their *"residual expression"* after orthogonalization.

Identifying cell type-specific transcriptional regulatory network (TRN)

Given *residual expression* vectors for each archetype, we can identify key regulatory circuits responsible for the observed transcriptional phenotype. We construct induced subgraphs of the global transcriptional regulatory network (TRN), which drive characteristic behavior of each cell type. First, we order all genes according to their residual expression for a given archetype. Then, for each transcription factor (TF), we identify the over-representation of its target genes (TGs) among top-ranked genes with respect to that archetype. To this end, we use minimum hypergeometric (mHG) p-value. This method is nonparametric, in the sense that we do not need to predefine a fixed cut. Let us represent the total number of genes by m. Given a set of target genes, of size T, we construct a binary vector of true positives (targets) as $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots \lambda_m] \in \{0,1\}^m$. Let the random variable Z denote the number of target genes among a fixed number of l top-ranked genes, if we distribute genes randomly. In this formulation, we can express the p-value in terms of the hypergeometric distribution:

$$\begin{aligned} p\text{-value}(Z = b_l(\lambda)) &= \text{Prob}(b_l(\lambda) \leq Z) \\ &= \text{HGT}(b_l(\lambda)|m, T, l) \\ &= \sum_{x=b_l(\lambda)}^{min(T,l)} \frac{\binom{T}{x}\binom{m-T}{l-x}}{\binom{m}{l}} \end{aligned} \tag{8}$$

where HGT is the tail of hypergeometric distribution and $b_l(\lambda) = \sum_{i=1}^{l} \lambda_i$ counts the total number of true positives in top-l observations. The drawback of this approach is that we still need a predefined cutoff value, l. To remedy this, Eden *et al.* [17] proposed a two-step process for computing the exact enrichment p-value, called *mHG p-value*, without the need for a predefined cutoff value of l. First, an optimal cutoff value is chosen among all possible values of $1 \leq l \leq N$. The computed value for this optimal cutoff is called the *minimum hypergeometric (mHG) score*, and is defined as:

$$\text{mHG}(\lambda) = \min_{1 \leq l \leq m} p\text{-value}(Z = b_l(\lambda)) \tag{9}$$

Next, a dynamic programming (DP) method is used to compute the exact p-value of the observed mHG score, in the state space of all possible λ vectors of size m having exactly T ones.

We use this formulation to identify significant transcription factors based on the number of target genes (TGs) with high residual expression. This, in turn, splits TGs of each TF into top vs bottom-ranked genes. We then select all significant TFs, together with their top-ranked target genes and construct a node-weighted induced subgraph of the global TRN, which represents the cell type-specific TRN.

Acknowledgements

This work is supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370, and by NSF awards BIO-1124962, CCF-1149756, IIS-1546488, and the DARPA SIMPLEX program. DG acknowledges partial support from the Sloan Foundation as well.

References

[1] C. Trapnell, “Defining cell types and states with single-cell genomics”, *Genome Research*, vol. 25, no. 10, pp. 1491–1498, 2015.

[2] A. M. Newman, C. L. Liu, M. R. Green, *et al.*, “Robust enumeration of cell subsets from tissue expression profiles.”, *Nature methods*, vol. 12, no. MAY 2014, pp. 1–10, 2015.

[3] A. Saadatpour, S. Lai, G. Guo, *et al.*, “Single-cell analysis in cancer genomics.”, *Trends in genetics : TIG*, vol. 31, no. 10, pp. 576–86, 2015.

[4] O. Stegle, S. A. Teichmann, and J. C. Marioni, “Computational and analytical challenges in single-cell transcriptomics”, *Nat Rev Genet*, vol. 16, no. 3, pp. 133–145, Mar. 2015.

[5] V. Plaks, C. D. Koopman, and Z. Werb, “Circulating tumor cells”, *Science*, vol. 341, no. 6151, pp. 1186–1188, 2013.

[6] A. Verfaillie, H. Imrichova, Z. K. Atak, *et al.*, “Decoding the regulatory landscape of melanoma reveals teads as regulators of the invasive cell state.”, *Nature communications*, vol. 6, p. 6683, 2015.

[7] E. Marco, R. L. Karp, G. Guo, *et al.*, “Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape”, *Proceedings of the National Academy of Sciences*, vol. 111, no. 52, E5643–E5650, 2014.

[8] A. Zeisel, A. B. M. Manchado, S. Codeluppi, *et al.*, “Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq”, *Science*, vol. 347, no. 6226, pp. 1138–42, 2015.

[9] D. Grün, A. Lyubimova, L. Kester, *et al.*, “Single-cell messenger rna sequencing reveals rare intestinal cell types.”, *Nature*, vol. 525, no. 7568, pp. 251–5, 2015.

[10] C. Xu and Z. Su, "Identification of cell types from single-cell transcriptomes using a novel clustering method", *Bioinformatics*, vol. 31, no. 12, pp. 1974–1980, 2015.

[11] Y. Korem, P. Szekely, Y. Hart, *et al.*, "Geometry of the gene expression space of individual cells", *PLOS Computational Biology*, vol. 11, no. 7, L. M. Iakoucheva, Ed., e1004224, 2015.

[12] Y. Hart, H. Sheftel, J. Hausser, *et al.*, "Inferring biological tasks using pareto analysis of high-dimensional data", *Nature Methods*, vol. 12, no. 3, pp. 233–235, 2015.

[13] Z. Ji and H. Ji, "Tscan: pseudo-time reconstruction and evaluation in single-cell rna-seq analysis", *Nucleic Acids Research*, vol. 44, no. 13, e117–e117, 2016.

[14] S. Mohammadi and A. Grama, "De novo identification of cell type hierarchy with application to compound marker detection", in *ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB)*, 2016.

[15] B. Wang, J. Zhu, E. Pierson, *et al.*, "Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning", Tech. Rep., 2016.

[16] I. Tirosh, B. Izar, S. M. Prakadan, *et al.*, "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq", *Science*, vol. 352, no. 6282, pp. 189–196, 2016.

[17] E. Eden, D. Lipson, S. Yogev, *et al.*, "Discovering motifs in ranked lists of dna sequences.", PhD thesis, Technion - Israel Institute of Technology, 2007, p. 77.

[18] E. Feige, S. Yokoyama, C. Levy, *et al.*, "Hypoxia-induced transcriptional repression of the melanoma-associated oncogene mitf.", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 43, E924–33, 2011.

[19] Y. Chen, A. B. Rabson, and D. H. Gorski, "Meox2 regulates nuclear factor-b activity in vascular endothelial cells through interactions with p65 and ib", *Cardiovascular Research*, vol. 87, no. 4, pp. 723–731, 2010.

[20] X. B. Sai, T. Makiyama, H. Sakane, *et al.*, "Tsg101, a tumor susceptibility gene, bidirectionally modulates cell invasion through regulating mmp-9 mrna expression", *BMC Cancer*, vol. 15, no. 1, p. 933, 2015.

[21] G. Guo, S. Luc, E. Marco, *et al.*, "Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire", *Cell Stem Cell*, vol. 13, no. 4, pp. 492–505, 2013.

[22] A. A. Pollen, T. J. Nowakowski, J. Shuga, *et al.*, "Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex", *Nature Biotechnology*, vol. 32, no. 10, pp. 1053–1058, 2014.

[23] Z.-P. Liu, C. Wu, H. Miao, *et al.*, "Regnetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse", *Database*, vol. 2015, bav095, 2015.

[24] L. Jiang, H. Chen, L. Pinello, *et al.*, "Giniclust: detecting rare cell types from single-cell gene expression data with gini index", *Genome Biology*, vol. 17, no. 1, p. 144, 2016.

[25] L. Haghverdi, F. Buettner, and F. J. Theis, "Diffusion maps for high-dimensional single-cell analysis of differentiation data", *Bioinformatics*, vol. 31, no. 18, pp. 2989–2998, 2015.

[26] E. Eisenberg and E. Y. Levanon, *Human housekeeping genes, revisited*, 2013.

[27] J. Schug, W.-P. Schuller, C. Kappen, *et al.*, "Promoter features related to tissue specificity as measured by shannon entropy.", *Genome biology*, vol. 6, no. 4, R33, 2005.

[28] N. Gillis and S. A. Vavasis, “Fast and robust recursive algorithmsfor separable nonnegative matrix factorization”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 4, pp. 698–714, 2014.

[29] ——, “Semidefinite programming based preconditioning for more robust near-separable nonnegative matrix factorization”, *SIAM Journal on Optimization*, vol. 25, no. 1, pp. 677–698, 2015.

[30] A. Cutler and L. Breiman, “Archetypal analysis”, *Technometrics*, vol. 36, no. 4, p. 338, 1994.