De novo identification of cell type hierarchy with application to compound marker detection

Shahin Mohammadi*
Department of Computer Science
Purdue University
West Lafayette, IN 47907
mohammadi@purdue.edu

Ananth Grama
Department of Computer Science
Purdue University
West Lafayette, IN 47907
ayg@cs.purdue.edu

ABSTRACT

Uncovering biochemical processes that drive the transformation of a totipotent cell into various cell types is essential to our understanding of living systems. This complex machinery determines how tissues differ in terms of their anatomy, physiology, morphology, and, more importantly, how various cellular control mechanisms contribute to the observed similarities/ differences. Recent single-cell experiments have shown that tissues/ cell types are related to each other through a complex hierarchical structure that blurs the defining lines of "cell type identity." Furthermore, seemingly identical cell types may exhibit different transcriptional characteristics, motivating new approaches to defining refined groupings of cells. These issues, in turn, strongly influence techniques for identification of key marker genes that characterize and/ or drive tissue-specific processes.

Motivated by emerging single-cell experiments, we propose a novel statistical approach and associated methods for cell-type classification and marker identification. We present a framework for identification and removal of the shared subspace (signature of housekeeping genes) of a given expression domain. We use this framework to reduce the effect of universally expressed genes and show that this reduction step enhances the signal-to-noise ratio (SNR) for known markers. We show that repeated application of subspace reduction within groups of cell types allows us to identify highly specific markers. Finally, we present a label propagation algorithm for automatically clustering cells and identifying putative groups of relevant cell types. We use the set of identified markers in each cluster to build tissue-specific transcriptional regulatory networks (tsTRNs), which are highly determinant of the cell type identity. Our framework allows one to start from a compendium of unknown cells, automatically identify groups of similar cell types, extract relevant markers, and construct accurate regulatory circuits for each group of putative cell types.

*Corresponding authors: mohammadi@purdue.edu

BCB '16 Oct 2–5, 2016, Seattle, WA, USA

DOI: 10.1145/1235

CCS Concepts

•**Mathematics of computing** → *Probabilistic inference problems;* •**Information systems** → *Clustering;* •**Theory of computation** → *Design and analysis of algorithms;*

Keywords

Cell type/tissue-specific expression; marker genes; subspace reduction; SVD; clustering; label propagation

1. INTRODUCTION

An embryonic stem cell encapsulates all of the genetic information needed to develop an individual; it differentiates into various cell types, which group together to shape tissues, combine to constitute organs, and assemble into organ systems. Various differentiated tissues/ cell types, while inheriting a similar genetic code, exhibit unique anatomical and physiological features. Traditionally, these cell types/ tissues have been classified using their high-level phenotypic characterizations, such as location and morphology. However, more recently, single-cell technologies have revealed an unprecedented heterogeneity among what were, until recently, believed to be identical cell types. This heterogeneity is achieved through systematic control of cellular machinery at different levels, including transcriptional, translational, and post-translational regulations, to orchestrate tissue-specific functions and dynamic responses to environmental stimuli.

Transcriptional regulation is among the best-studied aspects of this control. It is manifested in the observed differences in expression levels of genes across tissues. *Housekeeping genes* constitute the subset of the transcriptome that is universally expressed in human tissues. These genes are responsible for core cellular functions[1]–[3], and their corresponding pathways, are essential to all cells for their normal activity. However, they are not informative, with respect to the identity of cells, nor do they provide any power to classify cells into coherent groups of cell types. In contrast, certain genes are specifically or preferentially expressed in one, or a set of biologically relevant tissue types [2], [4]–[6]. These *marker genes* are critical for distinguishing various cells. In fact, *cell surface markers* have long been used to sort different subsets of immune cells. These genes play a crucial role in the physiology and the pathophysiology of human tissues. Many of the known disease genes are tissue-specific and are under/ over-expressed in the specific tissue(s) where the gene defect causes pathology [7], [8].

The use of transcriptomic profile of cells as a genome-

scale phenotype to identify their subtypes has attracted considerable attention. However, identifying transcriptionally-related cell types and their key marker genes remains a challenging task. One of the complicating factors in this paradigm is the hierarchical relationships among cell types. At the highest level, all cells are highly similar due to the expression of housekeeping genes. These genes are typically expressed at high levels and strongly impact the computed cell-cell distances (using any of the existing distance measures). After peeling this common layer, cell types split into groups with common functionality, which can be represented usng the *community affiliation graph* model [9]. Here, we can model common functionalities such as "affiliations", which are used to annotate cell types. However, these affiliations are not known a priori. Furthermore, cell types are not uniformly spaced and form a hierarchical structure linking them together. As we move deeper into this hierarchy, shared functionalities become more detailed, and distances among cell types reduce – necessitating use of rigorous statistical models and methods to assess the "proximity" of cell types.

In this paper, we propose an iterative, multi-step process to simultaneously identify groups of similar cell types as well as their characteristic marker genes that are specifically expressed within each group of cell types. The overall workflow of our method is shown in Figure 1. The two main operators in our framework are *subspace reduction*, in which we identify the unique signature of a given *expression domain*, and *clustering*, in which we group similar tissues/ cell types in the *reduced space* to define new *expression domains*. In this framework, we make an implicit assumption that genes do not work alone, but rather, as part of functional pathways. These pathways can be viewed as *barcodes* that uniquely identify their corresponding cell types/ tissues.

Motivated by these considerations, we develop a novel algorithm for *de novo* identification of cell types and their corresponding markers. We show that our subspace reduction step significantly enhances the signal-to-noise ratio (SNR) for markers, and that repeated application of reduction step within known groups of cells can identify their markers. Next, we show that, in the absence of known groupings, our method can automatically identify similar cell types using a clustering algorithm. We use this as a hierarchical prior for characterizing the expression domain of genes. Finally, we show that our method is able to reconstruct highly accurate models of tissue-specific transcriptional regulatory networks (tsTRN). Our framework is particularly well-suited for applications in single-cell analysis, in which the true identity of cell types, as well as their corresponding markers, is critical.

2. MATERIALS AND METHODS

2.1 Identifying the shared subspace among a group of tissues

A given set of tissues/ cell-types typically share a common set of genes/ pathways, while specializing through preferential genes that control and regulate this core shared set. We represent the *raw transcriptional signature* of these tissues using a matrix T, in which rows correspond to genes and columns correspond to various tissues. We are interested in finding the subspace of common genes, and to use it to adjust the transcriptional signatures. When the given set includes all, or majority of, human cell-types, the shared subspace represents the signature of housekeeping genes.

There are a number of methods for approximating this common signature in T, the simplest of which would be to compute the mean of its columns. An alternate approach involves decomposing T into sum of rank-one matrices, using methods such as *singular value decomposition (SVD)* or *non-negative matrix under-approximation (NMU)*. The general goal of these methods is to represent T as a sum of outer products of vectors. More formally, we write T as follows:

$$T = U_r \Sigma_r V_r = \sum_{i=1}^{r} \sigma_i u_i {v_i}^T, \tag{1}$$

where $r \leq \min(n_A, n_B)$ is the rank of the approximation. In the SVD formulation, u_i and v_i vectors are called left and right singular vectors, respectively. These vectors constitute an orthonormal basis, that is, both $u_i u_j^T = \delta_{ij}$ and $v_i v_j^T = \delta_{ij}$ for all i and j. Additionally, for any r, an SVD is the optimal rank-r approximation of T. When all entries of T are positive, Perron-Frobenius theorem ensures that all entries of the both left and right singular vectors are positive. However, the first residual matrix, $R_1 = M - \sigma_1 u_1 {v_1}^T$, can, and typically does, contain negative elements to ensure orthonormality. On the other hand, the *NMU* formulation does not ensure orthonormality, but, rather enforces an additional constraint on the optimization problem, which is that R_k should consist of only positive elements. Unfortunately, while SVD has an optimal solution, the additional non-negativity constraint of NMU makes its computation non-convex, though heuristics exist to approximate the solution.

In this work, we use a rank-one approximation of matrix T, that is $r = 1$, to identify a unique signature that closely represents the common signature in T. We use the first singular vector of matrix T, after z-score normalization, as a proxy for the housekeeping signature throughout our study.

2.2 Adjusting transcriptional signatures to control for the effect of shared subspace

Let us denote the transcriptional profile of the i^{th} tissue by $\mathbf{T}_i$. In order to compute the *raw transcriptional similarity* between each given pair of tissues, we compute the Pearson's correlation as follows:

$$r_{\mathbf{T}_i \mathbf{T}_j} = \frac{\sum_{k=1}^{n}(t_{ki} - \bar{t_i})(t_{kj} - \bar{t_j})}{\sqrt{\sum_{k=1}^{n}(t_{ki} - \bar{t_i})^2}\sqrt{\sum_{k=1}^{n}(t_{kj} - \bar{t_j})^2}} \tag{2}$$

where, n represents the total number of genes, t_{ki} and t_{kj} are the expression levels of the k^{th} genes in the i^{th} and j^{th} tissues, respectively. Similarly, $\bar{t_i}$ and $\bar{t_j}$ represent the average expression levels of genes in the corresponding tissues. Let $\mathbf{Z}_i$ denote the Z-score normalized version of $\mathbf{T}_i$, defined as $\frac{\mathbf{T}_i - \mu(\mathbf{T}_i)}{\sigma(\mathbf{T}_i)}$. We refer to $\mathbf{Z}_i$ as the *raw transcriptional signature* of tissue i. Using this formulation, we can simplify the raw transcriptional similarity as the normalized dot product of raw transcriptional signatures:

$$r_{\mathbf{T}_i \mathbf{T}_j} = \frac{\mathbf{Z}_i \mathbf{Z}_j}{n} \tag{3}$$

The raw transcriptional similarity of tissues is artificially inflated due to the ubiquitous expression of housekeeping genes across all tissues. To control for this effect, we first de-

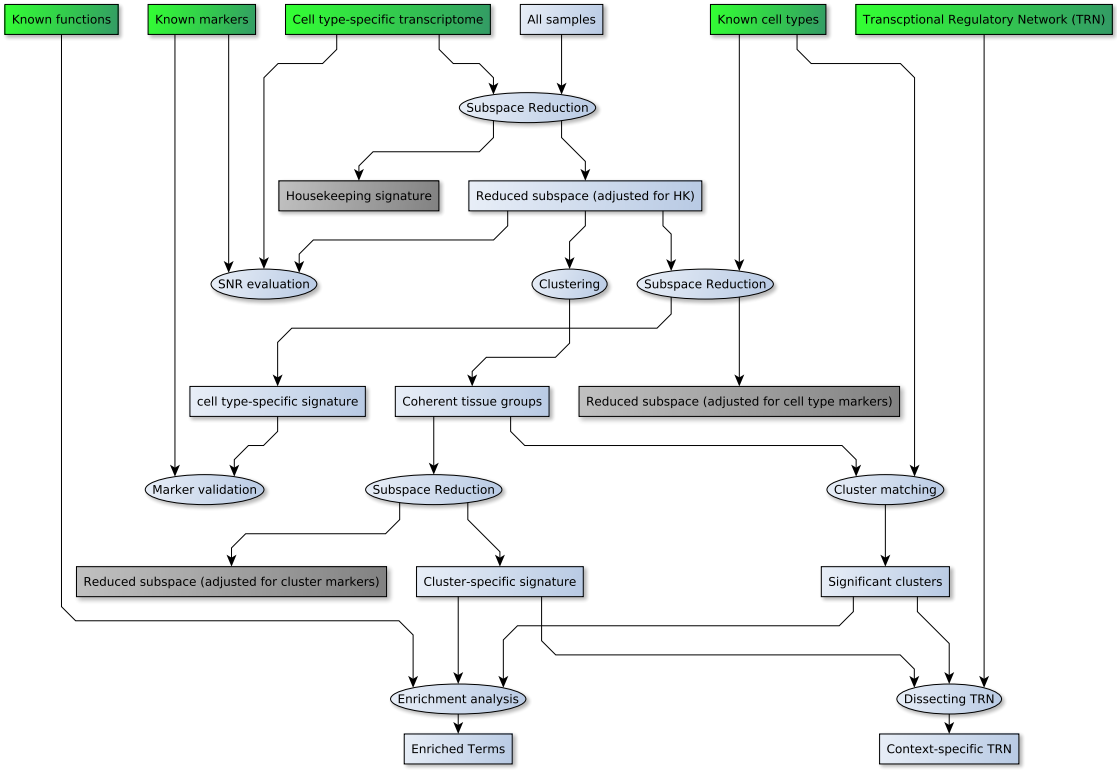


Figure 1: Overall workflow of our experiment

fine *housekeeping transcriptional signature*, denoted by vector $\mathbf{S}$, as the left singular vector of matrix $\mathbf{Z}$. Using this notation, we revise our similarity scores by computing the partial Pearson's correlation between $\mathbf{T_i}$ and $\mathbf{T_i}$, after controlling for the effect of $\mathbf{S}$ as follows:

$$r_{\mathbf{T}_i\mathbf{T}_j \bullet \mathbf{S}} = \frac{r_{\mathbf{T}_i\mathbf{T}_j} - r_{\mathbf{T}_i\mathbf{S}} r_{\mathbf{T}_j\mathbf{S}}}{\sqrt{1 - r^2_{\mathbf{T}_i\mathbf{S}}}\sqrt{1 - r^2_{\mathbf{T}_j\mathbf{S}}}} \tag{4}$$

As before, we can rewrite this using the Z-score formulation. Let us denote the adjusted transcriptional profile of tissue i as $\mathbf{Y}_i = \mathbf{Z}_i - r_{\mathbf{T}_i\mathbf{S}}\mathbf{S}$. We define the *adjusted transcriptional signature* of tissue i as $\hat{Z}_i = \frac{\mathbf{Y}_i - \mu(\mathbf{Y}_i)}{\sigma(\mathbf{Y}_i)}$. Finally, we can rewrite the adjusted transcriptional similarity as:

$$r_{\mathbf{T}_i\mathbf{T}_j \bullet \mathbf{S}} = \frac{\hat{\mathbf{Z}}_i \hat{\mathbf{Z}}_j}{n} \tag{5}$$

We use this approach to remove the shared subspace of a given set of expression profiles, and to construct the corresponding adjusted transcriptional signatures. Significantly positive transcriptional similarities in this framework are indicators of shared tissue-specific pathways. We use these adjusted transcriptional signatures in our study to identify marker genes. However, when applying methods that rely on the positivity of the input expression matrix, one can use the sigmoid transform of these scores as follows:

$$\hat{p}_{ki} = \frac{1}{1 + e^{-\hat{z}_{ki}}} \tag{6}$$

Please note that this transformation, when applied to the raw transcriptional signatures, is equivalent to the previously known softmax normalization:

$$\begin{aligned} p_{ki} &= \frac{1}{1 + e^{-z_{ki}}} \\ &= \frac{1}{1 + e^{-(\frac{t_{ki} - \mu(\mathbf{T}_i)}{\sigma(\mathbf{T}_i)})}} \end{aligned} \tag{7}$$

This normalization is known to remove the effect of outliers, while preserving a linear relationship for mid-range values.

2.3 Computing Signal-to-Noise Ratio (SNR)

Signal-to-Noise Ratio (SNR) is a commonly used measure for evaluating the quality of a desired *signal* by comparing the power of the signal to the power of (undesired) *noise*. We define the desired signal as the expression of marker genes in their corresponding tissue/ cell type of origin. Similarly, we define noise as the expression of the rest of the genes in that cell type. Lets assume there are k replicas of a given tissue/ cell type, and a total of n genes, represented in a matrix $\mathbf{A} \in \mathbb{R}^{n \times k}$. We are also given a subset $\mathcal{S}$ of rows that are designated as markers. We compute the power of signal as: $P_{signal} = \frac{\|A(\mathcal{S},:)\|_F^2}{|\mathcal{S}|}$. The numerator can be also expressed as $\| \mathbf{vec}(A(\mathcal{S},:)) \|_2^2$, where the **vec** operator vectorizes a matrix by stacking up its columns. Similarly, we can compute the power of noise as: $P_{signal} = \frac{\|A(\mathcal{S}',:)\|_F^2}{|\mathcal{S}'|}$, where $\mathcal{S}' = \{1..n\} \setminus \mathcal{S}$. Then, we can compute SNR as:

$$SNR = 10log_{10}(\frac{P_{signal}}{P_{noise}}), \quad (8)$$

which is in unit of decibels (dB).

2.4 Assessing the significance of of marker detection methods

We use the *hypergeometric* p-*value* as a statistical measure of the overlap among sets. A typical use case for this formulation is in over-representation analysis (ORA). The classical approach to this problem is to select a predefined cutoff l to identify top-ranked genes, and then to compute the enrichment p-value using the hypergeometric distribution. Let us denote the total number of gene products by N. Given a set of known gene annotations (true positives) of size A, we encode these annotations using a binary vector $\lambda = \lambda_1, \lambda_2, ...\lambda_N \in \{0,1\}^N$. Let the random variable T denote the number of positive genes in the target set, if we distribute genes randomly. In this formulation, the hypergeometric p-value is defined as:

$$\begin{aligned} p\text{-value}(T = b_l(\lambda)) &= Prob(b_l(\lambda) \leq T) \\ &= HGT(b_l(\lambda)|N, A, l) \\ &= \sum_{t=b_l(\lambda)}^{min(A,l)} \frac{C(A,t)C(N-A, l-t)}{C(N,l)} \end{aligned} \quad (9)$$

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 need a predefined cutoff value, l. To remedy this, Eden et al.[10] propose 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:

$$mHG(\lambda) = min_{1 \leq l \leq N} HGT(b_l(\lambda)|N, A, l) \quad (10)$$

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 N having exactly A ones (please refer to Eden et al.[10] for algorithmic details, and Eden[11] for an efficient implementation).

2.5 Combining individual p-values to compute a meta p-value

For cases in which we compute individual p-values for each tissue/ cell type, we need to combine them in order to define a meta p-value that can be used to assess each selection. To combine a set of computed p-values, we use the Fisher's method [12]. This method computes a statistic $\mathcal{S} = -2\sum_{i=1}^{k} ln(p_i)$ for a set of k given *p-values* p_i. Then, we can use χ^2 test with $2k$ degrees of freedom to assess the significance of the meta-analysis, assuming that p_is are independent.

3. RESULTS AND DISCUSSION

cell line origin	number of markers
brain	336
colo-rectal	72
kidney	158
liver	185
lung	56
ovary	35
stomach	77
urinary bladder	27
skin	133

Table 2: Number of markers for cancer cell lines dataset

In this section, we validate the following hypotheses: (i) a single level of adjustment (reducing the effect of housekeeping genes) enhances the signal-to-noise ratio (SNR); (ii) repeated application of the reduction process over groups of cell types allows us to recover cell type-specific markers; (iii) automatic identification of *putative cell types* using label propagation based clustering yields reliable grouping of cell types; and (iv) cluster-specific, adjusted signatures yield highly accurate models of cell type/ tissue-specific transcriptional regulatory circuits. All of these hypothesis are validated using known cell type groupings and markers.

3.1 Datasets

3.1.1 Gene expression profiles

In our experiments, we use two separate datasets derived from different technologies. The first dataset, which we will loosely refer to as *immune cell types*, is the expression profile of 38 distinct subpopulations of hematopoietic cells measured using Affymetrix GeneChip microarray [13]. This dataset consists of the gene expression of $12,074$ genes in a total of 211 samples. The second dataset contains a comprehensive compendium of 675 cancer cell lines [14]. The origin of these cell lines can be classified into 17 different tissues. We will focus on these 17 distinct groups, but collectively refer to this dataset as the *cancer cell lines* dataset.

3.1.2 Gold standard marker genes

To evaluate identified markers and the impact of adjustment, we collect marker genes from two independent studies. For immune cell types, we adopt the LM22 dataset from Newman *et al.* [15]. First, for each cell type in LM22, we identify genes that are highly expressed. Then, we compute the mean expression of these markers in each of the immune cell types in our dataset. We construct a weighted bipartite graph between cell types in these two datasets and identify matches using a maximum-weight bipartite matching algorithm [16], followed by manual assessment. Table 1 shows the final results for the immune marker set.

For the cancer cell lines dataset, we download the gold standard tissue-specific markers from the Human Protein Atlas (HPA) [17]. We manually matched ten different tissues of origin to the markers in HPA, and limited our focus on the markers that have both transcriptomic and proteomic evidence. Among these ten, *pancreas* markers were not significantly expressed in the pancreas-originated cell lines, and thus we removed this from our set. The final set consists of nine tissues, shown in Table 2.

LM22 cell type	mapped cell type	number of markers
B cells naive	Naive B-cells	118
B cells memory	Mature B-cell class able to switch	106
Plasma cells	Mature B-cells	109
T cells CD8	CD8+ Effector Memory	142
T cells CD4 naive	Naive CD4+ T-cell	121
T cells CD4 memory activated	CD4+ Effector Memory	107
NK cells activated	Mature NK cell_CD56+ CD16+ CD3-	109
Monocytes	Monocyte	104
Dendritic cells activated	Myeloid Dendritic Cell	121
Eosinophils	Eosinophill	159
Neutrophils	Granulocyte (Neutrophilic Metamyelocyte)	140

Table 1: Statistics of the immune cell type markers

3.1.3 Transcriptional regulatory network (TRN)

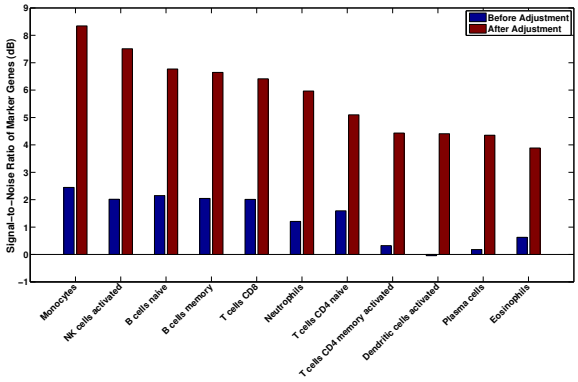
We collected transcription factor (TF) – target gene (TG) interactions from the RegNetwork database [18], which aggregates data from 25 different databases. This dataset contains a total of 151,214 regulatory interactions between 1,408 TFs and 20,230 TGs.

3.2 Adjusting for the effect of housekeeping genes enhances signal-to-noise ratio (SNR) for the known marker genes

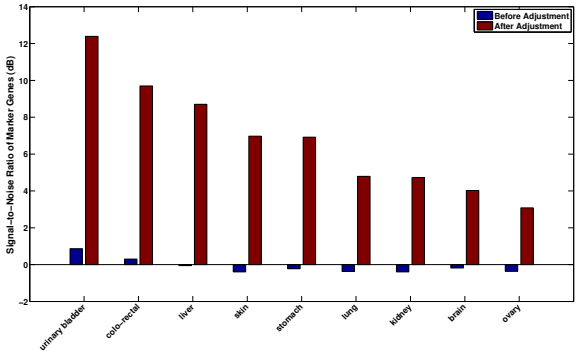
We hypothesize that the global expression of housekeeping genes, which are universally expressed genes that perform core cellular functions, masks the true signal from tissue-specific markers. Thus, adjusting for this common signature should enhance the signal-to-noise ratio (SNR) of marker detection methods. To systematically evaluate our hypothesis, we compute SNR using Equation 8, for immune cells and cancer cell line markers, respectively. The results of this adjustment are presented in Figure 2. In the two cases shown, we observe a significant improvement over the raw expressions. However, we note that in the cancer cell line dataset, there are cases in which the power of non-marker genes is stronger than the power of marker genes, thus the negative dB values. This effect is remedied in all cases after adjustment, which suggests that the proposed adjustment process deflates housekeeping gene expression effectively, but does not negatively influence the expression of marker genes.

3.3 Iterative application of adjustment process identifies markers that are comparable or better than the t-test

In this experiment, we quantify the extent to which highly expressed genes in the adjusted profile can be used to identify tissue/ cell type-specific markers. We apply the same adjustment process to each group of cells/tissues, after adjusting for housekeeping effect. The result is a single shared signature for each group of cell types/ tissues. We rank genes according to their expression level in this signature and assess the over-representation of known markers among the higher ranked elements in this list. We use mHG p-values, introduced in Section 2.4, to assess the significance of each case. Similarly, we compute the mHG p-values for results of one-sided and two-sided t-tests, which correspond to the most commonly used methods for identifying differentially expressed genes. Our final results are presented in Figure 3. For the immune cell dataset, the iterative adjustment process yields superior results in every single case. However,



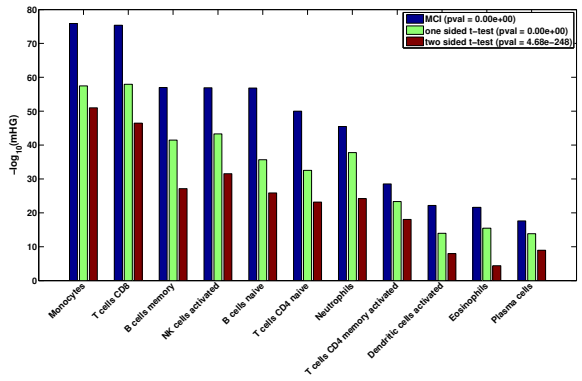
(a) Immune cell types



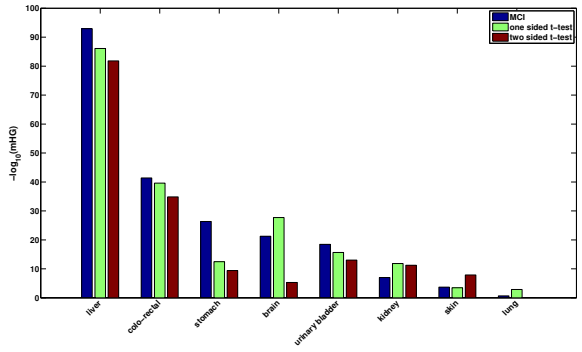
(b) Cancer cell lines

Figure 2: SNR enhancement for marker genes after the adjustment process

in the cancer cell line datasets, the results are more varied. In this case, we removed *ovary* from our study, since none of the methods had significant p-values. To systematically evaluate different methods, we use Fisher's method [12] to combine individual p-value into a *meta* p-value, the details of which are presented in Section 2.5. This results in the combined p-values of 2.5×10^{-197}, 2×10^{-185}, and 8.1×10^{-150}, for our method, one-sided t-test, and two-sided t-test, respectively. In summary, in both cases our method significantly outperforms the standard t-test but, *more importantly*, as we show in the next section, it does not depend on a predefined grouping and can automatically identify relevant expression domains.



(a) Immune cell types

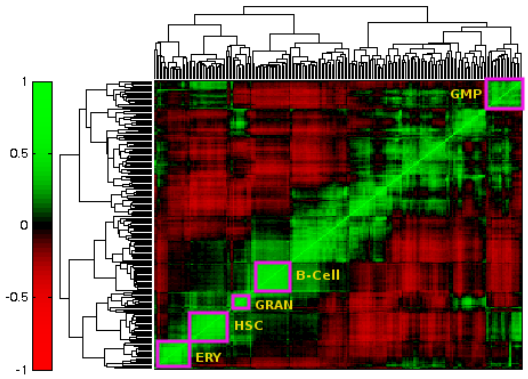


(b) Cancer cell lines

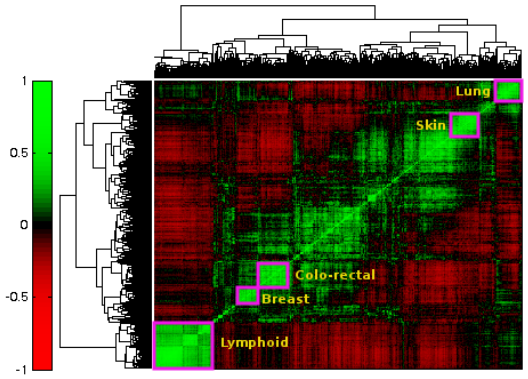
Figure 3: Significance of marker predictions using two-step adjustment process compared to the standard t-test. Most Common Identifier (MCI): shared subspace of each tissue/cell type after adjustment for HK genes.

3.4 Adjusted cell type signatures identify groups of similar tissues/ cell types

Having established that iterative application of the adjustment process can identify marker genes within given groups of cells, we now study whether these groups can be identified from the data directly. Given a compendium of cells, this would allow us to automatically identify major subgroups corresponding to cell types, as well as key marker genes associated with each group. In order to evaluate if such structure exists in the adjusted data, we perform bi-clustering on the similarity matrix between tissues/ cell types. We compute similarities using Pearson's correlation, after adjusting expression profiles for the effect of housekeeping genes. Figure 4 shows the clustered heat-map of samples in each of our datasets. Each coherent group of samples is marked according to the majority of cell types/ tissues in the group. For the immune cell dataset, *B-cell (mature)* and *hematopoietic stem cell (HSC)* are two of the largest coherent groups, followed by *erythrocyte (ERY), granulocyte/monocyte progenitor (GMP)*, and *granulocyte (GRAN)*. In the cancer cell line dataset, *lymphoid* tissues comprise the largest coherent group, followed by *lung, skin, colo-rectal*, and *breast* groups. These groups are the major *separable* clusters at the first level of hierarchy. The size of each group is related to the total number of samples for that tissue/ cell type, whereas consistency within the group is related to the homogeneity of cell types. For example, *lymphoid* tissues exhibit three separate subtypes in the heat-map, which correspond to *bone marrow, lymph node* and *blood*.



(a) Immune cell types



(b) Cancer cell lines

Figure 4: Heatmap of tissue/cell type similarities after the adjustment process for housekeeping genes

We use a recent method proposed by Gaiteri [19] to automatically identify these *separable* groups. This algorithm is a modification of the label propagation clustering that corrects for the global frequency of labels, which in turn allows it to identify more refined clusters. We compute similarity matrices before and after adjustment, and remove all negative entries after computing the correlation scores. For the immune cell types dataset, the normalized mutual information (NMI), a supervised measure of clustering quality given a ground truth, was 0.53 and 0.62 before and after adjustment, respectively. Similarly, NMI for the cancer cell line dataset was 0.37 and 0.47 before and after adjustment. This shows that the adjustment process enhances the quality of clustering in both cases. Next, we match each identified cluster to known groups in each dataset. We first construct a weighted bipartite graph between clusters on one side and

known groups of cell types on the other, by assigning a hypergeometric p-value to the size of their overlap. We then use a maximum-weight bipartite matching algorithm [16] to compute the best match for each cluster. We rank each tissue/ cell type according to the best matched cluster, i.e., how well identified clusters capture each group. Table 3 summarizes the set of tissues/ cell types in each dataset best matched to the identified set of clusters. Interestingly, all major separable groups in the cancer cell line dataset are captured by at least one cluster. In addition, brain and pancreatic tissues both have a corresponding cluster, even though in the heat-map, they were not distinguishable from the rest of tissues. On the other hand, for immune cell types, clusters cover a majority of separable cell types, with the exception of *GMP*, which is a heterogeneous group by itself consisting of a group of progenitor cells in the myeloid branch. Memory T-cells are also strongly connected in the heat-map, but are split into different groups, with the groups themselves being fairly homogeneous. We note that, in general, known tissues in the cancer cell lines dataset are better represented by their clusters than the immune cell types, in terms of their overlap p-value. We hypothesize that this phenomena is due to higher underlying similarity among immune cell types that is not separable using only one level of clustering.

In summary, label propagation applied to the similarity scores after adjustment for housekeeping genes can automatically identify groups of cell types/ tissues with coherent functions/ expression. These groups can be used as a hierarchical prior to define the expression domain of tissue/ cell type-specific genes and their corresponding pathways, as we demonstrate in the next section.

(a) Immune cell types

Celltype	$-log_{10}(p-val)$
Hematopoietic stem cell	8.36
Erythroid	7.63
Mature B-cell class able to switch	4.90
CD4+ Central Memory	4.34
Granulocyte (Neutrophil)	3.88
Basophils	3.60

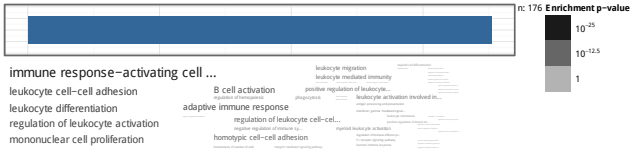
(b) Cancer cell lines

Cellline origin	$-log_{10}(p-val)$
lymphoid	120.02
skin	51.45
breast	38.86
colo-rectal	38.43
brain	13.17
lung	11.85
pancreas	11.75

Table 3: Significance of matching known groups to the identified clusters

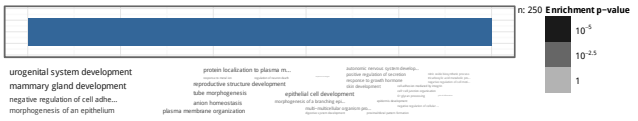
3.5 Putting the pieces together: automated identification of cell types and their characteristic markers

We have, thus far, shown that adjusted transcriptional signatures are capable of identifying highly accurate cell-type markers. Furthermore, being accurate representations of cell type/ tissue-specific functionality, these signatures are better suited to quantifying cell type-cell type and tissue-tissue similarities. These similarities, in turn, can be used to identify coherent groups of cell types/ tissues. Here, we show that highly expressed genes within identified clusters are enriched with tissue/ cell type-specific pathways. We select the top three clusters that correspond to top three best-covered tissues of origin in the cell lines dataset as our test cases. First, we apply the adjustment process over each cluster, instead of known groups. We then filter each cluster signature vector to select anything above z-score threshold of 1.96. We use these three genesets and performed GO enrichment analysis over each one of them using the *GOsummaries* package in R/Bioconductor [20]. This package relies on the g:Profiler [21] package to identify and summarize enriched terms using their hierarchical relationships, but also generates a word cloud of the final, simplified results. Figure 5 shows the enrichment for the top three clusters in the cancer cell line dataset, which are *lymphoid, skin,* and *breast*, respectively. We note that the annotations of each cluster are consistent with the matched pair of known groups. Furthermore, each cluster is highly enriched with respect to related tissue-specific functions. This validates the fact that the grouping/ marker detection process is able to automatically identify cell types/ tissues, and to identify highly specific markers.



(a) Lymphoid

(b) Skin



(c) Breast

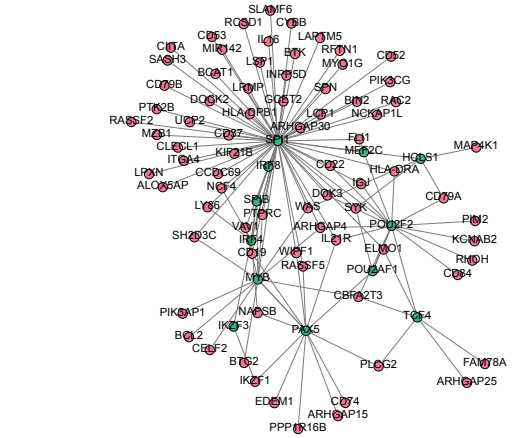
Figure 5: Enrichment of top-ranked genes in the top three clusters in cancer cell lines dataset

3.6 Adjusted signatures predict tissue-specific transcriptional regulatory networks

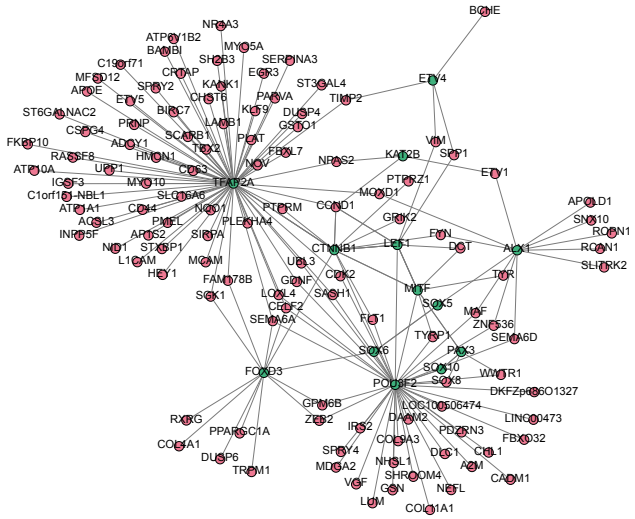
Tissue/ cell type-specific transcription factors (tsTFs) are significantly implicated in various human disorders [22], [23], including cancers [24]. Having established that adjusted signatures can be used to identify marker genes from among identified clusters, we now construct core regulatory networks responsible, in each tissue, for defining its identity. We focus on the same set of tissues as in Section 3.5. For each tissue, we first identify the set of transcription factors that are *highly expressed*, specifically in that tissue. We then

assign a p-value to each of these TFs by looking at their target genes. We identify how many total targets each TF has, how many of them are expressed (above z-score of 1.96), and how many total genes are expressed in the adjusted signature. Using these statistics, we compute the p-value of tissue-specificity for each selected TF using the tail of hypergeometric distribution. A TF is deemed signficant in a given tissue if it is specifically expressed highly in that tissue, after the iterative adjustment process, and has a significantly large number of targets that are also highly expressed. We identify a minimal set of 12, 14 and 8 TFs for *lymphoid, skin,* and *breast* tissues, respectively. We then construct the tissue-specific transcriptional regulatory network (tsTRN) as the bipartite graph consisting of the selected TFs, together with their highly expressed gene targets. For breast, *GRHL1* just has a self-loop, whereas, in skin network, *NDN* TF is only connected to *NGFR*. We exclude these two TFs from further study. Figure 6 shows three networks corresponding to the tsTRN of these tissues.

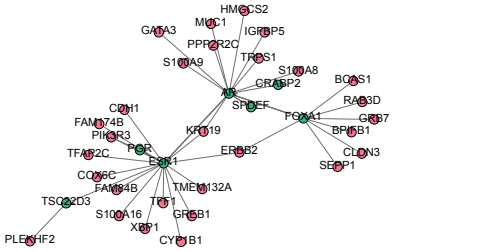
Functional enrichment analysis of identified TFs shows a very significant and very relevant set of functions. *Myb* is a known proto-oncogene and its over-expression plays a key role in development of chronic B-lymphocytic leukemia (B-CLL) [25]. On the other hand, *POU2F2, SPI1, MEF2C, MYB, IRF4, IRF8, IKZF3*, and *HCLS1* are all involved in the *hematopoiesis* (GO:0030097 p-val $= 3.6 \times 10^{-9}$). Among these genes, *SPI1* has the highest connectivity in the constructed lymphoid-specific TRN (Figure 6(a)). This TF regulates gene expression during myeloid and B-lymphoid cell development. In skin-specific TRN (Figure 6(b)), *TFAP2A* has the highest connectivity, but *CTNNB1* has a higher centrality. Interestingly, a subset of TFs in this network, *LEF1,CTNNB1*, and *ALX1*, are known to be involved in the *positive regulation of epithelial to mesenchymal transition* (p-val $= 3.2 \times 10^{-6}$). This suggests that the skin-specific network can be used to identify new targets for trans-differentiation. Finally, breast-specific TRN is centered around *Estrogen Receptor 1 (ESR1)*, *Androgen Receptor (AR)*, and *Forkhead Box A1 (FOXA1)* TFs. These TFs, together with *progesterone receptor (PGR)*, constitute the core of the *steroid hormone mediated signaling pathway* (p-val $= 9.4 \times 10^{-7}$), and essential for sexual development and reproductive function. In summary, these tsTRNs, identified automatically from the given cell type/ tissue-specific transcriptome, capture highly relevant functionalities that are fundamental to the core identity of each cell type. We conclude that, our framework can identify hypothesized groups of related cells, identify their common markers, and construct the underlying circuits that regulate the context-specific machinery.



(a) Lymphoid



(b) Skin



(c) Breast

Figure 6: Tissue-specific transcriptional regulatory network (tsTRN) of top 3 clusters identified in the cancer cell lines dataset

4. CONCLUDING REMARKS

Cell-type identities have traditionally been determined by their high-level characteristics, such as their location and morphology. However, recent advances in single cell analysis have uncovered significant biological variability, even within known cell types. This, in turn, provides both opportunities and challenges for redefining cell-type identity using its transcriptomics profile as the primary phenotype. However, this approach is complicated by the multi-resolution nature of groupings among cell types, in which the outer layers have a stronger signal, masking more specific, detailed signals shared between selected groups of cells.

We present a novel statistical framework for constructing a refined hierarchical prior of tissue similarities to organize them into a well-separated, reduced subspace. This organization is constructed from the raw expression profiles by first deflating the shared functionality attributed to housekeeping genes. We then identify groups of similar tissues in this adjusted space, and further contract each tissue group by iterative application of our SVD-based method to identify the common signature of each cluster. We use clus-

ter signatures to identify cell type/ tissue-specific markers that uniquely contribute to the identity of each putative cell type. Finally, we show through concrete examples that these markers are regulated by a highly specialized tissue-specific transcriptional regulatory network (tsTRN), which can greatly enhance our understanding of cellular differentiation.

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 Grant BIO 1124962.

References

[1] Z. Dezso, Y. Nikolsky, E. Sviridov, *et al.*, "A comprehensive functional analysis of tissue specificity of human gene expression.," *BMC biology*, vol. 6, p. 49, Jan. 2008.

[2] C.-W. Chang, W.-C. Cheng, C.-R. Chen, *et al.*, "Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis.," *PloS one*, vol. 6, no. 7, e22859, Jan. 2011.

[3] O. Souiai, E. Becker, C. Prieto, *et al.*, "Functional integrative levels in the human interactome recapitulate organ organization.," *PloS one*, vol. 6, no. 7, e22051, Jan. 2011.

[4] L. Wang, A. K. Srivastava, and C. E. Schwartz, "Microarray data integration for genome-wide analysis of human tissue-selective gene expression.," *BMC genomics*, vol. 11 Suppl 2, S15, Jan. 2010.

[5] F. M. G. Cavalli, R. Bourgon, W. Huber, *et al.*, "SpeCond: a method to detect condition-specific gene expression.," *Genome biology*, vol. 12, no. 10, R101, Jan. 2011.

[6] S. Teng, J. Y. Yang, and L. Wang, "Genome-wide prediction and analysis of human tissue-selective genes using microarray expression data.," *BMC medical genomics*, vol. 6 Suppl 1, S10, Jan. 2013.

[7] K.-I. Goh, M. E. Cusick, D. Valle, *et al.*, "The human disease network.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 21, pp. 8685–90, May 2007.

[8] K. Lage, N. T. Hansen, E. O. Karlberg, *et al.*, "A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 52, pp. 20 870–5, Dec. 2008.

[9] J. Yang and J. Leskovec, "Community-affiliation graph model for overlapping network community detection," in *12th IEEE International Conference on Data Mining, ICDM 2012, Brussels, Belgium, December 10-13, 2012*, 2012, pp. 1170–1175.

[10] E. Eden, D. Lipson, S. Yogev, *et al.*, "Discovering motifs in ranked lists of DNA sequences.," *PLoS computational biology*, vol. 3, no. 3, e39, Mar. 2007.

[11] E. Eden, "Discovering Motifs in Ranked Lists of DNA Sequences," PhD thesis, Technion - Israel Institute of Technology, 2007, p. 77.

[12] R. A. Fisher, *Statistical Methods for Research Workers*, ser. Cosmo study guides. Cosmo Publications, 1925.

[13] N. Novershtern, A. Subramanian, L. N. Lawton, *et al.*, "Densely interconnected transcriptional circuits control cell states in human hematopoiesis.," *Cell*, vol. 144, no. 2, pp. 296–309, 2011.

[14] C. Klijn, S. Durinck, E. W. Stawiski, *et al.*, "A comprehensive transcriptional portrait of human cancer cell lines," *Nature Biotechnology*, vol. 33, no. 3, pp. 306–312, 2014.

[15] A. M. Newman, C. L. Liu, M. R. Green, *et al.*, "Robust enumeration of cell subsets from tissue expression profiles," *Nature Methods*, no. MAY 2014, pp. 1–10, 2015.

[16] H. W. Kuhn and B. Yaw, "The hungarian method for the assignment problem," *Naval Res. Logist. Quart*, pp. 83–97, 1955.

[17] M. Uhlén, L. Fagerberg, B. M. Hallström, *et al.*, "Tissue-based map of the human proteome," 2015.

[18] 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.

[19] C. Gaiteri, M. Chen, B. Szymanski, *et al.*, "Identifying robust communities and multi-community nodes by combining top-down and bottom-up approaches to clustering," *Scientific Reports*, vol. 5, p. 16 361, 2015.

[20] R. Kolde, *Gosummaries: word cloud summaries of go enrichment analysis*, R package version 2.0.0, 2014.

[21] J. Reimand, T. Arak, and J. Vilo, "g:Profiler–a web server for functional interpretation of gene lists (2011 update).," *Nucleic acids research*, vol. 39, no. Web Server issue, W307–15, Jul. 2011.

[22] T. Raj, K. Rothamel, S. Mostafavi, *et al.*, "Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes.," *Science (New York, N.Y.)*, vol. 344, no. 6183, pp. 519–23, 2014.

[23] D. N. Messina, J. Glasscock, W. Gish, *et al.*, "An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression.," *Genome research*, vol. 14, no. 10B, pp. 2041–7, 2004.

[24] J. M. Vaquerizas, S. K. Kummerfeld, S. a. Teichmann, *et al.*, "A census of human transcription factors: function, expression and evolution.," *Nature reviews. Genetics*, vol. 10, no. 4, pp. 252–263, 2009.

[25] K. Vargova, N. Curik, P. Burda, *et al.*, "MYB transcriptionally regulates the miR-155 host gene in chronic lymphocytic leukemia.," *Blood*, vol. 117, no. 14, pp. 3816–25, 2011.