Bioinformatics, 31(10), 2015, 1584–1591
doi: 10.1093/bioinformatics/btv015
Advance Access Publication Date: 11 January 2015
Original Paper

OXFORD

Gene expression

CellCODE: a robust latent variable approach to differential expression analysis for heterogeneous cell populations

Maria Chikina[1,*], Elena Zaslavsky[2] and Stuart C. Sealfon[2]

[1]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15217, USA and [2]Department of Neurology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on August 25, 2014; revised on December 15, 2014; accepted on January 7, 2015

Abstract

Motivation: Identifying alterations in gene expression associated with different clinical states is important for the study of human biology. However, clinical samples used in gene expression studies are often derived from heterogeneous mixtures with variable cell-type composition, complicating statistical analysis. Considerable effort has been devoted to modeling sample heterogeneity, and presently, there are many methods that can estimate cell proportions or pure cell-type expression from mixture data. However, there is no method that comprehensively addresses mixture analysis in the context of differential expression without relying on additional proportion information, which can be inaccurate and is frequently unavailable.

Results: In this study, we consider a clinically relevant situation where neither accurate proportion estimates nor pure cell expression is of direct interest, but where we are rather interested in detecting and interpreting relevant differential expression in mixture samples. We develop a method, Cell-type COmputational Differential Estimation (CellCODE), that addresses the specific statistical question directly, without requiring a physical model for mixture components. Our approach is based on latent variable analysis and is computationally transparent; it requires no additional experimental data, yet outperforms existing methods that use independent proportion measurements. CellCODE has few parameters that are robust and easy to interpret. The method can be used to track changes in proportion, improve power to detect differential expression and assign the differentially expressed genes to the correct cell type.

Availability and implementation: The CellCODE R package can be downloaded at http://www.pitt.edu/∼mchikina/CellCODE/ or installed from the GitHub repository 'mchikina/CellCODE'.

Contact: mchikina@pitt.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Differential expression analyses are used widely in the study of human biology, but their utility is often limited by the extreme variability (and the resulting poor reproducibility) of human molecular measurements. One biological source of measurement variance is heterogeneity in sample composition. Human samples are often mixtures of multiple cell types with relative proportions that can vary several fold across samples. For example, in diseased brain, cell populations can change markedly, as some cell types die, whereas others proliferate (Kuhn *et al.*, 2011). In cancer samples, there may be different amounts of stromal tissue and different sub-populations of cancer cells that are molecularly distinct (Schwartz and Shackney,

2010; Yoshihara *et al.*, 2013). Cell-type variation is particularly pronounced in blood, where proportions of different cell types can vary 4-fold naturally (Adalsteinsson *et al.*, 2012; Shen-Orr *et al.*, 2010). Consequently, it is well established that mRNA and protein measurements from human blood, or blood derivatives such as peripheral blood mononuclear cells (PBMCs), are extremely variable.

Much effort has been devoted to the development of computational methods that can accurately model and analyze mixture samples. Existing approaches can computationally estimate proportions, pure expression states or both [see Gaujoux and Seoighe (2013) and Shen-Orr and Gaujoux (2013) for comprehensive reviews]. The methods range from simple matrix decomposition to complex iterative procedures. Importantly, these methods are largely focused on estimating physical quantities (pure expression states and proportions) rather than statistical ones (such as effects and interactions), and most are not designed for differential expression analysis. A recent R package unifying many of the existent methods lists only two (DSection and csSAM) that can work as differential expression pipelines, and both require independent cell proportion measurements as input (Gaujoux and Seoighe, 2013).

Shen-Orr *et al.* had showed that csSAM can be effective at addressing two questions. It can be used both to find differentially expressed (DE) genes, when standard statistical methods fail, and to assign the DE genes to the cell type in which their regulation has been altered. The latter is an important goal as disease-associated expression changes are expected to be cell-type specific and knowing which cells are affected by the disease state is important for interpreting the results.

In this study, we demonstrate that these two goals, improving the power to detect DE genes and assigning altered expression to the cell type of origin, can be effectively approached with latent variable analysis. Our Cell-type COmputational Differential Estimation (CellCODE) method is designed for DE analysis and requires no additional dataset-specific knowledge. In particular, we demonstrate that by relying on the data structure alone, we can correct for mixture variation and improve statistical power as well as or better than methods that use explicit cell proportion measurements. We also show that the CellCODE framework can be used to assign expression alterations to their cell type of origin with high accuracy. Our method is widely applicable and we demonstrate that it can be used to derive new insights from existing data.

2 Results

The biological complexity of quantitative measurements derived from mixture samples raises many challenges. The variation of mixture components from sample to sample induces large variance in gene expression measurements, making it difficult to detect relevant gene regulation. On the other hand, the cell-type proportions themselves may be different between the clinical groups, giving rise to many DE genes that do not correspond to any actual transcriptional regulation. Finally, for genes that are altered within a cell-type the source of transcriptional regulation is ambiguous, as many cell types are assayed together.

We propose a multi-step statistical framework that uses latent variable analysis to analyze differential expression from mixture samples. We first estimate a set of surrogate proportion variables (SPVs) by cross-referencing putative marker genes with the data correlation structure. These SPVs are then included in the differential expression analysis to improve the detection of bona fide regulated genes. Finally, the DE genes are assigned to the cell type in which they are regulated, evaluating interaction between genes and estimated SPVs.

2.1 CellCODE SPVs

To resolve the relative contributions of proportion changes and gene regulation in a statistical framework, our approach relies on estimating the relative differences in cell proportion (but not the actual numerical fractions) directly from molecular expression measurements. Our approach relies on using external reference datasets to determine which genes are likely to track cell-type abundance, i.e. marker genes, which has been successful previously (Kuhn *et al.*, 2011; Repsilber *et al.*, 2010). However, obtaining a set of reliable markers is not always a trivial task. In blood-derived mixtures, many genes express in more than one cell type, and even canonical surface proteins do not always provide good markers at the mRNA level (e.g. the CD4 marker of a T-cell subtype is not T-cell specific).

In general, exactly which genes reliably track which cell type is a product of a number of interacting factors (such as marker and isoform specificity, technical platform aspects, similarities among cell types and absolute cell-type abundance), which vary from dataset to dataset. To address this, we propose a data-dependent approach that finds marker genes that are reliable for the specific dataset being analyzed. Specifically, we use putative marker genes to guide the decomposition of the dataset structure into separate variance components that track mixture proportions. Our methodology is based on using singular value decomposition (SVD) to combine marker genes, a technique known as an eigengene summarization (Langfelder and Horvath, 2008), which effectively negates the contribution of any inconsistent marker assignments. Unlike previous approaches that rely on a few robust marker genes, our method has the advantage of using a larger list of putative marker genes, while requiring them to be correct only on average. This allows a permissive approach to marker selection, while improving the robustness of the result.

The CellCODE pipeline includes a visualization step that can be used to explore the structure of marker expression and ensure accurate estimation. After regressing out any known clinical variables, we compute the correlation coefficients of all marker genes of interest along with our computed SPVs. We expect that markers for the same cell type should correlate with each other, whereas markers for different cell types should be largely uncorrelated, resulting in a block-like correlation structure that is captured by the CellCODE SPVs. This visualization step also ensures that we have selected a set of cell types that can be reliably tracked as separate components by their marker sets. Some cell proportions may be fundamentally unmodelable, for example, if they are too rare or co-vary with another variable. Markers for such cell types would not form distinct correlated blocks.

We illustrate this approach by extracting SPVs from the Shen-Orr dataset using markers derived from the IRIS (Immune Response *In Silico*) (Abbas *et al.*, 2009) and DMAP (Differentiation Map) reference datasets (Novershtern *et al.*, 2011). As expected, we find that the correlation structure is block-like, although not all of the cell markers for an individual cell-type cluster together (Fig. 1). In fact, a number of selected markers cluster with genes from the wrong cell type, highlighting the potential inconsistencies that arise from differences in platforms and experimental conditions. Importantly, however, the CellCODE SPVs (denoted in black) always cluster with the majority of the cell-type markers. The CellCODE eigengene-based approach extracts the consensus correlation signature and is therefore robust to outliers.

We emphasize that because they are eigenvectors of an SVD, SPVs do not directly quantify cell-type proportions; in particular, SPVs will take on negative values. Nevertheless, SPV values should reflect the relative differences in cell-type composition. As such,

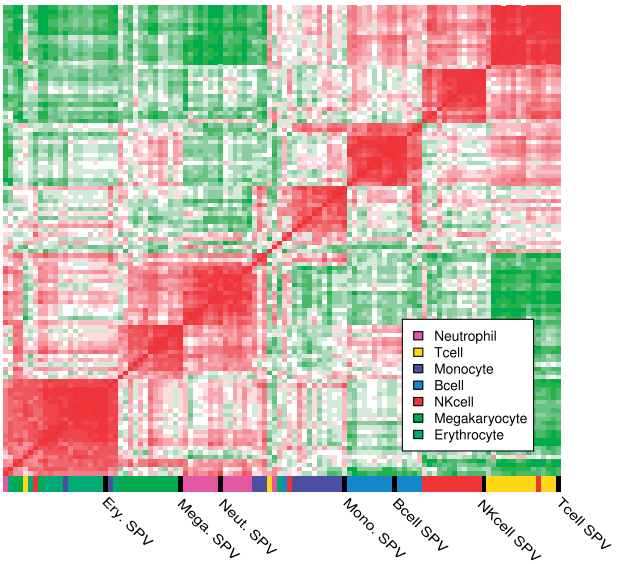

Fig. 1. Evaluating consistency of surrogate proportion estimates in the Shen-Orr dataset. The heatmap represents correlation coefficients between all pairs of marker genes with red (darker) representing high correlation and green (lighter) representing anti-correlation. Marker genes initially selected for a specific cell type are indicated by colors as shown in the key. The CellCODE SPVs (indicated by black) are also included. The heatmap is clustered with 1-ρ as a distance metric. Despite some apparent inconsistencies in marker assignments, distinct clusters of high correlation emerge for each cell type, and each SPV reliably associates with the correct cluster (Color version of this figure is available at *Bioinformatics* online.)

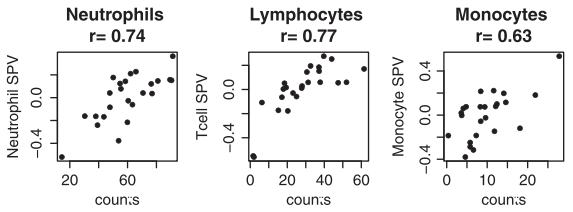

Fig. 2. CellCODE SPVs track Coulter counter measurements

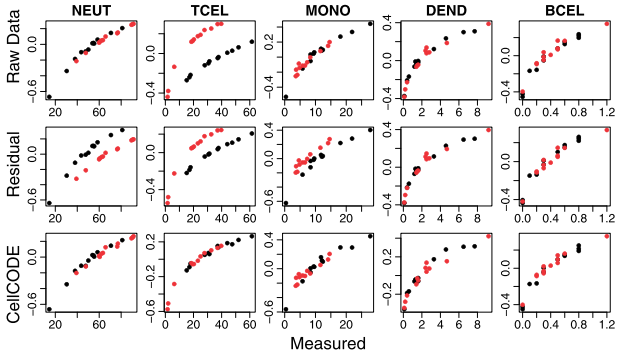

Fig. 3. Recovery of cell proportions from simulated expression data. The SPVs recovered using three different approaches are plotted against the true proportions used for simulation (x axis). We simulated two clinical groups plotted in red (grey) and black with global proportion differences (in neutrophils and T cells) and true transcriptional differences coming from the T-cell population. We specifically enforced that 30% of the T-cell markers are DE. Generally, all estimates track known proportions, even for very rare cell types. The relationship is non-linear due to log transformation of expression values. However, aside from providing high correlation, the ideal estimation procedure should be unbiased, resulting in red and black points falling on a single curve. Computing eigengenes with the raw expression values (first row) or expression normalized across clinical groups (second row) leads to biased proportion estimates. CellCODE (third row) is able to provide accurate estimates that track global proportion changes while being agnostic to transcriptional alterations within individual cell types (Color version of this figure is available at *Bioinformatics* online.)

SPVs should be well-correlated with true cell-type composition. We test this by making use of the Coulter counter measurements associated with the Shen-Orr dataset. Because Coulter counter analysis separates cells based on size, thus combining some molecularly distinct cell types, there is no direct correspondence between the cell types reported in this dataset and those for which CellCODE computes corresponding SPVs. Despite these limitations, when we plot the Coulter counter measurements for the three most abundant cell types against the relevant SPVs, we find that the two measures show good correlation (Fig. 2). Our technique also captures variation in other cell types, such as NK cells, B cells and non leukocyte cells that were not resolved by Coulter counter analysis.

Ideally, SPVs, which are eigengene-based summaries of cell-type marker genes, should not just correlate with true proportions, but should also be *consistent*, i.e. produce the same values for samples with the same cell-type proportions, independently of other sources of expression variation. This goal presents a potential challenge if marker genes themselves are subject to regulation within the cell type they represent.

To overcome these limitations, CellCODE employs a modification of the eigengene approach. Specifically, we only compute an eigengene summary for those genes that are estimated to not be regulated at the individual cell-type level. Our approach is related to the 'surrogate variable analysis' (SVA) two-step algorithm (Leek and Storey, 2007) (see Section 4 for details). As shown below, even in the presence of proportion differences between clinical groups and differential expression of 30% of marker genes, CellCODE was able to produce proportion estimates that are unbiased.

To explore this effect in a controlled manner, we formulate a pipeline for simulating a realistic mixture dataset. This pipeline relies on the Coulter counter cell-type proportions determined in Shen-Orr *et al.* and on experimentally determined pure-cell expression vectors from the IRIS study (see Section 4). We simulate two clinical groups, which differ in their proportion distributions and in marker gene expression at the individual cell-type level. We extract SPVs from the resulting data and plot them against true proportions for each clinical group. If the SPVs are indeed consistent, the plotted points should fall on a single functional curve, i.e. be agnostic to the clinical group. We find that while taking the naive eigengene approach or normalizing across groups produces bias, the CellCODE method is able to map proportions independently of clinical groups (Fig. 3).

2.2 CellCODE improves differential expression discovery

Analyzing differential expression in samples composed of diverse cell populations is a two-fold challenge. On the one hand, variation in mixture components increases measurement variance, thus reducing the power to detect small expression changes. On the other hand, when individual differences in cell proportions are asymmetrically distributed among the clinical groups, standard methodologies are prone to picking up false positives (genes whose expression values are altered, but that are not regulated on an individual cell-type level).

To investigate how the CellCODE approach can be harnessed to improve discovery of transcriptionally regulated genes, we employ

our simulation approach described above to create datasets with both cell-type proportion changes and individual cell-type expression changes. We simulate cell-type-specific expression differences occurring in different cell types, ranging from very frequent to very rare.

We begin by examining the performance of a simple *T*-test on this simulated dataset (Fig. 4). The number of altered genes in our simulation is a constant 10%, and the fold-change in expression is drawn from the same normal distribution (and thus the magnitude of change has constant expected value). Nonetheless, recovery of the DE genes by the *T*-test varies considerably with the relative abundance of the cell type involved. DE genes are easiest to recover when they are altered in a common cell type, such as neutrophils, and are very difficult to detect when the change is coming from a cell type that represents only 1% of the population, such as dendritic cells. As can be seen in Figure 4 for any method, the percentage of DE genes recovered is much lower for rare cell types. Because most genes are expressed in multiple cell types, any differential expression in rare cell types has a small effect on the overall expression level.

Though in mixture datasets many expression changes may be small in magnitude, we can improve their detection by accounting for the variance induced by cell-type proportion variation. The simplest statistical approach is to include the confounding variables as covariates, generating a modified estimate for the group effect. Indeed, as expected, including the same cell proportions that were used to simulate the dataset as covariates in our *T*-test (*T*-test + measured, Fig. 4) improved the detection of DE genes.

An alternative deconvolution approach to mixture data was suggested by Shen-Orr *et al.* This method uses independent proportion measurements to extract pure cell profiles from the mixture expression matrix. This approach produces a different *T* statistic for each cell type and a renormalized summary *T* statistic (where the deconvolved pure expression vectors are recombined in standard proportions). The method is equivalent to fitting interaction models without an intercept, which is a theoretically correct model of mixture data, but requires estimating more coefficients. In our simulation, neither the summary *T* statistic nor the cell-type-specific interaction coefficient, perform particularly well, and neither improves on the raw *T*-test (Fig. 4).

Finally, we consider the CellCODE approach, which neither requires nor utilizes independent proportion measurements. Instead, we first estimate the SPVs from marker genes and then use these estimates as covariates in our differential expression analysis. Even though CellCODE does not require independent knowledge of cell-type proportions, it still outperforms other methods at detecting differential expression. It may be surprising that plugging in CellCODE SPVs (which are accurate but not perfect) leads to better results than using the true proportions directly. This phenomenon results from the fact that the CellCODE SPVs directly model the existing data structure and consequently result in a better fit that explains more variance.

2.3 Assigning DE genes to cell type of origin

We have shown that the CellCODE method can be used to improve the detection of DE genes from mixture samples. We next examine how the same framework can be extended to determine which cell type is responsible for the differential expression.

Although our analysis suggests that interaction models are not well suited for improving the power to detect differential expression, they are useful for attributing the DE genes to their cell type of origin. Therefore, we investigated a sequential approach in which the first step employs CellCODE to detect DE genes and the second step assigns the DE genes to the cell type in which the genes are regulated.

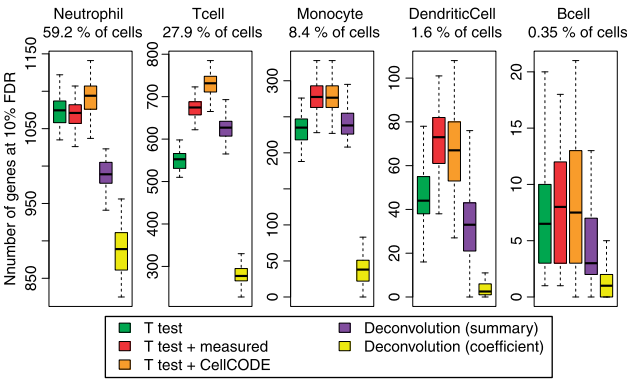

Fig. 4. Increasing differential expression detection power in mixture datasets. Mixture datasets were simulated by combining pure cell expression in different proportions. For half of the 24 samples simulated, one pure cell expression vector was altered to have 10% DE genes. Cell-type origin of differential expression was varied to create a range of simulated datasets. Each resulting dataset was ranked for differential expression using different methods, and the number of genes identified with a false discovery rate of 0.1 is shown as a boxplot distribution over 20 repeats of the simulation. The CellCODE method, which uses only the data structure, outperforms methods that use known cell proportions (Color version of this figure is available at *Bioinformatics* online.)

To evaluate assignment accuracy, we first select DE genes using the CellCODE approach, and then, in a subsequent step, test the accuracy of different methods at assigning those genes to a cell type. We first tested the total deconvolution method described by Shen-Orr *et al.* by assigning each DE gene to the cell type with maximal cell-type-specific *T* statistic. We find that once we separate the task of finding DE genes and assigning them to a cell type, the deconvolution method is effective for the second step. This method is able to correctly determine the cell type of origin for the majority of the detectable DE genes that are regulated in frequent and rare cell types (Fig. 5). The drawback of this method for our purposes is that it requires accurate independent knowledge of the relative frequencies of the different cell types, and thus cannot accept the CellCODE SPVs as input because they are not to scale.

To make use of CellCODE SPVs, we consider three cell-type assignment statistics that do not require correct scaling. The interaction *T*-test assigns genes based on maximal interaction coefficient between the clinical group and proportion variation and has been applied successfully to similar problems (Kuhn *et al.*, 2011; Repsilber *et al.*, 2010). The correlation test assigns each gene to cell type based on correlation with the proportion variable. Although this test ignores all information regarding clinical groups, it has been successfully used to annotate gene sets Bolen *et al.* (2011), and is very effective for rare cell types. In general, the interaction *T*-test works best at assigning regulated genes that are expressed in multiple cell types, whereas the correlation test assigns genes based on baseline expression specificity and works well for genes with exclusive expression. To capture both of these scenarios, we propose the *F*-test for the overall interaction model fit, which combines features of the interaction *T*-test and correlation metrics (see Supplementary Text for an in depth discussion).

Comparing the different assignment procedures, we find that the deconvolution method is the best at assigning genes regulated in the most common cell type (neutrophil) (Fig. 5). However, this is not indicative of higher accuracy, as this method is in general biased toward the most common cell type and thus produces a large number of neutrophil assignments under all simulation conditions.

Of the three methods that are able to use CellCODE rescaled covariates, we find that the *F*-test performs best. Additionally, for

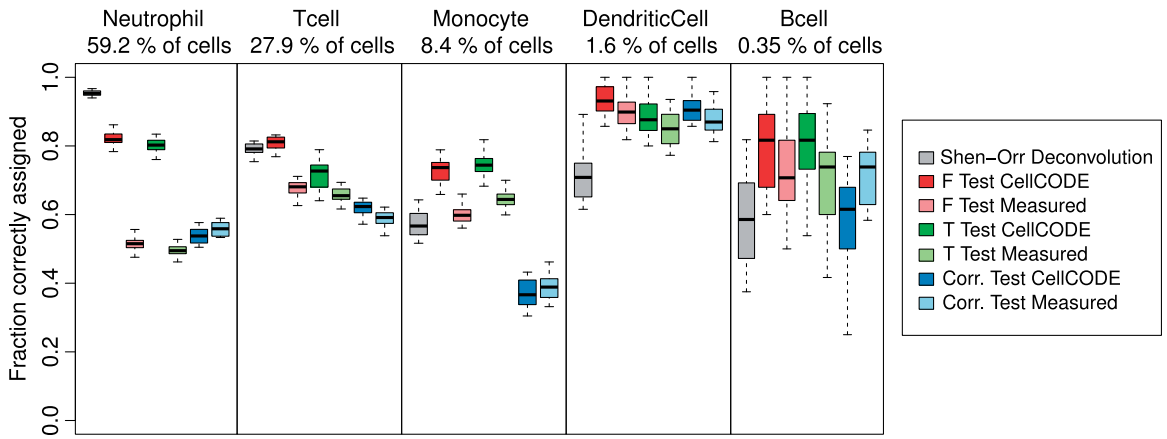

Fig. 5. Evaluating cell-type assignment methods using simulated data. Cell-type origin of differential expression is varied to create a range of simulated datasets. For each dataset, the set of DE genes is selected using the CellCODE approach (FDR 0.1) and is fixed for the subsequent analysis. These genes are assigned to the most likely cell type of origin using the different assignment methods. The fraction of correct assignments is plotted as a distribution boxplot for 20 independent repeats of the simulation. Methods that can accept rescaled covariates were evaluated using both the actual simulated cell proportions ('Measured', light colors) and the CellCODE SPVs (dark colors). Overall, we find that the *F*-test with CellCODE SPVs performs best (Color version of this figure is available at *Bioinformatics* online.)

all methods that are capable of using rescaled covariates, estimating SPVs from data structure was more effective than using known proportions (Fig. 5). Notably, this is true despite the fact that in our simulation the latter are exact values and are not even subject to the measurement error, which would occur in a real application.

In summary, we propose a multi-step pipeline for analyzing differential expression in mixture datasets. First, using independently obtained cell-type markers, we use the CellCODE approach to extract SPVs. The resulting SPVs can themselves be analyzed for differences between clinical groups. In the next step, the SPVs are used as covariates in a differential expression analysis. Genes that are predicted to be altered transcriptionally at the individual cell-type level can be assigned to the cell type of origin based on the *F*-test procedure.

2.4 Analysis of vaccination time course data

We demonstrate the utility of our method by applying it to a recent vaccination time-course study (Nakaya *et al.*, 2011). This study compares gene expression changes in PBMCs following the administration of two types of flu vaccines: the live attenuated influenza vaccine (LAIV) and the trivalent inactivated vaccine (TIV). Vaccine response produces large expression changes that result from both cell proportion changes and cell-specific transcriptional alterations. We apply CellCODE to resolve the sources of vaccine-related differential expression.

First we determine which cell proportion variation can be extracted from the data correlation structure. Using markers from the IRIS dataset, we extract separate components for neutrophils, monocytes, dendritic cells, NK cells, T cells, B cells and plasma cells (Supplementary Fig. S1). We find that some marker genes appear to be incorrectly assigned. Because the CellCODE SPVs track the main variance component, these have a negligible effect on the result. Importantly, we find a robust signature for plasma cells, which are particularly relevant to the study of vaccine biology but comprise a very small fraction (1–3%) of peripheral lymphocytes. Because of their relevance, the frequency of plasma cells was assayed independently in the original study. We find that our SPV estimates using expression data are in good agreement with the true proportions as measured by flow cytometry (Spearman rank correlation of 0.87, Supplementary Fig. S2).

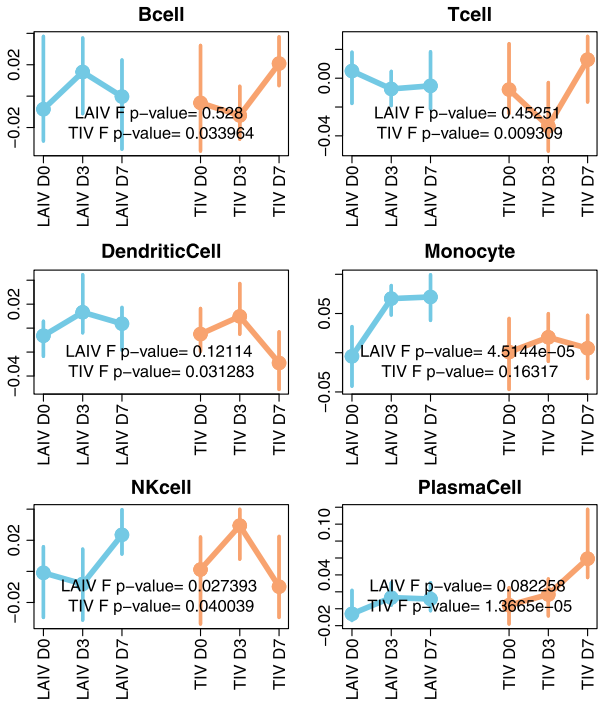

Fig. 6. Vaccine administration induces global changes in cell-type proportions. SPVs were extracted using the CellCODE method and evaluated for vaccine-related changes by comparing a model that captures individual variation only against one which includes post-vaccination day. The points and lines represent the median and interquartile range (IQR) of SPVs normalized to have mean 0 for each individual (which reduces variance without altering the trend). D0, D3 and D7 indicate day after vaccination (Color version of this figure is available at *Bioinformatics* online.)

The CellCODE SPVs explain a large fraction of the global gene expression changes observed with standard differential analysis by summarizing them as proportion changes. In particular, we find that the live vaccine causes an increase in the proportion of monocytes and the inactivated vaccine causes a large increase in the proportion of plasma cells (Fig. 6). After including the CellCODE SPVs as covariates in the differential expression analysis, the number of differentially regulated

genes is substantially reduced, as genes whose expression level simply tracks proportion changes are no longer included.

Those genes that are found to be DE after correcting for confounding proportion changes represent candidates for cell-type-specific transcriptional regulation. Using the CellCODE SPVs and *F*-test procedure to attribute these genes to specific cell types, we find that a large number of genes that are regulated in the LAIV time course are assigned to the T-cell population.

Further investigation of the top ranked T-cell DE genes suggest that the LAIV response regulates a T-cell-specific proliferation program. One top ranked gene is ZEB1, which is downregulated. ZEB1 is known to repress T-cell-specific IL2 expression, thereby inhibiting T-cell proliferation. We also find that BCLAF1, a pro-apoptotic gene is downregulated, while the expression BCL2, which antagonizes BCALF1, is increased. The direction of other gene expression changes is likewise consistent with a survival and proliferation program. As these changes were specific to the LAIV time course, the CellCODE analysis suggests that a unique T-cell transcriptional profile contributes to mechanistic differences between responses to the LAIV and TIV.

One goal of the original vaccine study was to define molecular determinants of vaccine efficacy. Efficacy of the TIV, as measured by antibody titers 28 days after vaccination, varied widely among individuals. The authors reported that the level of CAMK4 (Calcium/Calmodulin-Dependent Protein Kinase IV) mRNA on Day 3 post-vaccination was negatively correlated with antibody titers (a result also present in our reanalysis, Fig. 7A). The involvement of CAMK4 in regulating the antibody response was further confirmed by observations of a higher antibody titers in response to the TIV in mice lacking a functional CAMK4. Because antibodies are produced by cells from the B-cell lineage, it would be expected that the CAMK4 effect on antibody titers and its regulation involves B cells. We therefore applied the CellCODE pipeline to determine the cell type responsible, using the interaction *F*-test. Surprisingly, we find that CAMK4 is assigned to T cells (Fig. 7B). Various observations corroborate this prediction. In the IRIS dataset, CAMK4 is specific to T cells and RNAseq of human or mouse B cells confirms low B-cell expression [GSE39229: RPKM = 0.09 (Abraham *et al.*, 2013) and GSE49027: RPKM = 0.25]. Our computational analysis predicts that the effect CAMK4 exerts on antibody titers is not B-cell intrinsic and instead involves T-cell interactions. Although this hypothesis is strengthened by the dearth of CAMK4 transcripts in B cells, the original observation was made based solely on multivariate analysis of microarray mixture data, which demonstrates the potential power of our approach.

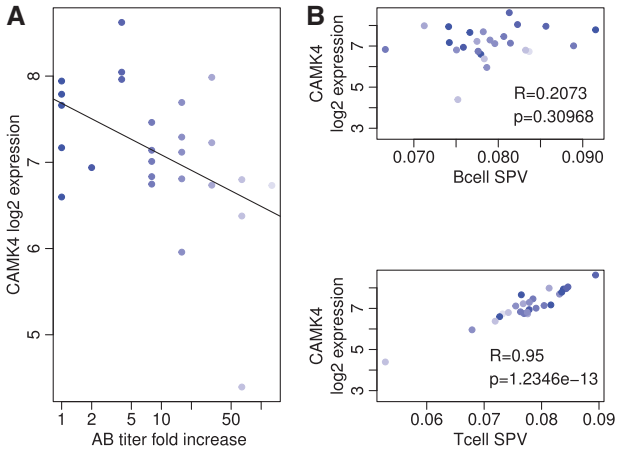

Fig. 7. CAMK4 effects antibody response through a T-cell-dependent mechanism. (**A**) Expression of CAMK4 on Day 3 negatively correlates with an increase in influenza-specific antibody titers. (**B**) CAMK4 expression correlates strongly with T-cell SPV but only slightly with B-cell SPV. Correlation with other cell-type proportions is negative (data not shown), suggesting that CAMK4 is T-cell specific. Colors denote the fold increase in antibody titers and are the same as in panel (**A**) (Color version of this figure is available at *Bioinformatics* online.)

3 Discussion

We propose a method for dissecting mixture datasets that is specifically designed for analyzing differential expression. Our method extracts cell proportion covariates from the data structure and estimates the confounding and interaction effects. We demonstrate the sensitivity and accuracy of this method on simulated data and its capacity to generate biological insights using experimental datasets.

Our method neither requires nor utilizes independent proportion measurements. Even when such measurements are available, because of various limitations in these data, they may not produce optimal covariates for statistical testing. For example, Coulter counter measurements can have an error of 5% or more for the rare cell types (Aulesa *et al.*, 2003). Additionally, some cell types that are counted may preferentially lose RNA due to processes such as apoptosis and, consequently, cell counts may not accurately reflect pooled mRNA composition. Most importantly, independent cell proportion measurements may not identify important molecularly distinct cell types. For example, the study by Shen-Orr *et al.* did not differentiate B-cells and T-cells, which cannot be distinguished by the Coulter principle alone. In contrast, CellCODE can capture variation from molecularly distinct cell types as long as some independent proportion variation in present and appropriate marker genes are available. Despite these limitations of the independent cell proportion measurements, it may be possible to develop a hybrid approach that makes use of this additional information. For example, a method that constrains the estimated latent variables to correlate with known covariates [see Mostafavi *et al.* (2013) for an example] might be adapted to extract mixture variation.

In summary, we propose a statistical framework for the analysis and interpretation of mixture samples that requires no dataset-specific prior knowledge yet performs comparably to or better than methods that use additional information. We demonstrate the general utility of evaluating interaction effects between known clinical groups and latent variables. In our case, the latent variables are cell-type proportions, but the methodology can be readily adapted to other latent sources of variation.

4 Methods

4.1 Data and processing

The DMAP dataset was downloaded in processed format from <http://www.broadinstitute.org/dmap/home>. All other datasets discussed in this article use Affymetrix arrays. They were processed using RMA with default parameters. The most highly expressed probe was chosen as a representative for a single gene. The values from the bg.params() method were used to compute a cutoff for non-expressed genes as $\mu + 2 * \sigma$, and genes with average value below this threshold were removed. For the vaccination data, we used the largest cohort corresponding to the 2008 vaccination year.

4.2 Simulated data

If we have n genes and m samples and the samples are composed of k cell types, then the n by m gene expression matrix $E_{n\times m}$ can be modeled as

$$E_{n\times m} = P_{n\times k} C_{k\times m} \quad (1)$$

Where P is a matrix of pure cell expression and C is a matrix of mixture proportions in each sample. In our simulation, we use the real cell-type proportions determined in Shen-Orr *et al.*, which capture the fact that some cell types are much more common than others. We also use experimentally determined pure-cell expression vectors from the IRIS study (Abbas *et al.*, 2009), GSE22886. These capture realistic expression values and real dependency structure because some cell types are more similar molecularly than others. The simulated data were derived from the IRIS dataset, whereas the Garvan [Jeffrey *et al.* (2006), GSE3928] dataset was used to define the cell-type-specific markers for CellCODE.

A simulated dataset was compiled as follows. A pure cell expression matrix on 5 cell types and 13 696 genes, P, was obtained from the IRIS data in linear (not log-transformed) space. An altered expression state, P^*, is simulated by altering the expression of 10% of the genes in one of five cell types by a factor drawn from $\mathcal{N}\{0,2\}$. The cell $C_{5\times 24}$ proportions were taken directly from Shen-Orr *et al.* and permuted so that when split into 2 groups of 12, the corresponding within group proportions were significantly different (Fig. 3). The control and disease datasets are simulated as $PC_{5\times\{1..12\}}$ and $PC_{5\times\{13..24\}}$. Noise, proportional to the gene expression mean, is added to the gene expression values (corresponding to constant variance in log space). The amount of noise is chosen so that the average variance explained by cell proportion variation was equal to 0.35, which is similar to that of real datasets. The final log transformed data serve as the input to all the methods discussed in this article. Functions used to simulate the data are part of the CellCODE package.

4.3 Marker genes

The general strategy for marker selection was to find genes whose expression value in one cell type exceeds that of all other cell types being considered by some defined threshold. For the IRIS and Garvan datasets, we used a cutoff of 2 in log 2 expression space. For the DMAP dataset, which has been normalized for batch effect and does not report true expression values, we found that a cutoff of 0.7 worked well. For ease of visualization, we also propose taking only the top most highly expressed marker genes. We find that adding more genes did not alter results but made the correlation structure more difficult to visualize. We found that restricting each marker list to 15 provides a good balance between capturing a robust consensus signature and interpretable visualization.

For analyzing real mixture datasets, we used the IRIS dataset to determine leukocyte markers and the DMAP dataset for other blood cell types. Many of the cells profiled in these experiments belong to the same lineage and have profiles that are too similar to yield marker genes with the expression difference we require. Therefore, we restrict our analysis to a subset of cell types that could produce a set of differentiating markers. In the IRIS dataset, we chose CD4 T cells to represent the T-cell population, naive B cells, 'Day 0' monocytes, stimulated dendritic cells, naive NK cells and plasma cells. For each mixture dataset analyzed, we relied on the visualization strategy to confirm a block-like correlation structure and ensure that SPVs tracked independent variance components.

We supply a function that implements our marker selection heuristics as part of the CellCODE R package, and code to reproduce the analysis of the Shen-Orr dataset, including marker selection, is supplied in the package vignette.

4.4 Proportion estimation

Estimating proportion variation from a pre-defined set of markers is complicated by the fact that some of the markers may themselves be transcriptionally altered within the cell type they represent. The goal then is to separate the variation that is due to proportion changes from that which is due to transcriptional alterations. Our method proceeds by evaluating all the genes for group effect with an F-test and discarding a fraction f_{exclude} of genes with the most changes. The eigengene from the remaining genes becomes the initial estimate.

To formalize, suppose that we have a set of genes M that are candidate markers of a cell type. Let E_M represent the subset of the expression matrix restricted to those marker genes. Then the corresponding SPV can be expressed as the first right singular vector (eigengene) of the following weighted SVD:

$$WE_M = UDV^T \quad (2)$$

The matrix W is diagonal with, $w_{ii} \in \{0,1\}$, where the value of 1 denotes the genes that are stable markers of cell type and are not themselves regulated.

We first determine W by sorting the marker genes in order of differential expression. Specifically, we evaluate the model fit, $g_i = \hat{g}_i + ay + \epsilon$, where in the simplest case of two clinical groups, g_i is the baseline expression level and y is a vector ($y_i \in \{0,1\}$) indicating group membership. Setting f_{exclude} fraction of genes with the greatest group effect to a weight of $w_{ii} = 0$ gives us the initial SPV estimate s_{initial} via Equation (2). Once we have the initial SPV estimate, we can reevaluate which marker genes have a group effect conditioned on proportion variation. This is achieved by including the initial SPV and the group–SPV interaction in out F-test. This second F-test compares the model $g_i = \hat{g}_i + bs_{\text{initial}} + ay + c(y \circ s_{\text{initial}}) + \epsilon$ [where $\circ$ denotes element-wise multiplication and $(y \circ s_{\text{initial}})$ specifies the interaction term] to $g_i = \hat{g}_i{}' + b's_{\text{initial}} + \epsilon'$. This new list of top DE marker genes is given a weight of 0 and the eigengene is recomputed.

This procedure is similar to the two-step implementation of SVA with an additional interaction term. However, because we only have a few marker genes to evaluate, we cannot rely on the local FDR calculation to select the genes to discard and instead a fraction f_{exclude} is specified a priori. Although f_{exclude} is technically a free parameter, we found that in simulations 0.3 was a reasonable choice as overestimating the fraction was not a problem, and unbiased values are obtained even when there are no DE markers. In a real dataset, the line between proportion changes and transcriptional changes may be blurred if the cell types being considered belong to the same lineage. In these cases, there may not be a unique correct solution, and it is likely that this procedure will be refined further as more real datasets are analyzed and interpreted.

4.5 Differential expression and cell-type assignment tests

Using the notation above, differential expression is evaluated as the T statistic for the a coefficient in the multiple regression denoted by $g_i = \hat{g}_i + ay + BS + \epsilon$, where S now represents a matrix of all the SPVs estimated from the data. The cell-type assignment relies on the interaction fit

$$g_i = \hat{g}_i + ay + b * s_j + c(y \circ s_j) + \epsilon \quad (3)$$

evaluated individually for all the SPVs s_j. For cell-type assignment based on a T-test, we take the maximal T statistic for the interaction coefficient c. For cell-type assignment based on the F-test (which we found performs best), we generate F statistics for each SPV by comparing the model in Equation (3) to a common null model $g_i = \hat{g}_i{}' + a'y + \epsilon'$. Note, the option of including additional SPVs as

non-interacting covariates in Equation (3) is also implemented, though we find that for smaller datasets it degrades the performance of the assignment metrics.

Funding

This work was supported by NIH Contract HHSN272201000054C and NIH award U54HG008540.

Conflict of Interest: none declared.

References

Abbas,A.R. *et al*. (2009) Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS One*, **4**, e6098.

Abraham,B.J. *et al*. (2013) Dynamic regulation of epigenomic landscapes during hematopoiesis. *BMC Genomics*, **14**, 193.

Adalsteinsson,B.T. *et al*. (2012) Heterogeneity in white blood cells has potential to confound DNA methylation measurements. *PLoS One*, **7**, e46705.

Aulesa,C. *et al*. (2003) Validation of the Coulter LH 750 in a hospital reference laboratory. *Lab. Hematol.*, **9**, 15–28.

Bolen,C.R. *et al*. (2011) Cell subset prediction for blood genomic studies. *BMC Bioinformatics*, **12**, 258.

Gaujoux,R. and Seoighe,C. (2013) Cellmix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics*, **29**, 2211–2212.

Jeffrey,K.L. *et al*. (2006) Positive regulation of immune cell function and inflammatory responses by phosphatase pac-1. *Nat. Immunol.*, **7**, 274–283.

Kuhn,A. *et al*. (2011) Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nat. Methods*, **8**, 945–947.

Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Leek,J.T. and Storey,J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, **3**, 1724–1735.

Mostafavi,S. *et al*. (2013) Normalizing RNA-sequencing data by modeling hidden covariates with prior knowledge. *PLoS One*, **8**, e68141.

Nakaya,H.I. *et al*. (2011) Systems biology of vaccination for seasonal influenza in humans. *Nat. Immunol.*, **12**, 786–795.

Novershtern,N. *et al*. (2011) Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, **144**, 296–309.

Repsilber,D. *et al*. (2010) Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC Bioinformatics*, **11**, 27.

Schwartz,R. and Shackney,S.E. (2010) Applying unmixing to gene expression data for tumor phylogeny inference. *BMC Bioinformatics*, **11**, 42.

Shen-Orr,S.S. and Gaujoux,R. (2013) Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Curr. Opin. Immunol.*, **25**, 571–578.

Shen-Orr,S.S. *et al*. (2010) Cell type-specific gene expression differences in complex tissues. *Nat. Methods*, **7**, 287–289.

Yoshihara,K. *et al*. (2013) Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.*, **4**, 2612.