Biostatistics (2013), **0**, 0, *pp.* 1–30
doi:10.1093/biostatistics/MIMOSA

Mixture Models for Single-Cell Assays with Application to Vaccine Studies

GREG FINAK[1*], ANDREW MCDAVID[1], PRATIP CHATTOPADHYAY[3], MARIA DOMINQUEZ[3], STEVE DE ROSA[1,2], MARIO ROEDERER[3], RAPHAEL GOTTARDO[1]

[1] *Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center (FHCRC), Seattle, WA*

[2] *HIV Vaccine Trials Network, Fred Hutchinson Cancer Research Center (FHCRC), Seattle, WA*

[3] *Vaccine Research Center, NIAID, NIH, 40 Convent Drive, Rm 5509, Bethesda, MD 20892*

gfinak@fhcrc.org

Summary

Blood and tissue are composed of many functionally distinct cell subsets. In immunological studies, these can only be measured accurately using single-cell assays. The characterization of these small cell subsets is crucial to decipher system level biological changes. For this reason, an increasing number of studies rely on assays that provide single-cell measurements of multiple genes and proteins from bulk cell samples. A common problem in the analysis of such data is to identify biomarkers (or combinations of biomarkers) that are differentially expressed between two biological conditions (*e.g.*, before/after stimulation), where expression is defined as the proportion of cells expressing that biomarker (or biomarker combination) in the cell subset(s) of interest. Here, we present a Bayesian hierarchical framework based on a beta-binomial mixture model for testing for differential biomarker expression using single-cell assays. Our model allows the inference to

*To whom correspondence should be addressed.

be subject specific, as is typically required when accessing vaccine responses, while borrowing strength across subjects through common prior distributions. We propose two approaches for parameter estimation: an empirical-Bayes approach using an Expectation-Maximization algorithm and a fully Bayesian one based on a Markov chain Monte Carlo algorithm. We compare our method against frequentist approaches for single-cell assays including Fisher's exact test, a likelihood ratio test, and basic log-fold changes. Using several experimental assays measuring proteins or genes at the single-cell level and simulated data, we show that our method has higher sensitivity and specificity than alternative methods. Additional simulations show that our framework is also robust to model misspecification. Finally, we also demonstrate how our approach can be extended to testing multivariate differential expression across multiple biomarker combinations using a Dirichlet-multinomial model and illustrate this multivariate approach using single-cell gene expression data and simulations.

Key words: Expectations-Maximization, Markov Chain Monte Carlo, Marginal Likelihood, Bayesian Modeling, Hierarchical Modeling, Immunology, Flow Cytometry, Single-Cell Gene Expression, MIMOSA

1. Introduction

Cell populations, particularly in the immune system, are never truly homogeneous; individual cells may be in different biochemical states that define functional but measurable differences between them. This single-cell heterogeneity is informative, but lost in assays that measure cell mixtures. For this reason, endpoints in vaccine and immunological studies are measured through a variety of assays that provide single-cell measurements of multiple genes and proteins. In the 1970s, single-cell analysis was revolutionized with the development of fluorescence-based flow cytometry (FCM). Since then, instrumentation and reagent advances have enabled the study of numerous cellular processes via the simultaneous single-cell measurement of multiple surface and intracellular biomarkers (up to 17 biomarkers). More recent technological development have drastically extended the capabilities of single-cell cytometry to measure dozens of simultaneous parameters (i.e. proteins, genes, cytokines, etc.) per cell (Bendall *and others*, 2011). Although cells sorted using well-established surface biomarkers may appear homogeneous, mRNA expression of other genes within these cells can be heterogeneous (Narsinh *and others*, 2011; Flatz *and others*, 2011) and could further characterize and subset these cells. A new technology based on microfluidic arrays combined with multiplexed polymerase chain reactions (PCR) can now be used to perform thousands of PCRs in a single device, enabling simultaneous, high-throughput gene expression measurements at the single-cell level across hundreds of cells and genes (Pieprzyk, 2009). While classic gene expression microarrays sum the expression from many individual cells, the intrinsic stochastic nature of biochemical processes results in relatively large cell-to-cell gene expression variability (van Oudenaarden, 2009). This heterogeneity may carry important information, thus single-cell expression data should not be analyzed in the same fashion as cell-population level data. Special treatment of single-cell level data, which preserves information about population heterogeneity, is warranted in general. For this reason, single-cell assays are an important tool in immunology, providing a functional and phenotypic snapshot of the immune system at a given

time. These assays typically measure multiple biomarkers simultaneously on individual cells in a heterogeneous mixture such as whole blood or peripheral blood mononuclear cells (PBMC), and are used for immune monitoring of disease, vaccine research, and diagnosis of haematological malignancies (Altman *and others*, 1996; Betts *and others*, 2006; Inokuma *and others*, 2007).

During analysis, cell level biomarker fluorescence intensities are typically thresholded as positive or negative so that subsets with different multivariate $+/-$ combinations can be obtained as Boolean combinations. For some assays (*e.g.*, flow cytometry), the positivity thresholds are set based on prior biological knowledge while for others, thresholds are given by the assay technology. This is the case for the Fluidigm technology where genes are recorded as absent (not expressed) or present (expressed) at the single-cell level. After this thresholding step, we obtain a Boolean matrix of dimension $N \times K$, where N is the number of cells recorded and K is number of biomarkers. Using this matrix, one can form 2^K putative cell subsets obtained as Boolean combinations. When K is large there is a combinatorial explosion of the number of subsets, and many of these might be small or even empty. A common statistical problem is, for a given biomarker combination, to identify subjects for whom the proportion of cells expressing that combination is significantly different between two experimental conditions (*e.g.*, before and after stimulation). Note that we use the term 'subject' throughout the paper, but the approaches described are general and can be applied to other experimental units (*e.g.*, animal studies).

A motivating example from vaccine research is the flow cytometric intracellular cytokine staining (ICS) assay, which is used to identify and quantify subjects' immune responses to a vaccine. Upon vaccination, antigen in the vaccine is taken up and presented to CD4 or CD8 T-cells via antigen presenting cells. While not all T-cells can recognize all antigens, those that recognize antigens in the vaccine become *activated* and produce a variety of cytokines, further promoting the immune response. After activation, this antigen-specific subpopulation proliferates and can persist in the immune system for some time providing *memory* that can more rapidly

recognize the same antigen again in the future (McKinstry *and others*, 2010). The antigen–specific T-cell subpopulations (i.e. the subset that can respond to one specific antigen) constitute a very small fraction of the total number of CD4 and CD8 T-cells. The ICS assay measures the number of antigen-specific T-cells in PBMC or whole blood by measuring cytokine production in response to activation following stimulation by an antigen that closely matches what was present in the original vaccine. Individual cells are labelled using fluorescently conjugated antibodies against cell–surface biomarkers (CD3, CD4, and CD8), used to subset T-cells, and functional biomarkers (cytokines) used to define antigen specific T-cells (Horton *and others*, 2007; De Rosa *and others*, 2004; Betts *and others*, 2006). A sufficiently large number of cells must be collected (on the order of 50,000 to 100,000 T–cells) to ensure that the rare cell populations can be detected. Subsequently, each individual cell is classified as either positive or negative for each maker based on predetermined thresholds, then the number of cells matching each subpopulation phenotype is counted.

These counts are compared between antigen stimulated and unstimulated samples from a subject to identify significant differences. Subjects who generate a response after stimulation are called *responders*, whereas subjects that do not show any differences are called *non-responders*. The comparisons between stimulated and unstimulated samples are typically done within a time point (*i.e.,* within pre–vaccination and within post–vaccination time points), rather than directly between pre–and post–vaccine time points to avoid confounding vaccine effect with date or time of blood draw; additionally not all vaccine trials collect baseline (pre–vaccination) samples. In many immunological studies, the size of the functionally distinct subpopulations (*i.e.*, the number of positive cells) is very low (on the order of 10s of cells, with effect sizes on the order of 10^{-4}, relative to the total number of cells), and real biological differences might be difficult to detect.

Although there is no standard approach to analyzing ICS assays, current methods range from ad-hoc rules based on log-fold changes (Trigona *and others*, 2003), to non–parametric meth-

ods (Sinclair *and others*, 2004), to exact tests of 2x2 contingency tables (*e.g.*, Fisher's exact test and χ^2 test) (Horton *and others*, 2007; Proschan and Nason, 2009; Peiperl *and others*, 2010; Nason, 2006). All of these methods test subjects separately, and no information is shared across observations even though one could expect some similarities across responders (or non-responders).

The framework developed in this paper, named MIMOSA (Mixture Models for Single Cell Assays), addresses these issues explicitly. In our model, cell counts are modelled by a binomial (or multinomial in the multivariate case) distribution and information is shared across subjects by means of a prior distribution placed on the unknown proportion(s) of the binomial (or multinomial) likelihood. In order to discriminate between responders and non-responders, the prior is written as a mixture of two beta (or Dirichlet in the multivariate case) distributions where the hyper-parameters for each mixture component are shared across subjects. This sharing of information helps regularize proportion estimates when the cell counts are small, which is typical with single-cell assays, and increases sensitivity and specificity when detecting responders. Because our framework is multivariate in nature, multiple cell subsets can be modelled simultaneously, which could help detect small biological changes that are spread out across multiple cell subsets (Nason, 2006). Our paper is organized as follows; Section 2 introduces the data and notations used in the paper. In Section 3, we present our model for testing differential biomarker expression in the univariate case. Section 4 compares our approach to alternative methods and tests the robustness of our model. In Section 5 we present a multivariate extension of our model that can be used to test multivariate biomarker differential expression and present some results using a single-cell gene expression data. Finally, in Section 6 we discuss our findings and future work.

2. Notation and Data

In the remainder of the paper, we use the following notation to describe our model. We assume that we observe cell counts from I subjects in two conditions: stimulated and un-stimulated. Each cell can either be positive or negative for a biomarker. Given a set of K biomarkers, the measured cells can be classified into 2^K positive/negative biomarker combinations. We denote by $n_{ik}^{(s)}$ and $n_{ik}^{(u)}$, $i = 1, \ldots, I, k = 1, \ldots, 2^K$, the observed counts for the 2^K combinations in the stimulated and un-stimulated samples, respectively. We denote by $N_i^{(s)} = \sum_k n_{ik}^{(s)}$ and $N_i^{(u)} = \sum_k n_{ik}^{(u)}$ the total number of cells measured for subject i in each sample, respectively. For ease of notation, we denote by $\mathbf{y}_i$ the vector of observed counts for subject i, *i.e.*, $\mathbf{y}_i = (\mathbf{n}_i^{(s)}, \mathbf{n}_i^{(u)})$ where $\mathbf{n}_i^{(s)} = \{n_{ik}^{(s)} : k = 1, \ldots, 2^K\}$ and $\mathbf{n}_i^{(u)} = \{n_{ik}^{(u)} : k = 1, \ldots, 2^K\}$. Finally, we define $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_I)$.

We consider two types of immunological single-cell assays: flow cytometry and single-cell gene expression, as described below.

Flow cytometry: The primary dataset used here is an ICS data set generated as part of a trial testing the GeoVax DNA and MVA (Modified Vaccinia Ankara) HIV vaccine in a prime-boost regimen (prime at zero and two months, boost at four and six months) (Goepfert *and others*, 2011). The goal of this data set was to assess the immune response to the vaccine across multiple antigen stimulations, time points, cytokines and T-cell subsets. Here, we analyze a subset of the data consisting of 98 subjects from the vaccine group at two time points: day 0 and day 182. Three cytokines (IFN-γ, TNFα and IL2) were measured at the single-cell level for each subject and time point, with and without stimulations with an antigen (here we focus on HIV Envelope peptide pool) matching part of the vaccine. For ease of presentation we restricted ourselves to the CD4+ T-cell subsets. Samples on day 0 were taken just before vaccination and no response is expected there. The corresponding samples can be used as negative controls. Conversely, day 182 (26 weeks) should be close to the immunogenicity peak, and many subjects are expected to respond,

for some cytokines at least. In this data set a median of 51K and 58K T–cells were collected for the stimulated and unstimulated samples, respectively (IQR $\approx$ 37K and 43K, respectively) were collected for the stimulated and unstimulated samples. The effect sizes (differences in the proportions of stimulated and unstimulated cells) are on the order of 10^{-4}, with the number of positive cells typically ranging from 1–70, with a few samples exhibiting large numbers of positive events.

Fluidigm single-cell gene expression: This is a single-cell gene expression data set of sorted CD8+ T-cells from sixteen subjects. T-cells isolated by flow cytometry from sixteen subjects were stimulated in blocks of four subjects with four different antigens (HIV Gag, HIV Nef, CMV pp65 tm10, and CMV pp65 nlv5) and gene expression post-stimulation measured at the single-cell level using the BioMark system (Fluidigm) 96×96 well arrays. The expression from the simulated samples was compared to paired, unstimulated controls. In this data set, we have approximately 90 single cells per subject per stimulation condition with 96 gene expression measurements per cell.

3. Differential expression with one biomarker

Datasets like the ones presented here are usually analyzed in a univariate fashion to avoid being underpowered due to the large number of combinations and the potential for very small cell counts in many of the combinations. By univariate, we mean that we have only one positive cell subset. This cell subset can be defined by considering the expression of one biomarker alone (marginalizing over all other measured biomarkers) such as A+ (*vs.* A−), or considering a specific positive biomarker combination (and marginalizing over everything else) such as A+ and/or B+ (*vs.* A−/B−). Without loss of generality, we treat the univariate case as a one biomarker case (*i.e.*, $K = 1$). In this case, for a given subject, the data can be summarized in a contingency table of +/− cell counts across the un-stimulated and stimulated samples as depicted in Table 1.

For a given subject and stimulation, we consider a biomarker to be differentially expressed if the proportion of positive cells in the stimulated samples is different from the number of positive cells in the un-stimulated sample. Subjects that show differential expression will be called responders for that biomarker. In this section, we are concerned with identifying differential expression one biomarker at a time, using a beta-binomial mixture model as described below.

3.1 *Beta-binomial model*

For a given subject i, the positive cell counts for the stimulated and un-stimulated samples are jointly modeled as follows:

$$(n_{si}|p_{si}) \sim \text{Bin}(N_i^{(s)}, p_i^{(s)}) \quad \text{and} \quad (n_i^{(u)}|p_i^{(u)}) \sim \text{Bin}(N_i^{(u)}, p_i^{(u)})$$

where $p_i^{(s)}, p_i^{(u)}$ are the unknown proportions for the stimulated and un-stimulated paired samples, respectively. In order to detect responding subjects, we consider two competing models:

$$\mathcal{M}_0 : p_i^{(u)} = p_i^{(s)} \quad \text{and} \quad \mathcal{M}_1 : p_i^{(u)} \neq p_i^{(s)}.$$

Under the null model, $\mathcal{M}_0$, there is no difference between the stimulated and un-stimulated samples, and the proportions are equal (yet the cell counts can differ). Under the alternative model, $\mathcal{M}_1$, there is a difference in proportions between the two samples and the subject i is a responder. In some studies, such as the ICS data used here, the proportion of positive cells is expected to only increase after stimulation, in which case the alternative model should be defined as $p^{(s)} > p^{(u)}$. This alternative parametrization is described in Appendix A of the Supplementary Material, and we refer to it as the one-sided model.

3.2 *Priors*

Our model shares information across all subjects using exchangeable Beta priors on the unknown proportions, as follows:

$$(p_i^{(u)}|z_i = 0) \sim \text{Beta}(\alpha^{(u)}, \beta^{(u)})$$

$$(p_i^{(u)}|z_i = 1) \sim \text{Beta}(\alpha^{(u)}, \beta^{(u)}) \quad \text{and} \quad (p_i^{(s)}|z_i = 1) \sim \text{Beta}(\alpha^{(s)}, \beta^{(s)}),$$

where z_i is an indicator variable equal to one if subject i is a responder, *i.e.*, $\mathcal{M}_1$ is true, and zero otherwise, and $\alpha^{(u)}, \beta^{(u)}, \alpha^{(s)}, \beta^{(s)}$ are unknown hyper-parameters shared across all subjects. Note that the parameters $\alpha^{(u)}$ and $\beta^{(u)}$ are explicitly shared across the two models, whereas $\alpha^{(s)}$ and $\beta^{(s)}$ are only present in the alternative model. Finally, we assume that $z_i \sim \text{Be}(w)$ are independent draws from a Bernoulli distribution with probability w, where w represents the (unknown) proportion of responders. It follows that marginally, *i.e.*, after integrating z_i, the $p_i^{(u)}$ and $p_i^{(s)}$ are then jointly distributed as a mixture of a one dimensional Beta distribution and a product of two Beta distributions (with a possible constraint), with mixing parameter w. Treating the z_i's as missing data, the unknown parameter vector $\boldsymbol{\theta} \equiv (\alpha^{(u)}, \beta^{(u)}, \alpha^{(s)}, \beta^{(s)}, w)$ can be estimated in an Empirical-Bayes fashion using Expectation-Maximization algorithm (Dempster *and others*, 1977) as described in Section 3.3. As an alternative, we also describe a fully Bayesian model, where the hyperparameters $\alpha^{(u)}, \beta^{(u)}$, $\alpha^{(s)}$, and $\beta^{(s)}$ are each given vague exponential priors with mean 10^3, and w is assumed to be drawn from a uniform distribution between 0 and 1. In this case, all parameters will be estimated via a Markov chain Monte Carlo algorithm as described in Section 3.3.

3.3 *Parameter estimation*

In our proposed EM and MCMC algorithms, we greatly simplify our calculations by directly utilizing the marginal likelihoods, L_0 and L_1, obtained after marginalizing $p_i^{(s)}$ and $p_i^{(u)}$ from the

null and alternative likelihoods. Given the conjugacy of the priors, the marginal likelihoods L_0 and L_1 are available in closed-forms (Appendix B, Supplementary Material), and are given by,

$$\mathrm{L}_0(\alpha^{(u)}, \beta^{(u)}|\mathbf{y}_i) = \binom{N_i^{(u)}}{n_i^{(u)}}\binom{N_i^{(s)}}{n_i^{(s)}} \cdot \frac{\mathrm{B}(n_i^{(s)} + n_i^{(u)} + \alpha^{(u)}, N_i^{(s)} - n_i^{(s)} + N_i^{(u)} - n_i^{(u)} + \beta^{(u)})}{\mathrm{B}(\alpha^{(u)}, \beta^{(u)})}$$

and

$$\mathrm{L}_1(\alpha^{(u)}, \beta^{(u)}, \alpha^{(s)}, \beta^{(s)}|\mathbf{y}_i) = \binom{N_i^{(u)}}{n_i^{(u)}}\binom{N_i^{(s)}}{n_i^{(s)}} \cdot \frac{\mathrm{B}(n_i^{(u)} + \alpha^{(u)}, N_i^{(u)} - n_i^{(u)} + \beta^{(u)})}{\mathrm{B}(\alpha^{(u)}, \beta^{(u)})} \cdot \frac{\mathrm{B}(n_i^{(s)} + \alpha^{(s)}, N_i^{(s)} - n_i^{(s)} + \beta^{(s)})}{\mathrm{B}(\alpha^{(s)}, \beta^{(s)})} \tag{3.1}$$

Above, B is the Beta function. Assuming that the missing data, $z_i, i = 1, \ldots, I$, are known, we define the complete data log-likelihood:

$$l(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) = \sum_i z_i l_0(\alpha^{(u)}, \beta^{(u)}|\mathbf{y}_i) + (1 - z_i) l_1(\alpha^{(u)}, \beta^{(u)}, \alpha^{(s)}, \beta^{(s)}|\mathbf{y}_i) + z_i \log(w) + (1 - z_i)\log(1 - w), \tag{3.2}$$

where l_0 and l_1 are the log marginal-likelihoods and $\boldsymbol{\theta} \equiv (\alpha^{(u)}, \beta^{(u)}, \alpha^{(s)}, \beta^{(s)}, w)$ is the vector of parameters to be estimated. In the one-sided case, the alternative prior specification must satisfy the constraint $p^{(s)} > p^{(u)}$, and the marginal likelihood derivation involves the calculation of a normalizing constant that is not available in closed-form but can easily be estimated. All calculations for the one-sided case are described in Appendix A of the Supplementary Material.

EM algorithm

Given an estimate of the model parameter vector $\tilde{\boldsymbol{\theta}} = \left\{\tilde{\alpha}^{(u)}, \tilde{\beta}^{(u)}, \tilde{\alpha}^{(s)}, \tilde{\beta}^{(s)}, \tilde{w}\right\}$ and the data $\mathbf{y}$, the E step consists of calculating the posterior probabilities of differential expression, defined by

$$\tilde{z}_i \equiv \Pr(z_i = 1|\mathbf{y}, \tilde{\boldsymbol{\theta}}) = \frac{\tilde{w} \cdot \mathrm{L}_1(\tilde{\alpha}^{(u)}, \tilde{\beta}^{(u)}, \tilde{\alpha}^{(s)}, \tilde{\beta}^{(s)}, |\mathbf{y}_i)}{(1 - \tilde{w}) \cdot \mathrm{L}_0(\tilde{\alpha}^{(u)}, \tilde{\beta}^{(u)}|\mathbf{y}_i) + \tilde{w} \cdot \mathrm{L}_1(\tilde{\alpha}^{(u)}, \tilde{\beta}^{(u)}, \tilde{\alpha}^{(s)}, \tilde{\beta}^{(s)}|\mathbf{y}_i)}.$$

The M-step then consist of optimizing the complete-data log-likelihood over $\boldsymbol{\theta}$ after replacing z_i by $\tilde{z}_i$ in (3.2). Straightforward calculations lead to $\tilde{w} = \sum_i \tilde{z}_i / I$, but unfortunately no closed form solutions exist for the remaining parameters. We use numerical optimization as implemented in R's *optim* function to estimate the remaining parameters (Ihaka and Gentleman, 1996). Starting

from some initial values, the EM algorithm iterates between the E and M steps until convergence. In our case, we initialize the z_i's using Fisher's exact test to assign each observation to either the null or alternative model components. We then use the estimated z_i's to estimate the $p_i^{(u)}$'s and $p_i^{(s)}$'s and use these to set the hyper-parameters to their method-of-moments estimates.

MCMC algorithm

We generated realizations from the posterior distribution via MCMC algorithms (Gelfand, 1996). All updates were done via Metropolis-Hastings sampling except for the z_i's and w that were performed via Gibbs samplings. Details about the algorithms are given in Appendix B of the Supplementary Material. We used the method of Raftery and Lewis (1992) and Raftery (1996) to determine the number of iterations, based on a short pilot run of the sampler. For each dataset presented here, we calculated that no more than about 100,000 iterations with 50,000 burn-in iterations was sufficient to estimate standard posterior quantities. To leave some margin, we used 200,000 iterations after 50,000 burn-in iterations for each dataset explored here.

4. Results

In this section, we apply our MIMOSA model to the data described in Section 2, and present the results of a simulation study based on the ICS data. We evaluated and compared the performance of MIMOSA against Fisher's exact test, the likelihood ratio test, and log fold-change by ROC (receiver operator characteristic) curve analysis and by comparing the *observed FDR* (false discovery rate) against the *nominal FDR* (expected false discovery rate) for each data set (Newton *and others*, 2004) where a false discovery (for the ICS data) is a day 0 sample (non–responder) that is incorrectly identified as a responder.

4.1 *ICS*

In the ICS data we considered observations at the day 0 time point as true negatives, and observations at the day 182 time point as true positives (potentially underestimating the sensitivity of all methods considered here due to real non-responders at day 182). The MIMOSA model has higher sensitivity and specificity than method compared here for discriminating vaccine responders and non-responders as shown by the ROC curves on Figure 1, panels A,C,E. At an FDR between 10-20%, MIMOSA would lead to about 20% more true positives being detected. Our comparisons also show that ranking based on log-fold change alone is not reliable and should not be used. In addition, MIMOSA gave estimates of the observed false discovery rate that are better or comparable to competing methods (Figure 1, panels B,D,F). Here we present the results based on IL2 and IFN-γ alone and the subset IL2 and/or IFN-γ that were used in the original study (Goepfert *and others*, 2011). These results are consistent for other cytokines and cytokine combinations (see Supplementary Figure 1).

We verified the assumption of a common distribution for the $p_i^{(u)}$ across subjects by examining the empirical estimates of $p_i^{(u)}$ and the density of the Beta distribution with parameters $\hat{\alpha}^{(u)}, \hat{\beta}^{(u)}$ fitted to ENV-1-PTEG stimulated CD4+/IFN–γ+ T–cell ICS data (Supplementary Figure 2).

4.2 *Single-cell gene expression*

We applied the MIMOSA model to a Fluidigm single-cell gene expression data set. We used the two-sided MIMOSA model because genes could be regulated upward or downward upon stimulation. In order to detect stimulation specific changes of expression, we fit our model to each gene within each stimulation. The results presented in Figure 2 show that MIMOSA identifies stimulation-specific differences in the proportions of cells expressing each gene while preserving inter-subject variability (Figure 2 A,B). These patterns are evident in the posterior probabilities (Figure 2 A) and preserved in the posterior estimates of the differences of proportions (Figure 2

B). A similar analysis using a two-sided Fisher's exact test and clustering the signed FDR adjust p-values (Figure 2 C) does not reveal any stimulation-specific patterns. For an FDR of 10%, Fisher's exact test identified 47 significant genes, while MIMOSA identified 50 significant genes. Both methods identified 39 genes in common.

4.3 *Simulation Studies*

We examined the performance of the constrained ($p^{(s)} > p^{(u)}$) and unconstrained ($p^{(s)} \neq p^{(u)}$) beta-binomial mixture models via simulations. For the simulation, we used hyper–parameters estimated from a one-sided MIMOSA model fit to ICS data (IL2 univariate) from the primary immunogenicity time point. We simulated data from this constrained model with 200 observations, a response rate of 60%, N = 1,000, 5,000, and 10,000 events, with ten independent realizations of data for each N. We fit the one-sided MIMOSA model to this data. We evaluated the sensitivity and specificity of the model's ability to correctly identify observations from the "responder" and "non-responder" groups through analysis of ROC curves, and compared against Fisher's exact test, the likelihood ratio test, and log fold-change. We repeated this procedure for the two-sided models fit to two-sided data (Figure 3 A,C). In addition, we examined the nominal *vs.* observed FDR to assess the ability of each method to properly estimate the FDR (Figure 3 B,D).

For both the constrained and unconstrained simulations, MIMOSA was superior to competing methods, including Fisher's exact test, with respect to sensitivity and specificity even at small values of N (Figure 3 A and C). Additionally, the estimated FDR for MIMOSA more closely reflected the nominal FDR compared to Fisher's exact test and competing methods (Figure 3, panels B, D).

To assess the sensitivity of the model to deviations from model assumptions, we repeated the simulations with the cell proportions drawn from truncated normal distributions with support

$(0, 1)$, rather than beta distributions. The means and variances of the truncated normal distributions were set to the maximum likelihood estimates of the beta distributions defined by the hyper–parameters α and β estimated from the ICS data set (see Supplementary Figure 3 panels C and D). Even under these departures from the model assumptions, the unconstrained MIMOSA model outperformed Fisher's exact test.

5. Differential expression across biomarker combinations

Our beta-binomial model described in Section 3.1 can be generalized to a Dirichlet-multinomial model to assess differential expression across multiple biomarker combinations. As described in the data section, we now have counts for each biomarker combination, denoted by $\mathbf{n}_i^{(s)} = \{n_{ik}^{(s)} : k = 1, \ldots, 2^K\}$ and $\mathbf{n}_i^{(u)} = \{n_{ik}^{(u)} : k = 1, \ldots, 2^K\}$.

5.1 *Model*

In our multivariate model, the beta distribution is replaced by a multinomial distribution, as follows:

$$(\mathbf{n}_i^{(u)}|\mathbf{p}_i^{(u)},) \sim \mathcal{M}(N_i^{(u)}, \mathbf{p}_i^{(u)}) \quad \text{and} \quad (\mathbf{n}_i^{(s)}|\mathbf{p}_i^{(s)}) \sim \mathcal{M}(N_i^{(s)}, \mathbf{p}_i^{(s)})$$

where $N_i^{\{s,u\}} = \sum_{k=1}^{2^K} n_{ik}^{\{s,u\}}$ are the number of cells collected and $\mathbf{p}_i^{(u)}$ and $\mathbf{p}_i^{(s)}$ are the unknown proportions for the un-stimulated and stimulated samples, respectively.

5.2 *Prior*

As in the one-biomarker case, we share information across subjects using an exchangeable prior on the unknown proportions. This time the beta priors are replaced by Dirichlet priors, such that

$$(\mathbf{p}_i^{(u)}|z_i = 0) \sim \text{Dir}(\boldsymbol{\alpha}^{(u)}),$$

$$(\mathbf{p}_i^{(s)}|z_i = 1) \sim \text{Dir}(\boldsymbol{\alpha}^{(s)}) \quad \text{and} \quad (\mathbf{p}_i^{(u)}|z_i = 1) \sim \text{Dir}(\boldsymbol{\alpha}^{(u)}),$$

where the indicator variable z_i is defined in Section 3.2, *i.e.*, $z_i \sim \text{Be}(w)$. As in the beta-binomial case, both an EM and MCMC algorithms can be used for parameter estimation. When using a fully Bayesian approach via MCMC, we use the same priors for $\boldsymbol{\alpha}^{\{u,s\}}$ and w as for the beta-binomial model.

5.3 *Parameter estimation*

Again, to simplify the estimation problem, we make use of the marginal likelihoods that can be obtained in closed forms (see Appendix C of the Supplementary Material). For the null component, the marginal likelihood L_0 is given by,

$$\text{L}_0(\boldsymbol{\alpha}^{(u)}|\mathbf{n}_i^{(s)}, \mathbf{n}_i^{(u)}) = \frac{\text{B}(\boldsymbol{\alpha}^{(u)} + \mathbf{n}_i^{(u)} + \mathbf{n}_i^{(s)})}{\text{B}(\boldsymbol{\alpha}^{(u)})} \cdot \frac{N_i^{(s)}!}{\prod_k n_{ik}^{(s)}!} \cdot \frac{N_i^{(u)}!}{\prod_k n_{ik}^{(u)}!},$$

where B is the 2^K-dimensional Beta function defined as $\text{B}(\boldsymbol{\alpha}) = \prod_k \Gamma(\alpha_k)/\Gamma(\sum_k \alpha_k)$. Similarly the marginal likelihood for the alternative model is given by

$$\text{L}_1(\boldsymbol{\alpha}^{(u)}, \boldsymbol{\alpha}_i^{(s)}|\mathbf{n}_i^{(s)}, \mathbf{n}_i^{(u)}) = \frac{\text{B}(\boldsymbol{\alpha}^{(u)} + \mathbf{n}_i^{(u)})\text{B}(\boldsymbol{\alpha}^{(s)} + \mathbf{n}_i^{(s)})}{\text{B}(\boldsymbol{\alpha}^{(s)})\text{B}(\boldsymbol{\alpha}^{(u)})} \cdot \frac{N_i^{(s)}!}{\prod_k n_{ik}^{(s)}!} \frac{N_i^{(u)}!}{\prod_k n_{ik}^{(u)}!}.$$

The estimation procedures (both EM and MCMC based) for the Dirichlet–multinomial distribution are the same as for the beta-binomial model except that the number of parameters to estimate is larger. We initialize the z_i in the EM algorithm with the positivity calls from the multivariate Fisher's exact test. In our experience, the performance of the EM algorithm greatly deteriorates for $K > 3$, and is more dependent on the initial values and can fail to converge in many instances. Although our MCMC algorithm is slightly more computational, it does not suffer from this problem and provides a robust alternative when K is large. More details about our multivariate MCMC algorithm is given in Appendix C of the Supplementary Material.

5.4 *Polyfunctionality in Fluidigm Single-Cell Gene Expression Data*

As a proof-of-concept, we applied our multivariate MIMOSA model for two specific genes in the Fluidigm data, namely CCR7 and GZMK. For this example, $K = 2$, and we have four possible combinations. In Figure 4 we show heatmaps of the counts of cells expressing all combinations of the CCR7 and GZMK genes in unstimulated and stimulated samples (Figure 4 A,B). Only CCL5 positive cells express BIRC3, and its expression increases upon stimulation. The typical approach to analyzing poly-functional populations from intracellular cytokine staining data (summing the counts over all possible polyfunctional cell populations as in IL2+ and/or IFN-γ+) would not be appropriate in this case, since changes in the counts of these different cell populations occur in both directions. That is, the number of BIRC3-/CCL5+ cells decreases upon stimulation, while the number of BIRC3+/CCL5+ cells increases. When marginalizing over these cell populations, no difference is apparent in any of the samples. In contrast, the multivariate MIMOSA model tests all polyfunctional cell subpopulations simultaneously, and detects significant differences between stimulated and unstimulated conditions in 13 of the 16 samples (Figure 4 D, black labels). Testing all combinations simultaneously is an advantage over performing multiple univariate tests on the subject combinations, which requires multiplicity adjustment and a potential loss of power.

Since the Fluidigm data set has a limited number of observations (100 cells and 16 samples), we could not look at more than two biomarkers at once. Therefore, we performed simulations in eight dimensions to assess the power of the multivariate MIMOSA model compared to Fisher's exact test and the likelihood ratio test on the resulting 2x8 tables (Figure 5 A-C). These results show that multivariate MIMOSA has significantly increased power to detect true differences in multivariate data, even with small counts and small effect sizes, and the model better fits the data than the competing standard approaches tested (Figure 5 B).

6. Discussion

Experimentalists already have access to a myriad of single-cell assays such as flow cytometry, mass cytometry and multiplexed quantitative-PCR. The development of effective statistical methods to detect differences in gene or protein expression at the single-cell level is becoming increasingly important as single-cell assays will become more routine and sequencing at the single-cell level eventually becomes practical (Ramsköld *and others*, 2012). Current approaches for single-cell assays such as the t-test, χ^2 test and Fisher's exact test are for the most part simplistic, and resulting inference can be quite sub-optimal, especially when the cell counts are small. Most importantly, these methods do not share information across samples, resulting in less power to detect true differences than empirical-Bayes and hierarchical modeling approaches, which are widely applied in the microarray literature (Kendziorski *and others*, 2003; Newton *and others*, 2001; Smyth *and others*, 2005; Gottardo *and others*, 2006). In addition, most of these methods are univariate in nature and inappropriate for high–dimensional, next–generation single-cell assays.

The MIMOSA model presented here uses a mixture model framework of beta-binomial or Dirichlet-multinomial distributions to model counts in experimental subjects across multiple conditions (*i.e.*, vaccine responders and non-responders). Information is shared across responders and non-responders through exchangeable beta or Dirichlet priors, increasing the power to detect true differences between treatment and control conditions compared to Fisher's exact test, even when the underlying model assumptions are violated (Figures 3 and Supplementary Figure 3). The univariate MIMOSA model based on the Beta-Binomial distribution allows us to constrain the alternative hypothesis to the case $p^{(s)} > p^{(u)}$, where the proportion of cells in the stimulated sample is strictly greater than the proportion of cells in the matched unstimulated sample. This has proven to be useful for the ICS data where stimulation induced changes are expected to be one-sided.

Our model is fit to dichotomized data (cells are either positive or negative). While this is the

standard approach to ICS data analysis, it is believed that the magnitude (fluorescence intensity) of the signal also carries information. Likewise for single-cell gene expression data, most methods have focused on differences in the continuous part of the signal, though some have addressed both changes in frequency of expression and expression levelMcDavid *and others* (2012). Extensions of the MIMSOA model, incorporating the continuous MFI or gene expression level are warranted in the future.

Although we used two single-cell assay platforms as motivating examples, our MIMOSA model can be applied to any type of single-cell assay where cells are dichotomized into positive and negative sets, counted and compared across different conditions. In the case of the Fluidigm data, most analysis methods have been focused on identifying differences in the continuous part of the signal ignoring cells that are undetected (*i.e.*, the gene is not expressed in the cell), or the information is used for pre-filtering (Flatz *and others*, 2011). The ability of MIMOSA to identify stimulation-specific expression patterns in single-cell gene expression data demonstrates not only the broader utility of the method, but importantly, also demonstrates that biologically relevant signal is present in the proportion of cells expressing each gene under different conditions (Figure 2 A-C).

Detecting differences in poly-functional cell populations (*i.e.*, identifying changes in cell populations that co-express multiple proteins, cytokines, or genes) is important in immunology, since it allows the identification of more precisely defined, more homogeneous cell populations (Milush *and others*, 2009). In the context of HIV, poly-functional cell populations have been shown to be correlated with long-term disease non-progression, while in the context of vaccination studies (e.g. in Leishmania) poly–functional responses have been correlated with protection from disease (Betts *and others*, 2006; Darrah *and others*, 2007; Precopio *and others*, 2007). In the ICS data used here, the stimulation is expected to increase only the number of antigen specific cells detected. Hence, if a specific cell subset expressing multiple biomarkers is being differentially ex-

pressed, differential expression based on the marginal cell counts should also be detected. As such, identifying poly-functional cytokine profiles from ICS data can be done in an iterative way. First, univariate tests on marginal populations are performed, and then specific cell subsets expressing the positive biomarkers detected are tested. However, this iterative (univariate) approach might not be satisfactory due to the large number of possible combinations that need to be tested, and a multivariate approach might be preferable. In that case, as others have pointed out, in order to have the most power to detect a true difference, the statistical test should be selected taking into account only the cytokine combinations of interest (Nason, 2006).

For two–sided changes, as with the Fluidigm data, changes in poly-functional cell populations are not always detectable when looking at the marginal populations (Figure 4 A-C). In this case, the use of multivariate model, as our Dirichlet-multinomial model, will become important to detect differential biomarker expression. Here, we have shown that MIMOSA has higher sensitivity and specificity than the competing methods to identify true differences between conditions in multivariate count data (Figure 4 A, and Figure 5 A,C), and the model generally provides a better fit to the single-cell assay count data obtained from studies with these types of experimental designs (Figure 5 B). Unfortunately, the limited number of samples in the Fluidigm data prevented us from looking at co-expression involving more than two genes. In the case of more than two biomarkers, the number of parameters to estimate for our Dirichlet–multinomial model is $2^{K+1}+1$, which is large even for moderate values of K. As an example, we would need both, a large number of subjects and a large number of events (cells) collected, to properly estimate the 33 parameters for $K = 4$.

A solution would be to explore alternative model parameterizations that could be used to reduce the number of required parameters. For example, one could assume that the hyper-parameters are constant across biomarker combinations, *i.e.*, $\alpha_k^{\{s,u\}} = \alpha^{\{s,u\}}$ for all k, and the number of parameters would be reduced to 3 for any K. As attractive as this might sound, such

a model would be unrealistic given that certain stimulations are known to induce expression of certain biomarkers more than others. More exploratory work will need to be done in this area once high dimensional single-cell level data with large number of samples become available.

All of the results presented here were obtained with a software implementation of the EM and MCMC MIMOSA models in R and C++, and is freely available from GitHub (`http://www.github.org/RGLab/MIMOSA`). An R package will soon be released as part of the Bioconductor project (Gentleman *and others*, 2004).

Supplementary Materials

Appendix A, B, and C and Supplementary Figures 1–3 referenced in Sections 2 and 3,4 and 5 are available in the Supplementary Material.

Acknowledgments

This work was supported by the Intramural Research Program of the National Institute of Allergy and Infectious Diseases (NIAID) and the National Institutes of Health (NIH) [R01 EB008400, U01 AI068635-01 to RG], grants from the Bill and Melinda Gates Foundation VISC (Vaccine Immunology Statistical Center) [grant numbers OPP38744, OPP1032317], a grant from the Bill & Melinda Gates Foundation to the CAVD (Collaboration for AIDS Vaccine Discovery) [grant number OPP1032325]. Study HVTN065 was conducted by the HIV Vaccine Trials Network (HVTN), and supported by the National Institute of Allergy and Infectious Diseases (NIAID). The the HVTN Laboratory Program is supported by the NIH [grant number UM1AI068618]. Funding was also provided by the Public Health Service from the NIH and the University of Washington Center for AIDS Research (CfAR), an NIH-funded program [grant numbers UM1 AI068618, P30 AI027757]. We thank the James B. Pendleton Charitable Trust for their generous equipment donation.

References

Altman, J D, Moss, P A, Goulder, P J, Barouch, D H, McHeyzer-Williams, M G, Bell, J I, McMichael, A J and Davis, M M. (1996, October). Phenotypic analysis of antigen-specific T lymphocytes. *Science (New York, NY)* **274**(5284), 94–96.

Bendall, SC, Simonds, EF, Qiu, P, Amir, ED, Krutzik, PO, Finck, R, Bruggner, RV, Melamed, R, Trejo, A and Ornatsky, OI. (2011). Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. *Science (New York, NY)* **332**(6030), 687.

Betts, Michael R, Nason, Martha C, West, Sadie M, De Rosa, Stephen C, Migueles, Stephen A, Abraham, Jonathan, Lederman, Michael M, Benito, Jose M, Goepfert, Paul A, Connors, Mark, Roederer, Mario *and others*. (2006, June). Hiv nonprogressors preferentially maintain highly functional hiv-specific cd8+ t cells. *Blood* **107**(12), 4781–4789.

Darrah, Patricia A, Patel, Dipti T, De Luca, Paula M, Lindsay, Ross W B, Davey, Dylan F, Flynn, Barbara J, Hoff, Søren T, Andersen, Peter, Reed, Steven G, Morris, Sheldon L, Roederer, Mario *and others*. (2007, July). Multifunctional TH1 cells define a correlate of vaccine-mediated protection against Leishmania major. *Nature Medicine* **13**(7), 843–850.

De Rosa, Stephen C, Lu, Fabien X, Yu, Joanne, Perfetto, Stephen P, Falloon, Judith, Moser, Susan, Evans, Thomas G, Koup, Richard, Miller, Christopher J and Roederer, Mario. (2004, November). Vaccination in humans generates broad t cell cytokine responses. *J Immunol* **173**(9), 5372–5380.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete

data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**(1), 1–38.

Flatz, Lukas, Roychoudhuri, Rahul, Honda, Mitsuo, Filali-Mouhim, Abdelali, Goulet, Jean-Philippe, Kettaf, Nadia, Lin, Min, Roederer, Mario, Haddad, Elias K, Sékaly, Rafick P *and others*. (2011, April). Single-cell gene-expression profiling reveals qualitatively distinct CD8 T cells elicited by different gene-based vaccines. *Proceedings of the National Academy of Sciences* **108**(14), 5724–5729.

Gelfand, AE. (1996). *Markov Chain Monte Carlo in practice, edited by Gilks WR*. Richardson S.

Gentleman, R, Carey, V, Bates, D, Bolstad, B, Dettling, M, Dudoit, S, Ellis, B, Gautier, L, Ge, Y and Gentry, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**(10), R80.

Goepfert, P A, Elizaga, M L, Sato, A, Qin, L, Cardinali, M, Hay, C M, Hural, J, DeRosa, S C, DeFawe, O D, Tomaras, G D, Montefiori, D C, Xu, Y, Lai, L, Kalams, S A, Baden, L R, Frey, S E, Blattner, W A, Wyatt, L S, Moss, B, Robinson, H L *and others*. (2011, January). Phase 1 Safety and Immunogenicity Testing of DNA and Recombinant Modified Vaccinia Ankara Vaccines Expressing HIV-1 Virus-like Particles. *J Infect Dis* **203**(5), 610–619.

Gottardo, R, Raftery, A E, Ka Yee, Y and Bumgarner, R. (2006, March). Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biometrics* **62**, 10–18.

Horton, H, Thomas, EP, Stucky, JA, Frank, I, Moodie, Z, Huang, Y, Chiu, YL, McElrath, MJ and De Rosa, SC. (2007). Optimization and validation of an 8-color intra-

cellular cytokine staining (ics) assay to quantify antigen-specific t cells induced by vaccination. *Journal of immunological methods* **323**(1), 39–54.

Ihaka, R and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**(3), 299–314.

Inokuma, Margaret, dela Rosa, Corazon, Schmitt, Charles, Haaland, Perry, Siebert, Janet, Petry, Douglas, Tang, Mengxiang, Suni, Maria A, Ghanekar, Smita A, Gladding, Daiva, Dunne, John F, Maino, Vernon C, Disis, Mary L *and others*. (2007, August). Functional T cell responses to tumor antigens in breast cancer patients have a distinct phenotype and cytokine signature. *J Immunol* **179**(4), 2627–2633.

Kendziorski, C. M., Newton, M. A., Lan, H. and Gould, M. N. (2003). On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22**(24), 3899–3914.

McDavid, Andrew, Finak, Greg, Chattopadyay, Pratip K., Dominguez, Maria, Lamoreaux, Laurie, Ma, Steven S., Roederer, Mario and Gottardo, Raphael. (2012, October). Data Exploration, Quality Control and Testing in Single-Cell qPCR-Based Gene Expression Experiments. *Bioinformatics*, 9.

McKinstry, K Kai, Strutt, Tara M and Swain, Susan L. (2010, May). The potential of CD4 T-cell memory. *Immunology* **130**(1), 1–9.

Milush, Jeffrey M, Long, Brian R, Snyder-Cappione, Jennifer E, Cappione, Amedeo J, York, Vanessa A, Ndhlovu, Lishomwa C, Lanier, Lewis L, Michaëlsson, Jakob and Nixon, Douglas F. (2009, November). Functionally distinct subsets of human NK cells and monocyte/DC-like cells identified by coexpression of CD56, CD7, and CD4. *Blood* **114**(23), 4823–4831.

Narsinh, Kazim H, Sun, Ning, Sanchez-Freire, Veronica, Lee, Andrew S, Almeida, Patricia, Hu, Shijun, Jan, Taha, Wilson, Kitchener D, Leong, Denise, Rosenberg, Jarrett, Yao, Mylene, Robbins, Robert C *and others*. (2011, March). Single cell transcriptional profiling reveals heterogeneity of human induced pluripotent stem cells. *Journal of Clinical Investigation* **121**(3), 1217–1221.

Nason, M. (2006, August). Patterns of Immune Response to a Vaccine or Virus as Measured by Intracellular Cytokine Staining in Flow Cytometry: Hypothesis Generation and Comparison of Groups. *Journal of Biopharmaceutical Statistics* **16**(4), 483–498.

Newton, M A, Kendziorski, C M, Richmond, C S, Blattner, F R and Tsui, K W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**(1), 37–52.

Newton, Michael A, Noueiry, Amine, Sarkar, Deepayan and Ahlquist, Paul. (2004, April). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics (Oxford, England)* **5**(2), 155–76.

Peiperl, Laurence, Morgan, Cecilia, Moodie, Zoe, Li, Hongli, Russell, Nina, Graham, Barney S, Tomaras, Georgia D, De Rosa, Stephen C, McElrath, M Juliana and the NIAID HIV Vaccine Trials Network. (2010, October). Safety and immunogenicity of a replication-defective adenovirus type 5 hiv vaccine in ad5-seronegative persons: A randomized clinical trial (hvtn 054). *PLoS ONE* **5**(10), e13579.

Pieprzyk, M. (2009, July). Fluidigm Dynamic Arrays provide a platform for single-cell gene expression analysis . *Nature Methods* **6**, iv.

Precopio, Melissa L, Betts, Michael R, Parrino, Janie, Price, David A, Gostick, Emma, Ambrozak, David R, Asher, Tedi E, Douek, Daniel C, Harari, Alexandre,

Pantaleo, Giuseppe, Bailer, Robert, Graham, Barney S, Roederer, Mario *and others*. (2007, June). Immunization with vaccinia virus induces polyfunctional and phenotypically distinctive CD8(+) T cell responses. *The Journal of experimental medicine* **204**(6), 1405–1416.

Proschan, Michael A and Nason, Martha. (2009, March). Conditioning in 2 x 2 tables. *Biometrics* **65**(1), 316–322.

Raftery, AE. (1996). *Markov Chain Monte Carlo in Practice - Walter R. Gilks, Sylvia Richardson, D. J. Spiegelhalter*, First CRC Press Reprint edition. 2000 N.W. Corporate Blvd. Boca Raton, Florida, 33431: Chapman Hall / CRC Press.

Raftery, Adrian E and Lewis, Steven M. (1992, November). [Practical Markov Chain Monte Carlo]: Comment: One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo. *STATISTICAL SCIENCE* **7**(4), 493–497.

Ramsköld, Daniel, Luo, Shujun, Wang, Yu-Chieh, Li, Robin, Deng, Qiaolin, Faridani, Omid R, Daniels, Gregory A, Khrebtukova, Irina, Loring, Jeanne F, Laurent, Louise C, Schroth, Gary P *and others*. (2012, July). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology* **30**, 777–82.

Sinclair, Elizabeth, Black, Douglas, Epling, C Lorrie, Carvidi, Alexander, Josefowicz, Steven Z, Bredt, Barry M and Jacobson, Mark A. (2004). CMV antigen-specific CD4+ and CD8+ T cell IFNgamma expression and proliferation responses in healthy CMV-seropositive individuals. *Viral immunology* **17**(3), 445–454.

Smyth, Gordon K, Michaud, Joëlle and Scott, Hamish S. (2005, May). Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics (Oxford, England)* **21**(9), 2067–2075.

Trigona, Wendy L, Clair, James H, Persaud, Natasha, Punt, Kara, Bachinsky, Margaret, Sadasivan-Nair, Usha, Dubey, Sheri, Tussey, Lynda, Fu, Tong-Ming and Shiver, John. (2003, July). Intracellular staining for HIV-specific IFN-gamma production: statistical analyses establish reproducibility and criteria for distinguishing positive responses. *Journal of interferon & cytokine research : the official journal of the International Society for Interferon and Cytokine Research* **23**(7), 369–377.

van Oudenaarden, Alexander. (2009). Nature, nurture, or just blind chance: Stochastic gene expression and its consequences. *Biophysical Journal* **96**(3, Supplement 1), 15a –.

Table 1. 2 x 2 contingency table of counts for biomarker positive and negative cells between stimulated (s) and unstimulated (u) conditions for a given subject i.

	Biomarker	
	Negative	Positive
Stimulated	$N_i^{(s)} - n_i^{(s)}$	$n_i^{(s)}$
Unstimulated	$N_i^{(u)} - n_i^{(u)}$	$n_i^{(u)}$

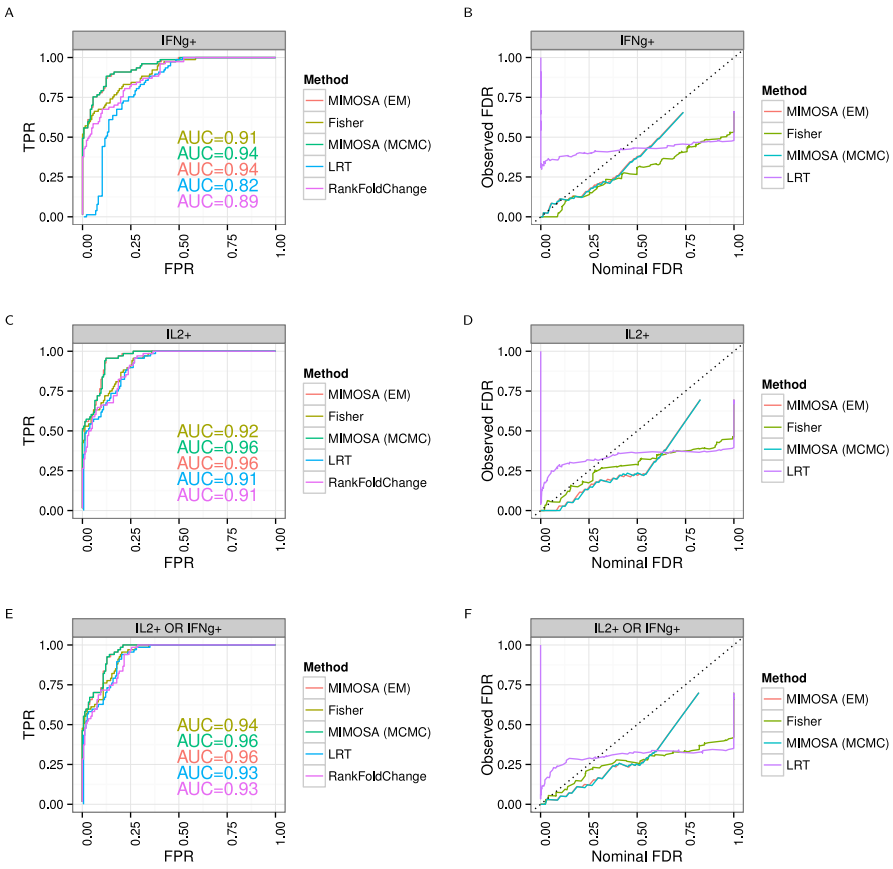

Fig. 1. Performance of MIMOSA (EM and MCMC implementations, one-sided model) and competing methods on ICS data from the example flow cytometry data set. Sensitivity and specificity (ROC analysis) as well as observed and nominal false discovery rates for positivity calls from CD4+ T-cells stimulated with A-B) ENV-1-PTEG and expressing IFN-γ or C-D) ENV-1-PTEG and expressing IL2. E-F) ENV-1-PTEG and expressing IFN-γ and/or IL2. ROC and FDR plots of other cytokine combinations can be found in Supplementary Figure 1.This figure appears in color in the electronic version of this article.

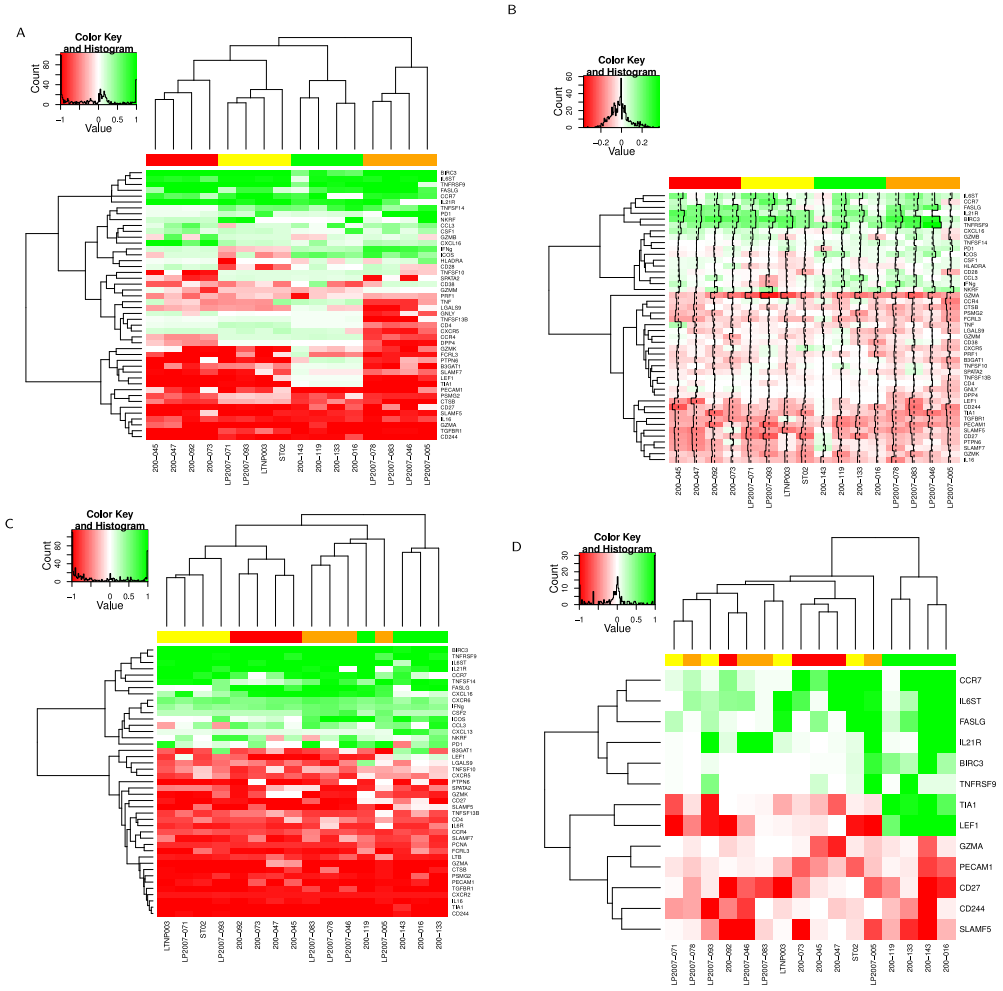

Fig. 2. Signed posterior probability and difference in empirical in proportions of single-cells expressing each gene on a 96x96 Fluidigm array. The posterior probability of response times the sign of the change in expression is shown in A) (red indicates a decrease, green an increase, relative to the control) for a model fit to each stimulation and gene separately. Columns and rows are clustered based on these signed posterior probabilities. B) The empirical differences in proportion of cells expressing a gene in the stimulated vs. control samples. Rows and columns are ordered as in A) for comparison. The traces show the deviations of each cell from zero. Colors along the columns denote different stimulations (green: CMV pp65 nlv5, red: HIV Gag, orange: HIV Nef, yellow: CMV pp65 tm10). C) Signed posterior probabilities for a model fit per gene but across all stimulations simultaneously. D) Clustering of the signed q-values from Fisher's exact test. Genes selected from Fisher's exact test at the 10% FDR level. This figure appears in color in the electronic version of this article.

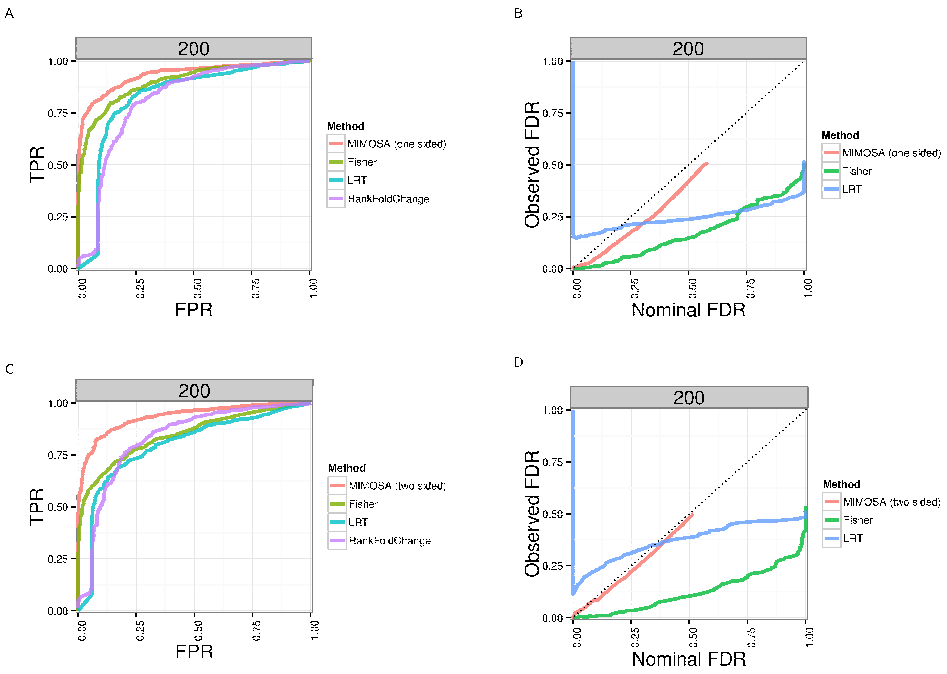

Fig. 3. Comparison of positivity detection methods on data simulated from the one-sided and two–sided models. Ten simulations were generated at an N of 50,000 total counts using hyper-parameter estimates from real ICS data (IFN-γ expressing CD4+ T-cells stimulated with ENV-1-PTEG from HVTN065) with a five-fold effect size between responder and non-responder components. A) Average ROC curve over the 10 simulated data sets (N=50,000), one–sided B) Average observed and nominal false discovery rate over 10 simulated data sets (N=50,000), one–sided. C) Average ROC curves, two–sided model. D) Average observed and nominal FDR, two–sided model. Curves are shown for MIMOSA, Fisher's exact test, the likelihood ratio test, and log fold-change. Results for MIMOSA fit to a model violating model assumptions, as well as other values of N are in Supplementary Figure 3, and for varying values of I (number of observations) are in Supplementary Figure 4. This figure appears in color in the electronic version of this article.

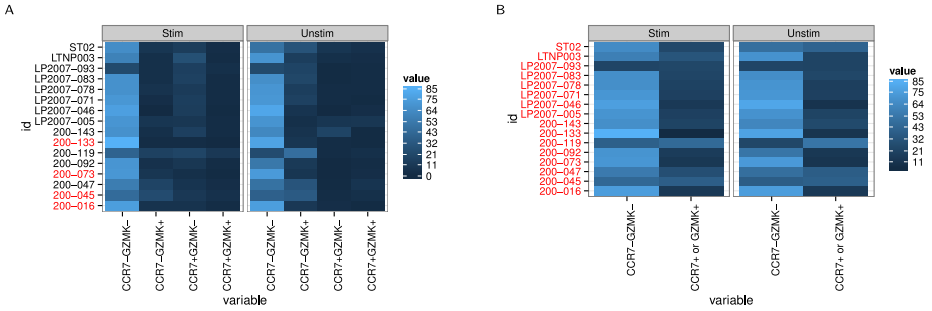

Fig. 4. Counts of cells expressing A) different combinations of CCR7 and GZMK genes in the unstimulated and stimulated conditions (+/+,+/-,-/+,-/-), and for B) the marginalized positive counts in stimulated and unstimulated conditions. No difference is observed from the marginalized counts, while multivariate MIMOSA detects a difference between stimulated and unstimulated conditions in 12 of 16 samples, while Fisher's test detects 9 of 16. Sample names highlighted in red identify those where MIMOSA did not detect a difference. This figure appears in color in the electronic version of this article.

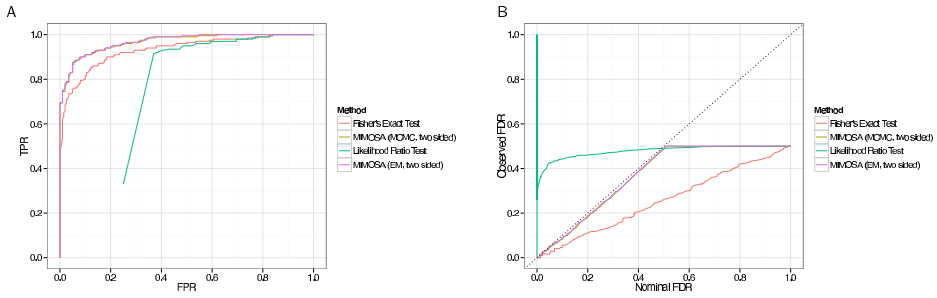

Fig. 5. Multivariate simulations from a two-sided model. Ten, eight-dimensional data sets were simulated from a two-sided model with an effect sizes of 2.5×10^{-3} and -2.5×10^{-3} in two of the eight dimensions (N=1,500). Multivariate MIMOSA was compared against Fisher's exact test, and the likelihood ratio test. A) Average ROC curves for the competing methods over 10 simulations. B) Average observed and nominal false discovery rate for each method over 10 simulations. This figure appears in color in the electronic version of this article.