BIOMETRICS , 1-19 DOI: 12.1234/j.12345-1234.2012.12345.x

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

Greg Finak[1,*], SC De Rosa[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

**email:* gfinak@fhcrc.org

***email:* sderosa@fhcrc.org

****email:* marior@mail.nih.gov

*****email:* rgotard@fhcrc.org

SUMMARY: Most biological samples (*e.g.*, blood and tissue) are composed of many functionally distinct cell subsets that 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 single-cell assays to 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 markers (or combinations of markers) that are differentially expressed between two biological conditions (*e.g.*, before/after vaccination), where expression is defined as the proportion of cells expressing that marker (or marker combination) in the cell subset(s) of interest. Here we present a Bayesian hierarchical framework based on beta-binomial mixture models for testing for differential marker expression using such single-cell assays. Our model allows the inference specific to each observation, while borrowing strength across observations 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 Fisher's exact test, a likelihood ratio test, and basic log-fold changes. Using several experimental immunological assays measuring proteins or genes at the single-cell level, we find that our method has higher sensitivity and specificity than alternative methods. Using a simulation study we 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 marker combinations using a multinomial-Dirichlet model.

KEY WORDS: Mixture modelling, hierarchical modelling, bayesian modelling, single-cell assays, immunology

This paper has been submitted for consideration for publication in *Biometrics*

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 markers (up to 17 markers). More recent technological development have drastically extended the capabilities of single-cell cytometry to measure dozens of simultaneous parameters per cell (Bendall et al., 2011). Although cells sorted using well-established surface markers may appear homogeneous, mRNA expression of other genes within these cells can be heterogeneous (Narsinh et al., 2011; Flatz et al., 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 markers simultaneously on individual cells in a heterogeneous mixture such as whole blood or peripheral blood cell mononuclease (PBMC), and are used for immune monitoring of disease, vaccine research, and diagnosis of haematological malignancies (Altman et al., 1996; Betts et al., 2006; Inokuma et al., 2007).

During analysis, cell level marker 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 markers. Using this matrix, one can form 2^K putative cell subsets. 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 marker combination, to identify individuals for whom the proportion of cells expressing that combination is significantly different between two experimental conditions (*e.g.*, before and after vaccination).

A motivating example from vaccine research is the flow cytometric intracellular cytokine staining (ICS) assay, which is used to identify and quantify individuals' 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 et al., 2010). These antigen-specific T-cell subpopulations 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 phenotypic markers (CD3, CD4, and CD8), used to subset T-cells, and functional markers (cytokines) used to define antigen specific T-cells (Horton et al., 2007; De Rosa et al., 2004; Betts et al., 2006). A sufficiently large number of cells must be collected 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 an individual to identify significant differences. individuals who generate a response after stimulation are called *responders* whereas individual that do not show any differences are called *non-responders*. In many immunological studies, the size of the functionally distinct subpopulations (*i.e.*, the number of positive cells) is very low (relative to the total number of cells), and real biological differences might be difficult to detect.

Although there is no standard approach to analyzing single-cell assays current methods range from ad-hoc rules based on log-fold changes, to permutation tests based on Hotelling's T^2 statistics, to exact tests of 2x2 contingency tables (*e.g.*, Fisher's exact test and χ^2 test) (Trigona et al., 2003; Sinclair et al., 2004; Horton et al., 2007; Nason, 2006; Peiperl et al., 2010; Proschan & Nason, 2009). All of these methods test observations separately, and

no information is shared across observations even though one could expect some similarities across responders (*resp.* non-responders).

The framework developed in this paper 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 individuals 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 individuals. This sharing of information across individuals helps regularize proportion estimates when the cell counts are small, which is typical with single-cell assays, and increase sensitivity and specificity when detecting responders. Because our framework is multivariate in nature, multiple cell subsets can be model simultaneously, which could help detect small biological changes that are spread out across multiple cell subsets (Nason, 2006).

2. Data structure and notation

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

Flow cytometry: The primary ICS data set is from a trial testing the GeoVax DNA and MVA vaccines in a prime-boost regimen with 120 individuals (98 vaccines and 20 placebo recipients, see Web Appendix A). The goal of this data set was to assess the immune response to the vaccine across multiple stimulations, time points, cytokines and T-cell subsets. Here, we analyze a subset of the data consisting of 98 individual from the vaccine group at two time points, day 0 and day 182. 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 individuals are expected to respond, for some cytokines at least. More details can be found in Web Appendix A.

Fluidigm single-cell gene expression: This is a single-cell gene expression data set of sorted CD8+ T-cells from sixteen individuals. T-cells isolated by flow cytometry from sixteen individuals were stimulated in blocks of four individuals with four different antigens (HIV Gag, HIV Nef, CMV pp65 tm10, 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 [1].

In the remainder of the paper, we use the following notation to describe our model. From this point on, we assume that we observe cell counts from I individuals in two conditions: stimulated and un-stimulated. Each cell can either be positive or negative for a marker. Given a set of K markers, the measured cells can be classified into 2^K positive/negative marker combinations. We denote by n_{sik} and n_{uik}, $k = 1 \dots 2^K$, the observed counts for the 2^K combinations in the stimulated and un-stimulated samples. We denote by $N_{si} = \sum_k n_{sik}$ and $N_{ui} = \sum_k n_{uik}$ the total number of cells measured for individual i in each sample. For ease of notations, we will denote by $\mathbf{y}_i$ the vector of observed counts for individual i, *i.e.* $\mathbf{y}_i = (\mathbf{n}_{si}, \mathbf{n}_{ui})$ where $\mathbf{n}_{si} = \{n_{sik} : k = 1, \dots, 2^K\}$ and $\mathbf{n}_{ui} = \{n_{uik} : k = 1, \dots, 2^K\}$. Finally, we define $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_I)$.

3. Differential expression with one marker

Datasets like the ones presented here are usually analyzed one marker at a times to avoid being underpowered due to the large number of combinations and the potential for very small cell counts in many of the combinations. As a consequence, we first consider the one marker case (*i.e.* $K = 1$) where marginal cell counts are analyzed separately. In this case, for

[1] Add a bit more info here, how many genes, tetramer sorted, etc, Mario et al. could you fill this in?

a given individual, the data can be summarized in a contingency table of $+/-$ cell counts across the un-stimulated and stimulated samples as depicted in Table 1.

[Table 1 about here.]

For a given individual and stimulation, we consider a marker 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. Individuals that show differential expression will be called responders for that marker. In this section, we shall be concerned with identifying differential expression one marker at a time, using a beta-binomial mixture model as described in what follows.

3.1 *Beta-binomial model*

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

$$(n_{si}|p_{si}) \sim \text{Bin}(N_{si}, p_{si}) \quad \text{and} \quad (n_{ui}|p_{ui}) \sim \text{Bin}(N_{ui}, p_{ui})$$

where p_{si}, p_{ui} are the unknown proportions for the stimulated and un-stimulated paired samples. In order to detect responding individuals we consider two competing models:

$$\mathcal{M}_0 : p_{ui} = p_{si} \quad \text{and} \quad \mathcal{M}_1 : p_{ui} \neq p_{si}.$$

Under the null model, $\mathcal{M}_0$, there is no difference between the stimulated and un-stimulated samples, and the proportions are equal. Under the alternative model, $\mathcal{M}_1$, there is a difference in proportions between the two samples and the individual i is a responder. In some study, 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 Web Appendix C and we refer to it as the one-sided model.

3.2 *Priors*

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

$$(p_{ui}|z_i = 0) \sim \text{Beta}(\alpha_u, \beta_u)$$

$$(p_{si}|z_i = 1) \sim \text{Beta}(\alpha_s, \beta_s) \quad \text{and} \quad (p_{ui}|z_i = 1) \sim \text{Beta}(\alpha_u, \beta_u)$$

where z_i is an indicator variable equal to one if individual 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 individuals. Note that the parameters α_u, β_u are explicitly shared across the two models, whereas α_s, β_s are only present in the alternative model. Finally, we assume that the z_i's are independent and identically distributed Bernoulli with probability w, where w represents the (unknown) proportion of responders. It follows that marginally, *i.e.* after integrating z_i, p_{ui} and p_{si} are jointly distributed as a mixture of a one dimensional and a two dimensional Beta distributions 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 and Expectation-Maximization (Dempster et al., 1977) algorithm as described in Section 3.3. As an alternative, we also describe a fully Bayesian model where the hyperparameters α_u, β_u and α_s, β_s are 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*

Our estimation algorithms make direct use of the marginal likelihoods, L_0 and L_1, obtained after integrating out the $p_{\{s,u\}i}$'s for the null and alternative models, which greatly simplify our calculations. Given the conjugacy of the priors, the marginal likelihoods L_0 and L_1 are

available in closed-forms (Web Appendix B), and are given by,

$$\mathrm{L}_0(\alpha_u, \beta_u|\mathbf{y}) = \prod_{i=1}^{P} \binom{N_{ui}}{n_{ui}} \binom{N_{si}}{n_{si}} \cdot \frac{\mathrm{B}(n_{si} + n_{ui} + \alpha_u, N_{si} - n_{si} + N_{ui} - n_{ui} + \beta_u)}{\mathrm{B}(\alpha_u, \beta_u)}$$

and

$$\mathrm{L}_1(\alpha_u, \beta_u, \alpha_s, \beta_s|\mathbf{y}) = \prod_{i=1}^{P} \binom{N_{ui}}{n_{ui}} \binom{N_{si}}{n_{si}} \cdot \frac{\mathrm{B}(n_{ui} + \alpha_u, N_{ui} - n_{ui} + \beta_u)}{\mathrm{B}(\alpha_u, \beta_u)} \cdot \frac{\mathrm{B}(n_{si} + \alpha_s, N_{si} - n_{si} + \beta_s)}{\mathrm{B}(\alpha_s, \beta_s)} \tag{1}$$

Assuming that the missing data, the z_i's, are known, we define the complete data log-likelihood as follows,

$$l(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) = \sum_i z_i l_0(\alpha_u, \beta_u|\mathbf{y}_i) + (1 - z_i) l_1(\alpha_u, \beta_u, \alpha_s, \beta_s|\mathbf{y}_i) + z_i \log(w) + (1 - z_i) \log(1 - w) \tag{2}$$

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 Web Appendix C.

3.4 *EM algoritm*

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 (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. 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_{ui}'s and p_{si}'s and use these to set the hyper-parameters to their method-of-moments estimates.

3.5 *MCMC algorithm*

Realizations were generated 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 done via Gibbs samplings. Details about the algorithms are given in Web Appendix B. We used the method of Raftery & Lewis (1992); Raftery (1996) to determine the number of iterations, based on a short pilot run of the sampler. For each dataset presented here, this suggested that a sample of no more than about 1,000,000 iterations with 50,000 burn-in iterations was sufficient to estimate standard posterior quantities. Guided by this, and leaving some margin, we used 2,000,000 iterations after 50,000 burn-ins 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. The performance of MIMOSA was evaluated and compared 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 for each data set.

4.1 *ICS*

Using the ICS data, an ROC (receiver operator characteristic) analysis was performed to assess the sensitivity and specificity of the one-sided MIMOSA model compared to a one-

sided Fisher's exact test, log fold-change, and a likelihood ratio test based on the MIMOSA model for identifying vaccine responders and non-responders. Observations at the day 0 time point were treated as true negatives, while observations at the day 182 time point were treated as true positives (potentially underestimating the sensitivity due to non-responders). We examined the CD4+ T-cell cytokine responses [2]. The MIMOSA model has higher sensitivity and specificity than Fisher's exact test, the likelihood ratio test, or log fold-change for discriminating vaccine responders and non-responders as shown by the ROC curves on Figure 1 A. In addition, MIMOSA gave estimates of the observed false discovery rate that are better or comparable to competing methods (Figure 1 B). These results are consistent for other cytokines (see Web Figure A).

[Figure 1 about here.]

4.2 *Single-cell gene expression*

We applied the MIMOSA model to Fluidigm single-cell gene expression. This time we used the two-sided MIMOSA model because genes could be up or down regulated upon stimulation. In order to detect stimulation specific changes of expression, our model was fit 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-individual 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 q-values (Figure 2 C) does not reveal any stimulation-specific patterns. Comparing the size of the gene sets, Fisher's exact test identified 47 significant genes while MIMOSA identified 50 significant genes, with 39 genes identified in common between the two methods.

[2]What about CD8+? Didn't look at CD8's

[Figure 2 about here.]

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. Using hyper parameters estimated from a one-sided MIMOSA model fit to data from ENV-1-stimulated, CD4-positive, IL2-expressing T-cells from the primary immunogenicity time point of the HVTN065 trial . We simulated data from this constrained model with 200 observations, a response rate of 60%, an N of 1,000, 5,000, 10,000, and 50,000 events, with ten independent realizations of data for each N. The one-sided MIMOSA model was fit to this data and the sensitivity and specificity of the model's ability to correctly identify observations from the "responder" and "non-responder" groups was evaluated through analysis of ROC curves, and compared against Fisher's exact test, the likelihood ratio test, and log fold-change. This procedure was repeated for the two-sided models fit to two-sided data (Figure 3 A-D). In addition, the nominal *vs.* observed FDR[3] was also examined to assess the ability of each method to properly estimate the FDR (Figure 3 A,B).

For both the constrained and unconstrained simulations, MIMOSA out-performed competing methods, including Fisher's exact test, with respect to sensitivity and specificity at all values of N (Figure 3 A and Web Figure B, panel A). Additionally, the estimated FDR for MIMOSA more closely reflected the nominal FDR compared to Fisher's exact test and competing methods (Figure 3 B and Web Figure B panel B).

[Figure 3 about here.]

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 on

[3]Need to define this earlier

$(0, 1)$, 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 α, β hyper parameters estimated from the HVTN065 data set (see Web Figure B 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 marker combinations

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

5.1 *Model*

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

$$(\mathbf{n}_{ui}|\mathbf{p}_{ui},) \sim \mathcal{M}(N_{ui}, \mathbf{p}_{ui}) \quad \text{and} \quad (\mathbf{n}_{si}|\mathbf{p}_{si}) \sim \mathcal{M}(N_{si}, \mathbf{p}_{si})$$

where $N_{\{s,u\}i} = \sum_{k=1}^{2^K} n_{\{s,u\}ik}$ are the number of cells collected and $\mathbf{p}_{ui}$ and $\mathbf{p}_{si}$ are the unknown proportions for the un-stimulated and stimulated samples.

5.2 *Prior*

As in the one-marker case, we share information across individuals using an exchangeable prior on the unknown proportions. This time the beta priors are replaced by Dirichlet priors, as follows,

$$(\mathbf{p}_{ui}|z_i = 0) \sim \text{Dir}(\boldsymbol{\alpha}_u)$$

$$(\mathbf{p}_{ui}|z_i = 1) \sim \text{Dir}(\boldsymbol{\alpha}_u) \quad \text{and} \quad (\mathbf{p}_{si}|z_i = 1) \sim \text{Dir}(\boldsymbol{\alpha}_s)$$

where the indicator variable z_i is as defined in Section 3.2, i.e. $z_i \sim \text{Be}(w)$ where w is the proportion of responders. 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 Web Appendix E). For the null component, the marginal likelihood L_0 is given by,

$$\text{L}_0(\boldsymbol{\alpha}_u|\mathbf{n}_s, \mathbf{n}_u) = \prod_{i=0}^{I} \frac{\text{B}(\boldsymbol{\alpha}_u + \mathbf{n}_{ui} + \mathbf{n}_{si})}{\text{B}(\boldsymbol{\alpha}_u)} \cdot \frac{N_{si}!}{\prod_k n_{sik}!} \cdot \frac{N_{ui}!}{\prod_k n_{uik}!}$$

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

$$\text{L}_1(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_s|\mathbf{n}_s, \mathbf{n}_u) = \prod_{i=0}^{I} \frac{\text{B}(\boldsymbol{\alpha}_u + \mathbf{n}_{ui})\text{B}(\boldsymbol{\alpha}_s + \mathbf{n}_{si})}{\text{B}(\boldsymbol{\alpha}_s)\text{B}(\boldsymbol{\alpha}_u)} \cdot \frac{N_{si}!}{\prod_k n_{sik}!} \cdot \frac{N_{ui}!}{\prod_k n_{uik}!}.$$

The estimation procedures (both EM and MCMC based) for the multinonial-Dirichlet are the same as for the beta-binomial model except that the number of parameters to be estimated is larger. The EM algorithm is now initialize with the multivariate Fisher's exact test[4]. In our experience, the performance of the EM algorithm greatly deteriorates when K becomes larger than three. It becomes 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 Web Appendix D.

[4]Greg is this correct?

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

As a proof-of-concept, we fitted our multivariate MIMOSA model looking at two specific genes in the Fluidigm data, namely BIRC3 and CCL5. In this case, $K = 2$ and we have four possible combinations. In Figure 4 we show heatmaps of the counts of cells expressing all combinations of the BIRC3 and CCL5 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)[5] would not be appropriate in this case, since changes in the counts of these different cell populations occur in both directions. That is to say, the number of BIRC3-/CCL5+ cells decreases upon stimulation and the number of BIRC3+/CCL5+ cells increases. When marginalizing over these cell populations, no difference is apparent in any of the samples. In contrast, multivariate MIMOSA tests all polyfunctional cell subpopulations simultaneously, identifying 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 individual combinations, which requires multiplicity adjustment and a potential loss of power.

[Figure 4 about here.]

Since the Fluidigm data has a limited number of observations (100 cells and 16 samples), we could not look at more than two makers at once. Therefore, we performed simulations in eight dimension to assess the power of the multivariate MIMOSA model compared to Fisher's exact test on the resulting 2x5 tables, as well as the likelihood ratio test (Figure 5 A-C). These results show that multivariate MIMOSA has significantly increased power to

[5] We have not discussed that yet. Perhaps we could add a sentence in the previous result section to discuss it and refer to the figure in supplementary material

detect true differences in multivariate data, even with small counts and small effect sizes, and the model is a better fit to the data than other standard approaches tested for analyzing such multivariate count data (Figure 5 B).

[Figure 5 about here.]

6. Discussion

Experimentalists have already access to a myriad of single-cell assays such as flow cytometry, mass cytometry and multiplexed quantitative-PCR, to name a few. Single-cell assays will become even more routine once sequencing at the single-cell level becomes practical (Ramsköld et al., 2012). As a consequence, the development of effective statistical methods to detect differences in gene or protein expression at the single-cell level is becoming increasingly important. Current approaches for single-cell assays are for the most part simplistic (t-test, χ^2 test, Fisher's exact test), and resulting inference can be quite unreliable 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 et al., 2003; Newton et al., 2001; Smyth et al., 2005). In addition, most of these methods are univariate in nature and inappropriate for high dimensional next generation single-cell assay.

The MIMOSA model presented here uses a mixture model framework of beta-binomial or Dirichlet-multinomial distributions to model counts in experimental individuals 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

Web Figure B). 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.

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 et al., 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 et al., 2009). In the context of HIV, these poly-functional cell populations have been shown to be correlated with good outcome such as vaccine protection and long-term disease non-progression (Betts et al., 2006; Darrah et al., 2007; Precopio et al., 2007). In the ICS data used here, the stimulation is expected to only increase the number of antigen specific cells detected. It follows that if a specific cell subset expressing multiple markers is being differentially expressed, 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 markers detected are then tested. Even in this situation, this iterative (univariate) approach might not be satisfactory due to the large number of possible combinations that needs to be tested, and a multivariate approach might be preferable. In this 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).

When changes are two-sided, 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 marker expression. Here we have shown that MIMOSA has higher sensitivity and specificity than these 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 arising 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 markers, there is a combinatorial explosion in the number of parameters involved in our multinomial-Dirichlet mixture model. The number of parameters to be estimated is $2^{K+1}+1$, which can become very large even for moderate values of K. As an example, $K = 4$ leads to 33 parameters, and this would require one to have a large number of individuals to properly estimate all parameters. A solution would be to explore alternative model parametrizations that could be used to reduce the number of required parameters. For example, one could assume that the hyper-parameters are constant across marker combinations, *i.e.* $\alpha_{\{s,u\}k} = \alpha_{\{s,u\}}$ 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 markers 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/finak/MIMOSA`). An R package will soon be released as part of the Bioconductor project[6].

Supplementary Materials

Web Appendices A, B, C, D, and E, and Web Figures A and B referenced in Sections 2 and 3,3.3, and 5 are available in the attached Web-based supplementary material.

Acknowledgments

We wish to acknowledge ...

References

Altman J. D., Moss P. A., Goulder P. J., Barouch D. H., McHeyzer-Williams M. G., Bell J. I., McMichael A. J., Davis M. M., 1996, Science (New York, NY), 274, 94

Bendall S. et al., 2011, Science (New York, NY), 332, 687

Betts M. R. et al., 2006, Blood, 107, 4781

Darrah P. A. et al., 2007, Nature Medicine, 13, 843

De Rosa S. C. et al., 2004, J Immunol, 173, 5372

Dempster A., Laird N., Rubin D., 1977, Journal of the Royal Statistical Society. Series B (Methodological), 1

Flatz L. et al., 2011, Proceedings of the National Academy of Sciences, 108, 5724

[6] Add reference

Gelfand A., 1996, Markov Chain Monte Carlo in practice, edited by Gilks WR. Richardson S.

Horton H. et al., 2007, Journal of immunological methods, 323, 39

Inokuma M. et al., 2007, J Immunol, 179, 2627

Kendziorski C., Newton M., Lan H., 2003, Statistics in ...

McKinstry K. K., Strutt T. M., Swain S. L., 2010, Immunology, 130, 1

Milush J. M. et al., 2009, Blood, 114, 4823

Narsinh K. H. et al., 2011, Journal of Clinical Investigation, 121, 1217

Nason M., 2006, Journal of Biopharmaceutical Statistics, 16, 483

Newton M. A., Kendziorski C. M., Richmond C. S., Blattner F. R., Tsui K. W., 2001, Journal of Computational Biology, 8, 37

Peiperl L. et al., 2010, PLoS ONE, 5, e13579

Pieprzyk M., 2009, Nature Methods

Precopio M. L. et al., 2007, The Journal of experimental medicine, 204, 1405

Proschan M. A., Nason M., 2009, Biometrics, 65, 316

Raftery A., 1996, Markov chain Monte Carlo in practice

Raftery A. E., Lewis S. M., 1992, STATISTICAL SCIENCE, 7, 493

Ramsköld D. et al., 2012, Nature Biotechnology

Sinclair E., Black D., Epling C. L., Carvidi A., Josefowicz S. Z., Bredt B. M., Jacobson M. A., 2004, Viral immunology, 17, 445

Smyth G. K., Michaud J., Scott H. S., 2005, Bioinformatics (Oxford, England), 21, 2067

Trigona W. L. et al., 2003, Journal of interferon & cytokine research : the official journal of the International Society for Interferon and Cytokine Research, 23, 369

van Oudenaarden A., 2009, Biophysical Journal, 96, 15a

Received August 2012.

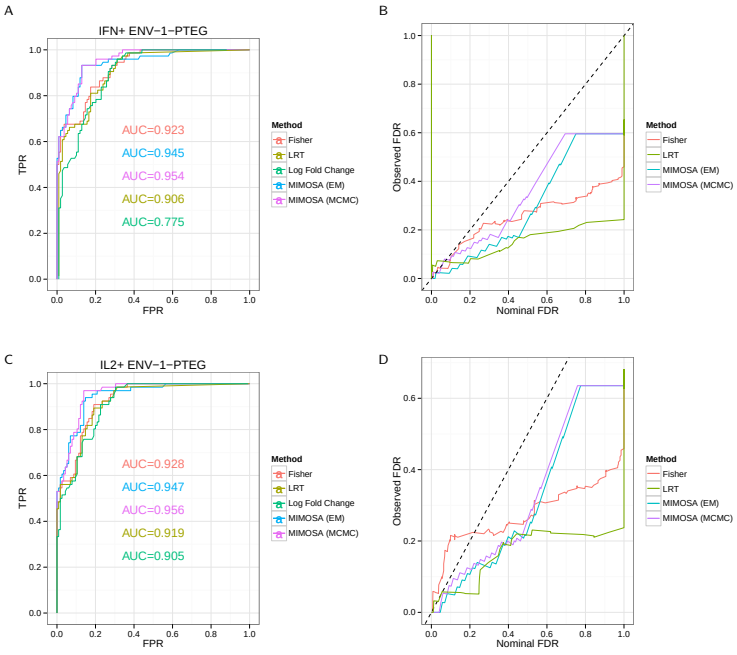

Figure 1. Performance of MIMOSA (EM and MCMC implementations, one-sided model) and competing methods on ICS data from HVTN065. Sensitivity and specificity (ROC analysis) as well as observed and nominal false discovery rates for positivity calls from CD4+ T-cells stimulated with A-B) ENV-1-PTEG and expressing IFNg or C-D) ENV-1-PTEG and expressing IL2. ROC and FDR plots of other cytokine combinations can be found in Web Figure A.

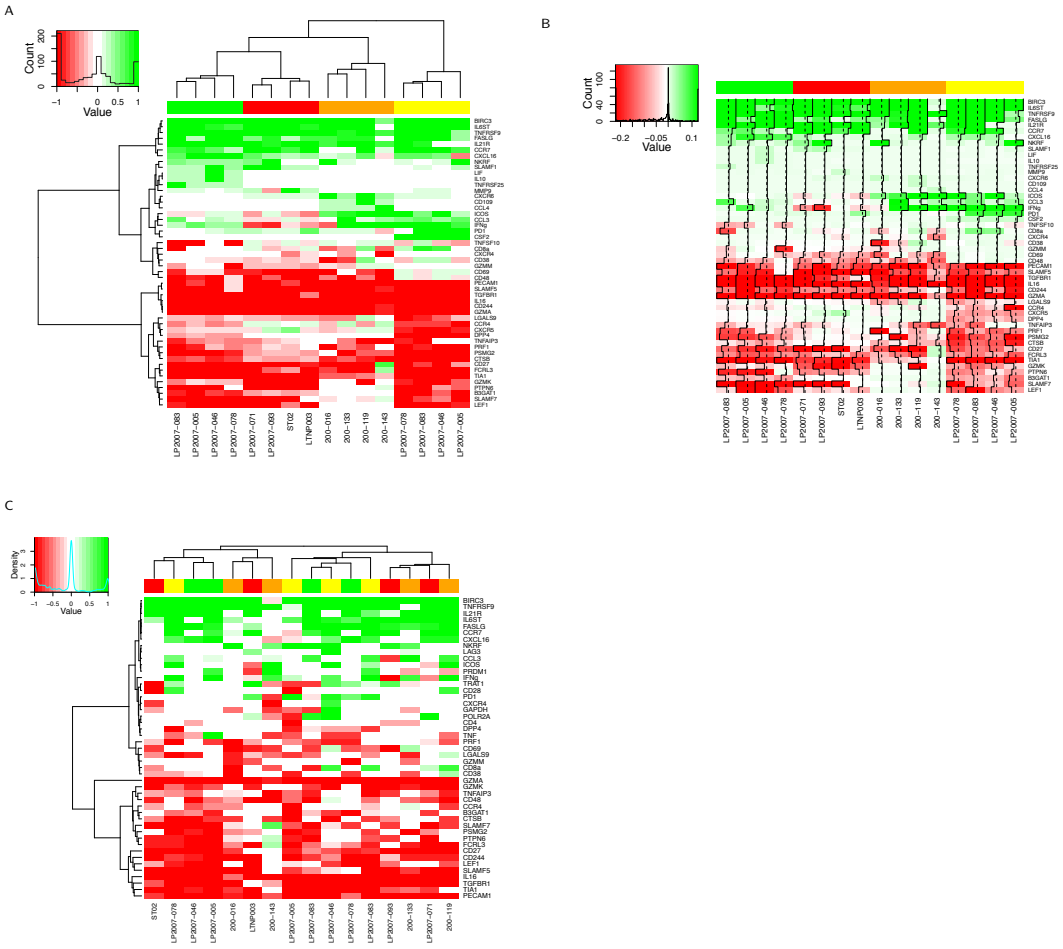

Figure 2. Signed posterior probability, difference and log-odds ratio of the proportion of single-cells expressing each gene on a 96x96 Fluidigm array. The posterior probability of response times the sign of the change in expression is shown in A) (red indicates a decrease, green an increase, relative to the control). Columns and rows are clustered based on these signed posterior probabilities. B) The posterior 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) Clustering of the signed q-values from Fisher's exact test. Genes selected from Fisher's exact test at the 10% FDR level.

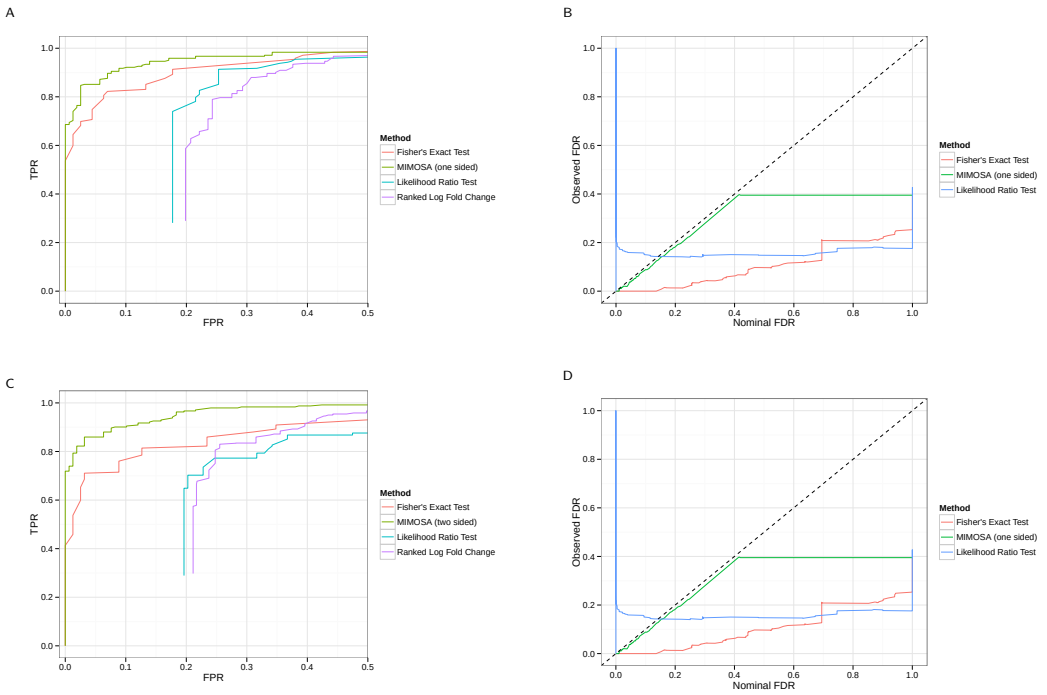

Figure 3. Comparison of positivity detection methods on data simulated from the one-sided and two–sided models. Ten simulations were generated at an N of 5,000 total counts using hyper-parameter estimates from real ICS data (IFNg expressing CD4+ T-cells stimulated with ENV-1-PTEG from HVTN065) with a five-fold effect size between responder and non-responder components. A) Average ROC curve over the 10 simulated data sets (N=5,000), one–sided B) Average observed and nominal false discovery rate over 10 simulated data sets (N=5,000), 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 Web Figure B.

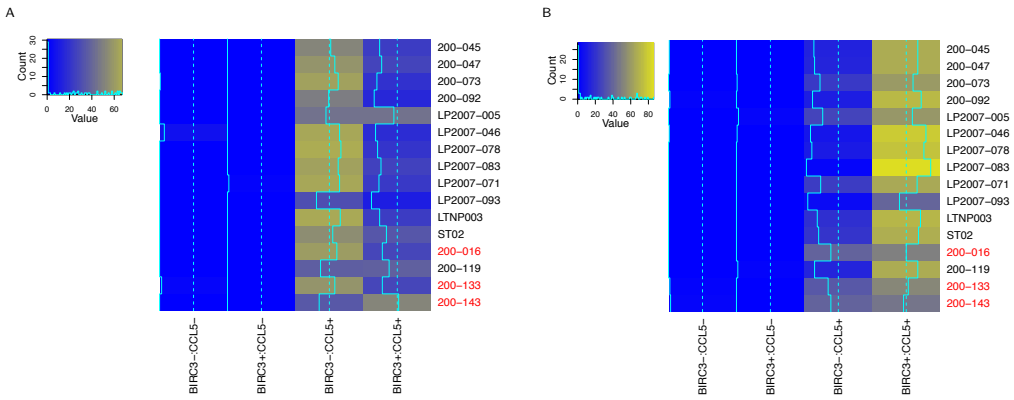

Figure 4. Counts of cells expressing different combinations of BIRC3 and CCL5 genes in the A) unstimulated and B) stimulated conditions. No difference is observed from the marginalized counts, while multivariate MIMOSA detects a difference between stimulated and unstimulated conditions in 13 of 16 samples. Sample names highlighted in red identify those where MIMOSA did not detect a difference.

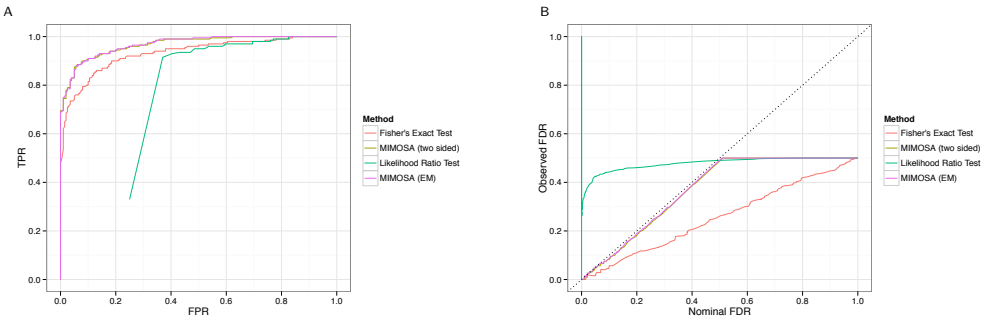

Figure 5. Multivariate simulations from a two-sided model. Ten, eight-dimensional data sets were simulated from a two-sided model with an effect sizes of 2.5×10^{-3} and -2.5×10^{-3} in two of the eight dimensions (N=1,500). Multivariate MIMOSA was compared against Fisher's exact test, and the likelihood 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.

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

	Marker	
	Negative	Positive
Stimulated	$N_{si} - n_{si}$	n_{si}
Unstimulated	$N_{ui} - n_{ui}$	n_{ui}