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

Greg Finak[1], SC De Rosa[2], and 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

June 18, 2012

Abstract

Immunological endpoints in vaccine trials are measured through a variety of assays that provide single–cell measurements of multiple genes and proteins in specific immunological cell populations. Using single–cell data, we consider the problem of identifying subjects where these cell populations exhibit differential responses under different experimental conditions. For example, in the intracellular cytokine staining assay from flow cytometry, individual cells are classified as either positive or negative for a marker based on a predetermined threshold. The assay is used to assess an individual's immune response to a vaccine by measuring the number of antigen–specific cells producing different cytokines in different T–cell subpopulations in response to different antigen stimulations. Individuals whose T–cells exhibit increased production of a cytokine in response to stimulation are termed "positive" for that cytokine, and multiple such "positivity calls" are used to identify vaccine responders. Here we present a framework based on mixtures of Beta–binomial or Dirichlet–Multinomial distributions for analyzing count data derived from such single–cell assays. Our method models cellular responses in a marker–specific manner, treating the responding and non–responding observations as separate components in the model. Cell counts from the different experimental conditions are modelled independently, while sharing information across responding and non–responding observations through empirical Bayes priors in order to increase the sensitivity and specificity of positivity calls. We compare our method against Fisher's exact test, a likelihood ratio test, and ranked log fold changes, and show how it can be extended to model multivariate (multiple markers) cellular responses. In simulations and applied to real data sets, we find that our method has higher sensitivity and specificity than Fisher's exact test and alternative methods.

1 Introduction

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. Although cells sorted using well-established surface markers may appear homogeneous, mRNA expression of other genes within these cells can be heterogeneous4,5 and could further characterize cell poly-functionality. 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 genes6. 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. 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 variables simultaneously on individual cells in a heterogeneous mixture such as whole blood. These variables are used to classify individual cells in the mixture into more homogeneous sub–populations based on phenotypic or functional differences. Such single–cell assays are used for immune monitoring of disease, vaccine research, and diagnosis of haematological malignancies [1–3].

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–cels 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 [4]. 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 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) and functional markers (cytokines) of the cell subpopulations of interest [2, 5, 6]. 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. Assessing a broad T cell response to a vaccine is particularly important in HIV vaccine trials, where the search for immune correlates of protection against HIV progression and infection is ongoing [5, 7–9].

Although there is no standard approach to analyzing ICS assay data 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[1]) [5, 10–12]. These methods generally test pairwise combinations of markers, raising questions of appropriate multiple testing adjustments, or they perform global tests of significance on multiple markers resulting in decreased power to detect small changes in subsets of cytokines [12, 13]. In the context of single-cell gene expression, very few methods have been proposed. To date, published analysis of Fluidigm data ignore cells where transcripts are undetected, focusing solely on continuous gene expression measurements. However, the proportion of single cells expressing individual genes also carries information and should be evaluated.

The framework developed in this paper addresses these issues explicitly. We present a multinomial-Dirichlet combinatorial mixture model framework for the analysis of single–cell assay data where multivariate measurement are made on individual cells. The model is used to identify observations where a significant difference exists between paired treatment and control samples with respect to the number of cells expressing different combinations of proteins, genes, or other measured properties of the cells. Importantly, our approach shares information across individuals (or subjects) by means of a Dirichlet prior distribution placed on the unknown cell population proportions of the mutinomial likelihood, which help increase sensitivity and specificity to detect rare antigen specific responding cell populations. The cell counts from the stimulated and unstimulated experiments are modeled independently, and different combinations of markers are represented as different mixture components in the framework. This approach allows us to omit certain combinations of markers that are not observable (i.e., two cytokines may never be co–expressed), or to explicitly represent combinations that are of interest (i.e., we can explicitly represent and test for a difference in a pair of cytokines in an experiment measuring many cytokines). This is a flexible approach that avoids the some of the drawbacks of global tests when only a few of the many measured markers show differences [12].

[1] make sure there is a reference for that

2 Data structure and notation

In this paper we consider two different immunological single-cell assays typically used in vaccine trials, one flow cytometry data set, and one single-cell gene expression data set.

Flow cytometry: This dataset is from a phase-I (safety and efficacy) trial of an adenoviral vector vaccine in individuals without prior immunity, measuring four cytokines via intracellular cytokine staining (ICS) in two cell populations from 20 individuals at two time points (zero and 28 days post-vaccination, see Supplementary Information S.1 for details) [14]. The statistical analysis of responder[2] and non-responder calls in the published trial is described in the original manuscript and outlined in the supplementary information (Supplementary Information **??**) [14]. The goal of this data set was to assess and quantify response rates of CD4 and CD8 T-cell populations to different antigens.

Fluidigm single-cell gene expression: This is a single-cell gene expression data set of flow-sorted CD8 T-cells from sixteen individuals. The T-cells from blocks of four individuals were stimulated with different antigens, gene expression measured at the single-cell level and compared to paired, unstimulated controls. [3].

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 where cell counts are marginalized and each marker is analyzed separately and $K = 1$. In this case, for 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.

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

[2] need to define this

[3] Add a bit more info here, how many genes, tetramer sorted, etc

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}

number of positive cells in the un-stimulated sample. Individuals that show differential expression for a given marker 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 One Marker

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

$$(n_{si}|p_{si}) \sim \text{Bin}(N_{si}, p_{si}) \quad \text{and} \quad (n_{ui}|p_{ui}) \sim \text{Bin}(N_{ui}, p_{ui}) \tag{1}$$

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:

$$\mathcal{M}_0 : p_{ui} = p_{si} \quad \text{and} \quad \mathcal{M}_1 : p_{ui} \neq p_{si}. \tag{2}$$

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.

3.2 Priors

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

$$(p_{ui}|z_i = 0) \sim \text{Beta}(\alpha_u, \beta_u) \tag{3}$$

$$(p_{si}|z_i = 1) \sim \text{Beta}(\alpha_s, \beta_s) \quad \text{and} \quad (p_{ui}|z_i = 1) \sim \text{Beta}(\alpha_u, \beta_u) \tag{4}$$

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 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 [15] algorithm as described in Section 3.3. As an alternative, we will also explore 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, to simplify our calculations. Given the conjugacy of the priors, the marginal likelihoods L_0 and L_1 are available in closed forms (Supplementary material), and are given by,

$$\mathrm{L}_0(\alpha_u, \beta_u|\mathbf{y}) = \prod_{i=1}^{P} \binom{N_{ui}}{n_{ui}} \binom{N_{si}}{n_{si}} \frac{\mathrm{B}(n_{si} + n_{ui} + \alpha_u, N_{si} - n_{si} + N_{ui} - n_{ui} + \beta_u)}{\mathrm{B}(\alpha_u, \beta_u)} \quad (5)$$

and

$$\mathrm{L}_1(\alpha_u, \beta_u, \alpha_s, \beta_s|\mathbf{y}) = \prod_{i=1}^{P} \binom{N_{ui}}{n_{ui}} \binom{N_{si}}{n_{si}} \frac{\mathrm{B}(n_{ui} + \alpha_u, N_{ui} - n_{ui} + \beta_u)}{\mathrm{B}(\alpha_u, \beta_u)} \frac{\mathrm{B}(n_{si} + \alpha_s, N_{si} - n_{si} + \beta_s)}{\mathrm{B}(\alpha_s, \beta_s)}. \quad (6)$$

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

$$l(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) = \sum_i z_i l_0(\alpha_u, \beta_u|\mathbf{y}_i) + (1-z_i) l_1(\alpha_u, \beta_u, \alpha_s, \beta_s|\mathbf{y}_i) + z_i \log(w) + (1-z_i)\log(1-w) \quad (7)$$

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.

3.3.1 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 (7). 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 algorithms 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.

3.3.2 MCMC algorithm

Realizations were generated from the posterior distribution via MCMC algorithms [16]. 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 Supplementary material. We used the method of Raftery and Lewis [17, 18] 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

The constrained model was applied to an ICS data set from a real–world vaccine trial in order to identify responders to antigen stimulation. The unconstrained model was applied to Fluidigm single–cell gene expression data to identify genes differentially expressed between stimulated and unstimulated conditions in populations of single–cells. We also performed simulation studies to assess the performance of the constrained and unconstrained models in a univariate and multivariate settings.

4.1 ICS

MIMOSA Outperforms Competing Methods on Vaccine Trial Data from Study HVTN065

We tested our method on ICS data from HVTN065, a trial testing the GeoVax DNA and MVA vaccines in a prime-boost regimen, with 98 individuals. We examined time points from day 0 and day 182 (the primary endpoint).We examined the CD4+ T–cell cytokine responses for ENV-1-PTEG and GAG-1-PTEG stimulated samples. An ROC (receiver operator characteristic) analysis was performed to assess the sensitivity and specificity of the constrained MIMOSA model compared to Fisher's exact test, ranked log fold change,

and a likelihood ratio test based on the MIMOSA model for identifying vaccine responders and non–responders.

4.2 Fluidigm

We applied the MIMOSA model to Fluidigm single–cell gene expression data from CD8+ T–cells from 16 individuals, under four different stimulation conditions, as well as unstimulated samples. The unconstrained MIMOSA model was fit separately to for each stimulation, and the posterior probabilities of response for each gene, the posterior differences in the proportion of cells expressing each gene between stimulated and unstimulated conditions, as well as the posterior log ratio of the proportion of cells expressing each gene in stimulated and unstimulated conditions were calculated and heat maps generated (Figure **??** A-C).

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 the model fit of the constrained model to data from Gag1 stimulated CD4–positive, IL2 expressing T–cells on day 28 from real data, we simulated data from the constrained model with 500 observations, a response rate of 60%, an N of 10K, 20K, 30K, 50K, 75K, 100K, and 150K events, and ten independent realizations for each N. The constrained 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" components was evaluated through ROC curve analysis and compared against Fisher's exact test, the likelihood ratio test, and ranked log fold change. (Figure 4). The nominal vs observed false discovery rate was also examined to assess the model fit.

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 4 A, B). The false discovery rate for MIMOSA more closely reflected the nominal false discovery rate than Fisher's exact test (Figure 4 C, 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 on $(0, 1)$, rather than beta distributions. The means and variances of the truncated normal distributions were set to the MLE estimates of the beta distributions defined by the α, β hyper parameters estimated from the HVTN065 data set (Figure **??**). Even under these departures from the model assumptions, the unconstrained MIMOSA model outperformed Fisher's exact test and performed about as well as the constrained model fit to constrained data.

5 Differential expression across marker combinations

Our beta-binomial model described in section **??** 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,

$$(\mathbf{n}_{ui}|\mathbf{p}_{ui},) \sim \mathcal{M}(N_{ui}, \mathbf{p}_{ui}) \quad \text{and} \quad (\mathbf{n}_{si}|\mathbf{p}_{si}) \sim \mathcal{M}(N_{si}, \mathbf{p}_{si}) \tag{8}$$

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 subjects using an exchangeable prior on the unknown proportions. This time the beta priors are replaced by Dirichlet priors, as follows,

$$\begin{aligned} &(\mathbf{p}_{ui}|z_i = 0) \sim \text{Dir}(\boldsymbol{\alpha}_u) \\ &(\mathbf{p}_{ui}|z_i = 1) \sim \text{Dir}(\boldsymbol{\alpha}_u) \quad \text{and} \quad (\mathbf{p}_{si}|z_i = 1) \sim \text{Dir}(\boldsymbol{\alpha}_s) \end{aligned} \tag{9}$$

where the indicator variable z_i is as defined in (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 supplementary material). For the null component, the marginal likelihood L_0 is given by,

$$\text{L}_0(\boldsymbol{\alpha}_u|\mathbf{n}_s, \mathbf{n}_u) = \prod_{i=0}^{I} \frac{\text{B}(\boldsymbol{\alpha}_u + \mathbf{n}_{ui} + \mathbf{n}_{si})}{\text{B}(\boldsymbol{\alpha}_u)} \cdot \frac{N_{si}!}{\prod_k n_{sik}!} \cdot \frac{N_{ui}!}{\prod_k n_{uik}!} \tag{10}$$

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

$$\mathrm{L}_1(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_s | \mathbf{n}_s, \mathbf{n}_u) = \prod_{i=0}^{I} \frac{\mathrm{B}(\boldsymbol{\alpha}_u + \mathbf{n}_{ui})\mathrm{B}(\boldsymbol{\alpha}_s + \mathbf{n}_{si})}{\mathrm{B}(\boldsymbol{\alpha}_s)\mathrm{B}(\boldsymbol{\alpha}_u)} \cdot \frac{N_{si}!}{\prod_k n_{sik}!} \cdot \frac{N_{ui}!}{\prod_k n_{uik}!}. \tag{11}$$

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. In our experience, the performance of the EM algorithm greatly deteriorates when K becomes larger than 2. The EM algorithms becomes very dependent on the initial values, and can even fail to converge when good initial values are provided. Although our MCMC algorithm is slightly more computational, it does not suffer from this problem and can be used even when K is large. More details about both algorithms are given in Supplementary material.

6 Discussion

The variety of single–cell assays being adopted by the immunology community is increasing. Flow cytometry, mass cytometry, ELISPOT, Fluidigm, and other single–cell assays can all be analyzed as single–cell count data. Development of effective statistical methods to detect differences in gene or protein expression at the single cell level is becoming increasingly important. Current approaches rely on asymptotic approximations (t–test, or χ^2 test), empirical or ad–hoc methods (2–fold change), or exact tests (Fisher's exact test) where model assumptions are generally not satisfied, all of which can lead to invalid conclusions about the data [5, 10, 11, 13, 19]. Most importantly, existing classical methods do not share information across samples, resulting in less power to detect true differences than empirical-Bayes and hierarchical modelling approaches, which are widely applied in the microarray literature [20–22].

The MIMOSA model presented here uses a mixture model framework of Beta–Binomial or Multinomial–Dirichlet distributions to model cell counts in non–responding and responding individuals. Information is shared across non–responders and across responders through an empirical–Bayes prior, increasing the power to detect true differences between treatment and control conditions compared to Fisher's exact test, even when model assumptions are violated (Figures 4 and **??**). The 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, our simulations show that we can relax this assumption to the unconstrained case where $p_s \neq p_u$ without compromising the observed false positive or false negative rates (Figures 4 and **??**), while saving on computation time.

Fix references to non-existent figure

The analysis of real–world ICS data from a vaccine trial demonstrated that the MIMOSA model outperformed Fisher's exact test in detecting individuals who were responders and non–responders to vaccine across multiple antigen stimulations and multiple cytokines (Figure **??**). Importantly, the MIMOSA model was fit within each antigen stimulation × cytokine × time point combination but the model was blinded to placebo and vaccine treatment. Despite this, MIMOSA demonstrated a lower false positive rate at day zero (increased specificity), where no response is expected, than Fisher's exact test (Figure **??** Gag2, IL2, CD4+ T–cells, day 0, treatment group T1). Similarly, none of the placebo treated samples were identified as responders by the MIMOSA model (Figure **??** treatment group C). MIMOSA also exhibited an increased sensitivity in real–world data compared to Fisher's exact test, as demonstrated by the higher response rates for 7 of the 8 stimulation/cytokine/cell type combinations at day 28 to which the model was fit, and where a response is expected to be observed. Although it is not possible to validate the responders identified at day 28 for all cases, the conclusion of increased sensitivity is consistent with both the performance of the model in simulations, and with the increased specificity observed on day 0 and in the placebo treated samples.

Fix references to non–existent figure

Polyfunctionality (identifying cells coexpressing multiple proteins, cytokines, or genes, is important in single cells assays since it allows the identification of functionally distinct subsets of cells in heterogeneous cell populations [23]. In vaccine trials, these polyfunctional cell subsets may be correlated with vaccine response or efficacy. Existing methods for identifying polyfunctional cytokine profiles from ICS data either test individual combinations of cytokines separately, perform global tests at the expense of decreased power when only specific differences are of interest, sometimes empirical in nature, and generally don't share information across observations [5, 10–13, 19]. As others have pointed out, in order to have the most power to detect a true difference, the statistical test selected to detect differences in various combinations of cytokines should take into consideration the cytokine combinations that are of interest [12]. A global test, one that tests for a difference between all proportions of cytokine expressing cells will have lower power than one testing for difference in a specific cytokine combination if only that specific combination is truly different. This approach is inherent in our mixture model framework, where each mixture component representing some pattern of cytokine expression that is of interest.

We show a proof of principle example with two components to identify responder and non–responder observations for the four–dimensional (two cytokine) case. Sharing of parameters across nested models in the two–cytokine case allows us to reduce the number of parameters. The model is easily extended. First, to multiple components to identify not just responders and non–responders, but responders with respect to each combination of cytokines. Second, extensions to multiple cytokines are straightforward, only requiring increasing the dimensionality of the Multinomial–Dirichlet distribution. With the increase in dimensionality there is a combinatorial explosion in the number of combinations of cy-

tokines, and the resulting number of parameters in the model. We present several possible approaches to reducing the number of parameters in the model by sharing the hyperparamters across different combinations of cytokines. A number of these strategies can be generalized to models with multiple components that allow us to identify observations with differences in specific cytokine combinations between treatment and control samples.

7 Conclusions

We have developed a combinatorial mixture model framework for identifying differences between treatment conditions in paired observations of cell counts from a variety of single–cell assays. The software is implemented in R and C++, and is freely available from GitHub (`http://www.github.org/finak/MIMOSA`).

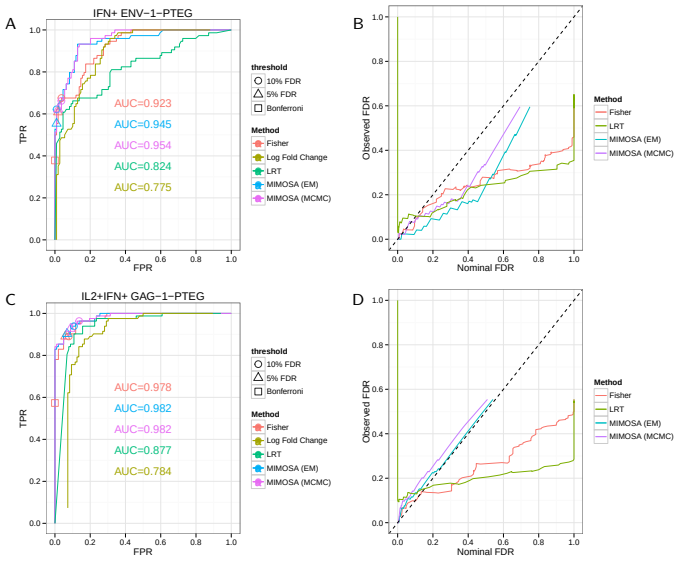

Figure 1: Performance of MIMOSA (EM and MCMC implementations, one–sided model) and competing methods on ICS data from HVTN065. Sensitivity and specificity (ROC analysis) as well as observed and nominal false discovery rates for positivity calls from CD4+ T–cells stimulated with A–B) ENV-1-PTEG and expressing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

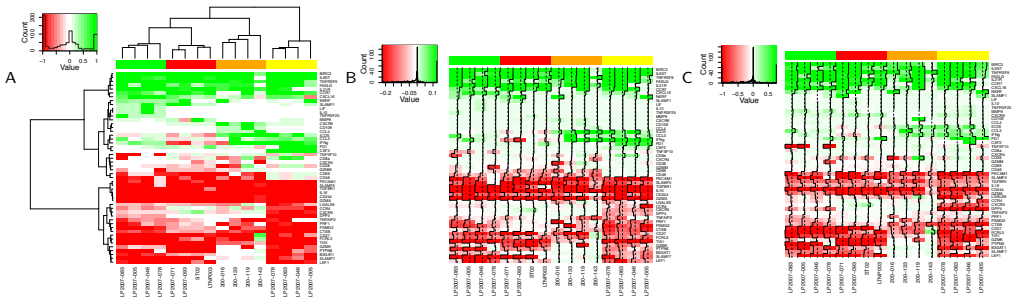

Figure 2: Signed posterior probability, difference and log-odds ratio of the proportion of single cells expressing each gene on a 96x96 Fluidigm array. The posterior probability of response times the sign of the change in expression is shown in A) (red indicates a significant decrease, green a significant increase, relative to the control). Columns and rows are clustered based on these signed posterior probabilities. B) The log_2 ratio of the proportion of cells expressing a gene in the stimulated vs. control samples. Rows and columns are ordered as in A) for comparison. C) The difference in the proportion of cells expressing each gene in the stimulated vs. control samples. Ordering of the rows and columns is preserved as in A). The traces show the deviations of each cell from zero.

Supplementary Information

S.1 HVTN065 Vaccine Trial ICS Dataset Description

HVTN065 is a phase 1 (safety and immunogenicity) trial of GeoVax HIV/AIDS DNA and MVA vaccine in 120 individuals (100 vaccinees, 20 placebo recipients, parts A and B). CD4 and CD8 T–cell epitope specific immune responses were measured via the ICS assay. Other humoral and cellular immune responses were measured via ELISA, and neutralizing antibody assays. Cytokines measured in the ICS assay included IFNg, TNFa, IL2, and IL4, and antigens included three Env, three Gag, and three Pol peptide pools. Results of the trial have been published [24].

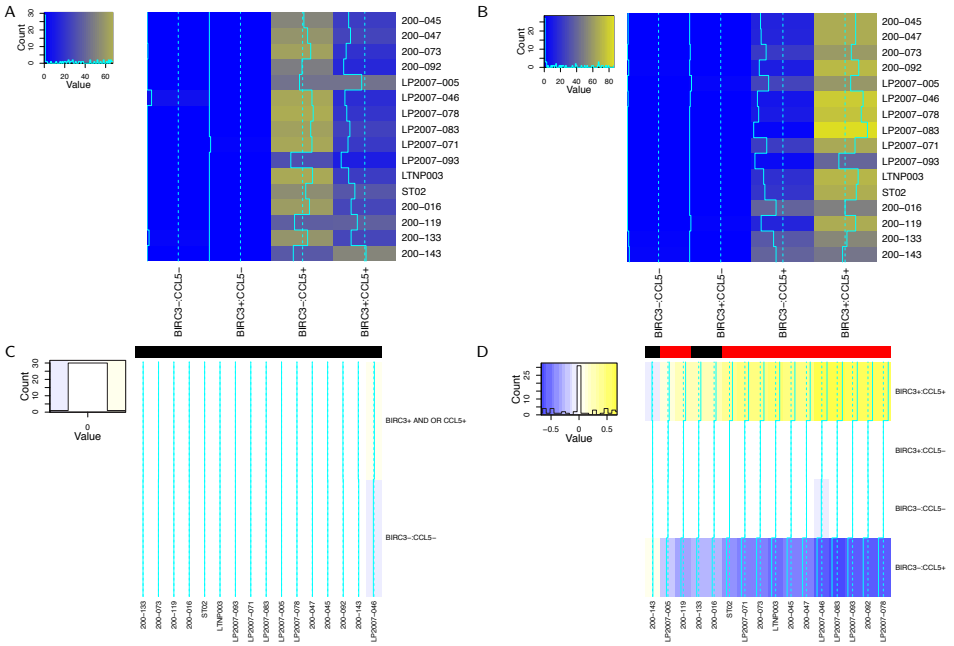

Figure 3: Counts of cells expressing different combinations of BIRC3 and CCL5 genes in the A) unstimulated and B) stimulated conditions. The posterior difference in proportions between stimulated and unstimulated samples fitting the C) marginalized counts D) multivariate combinations. No difference is observed from the marginalized counts, while multivariate MIMOSA detects a difference between stimulated and unstimulated conditions in 13 of 16 samples.

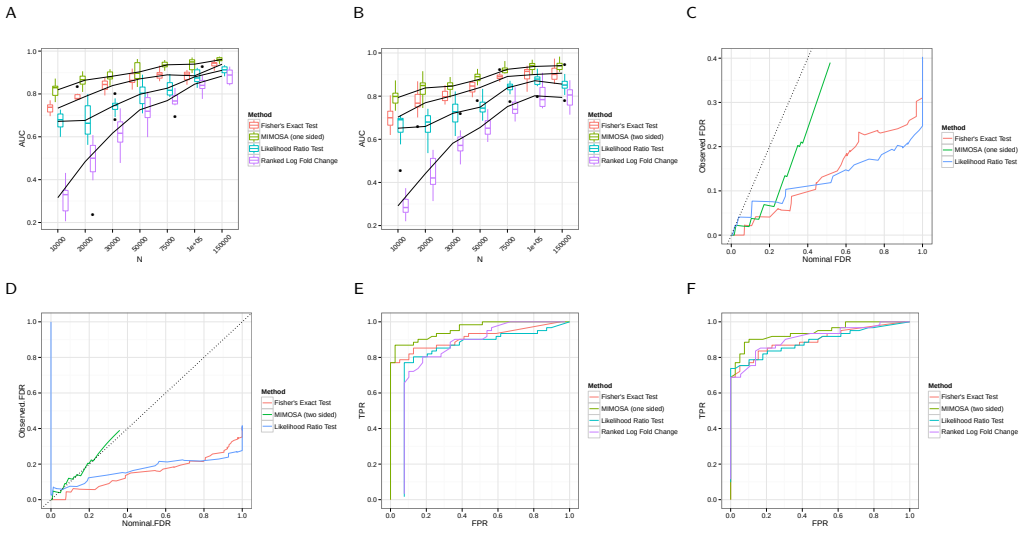

Figure 4: Comparison of positivity detection methods on data simulated from the one–sided model (top row) and the two–sided model (bottom row). Ten simulations were generated at each of seven increasing values of N (total counts) using hyper–parameter estimates from real ICS data. A) Boxplots of the area under the ROC curves for Fisher's exact test, MIMOSA (MCMC constrained model), likelihood ratio test, and the ranked log fold change. B) Boxplots for the unconstrained model. C) Observed and nominal false discovery rate for a representative simulated data set (N=100,000 counts) MIMOSA (one–sided), Fisher's exact test, and the likelihood ratio test. C) ROC curves for a representative simulated data set (N=100,000 counts). D) AUC boxplots, E) observed vs. nominal FDR, F) ROC analysis comparing MIMOSA (MCMC unconstrained model) and competing methods for two–sided simulated data.

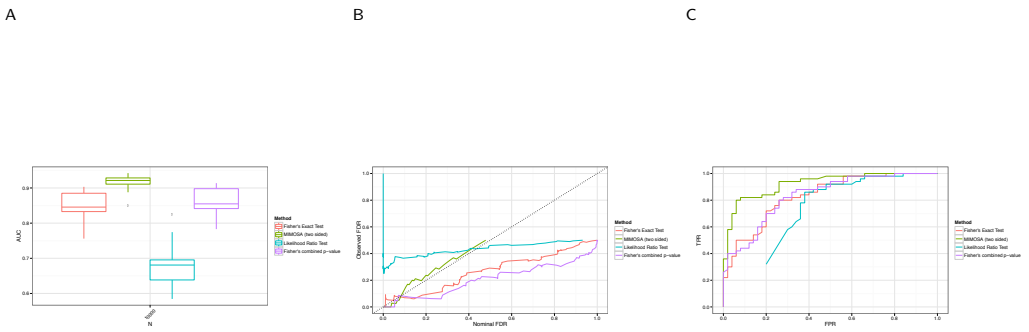

Figure 5: Multivariate simulations from a two–sided model. Ten, five-dimensional data sets were simulated from a two–sided model with an effect size of 2.5e-3 and -2.5e-3 in two of the five dimensions. Multivariate MIMOSA was compared against Fisher's exact test, the likelihood ratio test, and Fisher's combined p–value, combining Fisher's exact test run marginally on each of the five dimensions. A) Boxplots of AUCs (area under the curve) for the ten simulations. B) Observed and nominal false discovery rate for each method. C)ROC curves for the competing methods.

References

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

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

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

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

[5] H Horton, EP Thomas, JA Stucky, I Frank, Z Moodie, Y Huang, YL Chiu, MJ McElrath, and SC De Rosa. Optimization and validation of an 8-color intracellular cytokine staining (ics) assay to quantify antigen-specific t cells induced by vaccination. *Journal of immunological methods*, 323(1):39–54, 2007.

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

[7] S Plotkin. Correlates of protection induced by vaccination. *Clinical and Vaccine Immunology*, 2010.

[8] Jerome H Kim, Supachai Rerks-Ngarm, Jean-Louis Excler, and Nelson L Michael. Hiv vaccines: lessons learned and the way forward. *Current opinion in HIV and AIDS*, 5 (5):428–434, September 2010.

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

[10] Wendy L Trigona, James H Clair, Natasha Persaud, Kara Punt, Margaret Bachinsky, Usha Sadasivan-Nair, Sheri Dubey, Lynda Tussey, Tong-Ming Fu, and John Shiver. 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, July 2003.

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

[12] M Nason. 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, August 2006.

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

[14] Laurence Peiperl, Cecilia Morgan, Zoe Moodie, Hongli Li, Nina Russell, Barney S Graham, Georgia D Tomaras, Stephen C De Rosa, M Juliana McElrath, and the NIAID HIV Vaccine Trials Network. 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, October 2010.

[15] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[16] AE Gelfand. *Markov Chain Monte Carlo in practice, edited by Gilks WR.* Richardson S., 1996.

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

[18] AE Raftery. Markov Chain Monte Carlo in Practice - Walter R. Gilks, Sylvia Richardson, D. J. Spiegelhalter - Google Books. *Markov chain Monte Carlo in practice*, 1996.

[19] Marcus Dittrich and Paul V Lehmann. Statistical analysis of ELISPOT assays. *Methods Mol Biol*, 792:173–183, 2012.

[20] CM Kendziorski, MA Newton, and H Lan. On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles - Kendziorski - 2003 - Statistics in Medicine - Wiley Online Library. *Statistics in ...*, 2003.

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

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

[23] Jeffrey M Milush, Brian R Long, Jennifer E Snyder-Cappione, Amedeo J Cappione, Vanessa A York, Lishomwa C Ndhlovu, Lewis L Lanier, Jakob Michaëlsson, and Douglas F Nixon. 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, November 2009.

[24] Nilu Goonetilleke, Stephen Moore, Len Dally, Nicola Winstone, Inese Cebere, Abdul Mahmoud, Susana Pinheiro, Geraldine Gillespie, Denise Brown, Vanessa Loach, Joanna Roberts, Ana Guimaraes-Walker, Peter Hayes, Kelley Loughran, Carole Smith, Jan De Bont, Carl Verlinde, Danii Vooijs, Claudia Schmidt, Mark Boaz, Jill Gilmour, Pat Fast, Lucy Dorrell, Tomas Hanke, and Andrew J McMichael. Induction of multifunctional human immunodeficiency virus type 1 (HIV-1)-specific T cells capable of proliferation in healthy subjects by using a prime-boost regimen of DNA- and modified vaccinia virus Ankara-vectored vaccines expressing HIV-1 Gag coupled to CD8+ T-cell epitopes. *J Virol*, 80(10):4717–4728, May 2006.

[25] F Hahne, N LeMeur, RR Brinkman, B Ellis, P Haaland, D Sarkar, J Spidlen, E Strain, and R Gentleman. flowcore: a bioconductor package for high throughput flow cytometry. *BMC Bioinformatics*, 10(1):106, 2009.