Dear Dr. Molenberghs,

We thank you for taking the time to review our manuscript and appreciate your thoughtful comments and suggestions. We have now substantially revised our manuscript, and we believe to have addressed all comments. In what follows, our responses are highlighted in **bold**, while the reviewer's and associate editor's comments are in normal font.

Reviewer: 1

Comments to the Author

The authors present a beta-binomial mixture model for inferring differential expression from single-cell assay experiments. These assays are used to measure the presence or absence of combinations of predetermined biomarkers on a cell-by-cell basis and the challenge is to determine whether cells expressing these combinations are present in different proportions in cases compared to paired controls. Two inference approaches are presented: one uses the EM algorithm and a fully Bayesian alternative uses MCMC. The methods have been applied to two datasets with relevance to vaccine research and to simulated data. Comparisons have been made to frequentist approaches, namely Fisher's exact test, likelihood ratio test and mean log fold changes. A multivariate (Dirichlet-multinomial) version is also presented with which inference can be done using counts for all test markers simultaneously.

In my view the proposed model incorporates information-sharing across subjects in a sensible fashion and it appears to be a clear improvement over the alternatives considered. Here are a few specific comments.

- I found the use of _u and _s subscripts to label unstimulated and stimulated samples somewhat confusing in conjunction with _i to index the subjects (i = (1,...,I)) and _k to index the biomarker combinations (k = (1,...,K)). Perhaps an alternative notation to denote subjects as stimulated/unstimulated (e.g. bracketed superscripts) will make the manuscript easier to follow. Also, sometimes the authors have written "p_ui" and other times "p_iu"

We have made the changes to the notation as suggested, using bracketed, superscript s and u to denote stimulated and unstimulated subjects. The issue of p_ui and p_iu notation consistency has also been resolved.

- Could the authors provide a rough estimate of the magnitude of N for each dataset (and K for the Fluidigm dataset) early on in the manuscript. Also a rough idea of the magnitude of the p's would give a clearer sense of how many positive cells one would expect in the resting vs. activated state.

We have added text to the introduction of the data sets early in the manuscript indicating the magnitude of N for stimulated and unstimulated samples (approximately 50K for stimulated and 58K for unstimulated) in the HVTN data, and the size of K, the number of genes (96 genes) and the number of single cells (approximately 90 per subject per stimulation) in the Fluidigm data. We have also indicated the magnitude of the p's (on the order of 10^{-4}) as well as the order of the observed effect size, and the number of positive cells for the HVTN data set in the Data introduction section. We have added volcano plots of effect size vs posterior probability of response for each model fit to the HVTN data to the Supplementary Materials as Supplementary Figure 6.

- "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)" - if there are a small number of non-responders at day 182, then we can have confidence that method A is better than method B as long as the difference between the ROC curves is sufficiently large, but we can't have much confidence if the curves are very similar (a method that appears slightly worse might actually not be because the ground truth is wrong (by ground truth I mean what has been defined to be true, in this case that all subjects are responders at day 182)). Could the authors comment on what % of subjects they expect are true responders and assess, for different levels of inaccuracies in the ground truth, how these inaccuracies would translate to inaccuracies in the ROC curves?

We don't have an independent external measurement of the true response rate for day 182 vaccinees. Furthermore, the functional definition of vaccine responders is somewhat tautological, *i.e.,* a responder is any vaccine recipient detected as such at a post-vaccine time point. Since this depends on the sensitivity of the method applied, we can only reliably evaluate and control the sensitivity and specificity. In practice, any method that discriminates better between day 0 and day 182 placebo recipients or day 0 vaccinees vs. day 182 vaccinees is inherently preferred.

As a consequence of the above, in the manuscript, "true responders" were defined as any vaccine recipient at day 182, while non-responders were defined as all day 0 samples, and all samples from placebo recipients at day 182 (since those subjects should not have a vaccine response at any time point). In order to evaluate inaccuracies in this scheme we performed simulations where the true response rate in the vaccinee group at day 182 was varied from 20% to 100% but was assumed to be 100% in the evaluation, consistent with the assumptions above. The impact on the ROC curves was to shift them all down, while preserving the ordering of the methods. We have addressed this in the results and added an illustrative figure based on simulation in the Supplementary Material as Supplementary Figure 2.

- What posterior summary of the z_i's was used for the ranking?

The posterior mean of z_i, i.e. the posterior probability that individual *i* is a responder, is used to summarize the posterior z_i's.

- How was the FDR estimated using MIMOSA's posterior summaries (the Storey method cited uses to p- values only, does it not?)

We used the approach of Newton et al[1]. We have clarified it.

- For the simulation study, 200 individuals were used. I presume information-sharing for that many individuals results in a greater improvement than for fewer individuals - could the authors show results for different values of I?

We have performed additional simulations that examine the change in sensitivity as a function of I, the number of observations. Our method performs better than competing approaches down although the nominal false discovery rate tended to be an overestimate of the observed false discovery rate as the number of subjects decreased. We have included these additional simulations as Supplementary Figure 4.

- Could the authors comment on the effect of dichotomising the underlying signal and whether any importance should be given to the information lost in the thresholding?

Dichotomizing the underlying signal is the current standard approach to analyzing

ICS data. In the case of either flow or Fluidigm single-cell gene expression, dichotomizing may lead to some loss of information. In fluidigm, the continuous Ct values are known to carry information about the expression level and are particularly important for detecting real differences when the proportion of cells expressing a gene is high, and consequently the differences in proportions of cells between conditions are small. For ICS data, it is believed that the magnitude of fluorescence, conditional on positive expression, also carries information. This makes intuitive sense, as one would be more likely to believe that expression of several "positive" cells close to the thresholding boundary is more likely due to noisy fluctuations, compared to if those cells were expressing protein at levels well beyond the thresholding boundary. Extensions of our method in this direction, incorporating the continuous measurements, are certainly warranted in the future, however since dichotomization is the current standard approach, it makes sense as a starting point for modelling. We have added a paragraph in the discussion section.

- Given inference is performed on proportions, mightn't there be a need for normalisation approaches if certain subsets are much more highly expressed in one condition than the other? C.f. methods for RNA- seq normalisation. E.g. Robinson & Oshlack, 2010.

The normalization proposed by Robinson & Oshlack is meant to correct for biases due to technical limitations of high throughput sequencing, where differences in composition of total RNA between samples can impact the estimated proportions of reads for individual genes because all samples or genes are sequenced to the same depth. Oshlack states that their focus is in ensuring that genes with the same read counts across samples are not called DE. This is not an issue in our case for several reasons. First, we are looking at very sensitive single-cell assays where the individual cell is a natural unit of normalization, and we do not have issues with RNA composition.
We consider the univariate case first. Although we are counting cells, we do not care if the absolute number of positive cells for a subset is the same or different across samples. We are interested in whether the proportion of positive cells relative to the total number of cells differs. For the multivariate case, we are not interested in identifying which specific subset of cells differs in proportion relative to the total, only that a difference exists somewhere across the measured subsets. Thus, using Robinson's example, if there is a subset of cells in one sample that is not expressed in another sample, this would alter the proportions of all other measured cell subsets, but this is exactly what we wish to detect in our approach, since the total cell count is biologically meaningful (e.g., total CD4 or CD8 T-cells).

Reviewer: 2

Comments to the Author

Finak et al. present a method to analyse single-cell experiments with uni- or multivariate binary outcomes; the main application being vaccination studies, where one tests cells for the presence of specific biomarkers that indicate that the cell has recognized an antigen an is responding to it. Typically, cells from several subjects are analysed in this manner but, as the authors argue, the state of the art is to analyse the data from each subject independently. They suggest to use a hierarchical Bayesian model to share information across subjects to improve inferential power.

The method seems to be technically sound and the presented analyses of real and simulated data convincingly demonstrate the method's usefulness. This is a very decent piece of work that is definitely well suited for publication in Biostatistics. The writing style and presentation is overall good. The authors also made a commendable effort to explain the biological background for the benefit of

readers without an immunology background.

Nevertheless, I am left with a couple of question, which, I hope, should be straight-forward to address by clarifying the manuscript at a few places.

- The output of the inferential method is a call indicating whether the cells gathered from a specific subject at a specific time point show response to stimulation with an antibody, i.e., whether, for a specific sample, the probability of a cell showing a marker or marker combination is increased if the cell has been stimulated. The power increase is gained by sharing information across subjects. I am unclear, however, how and to which extent data is shared across time points.

Specifically, the abstract promises a method to "identify biomarkers [...] that are differentially expressed [...] e.g. before/after vaccination". Such a comparison between time points is not done in the paper: the inference is within a time point. This needs to be clarified. What actually is the role of the data from before vaccination? Is it just a negative control, i.e., any positive call is an indication of a problem? Or is it used as additional source of information on p_u?

In the analysis presented in the paper we do not compare pre-vaccination against post-vaccination time points against each other directly. Rather, we have paired samples for pre- and post-vaccination time points, as well as corresponding negative controls. Observations within each time point are compared against the corresponding negative controls, and we fit the model to observations from all time points simultaneously. Consequently all time points that are fit simultaneously will contribute to the priors for p_u. A subsequent evaluation of pre- and post-vaccination time points would involve comparing the positivity calls for paired samples at the two time points. However, data are not typically analyzed this way, as this could lead to confounding of date with vaccine effect (i.e. batch effects). Furthermore, vaccine trials do not always collect baseline samples. We feel that the sentence in the abstract was perhaps confusing. We have revised it, and give further justification in the paper why direct comparisons between time points are not typically performed.

- The description of a typical vaccination assay would be more concrete if some rough numbers were given. How many T cells are typically used per subject (thousands? or millions?) and which fraction reacts to the antigen (a few percent? one in a million?)? How drastic are the changes to this fraction upon stimulation in a responder? Is the strength of response in the presented experiments typical or low, i.e., is this typical or particularly challenging data to analyse?

These clarifications were also requested by the first reviewer, and we have provided this information in the introduction and background.

- For the concrete experiment, we also need more numbers: How many subjects? Which time points? What were the inferred values for p_u and p_s? (Maybe quote posterior mean and SD to give the reader a feeling for the values and uncertainties.)

We have provided this information as Supplementary Figure 7.

- The description in Section 4.1 is slightly confusing. Have I got it right in the following? the authors used only the data from day 0 and from day 182 (but no further time points), gave this to the model without revealing the assignment of the samples to the time points to the inferential scheme and then made the ROC curve by considering the time-point information as proxy for the true responder status of a sample?

The reviewer is correct in their interpretation. We have clarified the paragraph and

description as suggested. We have also added simulations in Supplementary Figure 2 that demonstrate what happends to the ROC and FDR curves when the assumption doesn't hold that all vaccinees at the post-vaccine timepoint are responders.

Maybe this paragraph needs a bit of rephrasing.

- In Figure 1, FDR is well controlled, but in Suppl. Fig. 1, all methods violate FDR control for some biomarker combinations. I am willing to believe that this can be explained by an unsuitability of the markers for the above pseudo-ROC scheme, but this needs to be discussed.

We have addressed this phenomenon further in the discussion. To summarize, in examining the raw data, for some markers the increase in the observed FDR for all methods is due to response--like signal in the "non-responder" category of samples (*i.e.,* these are likely false positives at the assay level).

- page 12, "and the subset IL2 and/or IFN-gamma that were used in the original study". It would be helpful to explain this ("and" or "or"?), and, more generally, which markers are biologically considered good indicators of responsiveness.

We have clarified this in the text. The and / or designation qualifies the sum of counts of the IFNg+/IL2+, IFNg-/IL2+, IFNg+/IL2- subsets of cells. It is the logical OR of cells expressing these markers. We have changed the nomenclature to IFNg OR IL2. While there are no good general indicators of responsiveness, the above subset was selected based on empirical evidence over many HIV vaccine studies. Generally, different types of vaccines are believed to induce different responses, and "good" markers of response are evaluated based on such empirical evidence.

- section 4.3: Instead of "subjects", "cells" and "markers", the section now talks about "observations" and "events". What is what?

We have clarified the nomenclature in this section to indicate that observations refer to "subjects" and events refer to "cells".

- Figure 1: The legend should list the methods in the same order as the list of AUC values.

We have updated the legend as requested.

- Figure 2a: Should we be impressed by the fact that the columns cluster according to the type of stimulation? Or is this expected because each stimulation had its own prior for p_u and p_s? The question of whether and how information is shared across samples from different stimulation is not discussed, but relevant here.

The reviewer is correct in noting that each stimulation is fit separately and thus has its own prior, resulting in the observed clustering by stimulation. The manner in which we fit the model to this data, we did not share information across stimulations. However, we argue here that our approach is the desired manner in which to proceed, since we are interested in an exploring this data set and identifying stimulation-specific effects even when the number of subjects is sparse. For comparison purposes, we have included the results of our model when all stimulations subjects are pooled together (Figure 2C). Interestingly, our model still leads to a good clustering of the subjects by stimulation group. We have also included the empirical proportions of cells expressing different genes in Figure 2B as a comparison. The pattern of these mirrors the pattern of posterior probabilities of response in the stimulation-specific model fit in Figure 2A. Fisher's test in

Figure 2D does not lead to a good resolution of stimulation specific differences amongst subjects, with the exception of the CMV pp65 nlv5 stimulation.

- page 3: "single-cell cytometry measure[s] [...] proteins, genes, cytokines" -- While technically not wrong, this sounds a bit "unprofessional" from a biology perspective. After all, a cytokine is a protein, and any protein is a gene's product.

We have corrected the terminology to refer to proteins and RNA gene expression , and later clarify the importance of cytokines specifically.

- Finally, not a question, rather a suggestion that we authors may or may not want to pursue: Eventually, we are not interested in knowing which subject is a responder but only which _fraction_ of subjects is a responder in order to quantify the efficacy of a vaccine. Could one get even better power by trying to infer only the fraction rather than the specific responder status of each subject, by introducing a further layer in the model?

The fraction of responders is modeled explicitly through the parameter, the mixing proportion between the non-responder and responder components of the model.

Associate Editor Comments to the Author:

The authors present a Bayesian mixture model for differential expression in relatively large-scale single-cell assays. The application is one that is becoming important currently, so the paper is timely, and the proposed model appears well-suited to the application.

There are two main criticisms of the paper: firstly that some parts are less clear than they could be. Both reviewers give several examples of confusing explanations.

Secondly there are several places where more precise details should be given. The reviewers give examples in the Results section where more results and explanations should be given. Also in the Introduction/Methods sections there should be more discussion of typical numbers found in these experiments (of cells, subjects etc.).

Specific questions from AE:

- Is it correct that in section 4.2 the binomial model was fit separately for each gene? But for all samples with the same stimulation simultaneously? It is not quite clear. In particular, as one reviewer asks, if models are run separately for each stimulation, then the different results shown in Figure 2 are to be expected?

The reviewer is correct in their observation, and while the result is expected, it is also desired since we wish to identify stimulation-specific differences in the data. See responses to reviewer who also asked about this issue for more details. We have updated the figures to highlight the reasons for this approach and addressed it in the discussion.

- Is the reason for better performance (under model mis-specification) than Fisher's exact test due to the information sharing across subjects? How few people would you have to have for the binomial model to be worse?

We have evaluated this with further simulation studies looking at smaller values of I (number of subjects). These results are presented in the Supplementary Figures. ROC curves show that our model performs better than competing methods down to at least 20 subjects (the smallest number we evaluated), however, the nominal FDR tends to be an

overestimate of the observed FDR as the number of subjects decreases.

1 Michael a Newton et al., "Detecting Differential Gene Expression with a Semiparametric Hierarchical Mixture Method.," *Biostatistics (Oxford, Engla*nd) 5, no. 2 (2004): 155–176, doi:10.1093/biostatistics/5.2.155.