Dear Tal,

Many thanks for your positive comments on our manuscript. We have revised the paper to address the reviewers' concerns. In particular, we have added explicit comparisons to a state-of-the-art clustering-based approach (CITRUS) using both simulated and real data, and show that our approach provides improved performance. We also propose a method for normalizing across batches of mass cytometry samples that have been barcoded separately or which lack multiplexing entirely, which extends the applicability of our approach. Finally, we now describe a novel strategy for facilitating biological interpretation of regions of the high-dimensional space with significant differences in abundance between samples. Our responses to each of the reviewers' specific comments are listed below, and our revisions are marked with red text in the manuscript.

We hope that the revised manuscript is of interest to you and to Nature Methods.

Regards,

John Marioni

**REVIEWER 1:**

**1. I find it surprising that the authors have not addressed batch effects anywhere in their manuscript (outside of acknowledging that it is a non-issue in their dataset). It’s nice that they have first demonstrated their method by picking a dataset where this is not an issue (because of barcoded multiplexing). However, Users of CyTOF are aware that batch effects (antibody staining, conjugation issues, signal normalization) are critical issues. If multiplexed samples were not available (which is almost always the case), naïve application of this method would simply nominate batch-driven subpopulations (in other words, LOTS of false positives). In fact clustering methods are often used for precisely this reason - to deliberately coarse grain the description to remove batch specific variation. Some commentary/analysis is critical if this method is to be generally useful for the CyTOF community, as multiplexing is often ruled out in many settings (e.g. clinical samples). Have the authors attempted analyzing any cases of non-multiplexed (i.e. independently collected) data?**

We have added a discussion on the impact of batch effects, and how they can be handled in the differential abundance framework in Section 2 of the Supplementary Materials. In particular, we describe several normalization approaches (quantile and range-based) for correcting intensities across batches of samples that have been barcoded separately, under the assumption that the only difference in the intensity distributions between batches is caused by the technical batch effect. We demonstrate this approach on a publicly available data set involving several batches of barcoded samples.

In the case where no barcoding is used at all (i.e., each individual sample has a “batch effect”), we note that the shifts in intensity will result in greater variability in abundances across replicate samples. Small shifts are automatically handled in our statistical model by a corresponding increase in the negative binomial dispersion. Larger shifts can be accommodated by increasing the hypersphere radius in proportion to the average size of the shift (estimated as the distance between replicate samples). Using simulations, we show that type I error control is maintained for modest shifts in intensity across samples, while increasing the hypersphere radius mitigates the loss of error control for larger shifts.

Of course, in a pathological setting with very large shifts, our method fails, e.g., if one subpopulation is (incorrectly) overlaid on top of another in high-dimensional space due to shifts in intensities across samples. However, importantly, clustering approaches would also be unable to handle such effects.

**2. The supplemental text in its current form is missing key details underlying the implementation of the statistical methods. Specifically details of the computation of the p-values, FDR calculations or implementation of GLMs are especially opaque. For example, the p-value calculation is described as follows... I feel this is all but utterly incomprehensible to even a person with a moderately advanced statistics background. The authors are admittedly piggybacking on a lot of existing work (e.g. edgeR, NB GLMs), but merely citing these papers does not suffice as it’s likely that researchers using mass cytometry are likely to have never heard of edgeR. I believe it’s absolutely imperative that the authors provide the basic steps of their analysis as a stepwise modus operandi (equations included), and cite only the most peripheral details. In the absence of this, it is very difficult to quickly follow exactly how the authors are performing their significance calculations (unless one digs through their code). This is equally important for those who wish to improve such methods in the future.**

We apologise for not more clearly describing our method. To address this, we have added substantially more detail to the description of the statistical methods in Sections 3 and 5 of the Supplementary Materials, where we elaborate on the theory underlying the methods in *edgeR* that we exploit in our manuscript as well as providing more information about the spatial FDR.

**3. A big problem in expression cytometry (flow or mass) is the varying dynamic ranges between different channels. Therefore I have concerns about the assumption of a common standard deviation (r0) across different channels, as I don’t think the authors have performed standardization (z-scoring) on the logicle-transformed marker intensities. I also wonder if the “1-log” range of expression variation is strictly true across all channels, or are some channels more or less varying (e.g. 2-log vs 0.5-log). In the absence of correction for dynamic range variation (if it exists), some markers will dominate others. Therefore, some commentary is necessary on this issue. I note that the assumption of a common standard deviation can be simply checked by plotting a histogram of each marker.**

This is an interesting point. We decided not to perform any standardization as differences in the range of expression variation may be biologically interesting. For example, in a population of T cells, CD25 may be expressed across a wide range of intensities reflecting the range of activation states. In contrast, CD3 (a classical T cell marker) is likely to be expressed across a very tight range. Likewise, even in a population of whole peripheral blood mononuclear cells where CD3 might be bimodal, expression of CD45 will be fairly uniformly high. Scaling these distributions to have the same standard deviation would reduce the interesting differences in CD25 expression and amplify the uninteresting differences in CD3 expression. Indeed, the fact that more highly variable markers will dominate the placement of hyperspheres is desirable, as it ensures that biological differences between subpopulations can be captured. We have added a discussion of this issue to Section 1.4 of the Supplementary Materials.

**4. The author’s point that the value of r0 will only affect power, but not error is convincing. However, I would still have liked to see how much power is impacted at values of r0 that are different from the default value. For example, versions of Fig. 5a for two different values of r0 would provide some insight into the impact on power. In other words, which DA subpopulations are detected/missed at different granularities.**

We agree that understanding how our method performs when the radius of the hyperspheres is altered is of interest. In particular, the relationship between the radius and the number of cells assigned to each hypersphere is an important factor, especially in high-dimensional space. For example, a 10% increase in the radius leads to a 17-fold increase in hypersphere volume for 30 dimensions. Thus, assessing the effect of moderate changes in the hypersphere volume, which corresponds to small changes in the radius, is more important than assessing the effect of moderate changes in the radius itself.

To this end, we repeated the real data analyses by altering the radius to double or halve the counts in each hypersphere (Section 1.2 of the Supplementary Materials). We observed that most of the results of our original analysis were generally recovered, indicating that our approach is robust to changes in the size of the counts per hypersphere. A moderate number of differences were also present – namely, 30% fewer hyperspheres were detected as being DA upon halving the counts in each hypersphere and 30% more hyperspheres were detected upon doubling the counts. We attributed this to loss or gain of power due to changes in the total number of cells within each hypersphere. Consequently, we do not feel that changes in granularity have a substantial effect – if it were an important factor, a substantial number of new hyperspheres should be detected with decreases in the radii (i.e., as spatial resolution improves), but this is not the case.

We have also added new simulations in Section 1.2 of the Supplementary Materials to demonstrate the effect of increasing the hypersphere radius in the presence of both DA and non-DA populations.

**5. I would strongly recommend that in addition to the code, the authors also provide a markdown file detailing a step-by-step use of their software on one of the datasets as a supplemental file. Providing such a template would greatly induce early adoption, and also provide a tutorial so that users can reproduce the key results in the manuscript and examine the impact of different parameters.**

We have added a Rmarkdown file to the GitHub repository referenced in the manuscript (https://github.com/MarioniLab/DAMethods2016/tree/master/real/workflow). This contains all steps of the workflow, starting from processing of the FCS files and ending at the t-SNE plot.

**(Minor) 1. I’m not sure if it’s accurate to say that the notion of DA is entirely novel. Levine et al., Cell, 2015, which the authors cite, do perform an abundance analysis in AML data (albeit within a clustering framework). Hence I think it’s only accurate to say that the notion of DA has always been tied to the notion of clustering in past works, which the authors have challenged.**

We agree that investigating DA is not novel and that the novelty of our approach is tied to not pre-clustering the data and to properly controlling for multiple testing in the high-dimensional space. To clarify this, we have added relevant citations to the Introduction.

**(Minor) 2. For the simulations in Fig 11, I was surprised at such low power of the naïve method. Did the authors try k-means clustering? And why not values of k lower than 20, as there are only three true “clusters” in the data (the two DA clusters and the non-DA population).**

Application of k-means clustering yielded results similar to those obtained with hierarchical clustering, so we did not include them in the plot for brevity. As for the choice of k, it is worthwhile noting the difference between empirically defined clusters and the true subpopulations. The two DA subpopulations are smaller than the non-DA subpopulation, such that clustering with k=3 would group the two adjacent DA subpopulations in the same cluster while breaking up the non-DA subpopulation into two chunks along its major axes of variation. We used large values of k to obtain small, precise clusters in the hope that at least one cluster would successfully represent each DA subpopulation (while including as few cells as possible from the non-DA subpopulation).

Nonetheless, we have repeated clustering with k=5, 50 and 500 to show the results across a wider range of values. We have also substantially simplified the simulation design to more clearly demonstrate the cause of the differences in performance between clustering and the hypersphere-based method. These new results are depicted in Supplementary Figure 17.

**(Minor) 3. The authors description of “composition effects” is not clear. I have read Supplementary section 2 multiple times and am not able to follow what problem is being addressed here. Is it the absence of certain biological subpopulations in some samples, which can cause large changes in counts? In their T/B cell example, it’s not clear to me why changes in B-cell subpopulations would create issues on the T cell side. Some rewording is necessary.**

We have changed the description of composition effects to provide more clarity in Section 4 of the Supplementary Materials.

**REVIEWER 2:**

**1. All the hyperspheres have a constant radius that depends solely on the number of markers used. However, since different markers exhibit different expression variation, low-expressed (or functional) markers may deserve smaller radiuses compared to highly expressed markers. Hence, there may be a need to scale the radius relevant for each marker to its variability in expression. Such strategy yields ellipsoids instead of hyperspheres, while the radius relevant for each marker is proportional to its variability across the samples. Such idea is discussed in Supp. 1.2, where the authors claim that the optimal choice of the radius depends on the estimated variability of the marker in each single subpopulations, and since such estimation is difficult they decided not to use it. Maybe a better strategy is to scale to the markers' expression across all the cells instead of in each subpopulation.**

As described above in response to Point 3 of reviewer 1, we decided not to perform any standardization as differences in the range of expression variation may be biologically interesting. High variability in the observed expression of a marker may represent interesting heterogeneity in the cell population, which may be compressed and lost if the intensity distribution was scaled to be the same as, e.g., a lowly-variable, constitutively expressed marker.

**2. The authors used hyperspheres with at least 5 cells to calculate differential abundance across conditions. Is it possible to justify this threshold?**

As mentioned in the text, it is difficult to detect significant differences at low counts, simply because there is not enough evidence to reject the null hypothesis. We chose a threshold of 5 to ensure that there are enough counts in each hypersphere so that rejection is possible if the hypersphere contains differences in abundance. Of course, lower thresholds can be used but this risks increasing the severity of the multiple testing correction. Our experience with negative binomial models on other types of biological count data (e.g., RNA-seq, ChIP-seq, Hi-C) suggests that a threshold from 5-10 usually yields a good compromise between retention of genuine differences and removal of the vast majority of unchanged features. We have added a discussion of this rationale to Section 3 of the Supplementary Materials.

**3. The authors suggest a strategy to find differential abundance with continuous predictors (time). However, in many cases the predictors are two or more discrete conditions. Is it possible to suggest a method to find differential abundance in such cases?**

This is certainly possible and a targeted use of the method. Implementation requires only a change in the design matrix that is passed to *edgeR*. We use generalized linear models (GLMs) in our statistical framework, which easily accommodate both simple and complex experimental designs with multiple explanatory factors. We now demonstrate this flexibility using a real data set containing unstimulated and IL-10 stimulated conditions (see Section 8 of the Supplementary Materials).

**4. The authors use partitions of similar volume in which the spatial FDR is controlled. They don't mention, however, the volume of each partition and how such volume was chosen. Additional effects of the partition volume on the differential abundance analysis should be examined. Furthermore, the effect of the partitions' positions on the spatial FDR is not examined.**

We thank the reviewer for highlighting our lack of clarity in this section of the manuscript. The partitions are purely hypothetical, and are only mentioned to help explain the concept of the spatial FDR. Importantly, the partitions are not actually used in the process of controlling the spatial FDR, and thus their volume, shape and orientation are irrelevant. We apologise for the confusion and have reworded Section 2.4 of the manuscript accordingly. Specifically, we provide a more intuitive explanation of the spatial FDR as the proportion of volume occupied by DA hyperspheres that corresponds to false positive hyperspheres. A more precise definition involving partitions has been moved to Section 5 of the Supplementary Materials, where we stress the hypothetical nature of the partitions.

**5. The authors argue that since there is redundancy between adjacent partitions not all the partitions are checked. The random sampling of partitions and the subsequent control of the FDR across the non-sampled partitions are not clear.**

We have removed this section, as it is unnecessarily confusing in light of the partitions being hypothetical.

**6. In Figure 4, the authors introduced differentially abundant populations into the MEF dataset. For simplicity, they calculated the spatial FDR in 2D space following application of PCA on the hyperspheres positions. Since they sought to estimate the performance of their method, the prompted test would be to make these calculation in the high dimensional space and not in the projected 2D space.**

We agree with the reviewer on this point. We have modified Figure 4 after repeating the calculations in high-dimensional space. We found that the spatial FDR is still controlled with our method, as expected.

**7. For validation, the authors used the MEF dataset, containing barcoded samples. Since the samples were barcoded, they didn't apply any normalization on this dataset. As far as I know, barcoded samples should be normalized to control the intrinsic drop in the signal during the run.**

The intensities in the debarcoded FCS files that we obtained from Cytobank were already normalized to account for changes in detector sensitivity (Zunder et al., 2015; see Methods, “Mass Cytometry Measurement and Data Processing”). Thus, further normalization is unnecessary.

**8. To further demonstrate the performance of the method it may be useful to apply on multiple mass-cytometry datasets.**

We agree and have now applied our approach to an additional public dataset that explored the effect of cytokine treatments on bone marrow mononuclear cells, and successfully detected a number of differentially abundant subpopulations. This is included in the manuscript as Supplementary Figure 16, with a brief discussion of the results in Section 8 of the Supplementary Materials.

**9. The authors mentioned that the existing methods don't perform DA analysis. However, the mentioned Citrus algorithm (Brugger et al) offers such analysis either by regression (glmnet, as described in the paper) or by SAM (significance analysis of microarrays).**

We have added a direct comparison to CITRUS for detecting differentially abundant subpopulations in our simulated and real data (Supplementary Figures 17-18). We demonstrate that our method provides a substantial improvement in performance over CITRUS for detection of DA subpopulations in both data sets. We note that only the SAM-based method is relevant here, as we are interested in testing for differential abundance between conditions rather than identifying features associated with each condition.

**10. Simulated-data validation: the authors generated a dataset by sampling cells from the same distribution or different distributions. (1) Detection of the DA clusters may be applied by other DA detection methods such as SAM. (2) There is a need to use much more clusters in the analysis to normalize the cell count in each cluster to the one in each hypersphere. Specifically, if we have 200,000 cells per sample and the maximal number of clusters used is 100, it means that each cluster includes approximately 2,000 cells, much more than the artificial DA populations. (3) The assignment of DA populations to the clusters is not clear. Since the clusters are much bigger than the DA populations, it makes sense their distance from these populations will be affected by non-DA cells, thereby affecting the assignment. A better strategy would be to calculate the proportion of the cells (from the DA populations) that were included in the detected DA clusters.**

For point 1, we have now compared the performance of our method to that of an existing method (CITRUS) that uses SAM, as mentioned in the previous comment.

For point 2, we have repeated the simulations using a much greater range of cluster numbers (5, 50, 500), the largest of which leaves ~40 cells in each cluster for a simulation involving 20000 cells. We have also simplified the simulation design to clarify the cause of the differences in performance between the clustering and hypersphere-based methods. These new results are depicted in Supplementary Figure 17. We note that using a large number of clusters ameliorates some of the performance disadvantages of clustering – however, the choice of an appropriate number of clusters is not obvious in complex data sets, and issues arise with FDR control and interpretation of the results. A short discussion about this has been added to Section 9 of the Supplementary Materials.

For point 3, we note that a common strategy for summarizing the cluster characteristics is to use the median intensity of each feature. Our assessment strategy reflects this practice. Indeed, even if a detected cluster included all the cells from a DA subpopulation, the cluster would not be considered to represent the DA subpopulation during interpretation, if it included many cells from other subpopulations. This is because the median intensities of the cluster would correspond to a point in the high-dimensional marker space that was far away from the location of the DA subpopulation.

**11. Real-data validation: the authors down-sampled 1,000 cells per sample and applied the DA analysis, as before. (1) Down-sampling is too small, and therefore may introduce sampling-noise to the results. (2) Similarly to (2) in the previous section: clusters' size should be scaled to the hyperspheres radius. (3) Similar to (3) in the previous section.**

For point 1, we chose 1000 cells to mimic the behaviour of CITRUS. We have since modified the manuscript in Section 2.6 to directly compare our approach to CITRUS on real data, rather than comparing to an approach of clustering followed by *edgeR* (which, as far as we know, has not actually been used in the literature). This resolves point 2, since CITRUS searches throughout all possible cluster sizes during significance testing. For point 3, we apply the same rationale as described in our response to point 3 in comment 10, above.

**According to the supplementary text section 1.1, "the expectation of the squared Euclidean distance between any two cells in this subpopulation is 2M〖r\_0〗^2". This is further validated in supplementary Figure 1, by plotting the median 1-NN distance across all cells. A more appropriate plot should describe the average distance instead of the median (as an approximation to the expectation).**

We have regenerated the plots using the mean instead of the median.

**REVIEWER 3**

**1. The authors emphasize that the hypersphere method surpasses clustering and give several justifications. While their logic is compelling, it is not clear to me that simply over clustering the data (as is commonly done in existing methods) will not alleviate this issue.**

We thank the reviewer for raising this point, and have now increased the number of clusters used in the simulations (using a range of 5, 50 and 500; also see our response to Reviewer 2 point 10). While using many clusters does indeed mitigate some of the disadvantages of clustering, it still lacks the detection power of hyperspheres for subtle DA subpopulations. We have added a discussion of this to Section 9 of the Supplementary Materials.

**2. Further, note that shifts in marker expression will often appear across samples (I believe the authors refer to this as bias). It is true that this effect disappears when samples are properly barcoded, as they are in Zunder et al, but in many (most?) studies barcoding all samples to be compared is either impossible or infeasible (eg, if samples are stained fresh and cannot be stored to enable cross-sample barcoding). It seems plausible that the hyper spheres will be \*more\* sensitive to such shifts than pre-clustering approaches, since slight shifts will merely smear a cluster from pooled data, whereas this approach may be overly sensitive in this case and detect a shift as a DA region.**

This is a good point. To address this, as described in our response to Reviewer 1 point 1, we have incorporated an additional adaptable radius into our method, which accounts for non-multiplexed datasets. We also propose a simple method where batches of samples that have been barcoded separately can have their intensities normalized for valid comparisons between samples from different batches. Both of these methods are described in Section 2 of the Supplementary Materials.

**3. Regarding Bruggner et al: As the authors point out in the discussion, detection of differential marker expression is effectively equivalent to DA detection. For this reason, Bruggner et al is a relevant point of comparison that wasn’t addressed in the paper directly.**

**i) I would like therefore to see a simple analysis: Starting with cytof data from 2 or more classes (say, 2 time points from Zunder et al, or any other 2 conditions in real life CyTOF data known to have some DA populations), compare results from the proffered approach with 1) Bruggner et al 2) SPADE and X-shift - since these latter approaches do not directly detect DA populations, the resulting clusters' counts can be fed into a statistical test such as Mann Whitney.**

As described in our response to Reviewer 2 comment 9, we have incorporated additional comparisons to the CITRUS method (Bruggner et al.) for both real and simulated data, and show that our method provides a substantial improvement in performance. However, we feel that it is beyond the scope of this manuscript to modify SPADE and X-shift such that they can perform differential abundance analyses, as these tools were not designed for this purpose.

**ii) Although the writing emphasizes use of hyperspheres as the main departure from current approaches, 2 important contributors to the performance are potentially the NB modeling and/or the spatial FDR. To fully explore this contributions of these components, I would like to see the results when SPADE and X-Shift clusters are modeled similarly with NB and spatial FDR, in comparison to hypersphere counts modeled with NB and spatial FDR. It is my understanding that all 3 components are important to the end result with this method, but the contribution of each is not shown clearly in the current manuscript.**

We agree that all three components are crucial to our model and contribute to its improved performance. However, we are proposing our full method as a novel differential abundance analysis approach. It would be interesting if our components were able to improve the performance of other methods, but showing that they do so seems like a substantial research project in itself.

**i) (minor, page 2) What are the DA methods used in flow and what prevents them from scaling for CyTOF? The authors mention this but don’t elaborate. Are you referring to manual gating?**

We are referring to manual gating schemes and have modified the Introduction to be more clear.

**ii) (minor, bc more or less addressed elsewhere page 3) - As I mention above,over clustering can circumvent the issues described here. You might explain why hypershperes are conceptually better.**

We have addressed these points in Section 9 of the Supplementary Materials.

**iii) Section 2.2. If I understand correctly, the radius will depend on the transformation method employed, for instance, it will be different if the asinh transform is selected. Please address this briefly.**

The justification for our default radius assumes that the intensities are at or close to a log10-scale, which is achieved by log10, arcsinh, and biexponential (based on logicle) transformations. Visualization on this scale seems to be fairly standard in the cytometry literature, where decades are frequently used to describe intensities. We have added a comment on this to Section 1.4 of the Supplementary Materials.

**iv) Section 2.3 (Major) - This section needs more clarification. Can you provide intuition as to why the NB model is useful in this context? I was unclear as to the dispersion - in your motivating example this is used to smooth across probes, etc, but in this data there is only one measurement per marker. Is dispersion referring to dispersion across replicates in this case?Although Figure 2b is beautiful and compelling, it is not clearly explained WHY your approach is, as claimed, more sensitive and specific.**

The NB model is widely used for all types of count data containing greater-than-Poisson variance. The dispersion provides a measure of the variability between counts corresponding to replicate observations. In this case, a large dispersion for a hypersphere indicates that its counts are highly variable across replicate samples. Smoothing is then performed to stabilize the dispersion estimate across hyperspheres. We have added a thorough description of the mathematical framework of NB GLMs as it relates to the hypersphere counts in Section 3 of the Supplementary Materials.

We stress that we are not estimating the dispersion for each marker – unlike applications for gene expression, where the dispersion is estimated from the variability of expression across samples for each gene. Rather, we are estimating the dispersion based on the variability of cell abundance across samples for each hypersphere. We have added a comment about this to Section 2.3 of the manuscript.

In addition, we have modified the simulations for Figure 2b and provided a more detailed explanation of how our approach compares to the Mann-Whitney test.

**v) Section 3.3 (Major) It is obvious that the authors went to great lengths in this section to provide intuition, but it was unfortunately not effective in my opinion. Explain WHY the weighted scheme provides a better FDR correction. The text is confusing, particularly because of the back and forth between partitions, which are both claimed to be the reason for superior performance, as well as claimed to be not used in practice but merely constructs (" We stress that the partitions described above are purely conceptual.”). I end up scratching my head. I gather that employing weighting based on the local density is the reason behind the improved performance. Please explain this more clearly in simple terms: What is actually happening in your FDR, and why does it lead to better performance.**

We appreciate the reviewer pointing out this confusion and have revised Section 2.4 of the manuscript to improve the clarity of the explanation. Briefly, the spatial FDR can be intuitively interpreted as the proportion of volume occupied by DA hyperspheres that corresponds to false positive hyperspheres. We have moved the discussion of the partitions into Section 5 of the Supplementary Materials, where we elaborate on some of the theoretical subtleties. In particular, we note that the partitions are required to define the concept of the spatial FDR (and for calculating the observed spatial FDR in simulations, where the true or false positives are known). However, the partitions are not actually required to control the spatial FDR in real data analyses.

**vi) Section 4.4. In section 4.4 the partitions are described again and they do seem to be in use, but the binning is performed in PCA space. It is not mentioned how much of the overall variance is captured in the first 2 PCs. Can you comment on this? How important is this matter to performance of this approach?**

We have modified the simulations to use partitions in the original high-dimensional space, thus circumventing the need for PCA.