Testing for differential abundance in mass cytometry data

Aaron T L Lun1, Arianne C Richard1,2 & John C Marioni1,3,4

When comparing biological conditions using mass cytometry data, a key challenge is to identify cellular populations that change in abundance. Here, we present a computational strategy to detect ‘differentially abundant’ populations by assigning cells to hyperspheres, testing for significant differences between conditions and controlling the spatial false discovery rate. Our method (http://bioconductor.org/packages/cydar) outperforms other approaches in simulations and finds novel patterns of differential abundance in real data.

Mass cytometry allows researchers to simultaneously characterize the expression of many (>30) protein markers in each of millions of cells1. Antibodies specific to markers of interest are conjugated to heavy-metal isotopes and used to stain a population of cells. Single-cell droplets are formed and vaporized to ionize the metals, and the quantity of each isotope bound to each cell is measured by time-of-flight mass spectrometry. The resolution of mass spectrometry avoids problems with spectral overlap that are frequently encountered in conventional flow cytometry with fluorescent markers. This means that more markers can be quantified for each cell; this improves the resolution of distinct subpopulations and enables deep phenotyping of cellular profiles in fields such as immunology, hematopoietic development and cancer2–6.

Mass cytometry can assay more markers than conventional flow cytometry, which increases the dimensionality of the data. This complicates analysis, as manual gating and visual examination of biaxial plots (commonly used in flow cytometry) are no longer feasible when multiple marker combinations have to be considered. Bespoke computational tools such as SPADE7 and X-shift8 address this by clustering cells into biologically relevant subpopulations based on the ‘intensity’ of each marker (i.e., the signal of the corresponding isotope in the mass spectrum) and quantifying the abundance of each subpopulation in the total cell pool. However, these approaches cannot directly address the important question of what differs between biological conditions in multi-condition experiments.

To this end, one strategy is to identify subpopulations that change in abundance between conditions9,10. For example, certain immune compartments are enriched or depleted upon drug treatment, and the composition of cell types changes during development. Detection of these differentially abundant (DA) subpopulations is useful, as it can provide insights into the cause or effect of the biological differences between conditions. Existing methods for DA analysis cluster cells from all samples into empirical subpopulations before checking each cluster for characteristics (e.g., marker intensities or cell abundance) that differ between conditions11,12. While intuitive, this approach is sensitive to the parametrization of the initial clustering step. Noisy data or poorly separated cells also introduce uncertainty into the cluster definitions13. This is particularly relevant for markers that are expressed across a range of intensities without clear changes in cellular density at subpopulation boundaries, such as CD38 and HLA-DR to mark activated T cells or CD24 and CD38 to define plasmablasts among B cells14. Ambiguity in clustering can affect the performance of the subsequent DA analysis if, for example, DA and non-DA subpopulations are erroneously clustered together.

Here, we present a computational strategy to perform DA analyses of mass cytometry data that does not rely on an initial clustering step (Fig. 1, Supplementary Software and http://bioconductor.org/packages/cydar). Our method allocates cells to hyperspheres, tests for differential abundance of cells between conditions for each hypersphere, and controls the false discovery rate (FDR) across the high-dimensional space. It can be used to robustly detect differentially abundant subpopulations or shifts in marker expression between conditions.

First, we assign cells from all samples to hyperspheres in the multidimensional marker space. Consider a mass cytometry data set with S samples and M markers. Each cell in each sample defines a point in the M-dimensional space, with coordinates defined by the cell’s marker intensities. We consider M-dimensional hyperspheres where each hypersphere is centered on an existing cell and has radius \( r = 0.5\sqrt{M} \) to offset the increasing sparsity of the data as the number of dimensions increases. All cells lying within a hypersphere are assigned to that hypersphere. We count the number of cells from each sample assigned to each hypersphere, yielding S counts per hypersphere. (Each cell can be counted multiple times if it is assigned to overlapping hyperspheres.) For each hypersphere, we also compute the median intensity of each marker across all cells in that hypersphere. This provides a median-based position for the hypersphere, representing a central point in M-dimensional space around which most of the cells in the hypersphere are located (see Supplementary Note 1, Supplementary Figs. 1–4 and Supplementary Table 1 for more details). We also assume

1Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. 2Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK. 3EMBL European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, UK. 4Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge, UK. Correspondence should be addressed to J.C.M. (marioni@ebi.ac.uk).

RECEIVED 12 AUGUST 2016; ACCEPTED 7 APRIL 2017; PUBLISHED ONLINE 15 MAY 2017; DOI:10.1038/NMETH.4295
BRIEF COMMUNICATIONS

that marker intensities are comparable across samples—strategies for handling sample-specific intensity shifts are described in Supplementary Note 2 and Supplementary Figures 5 and 6.

Next, we use the count data for each hypersphere to test for significant differences in cell abundance between conditions. The null hypothesis is that there is no change in the average counts between conditions within each hypersphere. Testing is performed with negative binomial generalized linear models (NB GLMs), which explicitly account for the discrete nature of counts and model overdispersion due to biological variability between replicate samples; and which can accommodate complex experimental designs involving multiple factors and covariates. We use the NB GLM implementation in edgeR\textsuperscript{15}, which was originally designed for analyzing read count data from RNA sequencing experiments. edgeR uses empirical Bayes shrinkage to share information across hyperspheres. This improves estimation of the dispersion parameter in the presence of limited replicates, increasing the reliability and power of downstream inferences (Supplementary Note 3, Supplementary Figs. 7 and 8). Indeed, edgeR is more powerful than the commonly used Mann–Whitney test for detecting differences in hypersphere counts in simulated data while still controlling the type-I error rate (Supplementary Fig. 9).

Finally, we use the hypersphere P values to control the FDR across the multidimensional space. Roughly speaking, the spatial FDR can be interpreted as the proportion of the total volume of DA hyperspheres (the union rather than sum of the individual volumes, due to overlaps between hyperspheres) that is occupied by false-positive hyperspheres. This is not equivalent to the FDR across the individual hyperspheres, as hypersphere density differs across the space. For example, the FDR across hyperspheres in Figure 1d is 25%; while the spatial FDR across volume is 50%. To control the spatial FDR, each hypersphere is weighted by the reciprocal of its density (calculated in terms of the neighboring hyperspheres). A weighted version of the Benjamin–Hochberg (BH) method\textsuperscript{16} is then applied to the P values and weights for all hyperspheres. If one were to split the high-dimensional space into nonoverlapping partitions of equal volume, the total weight of hyperspheres within each nonempty partition would be similar; i.e., each partition of the space makes a similar contribution to the BH correction, regardless of how many hyperspheres it contains. Thus, weighting allows the FDR to be controlled across volume, rather than across hyperspheres (see Supplementary Note 4 and Supplementary Fig. 10 for a more precise description of the spatial FDR.) We demonstrate that our weighting scheme successfully controls the spatial FDR in simulated data, whereas a naïve approach without weighting does not (Supplementary Fig. 11).

In the cydar package, we provide several options for interpreting and exploring DA hyperspheres after the statistical analysis. We can identify significant hyperspheres that are not redundant to—i.e., do not lie within a certain distance of—hyperspheres with smaller P values (Supplementary Note 5). The resulting subset of hyperspheres is small enough for detailed inspection of the marker intensities with a Shiny graphical interface (Supplementary Fig. 12) to characterize each hypersphere. A complementary approach is to perform dimensionality reduction on the positions of the putative DA hyperspheres, yielding a low-dimensional representation of the differential subspaces for plotting. The plot can be annotated based on marker intensities, incorporating biological expertise on the relationships between specific markers and cell types. These approaches allow biologically relevant subpopulations to be identified from the DA hyperspheres.

We demonstrate our approach by identifying differentially abundant populations among mouse embryonic fibroblasts (MEFs) (bearing Oct4–GFP, Nanog–GFP or Nanog–Neo reporters) that were reprogrammed to induced pluripotent stem cells\textsuperscript{17}. We assessed changes in abundance across samples collected at various timepoints during reprogramming; and we detected 7416, 5947 and 21532 DA hyperspheres in the Oct4–GFP, Nanog–GFP and Nanog–Neo time courses, respectively, at a spatial FDR of 5%. We applied t-SNE\textsuperscript{18} to the positions of detected hyperspheres to visualize them in a spatial context (Fig. 2 and Supplementary Figs. 13–18). In the Oct4–GFP analysis, we recovered previously identified DA subpopulations, including the three reprogramming endpoints, as well as distinct DA subpopulations that were not clearly characterized in the original analysis, such as a subpopulation of SC4-like cells with phosphorylated STAT3, AMPK and PLK1 that exhibited a nonlinear change in abundance over time (Supplementary Fig. 19 and Supplementary Note 6).

We also applied our method to another data set in order to examine the effect of interleukin 10 (IL–10) treatment on bone marrow

Figure 1 | A pipeline to determine differential cell population abundance from mass cytometry data. (a) Cells from samples 1 or 2 in this schematic are distributed across the multidimensional marker space (two markers are shown here for simplicity). Hyperspheres (yellow; h$_1$–h$_7$) centered on selected cells are constructed, and the number of cells from each sample inside each hypersphere is counted. (b) Counts for each hypersphere are tested for significant differences between samples. (c) Multiple testing correction is performed by controlling the spatial FDR. Positions of significant hyperspheres at a given spatial FDR threshold are visualized by dimensionality reduction (e.g., PCA). (d) The spatial FDR is roughly equivalent to the proportion of the volume occupied by false-positive hyperspheres. Each hypersphere has a median-based position (small circles) and occupies a volume of the high-dimensional space (shown as the dotted ring for two hyperspheres). The total occupied volume is the union of individual hypersphere volumes. Two groups of hyperspheres are shown—one containing true positives with genuine differences in abundance, the other containing false positives—that occupy a similar total volume V with different densities.
mononuclear cells (BMMCs) across five healthy donors. Importantly, this data set contained matched stimulated and unstimulated samples from each donor. This experimental design is easily accommodated by the GLM machinery in edgeR, highlighting the flexibility of our framework. We observed changes in abundance associated with phosphorylated STAT3 expression, consistent with the expected biology of IL-10, as well as several interesting DA subpopulations that were not identified by the original study (Supplementary Note 7 and Supplementary Figs. 20–21). More generally, shifts in marker intensity for signaling molecules or activation markers will cause changes in abundance that can be detected by the DA analysis (Supplementary Note 8 and Supplementary Fig. 22).

Finally, we compared our approach with CITRUS, a method that uses an initial clustering step for comparative analysis of mass cytometry data. We simulated a simple scenario involving two adjacent subpopulations with opposite changes in abundance between conditions (Supplementary Note 9 and Supplementary Fig. 23). These subpopulations were consistently detected as being differentially abundant by our hyperspheres-based method but not by CITRUS. We also tested the performance of CITRUS for detecting differentially abundant subpopulations across time in the MEF-reprogramming data set. CITRUS did not detect a number of subpopulations that were found by our method (Supplementary Fig. 24), nor did it detect any new subpopulations. This suggests that the use of hyperspheres, in combination with edgeR and the spatial FDR, can improve detection of subtle changes in abundance within complex subpopulations that are difficult to cluster.

As mass cytometry becomes more accessible, large-scale experiments containing many conditions and replicates are likely to become increasingly routine. Indeed, a growing number of studies are using mass cytometry in fields such as immunology, hematopoietic development and cancer. We anticipate that our differential abundance analysis pipeline will be useful to researchers planning to perform comparative studies with such data sets.

**METHODS**

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

**ACKNOWLEDGMENTS**

This work was supported by Cancer Research UK (core funding to J.C.M., award no. A17197), the University of Cambridge and Hutchison Whampoa Limited. J.C.M. was also supported by core funding from EMBL.

**AUTHOR CONTRIBUTIONS**

A.T.L.L. developed the analysis pipeline, tested it with simulations and applied it to the real data. A.C.R. interpreted the results to identify the DA subpopulations. J.C.M. provided direction and advice on method development and biological interpretation. All authors wrote and approved the final manuscript.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Ornatsky, O.I. et al. Anal. Chem. 80, 2539–2547 (2008).
2. Leijpold, M.D., Newell, E.W. & Maeker, H.T. Methods Mol. Biol. 1343, 81–95 (2015).
3. Leelatian, N., Diggins, K.E. & Irish, J.M. Cytometry A 87, 876–891 (2011).
4. Samusik, N., Good, Z., Spitzer, M.H., Davis, K.L. & Nolan, G.P. Nat. Methods 13, 493–496 (2016).
5. Gaudilliére, B. et al. Sci. Transl. Med. 6, 25ra131 (2014).
6. McCarthy, D.J., Chen, Y. & Smyth, G.K. Nucleic Acids Res. 40, 4288–4297 (2012).
7. Benjamini, Y. & Hochberg, Y. Scand. J. Stat. 24, 407–418 (1997).
8. Zunder, E.R., Lujan, E., Goltsev, Y., Wernig, M. & Nolan, G.P. Cell Stem Cell 16, 323–337 (2015).
9. Van der Maaten, L. & Hinton, G. J. Mach. Learn. Res. 9, 2579–2605 (2008).
In this section, we describe the processing of data from the MEF-reprogramming study\textsuperscript{17}. For processing of data from the BMMC study\textsuperscript{6}, see Supplementary Note 7 for details.

We obtained debarcoded flow cytometry standard (FCS) files for each time course from Cytobank (accession number 43324). We applied the logicle transformation\textsuperscript{19} to the marker intensities in each sample. The transformation parameters were estimated with the estimateLogicle function from the flowCore package\textsuperscript{20}, using pooled cells from all samples in each time course. (This avoids spurious differences from sample-specific transformation.) We gated out cell events with low intensities for the two DNA markers (Iridium-191 and Iridium-193), where the threshold was defined as three median absolute deviations below the median intensity for the pooled cells. We saved the transformed and gated intensities into new FCS files for processing with our pipeline. Only the intensities for relevant markers (i.e., no DNA, barcodes) were used for further analysis. Note that normalization of marker intensities between samples is not required for this data set, because the samples in each time course were barcoded and pooled for antibody staining and mass cytometry.

To compute time as a covariate. This yielded a log transformation to the counts for each hypersphere using a design matrix with three degrees of freedom. This provided 9, 11 and 10 residual degrees of freedom for dispersion estimation in the GLM deviance and stabilized the estimates by empirical Bayes shrinkage toward a second mean-dependent trend. Finally, we fitted an NB GLM to the NB dispersion estimates. We fitted an NB GLM to the counts for each hypersphere, using the transformed dispersion for each hypersphere and the log-transformed total number of cells as the offset for each sample. We estimated the NL dispersion from the GLM deviance and stabilized the estimates by empirical Bayes shrinkage toward a second mean-dependent trend. Finally, we used the NL F-test with a specified contrast to compute a \( P \) value for each hypersphere. Details of the statistical framework are provided in Supplementary Note 3.

For the time course analyses, we used a design matrix constructed from a B-spline basis matrix with a time covariate and three degrees of freedom. This provided 9, 11 and 10 residual degrees of freedom for dispersion estimation in the Oct4–GFP, Nanog–GFP and Nanog–Neo data sets, respectively. Contrasts were constructed to test whether all spline coefficients were equal to zero. This represents a null hypothesis that time has no effect on abundance. The use of splines accommodates both linear and nonlinear trends in abundance with respect to time.

For each hypersphere detected at a spatial FDR of 5\%, we defined the median-based position as a set of intensity values across all markers. These values were used to perform t-SNE via the Rtsne package (https://cran.r-project.org/web/packages/Rtsne), using a perplexity value of 10. To color the plot based on differential abundance, a GLM was fitted to the counts for each hypersphere using a design matrix with time as a covariate. This yielded a \( \log_{2} \)-fold change in abundance per day for each hypersphere, corresponding to a blue-to-red gradient from negative (blue) to positive (red) values. (We assumed a linear change in abundance over time for simplicity. This assumption did not affect the significance statistics, which were computed with a spline to account for nonlinear trends.) To color the plot based on marker intensity, the 1\textsuperscript{st} and 99\textsuperscript{th} percentiles of the intensities for all cells were computed for each marker. A linear gradient between these two percentiles was constructed using the viridis color scheme (https://cran.r-project.org/web/packages/viridis). Each hypersphere was then assigned a color based on the location of its median marker intensity on the gradient.

Using CITRUS to analyze the mouse embryonic fibroblast data. To run CITRUS (v0.08), the citrus.full command was used with the featureType argument set to “abundances” and the modelType argument set to “continuous” to identify changes in abundance over time. Downsampling was performed to 1,000 cells per sample; and the minimum cluster size was set to 5\%, based on the default settings. Detected clusters were defined as those reported at an FDR of 5\%, as reported by the SAM method. For each detected cluster, the median-based center was determined; and the hypersphere with the closest position to the cluster center in \( M \)-dimensional space was identified. Each cluster center was mapped onto the t-SNE plot of DA hyperspheres using the coordinates of the hypersphere closest to the cluster center. Note that a cluster center was not mapped if the distance to the closest hypersphere was greater than 0.5\( \sqrt{M} \). If an unmapped DA cluster was present, it was treated as being undetected by the hypersphere-based approach.

Implementation of cell-counting software. All simulation and analysis code was written in R. Methods in the cydar package were written in a combination of R and C++. Cell counting, nearest-neighbor detection and density estimation were performed using an approach similar to that in X-shift\textsuperscript{8}. Briefly, k-means clustering was performed on all cells, setting \( k = \sqrt{N} \), where \( N \) is the total number of cells. Let \( |j - t| \) denote the Euclidean distance between cell \( j \) and the center of cluster \( t \) in the \( M \)-dimensional marker space. Similarly, let \( |h - t| \) denote the distance between the centers of \( t \) and hypersphere \( h \). Both of these distances only need to be computed once per cell—in the latter case, this is because each hypersphere is centered on a cell. By applying the triangle inequality, a cell \( j \) in cluster \( t \) was only considered for assignment to a hypersphere \( h \) if \( r + \sqrt{|j - t|} \geq |h - t| \). In order to avoid unnecessary work, the distance between \( j \) and \( h \) was not computed for cells not satisfying this requirement. Similarly, \( j \) was only considered as a possible neighbor of a cell \( j' \) if \( d_{n} + \sqrt{|j - t|} \geq |j' - t| \), where \( d_{n} \) is the distance to the current \( n \)th nearest neighbor. (The value of \( d_{n} \) is continually updated once a closer \( n \)th nearest neighbor is identified.) This speeds up the pipeline while yielding the same results as a naïve approach that computes distances between every pair of cells. On a desktop machine, the analysis takes 10–20 min to run for each of the MEF-reprogramming time courses.

Code availability. Simulation and analysis code are accessible at http://github.com/MarioniLab/DAMethods2016. Methods in the DA analysis pipeline are publicly available in the mass cytometry for differential abundance analyses in R (cydar) package from
the open-source Bioconductor project at http://bioconductor.org/packages/cydar or in the Supplementary Software.

Data availability statement. All data sets used here are publicly available from Cytobank using accession number 43324 for the MEF study and 44185 for the BMMC study. Source data is available online for Figure 2.

19. Parks, D.R., Roederer, M. & Moore, W.A. Cytometry A 69, 541–551 (2006).
20. Hahne, F. et al. BMC Bioinformatics 10, 106 (2009).