Testing for differential abundance in mass cytometry data

Aaron T. L. Lun[1], Arianne C. Richard[1,2] and John C. Marioni[1,3,4,*]

[1] Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, United Kingdom
[2] Cambridge Institute for Medical Research, University of Cambridge, Cambridge, United Kingdom
[3] EMBL European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, United Kingdom
[4] Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge, United Kingdom

* Corresponding author: marioni@ebi.ac.uk

Abstract: When comparing biological conditions using mass cytometry data, one key challenge is to identify cellular populations that change in abundance. Here, we present a novel computational strategy for detecting these "differentially abundant" populations, by assigning cells to hyperspheres, testing for significant differences between conditions and controlling the spatial false discovery rate. The method's performance is established using simulations and real data where it finds novel patterns of differential abundance.

Mass cytometry allows researchers to simultaneously characterise the expression of many (> 30) protein markers in each of millions of cells[1]. 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, improving resolution of distinct subpopulations and enabling deep phenotyping of cellular profiles in fields such as immunology, haematopoietic development and cancer[2, 3, 4, 5, 6]. The ability of mass cytometry to assay more markers leads to a concomitant increase in the dimensionality of the data. This complicates the data analysis as manual gating and visual examination of biaxial plots (as commonly used in flow cytometry) are no longer feasible when multiple marker combinations have to be considered. To address this, bespoke computational tools such as SPADE[7] and X-shift[8] have been developed, focusing on 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 fail to directly address an important question of multiparameter multi-group experiments – namely, what differs between groups?

To this end, an alternative analytical strategy is to identify subpopulations that change in abundance between biological conditions[9, 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 conditions[11, 12]. While intuitive, this approach is sensitive to the parametrization of the initial clustering step. Uncertainty will be introduced into the cluster definitions when the data are noisy or the cells are not clearly separated[13]. 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 cells[14]. Ambiguity in clustering can affect the performance of the subsequent DA analysis, e.g., if DA and non-DA subpopulations are erroneously clustered together.

Here, we present a novel computational strategy to perform DA analyses of mass cytometry data (Figure 1) that does not rely on an initial clustering step. Firstly, we assign cells from all samples to hyperspheres in the multi-dimensional 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 its intensities. We consider M-dimensional hyperspheres where each hypersphere is centred 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 then assigned to that hypersphere. (Each cell can be counted multiple times if it is assigned to overlapping hyperspheres.) We count the number of cells from each sample assigned to each hypersphere, yielding S counts per hypersphere. For each marker, we also compute its median intensity for all cells in each 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 Figures 1-4 and Supplementary Table 1 for more details. We also assume that marker intensities are comparable across samples – some strategies for handling sample-specific intensity shifts are described in Supplementary Note 2 and Supplementary Figures 5-6.

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; model overdispersion due to biological variability between replicate samples; and can accommodate complex experimental designs involving multiple factors and covariates. We use the NB GLM implementation in the edgeR package[15], which was originally designed for analyzing read count data from RNA sequencing experiments. However, the same mathematical framework can be applied here to cell counts. In particular, 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. (See Supplementary Note 3 and Supplementary Figures 7-8 for more details.) 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 Figure 9).

Finally, we use the hypersphere *p*-values to control the false discovery rate (FDR) across the multi-dimensional space, i.e., the spatial FDR. To illustrate, consider the total volume occupied by the set of DA hyperspheres. (This is a union rather than a sum of the hypersphere volumes, due to overlaps between hyperspheres.) Roughly speaking, the spatial FDR can be interpreted as the proportion of this volume that is occupied by false positive hyperspheres. This is not equivalent to the FDR across the individual hyperspheres, due to the differences in the density of hyperspheres 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 neighbouring hyperspheres). A weighted version of the Benjamini-Hochberg (BH) method[16] is then applied to the *p*-values and weights for all hyperspheres. If one were to split the high-dimensional space into non-overlapping partitions of equal volume, the total weight of hyperspheres within each non-empty 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 Figure 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 Figure 11).

Several options are available for examining 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 graphical interface (Supplementary Figure 12) to characterise 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 is annotated based on examination of the marker intensities, incorporating biological expertise on the relationships between specific markers and cell types. This allows identification of biologically relevant subpopulations from the DA hyperspheres.

We demonstrate our approach using data from a study of mouse embryonic fibroblast (MEF) reprogramming[17]. In this study, three transgenic MEF reporter systems (*Oct4*-GFP, *Nanog*-GFP or *Nanog*-Neo) were reprogrammed to induced pluripotent stem cells. Samples were collected across various points of the reprogramming time course for each MEF reprogramming system. We applied our method to each time course to detect changes in abundance over time, defining putative DA hyperspheres as those detected at a spatial FDR of 5%. In this manner, we detected 7416, 5947 and 21532 DA hyperspheres in the *Oct4*-GFP, *Nanog*-GFP and *Nanog*-Neo time courses, respectively. We applied *t*-SNE[18] to the positions of detected hyperspheres to visualize them in a spatial context (Figure 2, Supplementary Figures 13-18). In the *Oct4*-GFP analysis, we recovered previously identified DA subpopulations, including the three reprogramming end points; as well as distinct DA subpopulations that were not clearly characterised in the original analysis, such as a subpopulation of SC4-like cells with phosphorylated STAT3, AMPK and PLK1 that exhibited a non-linear change in abundance over time (Supplementary Figure 19) – see Supplementary Note 6 for details.

We also applied our method on another data set examining the effect of interleukin 10 (IL-10) treatment on bone marrow mononuclear cells (BMMCs) across five healthy donors[6]. 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 (see Supplementary Note 7, Supplementary Figures 20-21 for details). More generally, shifts in marker intensity for signalling molecules or activation markers will cause changes in abundance that can be detected by the DA analysis (Supplementary Note 8, Supplementary Figure 22).

Finally, we compared our approach to CITRUS[12], an existing 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, Supplementary Figure 23). These subpopulations were consistently detected as being differentially abundant by our hypersphere-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 Figure 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, haematopoietic 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.

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.

Acknowledgements

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.

Competing financial interests

The authors declare no competing financial interests.

References

1. Ornatsky, O.I. *et al*. *Anal. Chem.* **80**, 2539–2547 (2008).
2. Leipold, M.D., Newell, E.W. & Maecker, H.T. *Methods Mol. Biol.* **1343**, 81–95 (2015).
3. Leelatian, N., Diggins, K.E. & Irish, J.M. *Methods Mol. Biol.* **1346**, 99–113 (2015).
4. Hansmann, L. *et al*. *Cancer Immunol. Res.* **3**, 650–660 (2015).
5. Bendall, S.C. *et al. Science* **332**, 687–696 (2011).
6. Levine, J.H. *et al*. *Cell* **162**, 184–197 (2015).
7. Qiu, P. *et al. Nat. Biotechnol.* **29**, 886–891 (2011).
8. Samusik, N., Good, Z., Spitzer, M.H., Davis, K.L. & Nolan, G.P. *Nat. Methods* **13**, 493–496 (2016).
9. Gaudillière, B. *et al. Sci. Transl. Med.* **6**, 255ra131 (2014).
10. Gaudillière, B. *et al. Cytometry A* **87**, 817–829 (2015).
11. Anchang, B. *et al. Nat. Protoc.* **11**, 1264–1279 (2016).
12. Bruggner, R.V., Bodenmiller, B., Dill, D.L., Tibshirani, R.J. & Nolan, G.P. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E2770–2777 (2014).
13. Ronan, T., Qi, Z. & Naegle, K.M. *Sci. Signal.* **9**, re6 (2016).
14. Finak, G. *et al. Sci. Rep.* **6**, 20686 (2016).
15. McCarthy, D.J., Chen, Y. & Smyth, G.K. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
16. Benjamini, Y. & Hochberg, Y. *Scand. J. Stat.* **24**, 407–418 (1997).

17. Zunder, E.R., Lujan, E., Goltsev, Y., Wernig, M. & Nolan, G.P. *Cell Stem Cell* **16**, 323–337 (2015).
18. Van der Maaten, L. & Hinton, G. *J. Mach. Learn. Res.* **9**, 2579- 2605 (2008).

Figure legends

Figure 1: Schematic of the differential abundance pipeline. (a) Cells from samples 1 or 2 are distributed across the multi-dimensional marker space (two markers shown here for simplicity). Hyperspheres (yellow, h_1-h_4) centred 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. This yields a *p*-value representing the evidence against the null hypothesis of no differences. (c) Multiple testing correction of hypersphere *p*-values 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.

Figure 2: Differentially abundant subpopulations in the *Oct4*-GFP time course, detected at a spatial FDR of 5%. (a) A *t*-SNE plot of the median positions of DA hyperspheres. Each point represents a hypersphere and is coloured according to its average log-fold change in abundance over time. Grey points represent hyperspheres with significant but non-linear changes in abundance. Subpopulations were annotated based on results in Zunder et al.[17], with additional distinguishing features for each subpopulation noted in parentheses. OSKM: reprogramming factors (OCT4, SOX2, KLF4, c-MYC), NE: non-expressing, MET: mesenchymal-epithelial transition, SC4: partially reprogrammed cell line, ESC: embryonic stem cells, mixed 4F: mixed stoichiometry of the OSKM factors. (b) The same plots

coloured by the median intensity of selected markers in each hypersphere. The colour range for each marker was bounded at the 1st and 99th percentiles of the intensities across all cells.

Online Methods

Data preparation

In this section, we describe the processing of data from the MEF reprogramming study[17]. For processing of data from the BMMC study[6], see Supplementary Note 7 for details.

We obtained de-barcoded flow cytometry standard (FCS) files for each time course from Cytobank (accession number 43324). We applied the logicle transformation[19] to the marker intensities in each sample. The transformation parameters were estimated with the estimateLogicle function from the flowCore package[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 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.

Statistical methods for testing differences

To compute *p*-values, hypersphere counts were analyzed using the quasi-likelihood (QL) method in edgeR. First, we filtered out hyperspheres with an average count below 5. This improves efficiency by removing tests without enough information to reject the null hypothesis. For the remaining hyperspheres, we fitted a mean-dependent trend to the NB dispersion estimates. We fitted a NB GLM to the counts for each hypersphere, using the trended dispersion for each hypersphere and the log-transformed total number of cells as the offset for each sample. We estimated the QL dispersion from the GLM deviance and stabilized the estimates by empirical Bayes shrinkage towards a second mean-dependent trend. Finally, we used the QL 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 3 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 non-linear trends in abundance with respect to time.

Visualizing the differential hyperspheres

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 colour 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 yields a $\log_2$-fold change in abundance per day for each hypersphere, corresponding to a blue-to-red gradient for negative-to-positive values respectively. (We assume a linear change in abundance over time for simplicity. This does not affect the significance statistics, which are computed with a spline to account for non-linear trends.) To colour the plot based on marker intensity, the 1st and 99th percentiles of the intensities for all cells were computed for each marker. A linear gradient between these two percentiles was constructed using the viridis colour scheme (https://cran.r-project.org/web/packages/viridis). Each hypersphere was then assigned a colour based on the location of its median marker intensity on the gradient.

Using CITRUS to analyze the MEF 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 “sam”. The family argument was set to “continuous” to identify changes in abundance over time. Downsampling was performed to 1000 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 a FDR of 5%, as reported by the SAM method. For each detected cluster, the median-based centre was determined and

the hypersphere with the closest position to the cluster centre in *M*-dimensional space was identified. Each cluster centre was mapped onto the *t*-SNE plot of DA hyperspheres using the coordinates of its closest hypersphere. Note that a cluster centre 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 were written in R. Methods in the cydar package were written in a combination of R and C++. Cell counting, nearest-neighbour detection and density estimation were performed using an approach similar to that in X-shift[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 centre of cluster t in the M-dimensional marker space. Similarly, let $|h - t|$ denote the distance between the centres 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 centred on a cell. By applying the triangle inequality, a cell j in cluster t was only considered for assignment to a hypersphere h if $r + |j - t| \geq |h - t|$. For cells not satisfying this requirement, the distance between j and h was not computed to avoid unnecessary work. Similarly, j was only considered as a possible neighbour of a cell j' if $d_n + |j - t| \geq |j' - t|$ where d_n is the distance to the current n^{th} nearest neighbour (where the value of d_n is continually updated once a closer n^{th} nearest neighbour 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 minutes 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 cydar package (mass CYtometry for Differential Abundance analyses in R) from the open-source Bioconductor project at http://bioconductor.org/packages/cydar, or by downloading the Supplementary Software associated with this paper.

Methods-only References

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).

Data availability

All data sets used here are publicly available from Cytobank (https://community.cytobank.org), using the accession number 43324 for the MEF study and 44185 for the BMMC study.

Supplementary Materials

The Supplementary Materials is a single PDF file that consists of Sections 1-9 and contains Supplementary Figures 1-24 and Supplementary Table 1.

Figure 1

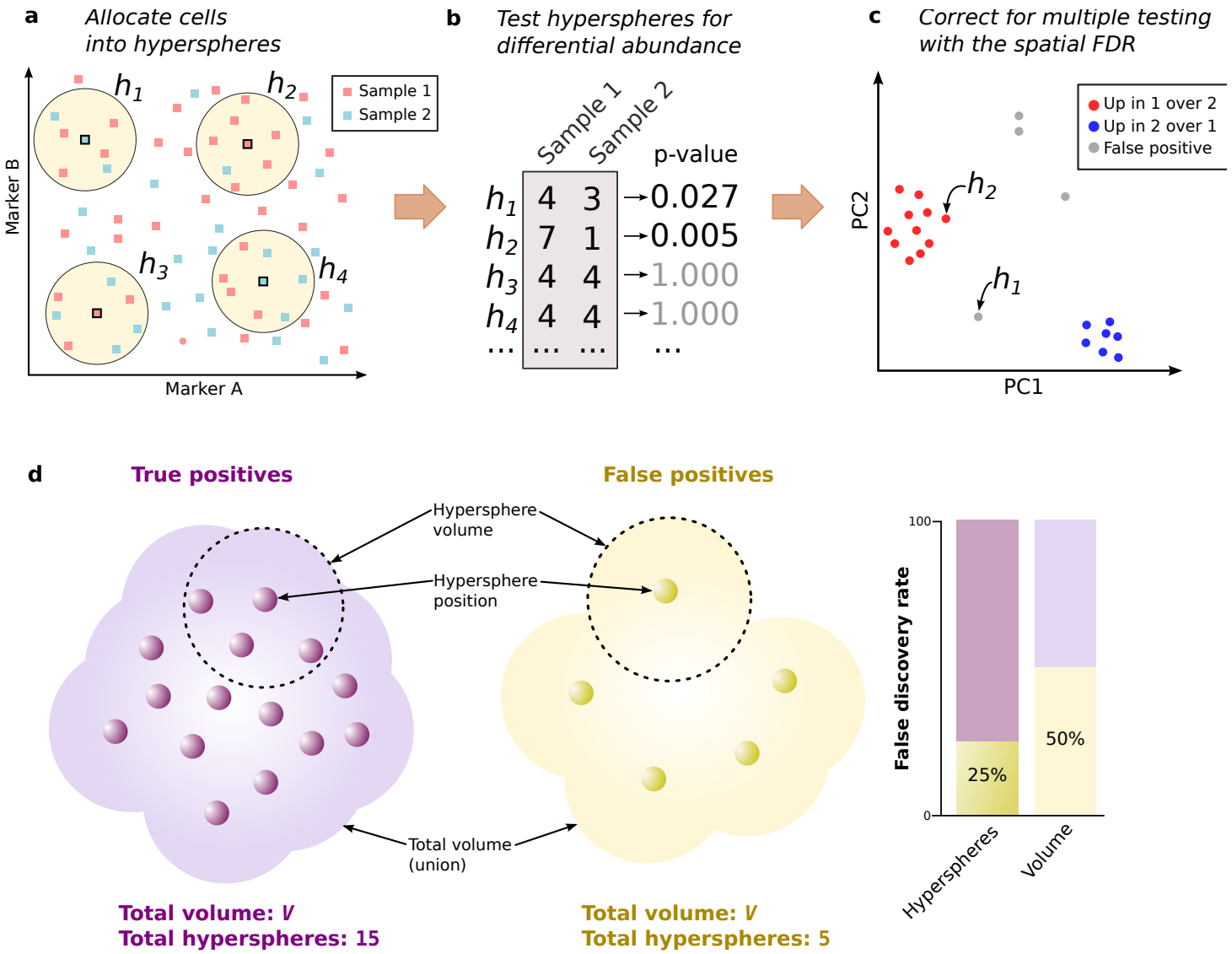
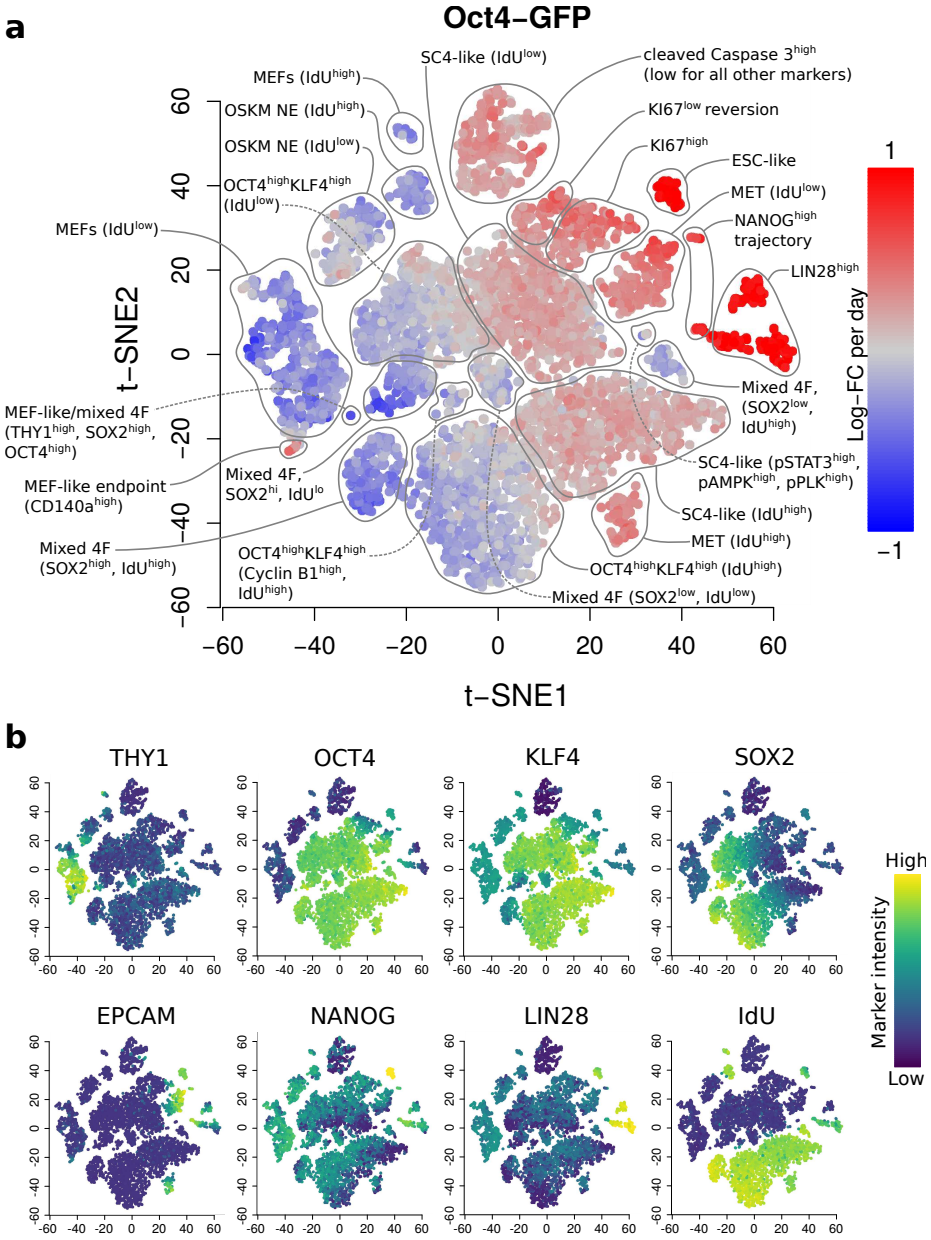

a
Allocate cells into hyperspheres
b
Test hyperspheres for differential abundance
c
Correct for multiple testing with the spatial FDR
Sample 1
Sample 2
Marker A
Marker B
p-value
Up in 1 over 2
Up in 2 over 1
False positive
PC1
PC2
d
True positives
False positives
Hypersphere volume
Hypersphere position
Total volume (union)
Total volume: V
Total hyperspheres: 15
Total volume: V
Total hyperspheres: 5
False discovery rate
100
0
25%
50%
Hyperspheres
Volume

Figure 2


a
Oct4−GFP
SC4-like (IdU^low)
cleaved Caspase 3^high
(low for all other markers)
MEFs (IdU^high)
OSKM NE (IdU^high)
KI67^low reversion
OSKM NE (IdU^low)
KI67^high
ESC-like
OCT4^high KLF4^high
(IdU^low)
MET (IdU^low)
NANOG^high
trajectory
MEFs (IdU^low)
LIN28^high
Mixed 4F,
(SOX2^low,
IdU^high)
MEF-like/mixed 4F
(THY1^high, SOX2^high,
OCT4^high)
SC4-like (pSTAT3^high,
pAMPK^high, pPLK^high)
Mixed 4F,
SOX2^hi, IdU^lo
SC4-like (IdU^high)
MEF-like endpoint
(CD140a^high)
MET (IdU^high)
Mixed 4F
(SOX2^high, IdU^high)
OCT4^high KLF4^high
(Cyclin B1^high,
IdU^high)
OCT4^high KLF4^high (IdU^high)
Mixed 4F (SOX2^low, IdU^low)
t−SNE2
t−SNE1
Log−FC per day
1
−1
b
THY1
OCT4
KLF4
SOX2
EPCAM
NANOG
LIN28
IdU
High
Marker intensity
Low