Note 1: Defining the hypersphere parameters

1.1 General comments

We assume that appropriate transformations (e.g., biexponential [1]) have been applied to the marker intensities for each cell. We also assume that normalization of marker intensities between samples, if necessary, has already been performed (see Note 2.2). In data sets where samples were barcoded and multiplexed for staining and processing [2], normalization of intensities is not required as the same technical biases should be present in all samples and thus should cancel out when comparing between samples. Assignment of cells to hyperspheres is then performed using the transformed (and normalized) intensities for each cell. Each hypersphere represents a portion of the M-dimensional space (i.e., a subspace) in which cells are counted.

It is important to stress that, while each hypersphere contains a subset of cells in the population, this should not be confused with a cell subpopulation. The latter refers to a biologically meaningful or functional subset of cells, while the former is simply an analytical construct used to quantify cell abundance.

1.2 Centring on existing cells

Each hypersphere is centred at a point defined by an existing cell. This is necessary because there are an infinite number of hyperspheres in the M-dimensional space. Centring on existing cells ensures that only non-empty hyperspheres are considered. In practice, we further reduce computational work by only constructing hyperspheres at every (randomly chosen) 10^{th} cell. This avoids redundant work in high-density subspaces where many of the resulting hyperspheres would be near-identical in terms of their positions and cell counts. Note that many adjacent hyperspheres will be overlapping, which has some implications for FDR control. This is demonstrated in Figure 1d and motivates the development of the spatial FDR (see Note 4.1).

1.3 Choosing the hypersphere radius

1.3.1 Overview

The radius r of each hypersphere dictates the trade-off between count size (the number of cells assigned to a hypersphere) and spatial resolution (the ability to distinguish between adjacent subspaces). Our choice of radius $r = 0.5\sqrt{M}$ increases with the number of markers M, to avoid problems with sparsity at high dimensions; and it acounts for the variability of the intensities for each marker, whereby cells with a 10-fold difference in marker expression are routinely assigned to the same hypersphere. (This is motivated by noting that technical or biological variability often results in an order of magnitude difference in observed expression within the same functional subpopulation [2, 3, 4].) This is described in more detail in the following text.

1.3.2 Formulating an expression for the radius

Consider a subpopulation where the intensities of each marker are distributed around some mean intensity with standard deviation r_0. The expectation of the squared Euclidean distance between any two cells in this subpopulation is $2r_0^2M$. Now, construct hyperspheres centred at each of the cells in the subpopulation. Setting the hypersphere radius as $r = r_0\sqrt{2M}$ ensures that most hyperspheres will contain a large proportion of cells from the same subpopulation. Indeed, a simple simulation with Normally-distributed intensities for 30 markers indicates that the majority of hyperspheres contain more than 50% of the subpopulation.

We further assume that the expression of a marker will differ by up to 10-fold across functionally similar cells, due to technical noise or biological variability within the same subpopulation. (Expression profiles for more variable markers like CD25 may consist of multiple internal subpopulations.) For typical analyses involving transformed intensities on the $\log_{10}$-scale, an interval of length 1 will span an order of magnitude in marker expression. When centred around the mean, this interval should contain most of the intensity distribution for a single subpopulation. According to Chebyshev's inequality, the interval defined by a distance of $r_0\sqrt{2}$ from the mean will contain at least 50% of the values from an arbitrary distribution. We equate these interval lengths to obtain a rough estimate for the standard deviation of the intensities, i.e., $2r_0\sqrt{2} = 1$. Applying this to the radius yields $r = 0.5\sqrt{M}$, meaning that each hypersphere will be able to include cells with 10-fold differences in expression for one or more markers. Thus, each hypersphere will

contain enough counts for further analysis, even when the intensities vary across an order of magnitude. We refer to the value of $r_0\sqrt{2}$ as the "tolerance" in the code, representing half the width of the intensity distribution for a single marker in a homogeneous population that is positive for that marker.

We verify the choice of formulation of r by examining the distance between each cell x and its neighbours in the first sample from each time course. The nearest neighbours of x are identified in the full M-dimensional space and represent other cells in the same subspace as x. As the number of markers increases, the distance to the neighbours increases at a rate that is roughly consistent with the square root function (Supplementary Figure 1). This justifies the use of the relation $r \propto \sqrt{M}$. With increasing M, the radius will also increase such that a hypersphere centred at x will still contain the neighbours of x, i.e., sensitivity of counting for cells in the same subspace is preserved. In contrast, cells in other subspaces will cease to be assigned to the hypersphere as they become separated from x on other markers, i.e., specificity of counting is improved.

For the *Oct4*-GFP and *Nanog*-Neo time courses, the radius is consistent with the average distance from each cell to its 10^{th} nearest neighbour. This means that around 10 cells from the subspace will be assigned to each hypersphere when $r_0\sqrt{2} = 0.5$. Combined with counts from the other samples in the time course, this provides enough information to reliably detect DA hyperspheres. In contrast, the radius in the *Nanog*-GFP time course is consistent with the distance of each cell to its 1^{st} nearest neighbour. This is because the first sample in the *Nanog*-GFP time course contains fewer than 3000 cells, while samples in the other time courses contain over 10000 cells. The sparsity of cells in this sample results in fewer neighbours at any given distance. For experiments with few cells across all samples, a larger r_0 may be required to obtain sufficient counts for further analysis. (Note that we keep $r_0\sqrt{2} = 0.5$ for *Nanog*-GFP, as the other samples in the time course contain 5-10-fold more cells. Thus, counts will be large enough to detect differences later on.)

1.3.3 Additional remarks on radius choice

The suitability of the default setting of $r_0\sqrt{2} = 0.5$ can be assessed for each data set by examining the distance from each cell to its nearest neighbours. Supplementary Figure 2 shows these distances for each cell in the *Oct4*-GFP reprogramming time course. Here, the default radius corresponds to the median of the distances from each cell to its 15^{th} nearest neighbour. This means that half of all hyperspheres will contain more than 15 cells, which is large enough for downstream hypothesis testing. In contrast, a smaller value of the radius (e.g., $r_0\sqrt{2} = 0.4$) would not even include the nearest neighbour for each cell on average. This would lead to a count of 1 for most hyperspheres, which is not sufficient for further statistical analysis. The neighbour distances can be used to gauge whether a larger radius should be used for a particular analysis, though it is not generally necessary to decrease the default radius when these distances are small. A reduction in the neighbour distances will naturally occur with an increase in the total number of cells per sample – decreasing the radius would undermine any improvement in power due to an increase in cell counts per hypersphere.

A consequence of working in high-dimensional space is that small changes to the radius will result in large changes to the counts. For example, in an experiment consisting of 30 markers, increasing the radius by 10% will increase the hypersphere volume (and the potential number of counted cells) by over 17-fold. In this respect, the analysis is quite sensitive to the choice of radius. However, a more relevant assessment of sensitivity is to consider changes in the counts per hypersphere. We repeated the DA analyses using radius values that approximately halved or doubled the median number of counted cells in each hypersphere relative to the default (Supplementary Figure 2). We observed that most of the DA hyperspheres from the default analysis were still detected in the repeated analyses (Supplementary Table 1). Over 60% of the DA hyperspheres were recovered with a smaller radius, while using a larger radius detected almost all of them (and an extra 35%). These differences are expected because larger hyperspheres have greater counts and more evidence to reject the null hypothesis, albeit at the cost of spatial resolution. Nonetheless, the general similarity in the results indicates that the analysis is robust to substantial changes in the size of the counts.

An obvious question is whether an "optimal" value of r_0 can be obtained for any given data set. This value depends on the variability of marker intensities in each subpopulation. For example, setting r_0 to the standard deviation of the marker intensities *for a single subpopulation* allows more cells associated with that subpopulation to be counted into the hypersphere, while reducing the chance of counting cells from adjacent subpopulations. However, the standard deviation of each subpopulation is not easy to estimate empirically, as we do not know the true number of subpopulations in the M-dimensional space. Indeed, a single optimal value may not exist, e.g., if different subpopulations or markers have different standard deviations.

A suboptimal choice of r_0 is not a critical concern for a DA analysis. A value of r_0 that is too small will reduce power to detect a DA subspace as the counts are too low. A value that is too large will also reduce power through the loss of spatial resolution – specifically, cells from non-differential subspaces will be included in the counts for a hypersphere in a differential subspace, reducing the log-fold change in abundance. In both cases, power is affected rather than error rate control. While loss of power is undesirable, the analysis will still be valid as any discoveries (that are made in spite of the diminished power) can be trusted. We note that "contamination" of a non-DA hypersphere with cells from differential subspaces is also possible for large r_0, which could lead to the detection of the former as a false positive. However, the effect of this contamination is mitigated by the fact that the position of each hypersphere is defined using the median intensities. Any contamination that substantively changes the counts for a hypersphere will also shift its position towards the differential subspace. Thus, a contaminated hypersphere that is erroneously classified as DA will be assigned to or near the differential subspace (based on its position), rather than a non-differential subspace (based on its centre). This reduces the detection rate of false positives in the non-differential subspace.

To demonstrate this effect, we set up a simulation where we examined the effect of increasing the radius on the position of each hypersphere (Supplementary Figure 3). When larger radii were used, hyperspheres centred on cells from a non-DA subpopulation exhibited larger log-fold changes in abundance (Supplementary Figure 4). This is consistent with increasing contamination from cells in the neighbouring DA subpopulation. However, the positions of the affected hyperspheres were also shifted towards the DA subpopulation. In most simulation scenarios, this shift was sufficiently large that the hyperspheres with large log-fold changes (> 1) were indistinguishable from hyperspheres centred on cells from the DA subpopulation itself. The only exception occurred when the size of the DA subpopulation was reduced, such that some of the non-DA hyperspheres had large log-fold changes but lay outside the subpopulation boundaries. Even in this case, a substantial shift in position towards the DA subpopulation was observed, mitigating the effect of misinterpreting these hyperspheres as part of a differential subspace. Using very large hypersphere radii also results in a decrease in the log-fold changes for hyperspheres centred on cells from the DA subpopulation. This is consistent with loss of power when cells from non-DA subspaces are included in the counts.

1.4 Using weighted medians to compute the hypersphere position

The median position is more appropriate than the hypersphere centre for characterising the location of the hypersphere, when the cells assigned to the hypersphere are not symmetrically distributed around the centre. When calculating the median-based position for a hypersphere h, each cell from sample s is assigned a weight of T_s^{-1} where T_s is the total number of cells in s. For each marker, the weighted median of intensities from all cells in h is computed. The set of weighted medians for all markers represents the coordinates of the position of h. This weighting scheme ensures that the calculation of the position is not dominated by large samples with many cells. Obviously, it converges to a simple median in data sets containing samples of similar size.

1.5 Choosing a transformation for the intensities

The interpretation of the default hypersphere radius assumes that the transformed intensities lie at or close to a $\log_{10}$ scale. We use the logicle (or biexponential) transformation [1] implemented in the flowCore package [5], which is linear around zero but approaches the $\log_{10}$ scale when applied to high values for the raw intensities. Other transformations can be used provided that this scale is approximately maintained. For example, the inverse hyperbolic sine function converges to the logarithmic function at high values.

In general, we do not scale the intensities to equalize the standard deviation of the intensity distribution across markers. This is because differences in the range of expression between markers may be biologically interesting and should be preserved. For example, in a population of T cells, CD25 may be expressed across a wide range of intensities reflecting a range of different activation states. In contrast, CD3 (a classical T cell marker) should be present at high intensities in all cells with low variance across the population. Scaling to equalize the standard deviations for all markers would compress the CD25 intensity distribution and compromise the resolution of potentially relevant subpopulations, while also amplifying irrelevant changes in CD3 expression. Without scaling, more highly variable markers will dominate the placement of hyperspheres. This is desirable as it ensures that biological differences between subpopulations can be captured.

One argument for rescaling intensities is to improve the resolution of subpopulations separated on low-variance markers, by increasing their dynamic range to be equal to that of high-variance markers. While this may provide some benefit, it is not necessary when small hyperspheres are used. Consider a marker m that is lowly expressed in one subpopulation and absent in all other subpopulations. As long as the level of expression in the positive subpopulation exceeds the variability due to biological/technical noise (as quantified by the hypersphere radius), the cells in that subpopulation will be allocated into a separate hypersphere, *regardless of whether the dynamic range of marker m is smaller than that of other markers*. In other words, the resolution is already sufficient for separating subpopulations prior to the statistical analysis. Furthermore, intensity scaling of low-variance markers can be detrimental as it "dilutes" the cells across a wider dynamic range. This reduces the number of cells per hypersphere and DA detection power.

Note 2: Strategies for dealing with technical intensity shifts

2.1 Overview

In barcoded experiments, technical effects causing shifts in marker intensity between samples are avoided [2]. This is due to use of multiplexed staining and cytometry, which ensures that any fluctuations in the experimental procedures (e.g., staining efficiency, cellular concentration, detector sensitivity) affect all samples equally. Although barcoding is becoming more common [6, 7], its application is not always possible. For example, if the number of samples is greater than the number of available barcodes, samples will need to be split into batches for separate barcoding and analysis. This may introduce intensity shifts between batches, where similar cells in different batches have different intensities due to technical factors. For hypersphere counting, these shifts are problematic as cells from the same subpopulation may no longer be counted into the same hypersphere across samples. This has some detrimental consequences for the downstream statistical analysis (see below). Here, we present some strategies for handling intensity shifts in several scenarios.

2.2 Intensity shifts between separately barcoded batches

2.2.1 Outline of the normalization method

Consider an experimental design containing multiple batches of separately barcoded samples, such that intensity shifts are present between but not within batches. Intensities can be normalized between batches by assuming that the pooled intensity distribution across samples in each batch should be similar between batches (i.e., if intensity shifting were not present). This is usually reasonable in experimental designs where each batch contains samples from the same or similar set of conditions. The aim of the normalization procedure is to transform the pooled intensity distributions for all batches towards some reference distribution, thus removing any intensity shifts or changes in the shape of the intensity distribution between batches.

2.2.2 Creating a weighted distribution of intensities

The first step is to create a weighted distribution of intensities for each batch.

1. Let the number of samples for condition c in batch b be denoted by S_{cb}. The average sample size $\bar{S}_c$ for condition c is computed by taking the mean of S_{cb} across all batches.

2. Denote the number of cells in sample s of condition c in batch b as T_{scb}. Each cell in s has a weight of

$$w_{scb} = \frac{\bar{S}_c}{S_{cb}T_{scb}} \ .$$

3. Cells are pooled across all samples within each batch. A weighted distribution of intensities is generated for each marker m, based on the intensities and weights for all cells in the batch. This is used to generate a batch- and marker-specific quantile function $q_{mb}(.)$ for the intensity distribution.

The weights ensure that the contribution of each condition is the same across batches, to accommodate experiments where S_{cb} differs between batches. They also ensure that the contribution of each sample is the

same within each batch, if T_{scb} varies between samples. Otherwise, the pooled intensity distributions may not be similar across batches if some samples or conditions contribute more to one batch than to others. Note that construction of $q_{mb}(.)$ is only performed using samples from conditions that are present in *all* batches. If a batch is missing a condition, no amount of weighting can overcome this complete lack of information.

2.2.3 Performing range normalization across batches

A simple approach is to scale the marker intensities so that the range of the weighted distribution is the same for each batch. We compute the minimum and maximum for each batch b as $q_{mb}(0.01)$ and $q_{mb}(0.99)$, i.e., the 1^{st} and 99^{th} percentiles of the reference distribution for marker m, respectively. (We use percentiles rather than taking the actual extremes to protect against outliers.) The reference minimum and maximum are defined as the average of $q_{mb}(0.01)$ and $q_{mb}(0.99)$, respectively, across all batches. For each batch b, a linear scaling function is defined that converts $q_{mb}(0.01)$ and $q_{mb}(0.99)$ to the reference minimum and maximum, respectively. This function is then applied to the intensities of m for all samples in b, yielding normalized intensities where differences in the location and spread of the distribution between batches are eliminated. However, this "range-based" normalization assumes that the technical differences between batches can be fully described as a linear transformation of intensities. More complex non-linear effects are not supported.

2.2.4 Performing warping normalization across batches

An alternative strategy is to exploit existing methods for automatic normalization of flow cytometry data. We adapt the approach used by the flowStats package (https://www.bioconductor.org/packages/flowStats).

1. For each batch b, randomly sample T_b cells with replacement from the weighted distribution of intensities, where T_b is the total number of cells across all samples in b. This is necessary as the flowStats implementation does not consider weights, which are instead reflected in the sampling probabilities.

2. Apply the normalization() function in flowStats to compute a monotonic warping function. Briefly, this identifies high-density landmarks (i.e., peaks) in the intensity distribution for each batch. Landmarks from all batches are pooled together and clustered based on their locations, i.e., intensity at the peak summit. A smooth monotonic "warping" function [8] is defined for each batch to align the location of each of its landmarks to that of other landmarks (from different batches) in the same cluster.

3. Apply the warping function for batch b to the intensities for marker m in all samples from b. This yields normalized intensity values that are comparable across batches.

This approach is more flexible than range-based normalization as it accommodates non-linear shifts across the intensity distribution. However, it relies on accurate identification of landmarks that can be matched across batches. This may not be possible in noisy data sets with strong inter-batch heterogeneity.

2.2.5 General comments about inter-batch normalization

Both of the normalization methods above require an experimental design where at least one condition has samples in every batch. Ideally, at least one sample from each condition would be present in each batch. This ensures that information from all samples will be used to construct the weighted distribution for each batch. In practice, a condition will not be used if it does not contain samples in all batches. This is not a problem as the scaling/warping functions, once constructed, can be applied to any set of intensities, allowing normalization of intensities for samples in all (used and unused) conditions. However, the obvious caveat is that intensities outside the range of values used to construct the functions may not be accurately normalized.

Both methods also assume that the systematic differences between batches are primarily driven by technical factors. Any changes in the intensities are considered to be uninteresting and normalized out. However, this may not be appropriate if batches are confounded with differences in the biological conditions. In the simplest case, an increase in marker intensity might be due to differences in staining efficiency between batches, or a genuine change in protein expression between conditions. Confounding factors can be avoided with good experimental design, though some subtleties need to be considered – see the example below.

2.2.6 Application of the normalization methods to real data

As a demonstration, we applied our normalization procedure to the BMMC data set generated by Levine *et al.* [9]. This data set consists of stimulated and unstimulated samples from each of multiple donors, where barcoding was performed within but not between individuals, i.e., each individual represents a batch. Before normalization, we observed large differences in the intensity distribution of some markers between corresponding samples from different individuals (Supplementary Figure 5). Part or all of these differences are likely caused by the technical effects of separate staining and cytometry between individuals. Shifts in the maximum intensity were successfully eliminated upon range-based normalization. Application of warping normalization also removed changes in the shape of the intensity distribution between individuals.

It is worth noting, however, that this experimental design is not ideal for our normalization methods. This is because the technical differences between batches are confounded by genuine biological differences between individuals. Subsequently, normalization to remove the former may also remove or distort the biological effects. A more suitable design would contain samples from multiple individuals in each batch, such that one could assume that there were no biological differences between the average of individuals within each batch. To mitigate any distortions of the underlying biology in this data set, we only apply range-based normalization as this is more restrained in how it corrects for the inter-batch differences. Nonetheless, the DA analysis following our normalization still yields sensible results, as described in Note 7.

2.3 Intensity shifts between non-barcoded samples

A more difficult situation is that of a data set containing samples with no barcoding at all. For an intensity shift between two samples in different conditions, it is impossible to determine whether the shift represents a technical or biological effect. Subsequently, normalizing the location of the intensity distributions risks discarding interesting biology. Instead, we use a different approach where the radius of the hyperspheres is increased by the size of the shift. Cells from the same subpopulation in different samples are more likely to be counted into the same expanded hypersphere, even after shifting of intensities between samples.

To calculate the magnitude of the intensity shift due to technical effects, we compute the mean intensity of each marker in each sample. This yields a vector of means $\mathbf{u}_m$ for each marker m. We fit a linear model to $\mathbf{u}_m$ for each m, using an appropriate design matrix that describes the experimental set-up. The residual variance of the fitted model provides an estimate of the average squared shift between replicate samples. As this component of the shift occurs between replicates, it is more likely to be technical in origin. We average the variance estimates across all markers to obtain v^2, the extra variance introduced by the shifting process. Stochastic intensity shifts between samples increase the squared Euclidean distances between cells (from the same subpopulation but in different samples) by $2v^2$ for each marker. Thus, the new hypersphere radius should be $\sqrt{2M(r_0^2 + v^2)}$ in order to be able to routinely assign such cells to the same hypersphere.

We tested this approach with simulations based on the reprogramming time courses. Simulated data were generated as described in Note 3.4, using only the first sampling scheme for simplicity. Then, for each sample and for each marker, we sampled a value from a Normal distribution with a mean of zero and a standard deviation ranging from 0 to 0.3. This value was added to the marker's intensities for all cells in that sample to mimic an intensity shift due to technical effects. We counted cells into hyperspheres with and without expansion of the radius, performed the DA analysis and computed the observed type I error rate. For small shifts, the type I error rate was controlled with both the default and expanded radii (Supplementary Figure 6). This is due to an increase in the negative binomial dispersion, which accounts for greater variability in counts between replicates when cells from the same subpopulation are shifted in or out of the hypersphere between samples. For larger shifts, increases in the dispersion were not sufficient to control the type I error rate with the default radius. This is because the movement of entire subpopulations in or out of hyperspheres yields cell counts that are not accurately modelled with the NB distribution. In contrast, the use of an expanded radius mitigates the loss of error control. This reduces the effect of the shifts on the cell counts for each hypersphere, which allows for smaller dispersions and ameliorates the inaccuracy of the NB model.

There are important caveats with this expansion approach, some of which are mentioned below. Firstly, we assume that the variance of the shifting process is the same for different markers. This may not be true if some antibodies or detector channels are more susceptible than others to technical effects. Secondly, we assume that shifts in location between replicates are wholly technical. This is unlikely to be true, given that biological variability will be present between, e.g., replicate mice or patients. Exact calculation of technical

shifts requires technical replicates, where the same sample is processed twice in separate rounds of staining and mass cytometry. Thirdly, as observed above, accurate control of the type I error rate is not actually achieved upon expansion. The improvement is only a mitigation of the more severe loss of control when the default radius is used. Also, we have not considered non-linear effects where the size of the shift is a function of the intensity, which may further reduce the accuracy of the method. Finally, expansion of the hypersphere radius leads to loss of spatial resolution, which can reduce power to detect changes in abundance. While these shortcomings are significant, the lack of barcoding in the data itself leaves few analytical options for rigorous quantitation across samples. In such cases, radius expansion seems to be the best strategy.

Note 3: A brief description of edgeR's statistical framework

3.1 Quasi-likelihood negative binomial generalized linear models

The following section largely paraphrases Lun *et al.* [10], with some minor adjustments. Consider a hypersphere h with count y_{hs} in each sample s. In edgeR, the count in each sample is modelled with a quasi-negative binomial distribution with mean μ_{hs}. (We will discuss the dispersion parameters of this distribution later.) The negative binomial distribution is well-suited to this application, as it is supported over all non-negative integers and can accurately model both small and large cell counts. Estimation of the dispersions also allows modelling of extra-Poisson variability in the counts from replicate samples. By combining this with generalized linear models, we can handle overdispersed count data from a variety of experimental designs.

Using a log-link generalized linear model with G coefficients [11], the mean is represented as

$$\log \mu_{hs} = \sum_{g=1}^{G} x_{sg}\beta_{hg} + o_{hs} \ ,$$

where β_{hg} is the hypersphere-specific value of coefficient g, x_{sg} is the sample-specific predictor for g, and o_{hs} is the hypersphere- and sample-specific offset. For simple one-way layouts, each coefficient represents the (log-transformed) average proportion of cells within a group, and each predictor specifies the group to which each sample belongs. More complex designs can also be used where coefficients represent blocking factors or real-valued covariates. The offset for each sample is defined as the log-transformed total number of cells, and ensures that differences in the numbers of cells between groups do not cause differences in β_{hg}. Estimation of β_{hg} for each h and g is performed by fitting the GLM to the count data for each hypersphere.

With the quasi-likelihood methods in edgeR, the mean-variance relationship of each count is modelled as

$$\mathrm{var}(y_{hs}) = \sigma_h^2(\mu_{hs} + \mu_{hs}^2\phi_h) \ ,$$

where σ_h^2 is the shrunken quasi-likelihood dispersion and ϕ_h is the trended negative binomial dispersion. Any increase in the observed variance of the counts will be modelled by an increase in these two dispersions. We stress that the dispersions model the variability of the cell counts across replicate samples for each hypersphere, *not* the variability of the marker intensities across cells. Thus, it is essential that the data set contains some level of replication for dispersion estimation (though this should already be standard procedure in any good experimental design). The two dispersions have different roles in this framework:

- The NB dispersion for each hypersphere is set to the fitted value of a mean-dispersion trend [11], i.e., $\phi_h = \phi(\mu_h)$ where ϕ is the mean-dependent trend function and μ_h is the average count across all samples for each hypersphere h. This increases the accuracy of the model by accounting for empirical mean-variance relationships. In contrast, using a fixed NB dispersion would assume a quadratic relationship.

- The raw QL dispersion for each hypersphere is estimated from the deviance of the fitted GLM. A separate mean-dependent trend is fitted to the raw QL dispersions against μ_h for all hyperspheres, and robust empirical Bayes (EB) shrinkage is performed to squeeze the raw estimates towards this trend [12, 13]. The resulting values are referred to as the shrunken QL dispersions σ_h^2. These improve the accuracy of the model by accounting for hypersphere-specific variability above/below that modelled by the trended NB dispersion. EB shrinkage stabilizes the estimates in the presence of limited replication.

Hypothesis testing is performed by formulating null hypotheses in terms of the various β_{hg}. For example, in a one-way layout with groups $g \in \{1, 2\}$, one could test for differential abundance between groups by testing the null hypothesis $\beta_{h1} = \beta_{h2}$ for each hypersphere. A similar approach can be used for more complex designs – for example, if a spline is fitted to the abundances with respect to time, the corresponding coefficients can be set to zero under the null to test for any time effect. The p-value for each hypersphere is calculated with the QL F-test, which accounts for the uncertainty in estimating the shrunken QL dispersions.

3.2 Filtering hyperspheres on their average abundances

edgeR assumes that the input data have already been filtered to remove hyperspheres with low average counts. Hyperspheres with very few cells do not have enough evidence to reject the null hypothesis, even if they did contain genuine changes in abundance. Discarding them reduces the total number of tests and mitigates the severity of the multiple testing correction [14]. We demonstrate this effect in Supplementary Figure 8, where the minimum p-value from a NB GLM is a function of the mean count. For a given p-value threshold, it is not possible to reject the null hypothesis if the mean count is too low. This motivates the use of filtering to remove these uninformative tests, though the exact choice of filter threshold depends on the specified p-value threshold (which, in practice, depends on the extent of the multiple testing correction).

By default, we use an average count threshold of 5 in our analysis, i.e., a hypersphere must contain an average of 5 cells or more across all samples to be retained. This assumes that the multiple testing correction is severe such that the p-values are scaled up by 10,000 prior to rejection at an adjusted p-value threshold of 0.05 (Supplementary Figure 8). For the MEF and BMMC data sets, the correction is fairly mild so lower filter thresholds could be used, e.g., to identify rare DA subpopulations. However, relaxing the filter may not result in the detection of more hyperspheres – while more hyperspheres will be retained, the effect of the multiple testing correction will increase concomitantly, potentially reducing power. The chosen filter threshold of 5 provides a good compromise between retention of potential DA hyperspheres and removal of the majority of low-abundance hyperspheres. This simplifies the interpretation of the results by focusing on detection of changes in cell subpopulations of modest-to-high abundance that are likely to constitute the major differences between conditions. It also reduces computational work by reducing the total number of tests, and avoids problems with discreteness at low counts when fitting mean-dispersion trends in edgeR.

3.3 Normalizing the hypersphere counts across samples

As mentioned above, we define the GLM offsets as the log-transformed total number of cells per sample. This means that GLM is effectively modelling the *proportion* of cells in each sample that are located inside each hypersphere. Thus, differences in the input quantities of cells between different samples will not result in spurious DA. Obviously, this means that global changes in abundance are not detected – for example, if the abundances of all cell types increased by the same fold-change between conditions, the proportions would remain the same. We disregard these global changes as they are inherently confounded with varying input quantities in most real experiments. We also ignore hypersphere-specific biases (i.e., systematic differences in abundance between hyperspheres caused by technical effects like staining efficiency) as these are expected to cancel out when comparing counts from the same hypersphere. The use of barcoding and multiplexing avoids introducing sample-specific biases related to differences in staining efficiency or machine behaviour.

Note that our strategy for normalizing the count data does not protect against composition effects. Composition effects refer to indirect changes in the proportion of cells assigned to a subspace, caused by changes to the total number of cells between conditions. For example, consider a situation where one subspace experiences a large increase in cell abundance between conditions, while the abundances in all other subspaces remain the same. The increase in the former drives an increase in the total number of cells in the affected condition. This results in a decrease in the *proportion* of cells that are assigned to the other subspaces, leading to differences being observed in hyperspheres that do not directly exhibit a change in abundance. From a mathematical perspective, this is not incorrect as the proportions *do* change when the total counts are altered. Nonetheless, detection of such subspaces may not be biologically relevant, so composition effects should ideally be removed prior to further analysis. Unfortunately, conventional strategies for normalizing these effects (from RNA-seq data analysis [15]) are not applicable here. If these methods were applied to the hypersphere counts, they would assume that most hyperspheres are not differentially abundant. This is

unlikely to be true in many settings, e.g., due to large-scale changes upon stimulation or activation.

Rather, alternative strategies are required to mitigate composition effects. The simplest approach is to gate out any high-abundance differential subpopulations. The total count can then be calculated from the remaining cells, which avoids introducing composition effects from the differential subpopulation. Identification of problematic subpopulations can be done before the DA analysis based on existing knowledge, or afterwards based on the top DA hyperspheres. For example, consider a mixed population of T and B cells. If these cells were analyzed together, a large increase in the number of B cells in one condition would result in a concomitant decrease in the *proportion* of T cells – even if the number of T cells did not change between conditions. This result would suggest that the T cell population decreases in abundance, which might be misleading. Instead, one can gate on CD3 or CD19 to isolate T or B cells, respectively, and then analyze each of the gated populations separately. This avoids detecting indirect changes in T-cell subpopulations due to changes in B-cell abundance, and vice versa. Note that this is only necessary for DA subpopulations with many cells, as changes in small subpopulations will not have a substantial effect on the total count.

Another approach is to test for differential abundance against a minimum log-fold change threshold. This avoids detecting small changes in abundance caused by composition effects. Such changes may be statistically significant but are unlikely to be biologically relevant [16], and are subsequently ignored. While this is a more general approach than gating, it only protects against small composition effects – large changes in the dominant subpopulation may induce large changes in abundance across all hyperspheres.

3.4 Assessing edgeR's performance on mass cytometry count data

We tested the performance of edgeR using simulations constructed from the MEF reprogramming study. For each MEF time course, we pooled cells from all associated samples. We generated new samples by randomly sampling cells from the pool. Each new sample contained the same number of cells as one of the original samples. We separated the new samples into two groups (i.e., biological conditions), counted cells into hyperspheres and tested for DA between groups using edgeR. Here, we used a design matrix with a one-way layout to fit a GLM to the counts for each hypersphere. A contrast was constructed to test for differences between groups, as described above. As a comparison, we also tested the performance of the Mann-Whitney test, which is often used to detect differential proportions in flow cytometry data [17] and has been applied to mass cytometry data for the same purpose [18]. Each count in each hypesphere was converted into a proportion of the total number of cells from the corresponding sample. The Mann-Whitney test was applied to these proportions to test for significant differences between groups, using the wilcox.test function in R.

We used two sampling schemes for this simulation. The first scheme involved sampling with replacement from the pool with a constant probability of sampling each cell. This is the simplest approach but assumes that only sampling noise contributes to variability between replicates. In practice, additional biological variability will be present due to differences in the composition of cell populations extracted from replicate animals or cultures. To represent this, each cell j in the pool was assigned a probability weight R_{js} for sample s. (R_{js} was sampled from a Gamma distribution with shape and rate set to 0.01. These parameters were chosen to yield NB dispersions of 0.5-1.5 per hypersphere, comparable to values observed in Supplementary Figure 7 for real data. In contrast, the first sampling scheme yields near-zero estimates.) The probability of sampling cell j in sample s is proportional to R_{js}, thus skewing selection towards a particular subset of cells in that sample. However, as the values of R_{js} differ between samples, the favoured subset will also be different for each replicate. Thus, the weighting introduces extra variability by changing the cell composition between replicates. Note that the null hypothesis is still true in this scheme – this is because $E(R_{js})$ is the same for all samples, which means that the average cell composition is the same between groups.

As all cells were sampled from the same pool, the null hypothesis of constant abundance should be true for each hypersphere. Thus, p-values from both methods should be uniformly distributed. We calculated the observed type I error rate as the proportion of p-values below a specified threshold $\alpha = 0.01$ or 0.05. For edgeR, we found that the observed type I error rate was close to or below the specified threshold for all simulation schemes and threshold values (Supplementary Figure 9a). Accurate type I error control indicates that the specificity of edgeR is maintained when applied to counts of cells assigned to hyperspheres. edgeR also routinely yields lower p-values than the Mann-Whitney test for hyperspheres with extreme log-fold changes in abundance between conditions (Supplementary Figure 9b). This is due to the loss of power from using ranks with small sample sizes in the latter. edgeR is more sensitive as its parametric model accounts

for the size of the change in abundance, yielding smaller p-values at the same log-fold changes. These results suggest that edgeR is appropriate for detecting differential abundance in mass cytometry data.

Note 4: Controlling the spatial FDR

4.1 A detailed explanation of the spatial FDR

Let us split the M-dimensional space into arbitrarily small non-overlapping partitions of similar volume. For example, these partitions might be pixels in two-dimensional space, or voxels in three-dimensional space. Each hypersphere is assigned to the partition containing its median-based position. The outcome of the test for differential abundance for the hypersphere is used as a proxy for the outcome of its assigned partition (Supplementary Figure 10). In the simplest scenario, a partition containing only one hypersphere will be represented by that hypersphere. If multiple hyperspheres are assigned to a partition, a single hypersphere is randomly selected as the representative. We define the FDR across the expected set of representative hyperspheres (defined as the set of all hyperspheres, where each hypersphere is weighted by the probability of being sampled from its partition) as the spatial FDR. By controlling this value below a specified threshold, we effectively control the FDR across partitions, and thus, the FDR across the volume of those partitions.

In practice, we do not need to explicitly define partitions in order to control the spatial FDR. Rather, the probability of sampling a representative hypersphere from a partition is inversely proportional to the density of assigned hyperspheres in that partition. Given that our definition above considers small partitions of similar size, the local density of each hypersphere can approximate the density of the partition to which it is assigned. Thus, instead of directly allocating hyperspheres into partitions and computing the partition density (which would involve arbitrary definitions of the shape, size and arrangement of partitions), we compute the local density of each hypersphere by applying a kernel density estimator to the positions of all tested hyperspheres. The weight of each hypersphere is defined as the reciprocal of its local density. This approximates the relative probability (scaled by some constant, which can be ignored) of sampling that hypersphere as a representative of its partition. Finally, the Benjamini-Hochberg method is applied to the hypersphere p-values with the associated weights. This controls the FDR across the expected set of representative hyperspheres.

We stress that the aforementioned partitions are only necessary for the theoretical definition of the spatial FDR. There is no need to actually construct these partitions when controlling the FDR in real analyses.

As described in Figure 1d, the spatial FDR can be roughly interpreted as the proportion of the volume occupied by DA hyperspheres that corresponds to false positives. This is based on considering partitions of similar size to the hyperspheres, such that the total volume of the partitions, and the proportion of which is false positive, is similar to that of the hyperspheres. (The corresponding assumption in our control procedure would be that the kernel bandwidth is similar to the hypersphere radius. However, the results are quite robust to the choice of bandwidth – see Supplementary Table 1 – so we will ignore this subtlety.) While a definition of the spatial FDR based on the hypersphere volumes is intuitive, it is difficult to implement. Computing the total volume of (overlapping) hyperspheres in high-dimensional space is not straightforward. It is also unclear how to define the "false positive volume" in the presence of overlaps between a false and true positive hypersphere. Our definition of the spatial FDR, while less intuitive, is easier to control.

4.2 Applying the weighted Benjamini-Hochberg procedure

Let each null hypothesis i be associated with a p-value $p_{(i)}$ and a weight of $w_{(i)}$. Assume that there are n null hypotheses, ordered such that $p_{(1)} < p_{(2)} < ... < p_{(n)}$. To control the FDR at some threshold α, a weighted BH procedure is applied [19] to reject any null hypothesis where the p-value is less than the threshold

$$\max_i \left\{ p_{(i)} : p_{(i)} \leq \alpha \frac{\sum_{l=1}^{i} w_{(l)}}{\sum_{l=1}^{n} w_{(l)}} \right\} .$$

To control the spatial FDR, we apply the weighted BH method to the hypersphere statistics. Each hypersphere corresponds to one null hypothesis, while its weight is defined as the reciprocal of its local density. Here, a decision must be made regarding the choice of kernel density estimator and bandwidth. For the latter, we compute the distance from each hypersphere position to its 50^{th} nearest neighbour. The

bandwidth is defined as the median of this distance across all hyperspheres. This ensures that, on average, around 50 neighbours will be available to stably calculate the local density for each hypersphere. We also use a tricube kernel to provide some robustness to the choice of bandwidth (Supplementary Table 1). This gives more weight to closer neighbours while reducing the influence of cells that fall just inside the bandwidth.

4.3 A discussion on the relevance of the spatial FDR

An obvious question is, why we do not directly count cells into equally sized and spaced partitions, rather than using hyperspheres? This would certainly simplify FDR control as the BH method could be applied directly to the p-values for the partitions. However, choosing a value for the spacing parameter is not straightforward. A value that is too small would be computationally impractical, while a value that is too large will sacrifice spatial resolution. Our approach avoids this problem by using hyperspheres centred on cells. This means that a hypersphere will generally be present at relevant parts of the high-dimensional space (i.e., those occupied by cells), while limiting the total number of hyperspheres to be proportional to the number of cells.

It is worth noting that we control the FDR in the expected set of representative hyperspheres, rather than the expectation of the FDR with respect to all possible sets of randomly sampled representatives. Formally speaking, control of the FDR would refer to controlling the latter value. The FDR across the expected set is used in our analysis because it is simpler to compute. It also approaches the expected FDR when considering small partitions, as strong correlations betwen hyperspheres in the same partition mean that the expected set will have a similar (frequency-weighted) distribution of p-values as any instance of the sampled set.

In terms of interpretation, controlling the spatial FDR may not be the same as controlling the FDR across the underlying subpopulations. For example, consider a scenario containing a DA subpopulation of large volume and a non-DA subpopulation of small volume. Assume that both subpopulations are detected after the DA analysis. In this situation, the observed spatial FDR would be small as the proportion of volume taken up by the second (false positive) subpopulation is low. By comparison, the observed FDR across subpopulations would be larger (50%) because one of the two subpopulations is a false positive. The FDR across subpopulations is appealing as it is more intuitive than the spatial FDR. However, it is difficult to control in general as it relies on the subpopulations being well-defined. Furthermore, the use of the spatial FDR is justified by the potential presence of further substructure within the larger DA subpopulation. This necessitates separate examination of each part of the first subpopulation, proportional to its volume.

A similar issue arises from the fact that visualization and interpretation of the differential subspaces is performed in low dimensions, e.g., with PCA or t-SNE plots. There is no guarantee that the FDR across, say, pixels in the two-dimensional space is equal to the spatial FDR computed in the original M-dimensional space. One could overcome this problem by controlling the spatial FDR using hypersphere coordinates in low-dimensional space. However, this level of statistical rigour seems unnecessary for data exploration. In addition, one of the aims of the DA analysis is to simplify data visualization by only processing the subset of DA hyperspheres. It is not clear how the FDR can be rigorously controlled across the low-dimensional space if dimensionality reduction is performed on hyperspheres that have been pre-selected for significance.

We note that the concept of the spatial FDR is analogous to the FDR across areas [20] or the size-weighted FDR across clusters [21] used in analyses of functional magnetic resonance imaging (fMRI) data. fMRI data analysis is a similar problem in that the signal of interest has a spatial component (2-3 dimensions) and the aim is to control the FDR across this space. However, the application of these methods to hypersphere-based p-values is not obvious – we do not use random fields for hypothesis testing, and we explicitly avoid clustering by using hyperspheres. This motivated the development of our own spatial FDR-controlling procedure.

4.4 Assessing the use of weights for spatial FDR control

We performed simulations based on the MEF reprogramming data to test the use of density-based weights. Each sample was constructed using the weighted sampling scheme described in Note 3.4, for an experimental design containing several replicates in each of two groups. We then added a further $T_s/10$ cells to each sample s, where T_s is the original number of cells in s. The additional cells were assigned marker intensities of zero for all samples in the first group, and intensities of 1 for all samples in the second group. This represents a differential subpopulation between groups where a subpopulation at $(0, 0, \ldots, 0)$ is lost and a subpopulation at $(1, 1, \ldots, 1)$ is gained. While more complex differential events can be simulated, we use this simple set-up

as we are not interested in the true differences, but rather, their effect on the detection of false positives.

We applied the BH method to the p-values for all hyperspheres, either directly or with density weights. Detected DA hyperspheres were defined as those with adjusted p-values below 0.05. To measure the observed spatial FDR across the detected hyperspheres, we partitioned the M-dimensional space into non-overlapping hypercubes with side lengths ranging from 0.2 to 1. Each DA hypersphere was assigned to the hypercube containing its median-based position. For each non-empty hypercube, we computed the proportion of its assigned hyperspheres that were not truly differential. The observed spatial FDR was defined as the mean of these proportions across all non-empty hypercubes. (This reflects the definition in Note 4.1. Each hypercube represents a partition of similar volume, from which one hypersphere is chosen as a representative. The mean proportion is an estimate of the expected proportion of representatives that are false positives.)

In our simulations, the BH method with density weights was able to control the observed spatial FDR close to or below the specified threshold (Supplementary Figure 11). In comparison, naïve application of the BH method (i.e., without weighting) failed to control the spatial FDR. This is because the naïve approach controls the FDR across hyperspheres, which is generally not equivalent to controlling the FDR across volume (Figure 1d). The weighting approach performs better as it explicitly controls for the latter feature.

Note 5: Detailed interpretation of non-redundant hyperspheres

While dimensionality reduction is useful for providing an overview of the data, it necessarily discards information that may actually be biologically relevant. To assist users with more detailed interpretation of individual hyperspheres, we provide an option to prune out redundant hyperspheres. First, hyperspheres are sorted by their p-value. A hypersphere is considered non-redundant if its position is more than 1 intensity unit away (in one or more dimensions) from another non-redundant hypersphere with a lower p-value. One unit represents an approximately 10-fold change in marker intensity between hyperspheres, which is large enough to warrant separate examination of those hyperspheres, though users can adjust this threshold as desired. The non-redundancy criterion is evaluated for all hyperspheres in order of increasing p-value, which ensures that DA hyperspheres are reported as non-redundant rather than their neighbouring non-DA counterparts. This procedure yields a small number of non-redundant hyperspheres that can be examined individually. For example, over 7000 hyperspheres were detected as significantly DA in the *Oct4*-GFP reprogramming time course, but only 325 were considered to be non-redundant. These can be investigated using a simple user interface like that shown in Supplementary Figure 12. This provides a clear view of the median intensity of each marker and facilitates the identification of the biological subpopulation represented by each hypersphere.

Note 6: Description of the MEF reprogramming study

6.1 Overview

We demonstrate our approach using data from a study of mouse embryonic fibroblast (MEF) reprogramming [4]. In this study, primary and secondary MEFs were reprogrammed to induced pluripotent stem cells. Primary MEFs expressing green fluorescent protein (GFP) from the endogenous *Oct4* locus (*Oct4*-GFP) were transduced with lentiviruses expressing a doxycycline-inducible suite of reprogramming factors. Secondary MEFs expressing either GFP or a neomycin resistance gene from the endogenous *Nanog* locus (*Nanog*-GFP or *Nanog*-Neo, respectively) already contained doxycycline-inducible reprogramming transgenes. For each MEF reprogramming system, a time course was constructed by taking samples at 13-15 timepoints between days 0 and 20 (for *Oct4*-GFP) or 30 (for *Nanog*-GFP or *Nanog*-Neo) after doxycycline-induced transgene expression. All samples from each time course were barcoded, stained with metal-conjugated antibodies and profiled by mass cytometry. The aim of the original data analysis – and of our re-analysis – was to detect subpopulations that change in abundance over time within each reprogramming system.

6.2 Annotating subpopulations in the *Oct4*-GFP time course

To annotate Figure 2 in the main text, we examined the marker intensities of each characterised subpopulation in Figure 3D of Zunder *et al.* [4] and identified the cluster with the most similar intensities in our

Supplementary Figure 13. The important markers for each subpopulation are listed below:

- MEFs were primarily defined as THY1^{high}mEF-SK4^{high}CD140a^{high} and OCT4^{low}SOX2^{low} with lower expression of KLF4 relative to other subpopulations. This was visually supported by the fact that the same cells were pS6^{high}β-Catenin^{high}IκBα^{high} in the original figure.
- The OSKM non-expressing population was identified as OCT4^{low}SOX2^{low} with lower KLF4. In addition, a majority of cells exhibited lower THY1 and CD140a expression compared to the neighbouring MEF population. This was further supported by low expression of pS6 in a majority of these cells, as well as reduced β-Catenin and IκBα expression and higher p53 expression relative to MEFs.
- A small subset of MEFs in Figure 2 exhibited very high CD140a expression. This represents OSKM non-expressing cells that reverted to a MEF-like endpoint state after doxycycline withdrawal.
- The OCT4^{high}KLF4^{high} population was identified as named, supported by high expression of SOX2.
- The SC4-like population was identified as KLF4^{high}OCT4^{high}CD73^{high}, additionally supported by high expression of KI67 and low expression of SOX2.
- Cells undergoing mesenchymal-epithelial transition were identified as EPCAM^{high}, supported by high expression of both β-Catenin and OCT4 in the majority of cells.
- KI67^{low} reverting cells were identified as KI67^{low}CD73^{high}. This annotation was supported by high expression of mEF-SK4 in parts of the population.
- The NANOG^{high} trajectory in the original study was a continuum of NANOG expression, starting from NANOG^{low} cells and progressing to NANOG^{high} cells. In our re-analysis, it manifests in Supplementary Figure 13 as several NANOG-intermediate subpopulations close to the LIN28^{high} and embryonic stem cell (ESC)-like subpopulations. The annotation was supported by high expression of EPCAM and SSEA1, as well as higher pS6 expression in some parts of the trajectory.
- ESC-like cells were identified as the population with the highest NANOG expression and high expression of SSEA1 and LIN28. This was supported by high levels of H3K9ac, H4Kac, CD54, pSRC and pERK.
- LIN28^{high} cells were identified as named, supported by high expression of CD24. They were further distinguished from ESC-like cells by having higher expression of CD140a and lower expression of NANOG.
- The KI67^{high} population was identified as named, supported by lower expression of OCT4.
- The mixed 4F population from the original study was highly heterogeneous and difficult to characterize. We identified it as mostly KLF4^{low}, containing subpopulations with intermediate c-Myc expression and both high and low SOX2 expression. However, we were unable to identify the OCT4^{low} subpopulations.

In summary, we were able to detect most of the previously defined subpopulations as being DA over the time course. This included intermediate populations as well as the reprogramming end points, i.e., the ESC-like cells, the mesendoderm-like LIN28^{high} cells, and THY1^{high}mEF-SK4^{high} cells that likely failed to reprogram and instead reverted to a MEF-like phenotype. In particular, the abundance of MEFs dropped over time while the abundance of ESC-like cells increased, consistent with the effects of reprogramming. Many of these subpopulations were further resolved into distinct subsets based on specific markers such as IdU.

For several subpopulations, we examined changes in abundance at critical junctures in the time course. Supplementary Figure 18 shows the effect of doxycycline-induced transgene expression and doxycycline withdrawal. In the former, the MEFs decrease in abundance while the mixed 4F population begins to increase, consistent with induction of some of the reprogramming factors. Upon withdrawal, the mixed 4F and SC4-like subpopulations decrease as cells progress towards the reprogramming endpoints. This highlights the flexibility of our analysis, where the data can be easily interrogated to study specific changes of interest.

We also identified changes in abundance of subpopulations that were not explicitly classified by Zunder *et al.* This included an increase in potentially apoptotic cells with cleaved Caspase-3; a non-linear change in abundance of a subpopulation of SC4-like cells with phosphorylated STAT3, AMPK and PLK1 (Supplementary Figure 19); and a decrease in abundance of a subpopulation of cells simultaneously expressing high levels of the MEF marker THY1 along with the reprogramming factors SOX2 and OCT4. Thus, our DA analysis method is able to identify significant changes in abundance even in small or transitional subpopulations.

6.3 Comparison to the Zunder *et al.* analysis

Our analytical pipeline offers several advantages over the original analysis of Zunder *et al.* [4]. The latter is mainly descriptive, whereas the error-controlling procedures in our analysis provide a greater degree of confidence as to whether the detected changes in abundance are genuine. Figure 2 also uses both visual dimensions to separate subpopulations, while the original analysis reserves one dimension for time. This enhances resolution of distinct subpopulations, albeit at the cost of temporal resolution. Of course, our approach is modular so alternative visualization schemes can be easily applied (on the significant hyperspheres, or the cells contained within them) to focus on particular aspects of the data. For example, methods like SPADE can be applied to organize the detected DA hyperspheres into a tree for visualization of lineages.

Note 7: Identifying subpopulations in the BMMC data set

Levine *et al.* [9] examined bone marrow aspirates from healthy donors and acute myeloid leukemia patients under various brief stimulation conditions. To demonstrate the general applicability of our method, we performed a limited re-analysis of this data set, with the aim of identifying changes in subpopulation abundance upon IL-10 treatment in the five healthy donors only. We obtained the raw FCS files for the relevant samples from Cytobank (accession 44185). Pre-processing of the marker intensities was performed as described for the MEF reprogramming data set, with some modifications. Specifically, cells with high outlier values for the viability marker were gated out to remove dead cells, and intensities were range-normalized across batches as described above. We then used our pipeline to test for differences in abundance between each treated sample and its donor-matched untreated control. Here, counts were modelled using an additive design for treatment with a donor blocking factor. We also increased $r_0\sqrt{2}$ to 0.55 to ensure that most hyperspheres contained at least 15-20 cells, based on a plot similar to Supplementary Figure 2. Finally, to simplify interpretation, we ran t-SNE separately on the significant hyperspheres with positive and negative log-fold changes. This yielded two sets of coordinates that were placed side-by-side in a single plot for convenient visualization.

Our DA analysis revealed several differential subpopulations at a spatial FDR of 5% – namely, a $\text{CD11b}^{\text{high}}$ $\text{CD64}^{\text{high}}$ population of monocytes, a $\text{CD19}^{\text{high}}$ population of B cells and a CD3^{high} population of T cells (Supplementary Figures 20-21). Within each subpopulation, $\text{pSTAT3}^{\text{high}}$ hyperspheres increased in abundance upon stimulation compared to the untreated samples, while $\text{pSTAT3}^{\text{low}}$ hyperspheres decreased in abundance. This represents an increase in pSTAT3 expression, consistent with the expected effects of IL-10 on mature immune cells. Similar to Levine *et al.*, we found that only mature bone marrow populations responded to IL-10 stimulation. No changes in abundance were observed in $\text{CD34}^{\text{high}}\text{CD3}^{\text{low}}\text{CD11b}^{\text{low}}\text{CD19}^{\text{low}}$ hyperspheres corresponding to progenitor populations. We also identified new subpopulations such as a subset of CD3^{high} cells that responded to IL-10 stimulation with increased pSTAT3 expression but did not express the T lineage marker CD7. Thus, our method is able to provide statistical rigour as well as additional biological insight.

Note 8: Detecting biological shifts in marker intensity

8.1 Overview

A complementary approach to DA analyses is to detect differential marker expression within the same subpopulation. Given a (manually or automatically defined) subpopulation, the average intensity of a particular marker can be compared between conditions [18, 22]. This detects changes in the expression of activation or signalling markers within a subpopulation. (In this note, we consider these as *biological* shifts in intensity, in contrast to the technical shifts discussed in Note 2.3.) Our pipeline is different in that it tests for changes in the abundance of cells, rather than their marker intensities. However, these two types of differential events are closely related. Consider a marker X in subpopulation Y that increases in expression between two conditions. In our analysis, this will manifest as the appearance of new subpopulation Y', separated from Y in the high-dimensional space by an increase in the intensity of X. This new subpopulation will then be detected as DA between conditions. We observed this effect in our re-analysis of the BMMC data set, where an increase in pSTAT3 levels upon IL-10 stimulation manifested as an increase in the abundances of $\text{pSTAT3}^{\text{high}}$

subpopulations and a concomitant decrease in the abundances of pSTAT3^{low} subpopulations. Thus, changes in intensity can be interpreted as changes in abundance for detection with a DA analysis pipeline.

8.2 Direct detection of intensity shifts between conditions

That said, it may be desirable to test for shifts in intensity directly, as this is easier to interpret than identifying opposing changes in two separate DA subpopulations. We implemented this functionality using the statistical methods in the limma package [23]. In this analysis, we separate our markers into two sets:

1. Markers to be tested for shifting. This typically includes markers that change in intensity *within a single cell* over the conditions being compared, e.g., during activation or stimulation.

2. Markers that will not be tested for shifting. This includes markers that do not change in intensity for each cell, e.g., when defining cell types. Here, testing for shifts has no clear biological meaning.

Hyperspheres are constructed using only the non-shifting markers in Set 2. Within each hypersphere, the median intensity of a shifting marker in Set 1 is computed across all cells in each sample. This yields S median intensities for each hypersphere where S is the number of samples. We perform weighted linear regression on the set of medians for each hypersphere, under the assumption that the errors are normally distributed. The weight for each median is set to the number of cells from that sample in the hypersphere. The precision of the sample median is approximately proportional to the number of values from which it is computed [24] – this ensures that a median calculated from a few cells is downweighted to reflect its instability.

Once the linear model is fitted to the medians, a variance estimate is obtained for each hypersphere. Note that this represents the variability of the median intensities across replicates, *not* the variability of the cell intensities within each sample, *nor* the variability of the cell counts across samples. Empirical Bayes shrinkage is performed to stabilize the variance estimates by sharing information between hyperspheres. Hypothesis testing is done using the moderated t-test, yielding a p-value against the null hypothesis that the median intensity is the same across conditions. In this manner, each hypersphere is tested for shifts in intensity for each marker in Set 1. The spatial FDR is then controlled using these p-values. This is reasonable as the only purpose of the spatial FDR is to account for variations in the density of hyperspheres. Thus, the procedure can be applied to any hypersphere-level p-values, be they for intensity shifts or changes in abundance.

It should be stressed that any markers to be tested for shifting (i.e., in Set 1) should *not* be used when assigning cells to hyperspheres (Set 2 only). Consider a marker X containing potential shifts in intensity between conditions. A hypersphere that is centred on a cell with high X expression would, by definition, contain neighbouring cells that also have high expression of X. This means that there would be little, if any, shifts in intensity for X within this hypersphere. Rather, the relevant comparison would be to cells with low expression of X, but these would be in a different hypersphere and not visible to the testing procedure.

8.3 Comparison of detection power

Of the two approaches to detecting intensity shifts – direct detection by linear modelling of medians, or through the DA framework – which should be used? To answer this question, we set up simple simulations with intensity shifts of varying size and scope. We generated intensity values for 20000 cells with 30 markers, each of which was sampled independently from a Normal(1, 0.5) distribution. These cells were randomly allocated to four samples, with two groups of two replicates each. From the cells in the second group, we randomly selected a proportion P and increased the intensity of the last marker by I for the selected cells. We used $P = 1$ and 0.1 as well as $I = \log_{10}(2)$ and $\log_{10}(40)$, i.e., a 2- or 40-fold increase in marker expression on the raw scale. In each simulation scenario, we set up hyperspheres on the first 29 markers and used the direct detection method to test for intensity shifts between groups. We considered the shifting event to be successfully detected if a hypersphere was detected at a spatial FDR of 5%, had a positive log-fold change in the median intensities for the last marker and contained at least one cell with shifted intensity. We also set up hyperspheres using intensities of all markers and performed a DA analysis between groups. Here, detection was considered to be successful if a hypersphere was detected at a spatial FDR of 5%, had a positive log-fold change in abundance and a median-based position within 0.5 intensity units of $1 + I$ for the last marker.

We observed that both methods were able to consistently detect shifts in intensity affecting the entire population of cells (Supplementary Figure 22). Even small shifts were detected by the DA approach – while

the shifted population in one group and the unshifted population in the other group still overlap substantially in M-dimensional space, the differences in cell density are sufficient to detect the shifting event. A more interesting scenario is that of a large shift in a subset of cells, where the robustness of the median becomes a disadvantage for the direct approach. Power is reduced because the median is not suited to capturing changes in a small subset of cells. In contrast, the DA approach simply defines new hyperspheres for the affected subset, allowing the shifting to be easily observed as a new subpopulation. Finally, both methods perform poorly for small shifts in a subset of cells. This is to be expected as the shift is weak and hard to detect.

These results suggest that the DA approach provides more power for detection of intensity shifts. It also allows detection of subpopulations with concomitant changes in multiple markers, e.g., if multiple signalling pathways are activated. This simply manifests as DA hyperspheres where the median-based positions have high intensity values for the appropriate combination of markers. On the other hand, the direct approach is more interpretable as it explicitly describes the changes in intensity of a marker in a given subpopulation. Thus, we have implemented both methods in our framework as each complements the other's capabilities.

Note 9: Comparing to approaches based on clustering

9.1 Overview

As mentioned earlier, existing methods for analyzing mass cytometry data involve an initial clustering step. This approach can be easily extended to DA analyses, where the number of cells in each cluster from each sample is counted, and the counts for each cluster are tested for differences between conditions. We compare our hypersphere-based method against two cluster-based approaches – a custom approach involving hierarchical clustering of cells followed by testing with edgeR; and the CITRUS software [25], which uses statistical methods developed for microarrays [26]. While testing for changes in abundance within clusters is intuitive, it is subject to the performance of the clustering algorithm [27, 28]. This is especially relevant for subpopulations that are not clearly separated from each other. For example, if a DA subpopulation is incorrectly clustered with a non-DA subpopulation, any change in the former will be masked by the latter. This will compromise detection power of the subsequent DA analysis. Cluster formation also depends upon the choice of algorithm and parameters [29, 30], which complicates the assessment of cluster reliability.

9.2 Simulation design and analysis

We simulated data for an experimental design involving 30 markers and two replicates in each of two conditions. For a population of 20000 cells, we sampled intensities for each marker from a Normal(1, 0.5) distribution. Each of these cells was randomly allocated to a sample, producing a non-DA population with equal representation in both conditions. We also simulated intensities for two subpopulations of 40 cells each, where the cells from each subpopulation were allocated to samples in one condition only. Marker intensities for cells in these subpopulations were sampled from a Normal(x, 0.3) distribution, where $x = 4$ for the first marker in the first subpopulation and $x = 2$ otherwise. This yields two small DA subpopulations that lie adjacent to each other but change in opposite directions between conditions (Supplementary Figure 23).

To perform the custom cluster-based DA analysis, all cells were used for complete-linkage hierarchical clustering based on the Euclidean distances between cells in the M-dimensional space. Clusters were defined by cutting the dendrogram with the cutree command in R to generate 5-500 clusters. For each cluster, the number of cells from each sample was counted. These counts were analyzed in edgeR to identify clusters with significant differences between conditions, as previously described. Correction for multiple testing was performed by directly applying the BH method to the cluster-level p-values. Detected clusters were defined at an FDR of 5%. To run CITRUS, the citrus.full command was used with family set to "classification", featureType set to "abundances" and modelType set to "sam". Downsampling was performed to 1000 cells per sample and the minimum cluster size was set to 5%. Clusters with significant differences between groups were detected at an FDR of 5%, as reported by the SAM method. For our hypersphere-based method, the DA analysis was performed as described and hyperspheres were detected at a spatial FDR of 5%.

For the cluster-based methods, the centre of each cluster in M-dimensional space was defined from the median intensity across its cells for each marker. We use the cluster centre as a summary of the location of the entire cluster, as this reflects the common use of the median marker intensity to characterise cell clusters in

practical applications [31, 25]. Each simulated DA subpopulation was considered to be successfully detected if the centre of a detected cluster was within $0.5\sqrt{M}$ of the true subpopulation centre. This assessment was repeated using the median-based positions of DA hyperspheres. For each method, we computed the percentage of simulation iterations in which each DA subpopulation was successfully detected.

We observed that these subpopulations were consistently detected as being differentially abundant by our hypersphere-based method but not by most of the cluster-based methods (Supplementary Figure 23). This is because clusters cannot be unambiguously defined in this scenario, such that a cluster corresponding to one of the DA subpopulations will include cells from the other subpopulation. Subsequently, the power to detect this cluster is reduced as the DA log-fold change in one direction is weakened by the contribution from the subpopulation changing in the other direction. This is likely to be problematic in situations where subpopulations are not clearly separated. Examples include gating strategies commonly used in immunophenotyping to characterise subsets of T and B cells [32], as well as protocols for isolation of haematopoietic stem cell and progenitor populations [33]. Indeed, many of the subpopulations in the MEF data set form a continuous trajectory over time [4], so suboptimal DA detection due to ambiguous cluster formation is not surprising (Supplementary Figure 24). In such cases, the use of hyperspheres may be more appropriate because each subspace is tested for differences, even if the underlying biological subpopulations are poorly defined.

9.3 Issues associated with overclustering

As Supplementary Figure 23 demonstrates, the use of a large number of clusters (i.e., "over-clustering") can mitigate the disadvantages of cluster-based methods compared to hyperspheres. At a certain number of clusters, the size of each cluster will be roughly similar to that of each hypersphere. This improves the spatial resolution of clustering and increases the power to detect small DA subpopulations. However, the need to explicitly define cluster boundaries still leads to some loss of power relative to our hypersphere-based method, particularly in cases where the separation between DA and non-DA subpopulations is not obvious.

There are also some practical issues with the use of over-clustering in routine analyses. The most obvious problem is an appropriate choice of the number of clusters k. If k is too small, resolution is lost, while if k is too large, the counts per cluster will be too low to reject the null hypothesis. In some respects, this is analogous to the choice of the hypersphere radius r. However, r scales with the number of markers and is largely agnostic to the true number of subpopulations in the data set. This is not the case with k, e.g., a choice of $k = 50$ in a data set with 10-20 subpopulations defined by 10 markers may not be sufficient when analyzing a more heterogeneous data set with 30 markers defining hundreds of distinct subpopulations.

While the DA analyses can be repeated with multiple k, integrating these analyses into a single set of results is challenging. This is due to the redundancies and dependencies between different k, which affects FDR control and interpretation of the results. For example, if a cluster is detected as DA, and all its internal subclusters are also detected, how many discoveries does that actually constitute? If some of the subclusters are false positives, what would the FDR be? Should all, some or none of the subclusters be reported and visualized? Even with a single k, the FDR across clusters is not easily interpretable after over-clustering, because we cannot assume that each cluster represents (accurately or otherwise) some biologically meaningful subpopulation. Rather, the interpretation would become closer to that of the spatial FDR, possibly requiring some additional work to explicitly control the spatial FDR if the volume of each cluster is not the same.

9.4 Comparing to CITRUS in other modes

The CITRUS software has several other modes of operation – one to identify features of subpopulations that associate with the conditions ("glmnet") and another to build a classifier ("pamr"). These modes are primarily intended for prediction and classification. In contrast, we perform formal hypothesis testing to identify significant changes (in abundance, or shifts in intensity) between conditions for explanatory purposes. These functions are not easily comparable as the aims of prediction and explanation are distinct from a statistical perspective [34]. For example, consider a treatment condition that increases the intensity of two markers, e.g., pSTAT3 and pSTAT5. Our approach would detect significant shifts between conditions in both of these markers. However, a prediction strategy would only be obliged to include one marker in the predictive model, as the other marker is redundant and unnecessary for distinguishing between conditions.

References

[1] D. R. Parks, M. Roederer, and W. A. Moore. A new "Logicle" display method avoids deceptive effects of logarithmic scaling for low signals and compensated data. *Cytometry A*, 69(6):541–551, 2006.

[2] E. R. Zunder, R. Finck, G. K. Behbehani, e. l. A. D. Amir, S. Krishnaswamy, V. D. Gonzalez, C. G. Lorang, Z. Bjornson, M. H. Spitzer, B. Bodenmiller, W. J. Fantl, D. Pe'er, and G. P. Nolan. Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell deconvolution algorithm. *Nat. Protoc.*, 10(2):316–333, 2015.

[3] O. I. Ornatsky, X. Lou, M. Nitz, S. Schafer, W. S. Sheldrick, V. I. Baranov, D. R. Bandura, and S. D. Tanner. Study of cell antigens and intracellular DNA by identification of element-containing labels and metallointercalators using inductively coupled plasma mass spectrometry. *Anal. Chem.*, 80(7):2539–2547, 2008.

[4] E. R. Zunder, E. Lujan, Y. Goltsev, M. Wernig, and G. P. Nolan. A continuous molecular roadmap to iPSC reprogramming through progression analysis of single-cell mass cytometry. *Cell Stem Cell*, 16(3):323–337, 2015.

[5] F. Hahne, N. LeMeur, R. R. Brinkman, B. Ellis, P. Haaland, D. Sarkar, J. Spidlen, E. Strain, and R. Gentleman. flowCore: a Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics*, 10:106, 2009.

[6] B. Gaudillière, G. K. Fragiadakis, R. V. Bruggner, M. Nicolau, R. Finck, M. Tingle, J. Silva, E. A. Ganio, C. G. Yeh, W. J. Maloney, J. I. Huddleston, S. B. Goodman, M. M. Davis, S. C. Bendall, W. J. Fantl, M. S. Angst, and G. P. Nolan. Clinical recovery from surgery correlates with single-cell immune signatures. *Sci. Transl. Med.*, 6(255):255ra131, 2014.

[7] B. Gaudillière, E. A. Ganio, M. Tingle, H. L. Lancero, G. K. Fragiadakis, Q. J. Baca, N. Aghaeepour, R. J. Wong, C. Quaintance, Y. Y. El-Sayed, G. M. Shaw, D. B. Lewis, D. K. Stevenson, G. P. Nolan, and M. S. Angst. Implementing mass cytometry at the bedside to study the immunological basis of human diseases: Distinctive immune features in patients with a history of term or preterm birth. *Cytometry A*, 87(9):817–829, 2015.

[8] J. O. Ramsay and B. W. Silverman. *Applied functional data analysis: methods and case studies*, volume 77. Springer, New York, 2002.

[9] J. H. Levine, E. F. Simonds, S. C. Bendall, K. L. Davis, e. l. A. D. Amir, M. D. Tadmor, O. Litvin, H. G. Fienberg, A. Jager, E. R. Zunder, R. Finck, A. L. Gedman, I. Radtke, J. R. Downing, D. Pe'er, and G. P. Nolan. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, 2015.

[10] A. T. Lun, Y. Chen, and G. K. Smyth. It's DE-licious: a recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR. *Methods Mol. Biol.*, 1418:391–416, 2016.

[11] D. J. McCarthy, Y. Chen, and G. K. Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, 40(10):4288–4297, 2012.

[12] S. P. Lund, D. Nettleton, D. J. McCarthy, and G. K. Smyth. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Stat. Appl. Genet. Mol. Biol.*, 11(5), 2012.

[13] B. Phipson, S. Lee, I. J. Majewski, W. S. Alexander, and G. K. Smyth. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Ann. Appl. Stat.*, 10(2):946–963, 2016.

[14] R. Bourgon, R. Gentleman, and W. Huber. Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci. U.S.A.*, 107(21):9546–9551, 2010.

[15] M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, 11(3):R25, 2010.

[16] D. J. McCarthy and G. K. Smyth. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*, 25(6):765–771, 2009.

[17] J. V. Watson. *Flow Cytometry Data Analysis*, chapter 4 - Significance testing and fit criteria. Cambridge University Press, 1992. Cambridge Books Online.

[18] G. K. Behbehani, N. Samusik, Z. B. Bjornson, W. J. Fantl, B. C. Medeiros, and G. P. Nolan. Mass cytometric functional profiling of acute myeloid leukemia defines cell-cycle and immunophenotypic properties that correlate with known responses to therapy. *Cancer Discov*, 5(9):988–1003, 2015.

[19] Y. Benjamini and Y. Hochberg. Multiple hypotheses testing with weights. *Scand. J. Stat.*, 24(3):407–418, 1997.

[20] M. P. Pacifico, C. Genovese, I. Verdinelli, and L. Wasserman. False discovery control for random fields. *J. Am. Stat. Assoc.*, 99(468):1002–1014, 2004.

[21] Y. Benjamini and R. Heller. False discovery rates for spatial signals. *J. Am. Stat. Assoc.*, 102(480):1272–1281, 2007.

[22] B. Anchang, T. D. Hart, S. C. Bendall, P. Qiu, Z. Bjornson, M. Linderman, G. P. Nolan, and S. K. Plevritis. Visualization and cellular hierarchy inference of single-cell data using SPADE. *Nat. Protoc.*, 11(7):1264–1279, 2016.

[23] G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, 3:Article3, 2004.

[24] P. R. Rider. Variance of the median of small samples from several special populations. *J. Am. Stat. Assoc.*, 55(289):148–150, 1960.

[25] R. V. Bruggner, B. Bodenmiller, D. L. Dill, R. J. Tibshirani, and G. P. Nolan. Automated identification of stratifying signatures in cellular subpopulations. *Proc. Natl. Acad. Sci. U.S.A.*, 111(26):E2770–2777, 2014.

[26] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.*, 98(9):5116–5121, 2001.

[27] M. K. Kerr and G. A. Churchill. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci. U.S.A.*, 98(16):8961–8965, 2001.

[28] T. Ronan, Z. Qi, and K. M. Naegle. Avoiding common pitfalls when clustering biological data. *Sci. Signal.*, 9(432):re6, 2016.

[29] S. Datta and S. Datta. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19(4):459–466, 2003.

[30] C. Wiwie, J. Baumbach, and R. Rottger. Comparing the performance of biomedical clustering methods. *Nat. Methods*, 12(11):1033–1038, 2015.

[31] P. Qiu, E. F. Simonds, S. C. Bendall, K. D. Gibbs, R. V. Bruggner, M. D. Linderman, K. Sachs, G. P. Nolan, and S. K. Plevritis. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.*, 29(10):886–891, 2011.

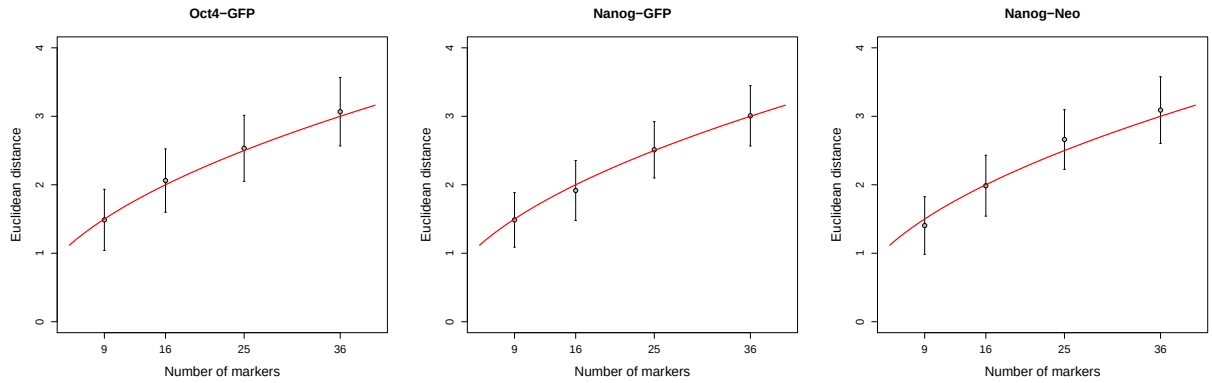
[32] G. Finak, M. Langweiler, M. Jaimes, M. Malek, J. Taghiyar, Y. Korin, K. Raddassi, L. Devine, G. Obermoser, M. L. Pekalski, N. Pontikos, A. Diaz, S. Heck, F. Villanova, N. Terrazzini, F. Kern, Y. Qian, R. Stanton, K. Wang, A. Brandes, J. Ramey, N. Aghaeepour, T. Mosmann, R. H. Scheuermann, E. Reed, K. Palucka, V. Pascual, B. B. Blomberg, F. Nestle, R. B. Nussenblatt, R. R. Brinkman, R. Gottardo, H. Maecker, and J. P. McCoy. Standardizing flow cytometry immunophenotyping analysis from the Human ImmunoPhenotyping Consortium. *Sci. Rep.*, 6:20686, 2016.

[33] N. K. Wilson, D. G. Kent, F. Buettner, M. Shehata, I. C. Macaulay, F. J. Calero-Nieto, M. Sanchez Castillo, C. A. Oedekoven, E. Diamanti, R. Schulte, C. P. Ponting, T. Voet, C. Caldas, J. Stingl, A. R. Green, F. J. Theis, and B. Gottgens. Combined single-cell functional and gene expression analysis resolves heterogeneity within stem cell populations. *Cell Stem Cell*, 16(6):712–724, 2015.

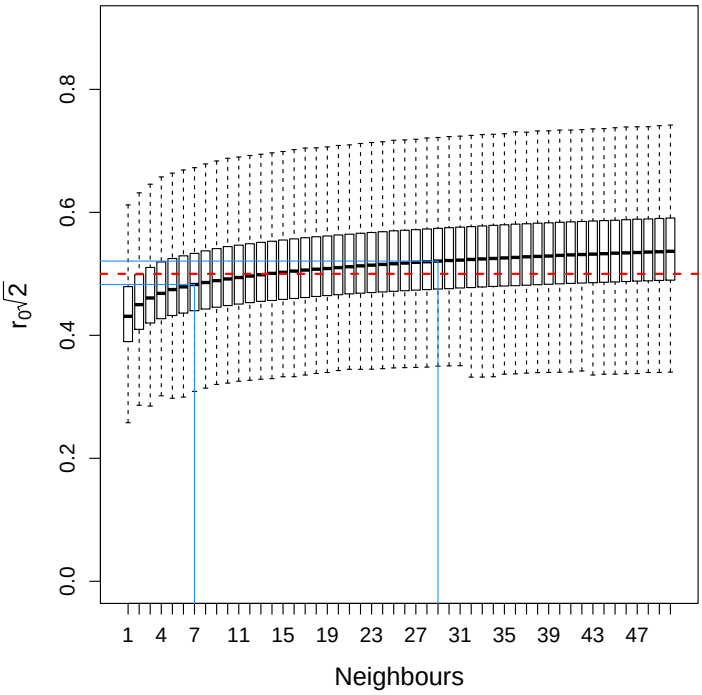
[34] G. Shmueli. To explain or to predict? *Statist. Sci.*, 25(3):289–310, 2010.

Supplementary Table 1: Effect of changes to the hypersphere radius on the DA analysis. The number of hyperspheres retained after filtering is shown, along with the number detected with significant differences at a spatial FDR of 5%. In each altered analysis, the number of DA hyperspheres gained or lost relative to the original analysis is reported. The effect of altering the bandwidth in the spatial FDR calculation was also tested, by defining the bandwidth using the 20^{th} (smaller) and 100^{th} nearest neighbour (larger).

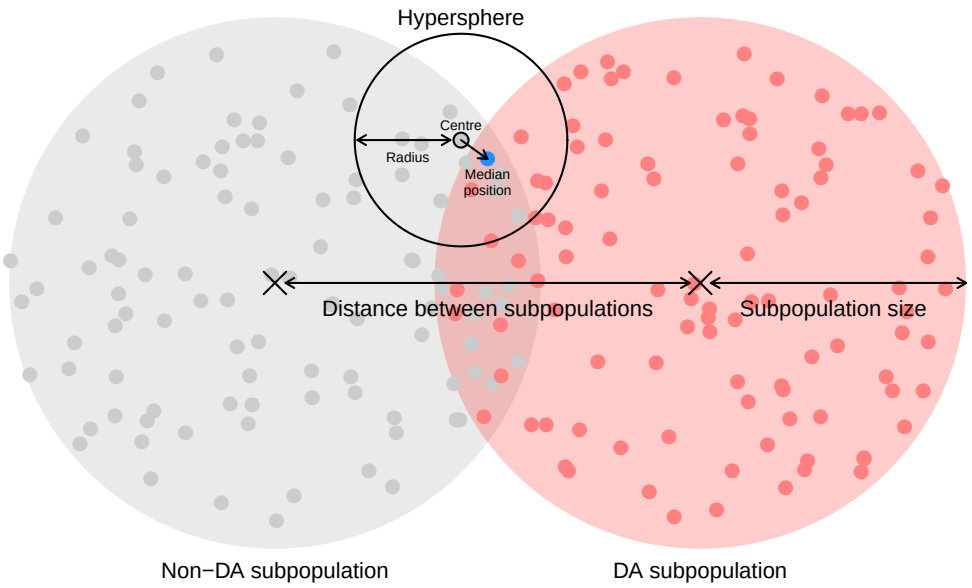
Dataset	**Statistic**	**Original**	**Value of $r_0\sqrt{2}$**		**Bandwidth**	
			0.48	*0.52*	*Smaller*	*Larger*
Oct4-GFP	Total	7720	4984	10799	7720	7720
	Significant	7416	4837	10242	7418	7414
	Gained	-	35	2916	2	0
	Lost	-	2614	90	0	2
Nanog-GFP	Total	6297	4137	8700	6297	6297
	Significant	5947	3917	8291	5947	5944
	Gained	-	63	2389	0	0
	Lost	-	2093	45	0	3
Nanog-Neo	Total	22043	15025	28944	22043	22043
	Significant	21532	14663	28271	21532	21532
	Gained	-	53	6809	0	0
	Lost	-	6922	70	0	0



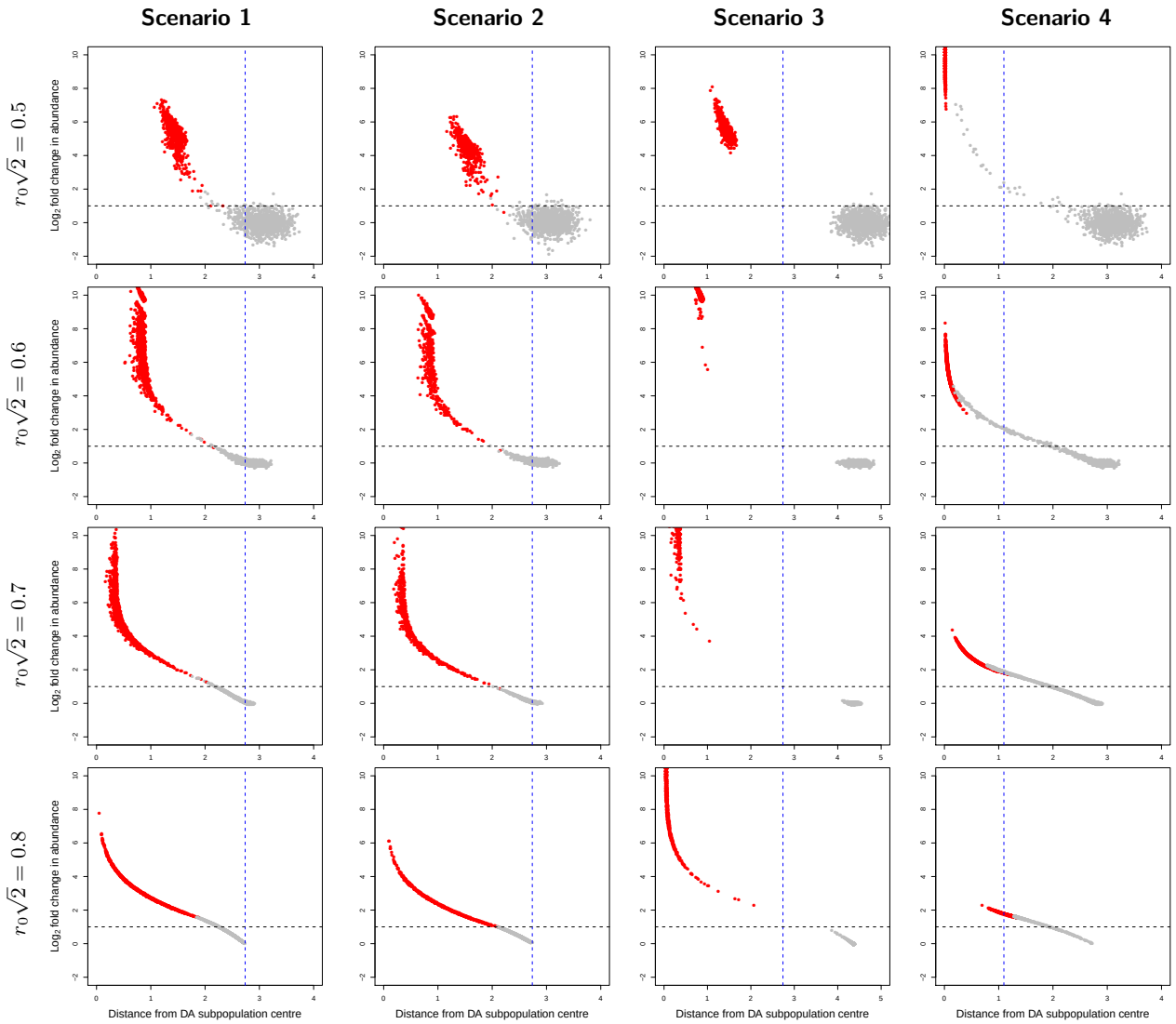
Supplementary Figure 1: Euclidean distance from each cell to its nearest neighbours as a function of the number of markers. For each cell in the first sample of each time course in the MEF reprogramming study, its nearest neighbours were identified in the full (36-dimensional) space. A subset of 9-25 markers were randomly chosen and the distance to each neighbour was recalculated for each cell in the reduced space. Here, the distance to the 10^{th} nearest neighbour is shown for *Oct4*-GFP and *Nanog*-Neo, while that to the 1^{st} nearest neighbour is shown for *Nanog*-GFP. Each point represents the mean distance and the error bar represents the sample standard deviation across all cells for a given number of markers. This calculation was also repeated using all 36 markers. The hypersphere radius used for each number of markers is shown in red.



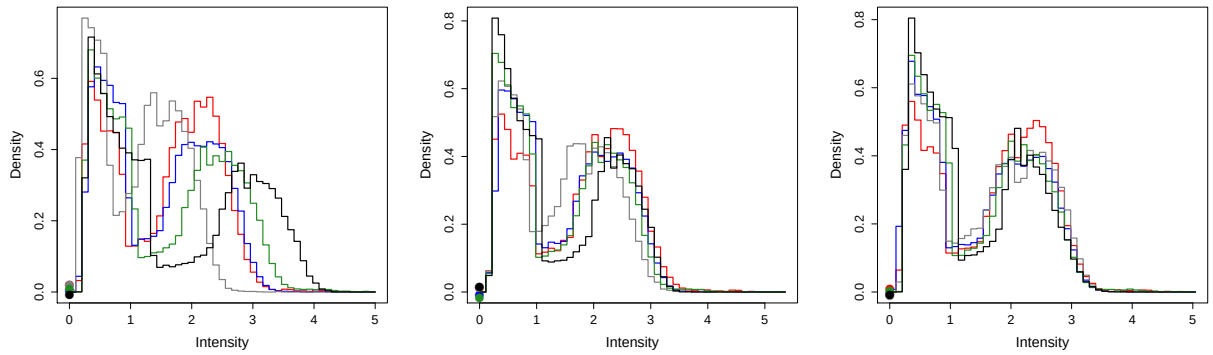
Supplementary Figure 2: The relationship between the radius (based on the choice of $r_0\sqrt{2}$) and the cell count for each hypersphere in the *Oct4*-GFP time course. For each hypersphere centred on a cell, the radius required to include a certain number of nearest neighbours is computed. The distribution of radii across all hyperspheres is shown as a boxplot for each neighbour. The red line represents the default $r_0\sqrt{2} = 0.5$. The blue lines mark the $r_0\sqrt{2}$ corresponding to median distances required to include 7 and 29 neighbours.



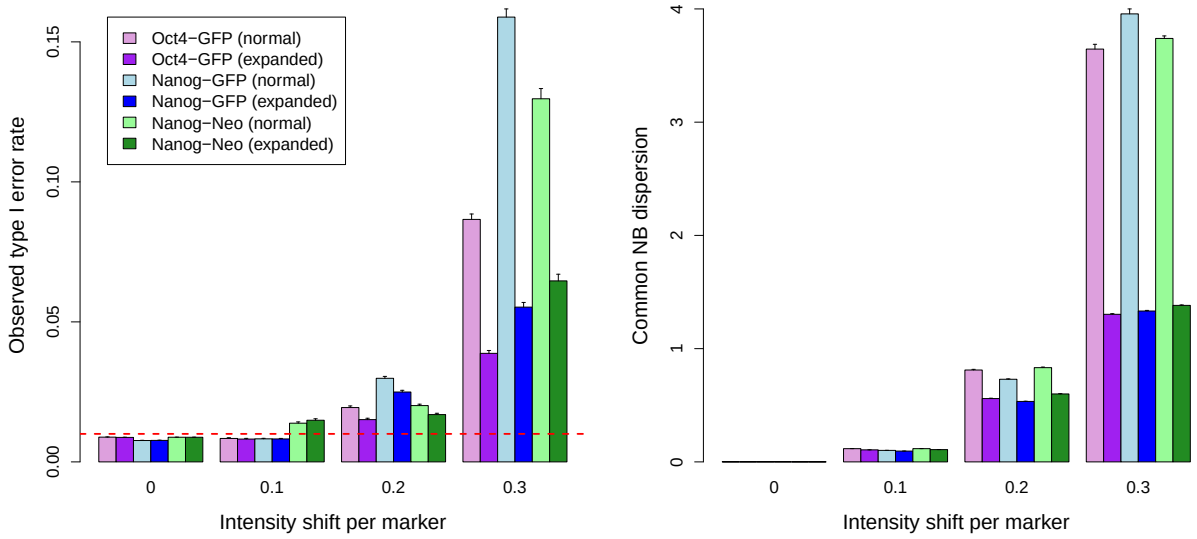
Supplementary Figure 3: Schematic of the simulation design used to examine the effect of increasing the hypersphere radius. Here, the experimental design consists of two samples from different conditions and 30 markers. Two subpopulations are present in the high-dimensional space, one of which is DA between samples (red) and the other is not (grey). Cells from each subpopulation (10000 each) are uniformly distributed within a 30-dimensional sphere with radius (referred to as "size", above) of length $0.5\sqrt{30}$, centred at the subpopulation mean (crosses). The means of the two subpopulations are separated by a distance of 0.5 in each dimension. Each hypersphere's position is calculated by taking the median in each dimension across all cells in the hypersphere. Note that the radii of the hyperspheres and subpopulations need not be identical.



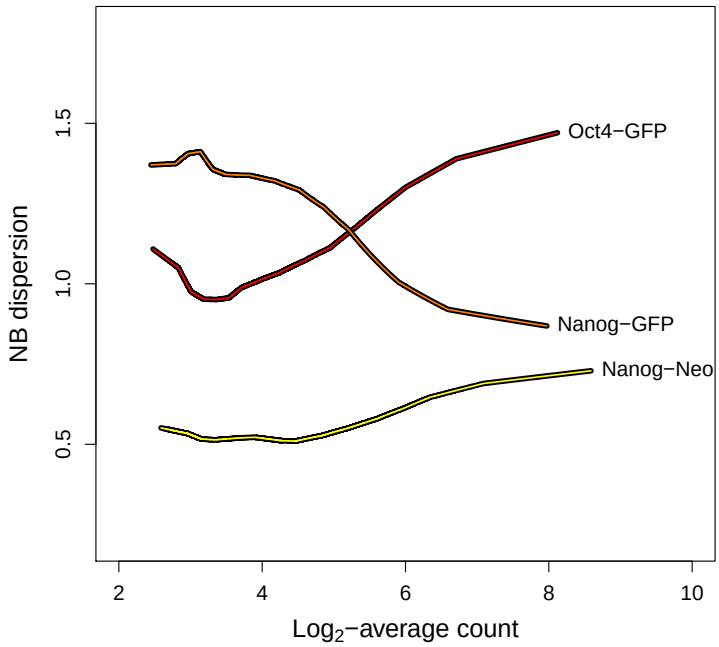
Supplementary Figure 4: Simulation results demonstrating the effect of expanding the radius on the median-based position of each hypersphere. For each hypersphere, the distance between its position and the mean of the DA subpopulation was computed and plotted against the log-fold change in cell counts between samples. Hyperspheres are coloured based on whether they were centred on cells from the DA (red) or non-DA (grey) subpopulations. The vertical dashed line marks the distance corresponding to the size of the DA subpopulation, while the horizontal dashed line represents a $\log_2$-fold change of 1. This process was repeated to create plots for a range of increasing hypersphere radii and for different simulation scenarios. Scenario 1 refers to the default scenario described in Supplementary Figure 3, while scenario 2 involves decreasing the number of cells in the DA subpopulation to 5000; scenario 3 increases the distance between subpopulations to 0.8 in all dimensions; and scenario 4 decreases the size of the DA subpopulation to $0.2\sqrt{30}$.



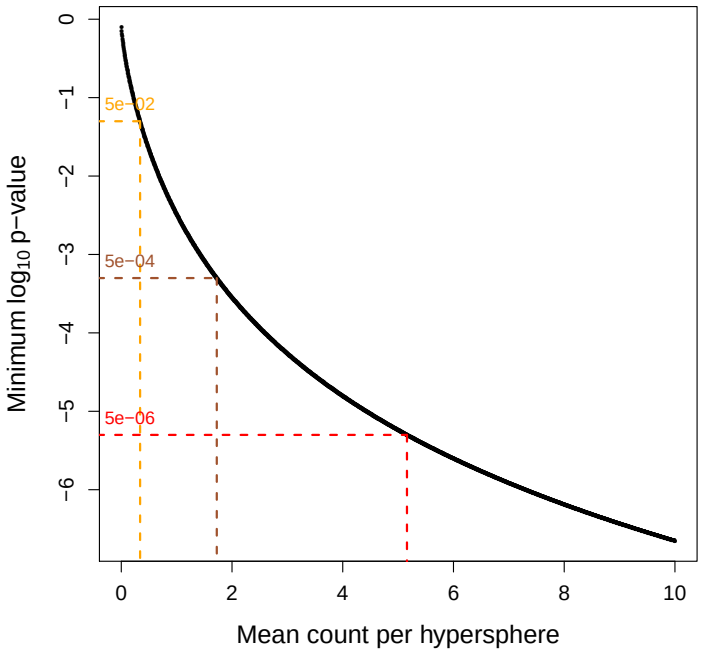
Supplementary Figure 5: Histograms of phosphorylated STAT3 intensities of BMMCs from five healthy, untreated individuals in the Levine *et al.* data set, before normalization (left), after range-based normalization (middle) and after warping normalization (right). Each curve represents a separate individual and is labelled with a different colour. Points represent the proportion of cells with zero intensities in each batch.



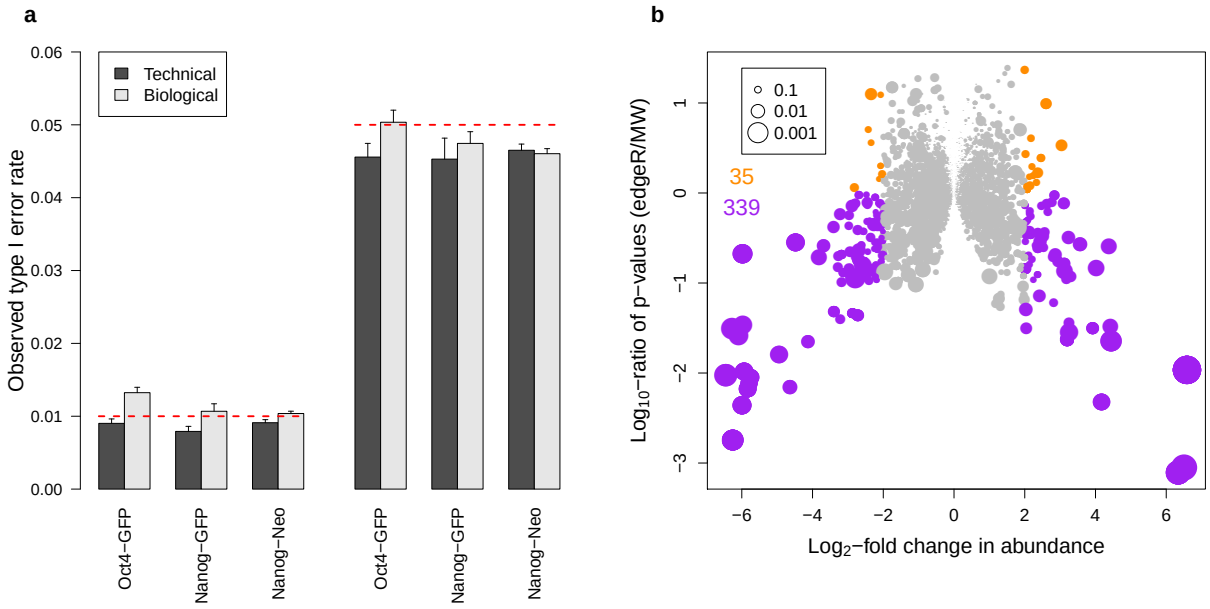
Supplementary Figure 6: Observed type I error rates in the simulation of intensity shifts, for DA analyses using either the default hypersphere radius ($r_0\sqrt{2} = 0.5$) or an expanded radius designed to accommodate the estimated shift. Simulations were based on the MEF reprogramming time courses, using values sampled from a Normal distribution with mean zero and standard deviation of 0 (no shift) to 0.3 (moderate shift) to shift marker intensities between samples. The observed type I error rate was calculated at a nominal threshold of 0.01 (left). The common dispersion estimate across all hyperspheres was also computed (right). All values represent the mean across 50 simulation iterations, with error bars representing the standard error.



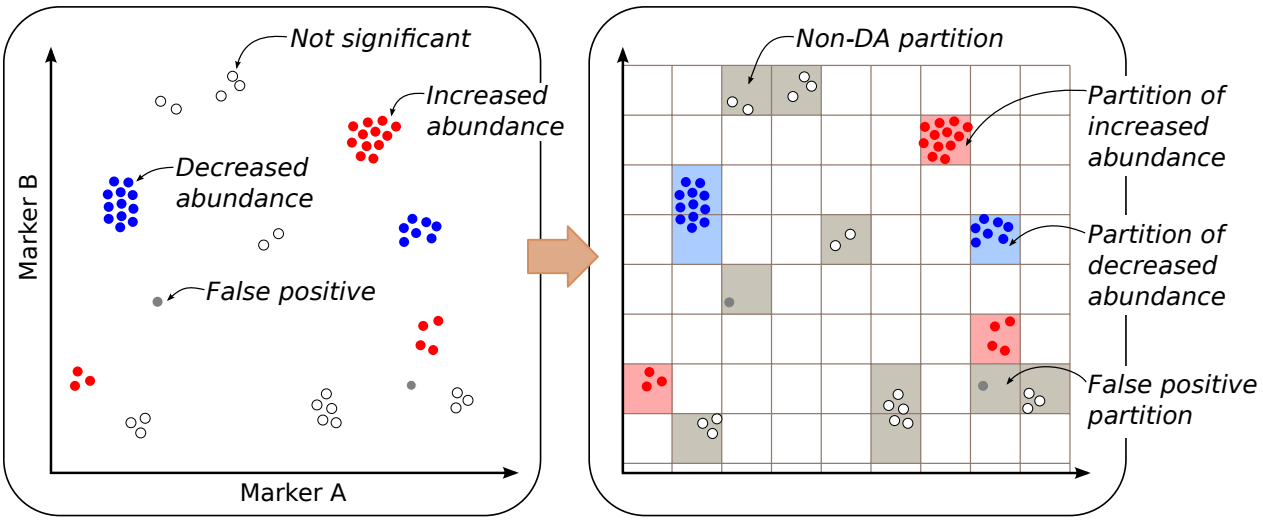
Supplementary Figure 7: NB dispersion estimates for all hyperspheres in each time course of the MEF reprogramming data set, plotted against the average number of cells across all samples in each hypersphere.



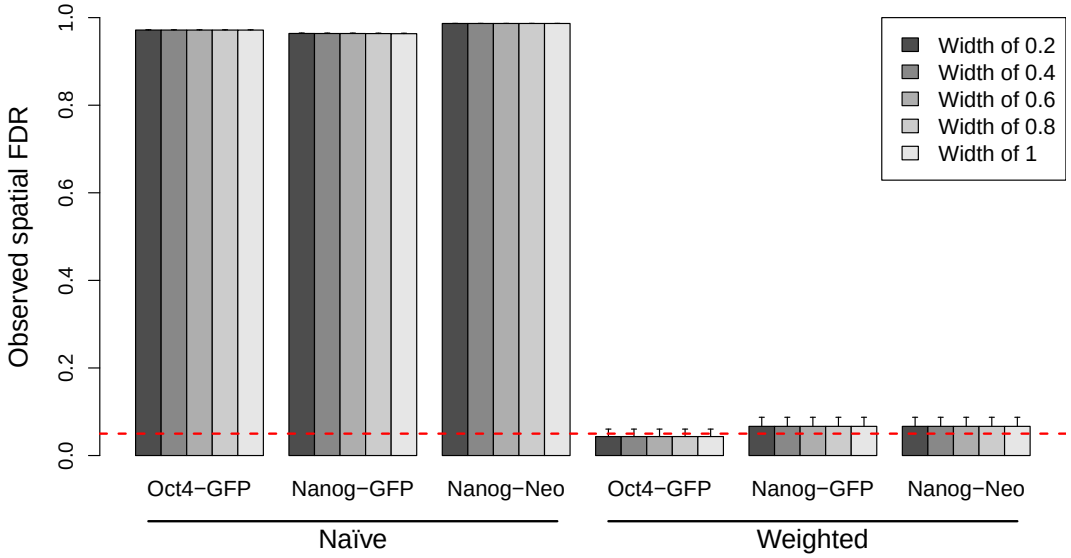
Supplementary Figure 8: Minimum p-value from edgeR as a function of the mean. This assumes an experimental design containing five replicates in each of two groups, where each replicate has the same total number of cells. edgeR was applied to test for significant differences between groups using a NB dispersion of 1. Dashed lines represent p-value thresholds and the smallest mean count required to achieve them.



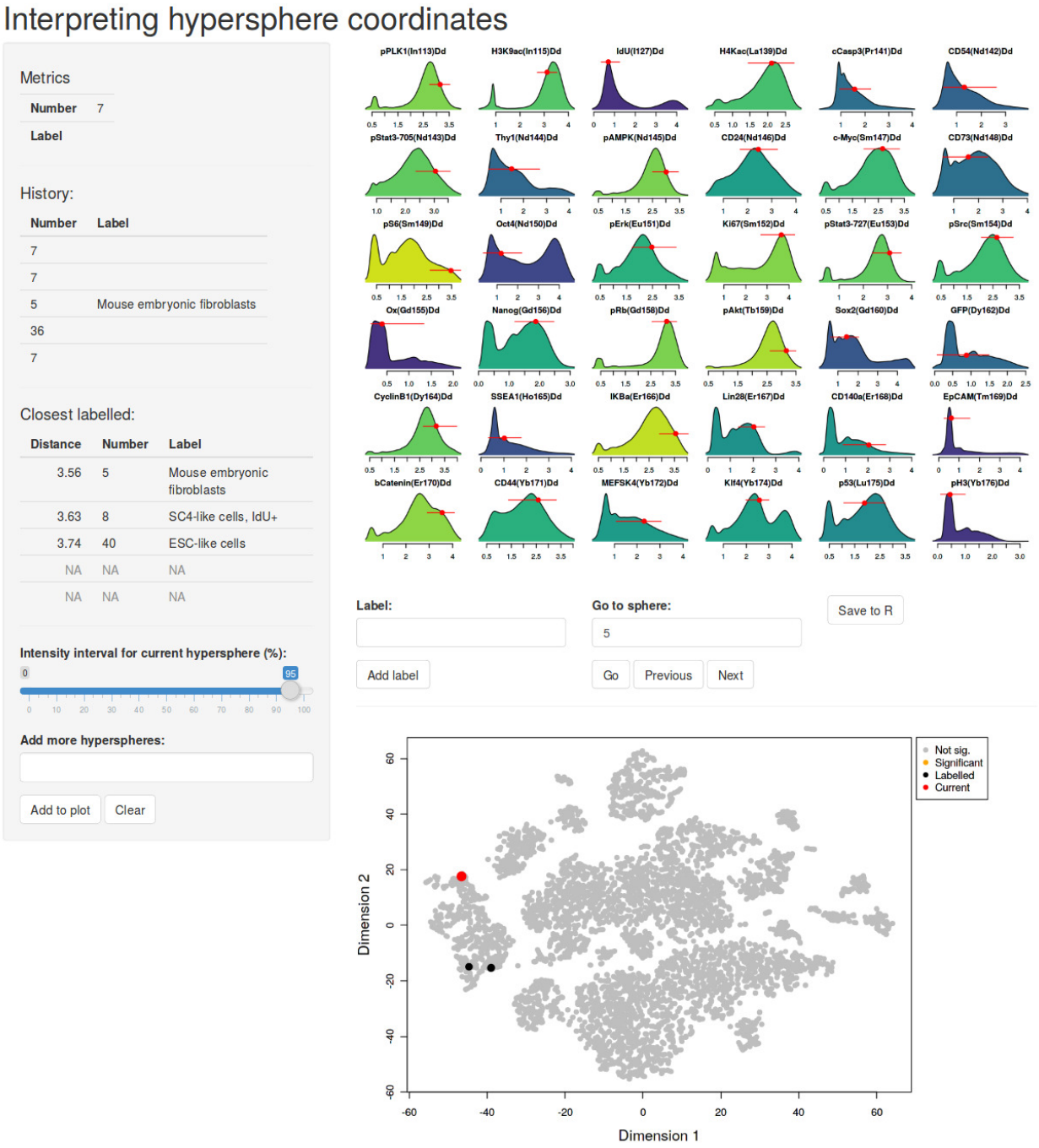
Supplementary Figure 9: Performance of edgeR for detecting DA hyperspheres, using simulations based on the MEF reprogramming time courses. (a) Observed type I error rates with edgeR for each simulated time course with technical or biological variability, constructed by simple or weighted sampling of cells respectively. Bar heights represent the mean of 10 simulation iterations, with error bars representing standard errors. The red lines denote the specified thresholds of 0.01 (left) or 0.05 (right). (b) Log_{10}-ratios of the p-values computed by edgeR against those from the Mann-Whitney (MW) test, plotted against $\log_2$-fold changes. Simulations were performed using weighted sampling for the *Oct4*-GFP time course, and each hypersphere was tested for differential abundance. Coloured points represent hyperspheres with absolute log-fold changes greater than two and p-values that are smaller in edgeR compared to the MW test (purple) or vice versa (orange). The number of such hyperspheres is also reported. The size of each point is determined by the edgeR p-value.



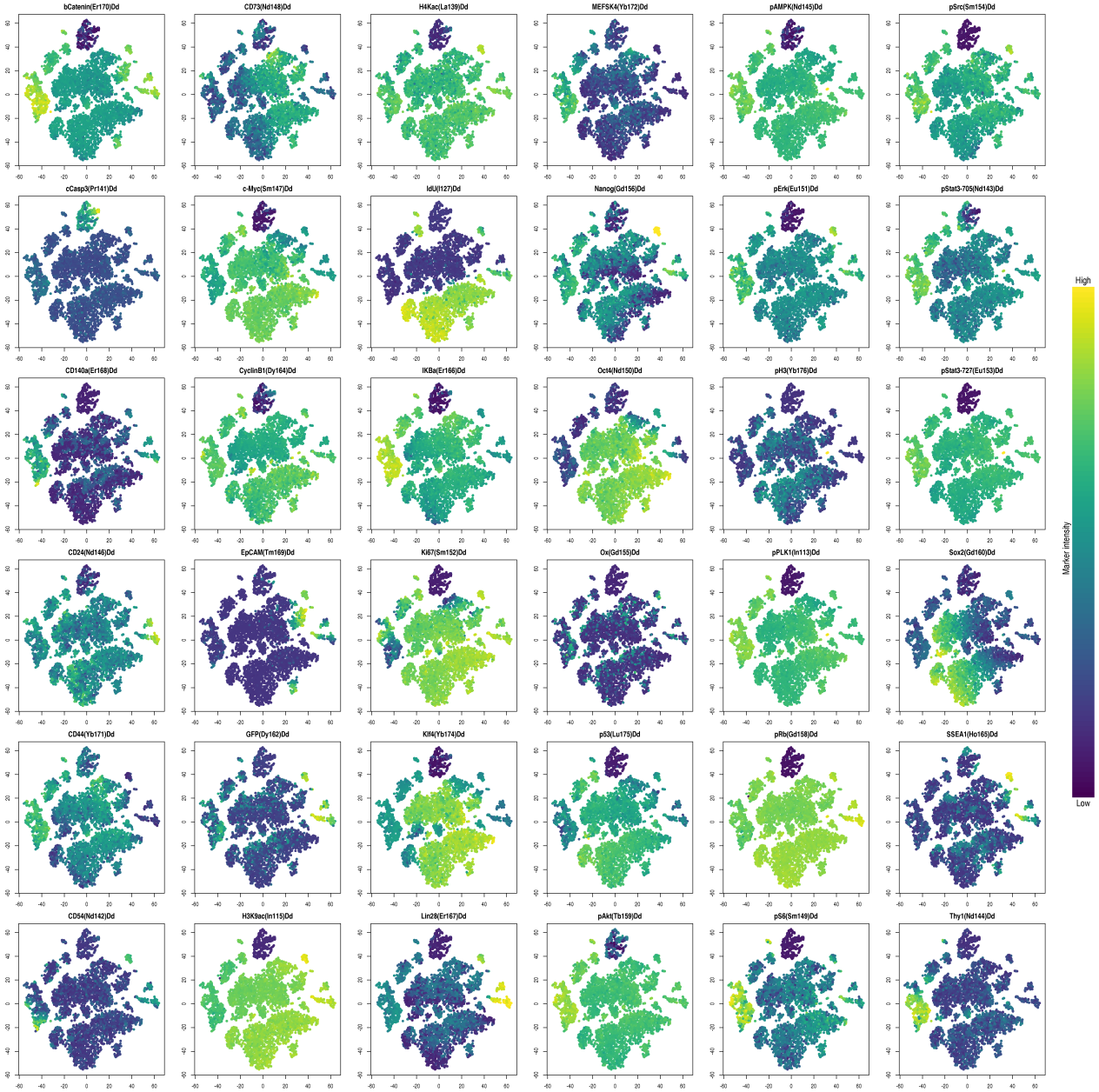
Supplementary Figure 10: Assignment of hyperspheres to hypothetical partitions. Each point represents the median-based position of a hypersphere in a data set with two markers, coloured based on its test outcome (red for significant increases in abundance between conditions, blue for significant decreases, white for no significant differences, and grey for false positives). Each hypersphere is assigned to a partition, and the test outcome of each partition is defined as that of its hyperspheres. For simplicity, all hyperspheres in a partition have the same outcome here. The aim is to control the spatial FDR across partitions, rather than the FDR across hyperspheres. In this example, the spatial FDR is 25% while the FDR across hyperspheres is 5%.



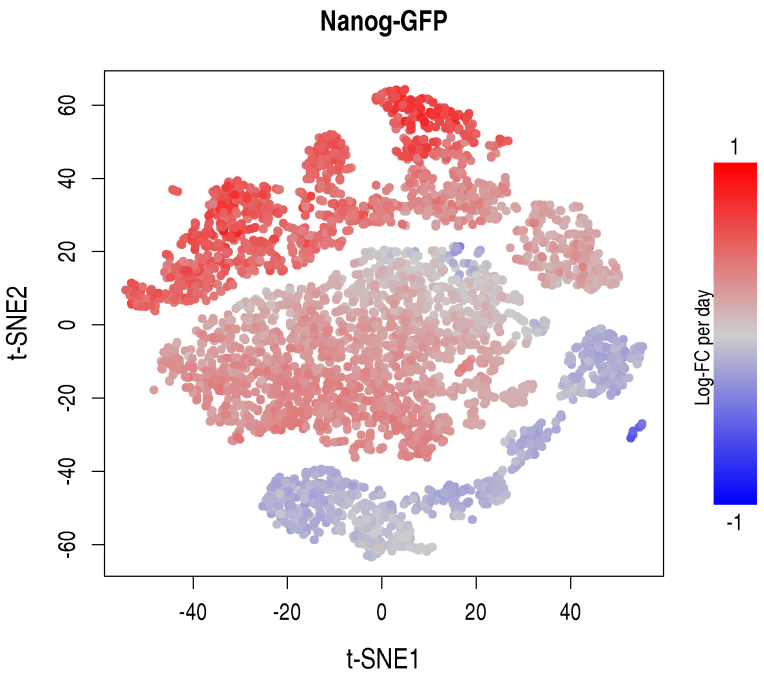
Supplementary Figure 11: Control of the spatial FDR using the naïve and weighted BH methods, in simulations based on the MEF reprogramming time courses. The observed spatial FDR was calculated in M-dimensional space using hypercubes of width 0.2 to 1. Bar heights represent the mean of 50 simulation iterations, and the error bars represent standard errors. The red line denotes the nominal threshold of 5%.



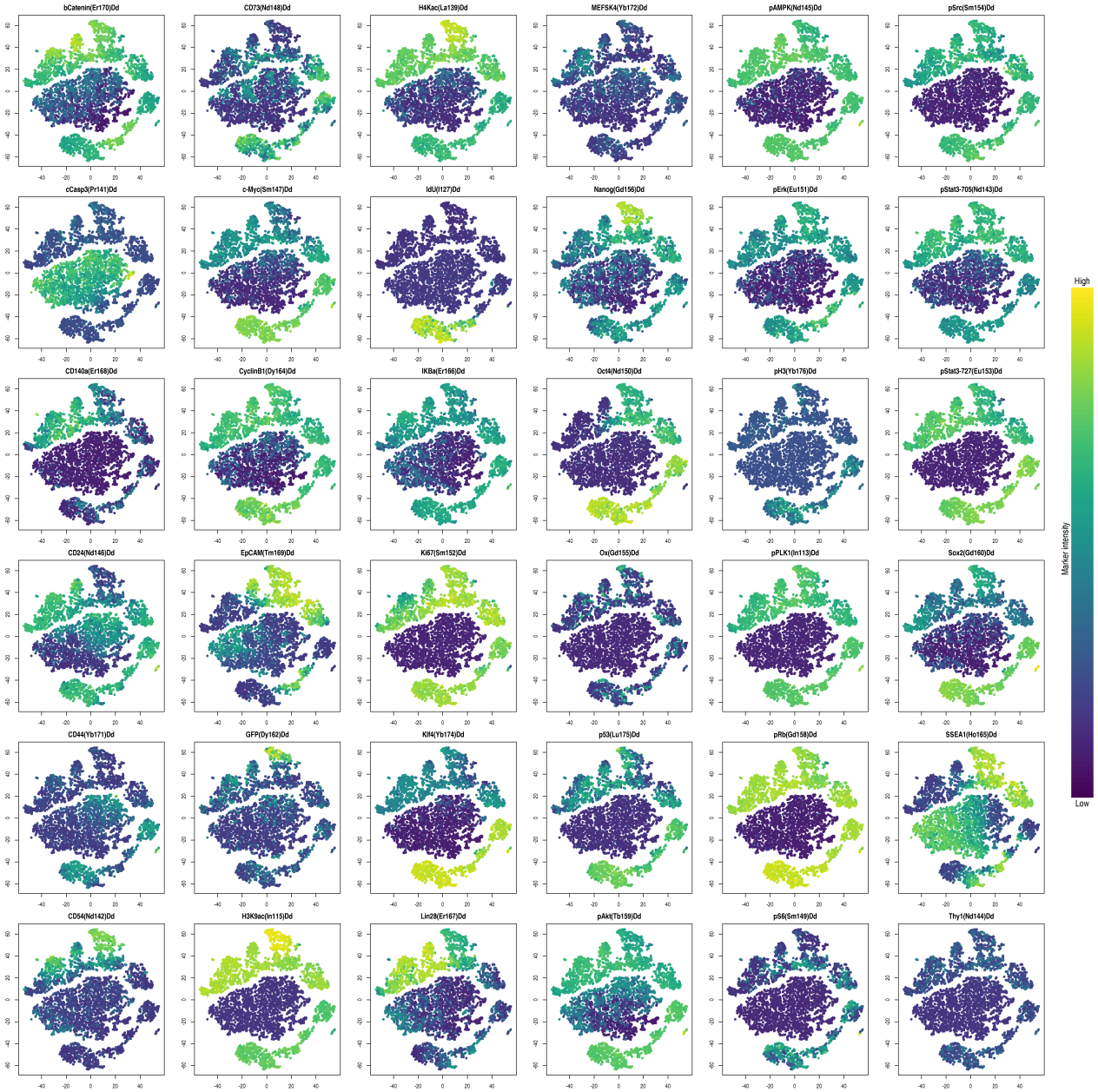
Supplementary Figure 12: A Shiny-based user interface for detailed interpretation of non-redundant hyperspheres. Each density curve represents the distribution of intensities across all cells for a particular marker. The red dot indicates the median intensity for the hypersphere being examined, while the interval contains 95% of all cells in the hypersphere. The area under the curve is coloured depending on whether the median is high (yellow) or low (purple) compared to the intensities for the majority of cells. Each hypersphere can be annotated based on its intensities (e.g., to describe the subpopulation that it represents), and these labels are stored in memory for future use. In addition, the closest non-redundant hyperspheres that have already been labelled are shown for each new hypersphere, to simplify further labelling of nearby hyperspheres in the high-dimensional space. Navigation through the set of hyperspheres is done based on the hypersphere number or by selecting locations in a dimensionality reduction plot (in this case, *t*-SNE is used).



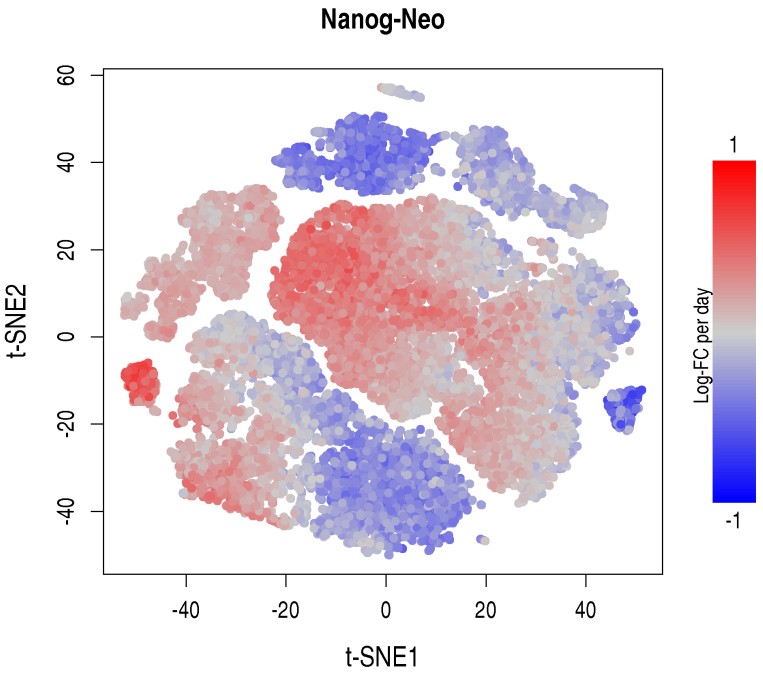
Supplementary Figure 13: Marker intensities of differentially abundant subpopulations in the *Oct4*-GFP time course, detected at a spatial FDR of 5%. Each plot corresponds to a marker while each point represents a hypersphere, coloured according to its median intensity for the corresponding marker. The boundaries of the colour range for each marker were set to the 1^{st} and 99^{th} percentiles of the intensities for all cells.



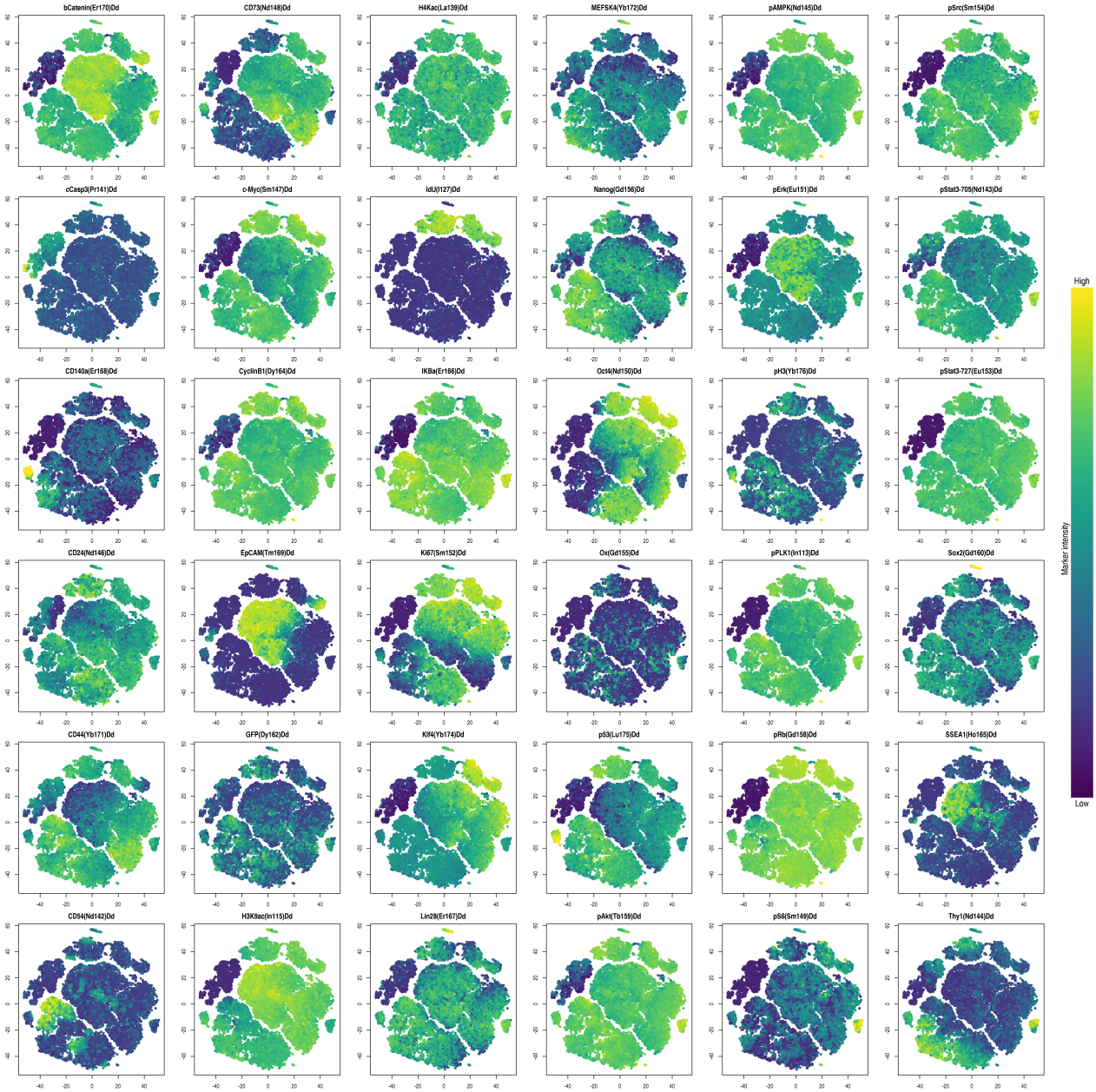
Supplementary Figure 14: Differentially abundant subpopulations in the *Nanog*-GFP time course, detected at a spatial FDR of 5%. Each point is a hypersphere and is coloured according to its log-fold change in abundance over time. Grey points are hyperspheres with significant but non-linear changes in abundance.



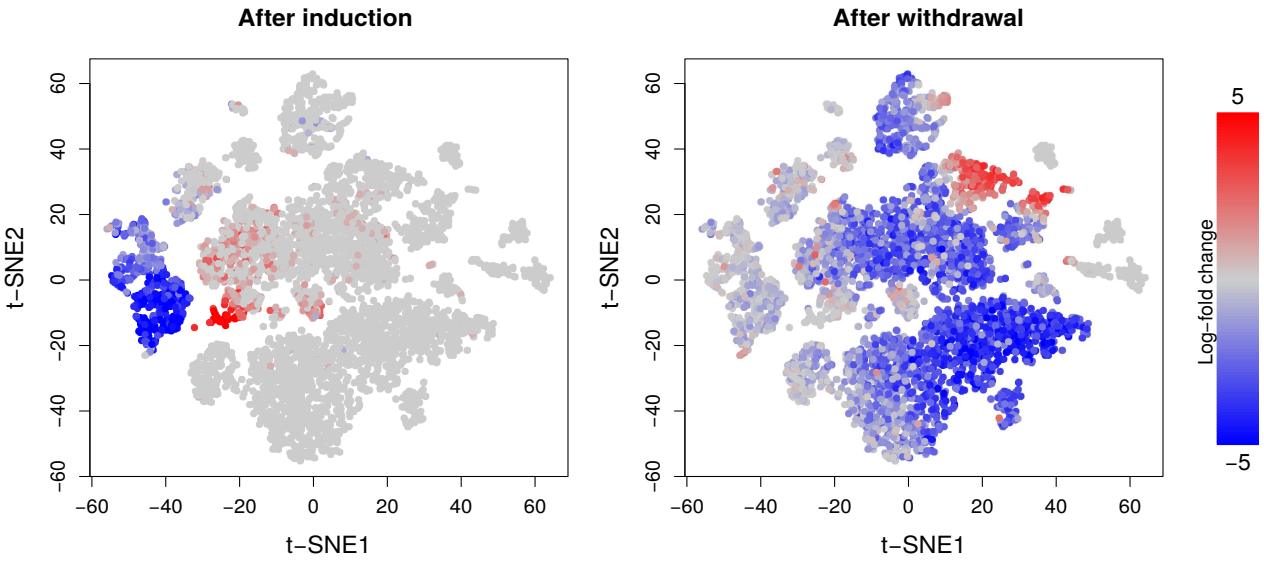
Supplementary Figure 15: Marker intensities of differentially abundant subpopulations in the *Nanog*-GFP time course, detected at a spatial FDR of 5%. Each plot corresponds to a marker while each point represents a hypersphere, coloured according to its median intensity for the corresponding marker.



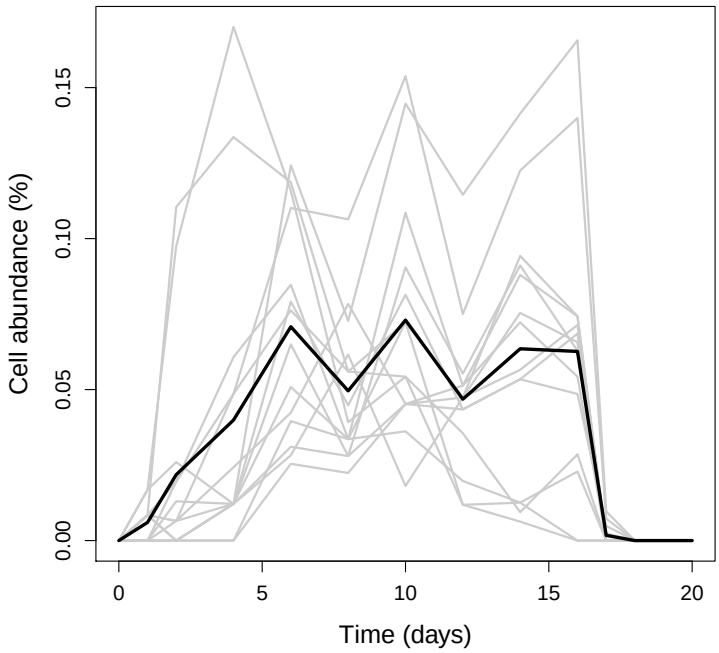
Supplementary Figure 16: Differentially abundant subpopulations in the *Nanog*-Neo time course, detected at a spatial FDR of 5%. Each point is a hypersphere and is coloured according to its log-fold change in abundance over time. Grey points are hyperspheres with significant but non-linear changes in abundance.



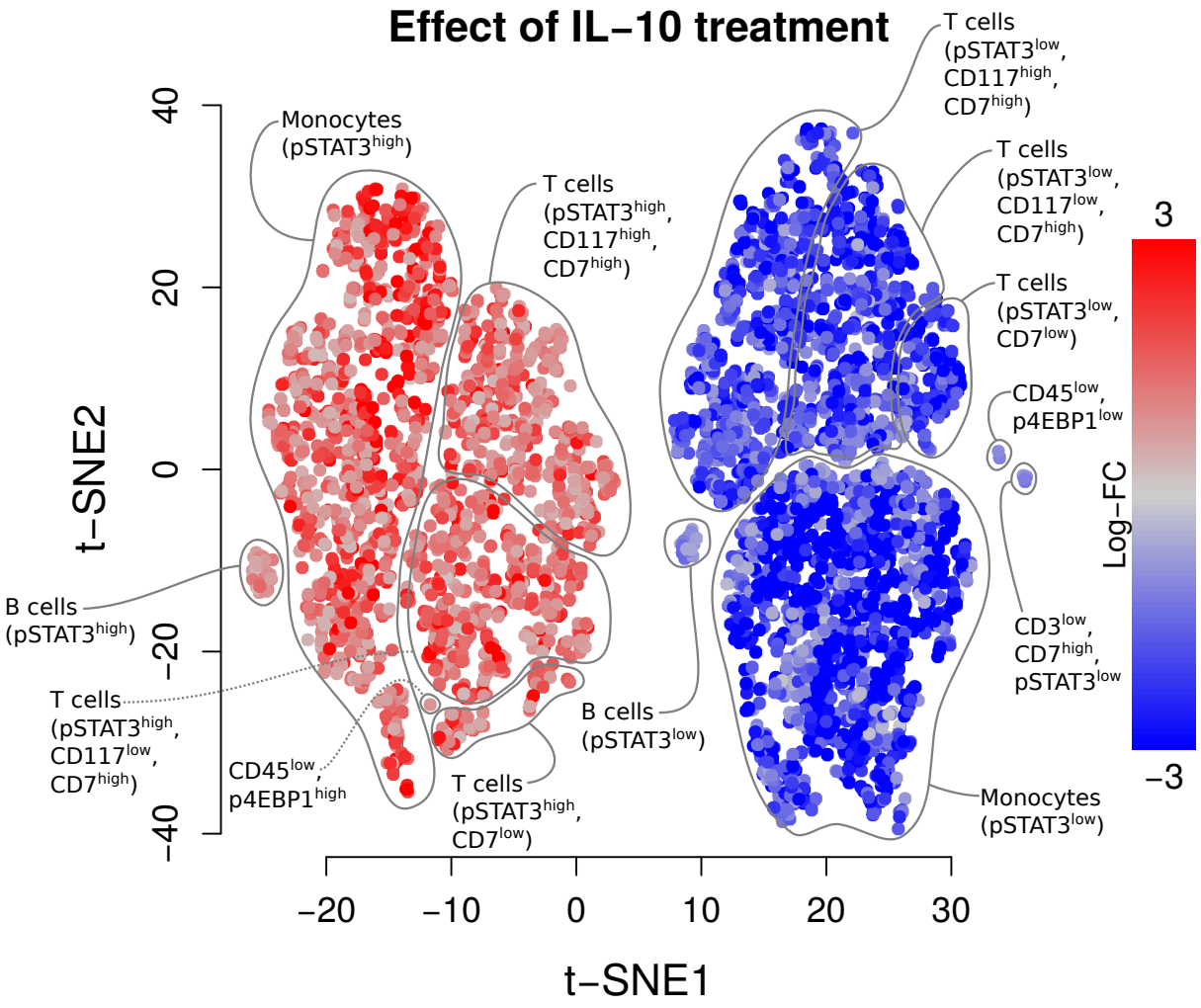
Supplementary Figure 17: Marker intensities of differentially abundant subpopulations in the *Nanog*-Neo time course, detected at a spatial FDR of 5%. Each plot corresponds to a marker while each point represents a hypersphere, coloured according to its median intensity for the corresponding marker.



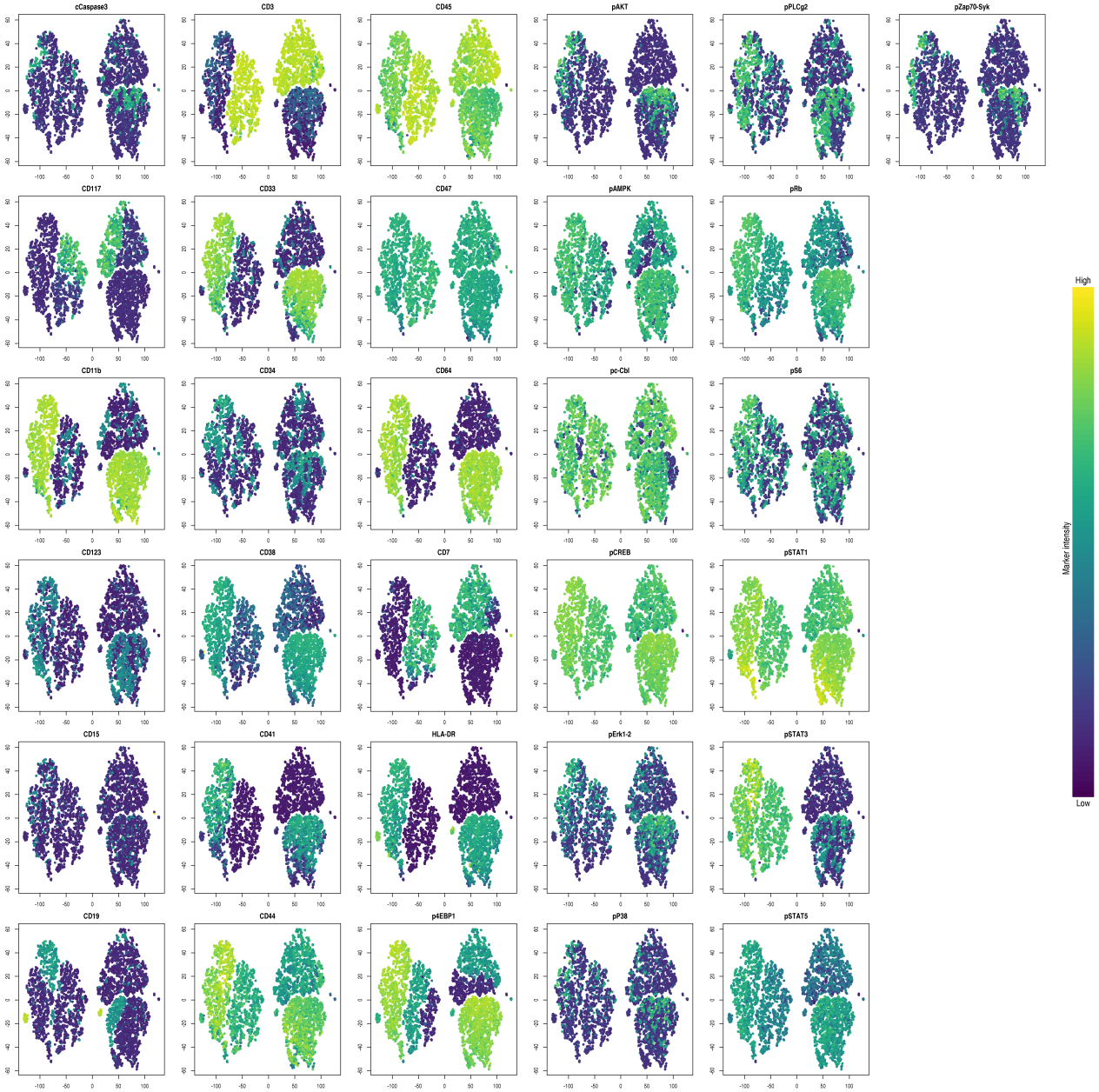
Supplementary Figure 18: Changes in cell abundance after transgene induction with doxycycline (left) or after doxycyline withdrawal (right) in the *Oct4*-GFP time course. Each point in the *t*-SNE plot represents a differential hypersphere, coloured based on the log-fold change during induction (day 0 to 1) or withdrawal (day 16 to 17). A prior of 3 was added to each count to stabilize the log-fold changes.



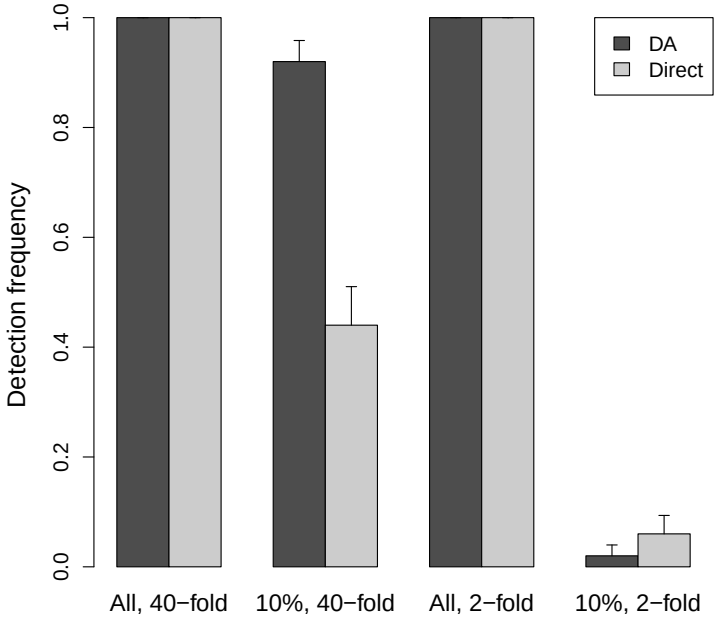
Supplementary Figure 19: Cell abundance for the novel SC4-like subpopulation with active STAT3, AMPK and PLK1 signalling, as a function of time in the *Oct4*-GFP reprogramming time course. Abundances are shown as a percentage of the total number of cells in each sample. Each line represents the abundance for each hypersphere representing the subpopulation, while the black line is the average across all hyperspheres.



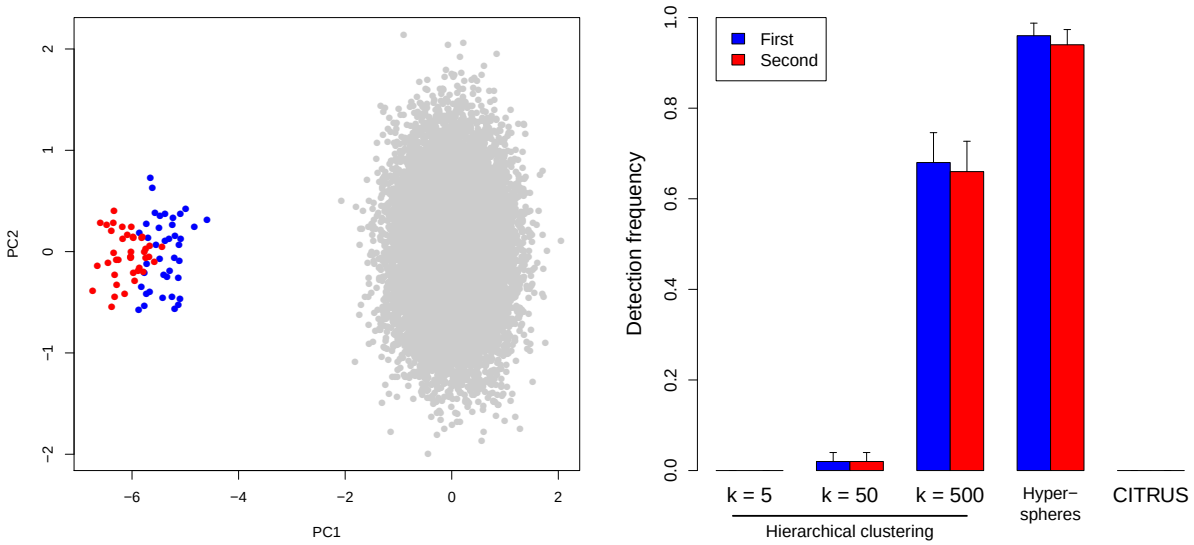
Supplementary Figure 20: Differentially abundant subpopulations in the BMMC data set upon treatment with IL-10, detected at a spatial FDR of 5% and visualized with a *t*-SNE plot. Each point represents a hypersphere and is coloured according to its average log-fold change in abundance upon treatment. Hyperspheres were annotated into subpopulations based on expression of lineage-specific markers and signalling molecules.



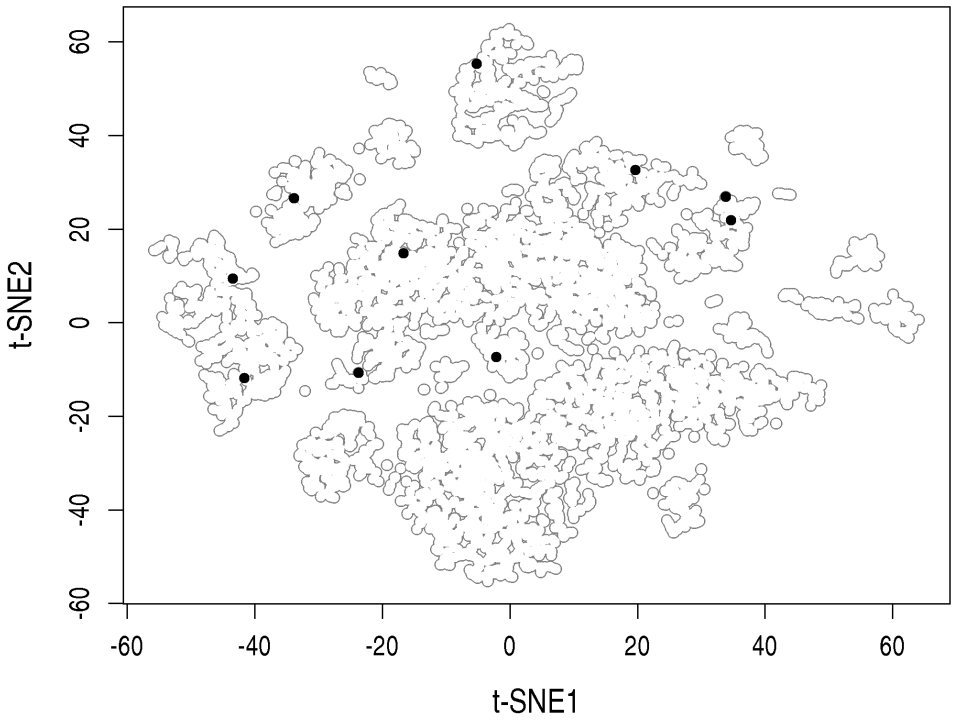
Supplementary Figure 21: Marker intensities of differentially abundant subpopulations upon IL-10 stimulation in the BMMC data set, detected at a spatial FDR of 5%. Each plot corresponds to a marker while each point represents a hypersphere, coloured according to its median intensity for the corresponding marker. The colour range for each marker was bounded at the 1^{st} and 99^{th} percentiles of intensities across all cells.



Supplementary Figure 22: Relative performance of direct detection and DA analyses on simulated data containing intensity shifts. Detection frequencies were defined as the number of simulation iterations (out of 50) in which the shifting was successfully detected, and error bars correspond to standard errors from a binomial distribution. Simulation scenarios were tested where all or 10% of cells shifted in intensity between conditions, with shifts ranging from 2- to 40-fold in size (in terms of the raw expression values).



Supplementary Figure 23: Relative performance of cluster-based DA analyses in simulated data. Left: a PCA plot to illustrate the simulation design, which contains a non-DA bulk of cells (grey) along with two small, adjacent DA subpopulations exclusive to the first (blue) or second conditions (red). Right: detection frequencies for each of the two DA subpopulations across 50 simulation iterations. Frequencies were defined as the proportion of iterations in which each subpopulation was detected, and error bars represent standard errors from a binomial distribution. DA analyses were performed using CITRUS, a custom approach based on hierarchical clustering with varying numbers of clusters k or the hypersphere-based approach.



Supplementary Figure 24: Detected clusters from a differential abundance analysis of the *Oct4*-GFP time course with CITRUS, mapped onto the *t*-SNE plot of DA hyperspheres. Each point represents the centre of a DA cluster that was detected at a FDR of 5%. DA hyperspheres are shown as a grey outline for comparison.