Kernel-Based Testing for Single-Cell Differential Analysis

Anthony Ozier-Lafontaine, Camille Fourneaux, Ghislain Durif, Céline Vallot, Olivier Gandrillon, Sandrine Giraud, Bertrand Michel, Franck Picard

Kernel-Based Testing for Single-Cell Differential Analysis

A. Ozier-Lafontaine*,[1], C. Fourneaux[2], G. Durif[2], C. Vallot[3,4], O. Gandrillon[2], S. Gonin-Giraud[2], B. Michel†*,[1], and F. Picard*†,[2]

[1]Nantes Université, Centrale Nantes, Laboratoire de mathématiques Jean Leray, F-44000, Nantes, France
[2]Laboratory of Biology and Modelling of the Cell, Université de Lyon, Ecole Normale Supérieure de Lyon, CNRS, UMR5239, Université Claude Bernard Lyon 1, Lyon, France
[3]CNRS UMR3244, Institut Curie, PSL University, Paris, France.
[4]Translational Research Department, Institut Curie, PSL University, Paris, France.

July 18, 2023

Abstract

Single-cell technologies have provided valuable insights into the distribution of molecular features, such as gene expression and epigenomic modifications. However, comparing these complex distributions in a controlled and powerful manner poses methodological challenges. Here we propose to benefit from the kernel-testing framework to compare the complex cell-wise distributions of molecular features in a non-linear manner based on their kernel embedding. Our framework not only allows for feature-wise analyses but also enables global comparisons of transcriptomes or epigenomes, considering their intricate dependencies. By using a classifier to discriminate cells based on the variability of their embedding, our method uncovers heterogeneities in cell populations that would otherwise go undetected. We show that kernel testing overcomes the limitations of differential analysis methods dedicated to single-cell. Kernel testing is applied to investigate the reversion process of differentiating cells, successfully identifying cells in transition between reversion and differentiation stages. Additionally, we analyze single-cell ChIP-Seq data and identify a subpopulation of untreated breast cancer cells that exhibit an epigenomic profile similar to persister cells.

1 Introduction

Thanks to the convergence of single-cell biology and massive parallel sequencing, it is now possible to create high dimensional molecular portraits of cell populations. This technological breakthrough allows for the measurement of gene expression [34, 25, 58], chromatin states [47], and genomic variations [15] at the single-cell resolution. These advances have resulted in the production of complex high dimensional data and revolutionized our understanding of the complexity of living tissues, both in normal and pathological states. Then, the field of single-cell data science has emerged, and new methodological challenges have arisen to fully exploit the potentialities of single-cell data, among which the statistical comparison of single-cell RNA sequencing (scRNA-Seq) datasets between conditions or tissues. This step has remained a prerequisite in the process to discriminate biological from technical variabilities and to assert meaningful expression differences. While most differential analysis methods primarily focus on expression data, similar methodological challenges have arisen in the comparative analysis of single cell epigenomic datasets, based for example on single cell chromatin accessibility assays (scATAC-Seq [42]) or single cell histone modifications profiling (e.g single-cell ChIP-Seq (scChIP-seq) [19], scCUT&Tag [4]). These approaches enable the mapping of chromatin states throughout the genome and their cell-to-cell variations at an unprecedented resolution [51, 6] . These single-cell epigenomic assays offer a quantitative perspective on regulatory processes, wherein cellular heterogeneity could drive cancer progression or the development of drug resistance for instance[36]. The identification of key epigenomic features by differential analysis in disease and complex eco-systems, will be key to understand regulatory principles of gene expression and identify potential drivers of tumor progression. Altogether, comparative analysis of single cell data sets,

*To whom correspondence should be addressed: anthony.ozier-lafontaine@ec-nantes.fr, and also Bertrand.Michel@ec-nantes.fr, franck.picard@ens-lyon.fr
†joint last authors

whatever their type, are an essential component of single cell data science, providing biological insights as well as opening therapeutic perspectives with the identification of biomarkers and therapeutic targets.

Differential Expression Analysis (DEA) is classically addressed by gene-wise two-sample tests designed to detect Differentially Expressed Genes (DEG) [11]. The generalized linear model (GLM) has been a powerful framework for linear parametric testing based on gene-expression summaries [32, 46, 45]. However, this parametric approach does not fully utilize the entire distribution of gene-expression that characterizes multiple transcriptional states. To achieve the full potential of differential analysis of scRNA-Seq data, DEA has been restated as a comparison between distributions. Distributional hypotheses were proposed to capture biologically relevant differences in univariate gene-expressions [29]. Initially, these tests were performed using Gaussian-based clustering, that was further challenged by distribution-free methods based on ranks or cumulative distribution functions [48, 14, 54]. While distribution-free approaches are flexible enough to capture the numerous complex alternatives encountered in DEA, their fully agnostic point of view does not benefit from the significant progress made in modeling scRNA-Seq distributions, which leads to a loss of statistical power. As a trade-off, we propose a distribution-free test based on a representation of the data that can take advantage of finely-tuned probabilistic modeling of scRNA-Seq data.

Single-cell technologies provide a unique opportunity to obtain a quantitative snapshot of the entire transcriptome, which contains crucial information about between-gene dependencies and underlying regulatory networks and pathways. Therefore, univariate DEA only captures a part of the biological differences and is unable to detect complex global modifications in the joint expression of groups of genes. To fully exploit the complexity of scRNA-Seq data, joint multivariate testing or differential transcriptome analysis should be performed, allowing for cell-wise comparisons. This strategy can be complementary to gene-wise approaches, as the detection of DEG should be interpreted in the context of global differences between conditions. The joint multivariate testing strategy seems also particularly suited to compare epigenomic data since it is well established that chromatin conformation can induce complex dependencies between sites occupancy [35]. From a distributional perspective, this involves complementing joint distribution-based analyses with analyses based on marginals. Another significant advantage of differential transcriptome analysis is that it can be restricted to targeted GRNs or pathways, allowing for differential network or pathway analyses [41]. So far, global approaches were mainly developed for differential abundance testing [9, 10, 7], or for the comparison of cell-type compositions. Graph-based methods have been proposed to address differential transcriptome analysis [41, 3], but they only derive a global p-value without any representation or diagnostic tool.

In recent years, there have been significant advancements in statistical hypothesis testing, alongside the emergence of single-cell technologies. One important breakthrough in hypothesis testing was achieved by Gretton et al. [16], who combined kernel methods with statistical testing. Kernel methods are widely used in supervised learning [50] and are based on the concept of embedding data in a feature space, allowing for non-linear data analysis in the input space. Popular dimension reduction techniques, such as tSNE and UMAP [33, 37], also use kernel-based embedding [55]. The distribution of the embedded data can be described using classical statistics such as means and variances, which can be applied in the feature space. Then the central concept of kernel-based testing is to rely on the Maximum Mean Discrepancy (MMD) test that compares the distance between mean embeddings of two conditions [40], allowing for non-linear comparison of two gene-expression distributions. Despite the significant potential of kernel-based testing, this approach has not yet been developed in single-cell data science.

In this work, we propose a new kernel-based framework for the exploration and comparison of single-cell data based on Differential Transcriptome/Epigenome Analysis. Our method relies on a normalized version of the Maximum Mean Discrepancy to account for the variability of the datasets. This results in a test statistic that can be interpreted as the distance between mean embeddings projected onto the kernel-Fisher discriminant axis (KFDA,[24]). Although KFDA was initially introduced as a non-linear classifier [39], it is a great example of how classifiers can be used for hypothesis testing [23, 31], and recent developments have demonstrated its optimality [22]. Here we show that the KFDA-witness function, which is the Fisher discriminant axis [30], can further be used for data exploration of scRNA-Seq and scChIP-seq data. Our method is available in a package called `ktest` [1] available in both R and Python, which offers many visualization tools based on the geometrical concepts from the Fisher Discriminant Analysis (FDA) to aid comparisons. We show the calibration and the power of our method compared with others on simulated [14] and multiple scRNA-Seq datasets [53]. Then we illustrate the power of the classification-based testing approach, that identifies sub-populations of cells based on expression and epigenomic data, that would not be detected otherwise. When applied to scRNA-Seq data, `ktest` reveals the heterogeneity in differentiating cell populations induced to revert toward an undifferentiated phenotype [59]. Our method also uncovers the epigenomic heterogeneity of breast cancer cells, revealing the

[1] `https://github.com/AnthoOzier/ktest`

pre-existence - prior to cancer treatment - of cells epigenomically identical to drug-persister cells, i.e the rare cells that can survive treatment.

As single-cell datasets grow larger and more complex, traditional testing methods may fail to capture subtle variations and accurately identify meaningful differences in molecular patterns. Here we show that kernel testing emerges as a promising approach to overcome these challenges, offering a robust and flexible framework. Kernel testing techniques are less sensitive to assumptions on data distribution than traditional methods, and can handle complex dependencies within and across cells. This capability is particularly relevant in the context of single-cell data, where inherent noise, sparsity, and heterogeneity pose unique challenges to accurate statistical inference. Overall, kernel testing represents a powerful tool for the differential analysis of single-cell data, enabling to uncover hidden patterns, and gain deeper insights into the intricate heterogeneities of cell populations.

2 Results

In the following we denote by $Y_1 = (Y_{1,1}, \dots, Y_{1,n_1})$ and $Y_2 = (Y_{2,1}, \dots, Y_{2,n_2})$ the gene expression measurements of G genes with distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ in conditions 1 and 2 on n_1 and n_2 cells respectively, $n = n_1 + n_2$. In the following, we will derive our method for expression data, but it can be generalized to any single-cell data. Then we suppose that

$$Y_{i,j} \sim \mathbb{P}_i, \quad i = 1, 2 \quad j = 1, \dots n_i.$$

Two-sample testing between distributions consists in challenging the null hypothesis $H_0 : \mathbb{P}_1 = \mathbb{P}_2$ by the alternative hypothesis $H_1 : \mathbb{P}_1 \neq \mathbb{P}_2$. To construct a non-linear test we consider the embeddings of the original data denoted by $\big(\phi(Y_{i,1}), \dots, \phi(Y_{i,n_i})\big)$ $(i = 1, 2)$, obtained using the feature map ϕ that maps the data into the so-called feature space $\mathcal{H}$ that is a reproducing kernel Hilbert space. The kernel provides a measure of the similarity between the observations, that turns out to be the inner product between the embeddings :

$$k(Y_{i,j}, Y_{i',j'}) = \langle \phi(Y_{i,j}), \phi(Y_{i',j'}) \rangle_{\mathcal{H}}.$$

Thanks to this relation, kernel methods are non-linear for the original data, but linear with respect to the embeddings in the feature space. They provide a non-linear dissimilarity between cells based either on the whole transcriptome or on univariate gene distributions. Kernel-based tests consist in the comparison of kernel mean embeddings of distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ [40], defined such that:

$$\forall i \in \{1, 2\}, \quad \mu_i = \mathbb{E}_{Y \sim \mathbb{P}_i} \left[\phi(Y)\right].$$

The initial contribution to kernel testing involved calculating the distance between kernel mean embeddings with the MMD statistic [17]. However, it is difficult to determine its null distribution, and since the MMD does not account for the variance of embedding, it has recently been show to lack optimality [22]. By utilizing a Mahalanobis distance to standardize the difference between mean embeddings, we can not only obtain an asymptotic chi-square distribution for the resulting statistic [23], but we can also take advantage of the kernel Fisher Discriminant Analysis (KFDA) framework that is typically used for non-linear classification. Therefore, we present two complementary perspectives on the KFDA testing framework: one based on a distance-based construction of the statistic and the other on the kernel FDA, which offers several visualization tools to highlight the main cell-wise differences between the two tested conditions.

2.1 Testing with a Mahalanobis distance between gene-expression embeddings

The squared distance between the kernel mean embeddings constitutes the so-called Maximum Mean Discrepancy statistic, such that:

$$\begin{aligned} \mathrm{MMD}^2(\mu_1, \mu_2) &= \|\mu_1 - \mu_2\|^2_{\mathcal{H}} \\ &= \mathbb{E}_{Y_1 \sim \mathbb{P}_1, Y_1' \sim \mathbb{P}_1} \left[k(Y_1, Y_1')\right] + \mathbb{E}_{Y_2 \sim \mathbb{P}_2, Y_2' \sim \mathbb{P}_2} \left[k(Y_2, Y_2')\right] \\ &- 2 \times \mathbb{E}_{Y_1 \sim \mathbb{P}_1, Y_2 \sim \mathbb{P}_2} \left[k(Y_1, Y_2)\right]. \end{aligned}$$

This statistic tests the between-class separation by comparing expected pairwise similarities between and within conditions 1 and 2. To account for the residual variability, we introduce the weighted Mahalanobis distance between mean embeddings,

$$\mathrm{D}^2(\mu_1, \mu_2) = \frac{n_1 n_2}{n} \|\Sigma_W^{-1/2} \left(\mu_1 - \mu_2\right)\|^2_{\mathcal{H}},$$

where Σ_W is the homogeneous within-group covariance of embeddings

$$\Sigma_W = \frac{n_1}{n}\Sigma_1 + \frac{n_2}{n}\Sigma_2,$$

with

$$\forall i \in \{1,2\}, \quad \Sigma_i = \mathbb{E}_{Y\sim\mathbb{P}_i}\left[(\phi(Y) - \mu_i)^{\otimes 2}\right],$$

the covariance operator within each condition ($\otimes$ stands for the outer product in the feature space). To avoid the singularity of Σ_W, we consider a regularized version of the kernel-based Mahalanobis distance, by approximating the within-covariance by its first T principal directions. This resumes to a kernel-PCA dimension-reduction step based on $\Sigma_{W,T}$ which catches the residual variability of expression data centered by condition. Then the corresponding regularized statistic is based on the estimated mean embeddings and covariances:

$$\forall i \in \{1,2\}, \quad \widehat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\phi(Y_{i,j}), \quad \widehat{\Sigma}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}(\phi(Y_{i,j}) - \widehat{\mu}_i))^{\otimes 2}.$$

The main computational complexity comes from the eigen-decomposition of $\widehat{\Sigma}_W = (n_1\widehat{\Sigma}_1 + n_2\widehat{\Sigma}_2)/n$ which requires $O(n^3)$ operations and results in the truncated covariance $\widehat{\Sigma}_{W,T} = \sum_{t=1}^{T}\widehat{\lambda}_t(\widehat{e}_t \otimes \widehat{e}_t)$, where $(\widehat{\lambda}_t)_{t=1:T}$ are the decreasing eigenvalues of $\widehat{\Sigma}_{W,T}$ and $(\widehat{e}_t)_{t=1:T}$ are the associated eigenfunctions referred by extension in the following as principal components. Then the empirical weighted Mahalanobis distance between the two mean-embeddings is :

$$\widehat{\mathrm{D}}_T^2(\widehat{\mu}_1, \widehat{\mu}_2) = \frac{n_1 n_2}{n}\left\|\widehat{\Sigma}_{W,T}^{-\frac{1}{2}}(\widehat{\mu}_2 - \widehat{\mu}_1)\right\|_{\mathcal{H}}^2.$$

This statistic follows a $\chi^2(T)$ asymptotically under the null hypothesis [24], which resumes to the Hotelling's test in the feature space. Using the asymptotic distribution for testing seems reasonable for scRNA-Seq data for which $n \geq 100$, otherwise, it is possible to test with a permutation procedure for small sample sizes. Our implementation runs in ~ 5 minutes for $n \sim 4000$, and the package proposes a sampling-based Nystrom approximation for larger sample sizes [57].

2.2 The kernel Fisher Discriminant analysis, a powerful tool for non-linear DEA

A major advantage of using the Mahalanobis distance between distributions is that the test statistic can be reinterpreted under the light of a classification problem, thanks to its connection with the Fisher Discriminant Analysis (FDA). This framework induces a powerful cell-wise visualization tool that allows to explore and understand the nature of the differences between transcriptomes. FDA is a linear classification method that consists in finding the linear axis that optimizes the discrimination between the two distributions. Intuitively, a direction is discriminant if the observations projected on it (i) do not overlap and (ii) are far from each other. Hence the best discriminant axis is found by maximizing the Fisher Discriminant Ratio, that models a trade-off between minimizing the overlap while maximizing the distance between the means of the two groups. By finding this linear axis in the feature space to classify the embeddings, we obtain a non-linear function that makes the two distributions linearly separable. Thus, in the feature space we denote by $h_T^\star$ the optimal axis that maximizes the truncated Fisher Discriminant Ratio :

$$h_T^\star = n \operatorname*{argmax}_{h\in\mathcal{H}} \frac{\langle h, \Sigma_B h\rangle_{\mathcal{H}}}{\langle h, \Sigma_{W,T} h\rangle_{\mathcal{H}}}.$$

where Σ_B is the between-group covariance capturing the part of the variance of the embeddings due to the difference between the two groups :

$$\Sigma_B = \frac{n_1 n_2}{n^2}(\mu_1 - \mu_2)^{\otimes 2}.$$

The numerator of the Fisher Discriminant Ratio captures the distance between the two mean embeddings on a given direction, to be maximized, and the denominator captures the variability of the embeddings projected on this direction, standing for a measure of the overlap, to be minimized. The discriminant axis $h_T^\star$ can be found in closed form from an analytical reasoning. The Mahalanobis distance then appears to be the maximal value of the ratio, which is the distance between the mean embeddings projected on $h_T^\star$:

$$\mathrm{D}_T^2 = n\frac{\langle h_T^\star, \Sigma_B h_T^\star\rangle_{\mathcal{H}}}{\langle h_T^\star, \Sigma_{W,T} h_T^\star\rangle_{\mathcal{H}}} = \frac{n_1 n_2}{n}\|\Sigma_{W,T}^{-1/2}(\mu_1 - \mu_2)\|_{\mathcal{H}}^2,$$

By relying on both the within-group and the between-group covariances, the FDA approach encompasses the total variability of the embeddings. We can interpret the projection of the embeddings on $h_T^\star$ in terms of similarity between the two groups. The extreme values of projected embeddings on the discriminant axis correspond to cells that contain the most significant information for distinguishing between conditions. Conversely, the central values of projected embeddings correspond to cells that do not contribute to the discrimination and hold less informative value. We will propose an illustration to show how this representation can be used to identify outliers or sub-populations.

Then non-linear testing turns out to be very powerful to detect complex alternatives, like the ones proposed in the context of distribution-based DEA [29]. We illustrate the discriminant axis by representing the four standard alternative hypotheses: differential mean (DE), differential proportions (DP), differential modality (DM) and differential both proportion and modality (DB) [29]. The DE, DP and DM alternatives are somehow easy to discriminate even with summary statistics because the distributions have different means, projecting the embeddings on the discriminant axis easily discriminates the two conditions. On the contrary, the DB alternative is the most difficult alternative to detect with many DEA approaches, because the two conditions share the same mean expression [14]. The discriminant axis acts as a powerful non-linear transformation of the expression data to make the two distributions easily separable (Fig. 1).

2.3 Kernel Choice

The design of appropriate kernels is an active field of research [2, 49]. In kernel-based testing, choosing an appropriate kernel has many objectives like capturing important data characteristics and showing sufficient power to distinguish between different alternatives. To this extent, the conclusions drawn in the feature space from the mean embeddings should apply to the initial distributions. In other words, it should be equivalent to test $\mu_1 = \mu_2$ for $\mathbb{P}_1 = \mathbb{P}_2$ which is not true in general. However, both are equivalent for a particular class of kernels called universal kernels, which has lead to theoretical and computational developments [52, 18, 49]. Fortunately, the Gaussian kernel fulfills this universality property. For two cells $\{(i,j),(i',j')\}$ and genes $g = 1, \ldots, G$, our developments will be based on k_{Gauss} defined such that :

$$k_{\text{Gauss}}(Y_{i,j}, Y_{i',j'}) = \exp\left(-\frac{1}{2\sigma^2}\sum_{g=1}^{G}(Y_{i,j}^g - Y_{i',j'}^g)^2\right).$$

This kernel can be used in both multivariate and univariate contexts. Once the Gaussian kernel has been chosen, the remaining question concerns the calibration of its bandwidth σ, which is done using the median heuristic [18, 49, 13]. We also propose to account for zero-inflation which is another important characteristic of scRNA-Seq data. This can be achieved by employing Fisher kernels, which adapt kernel methods to specific probabilistic generative models [26]. Considering a zero-inflated Gaussian distribution with π the proportion of additional zeros and $f_{\mu,\sigma}$ the Gaussian probability function, we show that the probability kernel between two zero-inflated Gaussian distributions of parameters (μ, σ, π) and (μ', σ, π') is (as detailed in the Methods Section):

$$\begin{aligned} k_{\text{ZI-Gauss}}(Y_{i,j}, Y_{i',j'}) &= \pi\pi' + \pi(1-\pi')f_{\mu',\sigma}(0) + (1-\pi)\pi' f_{\mu,\sigma}(0) \\ &+ (1-\pi)(1-\pi')k_{\text{Gauss}}(Y_{i,j}, Y_{i',j'}). \end{aligned}$$

2.4 Kernel testing is calibrated and powerful on simulated data

Simulations are required to compare the empirical performance of DE methods on controlled designs, to check their type-I error control and compare their power on targeted alternatives. Thanks to a very fruitful collaboration/data sharing with colleagues having developed a competing method [14], we challenged our kernel-based test with others on mixtures of zero-inflated negative binomial data reproducing the four Korthauer's alternatives [14] (as detailed in Material and Methods). Kernel testing was performed on the raw data using both the Gauss and the ZI-Gauss kernels. The type-I errors of the kernel test are controlled at the nominal levels $\alpha = 5\%$ and the performance increases with n (the asymptotic regime of the test is reached for $n \geq 100$). The kernel test is the best method for detecting the DB alternative, considered as the most difficult to detect, and it outperforms every other method in terms of global power excepted SigEMD. This gain in power can be explained by the non-linear nature of our method: despite the equality of means, the kernel-based transform of the data onto the discriminant axis allows a clear separation between distributions (Fig. 1). Our method shows its worst performances on the DP alternative, which is the only alternative for which all the values are covered by both conditions with different proportions. It shows that our method is particularly sensitive to alternatives where some values are occupied by one condition only (Fig. 2). Note that the Fisher kernel did not improve the

global performance, which indicates that the Gaussian kernel-based test is robust to zero inflation. This could also be due to the equality of the zero-inflation proportions between conditions. Finally, results on log-normalized data are similar.

2.5 Challenging DEA methods on experimental scRNA-Seq data

Differential analysis methods require validation through experimental data, typically by using a ground truth list of differentially expressed (DE) genes and an accuracy criterion. In this study, we examine the framework proposed by Squair et al. [53], which compared 14 DE analysis methods on 18 scRNA-Seq datasets. The authors proposed three main conclusions: *i*) replicate variability needs to be corrected, *ii*) single-cell DE methods are susceptible to false discoveries, and *iii*) pseudo-bulk methods are the most powerful. Pseudo-bulk methods involve applying DEA methods dedicated to bulk RNASeq to averaged scRNA-Seq. However, these conclusions are based on the use of bulk RNA sequencing DE genes as the ground truth, which inevitably favors pseudo-bulk methods designed to detect significant mean differences only. Hence, the study ignores genes with differential expression based on other characteristics, as shown in Korthauer's DB scenario [29]. Therefore, we propose to broaden the scope of this comparative study by comparing the outputs of different DE methods in a pairwise comparative manner, without relying on a reference ground truth list of DE genes. Based on pair-wise accuracies, Differential Analysis methods cluster into three groups of concordant groups that correspond to bulk, pseudo-bulk and single-cell based methods respectively (Fig. 3, top). As expected, bulk-based methods are separated from others, pseudo-bulk and single-cell methods performed the same, scRNA-Seq data being more similar. Kernel testing shows performance close to single-cell methods.

We demonstrate that kernel testing does not show the same bias as other single-cell DEA methods that tend to over-detect highly-expressed genes as mentioned in the original study (Fig. 3, bottom). By inspecting the distributional changes associated to genes considered as false-positive in the original study (with bulk-RNASeq genes as the ground truth), we show that they can in fact be interpreted as true positives. Many of them belong to the DB alternative (difference in both modalities and proportions, [29]), and were thus undetectable from bulk-RNAseq data and pseudo-bulk methods (Fig. S.7, left). Their classification in terms of false positives is then questionable, and kernel testing is clearly powerful to detect those alternatives on experimental data. Others present slight shifts in distribution and low zero proportions, these genes are correctly detected by the Fisher kernel adapted to zero-inflation (examples of such distribution shapes are shown in Fig. S.7, right).

2.6 Kernel testing reveals the heterogeneity of reverting cells

Single-cell transcriptomics has been widely used to investigate the molecular bases of cell differentiation, and has highlighted the stochasticity and dynamics of the underlying gene regulatory networks. The stochasticity of GRNs allows plasticity between cell states, and is a source of heterogeneity between cells along the differentiation path, which calls for multivariate differential analysis methods. We focus on the differentiation path of chicken primary erythroid progenitor cells (T2EC). A first study highlighted the existence of plasticity, i.e. the ability of cells induced into differentiation to reacquire the phenotypic characteristics of undifferentiated cells (e.g. starting self-renewing again), until a differentiation point of commitment (around 24H after differentiation induction) after which this phenotype was lost [44]. A second study investigated the molecular mechanisms underlying cell differentiation and reversion by measuring cell transcriptomes at four time points : undifferentiated T2EC maintained in a self-renewal medium (0H), then put in a differentiation-inducing medium for 24h (24H). The population was then split into a first population maintained in the same medium for 24h to achieve differentiation (48HDIFF), the second population was put back in the self-renewal medium to investigate potential reversion (48HREV) [59]. Cell transcriptomes were measured using scRT-qPCR on 83 genes selected to be involved in the differentiation process, as well as scRNA-Seq to complement the study by a non-targeted approach. Despite the strong global transcriptomic similarity between 0H and 48HREV cells, four DE genes were identified in the study (*RSFR*, *HBBA*, *TBC1D7*, *HSP90AA1*), interpreted as either a delay or as traces of engagement into differentiation of the 48HREV population, before returning to the self-renewal state. Hence, these first analyses suggested some heterogeneities between undifferentiated cells and reverted cells.

Our kernel-based test confirmed this heterogeneity by detecting a significant difference between undifferentiated cells (0H) and reverted cells (48HREV), both in scRT-qPCR and scRNA-Seq data (p-values $6.15\ 10^{-24}$ and $5.05\ 10^{-05}$ respectively), however considering the test statistic as a distance also confirmed the close proximity between these two conditions (Fig. 4.b and S.8.a). We assumed that population 48HREV was heterogeneous and contained reverted cells and non-reverted cells. A k-means clustering was unable to detect any particular cell cluster (Fig. S.9, middle). As the discriminant axis

provided by our framework represents a synthetic summary of the global transcriptomic differences between two cell populations, it allowed to highlight the existence of a sub-population of 48HREV cells (denoted 48HREV-1) that overlaps the distribution summary of 48HDIFF-cells (48HREV vs. 48HDIFF, Fig. 4.c). Interestingly, these cells also matched the distribution summary of 24H-cells (48HREV vs. 24H, Fig. 4.c), and were separated from the undifferentiated cells (48HREV vs 0H, Fig. 4.c). A similar sub-population was detected using scRNA-Seq data (48HREV vs. 48HDIFF Fig. S.8.b). According to our test, populations 48HDIFF and 48HREV-1 were very slightly different on scRT-qPCR data and similar on scRNA-Seq data (p-values 2.51 10^{-3} and 0.88 respectively). This slight difference may be explained by the targeted approach of scRT-qPCR that was based on a selection of 83 genes involved in differentiation and on the higher precision of the scRT-qPCR technology [59]. 48HREV-2 cells (48HREV cells after removing 48HREV-1 cells) were closer but still significantly different from 0H cells in both technologies (p-values 4.36 10^{-16} and 3.8 10^{-05} respectively). To describe these two sub-populations in terms of genes, we performed a k-means clustering on the averaged expression of genes over cells in populations 0H, 24H, 48HDIFF, 48HREV-1, 48HREV-2. We identified three and five gene clusters on the scRT-qPCR and the scRNA-Seq data respectively. These clusters can be separated in three groups (Fig 4.d and S.8.c): (*i*) genes activated during differentiation (scRT-qPCR cluster 2, scRNA-Seq clusters 0 and 2), e.g. hemoglobin related genes such as *HBA1* and *HBAD* (shown in Fig. S.8.d), (*ii*) genes deactivated during differentiation (scRT-qPCR cluster 1, scRNA-Seq cluster 3) e.g. genes involved in metabolism of self-renewing cells such as *LDHA* and *LY6E* (shown in Fig. S.8.d), and (*iii*) genes with no clear function pattern for which the expression levels did not change much during differentiation and reversion (scRT-qPCR cluster 0 and scRNA-Seq clusters 1 and 4). The p-value tables associated to each pair-wise univariate DE analysis with respect to each gene cluster are available online [2].

To conclude, our differential transcriptome framework showed that population 48HREV is composed of two sub-populations, which sheds light on new putative mechanisms driving differentiation and reversion processes. Whereas a population is only slightly different to undifferentiated cells (48HREV-2), a sub-population (48HREV-1) has remained engaged in differentiation. This difference could be either due to a delay in engaging the reversion process for some cells, or to cells having crossed the irreversible point of commitment. Furthermore, our method has identified cellular pathways which could be important either for cell plasticity or cell differentiation, and can guide design of further experiments. Overall, it could enhance our comprehension of how gene regulatory networks react to differentiation and reversion signals.

3 Towards a new testing framework for differential binding analysis in single-cell ChIP-Seq data

There is currently no dedicated method for the comparison of single-cell epigenomic profiles, existing studies often use non-parametric testing to compare epigenomic states and retrieve differentially enriched loci. The joint multivariate testing strategy seems particularly suited to compare epigenomic data since it is well established that chromatin conformation and natural spreading of histone modifications - in particular H3K27me3 [35] - can induce complex dependencies between sites occupancy. A recent study [36] has shown that the repressive histone mark H3K27me3 (trimethylation of histone H3 at lysine 27) is involved in the emergence of drug persistence in breast cancer cells. Drug persistence occurs when only a subset of cells, known as persister cells, survives the initial drug treatment, thereby creating a reservoir of cells from which resistant cells will emerge. The study identified a persister expression program involving genes such as *TGFB1* and *FOXQ1*, with H3K27me3 as a lock to its activation. Changes in H3K27me3 modifications at the single-cell level showed a consistent pattern in persister cells compared to untreated cells, in particular cells display recurrent losses of repressive histone methylation at a subset of genes of the persister expression program. However, this pattern was not necessarily maintained in cells that developed full resistance, suggesting the that part of the epigenomic features of persister cells might be transient. Moreover, analysis of untreated cells revealed heterogeneity within epigenomic profiles. Part of the population exhibited shared epigenomic features with persister cells, yet remaining distinguishable from them. This initial analysis suggested that a pool of untreated cells could contribute to the persister cell population later upon exposure to chemotherapy. However, unsupervised analyses were unable to clearly identify this pool of precursor cells.

We compared H3K27me3 scChIP-seq profiles between untreated and persister cells using kernel testing. Thanks to the discriminative approach, our framework offers a synthetic representation of the distributional differences between cell populations 5. Projecting cells on the kernelized discriminant axis reveals 3 sub-populations within the untreated cell population: Persister-Like (109 cells; 5% of untreated

[2] `https://github.com/AnthoOzier/kernel_testsDA.git`

cells), Intermediate (1124 cells; 57%), Naive (744 cells; 38%), with increasing distance to persister cells (Fig. 5). We then performed a differential analysis of H3K27me3 enrichment between persister cells and the $n = 109$ untreated cells that were the most similar to persister cells on the discriminant axis. Over the 6,376 tested regions, only 14 were significantly differentially enriched (p-value$< 10^{-3}$, Table S.1), suggesting that this sup-population of untreated cells is epigenomically very close to persister cells (with persister cells being hypo-methylated on these significant regions compared to persister-like cells). We then studied the differences between the three populations present in the untreated cell population, prior to any treatment. We performed differential analysis between the most distant untreated cells ('naive' vs 'intermediate'), and between 'intermediate' cells and 'persister-like' cells. We detected significant changes in repressive epigenomic enrichments, both losses and gains, that will need further functional testing to understand their potential role in drug-persistence. Altogether, our new kernel analytical framework shows that persister-like cells could exist prior to any treatment, and provides a novel level of appreciation of epigenomic heterogeneity - by revealing three sub-populations within treatment naive cell population. In addition, our method identifies small quantitative variations that are not detected by other methods and will need to be related to gene expression and other molecular features for further interpretation.

4 Conclusion

In this work we introduced the framework of kernel testing to perform differential analysis in a non-linear setting. This method compares the distribution of gene expression or epigenomic profiles through global or feature-wise comparisons, but can be extended to any measured single-cell features. Kernel testing has focused much attention in the machine learning community since it has the advantage of being non-linear, computationally tractable, and provides visualization combining dimension reduction and statistical testing. Its application to single-cell data is particularly promising, as it allows distributional comparisons without any assumptions about their shape. Moreover, using a classifier to perform discrimination-based testing has become popular [28], and allows powerful detection of population heterogeneities in both expression and epigenomics single-cell data. Our simulations show the power of this approach on specifically designed alternatives [29]. Furthermore, comparing kernel testing with other methods based on multiple scRNA-Seq data reveals its superior capability to identify distributional changes that go undetected by other approaches. Finally, the application of kernel testing to scRNA-Seq and scChIP-seq data uncovers biologically meaningful heterogeneities in cell populations that were not identified by standard procedures.

Perspectives of this work are numerous: we will first generalize the approach beyond the two-sample case and extend it to multiple sample comparisons. We are currently working on a more general design that considers multiple factors such as batch effects. The adaptability of kernel methods makes them particularly well-suited for spatial data, so we plan to extend the framework to include spatial data analysis. To bridge the gap between global and feature-wise approaches, we are actively developing sensitivity analysis methods. These methods will help us identify influential features that contribute to the rejection of the global hypothesis. By combining these findings, we can create a joint approach that considers the complex dependencies inherent in single-cell data, while still providing interpretable outputs based on feature-wise information. Finally, kernel testing also raises much theoretical efforts to further characterize its properties [22]. More than ever, single-cell data science appears at the convergence of many cutting-edge methodological developments in machine learning. As a result, these advancements will have significant implications for the old-tale of differential analysis, offering new avenues for progress and improvement.

5 Materials and Methods

5.1 Simulation settings

The comparison study on data simulated was performed on data following different mixtures of zero inflated negative binomial (ZINB) distributions [14]. The distribution parameters were chosen to reproduce the four Korthauer alternatives and two types of H_0 distributions. The performances were computed on 500 repetitions of a dataset composed of 1000 DE genes and 9000 non-DE genes. The DE genes are equaly separated in the four alternatives DE,DM, DP and DB. The non-DE genes are equally separated into an unimodal ZINB and a bimodal mixture of ZINB. The DE methods were applied on the raw data, type-I errors and powers were computed on the raw p-values while false discovery and true discovery rates were computed on the adjusted p-values, with the Benjamini-Hochberg correction [5]. The authors

also shared their p-values tables with us for their methods (cicdf-asymp and citcdf-perm) [14], MAST [12], scDD [29], SigEMD [56], DESingle [38] and SCDE [27].

5.2 Comparison of methods on published scRNA-Seq

The eighteen comparison datasets were downloaded from the Zenodo repository (`https://doi.org/10.5281/zenodo.5048449`) compiled by Squair and coauthors [53]. They consists of six comparisons of bone marrow mononuclear phagocytes from mouse, rat, pig and rabbit in different conditions [21], eight comparisons of naive and memory T cells in different conditions [8] and four comparisons of alveolar macrophages and type II pneumocytes between young and old mouses [1] and between patients with pulmonary fibrosis and control individuals [43]. More details on the datasets are in [53] or in the original studies. The preprocessing step consisted in filtering genes present in less than three cells and normalizing data with the Seurat function *NormalizeData*, as in the original comparative study [53]. This not very restrictive preprocessing was chosen in order to not introduce biaises in the analyses, and many genes would have been ignored form the analysis in real conditions. The Area Under the Concordance Curves (AUCC) scores were computed with the original scripts [53].

5.3 Fisher kernel for zero-inflated data

To derive the zero-inflated kernel, we consider a zero-inflated Gaussian distribution with π the proportion of additional zeros:

$$Y \sim \pi\delta_0(\bullet) + (1-\pi)\phi(\bullet;\mu,\sigma)$$

This distribution has a mixture representation, with Z standing for the binary variable of distribution $\mathcal{B}(\pi)$, such that

$$f_{\mu,\sigma,\pi}(x) = \mathbb{P}(Z=1)\delta_0(y) + \mathbb{P}(Z=0)\phi(y;\mu,\sigma)$$

We know the probability kernels for the Gaussian distributions:

$$k_{\text{Gauss}}(\mu,\mu') = \frac{1}{4\pi\sigma^2}\exp(-(\mu-\mu')^2/4\sigma^2)$$

and for the Bernoulli distribution:

$$k_{\mathcal{B}}\Big(\pi,\pi'\Big) = \pi\pi' + (1-\pi)(1-\pi').$$

To get the ZI-Gauss kernel, we compute the probability kernel $f_{\mu,\sigma,\pi}$ and $f_{\mu',\sigma,\pi'}$

$$\begin{aligned} k_{\text{ZI-Gauss}}\Big(f_{\mu,\sigma,\pi}, f_{\mu',\sigma,\pi'}\Big) &= \int_{y,y'}\sum_{z,z'}\Big(\mathbb{P}_\pi(Z=z)f_{\mu,\sigma}(y \mid Z=z)\Big)\Big(\mathbb{P}_{\pi'}(Z=z')f_{\mu',\sigma}(y' \mid Z=z')\Big)dydy' \\ &= \pi\pi' + \pi(1-\pi')f_{\mu',\sigma}(0) + (1-\pi)\pi' f_{\mu,\sigma}(0) + (1-\pi)(1-\pi')K_{\text{Gauss}}(\mu,\mu'), \end{aligned}$$

with $f_{\mu,\sigma}$ the Gaussian probability density function. In the simulations, the Fisher kernel was computed using the parameters of the Binomial distributions used to determine the drop-out rates of the simulated data (drawn uniformly in $[0.7, 0.9]$), the variance parameter σ was set as the median distance between the non-zero observations and the Gaussian means μ were set as the observed values. The truncation parameter choice as $t = 3$.

5.4 Reversion data

Details on the experiment and on the data can be found in the original paper [59]. The kernel-based testing framework was performed on the $\log(x+1)$ normalized RT-qPCR data and on the Pearson residuals of the 2000 most variable genes of the scRNA-Seq data obtained through the R package `sctransform` [20]. The truncation parameter of the global comparisons ($t = 10$ for both technologies) was chosen to be large enough for the discriminant analysis to capture enough of the multivariate information and to maximize the discriminant ratio. The truncation parameter retained for univariate testing ($t = 4$) was chosen according to the simulation study.

6 Acknowledgments

The research was supported by a grant from the Agence Nationale de la Recherche ANR-18-CE45-0023 SingleStatOmics, by the projects AI4scMed, France 2030 ANR-22-PESN-0002, and SIRIC ILIAD (INCA-DGOS-INSERM-12558). The authors would like to thank Boris Hejblum for sharing the simulated data,

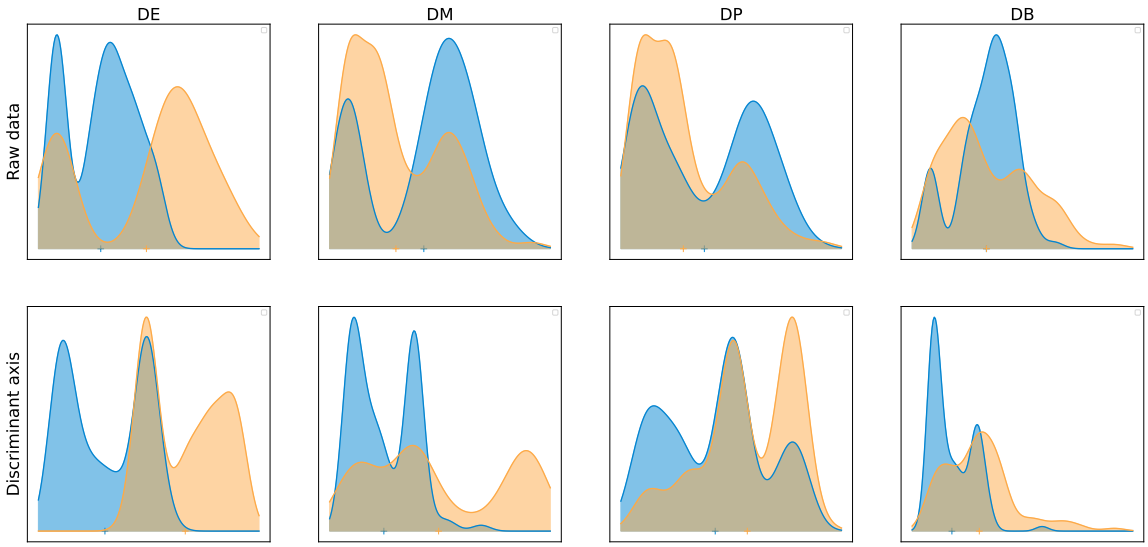

Figure 1: Top : Examples of distributions of the simulated data, DE : classical difference in expression, DM : difference in modalities, DP : difference in proportions, DB : difference in both modalities and proportions with equal means. Bottom : projection of cells on the discriminant axis ($T = 4$) for each alternative. The non-linear transform allows the separation of distributions on the discriminant axis.

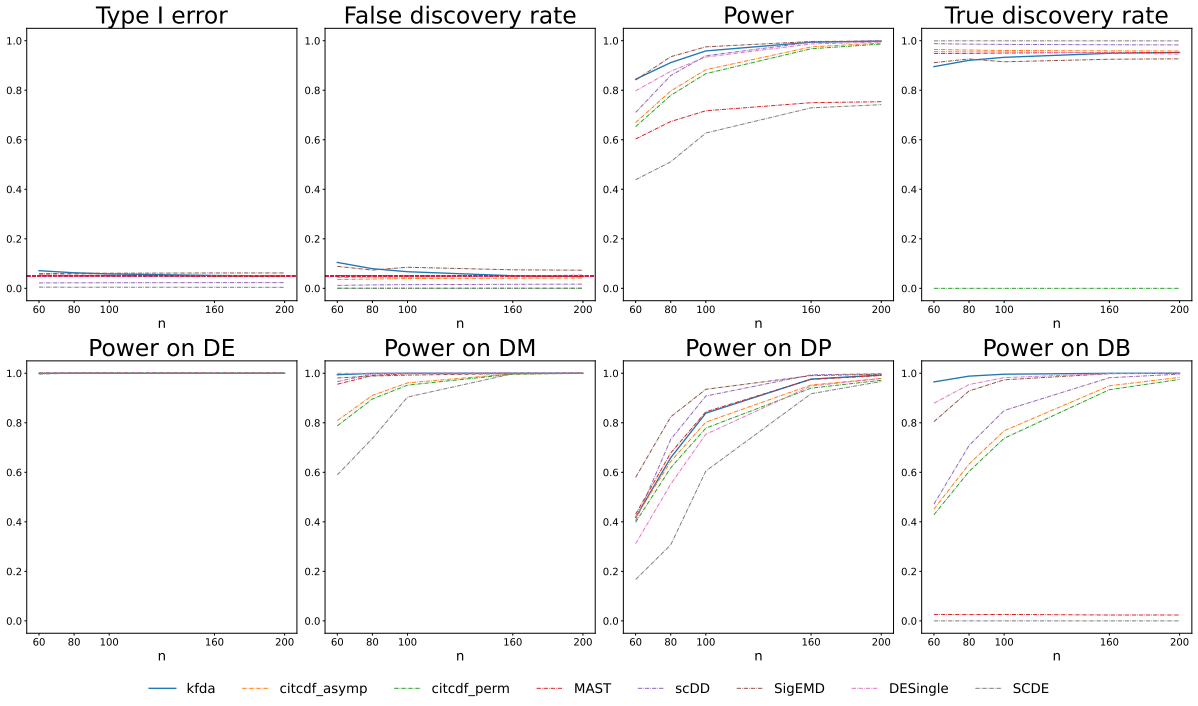

Figure 2: Comparison of DEA methods with respect to type-I errors and power. Top: Type-I errors are computed on raw p-values under H_0. False discovery Rate computed on Benjamini-Hochberg adjusted p-values. Power computed on raw p-values under H_1. True Discovery Rate computed on Benjamini-Hochberg adjusted p-values. Simulated data consists of 100 cells, 10000 genes (1000 DE, 9000 non-DE). Alternatives are simulated using DE : classical difference in expression (250 genes), DM : difference in modalities (250 genes), DP : difference in proportions (250 genes), DB : difference in both modalities and proportions with equal means (250 genes). Error rates are computed over 500 replicates.

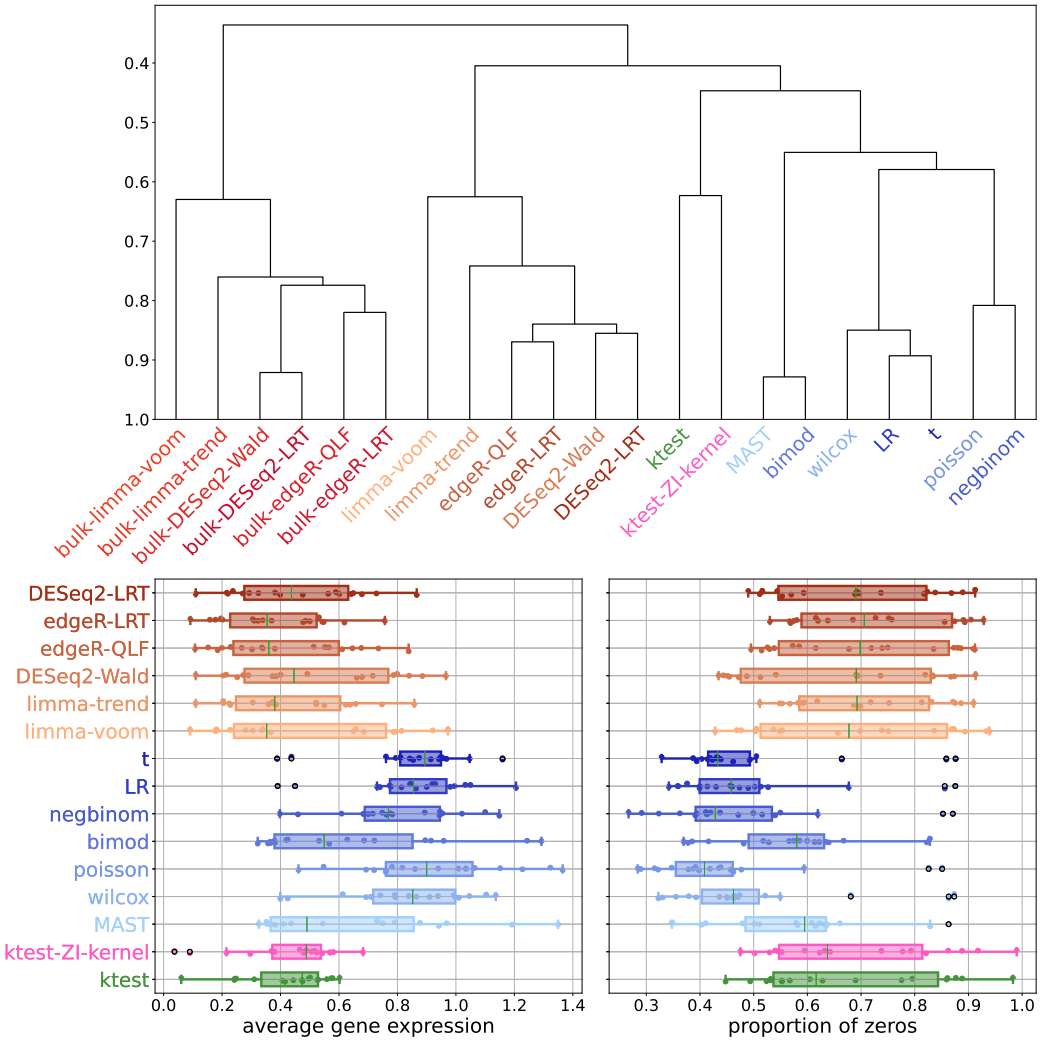

Figure 3: Top: Hierarchical clustering based on average AUCC scores computed between pairs of methods (over 18 datasets [53]). Bottom: Boxplot of the average expression (left) and proportion of zeros (right) of the top 500 DE genes for different DE methods (over 18 datasets [53]). Red: bulk methods, orange: pseudobulk methods, blue: single-cell methods. The truncation parameter is set to $T = 4$ for ktest.

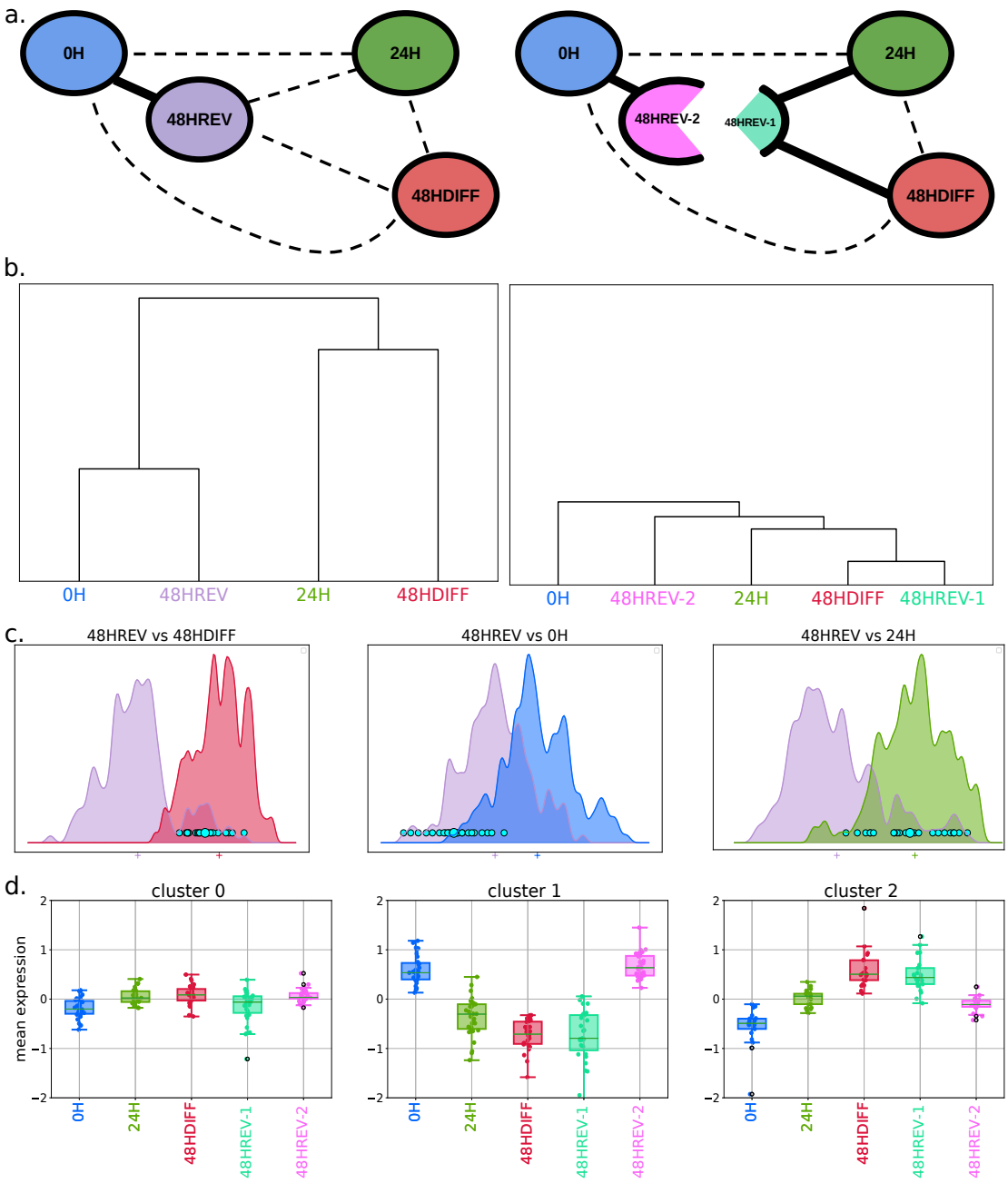

Figure 4: a: Summarized distance graphs between conditions before (left) and after (right) splitting condition 48HREV into populations 48HREV-1 and 48HREV-2. b: Trees from pairwise distances using our test statistic between conditions before (left) and after (right) splitting condition 48HREV into populations 48HREV-1 and 48HREV-2. c: Cell densities of compared conditions projected on the discriminant axis between conditions 48HREV and 48HDIFF (left), 48HREV and 0H (middle) and 48HREV and 24H (right) with highlighted population 48HREV-1. d : Boxplots of the mean expressions of the five populations 0H, 24H, 48HDIFF, 48HREV-1 and 48HREV-2 for the three genes clusters. a,b,c and d are obtained from scRT-qPCR data

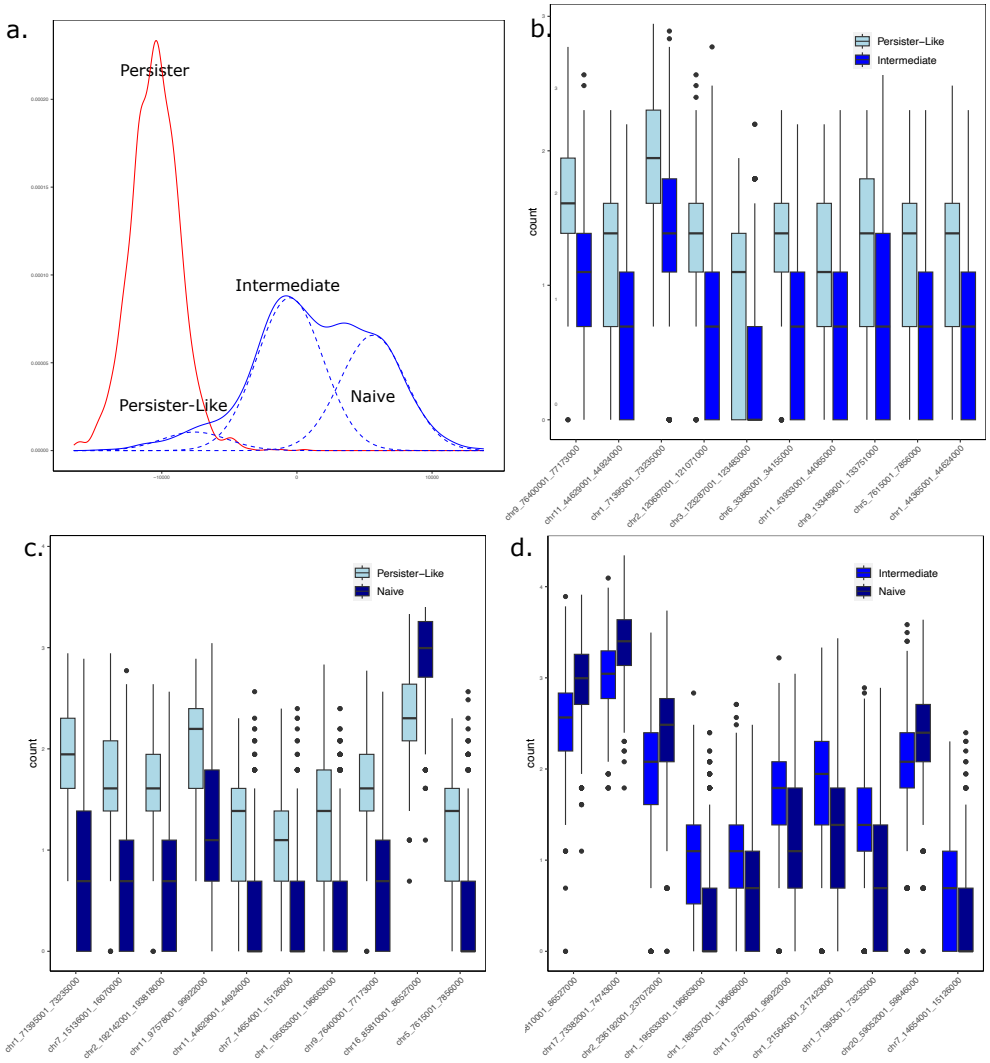

Figure 5: Differential analysis of scChIP-seq data on breast cancer cells. a. Cell densities of persister cells vs. untreated cells. Sub-populations of untreated cells were identified using 3-component mixture model, that revealed persister-like cells, intermediate and naive cells. b-c-d : boxplots of the top-10 differentially enriched H3K27me3 loci between the 3 sub-populations. Features are designated by the genomic coordinates of the ChIP-seq peaks. Corresponding overlapping genes are provided in Table S.1

François Gindraud for helping on the implementation of the kernel method, Stéphane Minvielle and Zaid Harchaoui for fruitful scientific discussions. This work was performed using HPC resources from GLiCID computing center.

S.1 Supplementary Material

S.1 Tuning the truncation hyperparameter

We use the simulation data to calibrate the hyperparameter of our method, i.e. the number T of principal directions of the within-covariance operator to retain to regularize the kernel-based Mahalanobis distance. The theoretical calibration of this hyperparameter still requires heavy mathematical developments, as shown by recent work [22]. However, these simulations provide a simple rule of thumb to choose it. Indeed, since T can be interpreted as the quantity of within-variance information used to describe the residual expression, increasing T will increase power in the detection of complex alternatives, at the price of increased type-I errors. In the simulations, Type-I errors of the kernel test remains at the nominal level $\alpha = 5\%$ until $T \leq 6$. with maximal power for $T = 4$ (Fig S.6). Interestingly, the test was completely unable to detect the DB alternative when $T = 1$. These results confirm that the truncation hyperparameter should be chosen as a trade-off between maximizing testing power while keeping the type-I errors controlled at the nominal level to ensure calibration. This motivates the choice of $T = 4$ for all the univariate DE analyses in the paper.

For multivariate analyses, we assumed that the meaningful information was contained in more than four principal directions of the within-covariance operator and chose to take a larger truncation parameter in order to take into account more of the multivariate information available. We then chose the truncation parameter $T = 10$ that maximized the discriminant ratio while being not too large to still ensure the calibration.

S.2 Kernel trick for the effective computation of the test statistic

In this section, we describe how to compute the test statistic $\widehat{D}_T^2(\widehat{\mu}_1, \widehat{\mu}_2)$ and the vector of projections of the embeddings onto the discriminant axis V, with $i \in \{1, 2\}$, $j \in \{1, \ldots, n_i\}$, and $V = (\langle h_T^\star, \phi(Y_{i,j})\rangle_{\mathcal{H}})_{i,j}$ for $T \in \{1, \ldots, n\}$. This computation relies on the kernel trick that consists in expressing every quantity of interest with respect to the gram matrix K containing every pair-wise evaluation of the kernel function, such that for $i, i' \in \{1, 2\}$, $K = (K_{i,i'})_{i,i'}$, where for $j \in \{1, \ldots, n_i\}$, $j' \in \{1, \ldots, n_{i'}\}$, $K_{i,i'} = (k(Y_{i,j}, Y_{i',j'}))_{j,j'}$. The computation has two steps. First, we determine a matrix K_W that has the same eigenvalues as the operator $\widehat{\Sigma}_W$, then we compute the quantities of interest with respect to K, the T first eigenvalues $(\widehat{\lambda}_t)$, $t \in \{1, \ldots, T\}$ and the associated unit eigenvectors $(u_t)_t$ of K_W. Let's denote by I_n the identity matrix of size n, J_n the matrix of size n full of 1, and $\mathbb{1}_n$ the vector of size n full of 1. Then for $i \in \{1, 2\}$, let $P_i = I_{n_i} - n_i^{-1} J_{n_i}$, $P = \mathrm{diag}(P_1, P_2)$ and $\omega = (\mathbb{1}_{n_1}, -\mathbb{1}_{n_2})' \in \mathbb{R}^n$. We can show that the matrix K_W is equal to $K_W = n^{-1}PKP$. Then we have :

$$\widehat{D}_T^2(\widehat{\mu}_1, \widehat{\mu}_2) = \frac{n_1 n_2}{n^2} \sum_{t=1}^{T} \widehat{\lambda}_t^{-2} (u_t' PK\omega)^2, \quad \text{and} \quad V = \frac{n_1 n_2}{n^2} \sum_{t=1}^{T} \widehat{\lambda}_t^{-2} (u_t' PK\omega) KPu_t.$$

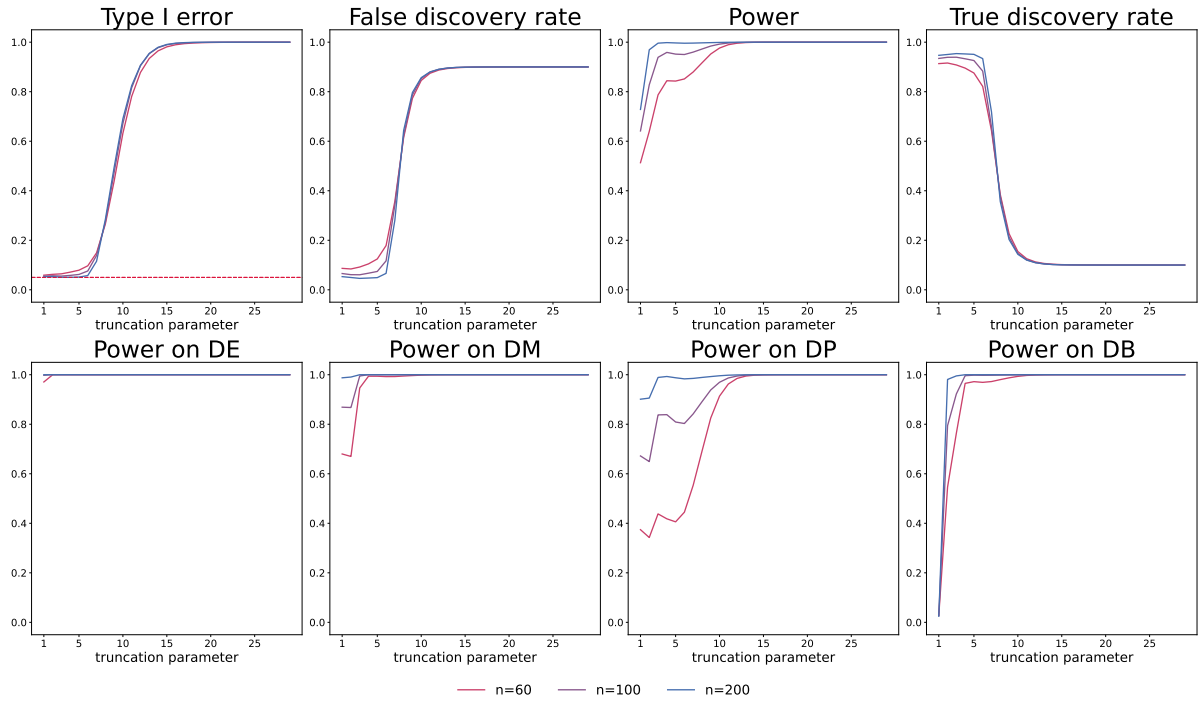

Figure S.6: Calibration of the truncation with respect to type-I errors and power. Top: Type-I errors are computed on raw p-values under H_0. False discovery Rate computed on Benjamini-Hochberg adjusted p-values. Power computed on raw p-values under H_1. True Discovery Rate computed on Benjamini-Hochberg adjusted p-values. Simulated data consists of 10000 genes (1000 DE, 9000 non-DE). Alternatives are simulated using DE : classical difference in expression (250 genes), DM : difference in modalities (250 genes), DP : difference in proportions (250 genes), DB : difference in both modalities and proportions with equal means (250 genes). Error rates are computed over 500 replicates.

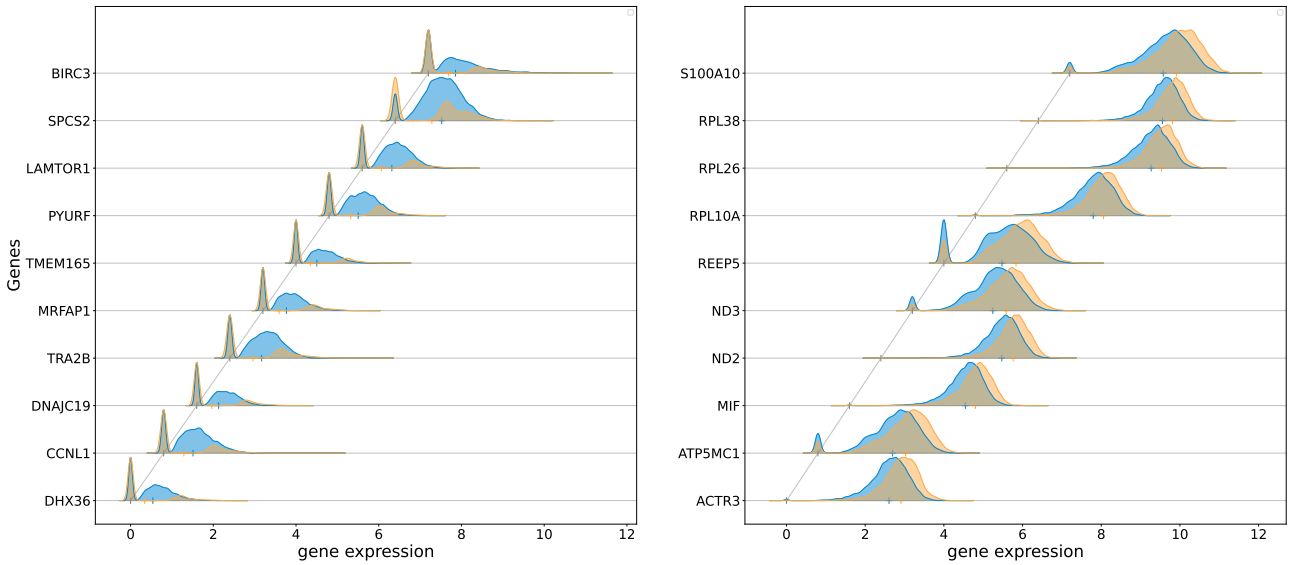

Figure S.7: Expression densities of the two compared conditions for genes considered as DE by ktest-ZI-kernel and the other single-cell DE methods and considered as non-DE by pseudo-bulk methods. Left: stimulated memory Th0 cells (blue, 4766 cells) vs control memory Th0 cells (orange, 3110 cells) from [8]. Right : pig cells stimulated with lipopolysaccharide (blue, 6605 cells) vs control pig cells (orange, 6148 cells) from [21].

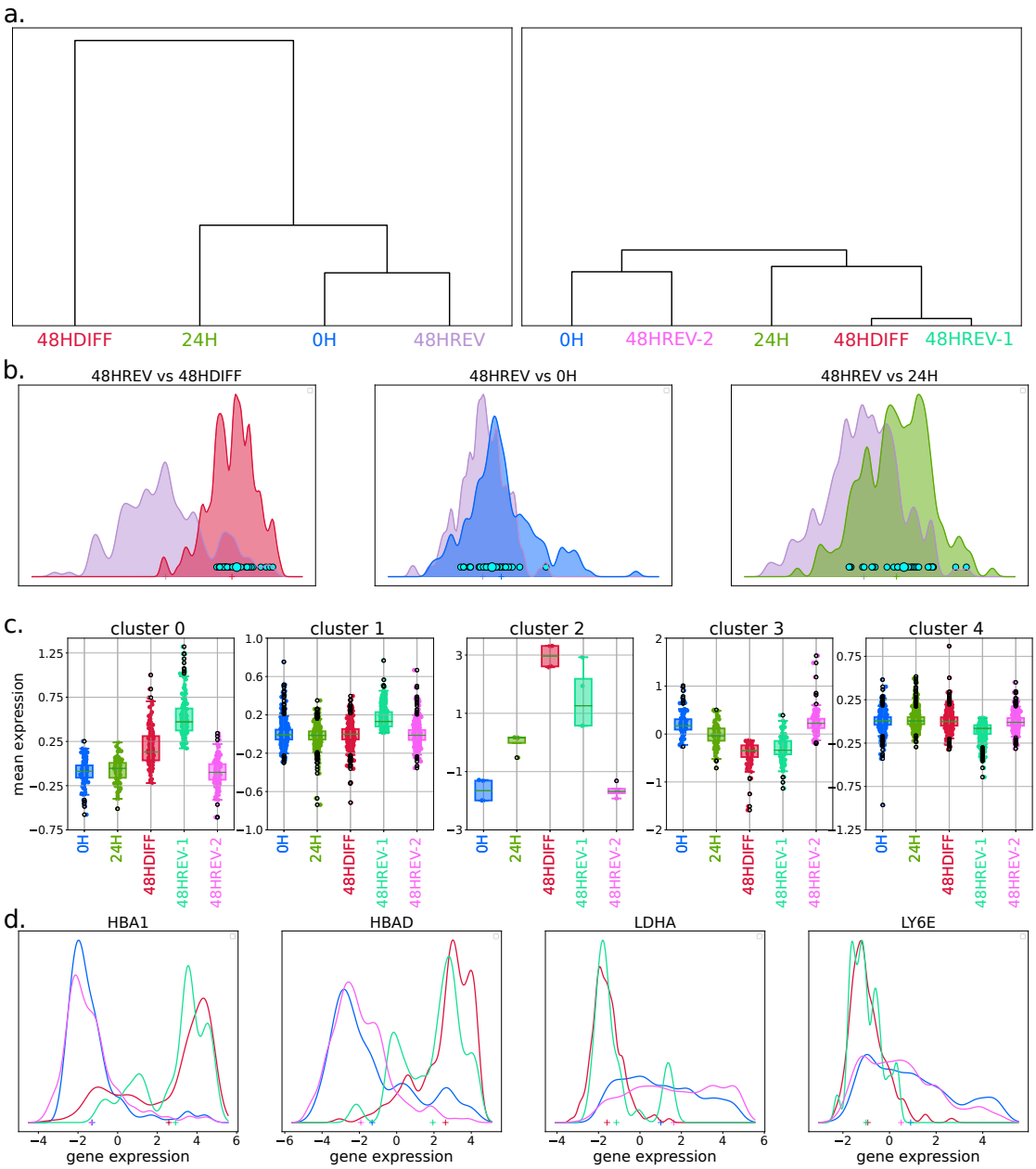

Figure S.8: a : Trees from pairwise distances using our test statistic between conditions before (left) and after (right) splitting condition 48HREV into populations 48HREV-1 and 48HREV-2. b : Cell densities of compared conditions projected on the discriminant axis between conditions 48HREV and 48HDIFF (left), 48HREV and 0H (middle) and 48HREV and 24H (right) with highlighted population 48HREV-1. c: Boxplots of the mean expressions of the five populations 0H, 24H, 48HDIFF, 48HREV-1 and 48HREV-2 for the five identified genes clusters. d : Examples of gene expression distributions in populations 48HREV-1 (turquoise) and 48HREV-2 (pink) compared to populations 0H (blue) and 48HDIFF (red). a,b,c and d are obtained from scRNA-Seq data

chr	start	end	$\widehat{D}_T^2$	average Persist.	average Persist.-Like	average log2FC	gene
chr9	133489001	133751000	123.60	0.77	1.25	-0.51	ADAMTSL2, DBH, SARDH
chr5	1832001	2740000	104.70	2.45	2.80	-0.43	IRX4
chr9	135509001	135802000	79.40	0.83	1.22	-0.43	PAEP, LCN1, OBP2A, SOHLH1, KCNT1, LCN9
chr9	134445001	135458000	72.50	1.67	2.12	-0.49	OLFM1, FCN2, FCN1, COL5A1
chr9	76400001	77173000	51.00	1.21	1.59	-0.41	GCNT1
chr14	23331001	23355000	50.20	0.05	0.12	-0.09	SLC22A17
chr9	136123001	136206000	49.10	0.22	0.42	-0.22	LHX3
chr12	129982001	130786000	48.10	1.05	1.41	-0.37	RIMBP2, PIWIL1
chr3	123287001	123483000	45.20	0.69	0.91	-0.26	ADCY5
chr22	47950001	49760000	43.30	2.48	2.78	-0.38	FAM19A5

chr	start	end	$\widehat{D}_T^2$	average Persist.-Like	average Interm.	average log2FC	gene
chr9	76400001	77173000	151.10	1.59	0.93	-0.64	GCNT1
chr11	44629001	44924000	107.70	1.19	0.65	-0.50	TSPAN18
chr1	71395001	73235000	102.10	1.94	1.37	-0.51	NEGR1
chr2	120687001	121071000	99.80	1.31	0.76	-0.48	GLI2
chr3	123287001	123483000	94.80	0.91	0.47	-0.43	ADCY5
chr6	33863001	34155000	90.30	1.24	0.72	-0.48	GRM4
chr11	43933001	44065000	79.80	1.04	0.61	-0.40	ACCSL
chr9	133489001	133751000	78.50	1.25	0.79	-0.41	SARDH, DBH, ADAMTSL2
chr5	7615001	7856000	78.20	1.20	0.69	-0.42	C5orf49
chr1	44365001	44624000	73.70	1.20	0.77	-0.40	RNF220

chr	start	end	$\widehat{D}_T^2$	average Interm.	average Naive	average log2FC	gene
chr16	85810001	86527000	466.50	2.52	2.95	0.50	FOXF1-IRF8
chr17	73382001	74743000	337.20	3.03	3.38	0.41	GPRC5C,CD300A,TTYH2, DNAI2, SDK2, RPL38, GPR142, CD300C, CD300LD, CD300LB, RAB37, KIF19, BTBD17, CD300LF, CD300E
chr2	236192001	237072000	314.50	1.99	2.41	0.48	IQCA1,ASB18
chr1	195633001	196663000	249.90	0.92	0.47	-0.51	KCNT2,CFH
chr1	189337001	190666000	235.80	1.03	0.58	-0.49	BRINP3
chr11	97578001	99922000	229.30	1.65	1.18	-0.55	CNTN5
chr1	215645001	217423000	221.10	1.79	1.30	-0.59	ESRRG,USH2A
chr1	71395001	73235000	215.20	1.37	0.91	-0.52	NEGR1
chr20	59052001	59846000	213.30	2.05	2.36	0.35	EDN3,PHACTR3
chr7	14654001	15126000	209.30	0.70	0.34	-0.37	DGKB

chr	start	end	$\widehat{D}_T^2$	average Persist.-Like	average Naive	average log2FC	gene
chr1	71395001	73235000	292.20	1.94	0.91	-1.05	NEGR1
chr7	15136001	16070000	250.40	1.65	0.70	-0.95	MEOX2,AGMO
chr2	192142001	193818000	237.00	1.64	0.74	-0.90	TMEFF2
chr11	97578001	99922000	230.30	2.04	1.18	-0.86	CNTN5
chr11	44629001	44924000	217.10	1.19	0.39	-0.77	TSPAN18
chr7	14654001	15126000	203.80	1.09	0.34	-0.72	DGKB
chr1	195633001	196663000	201.50	1.31	0.47	-0.83	KCNT2,CFH
chr9	76400001	77173000	199.90	1.59	0.68	-0.91	GCNT1
chr16	85810001	86527000	199.30	2.29	2.95	0.93	FOXF1,IRF8
chr5	7615001	7856000	188.20	1.20	0.43	-0.73	C5orf49

Table S.1: Differential analysis of sc-chIPseq data: top-10 differential regions for pairwise comparisons between persister cells and the three sub-populations of untreated cells. Adjusted p-values are $< 10^{-3}$ (Bonferroni correction). The last Gene column corresponds to the genes overlapping the regions.

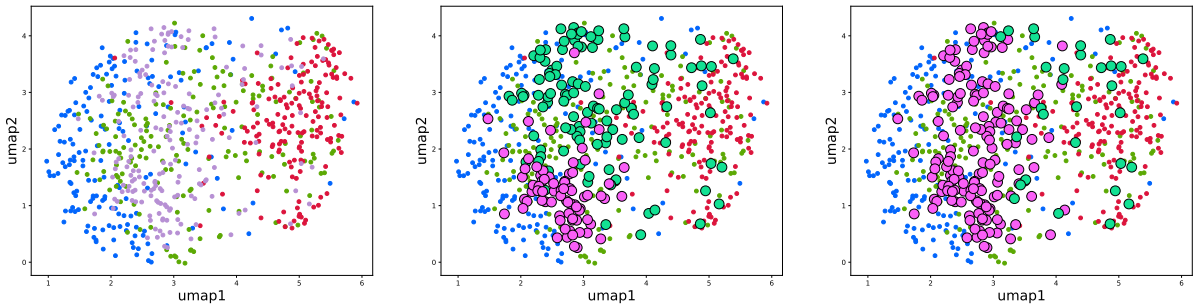

Figure S.9: Left: Umap representation of the four conditions from scRNA-Seq data (0H (blue), 24H (green) 48HDIFF (red) and 48HREV (purple)). Middle : highlight of the 2 groups of 48HREV identified through a k-means algorithm. Right : The two groups 48HREV-1 (turquoise) and 48HREV-2 (pink) identified on the discriminant axis associated to the truncation parameter $t = 10$.

References

[1] I. Angelidis, L. M. Simon, I. E. Fernandez, M. Strunz, C. H. Mayr, F. R. Greiffo, G. Tsitsiridis, M. Ansari, E. Graf, T.-M. Strom, M. Nagendran, T. Desai, O. Eickelberg, M. Mann, F. J. Theis, and H. B. Schiller. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nature Communications*, 10(1):963, Feb. 2019.

[2] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Twenty-first international conference on Machine learning - ICML '04*, page 6, Banff, Alberta, Canada, 2004. ACM Press.

[3] T. Banerjee, B. B. Bhattacharya, and G. Mukherjee. A Nearest-Neighbor Based Nonparametric Test for Viral Remodeling in Heterogeneous Single-Cell Proteomic Data. *arXiv:2003.02937 [stat]*, June 2020. arXiv: 2003.02937.

[4] M. Bartosovic, M. Kabbe, and G. Castelo-Branco. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nature Biotechnology*, 39(7):825–835, July 2021.

[5] Benjamini et Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing on JSTOR, 1995.

[6] J. D. Buenrostro, B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, and W. J. Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, July 2015.

[7] M. Büttner, J. Ostner, C. L. Müller, F. J. Theis, and B. Schubert. scCODA is a Bayesian model for compositional single-cell data analysis. *Nature Communications*, 12(1):6876, Nov. 2021.

[8] E. Cano-Gamez, B. Soskic, T. I. Roumeliotis, E. So, D. J. Smyth, M. Baldrighi, D. Willé, N. Nakic, J. Esparza-Gordillo, C. G. C. Larminie, P. G. Bronson, D. F. Tough, W. C. Rowan, J. S. Choudhary, and G. Trynka. Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4+ T cells to cytokines. *Nature Communications*, 11(1):1801, Apr. 2020.

[9] Y. Cao, Y. Lin, J. T. Ormerod, P. Yang, J. Y. Yang, and K. K. Lo. scDC: single cell differential composition analysis. *BMC Bioinformatics*, 20(19):721, Dec. 2019.

[10] E. Dann, N. C. Henderson, S. A. Teichmann, M. D. Morgan, and J. C. Marioni. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature Biotechnology*, 40(2):245–253, Feb. 2022.

[11] S. Das, A. Rai, and S. N. Rai. Differential Expression Analysis of Single-Cell RNA-Seq Data: Current Statistical Approaches and Outstanding Challenges. *Entropy (Basel, Switzerland)*, 24(7):995, July 2022.

[12] G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic, P. S. Linsley, and R. Gottardo. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16(1):278, Dec. 2015.

[13] D. Garreau, W. Jitkrittum, and M. Kanagawa. Large sample analysis of the median heuristic, Oct. 2018. arXiv:1707.07269 [math, stat].

[14] M. Gauthier, D. Agniel, R. Thiébaut, and B. P. Hejblum. Distribution-free complex hypothesis testing for single-cell RNA-seq differential expression analysis, Nov. 2021. bioRxiv doi: 10.1101/2021.05.21.445165.

[15] C. Gawad, W. Koh, and S. R. Quake. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 17(3):175–188, Mar. 2016.

[16] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A Kernel Method for the Two-Sample-Problem. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

[17] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

[18] A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[19] K. Grosselin, A. Durand, J. Marsolier, A. Poitou, E. Marangoni, F. Nemati, A. Dahmani, S. Lameiras, F. Reyal, O. Frenoy, Y. Pousse, M. Reichen, A. Woolfe, C. Brenan, A. D. Griffiths, C. Vallot, and A. Gérard. High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nature genetics*, 51(6):1060–1066, June 2019.

[20] C. Hafemeister and R. Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):296, Dec. 2019.

[21] T. Hagai, X. Chen, R. J. Miragaia, R. Rostom, T. Gomes, N. Kunowska, J. Henriksson, J.-E. Park, V. Proserpio, G. Donati, L. Bossini-Castillo, F. A. Vieira Braga, G. Naamati, J. Fletcher, E. Stephenson, P. Vegh, G. Trynka, I. Kondova, M. Dennis, M. Haniffa, A. Nourmohammad, M. Lässig, and S. A. Teichmann. Gene expression variability across cells and species shapes innate immunity. *Nature*, 563(7730):197–202, Nov. 2018.

[22] O. Hagrass, B. K. Sriperumbudur, and B. Li. Spectral Regularized Kernel Two-Sample Tests, Dec. 2022. arXiv:2212.09201 [cs, math, stat].

[23] Z. Harchaoui, F. Bach, O. Cappe, and E. Moulines. Kernel-Based Methods for Hypothesis Testing: A Unified View. *IEEE Signal Processing Magazine*, 30(4):87–97, July 2013.

[24] Z. Harchaoui, F. Vallet, A. Lung-Yut-Fong, and O. Cappe. A regularized kernel-based approach to unsupervised audio segmentation. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1665–1668, Taipei, Taiwan, Apr. 2009. IEEE.

[25] D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, and I. Amit. Massively parallel single cell RNA-Seq for marker-free decomposition of tissues into cell types. *Science (New York, N.Y.)*, 343(6172):776–779, Feb. 2014.

[26] T. Jebara, R. Kondor, and A. Howard. Probability Product Kernels. *Journal of Machine Learning Research*, 5(Jul):819–844, 2004.

[27] P. V. Kharchenko, L. Silberstein, and D. T. Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742, July 2014.

[28] I. Kim, A. Ramdas, A. Singh, and L. Wasserman. Classification accuracy as a proxy for two-sample testing. *The Annals of Statistics*, 49(1):411–434, 2021.

[29] K. D. Korthauer, L.-F. Chu, M. A. Newton, Y. Li, J. Thomson, R. Stewart, and C. Kendziorski. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biology*, 17(1):222, Dec. 2016.

[30] J. M. Kübler, W. Jitkrittum, B. Schölkopf, and K. Muandet. A Witness Two-Sample Test. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 1403–1419. PMLR, May 2022. ISSN: 2640-3498.

[31] D. Lopez-Paz and M. Oquab. Revisiting Classifier Two-Sample Tests, Mar. 2018. arXiv:1610.06545 [stat].

[32] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, Dec. 2014.

[33] L. v. d. Maaten and G. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[34] E. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. Bialas, N. Kamitaki, E. Martersteck, J. Trombetta, D. Weitz, J. Sanes, A. Shalek, A. Regev, and S. McCarroll. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214, May 2015.

[35] R. Margueron, N. Justin, K. Ohno, M. L. Sharpe, J. Son, W. J. Drury, P. Voigt, S. Martin, W. R. Taylor, V. De Marco, V. Pirrotta, D. Reinberg, and S. J. Gamblin. Role of the polycomb protein Eed in the propagation of repressive histone marks. *Nature*, 461(7265):762–767, Oct. 2009.

[36] J. Marsolier, P. Prompsy, A. Durand, A.-M. Lyne, C. Landragin, A. Trouchet, S. T. Bento, A. Eisele, S. Foulon, L. Baudre, K. Grosselin, M. Bohec, S. Baulande, A. Dahmani, L. Sourd, E. Letouzé, A.-V. Salomon, E. Marangoni, L. Perié, and C. Vallot. H3K27me3 conditions chemotolerance in triple-negative breast cancer. *Nature Genetics*, 54(4):459–468, Apr. 2022.

[37] L. McInnes, J. Healy, N. Saul, and L. Großberger. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861, Sept. 2018.

[38] Z. Miao, K. Deng, X. Wang, and X. Zhang. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics*, 34(18):3223–3224, Sept. 2018.

[39] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers. Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468)*, pages 41–48. Ieee, 1999.

[40] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017. arXiv: 1605.09522.

[41] S. Mukherjee, D. Agarwal, N. R. Zhang, and B. B. Bhattacharya. Distribution-Free Multisample Tests Based on Optimal Matchings With Applications to Single Cell Genomics. *Journal of the American Statistical Association*, 117(538):627–638, Apr. 2022.

[42] S. Pott and J. D. Lieb. Single-cell ATAC-seq: strength in numbers. *Genome Biology*, 16(1):172, Aug. 2015.

[43] P. A. Reyfman, J. M. Walter, N. Joshi, K. R. Anekalla, A. C. McQuattie-Pimentel, S. Chiu, R. Fernandez, M. Akbarpour, C.-I. Chen, Z. Ren, R. Verma, H. Abdala-Valencia, K. Nam, M. Chi, S. Han, F. J. Gonzalez-Gonzalez, S. Soberanes, S. Watanabe, K. J. N. Williams, A. S. Flozak, T. T. Nicholson, V. K. Morgan, D. R. Winter, M. Hinchcliff, C. L. Hrusch, R. D. Guzy, C. A. Bonham, A. I. Sperling, R. Bag, R. B. Hamanaka, G. M. Mutlu, A. V. Yeldandi, S. A. Marshall, A. Shilatifard, L. A. N. Amaral, H. Perlman, J. I. Sznajder, A. C. Argento, C. T. Gillespie, J. Dematte, M. Jain, B. D. Singer, K. M. Ridge, A. P. Lam, A. Bharat, S. M. Bhorade, C. J. Gottardi, G. R. S. Budinger, and A. V. Misharin. Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine*, 199(12):1517–1536, June 2019.

[44] A. Richard, L. Boullu, U. Herbach, A. Bonnafoux, V. Morin, E. Vallin, A. Guillemin, N. Papili Gao, R. Gunawan, J. Cosette, O. Arnaud, J.-J. Kupiec, T. Espinasse, S. Gonin-Giraud, and O. Gandrillon. Single-Cell-Based Analysis Highlights a Surge in Cell-to-Cell Molecular Variability Preceding Irreversible Commitment in a Differentiation Process. *PLOS Biology*, 14(12):e1002585, Dec. 2016.

[45] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, Apr. 2015.

[46] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, Jan. 2010.

[47] A. Rotem, O. Ram, N. Shoresh, R. A. Sperling, A. Goren, D. A. Weitz, and B. E. Bernstein. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology*, 33(11):1165–1172, Nov. 2015.

[48] R. Schefzik, J. Flesch, and A. Goncalves. Fast identification of differential distributions in single-cell RNA-sequencing data with waddR. *Bioinformatics*, 37(19):3204–3211, Oct. 2021.

[49] A. Schrab, I. Kim, M. Albert, B. Laurent, B. Guedj, and A. Gretton. MMD Aggregated Two-Sample Test, June 2022. arXiv:2110.15073 [cs, math, stat].

[50] J. Shawe-Taylor and N. Cristianini. Kernel Methods for Pattern Analysis, June 2004. Cambridge University Press, New York, NY, USA.

[51] E. Shema, B. E. Bernstein, and J. D. Buenrostro. Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. *Nature Genetics*, 51(1):19–25, Jan. 2019.

[52] C.-J. Simon-Gabriel and B. Schölkopf. Kernel Distribution Embeddings: Universal Kernels, Characteristic Kernels and Kernel Metrics on Distributions. *Journal of Machine Learning Research*, 19(44):1–29, 2018.

[53] J. W. Squair, M. Gautier, C. Kathe, M. A. Anderson, N. D. James, T. H. Hutson, R. Hudelle, T. Qaiser, K. J. E. Matson, Q. Barraud, A. J. Levine, G. La Manno, M. A. Skinnider, and G. Courtine. Confronting false discoveries in single-cell differential expression. *Nature Communications*, 12(1):5692, Sept. 2021.

[54] S. Tiberi, H. L. Crowell, P. Samartsidis, L. M. Weber, and M. D. Robinson. distinct: a novel approach to differential distribution analyses, Apr. 2022. bioRxiv doi: 10.1101/2020.11.24.394213.

[55] H. Van Assel, T. Espinasse, J. Chiquet, and F. Picard. A Probabilistic Graph Coupling View of Dimension Reduction. *Advances in Neural Information Processing Systems*, 35:10696–10708, Dec. 2022.

[56] T. Wang and S. Nabavi. SigEMD: A powerful method for differential gene expression analysis in single-cell RNA sequencing data. *Methods*, 145:25–32, Aug. 2018.

[57] C. K. I. Williams and M. Seeger. Using the Nystrom Method to Speed Up Kernel Machines. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 682–688. MIT Press, 2001.

[58] G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, Jan. 2017.

[59] S. Zreika, C. Fourneaux, E. Vallin, L. Modolo, R. Seraphin, A. Moussy, E. Ventre, M. Bouvier, A. Ozier-Lafontaine, A. Bonnaffoux, F. Picard, O. Gandrillon, and S. Gonin-Giraud. Evidence for close molecular proximity between reverting and undifferentiated cells. *BMC Biology*, 20(1):155, July 2022.