Kernel-based testing for single-cell differential analysis

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

†B. Michel and F. Picard are joint last authors.

*Correspondence:
anthony.ozier-lafontaine@ec-nantes.fr; Bertrand.Michel@ec-nantes.fr; franck.picard@ens-lyon.fr

[1] Nantes Université, Centrale Nantes, Laboratoire de Mathématiques Jean Leray, CNRS UMR 6629, 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

Abstract

Single-cell technologies offer insights into molecular feature distributions, but comparing them poses challenges. We propose a kernel-testing framework for non-linear cell-wise distribution comparison, analyzing gene expression and epigenomic modifications. Our method allows feature-wise and global transcriptome/epigenome comparisons, revealing cell population heterogeneities. Using a classifier based on embedding variability, we identify transitions in cell states, overcoming limitations of traditional single-cell analysis. Applied to single-cell ChIP-Seq data, our approach identifies untreated breast cancer cells with an epigenomic profile resembling persister cells. This demonstrates the effectiveness of kernel testing in uncovering subtle population variations that might be missed by other methods.

Keywords: Single cell transcriptomics, Single cell epigenomics, Differential analysis, Kernel methods

Background

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 [25, 33, 56], chromatin states [45], and genomic variations [14] 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

Ozier-Lafontaine *et al. Genome Biology* (2024) 25:114

Page 2 of 21

accessibility assays (scATAC-Seq [40]) or single-cell histone modifications profiling (e.g., single-cell ChIP-Seq (scChIP-Seq) [18], scCUT &Tag [4]). These approaches enable the mapping of chromatin states throughout the genome and their cell-to-cell variations at an unprecedented resolution [6, 49]. 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[35]. The identification of key epigenomic features by differential analysis in disease and complex ecosystems will be key to understand regulatory principles of gene expression and identify potential drivers of tumor progression. Altogether, comparative analysis of single-cell datasets, 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 [31, 43, 44]. 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 [28]. Initially, these tests were performed using Gaussian-based clustering that was further challenged by distribution-free methods based on ranks or cumulative distribution functions [13, 46, 53]. 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 that can still account for certain characteristics of the data, such as a potentially high proportion of zeros.

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 [34]. 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 [39]. So far, global approaches were mainly developed for differential abundance testing [7, 9, 10], or for the comparison of cell-type compositions.

Ozier-Lafontaine *et al. Genome Biology*   (2024) 25:114

Page 3 of 21

Graph-based methods have been proposed to address differential transcriptome analysis [3, 39], 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. [15], who combined kernel methods with statistical testing. Kernel methods are widely used in supervised learning [48] 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 [32, 36], also use kernel-based embedding [54]. 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 [38], 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 the Kernel Fisher discriminant analysis (KFDA) approach introduced by [24]. KFDA is 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. Although KFDA was initially introduced as a non-linear classifier [37], it is a great example of how classifiers can be used for hypothesis testing [22, 30], and recent developments have demonstrated its optimality [21]. Here, we show that the KFDA-witness function, which is the Fisher discriminant axis [29], 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. While originally designed for a two-sample framework, our method can be extended to accommodate multiple group comparisons. Furthermore, we discuss its applicability and extension to more complex experimental designs. We show the calibration and the power of our method compared with others on simulated [13] and multiple scRNA-Seq datasets [51]. 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 [57]. Our method also uncovers the epigenomic heterogeneity of breast cancer cells, revealing the pre-existence——prior to cancer treatment——of cells epigenomically identical to drug-persister cells, i.e., the rare cells that can survive treatment.

---

[1] https://github.com/LMJL-Alea/ktest

Ozier-Lafontaine *et al. Genome Biology*   (2024) 25:114

Page 4 of 21

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.

Results

In the following, we denote by $Y_1 = (Y_{1,1}, \ldots, Y_{1,n_1})$ and $Y_2 = (Y_{2,1}, \ldots, Y_{2,n_2})$ 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

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

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 $(\phi(Y_{i,1}), \ldots, \phi(Y_{i,n_i}))$ $(i = 1, 2)$, 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:

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

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 $\mathbb{P}_2$ [38], defined such that:

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

The initial contribution to kernel testing involved calculating the distance between kernel mean embeddings with the MMD statistic [16]. 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 [21]. 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 [22], 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

Ozier-Lafontaine *et al. Genome Biology*   (2024) 25:114

Page 5 of 21

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.

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:

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

This statistic tests the between-class separation by comparing expected pairwise similarities between and within conditions 1 and 2 (a full derivation is proposed in Additional file 1: Supplementary Material). To account for the residual variability, we introduce the weighted Mahalanobis distance between mean embeddings,

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

where $\Sigma_{W,T}$ contains the first T principal directions of the homogeneous within-group covariance of embeddings defined such as:

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

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). Regularization is indeed necessary to prevent the singularity of Σ_W. One potential approach is to introduce ridge regularization; however, this leads to a complex distribution of the test statistic under the null hypothesis, with limited interpretability [23]. An alternative regularization strategy consists in considering $\Sigma_{W,T}$ which involves a kernel-PCA dimension-reduction step to capture the residual variability of expression data centered by condition. Then, the corresponding regularized statistic is based on the estimated mean embeddings and covariances:

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

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 :

Ozier-Lafontaine *et al. Genome Biology*   (2024) 25:114

Page 6 of 21

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

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 min for $n \sim 4000$, and the package proposes a sampling-based Nystrom approximation for larger sample sizes [55].

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 :

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

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 $h_T^\star$:

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

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 [28]. 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) [28]. 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 [13]. The discriminant axis acts as a powerful non-linear transformation of the expression data to make the two distributions easily separable (Fig. 1). For the sake of simplicity, we presented our method in the two-sample setting, but we also propose a generalization to multiple groups comparisons provided in Additional file 1: Supplementary Material.

Kernel choice

The design of appropriate kernels is an active field of research [2, 47]. 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 [17, 47, 50]. 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 :

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

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 that consists in choosing $\widehat{\sigma}^2 = \text{median}\left(\sum_g \left(Y_{i,j}^g - Y_{i',j'}^g\right)^2, (i,i') \in \{1,2\}^2, j \in \{1, \ldots, n_i\}, j' \in \{1, \ldots, n_{i'}\}\right)$ [12, 17, 47]. Depending on the sequencing technology [52], scRNA-Seq data may contain a fraction of zeros (especially for non-UMI data like Smart-Seq, for instance), which could impact the calibration of the kernel's bandwidth if not properly considered. Therefore,
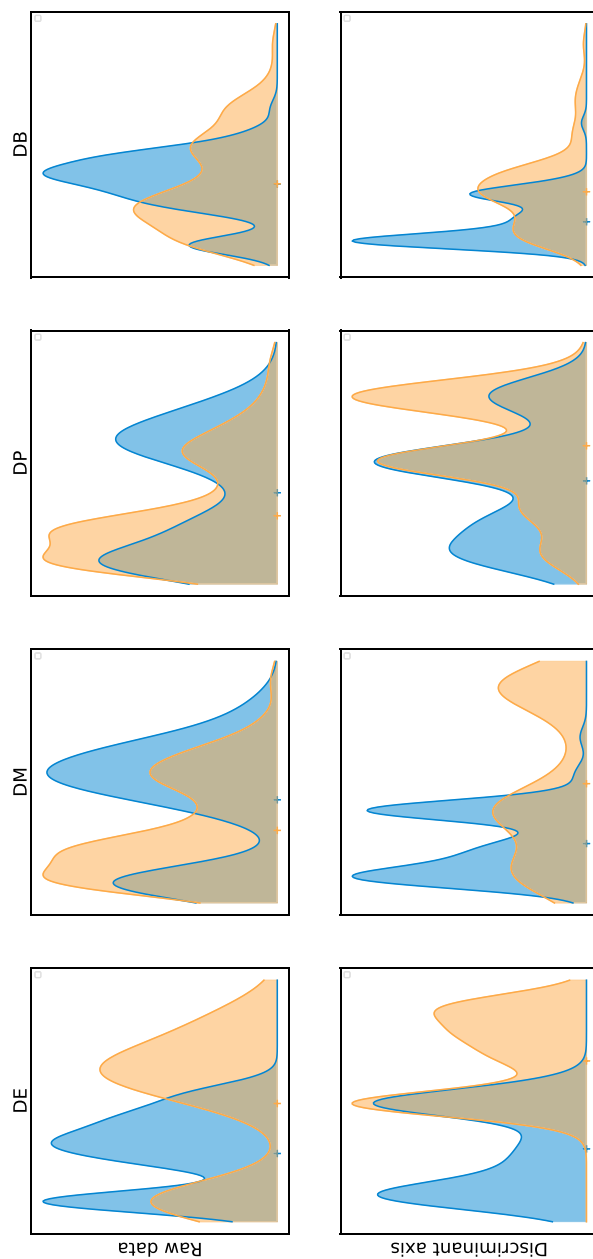
Fig. 1 Top: Examples of distributions of the simulated data, DE, classical difference in expression; DM, difference in modalities; DP, difference in proportions; DB, difference in both modalities and proportions with equal means. Bottom: Projection of cells on the discriminant axis ($T = 4$) for each alternative. The non-linear transform allows the separation of distributions on the discriminant axis

Ozier-Lafontaine *et al. Genome Biology*   (2024) 25:114

Page 9 of 21

we propose a two-compartment kernel based on probability product kernels [26]. Let π_i represent the proportion of zeros in condition i, and $f_{\mu,\sigma}$ denote the Gaussian probability function. We introduce a zero-inflated Gaussian kernel (details in the "Methods" section):

$$k_{\text{ZI-Gauss}}(Y_{i,j}, Y_{i',j'}) = \pi_i \pi_{i'} + \pi_i(1 - \pi_{i'})f_{\mu_{i'},\sigma}(0) + (1 - \pi_i)\pi_{i'}f_{\mu_i,\sigma}(0)$$
$$+ (1 - \pi_i)(1 - \pi_{i'})k_{\text{Gauss}}(Y_{i,j}, Y_{i',j'}),$$

so that the bandwidth is calibrated on non-zero entries only. Finally, in our method comparisons, we will explore the `ktest` framework with a linear kernel to highlight the advantages of non-linearity. For this illustration, we consider the standard scalar product:

$$k_{\text{linear}}(Y_{i,j}, Y_{i',j'}) = \sum_{g=1}^{G} Y_{i,j}^g \times Y_{i',j'}^g.$$

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. We challenged our kernel-based test with six standard DEA methods (Table S1) on mixtures of zero-inflated negative binomial data reproducing the DE, DM, DP, and DB alternatives [13] (as detailed in Material and Methods). Kernel testing was performed on the raw data using the Gauss and ZI-Gauss kernels, but we also considered the linear kernel (scalar product) to illustrate the interest of a non-linear method. 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 Gauss-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). This is well illustrated by the global lack of power of the test based on the linear kernel (especially on the DB alternative). The Gaussian kernel 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 ZI-Gauss 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. We also checked that the median heuristic was a reasonable choice for the bandwidth parameter (Fig. S2), as it established a good type I/power trade-off. Note that when the bandwidth of the Gaussian kernel increases, the truncation parameter should be calibrated accordingly to reach the same type I/power performance.
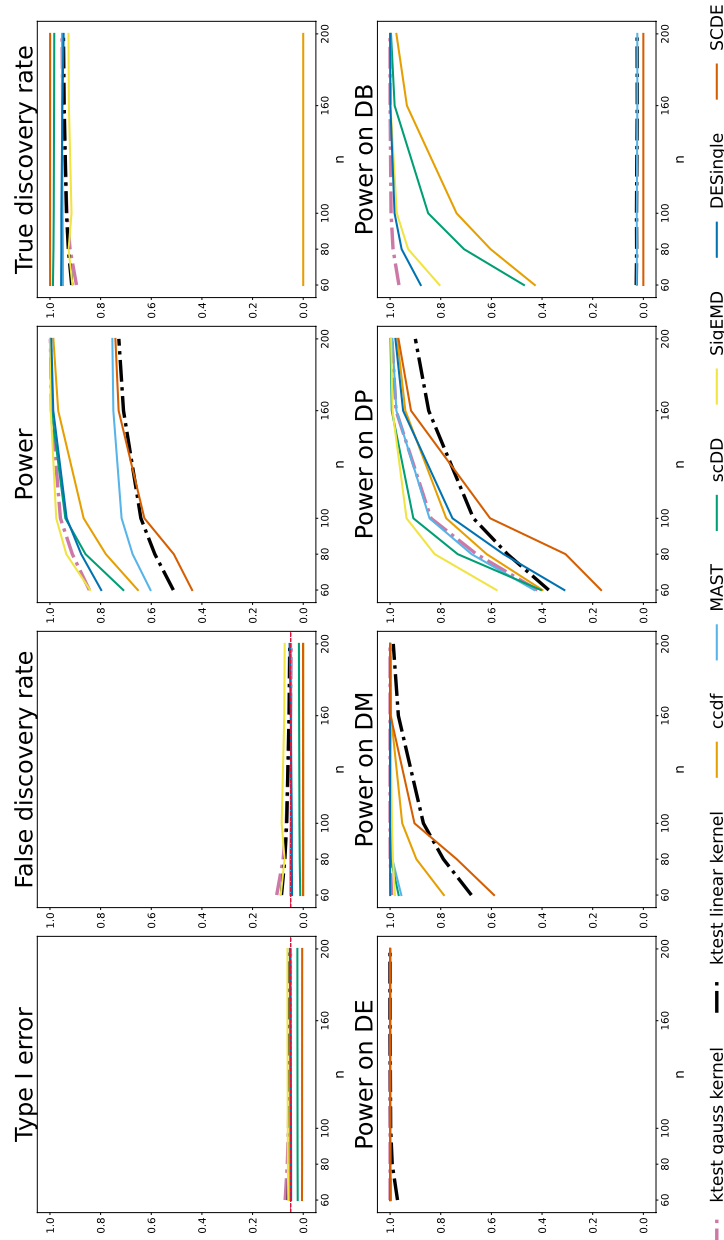
Ozier-Lafontaine *et al. Genome Biology*   *(2024) 25:114*

Page 10 of 21

Fig. 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. The truncation parameter is set to $T = 4$ for the Gauss-kernel

Ozier-Lafontaine *et al. Genome Biology*   (2024) 25:114

Page 11 of 21

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. [51], which compared 14 DE analysis methods (Table S2) 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 RNA-Seq 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 [28]. 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 are clustered together because they are trained on scRNA-Seq data. Kernel testing shows performance close to single-cell methods. Kernel testing emerges as a third approach, aligning more closely with single-cell methods. Its top differentially expressed (DE) genes exhibit characteristics akin to those of pseudo-bulk methods in terms of average expression and the proportion of zeros. Notably, kernel testing diverges from other single-cell DEA methods, which typically identify highly-expressed genes, as highlighted in the original study (Fig. 3, bottom). It is noteworthy that when the kernel method employs a linear kernel, its performances are close to those of the *t*-test and likelihood-ratio test, illustrating the interest of a non-linear procedure. By inspecting the distributional changes associated to genes considered as false-positive in the original study (with bulk-RNA-Seq 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, [28]) and were thus undetectable from bulk-RNA-Seq data and pseudo-bulk methods (Fig. S3, 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 ZI-Gauss kernel (examples of such distribution shapes are shown in Fig. S3, right). Finally, we compared the computational time of competing methods, illustrating the quadratic complexity of `ktest` (Fig. S3), which still remains reasonable for complete transcriptomes.

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
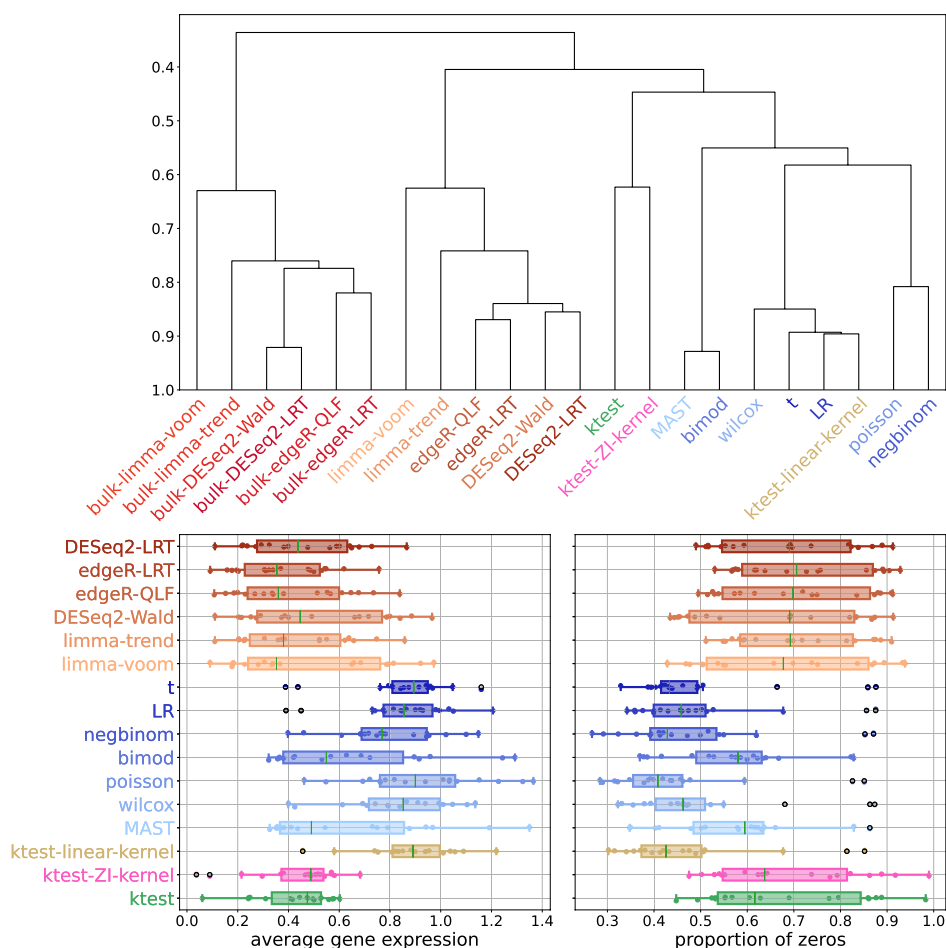
Fig. 3 Top: Hierarchical clustering based on average AUCC scores computed between pairs of methods (over 18 datasets [51]). Bottom: Boxplot of the average expression (left) and proportion of zeros (right) of the top 500 DE genes for different DE methods (over 18 datasets [51]). Red: bulk methods, orange: pseudo-bulk methods, blue: single-cell methods. The truncation parameter is set to $T = 4$ for `ktest` (only univariate tests were performed)

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 [42]. A second study investigated the molecular mechanisms underlying cell differentiation and reversion by measuring cell transcriptomes at four time points (Fig. 4a): undifferentiated T2EC maintained in a self-renewal medium (0H), then put in a differentiation-inducing medium for 24 h (24H). The population was then split into a first population maintained in the same medium for 24 h to achieve differentiation (48HDIFF); the second population was put back in the self-renewal medium to investigate potential reversion (48HREV) [57]. Cell transcriptomes were measured using scRT-qPCR on 83 genes selected to be involved in the differentiation process as
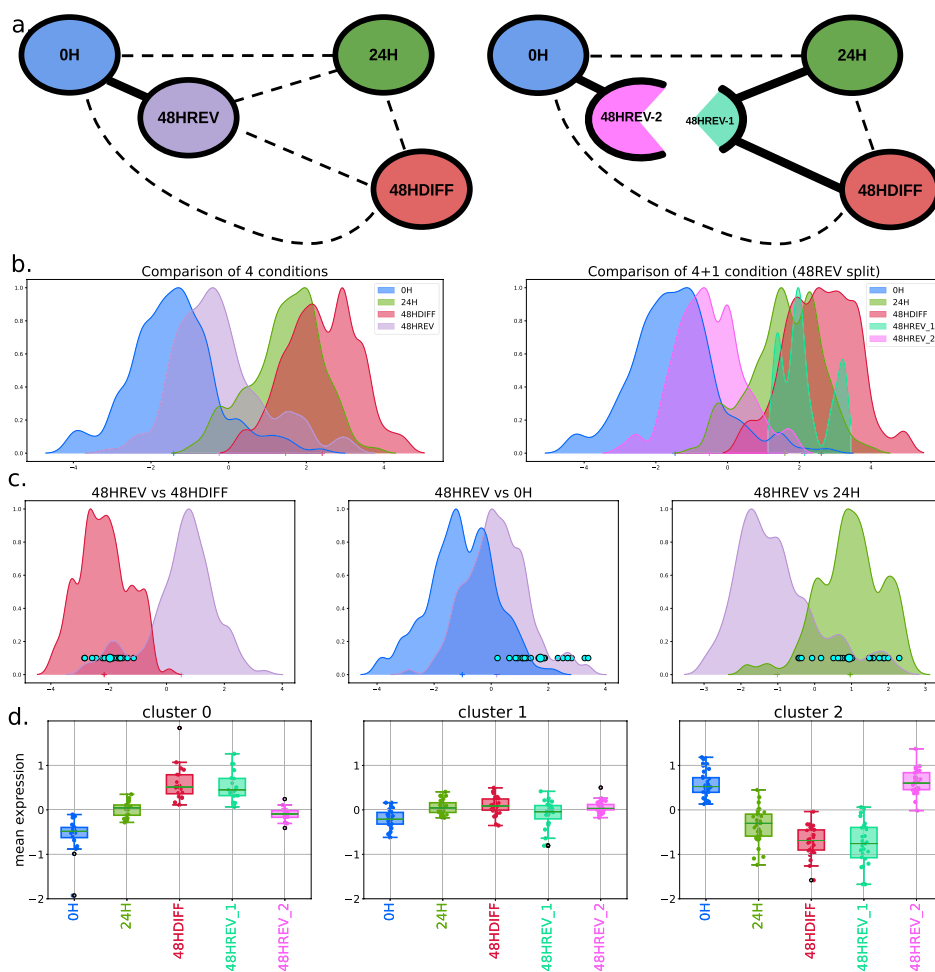
Fig. 4 a Summarized distance graphs between conditions before (left) and after (right) splitting condition 48HREV into populations 48HREV-1 and 48HREV-2. **b** Cell densities of all compared conditions, before (left) and after (right) splitting condition 48HREV **c** Cell densities of compared conditions 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 variation of the gene expression along 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. The multivariate differential expression analysis was performed with $T = 10$

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.

Since the experiments were conducted on eight independent batches, our analysis began by assessing the significance of the batch effect using the multigroup kernel-based test. Both scRT-qPCR and scRNA-Seq data exhibited a significant effect (*p*-values of 3.18×10^{-78} and 1.26×10^{-85}, respectively). To address this, we corrected the data embedding by applying the mean embedding of the batch effect, resulting in a non-linear normalization with respect to the batch (details in Additional file 1: Supplementary

Material). Then, we conducted a new test to compare the batch-corrected distribution of gene expressions between biological conditions (differentiation time). The multi-group kernel test first confirmed heterogeneity among conditions in both scRT-qPCR and scRNA-Seq (*p*-values of 0 and 3.64×10^{-142}, respectively). The 4-group discriminant analysis yielded three discriminant axes that represent the global heterogeneities of the data. Notably, the first discriminant axis associated with the global 4-group comparison ordered the four conditions according to the time of differentiation (Figs. 4b and S6), while subsequent axes provided less pronounced information (Fig. S5). We then employed `ktest` for pair-wise comparisons between conditions, confirming a significant difference between undifferentiated cells (0H) and reverted cells (48HREV) in both scRT-qPCR and scRNA-Seq data (*p*-values of 4.55×10^{-23} and 7.39×10^{-06}, respectively). However, considering the test statistic as a distance also confirmed the close proximity between these two conditions (Figs. S4 and S5).

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. S7, 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. 4c). Interestingly, these cells also matched the distribution summary of 24H-cells (48HREV vs. 24H, Fig. 4c) and were separated from the undifferentiated cells (48HREV vs 0H, Fig. 4c). A similar sub-population was detected using scRNA-Seq data (48HREV vs. 48HDIFF Fig. S6b). 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 $4.73\ 10^{-5}$ and 0.80 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 [57]. 48HREV-2 cells (48HREV cells after removing 48HREV-1 cells) were closer but still significantly different from 0H cells in both technologies (*p*-values $4.48\ 10^{-17}$ and $3.98\ 10^{-05}$ respectively). To describe these two sub-populations in terms of genes, we performed a *k*-means clustering on the averaged centered expressions of genes over cells in populations 0H, 24H, 48HDIFF, 48HREV-1, and 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 (Figs. 4d and S6c): (*i*) genes activated during differentiation (scRT-qPCR cluster 0, scRNA-Seq clusters 2 and 3), e.g., hemoglobin related genes such as *HBA1* and *HBAD* (shown in Fig. S6d); (*ii*) genes deactivated during differentiation (scRT-qPCR cluster 2, scRNA-Seq cluster 0), e.g., genes involved in metabolism of self-renewing cells such as *LDHA* and *LY6E* (shown in Fig. S6d); and (*iii*) genes with no clear function pattern for which the expression levels did not change much during differentiation and reversion (scRT-qPCR cluster 1 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].

---

[2] https://github.com/AnthoOzier/ktest_experiment_genome_biology_2024

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.

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 [34]——can induce complex dependencies between sites occupancy. A recent study [35] 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 Fig. 5. Projecting cells on the kernelized discriminant axis reveals 3 sub-populations within the untreated cell population: persister-like (109 cells; 5% of untreated cells), intermediate (1124 cells; 57%), and 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 6376 tested regions, only
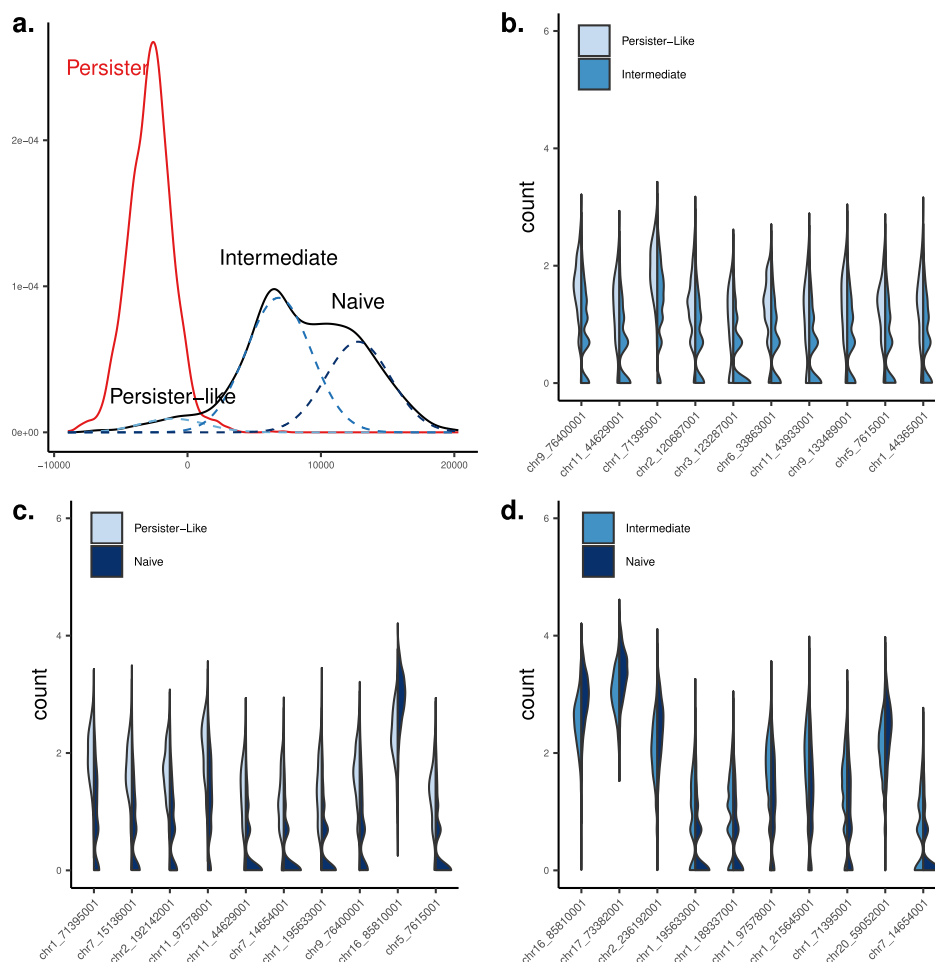
Fig. 5 Differential analysis of scChIP-Seq data on breast cancer cells. **a** Cell densities of persister cells vs. untreated cells. Sub-populations of untreated cells were identified using 3-component mixture model, that revealed persister-like cells, intermediate, and naive cells. **b–d** violin plots 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 S3. Multivariate (**a**) and univariate analyses (**b–d**) were performed with $T = 5$

14 were significantly differentially enriched (p-value$< 10^{-3}$, Table S3), 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.

Conclusions

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 [27] 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 [28]. 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. We also demonstrate the applicability of kernel-testing for multiple group comparisons and two-factor designs. Our plan is to fully develop this approach, providing a comprehensive mathematical framework that facilitates the study of any complex design, including model validation and contrast testing, for instance. 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.

Methods

Simulation settings

The comparison study on data simulated was performed on data following different mixtures of zero inflated negative binomial (ZINB) distributions [13]. 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]. Compared methods are shown in Table S1.

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 [51]. They consists of six comparisons of bone marrow mononuclear phagocytes from mouse, rat,

pig, and rabbit in different conditions [20], 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 [41]. More details on the datasets are in [51] 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 [51]. This not very restrictive preprocessing was chosen in order to not introduce biases 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 [51].

Zero-inflated Gaussian kernel

Our method is non-parametric, meaning we do not assume a specific distribution for the data. In this context, we propose to derive a kernel that is tailored to a high proportion of zeros. To achieve this, we propose to develop the zero-inflated Gauss kernel, which involves considering a zero-inflated Gaussian distribution with π the proportion of additional zeros:

$$X \sim \pi \delta_0(\bullet) + (1 - \pi) f_{\mu,\sigma}(\bullet),$$

with $f_{\mu,\sigma}$ the Gaussian probability density function. It is important to note that this does not imply that we assume the data to follow a zero-inflated Gaussian distribution. This representation serves merely as a methodological framework for deriving the new kernel. This distribution has a mixture representation, with Z standing for the binary variable of distribution $\mathcal{B}(\pi)$, such that

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

We know the probability kernels for the Gaussian part of the model:

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

and for the Bernoulli distribution:

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

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

$$k_{\text{ZI-Gauss}}(f_{\mu,\sigma,\pi}, f_{\mu',\sigma,\pi'}) = \int_x \left(\sum_z \left(\mathbb{P}_\pi(Z = z) f_{\mu,\sigma,\pi}(x \mid Z = z) \right) \right) \left(\sum_z \left(\mathbb{P}_{\pi'}(Z = z) f_{\mu',\sigma,\pi'}(x \mid Z = z) \right) \right) dx$$

$$= \pi\pi' + \pi(1 - \pi') f_{\mu',\sigma}(0) + (1 - \pi)\pi' f_{\mu,\sigma}(0) + (1 - \pi)(1 - \pi') k_{\text{Gauss}}(\mu, \mu'),$$

In the simulations, the ZI-Gauss 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.

Ozier-Lafontaine *et al. Genome Biology*   (2024) 25:114

Page 19 of 21

Reversion data

Details on the experiment and on the data can be found in the original paper [57]. 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` [19]. For both datasets, we corrected for the batch effect in the feature space. The gene clusters were computed on the data after correcting for the batch effect in the input space. 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.

sc-chIP-Seq data

Single-cell chIP-Seq data correspond to a count matrix of unique reads mapping to the genome binned into H3K27me3 previously identified peaks [35]. This matrix was filtered for cells with a minimum coverage of 3,000 unique reads and a maximum coverage of 10,000 reads. Top 5% covered cells were furthered filtered out, as potential doublets.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03255-1.

Additional file 1: Supplementary Material.

Additional file 2: Table S2. DEA methods compared on the sc-RNASeq datasets (Fig. 3). **Table S1.** DEA methods compared in the simulation study (Fig. 1). **Table S3.** 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 < 0.001 (Bonferroni correction). The last Gene column corresponds to the genes overlapping the regions.

Additional file 3: Review history.

Availability of data and materials
The data used to compare methods are available from the Zenodo repository (https://doi.org/10.5281/zenodo.5048449) as compiled by Squair and coauthors [51]. Reversion scRT-qPCR data are available in the SRA repository number SRP076011 and fully described in the original publication [57]. Single-cell chIP-Seq data can be found on GEO with the accession number GSE164385 [35]. Our code and material are available on HAL https://hal.science/hal-04547380 and Zenodo https://doi.org/10.5281/zenodo.10974453 under a CC BY 4.0 license.

Declarations

Peer review information
Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history
The review history is available as Additional file 3.

References

1. Angelidis I, Simon LM, Fernandez IE, Strunz M, Mayr CH, Greiffo FR, Tsitsiridis G, Ansari M, Graf E, Strom T-M, Nagendran M, Desai T, Eickelberg O, Mann M, Theis FJ, Schiller HB. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. Nat Commun. 2019;10(1):963. Number: 1 Publisher: Nature Publishing Group.
2. Bach FR, Lanckriet GRG, Jordan MI. Multiple kernel learning, conic duality, and the SMO algorithm. In: Proceedings of the twenty-first international conference on machine learning, ICML '04. New York: Association for Computing Machinery; 2004. p. 6
3. Banerjee T, Bhattacharya BB, Mukherjee G. A nearest-neighbor based nonparametric test for viral remodeling in heterogeneous single-cell proteomic data. Ann Appl Stat. 2020;14(4):1777–805.
4. Bartosovic M, Kabbe M, Castelo-Branco G. Single-cell CUT &Tag profiles histone modifications and transcription factors in complex tissues. Nat Biotechnol. 2021;39(7):825–35.
5. Benjamini et Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing on JSTOR. 1995.
6. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015;523(7561):486–90.
7. Büttner M, Ostner J, Müller CL, Theis FJ, Schubert B. scCODA is a Bayesian model for compositional single-cell data analysis. Nat Commun. 2021;12(1):6876. Number: 1 Publisher: Nature Publishing Group.
8. Cano-Gamez E, Soskic B, Roumeliotis TI, So E, Smyth DJ, Baldrighi M, et al. Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4+ T cells to cytokines. Nat Commun. 2020;11(1):1801.
9. Cao Y, Lin Y, Ormerod JT, Yang P, Yang JY, Lo KK. scDC: single cell differential composition analysis. BMC Bioinformatics. 2019;20(19):721.
10. Dann E, Henderson NC, Teichmann SA, Morgan MD, Marioni JC. Differential abundance testing on single-cell data using k-nearest neighbor graphs. Nat Biotechnol. 2022;40(2):245–53.
11. Das S, Rai A, Rai SN. Differential expression analysis of single-cell RNA-Seq data: current statistical approaches and outstanding challenges. Entropy (Basel, Switzerland). 2022;24(7):995.
12. Garreau D, Jitkrittum W, Kanagawa M. Large sample analysis of the median heuristic. 2018. arXiv preprint arXiv:1707.07269.
13. Gauthier M, Agniel D, Thiébaut R, Hejblum BP. Distribution-free complex hypothesis testing for single-cell RNA-seq differential expression analysis. bioRxiv 2021.05.21.445165 (2021). https://doi.org/10.1101/2021.05.21.445165.
14. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nat Rev Genet. 2016;17(3):175–88.
15. Gretton A, Borgwardt K, Rasch M, Schölkopf B, Smola A. A kernel method for the two-sample-problem. In: Advances in Neural Information Processing Systems, vol. 19. Cambridge: MIT Press; 2006. p. 513–20.
16. Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A. A kernel two-sample test. J Mach Learn Res. 2012;13(25):723–73.
17. Gretton A, Sriperumbudur B, Sejdinovic D, Strathmann H, Balakrishnan S, Pontil M, et al. Optimal kernel choice for large-scale two-sample tests. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12). Red Hook, NY: Curran Associates Inc.; 2012. p. 1205–13.
18. Grosselin K, Durand A, Marsolier J, Poitou A, Marangoni E, Nemati F, et al. High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. Nat Genet. 2019;51(6):1060–6.
19. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019;20(1):296.
20. Hagai T, Chen X, Miragaia RJ, Rostom R, Gomes T, Kunowska N, et al. Gene expression variability across cells and species shapes innate immunity. Nature. 2018;563(7730):197–202.
21. Hagrass O, Sriperumbudur BK, Li B. Spectral regularized kernel two-sample tests. 2022. arXiv:2212.09201 [cs, math, stat].
22. Harchaoui Z, Bach F, Cappe O, Moulines E. Kernel-based methods for hypothesis testing: a unified view. IEEE Signal Process Mag. 2013;30(4):87–97.
23. Harchaoui Z, Bach FR, Moulines E. Testing for homogeneity with kernel fisher discriminant analysis. Stat. 2008;1050:7.
24. Harchaoui Z, Vallet F, Lung-Yut-Fong A, Cappe O. A regularized kernel-based approach to unsupervised audio segmentation. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. Taipei: IEEE; 2009. pp. 1665–8
25. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively parallel single cell RNA-Seq for marker-free decomposition of tissues into cell types. Science (New York, N.Y.). 2014;343(6172):776–9.
26. Jebara T, Kondor R, Howard A. Probability product kernels. J Mach Learn Res. 2004;5(Jul):819–44.

Ozier-Lafontaine *et al. Genome Biology*   (2024) 25:114

Page 21 of 21

27. Kim I, Ramdas A, Singh A, Wasserman L. Classification accuracy as a proxy for two-sample testing. Ann Stat. 2021;49(1):411–34.
28. Korthauer KD, Chu L-F, Newton MA, Li Y, Thomson J, Stewart R, Kendziorski C. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biol. 2016;17(1):222.
29. 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. PMLR; 2022. pp. 1403–19. ISSN: 2640-3498.
30. Lopez-Paz D, Oquab M. Revisiting classifier two-sample tests. 2018. arXiv preprint arXiv:1610.06545.
31. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
32. Maaten LVD, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(86):2579–605.
33. Macosko E, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161(5):1202–14.
34. Margueron R, Justin N, Ohno K, Sharpe ML, Son J, Drury WJ, et al. Role of the polycomb protein Eed in the propagation of repressive histone marks. Nature. 2009;461(7265):762–7.
35. Marsolier J, Prompsy P, Durand A, Lyne A-M, Landragin C, Trouchet A, et al. H3K27me3 conditions chemotolerance in triple-negative breast cancer. Nat Genet. 2022;54(4):459–68.
36. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. J Open Source Softw. 2018;3(29):861.
37. Mika S, Ratsch G, Weston J, Scholkopf B, Mullers KR. Fisher discriminant analysis with kernels. In: Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop, Madison, 23–25 August. Piscataway: IEEE; 1999. p. 41–8.
38. Muandet K, Fukumizu K, Sriperumbudur B, Schölkopf B. Kernel mean embedding of distributions: a review and beyond. Found Trends® Mach Learn. 2017;10(1-2):1–141. arXiv: 1605.09522.
39. Mukherjee S, Agarwal D, Zhang NR, Bhattacharya BB. Distribution-free multisample tests based on optimal matchings with applications to single cell genomics. J Am Stat Assoc. 2022;117(538):627–38.
40. Pott S, Lieb JD. Single-cell ATAC-seq: strength in numbers. Genome Biol. 2015;16(1):172.
41. Reyfman PA, Walter JM, Joshi N, Anekalla KR, McQuattie-Pimentel AC, Chiu S, et al. Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. Am J Respir Crit Care Med. 2019;199(12):1517–36.
42. Richard A, Boullu L, Herbach U, Bonnafoux A, Morin V, Vallin E, et al. Single-cell-based analysis highlights a surge in cell-to-cell molecular variability preceding irreversible commitment in a differentiation process. PLoS Biol. 2016;14(12):e1002585.
43. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
44. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
45. Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, Weitz DA, Bernstein BE. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nat Biotechnol. 2015;33(11):1165–72.
46. Schefzik R, Flesch J, Goncalves A. Fast identification of differential distributions in single-cell RNA-sequencing data with waddR. Bioinformatics. 2021;37(19):3204–11.
47. Schrab A, Kim I, Albert M, Laurent B, Guedj B, Gretton A. MMD aggregated two-sample test. 2022. arXiv preprint arXiv:2110.15073.
48. Shawe-Taylor J, Cristianini N. Kernel methods for pattern analysis. New York: Cambridge University Press; 2004.
49. Shema E, Bernstein BE, Buenrostro JD. Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. Nat Genet. 2019;51(1):19–25.
50. Simon-Gabriel C-J, Schölkopf B. Kernel distribution embeddings: universal kernels, characteristic kernels and kernel metrics on distributions. J Mach Learn Res. 2018;19(44):1–29.
51. Squair JW, Gautier M, Kathe C, Anderson MA, James ND, Hutson TH, et al. Confronting false discoveries in single-cell differential expression. Nat Commun. 2021;12(1):5692.
52. Svensson V. Droplet scRNA-seq is not zero-inflated. Nat Biotechnol. 2020;38(2):147–50. Number: 2 Publisher: Nature Publishing Group.
53. Tiberi S, Crowell HL, Samartsidis P, Weber LM, Robinson MD. distinct: a novel approach to differential distribution analyses. Ann Appl Stat. 2023;17(2):1681–700.
54. Van Assel H, Espinasse T, Chiquet J, Picard F. A probabilistic graph coupling view of dimension reduction. Adv Neural Inf Process Syst. 2022;35:10696–708.
55. Williams CKI, Seeger M. Using the Nystrom method to speed up kernel machines. In: Leen TK, Dietterich TG, Tresp V, editors. Advances in Neural Information Processing Systems 13. Cambridge: MIT Press; 2001. p. 682–8.
56. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049.
57. Zreika S, Fourneaux C, Vallin E, Modolo L, Seraphin R, Moussy A, et al. Evidence for close molecular proximity between reverting and undifferentiated cells. BMC Biol. 2022;20(1):155.

Publisher's Note