Distribution-free complex hypothesis testing for single-cell RNA-seq differential expression analysis

Marine GAUTHIER [1,2,*], Denis AGNIEL [3,4], Rodolphe THIÉBAUT [1,2,5],

Boris P. HEJBLUM [1,2]

[1] *University of Bordeaux, INSERM U1219 Bordeaux Population Health Research Center, INRIA SISTM, F-33000 Bordeaux, France.* [2] *Vaccine Research Institute, F-94000 Créteil, France.* [3] *Rand Corporation, Santa Monica (CA), USA.* [4] *Harvard Medical School, Boston (MA), USA.* [5] *CHU de Bordeaux, Bordeaux, F-33000 France.*

marine.gauthier@u-bordeaux.fr

Summary

State-of-the-art methods for single-cell RNA-seq (scRNA-seq) Differential Expression Analysis (DEA) often rely on strong distributional assumptions that are difficult to verify in practice. Furthermore, while the increasing complexity of clinical and biological single-cell studies calls for greater tool versatility, the majority of existing methods only tackle the comparison between two conditions. We propose a novel, distribution-free, and flexible approach to DEA for single-cell RNA-seq data. This new method, called `ccdf`, tests the association of each gene expression with one or many variables of interest (that can be either continuous or discrete), while potentially adjusting for additional covariates. To test such complex hypotheses, `ccdf` uses a conditional

*To whom correspondence should be addressed.

independence test relying on the conditional cumulative distribution function, estimated through multiple regressions. We provide the asymptotic distribution of the `ccdf` test statistic as well as a permutation test (when the number of observed cells is not sufficiently large). `ccdf` substantially expands the possibilities for scRNA-seq DEA studies: it obtains good statistical performance in various simulation scenarios considering complex experimental designs (*i.e.* beyond the two condition comparison), while retaining competitive performance with state-of-the-art methods in a two-condition benchmark.

Key words: single-cell, conditional cumulative distribution function, conditional independence test, differential expression analysis

1. Introduction

Single-cell RNA-Sequencing (scRNA-seq) makes it possible to simultaneously measure gene expression levels at the resolution of single cells, allowing a refined definition of cell types and states across hundreds or even thousands of cells at once. Single-cell technology significantly improves on bulk RNA-sequencing, which measures the average expression of a set of cells, mixing the information in the composition of cell types with different expression profiles. New biological questions such as detection of different cell types or cellular response heterogeneity can be explored thanks to scRNA-seq, broadening our comprehension of the features of a cell within its microenvironment (Eberwine *and others*, 2014).

Several challenges arise from the sequencing of the genetic material of individual cells like in transcriptomics (see Lähnemann *and others* (2020) for a thorough and detailed review). Differential expression analysis (DEA) is a major field of exploration to better understand the mechanisms of action involved in cellular behavior. A gene is called differentially expressed (DE) if its expression is significantly associated with the variations of a factor of interest. Single-cell data

have different features from bulk RNA-seq data that require special consideration for developing DEA tools. Indeed, scRNA-seq measurements display large proportions of observed zeros, called "dropouts". However, this term may refer to two different types of zeros: either biological zeroes, originating from biologically-true absence of expression in the cell ; or technical zeroes, where a gene is expressed but not detected by the sequencing technology (Lähnemann *and others*, 2020), due to technical limitations such as the scRNA-seq platform used or the sequencing depth. Furthermore, both technical noises and biological differences between cells in the same sample may generate intricate variations, for instance, coming from the difference of subgroup responses to treatment.

The large amount of single-cell measurements provides an opportunity to estimate and characterize the distribution of each gene expression and to compare it under different conditions in order to identify DE genes. In fact, scRNA-seq distributions usually show complex patterns. Therefore, the `scDD` method (Korthauer *and others*, 2016) condenses the difference in distribution between two conditions into four categories: the usual difference in mean, the difference in modality, the difference in proportions and the difference in both mean and modality. Since scRNA-seq data analysis lay unique challenges, new statistical methodologies are needed.

Several strategies making strong distributional assumptions on the data have been proposed to perform single-cell DEA. MAST (Finak *and others*, 2015) and SCDE (Kharchenko *and others*, 2014) are two well-know differential methods, the former using a two-part generalized linear model to take into account both the dropouts and the non-zero values by making a Gaussian assumption of each gene and the latter relying on a Bayesian framework combined with a mixture of Poisson and negative binomial distributions. `scDD` (Korthauer *and others*, 2016) makes use of a Bayesian modeling framework to detect differential distributions and then to classify the gene into four differential patterns using Gaussian mixtures. `DEsingle` (Miao *and others*, 2018) proposes a zero-inflated negative binomial (`ZINB`) regression followed by likelihood-ratio test to compare

two samples. `D³E` (Delmans and Hemberg, 2016) also applies a likelihood-ratio test after fitting a Poisson-Beta distribution.

Yet, Risso *and others* (2018) have advanced that scRNA-seq data are zero-inflated and have proposed to use zero-inflated negative binomial models, while Svensson (2020) have argued that the number of zero values is consistent with usual count distributions. Then, Choi *and others* (2020) illustrated that scRNA-seq data are zero-inflated for some genes but "this does not necessarily imply the existence of an independent zero-generating process such as technical dropout". In fact, biological information (e.g. cell type and sex) may explain it. The authors also discourage imputation as zeros can contain relevant information about the genes. In addition, Townes *and others* (2019) argument that single-cell zero inflation actually comes from normalization and log-transformation. The distribution and the sparsity of scRNA-seq data remains difficult to model, and – as there is no consensus on which model is the best one – it is of utmost importance to develop general and flexible methods for analyzing scRNA-seq data which do not require strong parametric assumptions.

Fewer distribution-free tools have been developed to model single-cell complex distributions without making any parametric assumption. EMDomics (Nabavi *and others*, 2016) and more recently SigEMD (Wang and Nabavi, 2018) are two non-parametric methods based on the Wassertein distance between two histograms, the latter including data imputation to handle the problem of the great number of zero counts. D^3E offers in addition the possibility to perform either the Cramer-von Mises test or the Kolmogorov-Smirnov test to compare the expression values' distributions of each gene, thus delivering a distribution-free option. In a comparative review, Wang *and others* (2019) illustrated that non-parametric methods, i.e. distribution-free, perform better in distinguishing the four differential distributions. Recently, Tiberi *and others* (2020) presented `distinct`, a hierarchical non-parametric permutation approach using empirical cumulative distribution functions comparisons. The method requires biological replicates (at

least 2 samples per group) and allows adjustment for covariates but only tackles the comparison between two conditions to our knowledge.

However, the limitations of these state-of-the-art methods for scRNA-seq DEA are many. The approaches based on a distributional assumption face methodological issues, as they rely on strong distributional assumptions that are difficult to test in practice. In fact, a deviation from the hypothesized distribution will translate into erroneous p-values and may lead to inaccurate results. While the increasing complexity of clinical and biological studies calls for greater tools versatility, the majority of existing methods, whether parametric or non-parametric, cannot handle data sets with a complex design, making them very restrictive. In fact, the most commonly used methods remain in the traditional framework of DEA and only tackle the comparison between two conditions. One might be interested in the genes differentially expressed across several conditions (e.g. more than two cell groups, multi-arm...) or in testing the genes differentially expressed according to a continuous variable (e.g. cell surface markers measured by flow cytometry...). In particular, the identification of surrogate biomarkers from transcriptomic measurements is becoming an emerging field of interest, especially in cancer therapy (Wang *and others*, 2007) or in new immunotherapeutic vaccines (Cliff *and others*, 2004). Gene expression could be used to compare treatments in observational settings. Yet in such cases, adjusting for some technical covariates or some confounding factors is paramount to ensure the validity of an analysis, as this external influence can impact the outcome as well as the dependent variables and thus generates spurious results by suggesting a non-existent link between variables.

Overall, the need of testing the association between gene expression and the variables of interest, potentially adjusted for covariates, makes an additional motivation for developing suitable tools.

The complex hypothesis we aim to test consists in performing a conditional independence test (CIT). A CIT broadens the classical independence test by testing for independence between two

variables given a third one, or a set of additional variables (see Figure 1). Two random variables X and Y are conditionally independent given a third variable Z if, and only if, $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$. As described in Li and Fan (2020), many CIT have been developed previously and are readily available such as discretization-based tests (Margaritis (2005), Huang *and others* (2010)), metric-based tests (Runge (2018), Su and White (2007), Huang *and others* (2016)), permutation-based two-sample tests (Doran *and others* (2014), Gretton *and others* (2012), Sen *and others* (2017)), kernel-based tests (Muandet *and others* (2017), Li *and others* (2009)) and regression-based tests (see Li and Fan (2020) for a short review). Yet, these CIT either suffer from the curse of dimensionality or are hardly applicable to a large number of observations (Muandet *and others* (2017), Zhang *and others* (2011)). Zhou *and others* (2020) have converted the conditional independence test into an unconditional independence test and then used the Blum–Kiefer–Rosenblatt correlation (Blum *and others*, 1961) to develop an asymptotic test. Yet, the latter cannot be applied when X is discrete. Those limitations make these tests impractical in our context of scRNA-seq DEA and thus require adaptation.

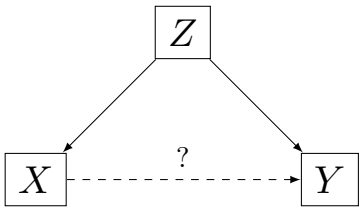

Fig. 1. Conditional dependence graph (Li and Fan, 2020)

Performing DEA necessarily involves performing as many independent tests as there are genes. The variables of interest may be either discrete or continuous, while the number of covariates to condition upon may also increase. Consequently there is an urgent need for a CIT that is both flexible and fast. Here, we propose a novel, distribution-free, and flexible approach, called `ccdf`, to test the association of gene expression to one or several variables of interest (continuous

or discrete) potentially adjusted for additional covariates. Because of the current limitations of existing CIT and the growing interest in testing differences in distribution, we make use of a CIT based on the conditional cumulative distribution function (CCDF), estimated by a regression technique. We derive the asymptotic distribution of the `ccdf` test statistic, which does not rely on the underlying distribution of the data, as well as a permutation test to ensure a good control of type I error and FDR, even with a limited sample size.

Section 2 describes our proposed method with both asymptotic and permutation tests. Section 3 presents several simulation scenarios to illustrate the good performances of `ccdf` when we consider complex designs (*i.e.* beyond the two condition comparison), while a benchmark in the two condition case shows our method retains similar performance in terms of statistical power compared to competitive state-of-the-art methods. Section 4 compares the performances of our method with several methods on a "positive data set" that included differentially expressed genes as well as a "negative data set" without any differentially expressed gene. Section 5 discusses the strengths and limitations of our proposed approach.

2. Method

In this section, we propose a new, easy-to-use, fast, and flexible test for scRNA-seq DEA. We give both its asymptotic distribution as well as a permutation approach to obtain valid p-values in small samples.

2.1 *Conditional independence test*

2.1.1 *Null hypothesis.* Testing the association of Y, namely the expression of a gene, with a factor or a group of factors of interest X either discrete (multiple comparisons) or continuous given covariates Z is equivalent to test conditional independence between Y and X knowing Z:

$$H_0 : Y \perp X \mid Z \tag{2.1}$$

Several statistical tools can be used to characterize the probability law of a random variable, such as the characteristic function, the probability density distribution or the cumulative distribution function. The characteristic function is not often used in practice, due to its relative analytical complexity. The probability density distribution, while more popular, remains difficult to estimate in practice when the number of variables increases due to the increasing number of bandwidths to be optimized. This curse of dimensionality quickly leads to classical computational problems because of complexity growth. As for the cumulative distribution function, its estimation does not require any parameter akin to these bandwidths, making it an efficient tool in high-dimensional non-parametric statistics. From this point, we built a general DEA method including an estimation of the CCDF based on regression technique.

The conditional independence test we propose is based on CCDFs. Indeed, if a group of factors is associated with the expression of a gene, the immediate consequence is that the CCDF of the gene expression would be significantly different from the marginal cumulative distribution, which overlooks this conditioning. Thus, the null hypothesis can be written as:

$$H_0 : F_{Y|X,Z}(y, x, z) = F_{Y|Z}(y, z) \tag{2.2}$$

where the CCDF of Y given X and Z is defined as $F_{Y|X,Z}(y, x, z) = \mathbb{P}(Y \leqslant y \mid X = x, Z = z)$. If there are no covariates, the conditional independence test turns into a traditional independence test as we test the null hypothesis $Y \perp X$ which is equivalent to test $F_{Y|X}(y, x) = F_Y(y)$.

2.1.2 *Test statistic.* In this section, we propose a general test statistic for testing the null hypothesis of conditional independence that is easy to compute. We denote $\boldsymbol{Y}^g = (Y_1^g, ..., Y_n^g)$ an outcome vector (i.e. normalized read counts for gene g in n cells) and $X^g = (\boldsymbol{X}_1^g, ..., \boldsymbol{X}_n^g)$ a $s \times n$ matrix encoding the condition(s) to be tested that can be either continuous or discrete. One may want to add exogenous variables, which are not to be tested but upon which it is necessary to adjust the model. Let $Z^g = (\boldsymbol{Z}_1^g, ..., \boldsymbol{Z}_n^g))$ be a $r \times n$ matrix for continuous or discrete covariates

to take into account. For the sake of simplicity, we drop the notation g in the remainder as we refer to gene-wise DEA.

We have $\boldsymbol{Y} \in [\zeta_{\min}, \zeta_{\max}]$ for some known constants ζ_{min}, ζ_{max}. Let $\zeta_{\min} \leqslant \omega_1 < \omega_2 < ... < \omega_p < \zeta_{\max}$ is a sequence of p ordered and regular thresholds. For each ω_j with $j = 1, ..., p$, the CCDF $F_{Y|X,Z}(\omega_j \mid x, z)$ may be written as a conditional expectation:

$$F_{Y|X,Z}(\omega_j \mid x, z) = \mathbb{E}\left[\mathbb{1}_{\{Y \leqslant \omega_j\}} | X = x, Z = z\right] = \mathbb{E}\left[\widetilde{Y}_{ij} | X_i = x, Z_i = z\right]$$

where $\widetilde{Y}_{ij} = \mathbb{1}_{\{Y_i \leqslant \omega_j\}}$ is a binary random variable that is 1 if $Y_i \leqslant \omega_j$ and 0 otherwise. We propose to estimate these conditional expectations through a sequence of p working models:

$$g\left(\mathbb{E}\left[\widetilde{Y}_{ij} | X_i, Z_i\right]\right) = \beta_{0j} + \boldsymbol{\beta}_{1j}\boldsymbol{X}_i + \boldsymbol{\beta}_{2j}\boldsymbol{Z}_i, \quad \forall i = 1, ..., n \tag{2.3}$$

where $\boldsymbol{\beta}_{1j} = (\beta_{1j1}, ..., \beta_{1js})$ is the vector of size s referring to the regression of $\widetilde{Y}_{ij}$ onto $\boldsymbol{X}_i$ and $\boldsymbol{\beta}_{2j}$ is the vector of size r referring to the regression of $\widetilde{Y}_{ij}$ onto $\boldsymbol{Z}_i$, for the fixed thresholds $\omega_1, \omega_2, ..., \omega_p$. If X has no link with Y given Z, then we expect that $\boldsymbol{\beta}_{1j}$ will be null. So, we aim to test:

$$H_0 : \boldsymbol{\beta}_{1j} = \mathbf{0}, \quad j = 1, ..., p \tag{2.4}$$

Although, there are many different test statistics associated with the null hypothesis (2.4), we propose to use the following test statistic that can be written as

$$D = n\sum_{j=1}^{p}\sum_{k=1}^{s}\boldsymbol{\beta}_{1jk}^2. \tag{2.5}$$

2.1.3 *Estimation and asymptotic distribution.* In this section, we describe how to estimate β_{1j}, which allows the computation of the test statistic (2.5).

While in principle any link function $g(\cdot)$ could be selected for the models (2.3), we select the identity link $g(y) = y$ for its computational simplicity, and we compute coefficient estimates using ordinary least squares (OLS). Because our approach requires p models for each of possibly

thousands of genes, speed is of utmost importance, so we use OLS. Other selections of $g(\cdot)$ are of course possible and could be explored at the cost of additional computation time.

We show in the Appendix that using OLS to estimate (2.3), $\widehat{\boldsymbol{\beta}}_{1j}$ can be expressed by

$$\widehat{\boldsymbol{\beta}}_{1j} = n^{-1} \sum_{i=1}^{n} \boldsymbol{h}_i \tilde{Y}_{ij} \tag{2.6}$$

where $\boldsymbol{h}_i$ is a function of the design matrix W with ith row $\boldsymbol{W}_i = (1, \boldsymbol{X}_i, \boldsymbol{Z}_i)$. These estimates may be plugged into (2.5) to obtain the estimated test statistic $\widehat{D}_n = n \sum_{j=1}^{p} \sum_{k=1}^{s} \widehat{\boldsymbol{\beta}}_{1jk}^2$.

Because of (2.6), we furthermore show in the Appendix that the asymptotic distribution of the test statistic may be approximated by a mixture of χ_1^2 random variables:

$$\widehat{D} = \tilde{\boldsymbol{u}}^{\intercal} A \tilde{\boldsymbol{u}} + o_p(1) = \sum_{j=1}^{ps} a_j \tilde{u}_j^2 + o_p(1) \tag{2.7}$$

where $\tilde{\boldsymbol{u}} = (\tilde{u}_1, ..., \tilde{u}_{ps}) \sim N(0, I)$ are standard multivariate normal random variables and a_j are the eigenvalues of $\Sigma = \text{cov}(\sqrt{n}\widehat{\boldsymbol{\gamma}}_1)$ where $\widehat{\boldsymbol{\gamma}}_1$ is a vectorized version of $\widehat{\boldsymbol{\beta}}_1 = (\widehat{\boldsymbol{\beta}}_{11}, \widehat{\boldsymbol{\beta}}_{12}, ..., \widehat{\boldsymbol{\beta}}_{1p})$, the $s \times p$ matrix then we deduce from (2.6) formed by concatenating the s rows of $\widehat{\boldsymbol{\beta}}_1$ one after another.

We may then compute p-values by comparing the observed test statistic $\widehat{D}_n$ to the distribution of $\sum_{j=1}^{ps} \widehat{a}_j \chi_1^2$, where $\widehat{a}_j$ is an estimate of a_j based on a consistent estimator for Σ (see Appendix for details). Therefore, this allows us to derive a p-value for the significance of a gene with regards to the variable(s) to be tested. In practice, we make use of Davies (1980) approximation to compute p-values for the mixture of χ^2s, implemented in the `CompQuadForm` R package (Duchesne and De Micheaux, 2010). Note that we obtain a simple limiting distribution without relying on any distributional assumptions on the gene expression. In fact, based on the results in Li and Duan (1989), our test will be asymptotically valid as long as there exist any $g(\cdot)$ and any $\beta_{0j}, \beta_{1j}, \beta_{2j}$ such that (2.3) holds. Lastly, the great number of tests requires the Benjamini–Hochberg (Benjamini and Hochberg, 1995) correction afterwards, which is automatically applied to the raw p-values in the R package of `ccdf`.

2.1.4 *Permutation test* Permutation tests are a simple way to obtain the sampling distribution for any test statistic, under the null hypothesis that there is no link between the outcome Y and the variable X. The observations of X can then be shuffled. Permutation tests are recommended when the number of observations is too small, so that the asymptotic distribution can not be assumed to hold. When the sample size n is low, we propose to perform permutations to estimate the empirical distribution of $\widehat{D}_n$ (2.5) under the null hypothesis (2.4). We distinguish two cases: i) testing the association between Y and X without any covariate and ii) testing the association between Y and X given a covariate Z.

i) In the absence of covariates. Under the null hypothesis, Y and X are independent, so the observations of X are exchangeable. Hence, we can randomly permute the observations of X.

ii) In the presence of covariates. When we need to perform a conditional independence testing with a covariate Z, the observations of X are not exchangeable without conditioning on Z. Indeed, if we randomly permute the observations of X, we break not only the link between X and Y but also the link between X and Z. To preserve the dependency between X and Z, we are facing two cases: (a) if Z is a categorical variable and (b) if Z is continuous. In case (a), we randomly switch X within the groups defined by the categories of Z. Under scenario (b), the permutations become tricky. The idea is to permute two observations of X only if the two corresponding observations of Z are close. To do so, a conditional permutation strategy based on the distance between the observations of Z is described in the supplementary material. Following the appropriate method, we can permute the observations of X.

Under i), we are able to compute the test statistic (2.5) from the observations of Y and the permuted observations of X while under ii), the test statistic is obtained from the observations of Y and Z as well as the permuted observations of X. Then, under the null hypothesis, B permutation-based test statistics $\{D_1^*, ..., D_B^*\}$ are calculated and p-values are computed as $\widehat{p} = \frac{1}{1+B}\left(1 + \sum_{b=1}^{B} \mathbb{1}_{\{\widehat{D}_n \leqslant D_b^*\}}\right)$, to avoid getting zero p-values (Phipson and Smyth, 2010), where D_b^*

is the test statistic obtained in the b-th permutation. Finally, we apply the Benjamini–Hochberg correction to the raw p-values to obtain FDR-adjusted p-values.

2.2 *Practical considerations for computational speed up*

2.2.1 *Adaptive permutations* The disadvantage of using permutations could be the onerous computation times, especially when dealing with large sample size, which is more often encountered in single-cell DEA than in bulk DEA. The software computes 1,000 permutations by default for all the genes, but an adaptive procedure may provide similar accuracy at much lower computational cost. When calculation times appear to be too excessive, the user can switch to adaptive permutations. According to some pre-defined rules, the number of permutations is increased at each step to get sufficient numerical precision on the p-values only for certain genes. By default, the method computes 100 permutations for all genes, then we add 150 permutations for the genes with a p-value less than 0.1, bringing the total number of permutations for these genes to 250. Then, for the genes with an associated p-values less than 0.05, we perform 250 permutations more and finally the genes with a p-values less than 0.01, we add 500 permutations to reach 1,000 permutations for a reduced bunch of genes. If the computation times are still too long, the user can choose the number of p-values thresholds and the different limit values. The number of permutations executed at each step is also configurable.

2.2.2 *Evaluation thresholds* Selecting the thresholds $\omega_1, \omega_2, \ldots, \omega_p$ where the CCDF is evaluated may be difficult in practice. If too few thresholds are selected, then important changes in $F_{Y|X,Z}(\omega_j \mid x, z)$ may not be detected. One could instead select the thresholds to match the unique observations of Y, even though this selection technically violates the assumption that the thresholds are fixed and independent of the data. Yet, this technical violation does not appear to adversely affect the performance of our approach in simulations (see Section 3), and `ccdf` selects

the thresholds to match the unique observations of Y by default.

However, there may be an important computational cost to selecting so many thresholds: the number of linear regressions required to estimate all β_1s is then equal to the sample size. So when analyzing data from a large number of cells, one gets an equally large number of regressions to estimate along with a large matrix Σ, significantly increasing the computation time. One solution to reduce computation times (both for the asymptotic test and the permutation test) is to decrease the number of evaluated thresholds and thus the number of estimated β_1s as well as the dimension of Σ.

Instead of going through all the unique values of Y, one can choose a regular sequence of thresholds. Since single-cell RNA-seq data are count data, we propose spacing these thresholds according to a logarithmic scale, i.e., to better focus on the values where the CCDFs will not be too close to 1. This way, `ccdf` statistical power is maximized as variations in distributions are more likely to appear for smaller values (this point is all the more important as the number of thresholds is small).

3. Simulation study

3.1 *Comparisons with state-of-the-art methods in the two conditions case*

We compared the performance of our method `ccdf` with three state-of-the-art methods, `MAST`, `scDD` and `SigEMD`, to find differentially distributed (DD) genes. We have selected methods implemented on R that have shown good results in Wang *and others* (2019) benchmarks and specially designed for single-cell data (excluding methods for bulk RNA-seq like `edgeR` and `DESeq2`). Also, we considered methods that use normalized (and therefore continuous) counts as input, so that the results of our simulations are comparable. We generated simulated count data from negative binomial distributions and mixtures of negative binomial distributions. Since `MAST`, `scDD` and `SigEMD` require continuous input, we transformed the counts into continuous values while pre-

serving as much as possible the count nature of the data (*i.e.* negative binomial assumption). To do so, we added a small Gaussian noise with mean equal to 0 and variance equal to 0.01 to the simulated counts. 500 simulated datasets were generated including 10,000 genes each, of which 1,000 are differentially expressed under a two conditions setting for 7 different sample sizes n (20, 40, 60, 80, 100, 160, 200). The observations were equally divided into the two groups. Korthauer *and others* (2016) classified four different patterns of unimodal or multi-modal distributions:

- differential expression (DE): two unimodal distributions with a different mean in each condition.
- differential proportion (DP): two bimodal distributions with equal component means across conditions; the proportion in the low mode is 0.3 for condition 1 and 0.7 for condition 2.
- differential modality (DM): one unimodal distribution in condition 1 and one bimodal distribution in condition 2 with one overlapping component. Half of the cells in condition 2 belongs to each mode.
- both differential modality and different component means within each condition (DB): one unimodal distribution in condition 1 and 1 bimodal distribution in condition 2. The distributions have no overlapping components. The mean of condition 1 is half-way between the overall means in condition 2. Half of the cells in condition 2 belongs to each mode.

We used these four differential distributions to create our own simulations. Specifically, we simulated 250 DD genes, 250 DM genes, 250 DP genes and 250 DB genes. Plus, 9,000 non-differentially expressed genes are simulated according to two non-differential scenarios (see Supplementary Materials for the simulation settings).

Figure 2 shows the Monte-Carlo estimation over the 500 simulations of the type-I error and the statistical power as well as the false discovery rate and true discovery rate, according to increasing samples sizes. The type-I error is computed as the number of significant genes among

the true negative and the power as the number of significant genes among the true positive. After Benjamini-Hochberg correction for multiple testing, the FDR is computed as the number of false positives among the genes declared DD and the TDR as the number of true positives among the genes declared DD. The p-value nominal threshold is fixed to 5%. The three state-of-the-art methods as well as `ccdf` exhibit good control of type-I error and no inflation of FDR. The four methods present a high overall True Discovery Rate. However, `MAST` shows a lack of power of about 23% for a sample size equal to 200. The True Positive rate (after Benjamini-Hochberg correction) for each scenario for all the methods is shown Figure 3. The three leading methods perform well in finding the traditional difference in mean (DE) as soon as a number of 60 observations is reached. `ccdf` is less powerful for a sample size of 20 and 40 cells but shows the same performances with a larger sample size. The difference in modality (DM) is well detected by all the methods with a slight advantage for `SigEMD` in low sample sizes (from 20 to 60 cells). The difference in proportion (DP) is not favorable for the asymptotic test of `ccdf` until 160 observations but the permutation test exhibits higher power. The asymptotic test requiring a sufficiently large number of observations to converge, the permutation test is more efficient for a lower number of cells. `SigEMD` and `scDD` are the most effective in detecting DP genes. Even though `MAST` shows good power for DE, DM and DP genes, it fails to detect DB genes. In fact, `MAST` is designed to detect difference in the overall mean (traditional differential expression), which is absent in DB scenario. The difference in modality and in different component means is then overlooked as expected with its parametric model. `SigEMD`, `scDD` and both `ccdf`'s tests present competitive performances. Generally speaking, `ccdf` retains competitive performance with the state-of-the-art in this two condition benchmark, which makes it a method particularly adapted to traditional DEA. Even though `ccdf` is a non-parametric method, the asymptotic test is relatively reasonable in terms of computation times, especially compared to `SigEMD` (see Table 1). If the computation times seem too large, we recommend to use the adaptive thresholds strategy explained section 2.2.2

and in particular for the permutation test, we recommend to switch to the adaptive permutations described section 2.2.1.

Table 1. Computation times for the state-of-the-art methods and `ccdf` in the two condition case, for $n = 100$, using 16 cores

Method	Computation times in minutes
`ccdf` asymptotic test	3
`ccdf` permutation test	1385
`ccdf` permutation test with adaptive permutations	750
`SigEMD`	1680
`MAST`	1.4
`scDD`	0.05

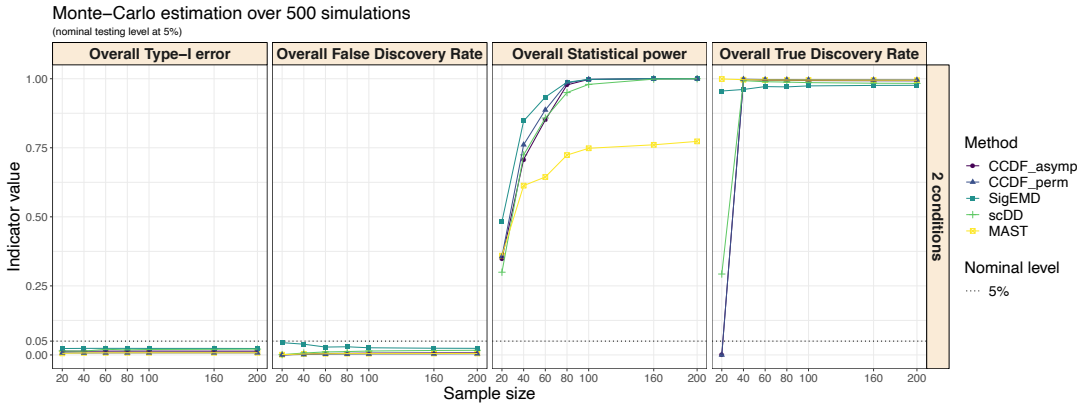

Fig. 2. **Overall Type-I error, Power, FDR and TDR under the 2 conditions case with increasing sample size**. For `ccdf`, we perform the asymptotic test and the permutation test.

3.2 *Multiple comparisons*

This second scenario deals with a multiple comparisons design where cell observations were split into 4 different groups. Count data were generated from negative binomial distributions and mixtures of negative binomial distributions and transformed into continuous values as in section 3.1. 500 simulated datasets were generated including 10,000 genes each, of which 1,000 are differentially expressed under a four conditions setting for 7 different sample sizes n (40, 80, 120, 160,

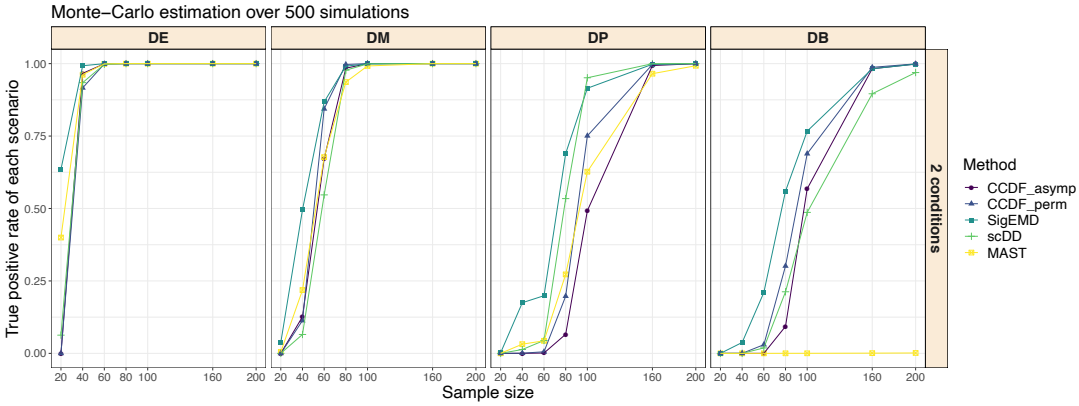

Fig. 3. **True positive rate under the 2 conditions case for the four DD scenarios with increasing sample size.** DE: difference expression in mean. DM: difference in modality. DP: difference in proportion. DB: both differential modality and different component means within each condition. For `ccdf`, we perform the asymptotic test and the permutation test.

200, 320, 400). The observations were then equally divided into four groups. Instead of generating two distributions for each gene as in section 3.1, we created four distributions and therefore new DD scenarios for this specific DEA simulation: multiple DE, multiple DP, multiple DM and multiple DB (more details in the Supplementary Materials). The non-differentially genes are also simulated in a specific fashion described in the Supplementary Material. The idea of differential distribution patterns was converted into a multiple differential distribution setting. For example, the multiple DD scenario consists in four distributions with four different means. Since the other approaches can not handle this type of design, only `ccdf` with the asymptotic test and the permutation test is run.

Figure 5 shows that both versions of `ccdf` have great power to identify DE and DM genes as the sample size increases. DP and DB genes require larger number of cells to achieve sufficient power (from $n > 200$).

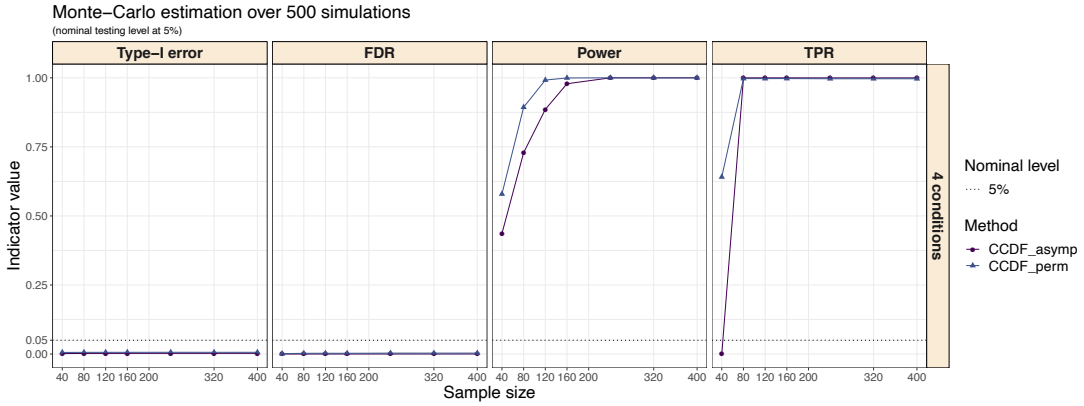

Fig. 4. **Overall Type-I error, Power, FDR and TDR for `ccdf` under the 4 conditions case with increasing sample size**. `ccdf` is the only method capable of dealing with more than 2 conditions. We perform the asymptotic test and the permutation test.

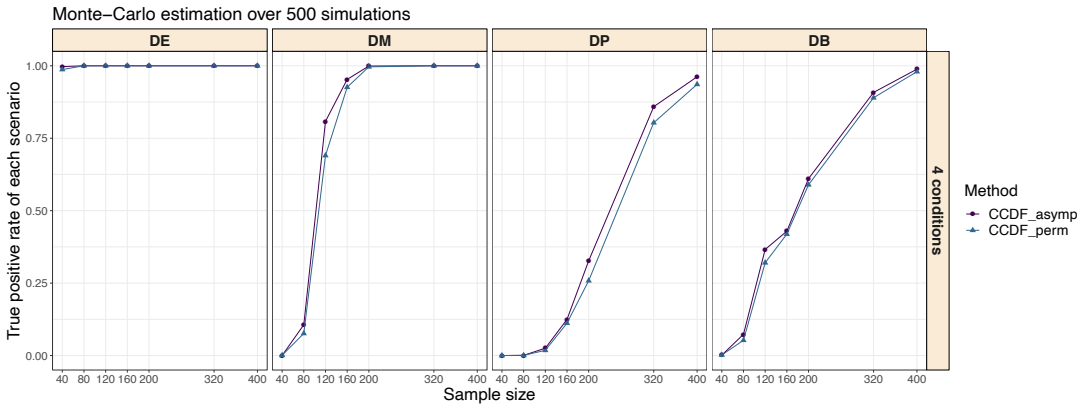

Fig. 5. **True positive rate under the 4 conditions case for the four DD scenarios with increasing sample size.** DE: difference expression in mean. DM: difference in modality. DP: difference in proportion. DB: both differential modality and different component means within each condition.`ccdf` is the only method capable of dealing with more than 2 conditions. We perform the asymptotic test and the permutation test.

3.3 *Two conditions comparison given a covariate Z*

As represented in Figure 1, a potential confounding covariate Z can interfere the test of the link between the outcome Y and the variable X. Then, it is needed to adjust to the confounding variable by carrying out a CIT. We aim to emphasize the importance of taking into account Z, thanks to our approach, by showing the erroneous results obtained not doing so. For this

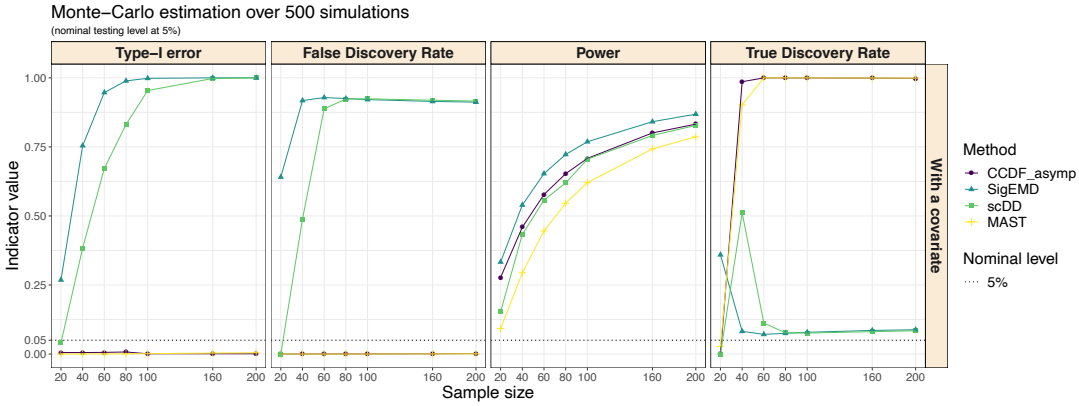

Fig. 6. **Overall Type-I error, Power, FDR and TDR under the 2 conditions comparison given a confounding variable with increasing sample size**. For `ccdf`, we perform the asymptotic test and the permutation test.

purpose, we simulated a confounding variable Z from a Normal distribution. The values of the variable to be tested X depends on the quartile of Z, creating a strong link between X and Z. Y is constructed from X for DE genes and from Z for non-DE genes, the last case is particularly misleading if one do not take into account Z (see simulation settings in Supplementary Material).

We simulated 10,000 genes of which 1,000 are differentially expressed for several sample sizes n (20, 40, 60, 80, 100, 160, 200). `ccdf` permutation test was excluded in this simulation scheme because of the large amount of time to compute. In fact, when we have to adjust for a covariate, the permutation test is not suited to such large sample sizes and number of genes because of the underlying permutation strategy.

`MAST` is able to adjust for covariates like `ccdf` so we expect good performances from both of them. Conversely, `scDD` and `SigEMD` can not control for counfounding variables that might impact the detected DE genes. The results of the benchmark between `MAST`, `scDD`, `SigEMD` and `ccdf` are depicted in Figure 6. Under the alternative hypothesis, we created a large difference in the mean of the normal distributions between the two conditions in order to make DE genes easier to identify. We see therefore that `scDD`, `SigEMD` and `ccdf` exhibit high power at all sample

sizes whereas `MAST` tends to be less powerful for a given size. As expected, the link between Y and Z is very confusing for `scDD`, `SigEMD` which interpret it as a link between Y and X, since X is constructed from Z. Consequently, we observe a consistent rise of Type-I error. `ccdf` and `MAST` perfectly control the Type-I error. In addition, `scDD` and `SigEMD` suffer greatly in terms of TDR as the number of cells is increasing. In fact, fewer real discoveries are found meaning that more genes are identified as significant while being actually false positive, which is in line with the drastically inflated FDR. Dealing with this complex design, `ccdf` outperforms the leading methods by controlling the Type-I error as well as the FDR and by providing a powerful test. This simulation study highlights the importance of taking into account the confounding variables that may exist. Otherwise, DEA may lead to inaccurate results and a potential huge amount of false positives.

It is worth mentioning that `ccdf` can adjust for more than one covariate using the asymptotic test while preserving relatively fast calculation times. The permutation test is for now limited to one adjustment variable and the computation times are obviously increased due to the permutation strategy.

4. Real data set analysis

To be in a more realistic context with a greater number of zeros, a positive control data set and a negative control data set described in Wang *and others* (2019) were used in order to compare the performances of several methods. The genes with a variance equal to zero were removed from the datasets and counts were converted into log-transformed transcript per millions (TPM) values.

4.1 *Positive control dataset*

Islam *and others* (2011) dataset includes 22,928 genes measured across 48 mouse embryonic stem cells and 44 mouse embryonic fibroblasts. We considered 1,000 genes validated through qRT-PCR

Table 2. Number of detected DE genes, and sensitivities of the state-of-the-art methods and `ccdf` tools using positive control real data for an adjusted p-value of 0.05

Method	Number of DE genes	(TP/1000 gold standard)
`ccdf`	9,353	0.734
`SigEMD`†	3,702	0.488
`scDD`†	2,638	0.351
`MAST`†	734	0.198

† results from Wang *and others* (2019).

experiments as a gold standard gene set in the same fashion as Wang *and others* (2019) in order to compute true positive rate. A gene is defined as a true positive if it is found as DE by the method and belongs to the gold standard set (Moliner *and others*, 2008). For `ccdf`, `SigEMD`, `scDD` and `MAST`, the number of DE genes for an adjusted p-value of 0.05 and the number of true positive over the 1,000 gold standard genes are given in Table 2.

`ccdf` leads to the highest true positive rate (0.706) and enables to identify 21.8% more genes in common with the top 1,000 genes compared to `SigEMD` (0.488). `scDD` shows a true positive rate of 0.351 and `MAST` exhibits the lowest rate (0.198) of all the tools.

4.2 *Negative control dataset*

To get false positive rate, we used the dataset from Grün *and others* (2014). We selected 80 samples under the same condition. To create 10 datasets, we randomly divided these 80 cells into 2 groups of 40 cells. As there is no difference between the two groups, not a single gene is to be found in each dataset. Performance evaluations are compared across methods in Table 2.

`ccdf` and `MAST` do not detect any genes which was expected. Although all the cells are under the same condition, `scDD` and `SigEMD` identified respectively 5 and 50 DE genes.

Table 3. Number of the detected DE genes and false positive rates of the state-of-the-art methods and `ccdf` using negative control real data for an adjusted p-value of 0.05

Method	Number of detected DE genes	False positive rate
`ccdf`	0	0
`scDD`†	5	0.0007
`MAST`†	0	0
`SigEMD`†	50	0.007

† results from Wang *and others* (2019).

5. Discussion

We propose a new framework for performing CIT, with an immediate application to differential expression analysis of scRNA-seq data. Our approach can accommodate complex designs, e.g. with more than two experimental conditions or with continuous responses while adjusting for several additional covariates. `ccdf` is capable of distinguishing differences in distribution by using a CIT based on the estimation of CCDFs through a linear regression. The resulting asymptotic test is attractive due to the low computation times, especially dealing with a high number of observations. Yet, for small samples sizes (e.g. due to experimental design or cost limitations in data acquisition), we cannot always rely on an asymptotic test. Consequently, a permutation test is proposed in such cases. Performing permutations is obviously time consuming, but as it is necessary only for small sample sizes, computation times remain reasonable in such settings. Nevertheless, easy parallelization of the permutation test can alleviate this problem. Furthermore, an adaptive procedure, for both the number of permutations and the number of thresholds, is implemented in order to accelerate `ccdf` while preserving a sufficient statistical power along with numerical precision for the lowest p-values. Per-gene asymptotic tests can also be computed in parallel to speed up computations. The proposed approach has been fully implemented in the user-friendly R package `ccdf` available on Github and soon to be on CRAN.

While `ccdf` can be applied to many types of data thanks to its flexibility, it has been specifically tailored for the need of scRNA-seq data DEA. In the simulation study, `ccdf` exhibits great

results and versatility in complex designs such as multiple conditions comparison, and also allows to analyze data when the experiment design includes confounding variables while most of the competing methods cannot. Finally `ccdf` maintains great power while ensuring an effective control of FDR.

Tiberi *and others* (2020) recently proposed non-parametric permutation approach that also compares empirical CDF. While `distinct` shares some common ideas with `ccdf`, the two approaches widely differs in their test statistics and capabilities. On the one hand `distinct` requires multiple samples and only addresses the two groups comparison, while on the other hand `ccdf` provides an asymptotic test and can accommodate more complex experimental designs.

The number of evaluating thresholds considered in `ccdf` directly impacts both its computation time but also potentially its statistical power. To optimize this trade-off and to maintain the performances while reducing the computational cost, we propose to use a sequence of evenly spaced thresholds alongside the log-scale of the observations. Besides, instead of choosing a regular sequence of thresholds, one can prefer to manually chose the different thresholds at which CCDFs are computed and compared, e.g. using *prior* knowledge to emphasize some areas of the distribution.

We did not discuss pre-processing normalization of the data, a corner stone of bulk and single-cell RNA-seq data analysis. Many normalizations have been explored such as Transcript-per-million (TPM), `scran` (Lun *and others*, 2016), `SCnorm` (Bacher *and others*, 2017) or BASiCS (Vallejos *and others*, 2015) (see Lytal *and others* (2020) for a comparative review of different normalizations for scRNA-seq data). As a distribution-free approach, `ccdf` can support any normalization method, and this choice is ultimately left to the responsibility of the user.

Although the linear regression estimation of the CCDF is efficient, different parameterizations can be considered (e.g. logistic regression which would be more natural). Currently `ccdf` cannot analyse multi-sample data, such as biological replicates or repeated measurements. It would be

straightforward to incorporate this structure directly into the working model in the regression equation (2.3). `ccdf` would thus be able to perform multi-sample analysis while keeping the advantage of its asymptotic test, at the cost of an increased computational burden. Besides, other test statistics could be investigated based on the same CIT.

6. Software

The proposed method has been implemented in an open-source R package called `ccdf`, available at https://github.com/Mgauth/ccdf. All R scripts used for the simulations and the real data set analyses in this article were performed using R v3.6.3 and can be found at the same GitHub repository.

Acknowledgments

MG is supported within the Digital Public Health Graduate's school, fundedby the PIA 3 (Investments for the Future - Project reference: 17-EURE-0019). The project is supported through SWAGR Inria Associate-Team from the Inria@SiliconValley program. Computer time for this study was provided by the computing facilities MCIA (*Mésocentre de Calcul Intensif Aquitain*) of the Université de Bordeaux and of the Université de Pau et des Pays de l'Adour. BH & MG thank Franck Picard for helpful discussion about this work.

Conflict of Interest: None declared.

References

Bacher, Rhonda, Chu, Li-Fang, Leng, Ning, Gasch, Audrey P, Thomson, James A, Stewart, Ron M, Newton, Michael and Kendziorski, Christina. (2017). Scnorm: robust normalization of single-cell rna-seq data. *Nature methods* **14**(6), 584.

Benjamini, Yoav and Hochberg, Yosef. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**(1), 289–300.

Blum, Julius R, Kiefer, Jack and Rosenblatt, Murray. (1961). Distribution free tests of independence based on the sample distribution function. *The annals of mathematical statistics* **32**(2), 485–498.

Choi, K., Chen, Y. and Skelly, D.A. et al. (2020). Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. *Genome Biology* **21**(183).

Cliff, Jacqueline M, Andrade, Iryna NJ, Mistry, Rohit, Clayton, Christopher L, Lennon, Mark G, Lewis, Alan P, Duncan, Ken, Lukey, Pauline T and Dockrell, Hazel M. (2004). Differential gene expression identifies novel markers of cd4+ and cd8+ t cell activation following stimulation by mycobacterium tuberculosis. *The Journal of Immunology* **173**(1), 485–493.

Davies, Robert B. (1980). The distribution of a linear combination of $\chi 2$ random variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **29**(3), 323–333.

Delmans, Mihails and Hemberg, Martin. (2016). Discrete distributional differential expression (d 3 e)-a tool for gene expression analysis of single-cell rna-seq data. *BMC bioinformatics* **17**(1), 110.

Doran, Gary, Muandet, Krikamol, Zhang, Kun and Schölkopf, Bernhard. (2014). A permutation-based kernel conditional independence test. In: *Uncertainty In Artificial Intelligence: Proceedings of the Thirtieth Conference*, UAI'14. Arlington, Virginia, USA: AUAI Press. p. 132–141.

Duchesne, Pierre and De Micheaux, Pierre Lafaye. (2010). Computing the distribution

of quadratic forms: Further comparisons between the liu–tang–zhang approximation and exact methods. *Computational Statistics & Data Analysis* **54**(4), 858–862.

Eberwine, James, Sul, Jai-Yoon, Bartfai, Tamas and Kim, Junhyong. (2014). The promise of single-cell sequencing. *Nature methods* **11**(1), 25–27.

Finak, Greg, McDavid, Andrew, Yajima, Masanao, Deng, Jingyuan, Gersuk, Vivian, Shalek, Alex K, Slichter, Chloe K, Miller, Hannah W, McElrath, M Juliana, Prlic, Martin *and others*. (2015). Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology* **16**(1), 1–13.

Gretton, Arthur, Borgwardt, Karsten M, Rasch, Malte J, Schölkopf, Bernhard and Smola, Alexander. (2012). A kernel two-sample test. *The Journal of Machine Learning Research* **13**(1), 723–773.

Grün, Dominic, Kester, Lennart and Van Oudenaarden, Alexander. (2014). Validation of noise models for single-cell transcriptomics. *Nature methods* **11**(6), 637–640.

Huang, Meng, Sun, Yixiao and White, Halbert. (2016). A flexible nonparametric test for conditional independence. *Econometric Theory* **32**(6), 1434–1482.

Huang, Tzee-Ming *and others*. (2010). Testing conditional independence using maximal nonlinear conditional correlation. *The Annals of Statistics* **38**(4), 2047–2091.

Islam, Saiful, Kjällquist, Una, Moliner, Annalena, Zajac, Pawel, Fan, Jian-Bing, Lönnerberg, Peter and Linnarsson, Sten. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex rna-seq. *Genome research* **21**(7), 1160–1167.

Kharchenko, Peter V, Silberstein, Lev and Scadden, David T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature methods* **11**(7), 740–742.

Korthauer, Keegan D, Chu, Li-Fang, Newton, Michael A, Li, Yuan, Thomson, James, Stewart, Ron and Kendziorski, Christina. (2016). A statistical approach for identifying differential distributions in single-cell rna-seq experiments. *Genome biology* **17**(1), 222.

Lähnemann, David, Köster, Johannes, Szczurek, Ewa, McCarthy, Davis J, Hicks, Stephanie C, Robinson, Mark D, Vallejos, Catalina A, Campbell, Kieran R, Beerenwinkel, Niko, Mahfouz, Ahmed *and others*. (2020). Eleven grand challenges in single-cell data science. *Genome biology* **21**(1), 1–35.

Li, Chun and Fan, Xiaodan. (2020). On nonparametric conditional independence tests for continuous variables. *Wiley Interdisciplinary Reviews: Computational Statistics* **12**(3), e1489.

Li, Ker-Chau and Duan, Naihua. (1989). Regression Analysis Under Link Violation. *The Annals of Statistics* **17**(3), 1009 – 1052.

Li, Qi, Maasoumi, Esfandiar and Racine, Jeffrey S. (2009). A nonparametric test for equality of distributions with mixed categorical and continuous data. *Journal of Econometrics* **148**(2), 186–200.

Lun, Aaron TL, Bach, Karsten and Marioni, John C. (2016). Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome biology* **17**(1), 75.

Lytal, Nicholas, Ran, Di and An, Lingling. (2020). Normalization methods on single-cell rna-seq data: An empirical survey. *Frontiers in Genetics* **11**, 41.

Margaritis, Dimitris. (2005). Distribution-free learning of bayesian network structure in continuous domains. In: *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 2*, AAAI'05. AAAI Press. p. 825–830.

Miao, Zhun, Deng, Ke, Wang, Xiaowo and Zhang, Xuegong. (2018). Desingle for detecting three types of differential expression in single-cell rna-seq data. *Bioinformatics* **34**(18), 3223–3224.

Moliner, Annalena, Ernfors, Patrik, Ibanez, Carlos F and Andäng, Michael. (2008). Mouse embryonic stem cell-derived spheres with distinct neurogenic potentials. *Stem cells and development* **17**(2), 233–243.

Muandet, K., Fukumizu, K., Sriperumbudur, B. and Schölkopf, B. (2017). *Kernel Mean Embedding of Distributions: A Review and Beyond*.

Muandet, Krikamol, Fukumizu, Kenji, Sriperumbudur, Bharath and Schölkopf, Bernhard. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning* **10**(1-2), 1–141.

Nabavi, Sheida, Schmolze, Daniel, Maitituoheti, Mayinuer, Malladi, Sadhika and Beck, Andrew H. (2016). Emdomics: a robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. *Bioinformatics* **32**(4), 533–541.

Phipson, B. and Smyth, G. (2010). Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology* **9**.

Risso, Davide, Perraudeau, Fanny, Gribkova, Svetlana, Dudoit, Sandrine and Vert, Jean-Philippe. (2018). A general and flexible method for signal extraction from single-cell rna-seq data. *Nature communications* **9**(1), 1–17.

Runge, Jakob. (2018). Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In: *International Conference on Artificial Intelligence and Statistics*. pp. 938–947.

Sen, Rajat, Suresh, Ananda Theertha, Shanmugam, Karthikeyan, Dimakis, Alexandros G and Shakkottai, Sanjay. (2017). Model-powered conditional independence test. In: *Advances in neural information processing systems*. pp. 2951–2961.

Su, Liangjun and White, Halbert. (2007). A consistent characteristic function-based test for conditional independence. *Journal of Econometrics* **141**(2), 807–834.

Svensson, Valentine. (2020). Droplet scrna-seq is not zero-inflated. *Nature Biotechnology* **38**(2), 147–150.

Tiberi, Simone, Crowell, Helena L, Weber, Lukas M, Samartsidis, Pantelis and Robinson, Mark D. (2020). distinct: a novel approach to differential distribution analyses. *bioRxiv*.

Townes, F William, Hicks, Stephanie C, Aryee, Martin J and Irizarry, Rafael A. (2019). Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome biology* **20**(1), 1–16.

Vallejos, Catalina A, Marioni, John C and Richardson, Sylvia. (2015). Basics: Bayesian analysis of single-cell sequencing data. *PLoS Comput Biol* **11**(6), e1004333.

Wang, Tianyu, Li, Boyang, Nelson, Craig E and Nabavi, Sheida. (2019). Comparative analysis of differential gene expression analysis tools for single-cell rna sequencing data. *BMC bioinformatics* **20**(1), 40.

Wang, Tianyu and Nabavi, Sheida. (2018). Sigemd: A powerful method for differential gene expression analysis in single-cell rna sequencing data. *Methods* **145**, 25–32.

Wang, Xi-De, Reeves, Karen, Luo, Feng R, Xu, Li-An, Lee, Francis, Clark, Edwin and Huang, Fei. (2007). Identification of candidate predictive and surrogate molecular

markers for dasatinib in prostate cancer: rationale for patient selection and efficacy monitoring. *Genome biology* **8**(11), 1–11.

Zhang, K., Peters, J., Janzing, D. and Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In: *Uncertainty in Artificial Intelligence: Proceedings of the Twenty-seventh Conference*. AUAI Press. pp. 804–813.

Zhou, Yeqing, Liu, Jingyuan and Zhu, Liping. (2020). Test for conditional independence with application to conditional screening. *Journal of Multivariate Analysis* **175**, 104557.