Bioinformatics Advances, 2022, 1–13
https://doi.org/10.1093/bioadv/vbac003
Advance Access Publication Date: 11 February 2022
Original Paper

Genome analysis

Compositional Data Analysis using Kernels in mass cytometry data

Pratyaydipta Rudra [1,*], Ryan Baxter[2], Elena W.Y. Hsieh[2,3] and Debashis Ghosh [4]

[1]Department of Statistics, Oklahoms State University, Stillwater, OK 74078, USA, [2]Department of Immunology and Microbiology, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA, [3]Department of Pediatrics, Section of Allergy and Immunology, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA and [4]Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA

*To whom correspondence should be addressed.

Received on August 23, 2021; revised on December 6, 2021; editorial decision on December 11, 2021; accepted on January 12, 2022

Abstract

Motivation: Cell-type abundance data arising from mass cytometry experiments are compositional in nature. Classical association tests do not apply to the compositional data due to their non-Euclidean nature. Existing methods for analysis of cell type abundance data suffer from several limitations for high-dimensional mass cytometry data, especially when the sample size is small.

Results: We proposed a new multivariate statistical learning methodology, Compositional Data Analysis using Kernels (CODAK), based on the kernel distance covariance (KDC) framework to test the association of the cell type compositions with important predictors (categorical or continuous) such as disease status. CODAK scales well for high-dimensional data and provides satisfactory performance for small sample sizes ($n < 25$). We conducted simulation studies to compare the performance of the method with existing methods of analyzing cell type abundance data from mass cytometry studies. The method is also applied to a high-dimensional dataset containing different subgroups of populations including Systemic Lupus Erythematosus (SLE) patients and healthy control subjects.

Availability and implementation: CODAK is implemented using R. The codes and the data used in this manuscript are available on the web at http://github.com/GhoshLab/CODAK/.

Contact: prudra@okstate.edu

Supplementary information: Supplementary data are available at *Bioinformatics Advances* online.

1 Introduction

Recent developments in single-cell-based technologies, such as mass cytometry (i.e. cytometry by time-of-flight, CyTOF), have led to the need for computational and analytic approaches that can accommodate the high dimensionality and single-cell granularity. The analysis of CyTOF data can elucidate novel disease biomarkers and mechanisms of the underlying immunopathology, leading to improved treatments and prognostic measures.

Mass cytometry allows the simultaneous detection of more than 40 proteins per cell in hundreds of thousands of cells per sample (Bendall *et al.*, 2011; Saeys *et al.*, 2016). The data are often clustered into cell subpopulations first, which can then be used to answer scientific questions regarding the abundance of cell types and expressions of specific parameters (e.g. activation markers, signaling proteins, cytokines) across groups, such as disease and control groups, pre- and post-treatment groups, or samples that are stimulated or not. There have been significant research on clustering procedures with these high-dimensional datasets [see Aghaeepour *et al.* (2013) and Weber and Robinson (2016) for a review]. We will focus on the downstream statistical analysis after the clustering has been performed. The statistical questions about the tree-structured cell population data (e.g. Fig. 1) can be visualized in two layers. First, it is clinically interesting to know if the abundance of the cell subpopulations is different across two or more groups and/or conditions. Given the proportion of cell types for each sample, the next question is whether there is any differential expression of activation markers, signaling proteins or cytokines (functional measurements of the cell populations studied). The latter is also known as 'cell state' analysis while the former is called 'cell type' analysis (Weber *et al.*, 2019). In this paper, we focus on the analysis of cell type.

While there are a variety of methods that test differential cell type abundance (Arvaniti and Claassen, 2017; Bruggner *et al.*, 2014; Lun *et al.*, 2017; Weber *et al.*, 2019), several of them suffer from limitations such as

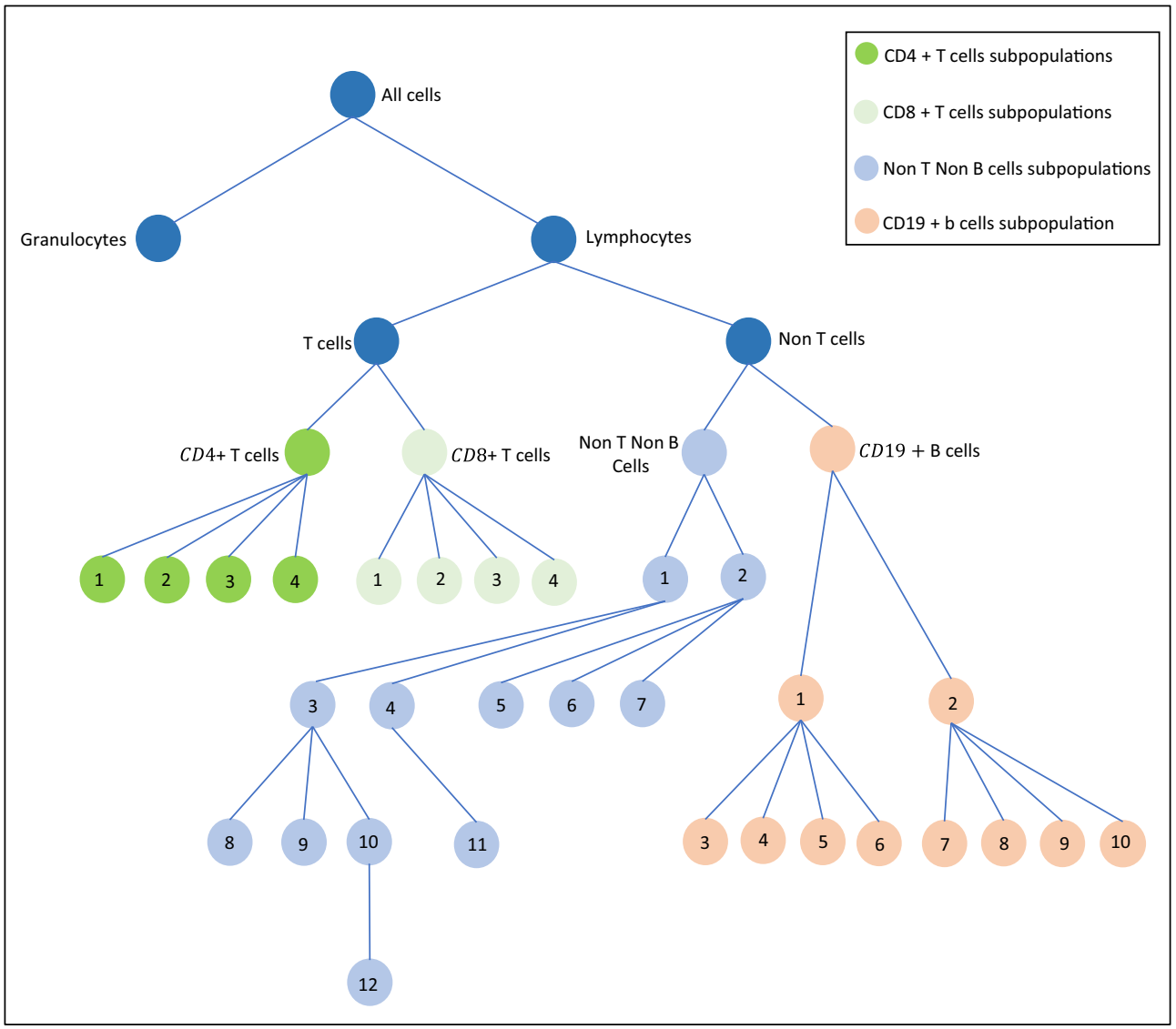

Fig. 1. Hierarchical tree structure of the cell types from the SLE study 2. See Supplementary Materials for a full list of the cell subpopulations. It is of interest to (i) test if the compositional profile of the cell types is associated with the disease groups, i.e. if there is differential abundance of any of the cell types between SLE patients and controls; and (ii) if yes, which cell types contribute the most to the association

1. Difficulty of interpretation due to not clearly distinguishing between cell type (abundance/frequency) and cell state (activation, function) questions.
2. No way of testing statistical significance for individual features (cell types).
3. Computational scalability.

See Weber *et al.* (2019) for a more detailed discussion on these limitations. The two approaches that overcome these limitations are (i) a GLMM-based approach using logistic mixed model (Nowicka *et al.*, 2017) and (ii) the diffcyt method by Weber *et al.* (2019). These approaches effectively perform mass univariate analyses for each cell type. This ignores correlations between cell types as well as increases the multiple testing burden. In this paper, we propose a statistical framework that can test for the association of the multidimensional cell-type profile (or, 'cell-type abundance', to be used interchangeably) with the predictor variables (e.g. disease groups). Starting the analysis with the test of this global null hypothesis avoids the problem of multiple comparisons and also accounts for the correlation structure present in the data.

This multivariate approach can often help us better understand the biological functions of immune cell types. In the context of understanding disturbances to normal immune phenotype and function (i.e. a disease process, iatrogenic intervention), the interdependent relations between different cell types and their function need to be addressed collectively. If we evaluated statistical significance for changes in each single-cell population, independent of the frequency changes in other cellular populations, it would not accurately depict the immunopathology, as the immune system is dynamic and changes in one compartment directly (or indirectly) affect the other. This is also the reason why 'biomarker panels' are currently used instead of single disease biomarkers, since single parameters do not accurately prognosticate disease progression or response to therapy and cannot account for the underlying pathology (no matter how significant that single parameter is). Lastly, the field of multiomic and systems immunology is accelerating our understanding of disease processes because of the capability to assess multiple processes simultaneously, depicting a closer view of the disease process. Hence, statistical analyses must account for this multiplexed capability as opposed to addressing single differences.

1.1 Motivating data

1.1.1 Study 1

This example originates from a previous CyTOF study conducted by O'Gorman *et al.* (2017) to understand single-cell phenotypic and functional characterization of pediatric Systemic Lupus Erythematosus (SLE) patients and healthy controls. SLE patients with presentation prior to age 18-years old and age and gender-matched controls were recruited for the study. Peripheral blood was collected at initial diagnosis prior to the initiation of any treatment. Blood samples were fixed either immediately after collection (T0); or after incubation at 37°C with a protein transport inhibitor cocktail for 6 h (T6). The final dataset after all filtering steps contained 28 single-cell data files (7 patients and 7 controls, and 2 stimulation conditions, T0 and T6, per subject). Single-cell data for every subject were manually gated to obtain the hierarchical clustered cell type data. It is of interest to test if the cell-type compositions are different (i) across the disease groups (ii) across the stimulation conditions.

1.1.2 Study 2

A second study comparing SLE patients and healthy controls is currently being conducted by the authors with a larger number of participants (see Supplementary Section S5, for a partial analysis of this data using our method). The hierarchical clustering structure for this study is shown in Figure 1. Due to the hierarchical nature of the data, cell-type abundance can be defined in many different ways by choosing different parent and children nodes of the tree. For example, it might be of interest to consider the abundance of all cell types from the terminal nodes as a fraction of the lymphocytes, but one may also want to test, e.g. the abundance of the B-cell subpopulations as a fraction of the B-cells. Our proposed multivariate approach can be used to conduct the test to answer each of these questions separately.

1.2 Statistical challenge

Data on cell-type abundance is compositional by nature, i.e. the sum of the cell-type proportions add up to one. In mathematical notation, if $P \equiv (P_1, P_2, \ldots, P_q)$ denotes the cell-type abundance of the q cell types, then

$$P \in \mathcal{S}^q = \left\{ p = (p_1, p_2, \ldots, p_q) : \sum_{i=1}^{q} p_i = 1 \right\},$$

where $\mathcal{S}^q$ is the q-dimensional simplex (Aitchison, 1982). Due to this, the classical statistical models for non-compositional data are not appropriate for testing differential abundance in mass cytometry.

Although there is a rich literature of statistical methods for compositional data analysis (Pawlowsky-Glahn *et al.*, 2015), most of the traditional methods of compositional data analysis (often based on multinomial or Dirichlet distributions) test the association of the predictors with the individual components one at a time. A multivariate approach accounting for the correlation among the components is expected to perform better and have higher statistical power due to a decreased burden of multiple testing. The advantage of using such multivariate approaches for genomic association studies is well documented (e.g. Broadaway *et al.*, 2016; Wu *et al.*, 2011). The authors are unaware of an appropriate multivariate approach to test association in high-dimensional cell-type abundance data arising from mass cytometry.

The traditional generalized linear models cannot be used here due to the presence of overdispersion typical for these data. The state-of-the-art methods to analyze differential abundance for mass cytometry data are based on classical generalized linear mixed models (GLMM; Nowicka *et al.*, 2017) with the help of 'observation-level random effects' (OLRE), or based on newer developments such as edgeR, limma or voom (Law *et al.*, 2014; Ritchie *et al.*, 2015; Robinson *et al.*, 2010). The recently developed diffcyt methods (Weber *et al.*, 2019) using the above approaches are shown to perform well for mass cytometry data, but they have the same limitation of approaching the problem in an univariate manner. Also, it has been shown that the statistical tests based on the GLMM approach often leads to anticonservative results, especially in small samples (Bolker *et al.*, 2009; Forstmeier *et al.*, 2017; Silk *et al.*, 2020) which is typical for clinical studies using CyTOF data. The newer methods such as edgeR or voom do not provide theoretical guarantee of type-I error control either. These methods have been shown to have inflated type-I error rate when applied to other types of data (Datta and Nettleton, 2014; Hawinkel *et al.*, 2019; Rocke *et al.*, 2015; Vestal *et al.*, 2020). In this paper, we propose a multivariate statistical framework, 'Compositional Data Analysis using Kernels' (CODAK), based on kernel distance covariance (KDC; Hua and Ghosh, 2015) to quantify and test the association of predictors such as grouping or application of drugs with the composition profile of cell types. This association test can often be used as the test of differential abundance, but we present the method in general so that it can be used for more general cases (e.g. continuous predictor) besides the two-group comparison. We also propose two approaches of covariate adjustment under this framework and suggest some follow-up methods to understand which components of the composition are most responsible for the association. We illustrate the performance of our methods using extensive simulation studies and also apply it to high-dimensional mass cytometry dataset we collected on SLE patients and healthy control subjects. Analysis of the data revealed clinically relevant patterns such as differential cell type abundance between the disease and the control group.

2 Materials and methods

2.1 The kernel distance covariance framework

Distance covariance/correlation is a method to quantify and test for association between random variables of arbitrary dimensions (Székely *et al.*, 2007, 2009). It is powerful against any form of lack of independence. The distance covariance approach is closely related to the kernel-based approaches using Hilbert Schmidt Independence Criterion (HSIC; Gretton *et al.*, 2007). The equivalence of the two approaches was shown by Sejdinovic *et al.* (2013) and Shen and Vogelstein (2020). Hua and Ghosh (2015) discussed the equivalence in the context of genetic association studies and introduced the term 'kernel distance covariance' (KDC).

For n measurements on two multidimensional random variables $X_{1\times p}$ and $Y_{1\times q}$, let us denote the observation from ith sample unit as $(X^{(i)}, Y^{(i)})$. Define the matrices $K = (k_{ij})$ and $L = (l_{ij})$ as

$$\begin{aligned} k_{ij} &= k(Y^{(i)}, Y^{(j)}), \\ l_{ij} &= l(X^{(i)}, X^{(j)}), \end{aligned} \tag{1}$$

where k and l are the appropriate kernel functions measuring the similarity of pairs of observations. The KDC statistic is defined as

$$KDC_n = \frac{1}{n^2} trace(KHLH), \tag{2}$$

where $H = I_n - 1_n 1_n^T/n$ is the centering matrix, I_n being the identity matrix of dimension n and 1_n being the $n \times 1$ vector with each element equal to 1.

For our application of the KDC approach, suppose we have a (potentially multivariate) predictor $X = (X_1, X_2, \ldots, X_p)$ and that the cell-type abundance is given by $P = (P_1, P_2, \ldots, P_q)$, where $\sum_{k=1}^{q} P_k = 1$. A key aspect of the KDC approach is the choice of the kernels k and l. Some common choices of kernels in association studies are the linear kernel, polynomial kernel and Gaussian kernel (Schölkopf *et al.*, 2004). However, these do not directly apply to cell-type abundance compositional data since the data belongs to a simplex. For CODAK, we propose a kernel using Aitchison distance (Aitchison, 1982) as an appropriate kernel for measuring similarity between two compositions. Let the cell-type composition for the ith sample be $P^{(i)} = (P_{i1}, P_{i2}, \ldots, P_{iq})$. Then the similarity between the

composition profiles of the ith and the jth sample can be given by (Gower, 1966)

$$k(P^{(i)}, P^{(j)}) = -\frac{1}{2} H D^2 H, \tag{3}$$

where the $(i,j)^{th}$ element of the matrix D^2 is $d^2(P^{(i)}, P^{(j)})$, square of the Aitchison distance (AD) between $P^{(i)}$ and $P^{(j)}$. The AD is defined by

$$d(P^{(i)}, P^{(j)}) = \sqrt{\frac{1}{2q} \sum_{r=1}^{q} \sum_{s=1}^{q} \left\{ \ln\left(\frac{P_{ir}}{P_{is}}\right) - \ln\left(\frac{P_{jr}}{P_{js}}\right) \right\}^2} \tag{4}$$

$$= \sqrt{\sum_{r=1}^{q} \left\{ \ln\left(\frac{P_{ir}}{g(P^{(i)})}\right) - \ln\left(\frac{P_{jr}}{g(P^{(j)})}\right) \right\}^2}, \tag{5}$$

where $g(P)$ is the geometric mean of $P_1, P_2, \ldots, P_q$. Alternatively, one can use a Gaussian kernel (Schölkopf *et al.*, 2004) with AD as

$$k(P^{(i)}, P^{(j)}) = \exp\{-d(P^{(i)}, P^{(j)})/\gamma\}, \tag{6}$$

where $d(P^{(i)}, P^{(j)})$ is the AD between $P^{(i)}$ and $P^{(j)}$ and γ is a tuning parameter which is often chosen as the median distance. Our empirical results show that the performance of these two kernels [Equations (3) and (6)] are nearly identical (Supplementary Fig. S6).

Note that AD is same as the Euclidean distance calculated after applying the centered log-ratio (clr) transformation (discussed later in section) to the compositions. The AD has some desirable properties such as scale invariance, permutation invariance, perturbation invariance and subcompositional dominance [see Martín-Fernández *et al.* (1998) for a detailed discussion]. For example, scale invariance ensures that rescaling each composition does not change the distance, and subcompositional dominance ensures that the distance between two subcompositions can only be smaller than the distance between the original compositions. Neither of these two properties is satisfied by the Euclidean distance (Martín-Fernández *et al.*, 1998). These properties are important for using distance-based methods, and failing to ensure them can result in reduced statistical power (Gloor *et al.*, 2017; Pawlowsky-Glahn and Buccianti, 2011). We have also demonstrated this in our simulation studies (Section 3).

Another commonly used distance for compositional data is the Bray–Curtis (BC) distance (Bray and Curtis, 1957). Therefore, we have also implemented a kernel based on the BC distance by replacing the AD in Equation (3) by the BC distance $d_{BC}(P^{(i)}, P^{(j)})$ defined as

$$d_{BC}(P^{(i)}, P^{(j)}) = \frac{\sum_{k=1}^{q} |P_{ik} - P_{jk}|}{\sum_{k=1}^{q} (P_{ik} + P_{jk})}. \tag{7}$$

However, the BC distance is not a proper distance (Greenacre and Primicerio, 2014) and therefore the kernel defined based on it may not be positive semidefinite. It is a common practice to modify the BC kernel matrix by changing the negative eigen values to 0 (Zhao *et al.*, 2015), but this may lead to inferior results.

In some other fields such as microbiome data analysis where the Operational Taxonomic Units (OTU) are related by a phylogenetic tree, it is important to account for the hierarchical structure of the phylogenetic tree to capture the degree of evolutionary divergence between neighboring bacterial groups. Different versions of Unifrac distance have been considered in the microbiome data analysis literature for this purpose (Chen *et al.*, 2012; Lozupone and Knight, 2005; Lozupone *et al.*, 2007; Wong *et al.*, 2016). Other approaches to account for hierarchical structures have been studied by Silverman *et al.* (2017), Wang and Zhao (2017a, 2017b) and Wang *et al.*, (2020). While it is possible to extend CODAK to incorporate distances that account for a hierarchical tree structure, for instance, using weighted Unifrac (Lozupone *et al.*, 2007) or generalized Unifrac distance (Chen *et al.*, 2012), doing so is less relevant for CyTOF as there is no natural equivalent of phylogenetic distance. Therefore, we have not used such a distance in favor of the simplicity of interpretation and follow-up analysis for the immunologist.

For most applications, $X^{(i)}$ is univariate and for the remainder of the paper we treat it as such. We use a linear kernel for continuous predictors defined by

$$l(X^{(i)}, X^{(j)}) = X^{(i)} X^{(j)}, \tag{8}$$

and a kernel based on Hamming distance for binary predictors given by

$$l(X^{(i)}, X^{(j)}) = \exp(-|X^{(i)} - X^{(j)}|). \tag{9}$$

Since the Hamming distance is a distance metric, we can use standard arguments (Sejdinovic *et al.*, 2013; Shen and Vogelstein, 2020) to show that (9) is a proper kernel. We use a permutation method to obtain the null distribution of the KDC statistic. We have also explored the performance of a Gaussian kernel for X and found it to have relatively worse statistical power (Supplementary Fig. S7).

The intuition for the KDC approach for the motivating data (Section 1.1.1) is demonstrated in Figure 2. The figure shows the densities for the kernel similarity measures when comparing the two disease groups (first panel) and when comparing the two stimulation conditions (second panel). The red curve shows the within-group (or condition) similarity and the blue curve shows between-group (or condition) similarity. It is often the case that the cell type compositions are different in two disease groups (SLE and healthy control). Loosely speaking, this is same as saying that the similarity in the compositional profiles between observations from the same disease group is likely to be more compared to that between observations from different disease groups. This is reflected in the first panel where the red 'same group' curve is located to the right of the blue 'different group' curve. On the other hand, the compositions are less likely to vary across stimulation conditions, which is why the curves in the second panel are more similar.

2.2 Adjusting for covariates

Our framework CODAK allows for covariate adjustment using two different approaches. Suppose we have a set of covariates $Z = (Z_1, Z_2, \ldots, Z_K)$. Denote the value of Z from the ith subject as $Z^{(i)}$. In the first approach, we use the additive log-ratio transformation (alr) commonly used in compositional data analysis (Aitchison, 1982; Aitchison *et al.*, 2000; Pawlowsky-Glahn *et al.*, 2015) to transform the data to $Y_{n\times(q-1)} = alr(P_{n\times q})$, perform covariate adjustment to compute the residuals $Y(Z) = (I_n - H_Z)Y$ using linear regression, and transform the residuals back to the simplex $\mathcal{S}^q$ using the inverse alr-transformation: $P(Z) = alr^{-1}(Y(Z))$. Here, H_Z is the hat matrix obtained from the design matrix of Z and I_n is the identity matrix. We can then also compute the residuals $X(Z) = (I_n - H_Z)X$ by regressing X on Z, and use CODAK on $P(Z)$ and $X(Z)$. The alr-transformation and the inverse alr-transformation for a single compositional observation are given by

$$Y = alr(P) = \left[\ln\left(\frac{P_1}{P_q}\right), \ln\left(\frac{P_2}{P_q}\right), \ldots, \ln\left(\frac{P_{q-1}}{P_q}\right) \right] \tag{10}$$

and

$$P = alr^{-1}(Y) = \mathcal{C}[\exp(Y_1), \exp(Y_2), \ldots, \exp(Y_{q-1}), 1]. \tag{11}$$

Here, $\mathcal{C}$ is the closure operator which divides each component of the vector by the sum to ensure the constant sum 1 of the resulting vector. We can skip this technicality of converting $P = alr^{-1}(Y)$ to a composition such that $\sum_{i=1}^{q} P_i = 1$ due to the scale invariance property of the AD. Note, however, that the alr-transformation is asymmetric in the components, and the alr-coordinates are non-isometric in nature (Pawlowsky-Glahn and Buccianti, 2011). Alternatively, one may use the centered log-ratio (clr) transformation and its inverse given by

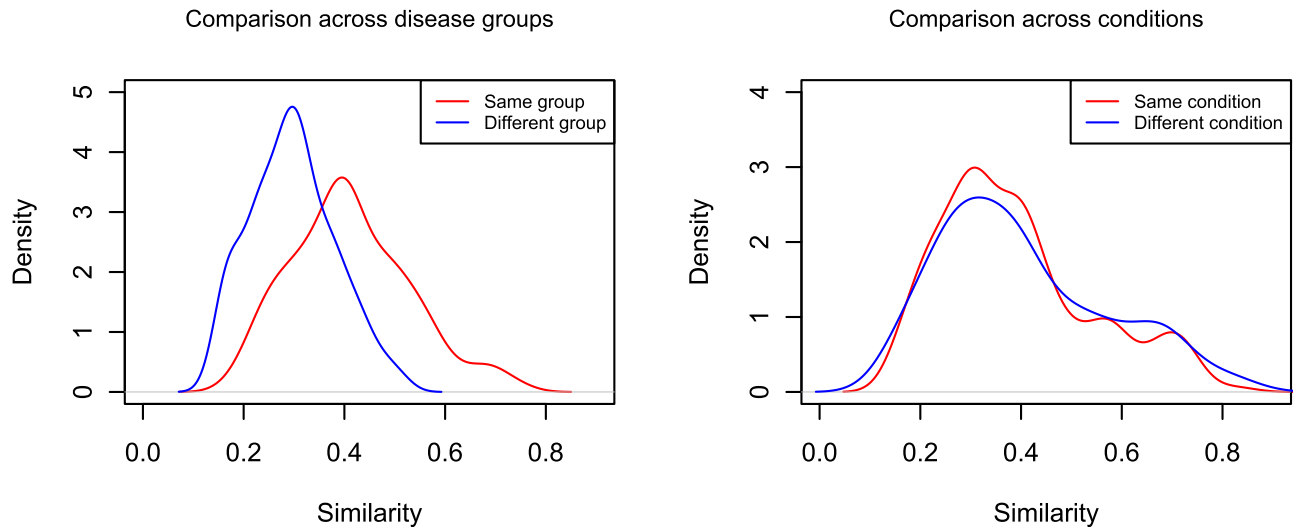

Fig. 2. Motivation of the KDC approach for the first SLE study: the densities for the kernel similarity measures are plotted when comparing the two disease groups (first panel) and when comparing the two stimulation conditions. The red curve shows the similarity of observations within the same group (conditions) and the blue curve shows the similarity of observations between groups (conditions)

$$Y = clr(P) = \left[\ln\left(\frac{P_1}{g(P)}\right), \ln\left(\frac{P_2}{g(P)}\right), ..., \ln\left(\frac{P_q}{g(P)}\right)\right] \tag{12}$$

and

$$P = clr^{-1}(Y) = \mathcal{C}[\exp(Y_1), \exp(Y_2), \ldots, \exp(Y_q)], \tag{13}$$

where $g(P)$ is the geometric mean of $P_1, P_2, \ldots, P_q$. The clr-transformed vector Y is symmetrical in the components, but belongs to a subset of $\mathbb{R}^q$ due to the constraint $\sum_{i=1}^{q} Y_i = 0$. One can obtain an orthonormal coordinate system based on the clr-transformed vector, but it is not necessary for our purpose, and it can be shown that alr and clr have identical result for our covariate adjustment (see Supplementary Section S7 for a simple proof). We present the results using the alr-based approach (CODAK-alr) in Section 3.

A second approach, motivated by Zhan *et al.* (2015), is developed when the covariates are categorical. This approach is based on the idea of stratified kernel, which, in our context, is defined as

$$k(P^{(i)}, P^{(j)}) = \begin{cases} e^{-d(P^{(i)}, P^{(j)})/\gamma}, & \text{if } Z^{(i)} = Z^{(j)} \\ 0, & \text{otherwise} \end{cases}. \tag{14}$$

The kernel l is defined similarly. The method essentially considers two observations i and j to have (potentially) non-zero similarity $k(P^{(i)}, P^{(j)})$ only if they are from the same stratum, where the strata are defined by the values of the categorical covariates. We can then proceed to use CODAK using Equation (2). It can be shown that the stratified kernel defined in this manner is strictly positive definite (Park *et al.*, 2012; Zhan *et al.*, 2015). It should be noted that this method reduces the effective sample size by only considering similarity within strata and therefore may lose power, especially in the presence of multiple categorical covariates. However, in many applications (e.g. the motivating problem), we only have one categorical covariate. The stratified kernel approach (CODAK-sk) can be effective in such cases, and has been empirically shown to be robust against slight violation of the independence of sample observations (Section 3.3).

One important consideration for permutation test in the presence of covariates is the choice of permutation method. Kennedy and Cade (1996) suggested residualizing both Y and X on Z and using a permutation test for the association between $Y(Z)$ and $X(Z)$ as discussed above. However, several studies have found the Kennedy and Cade method to be anticonservative (Anderson and Legendre, 1999; Winkler *et al.*, 2014) for linear and generalized linear models. These studies found that the alternative approach of Freedman and Lane (1983) achieves better control of type-I error. We have explored both the Kennedy and Cade (KC) method and the Freedman and Lane (FL) method here.

2.3 A measure of effect size

A measure of estimated effect size can be obtained by using the idea of distance correlation (Székely *et al.*, 2007). Following Sejdinovic *et al.* (2013), the distance correlation (dcor) between the variables P and X, in our context, can be defined as

$$dcor(P, X) = \frac{KDC(P, X)}{\sqrt{KDC(P, P)KDC(X, X)}}, \tag{15}$$

$$= \frac{\text{trace}(KHLH)}{\sqrt{\text{trace}(KHKH)\ \text{trace}(LHLH)}}. \tag{16}$$

This is equivalent to the squared distance correlation from Székely *et al.* (2007) if both K and L are L_2-distance kernels (Hua and Ghosh, 2015). Since the L_2-distance is not the appropriate distance for compositional data for reasons discussed earlier, the 'original' distance correlation (Székely *et al.*, 2007) should not be used in this case. We include some results on application of the original distance correlation method for this type of data in Supplementary Figure S8.

2.4 Follow-up methods for individual components

One criticism of multivariate approaches is the difficulty of interpretation for individual components of the composition profile. For example, as a follow up to differentially abundant cell-type compositions, it is often of interest to understand the cell types that are the top contributors to the differential abundance. While one can use the traditional effect sizes such as odds ratio or average difference in proportions for each cell type, we provide two kernel-based solutions here.

2.4.1 Leave-one-out approach

One intuition is that if a component c of the composition contributes to the association of the compositional profile P with a predictor X, then dropping that component from the compositional profile should lead to a decrease in the distance correlation. Therefore, we can compute the following leave-one-out (LOO) statistic for every component $c \in \{1, 2, \ldots, q\}$ and rank them in order of their values:

$$dcor_{LOO} = \max\{0, dcor(P, X) - dcor(P_{-c}, X)\}, \quad (17)$$

where $P_{-c} = \mathcal{C}(P_1, P_2, \ldots, P_{c-1}, P_{c+1}, \ldots, P_q)$. It is obvious from Equation (4) that the AD between two compositions $P^{(i)}$ and $P^{(j)}$ reduces when a component is excluded (subcompositional dominance). Further, it can be shown (see Supplementary Section S6 for a simple proof) that the reduction is maximized when the component c with the $\log\left(\frac{P_{ic}}{P_{jc}}\right)$ value farthest from the mean is excluded, i.e.

$$\underset{c}{\text{argmin}}\ d(P^{(i)}_{-c}, P^{(j)}_{-c}) = \underset{c}{\text{argmax}} \left| \ln\left[\frac{P_{ic}}{P_{jc}}\right] - \ln\left[\frac{g(P^{(i)})}{g(P^{(j)})}\right] \right|. \quad (18)$$

This further strengthens the above intuition since it is clear that the component with the highest true ratio of abundance in the two-groups contributes the most to the AD between the composition profiles of observations in different groups. Therefore, we propose picking the top cell-type candidates based on the ranking of this LOO statistic defined in Equation (17).

2.4.2 Weighted distance correlation approach

A second approach is motivated by Wen *et al.* (2020) and based on weighted Aitchison distance (Egozcue and Pawlowsky-Glahn, 2016). Following Wen *et al.* (2020), we define weighted distance correlation between P and X by modifying the kernel k in Equation (6) to use the weighted AD $d_w(P^{(i)}, P^{(j)})$ instead of $d(P^{(i)}, P^{(j)})$, where the weighted AD is defined as

$$d_w(P^{(i)}, P^{(j)}) = \sqrt{\sum_{r=1}^{q} w_r \left\{ \ln\left(\frac{P_{ir}}{g_w(P^{(i)})}\right) - \ln\left(\frac{P_{jr}}{g_w(P^{(j)})}\right) \right\}^2}, \quad (19)$$

where $g_w(P^{(l)}) = \exp\left(\frac{1}{s_w}\sum_{k=1}^{q} w_k \log P_{lk}\right)$ and $s_w = \sum_{k=1}^{q} w_k$. A set of weights that maximize the weighted distance correlation between P and X can then be obtained. Following Wen *et al.* (2020), we simplify the optimization problem by defining

$$w_k = \frac{\beta_k^{\gamma}}{\sqrt{\sum_{r=1}^{q} \{\beta_r^{\gamma}\}^2}}, \quad (20)$$

and by subsequently optimizing for $\gamma > 0$, where β_k is the distance correlation between P_k and X. Larger values of γ make the weights of the components with the strongest association with X contribute more to the weighted AD and subsequently to the KDC statistic. These weights can then be ranked in order of magnitude to indicate the top cell-type candidates contributing to the association.

2.5 Implementation

CODAK is implemented using statistical software package R, and the codes are available at http://github.com/GhoshLab/CODAK/. The time taken on a single computer to run CODAK for a mass cytometry data of usual size ranges from less than a second to ~90 s depending on number of permutations used for the test (10^3 to 10^6). A comparison of the computation time of CODAK with commonly used methods for CyTOF cell-type abundance testing is provided in Supplementary Table S1.

2.6 Relationship between CODAK and other kernel/distance-based approaches

The kernel distance covariance method of testing has previously been shown to be closely related to some other approaches. The kernel machine regression (KMR) approach (Kwee *et al.*, 2008; Liu *et al.*, 2007) uses a semiparametric linear model where the relationship of the predictor variable X with the response Y is modeled through a function $h(.)$ which is estimated from the data. The null hypothesis of interest is $H_0 : h = 0$. The form of the function $h(.)$ is determined by a user-specified kernel function $K(.,.)$ and the test is often done using a score test. Hua and Ghosh (2015) showed that this test using the KMR approach is equivalent to the KDC approach when the kernels chosen for Y and X in the KDC approach are K and the linear kernel, respectively. The KMR approach has been used in many omics applications (e.g. Maity *et al.*, 2012; Wu *et al.*, 2011). One such application was developed for compositional data in a different set up by Zhao *et al.* (2015). They proposed MiRKAT, a test of association for microbiome abundance data. In their model, they used different forms of Unifrac distances (Lozupone and Knight, 2005) and BC distance (Bray and Curtis, 1957), but did not use AD. However, the MiRKAT method can in principle be used with AD.

Another related approach is PERMANOVA, developed by McArdle and Anderson (2001), which tests the association between the two variables using a MANOVA-like F-statistic while obtaining the P-values using permutations. Although it was originally developed for the one-way ANOVA case, several extensions have been proposed to use quantitative variables and covariates (Anderson, 2014). PERMANOVA is a distance-based approach where the user can choose any appropriate distance including AD. Pan (2011) showed that the KMR approach (when using permutation test) and PERMANOVA are equivalent when there are no covariates. Therefore, the KDC approach with linear kernel for the predictor, the KMR approach and the PERMANOVA approach are all equivalent in the no covariate case, and CODAK can be thought of as either of these approaches applied to compositional data from mass cytometry through an appropriate AD kernel.

However, these methods are not all equivalent when adjusting for covariates as we show using our simulation results. In particular, based on the discussion of Hua and Ghosh (2015), the KDC approach is equivalent to KMR if we only adjust the response variable Y for the presence of Z, the covariate. However, adjusting only Y and not X in a permutation test can lead to suboptimal results (Kennedy and Cade, 1996). Therefore, the covariate adjustment procedure of CODAK is not equivalent to PERMANOVA or MiRKAT, and we explore the relative performances of them in section.

Both PERMANOVA and MiRKAT provide an R^2-like measure of effect size. The R^2 of MiRKAT using AD is equivalent to the square of our dcor measure [Equation (15)]. The R^2 of PERMANOVA is not equivalent to these (see Supplementary Section S2 for more discussion).

3 Results

3.1 Simulation studies

3.1.1 Description of the simulations

We conducted simulations to compare type-I error and power performance of the different methods for the following scenarios: (i) no covariate (binary or continuous predictor) (ii) with covariate adjustment (binary or continuous predictor) (iii) with covariate adjustment for repeated measures (binary predictor). We report the results for the binary predictor (i.e. two-group comparisons) here (Fig. 3) and the results for continuous predictor in Supplementary Figures S2 and S4. The third scenario was explored to understand the robustness of the methods for the violation of the independence assumption.

The 'true' effect sizes of multivariate parameter vectors (as used in the simulations) are difficult to illustrate using power plots. Instead, measures such as the true maximum log odds ratio (OR) or the true Aitchison distance can be used in place of effect size, which we explored separately in Figure 4. In order to demonstrate the comparison of statistical power, for each of the above scenarios, we considered four cases: (a) no association (null hypothesis) (b) all cell types have some small association with the predictor (e.g. small difference in abundance of every cell type across two groups), (c) 50% of the cell types have small associations with the predictor (d) 25% of the cell types have larger associations with the predictor. The 'small association' is defined as the 0.2 percentage point difference, and 'larger difference' as 0.4 percentage point difference between the two groups. For scenario (ii), the effect of the covariate Z is simulated similarly to the effect of the predictor in case (b). We used

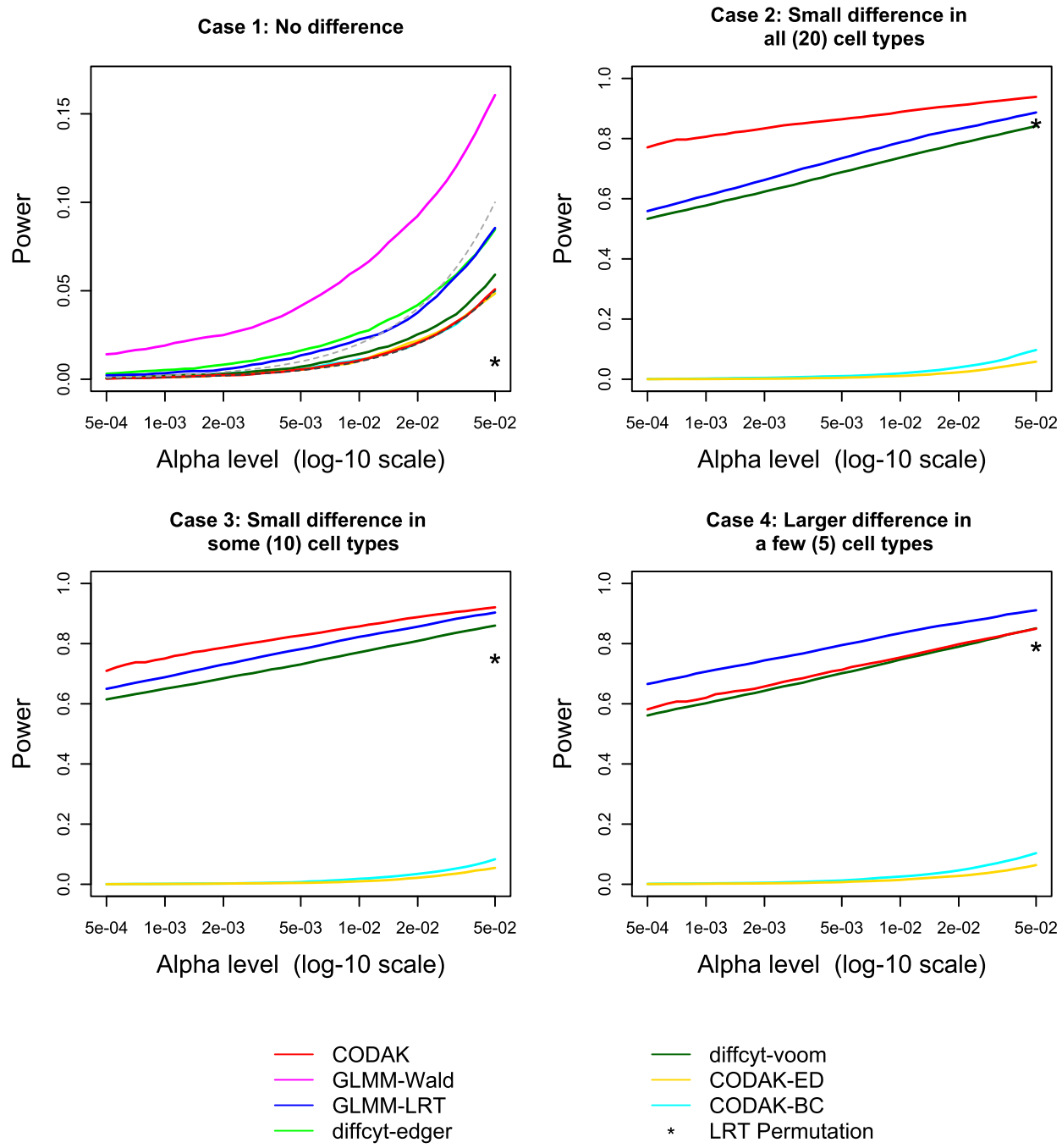

Fig. 3. Comparison of statistical power for binary predictor adjusting for a binary covariate. The black dashed line in the first plot shows the nominal level α and the gray dashed line shows two times α. Only the methods with reasonable control of type-I error are shown in the other three plots

$N = 10\,000$ simulations for each case and plotted the estimated size and power against different choices of the nominal level α.

We simulated data for $2n = 24$ subjects, i.e. $n = 12$ per group for the two-group comparison. Using a multinomial logit model, we generated 30 000–50 000 cells per individual sample with probability depending on the value of the predictor for a cell to be one of the 20 cell types. An OLRE term was included in the multinomial logit model to model the overdispersion present in real CyTOF datasets. The parameters, such as probabilities in the control group (ranging from 0.002 to 0.15), and the standard deviation of the random effects term ($\sigma_b = 0.2$) were chosen based on the estimates obtained from real data (Study 1.1.1). For scenario (iii), we used an additional subject-specific random effect term to induce the correlation in the repeated measures structure.

3.2 Competing methods

Besides using CODAK with the AD kernel, we included CODAK with Bray-Curtis distance (CODAK-BC) and Euclidean distance (CODAK-ED) for the purpose of comparison. We also compared the results using the original distance correlation (Székely *et al.*, 2007), but did not include it in Figures 3–6 since its performance was nearly identical to CODAK-ED (see Supplementary Fig. S8 for a comparison). We used 10 000 permutations for each CODAK

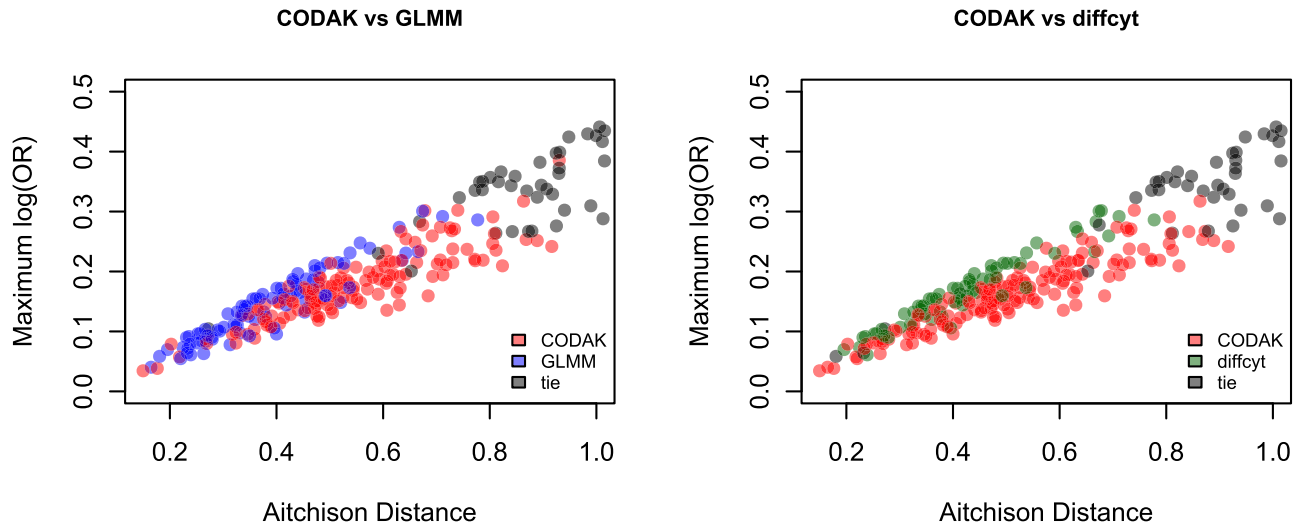

Fig. 4. Comparison of CODAK with GLMM and diffcyt-voom for various effect sizes. For every simulation scenario, the true Aitchison distance (AD) and true maximum log odds ratio are plotted. The colors represent the method with a higher statistical power for that scenario. It is evident that CODAK favors higher AD while the other methods favor strong effects for individual components. Scenarios $AD > 1$ or $|\log(OR)| > 0.5$ are not shown since all methods had perfect power in such cases

method. To compare the performance of CODAK with the commonly used methods for testing cell type abundance in mass cytometry, we considered two GLMM-based methods and two methods from the diffcyt package. The GLMM-based methods were both logistic mixed effects models to test the association of the predictor with the abundance of each cell type (followed by adjustments for multiple testing using Bonferroni method), the difference among the two methods being the test used: Wald test or likelihood ratio test (LRT). The methods for diffcyt were also conducted per cell type (followed by adjustments for multiple testing) using either the edgeR or the voom method [as suggested by Weber *et al.* (2019)]. It should be noted that there is also a GLMM test option in the diffcyt package which is essentially same as the Wald test we considered. Since both the GLMM-based methods resulted in inflated type-I error, we also applied a resampling method using the LRT statistic where the LRT statistic was calculated using the logistic mixed model and a permutation method was used to obtain the *P*-value instead of using the asymptotic distribution of the LRT statistic. Due to high computational time, we only used this method for one of our simulation scenarios and we only reported the power/type-I error rate for $\alpha = 0.05$.

When adjusting for covariates, we included them directly in the model for the GLMM or diffcyt methods. We have also compared PERMANOVA (McArdle and Anderson, 2001) and MiRKAT (Zhao *et al.*, 2015) using AD since these methods are not equivalent to CODAK when adjusting for covariates. R-packages MiRKAT (Zheng *et al.*, 2017) and vegan (Oksanen *et al.*, 2007) were used. For scenario (iii) with repeated measures, when fitting the GLMM models, we used the known information on which observations are coming from the same individual subject to fit multilevel models accounting for the repeated measures structure. On the other hand, CODAK did not use this information.

3.3 Results from the simulation studies

The results for binary predictor (Fig. 3) shows that all the three CODAK methods controlled the type-I error at the target level, but none of the other methods could do so. The type-I error control of diffcyt-voom was near satisfactory and the LRT method among the GLMM methods lead to less inflation of the type-I error. Therefore, we choose one from each of the approaches of the competing methods and only show the diffcyt-voom and GLMM-LRT for the subsequent power comparisons. However, the comparison of power of GLMM-LRT and the other two methods are still unfair since the size of GLMM-LRT is approximately twice the as large as the target α level for the range of values of α considered. The power comparison shows that CODAK (using the default distance AD) performed better than the existing univariate methods in the case with small differences in all cell types (case b). With the reduction in the number of associated cell types, the relative advantage of CODAK compared to the univariate methods gradually diminished. The resampling-based LRT method controlled the type-I error at the nominal level, but failed to achieve high power. It is also extremely computationally intensive. CODAK using BC or Euclidean distance failed to provide satisfactory power. This clearly demonstrates that CODAK with AD has far superior performance compared to these distances. Hence, these distances were not used for the next simulation scenarios. Similar results were obtained for continuous predictor (see Supplementary Fig. S2).

The results for binary predictor in the presence of a covariate (Fig. 3) lead to similar findings. CODAK-alr had slightly inflated type-I error here when we used the KC method of permutations (Kennedy and Cade, 1996), but it was comparable to diffcyt-voom. Using the FL method of permutations (Freedman and Lane, 1983) for CODAK-alr resulted in good control of type-I error, but slightly less power compared to the KC method. This is consistent with previous findings for linear models (Anderson and Legendre, 1999). CODAK-sk also controlled of type-I error at the nominal level. However, CODAK-sk had slightly inferior power performance compared to CODAK-alr. This is not unexpected, as discussed in section. Both PERMANOVA and MiRKAT appeared to be somewhat conservative and had less statistical power. The statistical power of MiRKAT was especially poor. It is likely to be due to the fact that it essentially adjusts the effect of the covariate only on the response variable. Such behavior of covariate adjustment methods has previously been discussed in the literature (Kennedy and Cade, 1996).

The simulation results from scenario (iii) are shown in Figure 6. CODAK-sk achieved reasonable control of type-I error under slight misspecification of the independence assumption (subject-specific random effect variance = 0.1) and the power was also comparable to its performance in other cases. None of the other methods (including CODAK-alr) controlled the type-I error rate. We also verified that when the amount of dependence was increased, the control of type-I error by CODAK-sk gradually worsened (Supplementary Fig. S5).

To better understand the situations where CODAK (or other univariate methods) have advantage, we plotted the true AD and maximum log(OR) for each simulation [from scenario (i)] and color coded according to which method had higher power (Fig. 4). It is not unexpected that the GLMM and diffcyt-voom appeared to perform better than CODAK when maximum log(OR) was high but

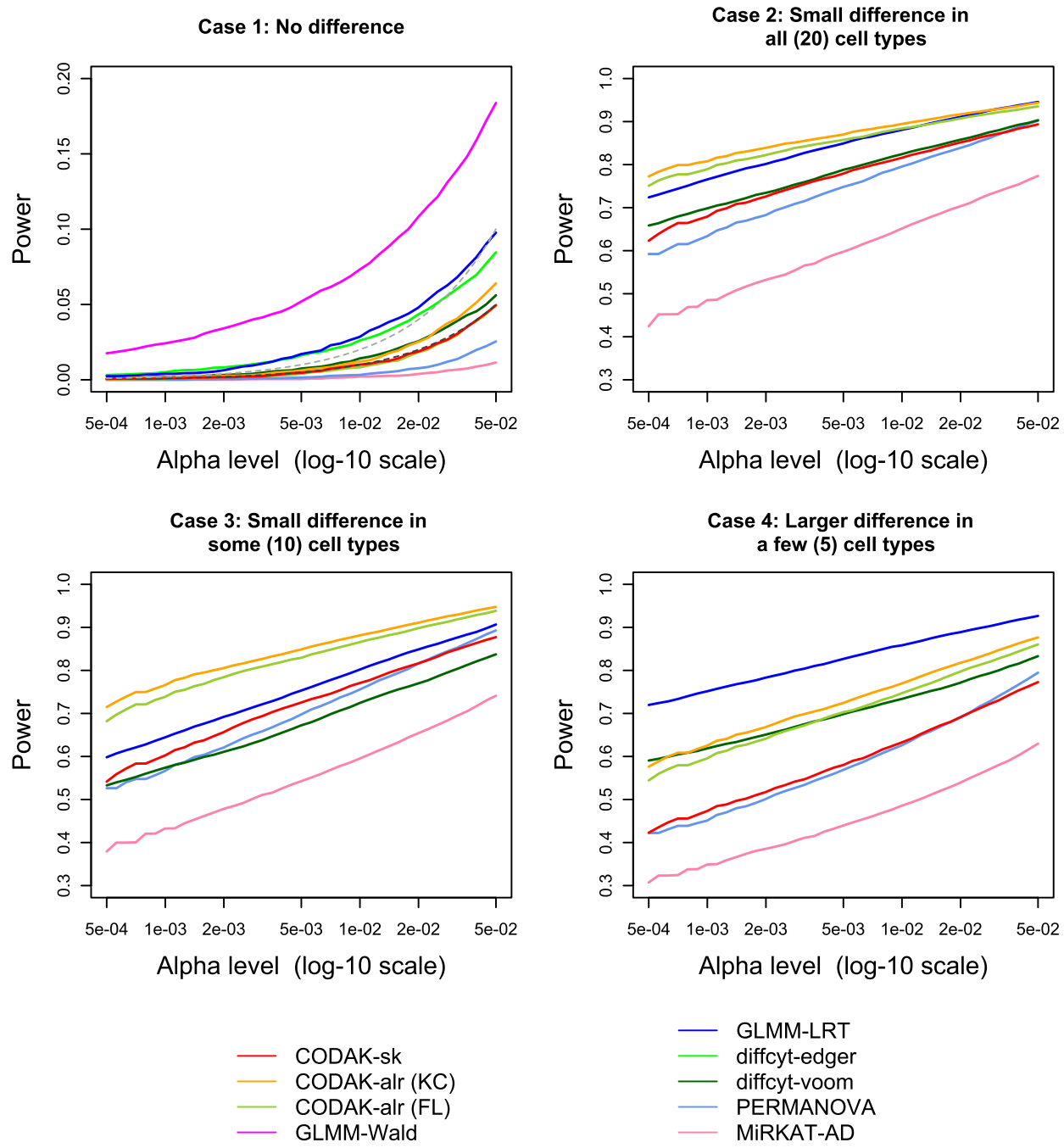

Fig. 5. Comparison of statistical power for binary predictor. The black dashed line in the first plot shows the nominal level α and the gray dashed line shows two times α. Only the methods with reasonable control of type-I error are shown in the other three plots. The power for the LRT-permutation is shown for one choice of α due to the high computation time

the AD was not (blue/green points). These were the cases where a smaller number of cell types are associated and the association is strong. CODAK performed better for effects that were relatively weaker, but more spread out across the cell types (red points). All methods performed well (black points) when there was strong association for several cell types, or very strong association for fewer cell types.

Finally, we compared the ranking of the cell types for the different follow-up methods. For every simulation, we calculated the rank correlation of the true log(OR) with the rank of the cell types obtained using the LOO and the weighted dcor methods. The median of the rank correlations was 0.49 for the LOO method and 0.41 for the weighted dcor method. These are reasonably high considering that the rank correlation between estimated log(OR) and the true log(OR) was 0.42. The median number of 'true' top-5 cell types in the top-5 list provided by these methods was 3 for all the three methods [LOO, weighted dcor, and estimated log(OR)]. A histogram of the number of cell types (among top five) identified by the different methods is shown in Supplementary Figure S10, which shows that both LOO and weighted dcor performed close to the 'true' model using the log(OR), with the LOO method performing slightly better. Based on these results, we can conclude that the performance of the LOO-based follow-up method to rank the contribution of the individual cell types is satisfactory and should be preferred over the weighted dcor method.

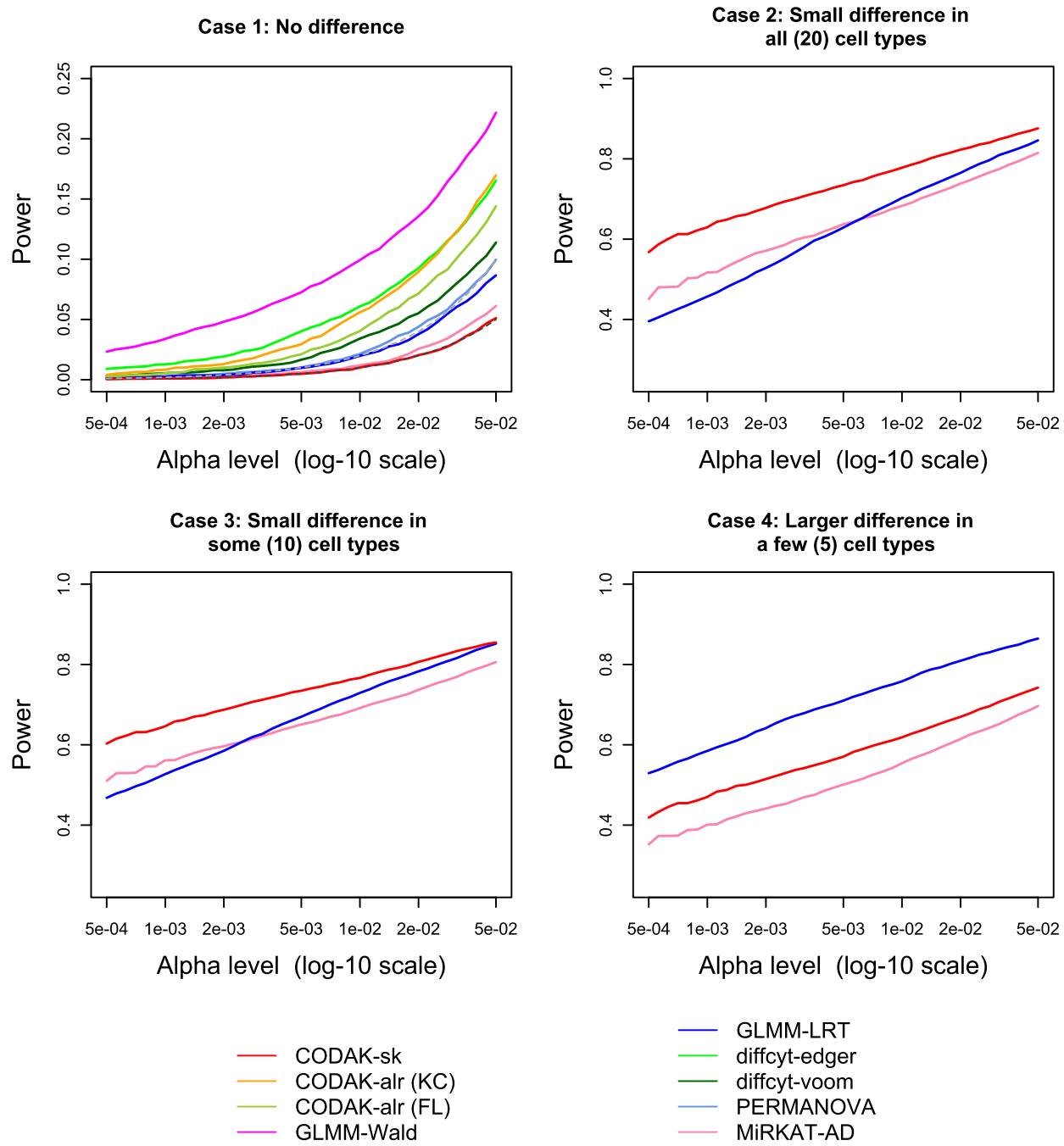

Fig. 6. Comparison of statistical power for binary predictor adjusting for a binary covariate with repeated measures. The black dashed line in the first plot shows the nominal level α and the gray dashed line shows two times α. Only the methods with reasonable control of type-I error are shown in the other three plots

3.4 Real data

In order to study the performance of CODAK on real data we applied it on the SLE data from Section 1.1.1. SLE is a systemic inflammatory disease in which multiple immunological derrangements have been described—Toll-like receptor signaling defects, over activation of neutrophil subsets, decreased regulatory T cell frequency due to T cell tolerance defects, excessive type-I interferon downstream activation, and autoantibody production by pathogenic B-cells, to name a few (Crow, 2014; Zharkova *et al.*, 2017). Therefore, one should expect that study of a single or a couple immune cell subsets in isolation would not fully depict the immunopathology. Recent publications have supported the multiimmune component etiology of the disease, and in fact, different disease phenotypes may be correlated with specific immune profiles that encompass the integrated immunological picture (Nehar-Belaid *et al.*, 2020). Therefore, our work analyzing multiple immune cell subsets and their downstream cytokine production, as a composite integrated approach, most accurately addresses the basic science and clinical questions.

Each of the datasets described in Section 1.1 were generated via Fluidigm mass cytometer instruments. Resultant FCS files were bead normalized to account for instrument calibration. The final cell-type

abundance proportion matrix after the data filtering steps (debarcoding, bead normalization, batch adjustment/normalization) had 28 rows and 12 columns, where each row corresponded to an observation $P^{(i)} \in \mathbb{S}^{12}$ and each column to a cell type. These different cell types were CD4+ T-cells, CD8+ T-cells, CD27-hi B-cells, CD27-lo B-cells, Basophils, CD14-hi monocytes, CD16-hi monocytes, CD56-hi NK-cells, CD56-mid NK-cells, cDCs, pDCs and other cells. The predictor of interest was the disease status (SLE or healthy control), and a potential covariate was the stimulation condition (T0 or T6). Applying the CODAK test on the full dataset with 10^6 permutations resulted in a distance correlation 0.6966 and a P-value $< 10^{-6}$. When adjusted for the covariate stimulation condition (using CODAK-sk), the distance correlation was 0.7974 and the P-value $< 10^{-6}$. However, one needs to be cautious due to the fact that the study contains two repeated measures on each individual subject which leads to the violation of the independence assumption. Separate testing for conditions T0 and T6 still resulted in statistically significant results ($P = 0.0006$ for both T0 and T6).

We also obtained the $dcor_{LOO}$ statistic for every cell type. The results are shown in Figure 7. The top three contributors for both conditions were CD27-lo B-cells, conventional dendritic cells (cDCs) and CD16-hi monocytes. These results are consistent with previously published literature (O'Gorman *et al.*, 2015, 2017; Rodríguez-Bayona *et al.*, 2010). Pathogenic B-cells have long been implicated in the pathogenesis of SLE, including expansion of CD21-lo B-cells, which are also most often CD27-lo (Dörner *et al.*, 2011). Monocytes and dendritic cells constitute key elements of a proinflammatory innate immune response and they are significantly influenced by the circulating cytokine environment. In the study group shown, patients demonstrated significantly increased proinflammatory serum cytokines (data not shown), which likely accounts for the expanded CD16-hi monocytes and cDCs subpopulations (Steinbach *et al.*, 2000).

Please see Supplementary Section S5 for a second data analysis example containing a continuous predictor (Section 1.1.2).

4 Discussion

We proposed an appropriate kernel to quantify similarities in the cell-type abundance profiles for mass cytometry data using the AD. The proposed kernel has the desirable properties of the AD, and also performs well for both simulated and real datasets. Unlike some existing methods, our framework CODAK can also adjust for covariates, both categorical and continuous.

One common issue for testing cell-type abundance for CyTOF is that the number of cell types can often be large and the sample sizes are often small. Many statistical tests, e.g. the tests for the GLMM, are asymptotic tests that do not perform well for such small sample scenarios. In particular, these tests are often anticonservative while resampling versions of these tests can be computationally intensive. We have shown that other tests not specifically requiring large samples in theory can also be anticonservative. Our simulation studies demonstrate that CODAK has far superior type-I error control and competitive power when compared to the state-of-the-art methods. CODAK is also non-parametric in nature, and therefore expected to be more robust compared to existing parametric models against model violations such as violation of distributional assumptions. In one example, we even show that the CODAK-sk method is somewhat robust against violation of the independence assumption and can be used, with caution, for repeated measures data if the repeated measures are expected to be only mildly correlated.

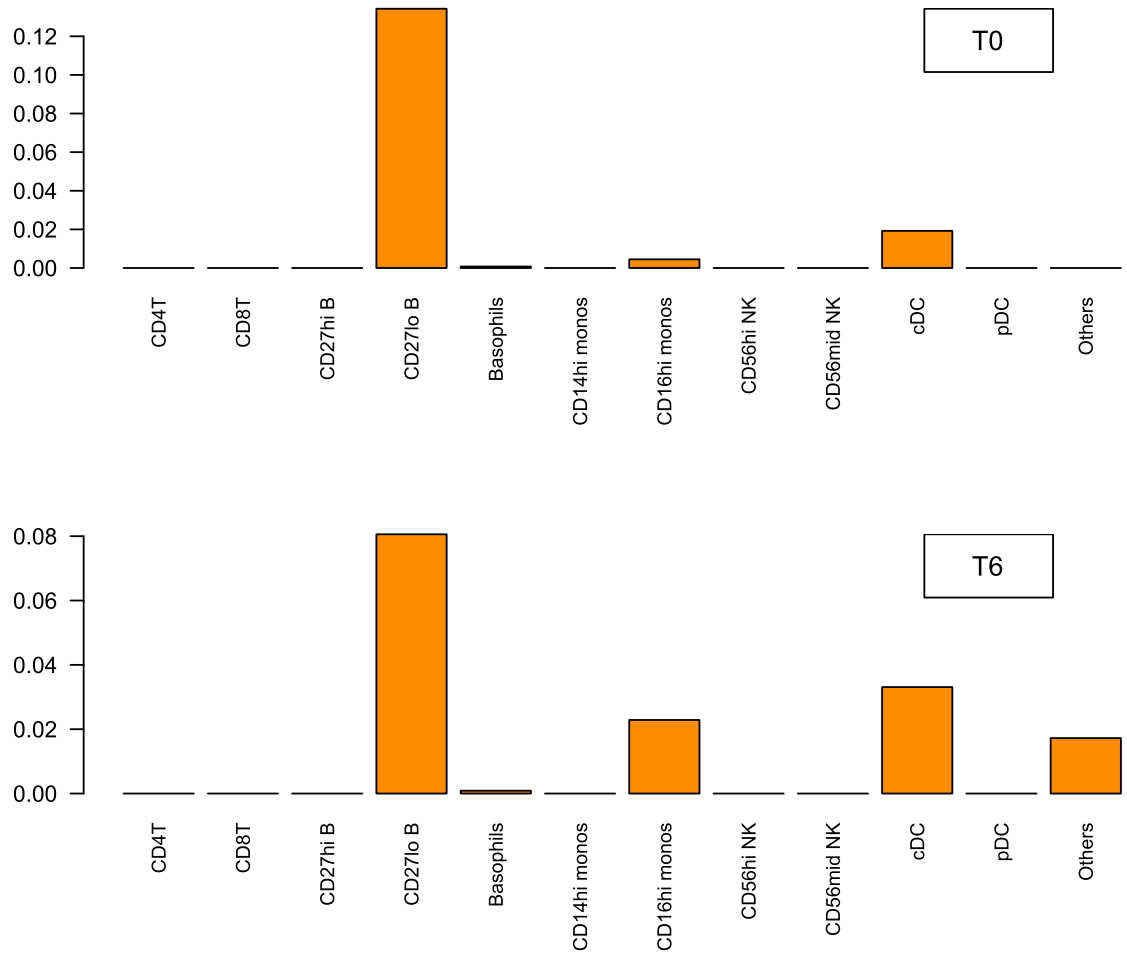

Fig. 7. The values of the $dcor_{LOO}$ statistic for testing the difference in cell type abundance when comparing SLE versus healthy controls at T0 and T6

Another important feature of CODAK is that it tests for the global null hypothesis of association of the predictors with the full compositional profile of the cell types, and therefore does not suffer from the burden of multiple testing. Such gain is most significant when a number of cell types are associated with the predictor variable. Figure 4 provides an insight into when this approach is most advantageous compared to univariate testing using GLMM or diffcyt-voom. CODAK is also able to test association for continuous predictors while some other methods (e.g. diffcyt) do not provide an easy way to do so.

In this article, we have shown the ability of kernel machines to have high power for testing for compositional associations with clinical outcomes, which mirrors its success in other settings (e.g. Rudra *et al.*, 2018; Wu *et al.*, 2011). Nevertheless, because the methodology operates at the level of samples and not individual features, one criticism is its interpretability. As suggested in Section 2.3, one can follow up using existing component-level tests and use gatekeeper type methods to adjust for multiple comparisons. We also suggested two kernel-based approaches to rank the individual features in order of their contribution to the overall association. Another finding is that the proposed method is powerful when there are small differences in the abundance of several cell types. By contrast, if only a small subset of cell types has a difference, the simulations show that our method loses power. However, in immunological disease processes, most pathology affects multiple immune cell subsets and different downstream functional effects (Galbraith *et al.*, 2021; Waugh *et al.*, 2019).

Although we have developed the CODAK framework with CyTOF data in mind, it can potentially be applied more generally for other single-cell platforms such as single-cell RNASeq data. However, it requires more exploration to ensure that the approach performs well for these other data types for which the nature of overdispersion, zero-inflation and data dimensionality might be different compared to CyTOF data. We plan to explore it separately in the future.

Finally, CODAK using the AD kernel cannot handle zero proportions. In our experience, it is uncommon to have zeros in the CyTOF cell-type abundance data compared to some other data types such as microbiome data. However, it is not impossible to have a small number of zeros in the compositional cell-type abundance data from CyTOF. In such situations the user can either use (i) the BC kernel (ii) an existing zero-imputation method to replace the zeros before analysis. Since the performance of the BC kernel was not satisfactory, we suggest using the zero-imputation strategy. Several methods of zero-imputation for compositional data exist in the literature [Martín-Fernández *et al.* (2003, 2012), see Martín-Fernandez *et al.* (2011) and Pawlowsky-Glahn and Buccianti (2011) for a detailed discussion]. The simplest method in the literature is to add a pseudocount, i.e. a small number, often 1, to a zero count (Mandal *et al.*, 2015; Xia *et al.*, 2013). We have performed a simulation study to explore the performance of CODAK after adding a pseudocount, and we found that it controls the type-I error and loses only a small amount of power compared to the situation with no zeros (Supplementary Fig. S9). We believe that more complicated kernel-based approaches modeling the probabilities of zeros is beyond the scope of this paper.

5 Conclusion

We have developed a statistical framework based on kernel distance covariance to test association between compositional profiles of cell type abundance with important predictors for mass cytometry data. Our framework can scale up well for high-dimensions and performs well even in small samples. We also proposed methods for covariate adjustment as well as follow-up methods for finding the top cell types contributing to the association. Using extensive simulation studies, the method has been shown to perform well compared to the existing methods under many scenarios. We also demonstrated the performance of the method in real mass cytometry datasets. The approach has further potential to find application for more complex applications such as immunogenomics for multidimensional predictors. With rising applications of CyTOF, our framework provides an important contribution toward the analysis of high-dimensional cell-type abundance data.

Author contributions

P.R. and D.G. conceived the methodology, R.B. and E.H. conducted the experiment(s) and collected the data, P.R. conducted data analysis and simulations. P.R. wrote the manuscript. All authors reviewed the manuscript.

Funding

This work was supported in part by funds from the National Institute of Arthritis and Musculoskeletal and Skin Diseases [K23AR070897, E.H. and R.B.], Grohne-Stepp Endowment from the University of Colorado Cancer Center (D.G.) and the Boettcher Foundation Webb-Waring Biomedical research grant (E.H. and R.B.).

Conflict of Interest: none declared.

Acknowledgement

The authors thank the anonymous reviewers for their valuable suggestions.

References

Aghaeepour,N. *et al.* (2013) Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods*, **10**, 228–238.

Aitchison,J. (1982) The statistical analysis of compositional data. *J. R. Stat. Soc. B*, **44**, 139–160.

Aitchison,J. *et al.* (2000) Logratio analysis and compositional distance. *Math. Geol.*, **32**, 271–275.

Anderson,M.J. (2014) Permutational multivariate analysis of variance (PERMANOVA). *Wiley Statsref*, 1–15.

Anderson,M.J. and Legendre,P. (1999) An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *J. Stat. Comput. Simul.*, **62**, 271–303.

Arvaniti,E. and Claassen,M. (2017) Sensitive detection of rare disease-associated cell subsets via representation learning. *Nat. Commun.*, **8**, 14825–14810.

Bendall,S.C. *et al.* (2011) Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, **332**, 687–696.

Bolker,B.M. *et al.* (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol. Evol.*, **24**, 127–135.

Bray,J. and Curtis,J. (1957) An ordination of upland forest communities of southern Wisconsin. *Ecol. Monogr.*, **27**, 325–349.

Broadaway,K.A. *et al.* (2016) A statistical approach for testing cross-phenotype effects of rare variants. *Am. J. Hum. Genet.*, **98**, 525–540.

Bruggner,R.V. *et al.* (2014) Automated identification of stratifying signatures in cellular subpopulations. *Proc. Natl. Acad. Sci. USA*, **111**, E2770–E2777.

Chen,J. *et al.* (2012) Associating microbiome composition with environmental covariates using generalized Unifrac distances. *Bioinformatics*, **28**, 2106–2113.

Crow,M.K. (2014) Type I interferon in the pathogenesis of lupus. *J. Immunol.*, **192**, 5459–5468.

Datta,S. and Nettleton,D. (2014) *Statistical Analysis of Next Generation Sequencing Data*. Springer.

Dörner,T. *et al.* (2011) Mechanisms of B cell autoimmunity in SLE. *Arthritis Res. Ther.*, **13**, 243–212.

Egozcue,J.J. and Pawlowsky-Glahn,V. (2016) Changing the reference measure in the simplex and its weighting effects. *Aust. J. Stat.*, **45**, 25–44.

Forstmeier,W. *et al.* (2017) Detecting and avoiding likely false-positive findings—a practical guide. *Biol. Rev.*, **92**, 1941–1968.

Freedman,D. and Lane,D. (1983) A nonstochastic interpretation of reported significance levels. *J. Bus. Econ. Stat.*, **1**, 292–298.

Galbraith,M.D. *et al.* (2021) Seroconversion stages covid19 into distinct pathophysiological states. *eLife*, **10**, e65508.

Gloor,G.B. *et al.* (2017) Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.*, **8**, 2224.

Gower,J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325–338.

Greenacre,M. and Primicerio,R. (2014) *Multivariate Analysis of Ecological Data*. Fundacion BBVA, Bilbao, Spain.

Gretton,A. *et al.* (2007) A kernel statistical test of independence. In: *NIPS*, Curran Associates, Inc., New York, Vol. **20**, pp. 585–592.

Hawinkel,S. *et al.* (2019) A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Brief. Bioinform.*, **20**, 210–221.

Hua,W.-Y. and Ghosh,D. (2015) Equivalence of kernel machine regression and kernel distance covariance for multidimensional phenotype association studies. *Biometrics*, **71**, 812–820.

Kennedy,P.E. and Cade,B.S. (1996) Randomization tests for multiple regression. *Commun. Stat.*, **25**, 923–936.

Kwee,L.C. *et al.* (2008) A powerful and flexible multilocus association test for quantitative traits. *Am. J. Hum. Genet.*, **82**, 386–397.

Law,C.W. *et al.* (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**, R29.

Liu,D. *et al.* (2007) Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*, **63**, 1079–1088.

Lozupone,C. and Knight,R. (2005) Unifrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, **71**, 8228–8235.

Lozupone,C.A. *et al.* (2007) Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.*, **73**, 1576–1585.

Lun,A.T. *et al.* (2017) Testing for differential abundance in mass cytometry data. *Nat. Methods*, **14**, 707–709.

Maity,A. *et al.* (2012) Multivariate phenotype association analysis by marker-set kernel machine regression. *Genet. Epidemiol.*, **36**, 686–695.

Mandal,S. *et al.* (2015) Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.*, **26**, 27663.

Martín-Fernández,J. *et al.* (1998) Measures of difference for compositional data and hierarchical clustering methods. In: *Proceedings of IAMG*, De Frede Editore, Napoli, Vol. **98**, pp. 526–531.

Martín-Fernández,J.A. *et al.* (2003) Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Math. Geol.*, **35**, 253–278.

Martín-Fernandez,J.A. *et al.* (2011) Dealing with zeros. In: Pawlowsky-Glahn,V. and Buccianti,A. (eds.), *Compositional Data Analysis: Theory and Applications*, Wiley-Blackwell, Chichester, UK, pp. 43–58.

Martín-Fernández,J.A. *et al.* (2012) Model-based replacement of rounded zeros in compositional data: classical and robust approaches. *Comput. Stat. Data Anal.*, **56**, 2688–2704.

McArdle,B.H. and Anderson,M.J. (2001) Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, **82**, 290–297.

Nehar-Belaid,D. *et al.* (2020) Mapping systemic lupus erythematosus heterogeneity at the single-cell level. *Nat. Immunol.*, **21**, 1094–1106.

Nowicka,M. *et al.* (2017) CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets. *F1000Research*, **6**, 748.

O'Gorman,W.E. *et al.* (2015) Single-cell systems-level analysis of human toll-like receptor activation defines a chemokine signature in patients with systemic lupus erythematosus. *J. Allergy Clin. Immunol.*, **136**, 1326–1336.

O'Gorman,W.E. *et al.* (2017) Mass cytometry identifies a distinct monocyte cytokine signature shared by clinically heterogeneous pediatric sle patients. *J. Autoimmunity*, **81**, 74–89.

Oksanen,J. *et al.* (2007) The vegan package. *Commun. Ecol. Package*, **10**, 719.

Pan,W. (2011) Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genet. Epidemiol.*, **35**, 211–216.

Park,I.M. *et al.* (2012) Strictly positive-definite spike train kernels for point-process divergences. *Neural Comput.*, **24**, 2223–2250.

Pawlowsky-Glahn,V. and Buccianti,A. (2011) *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons, Wiley, Chichester, UK.

Pawlowsky-Glahn,V. *et al.* (2015) *Modeling and Analysis of Compositional Data*. John Wiley & Sons, Wiley, Chichester, UK.

Plantinga A. *et al.* (2017) MiRKAT: Microbiome Regression-Based Analysis Tests. R package version 1.2.1. https://CRAN.R-project.org/package=MiRKAT.

Ritchie,M.E. *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.

Robinson,M.D. *et al.* (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Rocke,D.M. *et al.* (2015) Controlling false positive rates in methods for differential gene expression analysis using RNA-seq data. *BioRxiv*, page 018739. doi: https://doi.org/10.1101/018739.

Rodríguez-Bayona,B. *et al.* (2010) Decreased frequency and activated phenotype of blood CD27 IgD IgM B lymphocytes is a permanent abnormality in systemic lupus erythematosus patients. *Arthritis Res. Ther.*, **12**, R108–R110.

Rudra,P. *et al.* (2018) Testing cross-phenotype effects of rare variants in longitudinal studies of complex traits. *Genet. Epidemiol.*, **42**, 320–332.

Saeys,Y. *et al.* (2016) Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nat. Rev. Immunol.*, **16**, 449–462.

Schölkopf,B. *et al.* (2004) *Kernel Methods in Computational Biology*. MIT Press.

Sejdinovic,D. *et al.* (2013) Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Stat.*, **41**, 2263–2291.

Shen,C. and Vogelstein,J.T. (2020) The exact equivalence of distance and kernel methods in hypothesis testing. *AStA Adv. Stat. Anal.*, 1–19.

Silk,M.J. *et al.* (2020) Perils and pitfalls of mixed-effects regression models in biology. *PeerJ*, **8**, e9522.

Silverman,J.D. *et al.* (2017) A phylogenetic transform enhances analysis of compositional microbiota data. *eLife*, **6**, e21887.

Steinbach,F. *et al.* (2000) Monocytes from systemic lupus erythematous patients are severely altered in phenotype and lineage flexibility. *Ann. Rheumatic Dis.*, **59**, 283–288.

Székely,G.J. *et al.* (2007) Measuring and testing dependence by correlation of distances. *Ann. Stat.*, **35**, 2769–2794.

Székely,G.J. *et al.* (2009) Brownian distance covariance. *Ann. Appl. Stat.*, **3**, 1236–1265.

Vestal,B.E. *et al.* (2020) MCMSeq: Bayesian hierarchical modeling of clustered and repeated measures RNA sequencing experiments. *BMC Bioinform.*, **21**, 1–20.

Wang,S. *et al.* (2020) Optimal estimation of Wasserstein distance on a tree with an application to microbiome studies. *J. Am. Stat. Assoc.*, **116**, 1237–1253.

Wang,T. and Zhao,H. (2017a) A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. *Biometrics*, **73**, 792–801.

Wang,T. and Zhao,H. (2017b) Structured subcomposition selection in regression and its application to microbiome data analysis. *Ann. Appl. Stat.*, **11**, 771–791.

Waugh,K.A. *et al.* (2019) Mass cytometry reveals global immune remodeling with multi-lineage hypersensitivity to type I interferon in down syndrome. *Cell Rep.*, **29**, 1893–1908.

Weber,L.M. and Robinson,M.D. (2016) Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry Part A*, **89**, 1084–1096.

Weber,L.M. *et al.* (2019) diffcyt: differential discovery in high-dimensional cytometry via high-resolution clustering. *Commun. Biol.*, **2**, 1–11.

Wen,C. *et al.* (2020) Genome-wide association studies of brain imaging data via weighted distance correlation. *Bioinformatics*, **36**, 4942–4950.

Winkler,A.M. *et al.* (2014) Permutation inference for the general linear model. *Neuroimage*, **92**, 381–397.

Wong,R.G. *et al.* (2016) Expanding the unifrac toolbox. *PLoS One*, **11**, e0161196.

Wu,M.C. *et al.* (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.

Xia,F. *et al.* (2013) A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, **69**, 1053–1063.

Zhan,X. *et al.* (2015) Kernel approaches for differential expression analysis of mass spectrometry-based metabolomics data. *BMC Bioinform.*, **16**, 1–13.

Zhao,N. *et al.* (2015) Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *Am. J. Hum. Genet.*, **96**, 797–807.

Zharkova,O. *et al.* (2017) Pathways leading to an immunological disease: systemic lupus erythematosus. *Rheumatology*, **56**, i55–i66.