The Annals of Applied Statistics
2020, Vol. 14, No. 4, 1777–1805
https://doi.org/10.1214/20-AOAS1362

A NEAREST-NEIGHBOR BASED NONPARAMETRIC TEST FOR VIRAL REMODELING IN HETEROGENEOUS SINGLE-CELL PROTEOMIC DATA

BY TRAMBAK BANERJEE[1], BHASWAR B. BHATTACHARYA[2] AND GOURAB MUKHERJEE[3]

[1]*Analytics, Information and Operations Management, School of Business, University of Kansas, trambak@ku.edu*
[2]*Department of Statistics, The Wharton School, University of Pennsylvania, bhaswar@wharton.upenn.edu*
[3]*Department of Data Sciences and Operations, Marshall School of Business, University of Southern California, gmukherj@marshall.usc.edu*

An important problem in contemporary immunology studies based on single-cell protein expression data is to determine whether cellular expressions are remodeled postinfection by a pathogen. One natural approach for detecting such changes is to use nonparametric two-sample statistical tests. However, in single-cell studies direct application of these tests is often inadequate, because single-cell level expression data from processed uninfected populations often contain attributes of several latent subpopulations with highly heterogeneous characteristics. As a result, viruses often infect these different subpopulations at different rates, in which case the traditional nonparametric two-sample tests for checking similarity in distributions are no longer conservative. In this paper, we propose a new nonparametric method for *Testing Remodeling under Heterogeneity* (TRUH) that can accurately detect changes in the infected samples compared to possibly heterogeneous uninfected samples. Our testing framework is based on composite nulls and is designed to allow the null model to encompass the possibility that the infected samples, though unaltered by the virus, might be dominantly arising from underrepresented subpopulations in the baseline data. The TRUH statistic, which uses nearest neighbor projections of the infected samples into the baseline uninfected population, is calibrated using a novel bootstrap algorithm. We demonstrate the nonasymptotic performance of the test via simulation experiments and also derive the large sample limit of the test statistic which provides theoretical support toward consistent asymptotic calibration of the test. We use the TRUH statistic for studying remodeling in tonsillar T cells under different types of HIV infection and find that, unlike traditional tests which do not have any heterogeneity correction, TRUH based statistical inference conforms to the biologically validated immunological theories on HIV infection.

1. Introduction. In many contemporary scientific methodologies it is extremely difficult, even in well-regulated laboratory experiments, to simultaneously control the multitude of factors that give rise to heterogeneity in the population (Chapter 3 of Holmes and Huber (2018)). Nevertheless, these experiments are very powerful, and are often our only recourse to study several interesting biological phenomena. For example, in single-cell proteomic and genomic studies (Jia et al. (2017), Jiang et al. (2018), Shi and Huang (2017), Wang et al. (2018)), it is now well understood that there is high heterogeneity in cellular responses from controlled cell population. Statistical tests are often used on these datasets to determine differences between the case and control samples. The presence of heterogeneity greatly complicates statistical inference and direct application of existing two-sample testing methods,

Received November 2019; revised May 2020.
Key words and phrases. Single-cell virology, immunology, two-sample tests, viral remodeling, homogeneous Poisson process, nearest neighbors, HIV infection, mass cytometry.

without modulating for the latent heterogeneity in the samples, may lead to erroneous statistical decisions and scientific consequences. The problem of testing similarity in the distributions of two samples under heterogeneity arises in a host of modern immunology research set-ups where heterogeneous protein expression datasets collected at single-cell resolution are analyzed to detect viral perturbation. To provide a rigorous statistical hypothesis testing framework for these immunology studies, we consider a composite null hypothesis that allows mixture expression distributions in cases and controls with the mixture having same components but potentially different mixing proportions; the alternative hypothesis contains scenarios where at least one of the mixture components is actually different between the cases and the controls. We develop a new nonparametric testing procedure based on nearest-neighbor distances that can accurately detect if there are differences between the case and control samples in the presence of unknown heterogeneity in the data-generation process. We next provide the background of the problem through an immunology study on human immunodeficiency virus (HIV) infection in tonsillar cells.

1.1. *Phenotypic profiling of T cells under HIV infection.* In single-cell immunology, phenotypic profiling of immune cells under the influence of a target virus, such as the HIV (Cavrois et al. (2017)), the varicella zoster virus (VZV) (Sen et al. (2014)) or the rotavirus (RV) (Sen et al. (2012)), is a critical research endeavor. It enhances understanding of which subsets of cells are most or least susceptible to infection, leading to new insights regarding the magnitude of viral persistence which is crucial in the development of life saving drugs (Sen, Mukherjee and Arvin (2015)). Mass cytometry based techniques (Bendall et al. (2011), Giesen et al. (2014)) are popularly used for generating proteomic datasets for such phenotypic analysis. These techniques can simultaneously measure around 50 protein expressions on individual cells. In this paper we provide a rigorous statistical analysis for testing if there are any HIV induced changes in the proteomic expressions of tonsillar T cells, which are a type of lymphocyte that plays a central role in the immune response, based on the dataset generated in Cavrois et al., 2017.

Figure 1 presents a schematic representation of the experimental set-up used for generating single-cell level proteomic expression data of HIV-infected T cells using cytometry by time of flight (CyTOF) technique. Tonsillar T cells from four healthy donors were infected with two variants of a HIV viral strain: `Nef rich HIV` and `Nef deficient HIV`. `Nef` (negative regulatory factor) is a protein encoded by HIV which enhances virus replication in the host cell by protecting infected cells from immune surveillance. We study the differential impact of these two variants on the immune cells. The healthy cells were cultured and processed into three batches for each donor. For each patient, one among the three batches

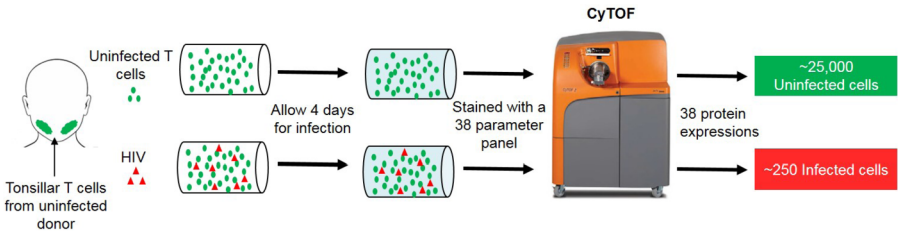

FIG. 1. *Schematic representation of the experimental design associated with the phenotypic analysis of HIV-infected CD4+ T cells using mass cytometry. Tonsillar T cells from a healthy donor (represented by green circles) are infected with the* `Nef rich` *or the* `Nef deficient` *HIV virus (represented by red triangles). These cells were then phenotyped in a* 38 *parameter panel after allowing four days for infection. The resulting data has* 38 *protein expressions for approximately* 25,000 *uninfected cells, and the number of virally infected cells was around* 250.

was randomly selected and phenotyped to generate the expression data of the uninfected population, while the other two batches were contaminated with the `Nef rich HIV` and the `Nef deficient HIV`, respectively, and phenotyped after *four* days of infection. All the batches where phenotyped using multiparameter CyTOF panel which contained 35 surface markers and *three* viral markers. These are special proteins attached to the cell membrane. After leaving out dead cells from each run of the CyTOF experiment, we had 38 protein expressions for approximately 25,000 uninfected cells. Virus infected cells in the contaminated population were marked based on the expression of the viral markers, and it was found that the number of virally infected cells in the batch subjected to HIV infection was around 250. These cells constitute the infected cell population.

1.2. *Viral remodeling*. If the virus changes the expression of any of the surface markers, which are proteins attached to the cell membrane, then the cell is said to have undergone viral remodeling of its phenotypic characteristics (Sen et al. (2014)). A virally remodeled cell will have aberrant intercellular activities, therefore detecting the presence of remodeling is a fundamental step toward understanding the mechanism of pathogenesis and disease progression. Detecting remodeling translates to testing if there is enough evidence in the data to reject the null hypothesis that the joint distribution of all the surface proteins is same between the uninfected and virus infected sample. A natural approach for this problem is to invoke nonparametric two-sample testing methods to see if there is enough evidence to support the alternative hypothesis that the virus has changed the distribution of least one of the subpopulations. However, for single-cell level expression data the hypothesis test described above is particularly difficult because of the following two reasons: (a) the presence of *heterogeneity* in the uninfected population, and (b) due to the phenomenon of *preferential infection*. Single-cell resolution expression data from processed uninfected population often contains attributes from several latent subpopulations with highly heterogeneous characteristics. This subpopulation level heterogeneity in the uninfected (also referred to as the control or baseline) samples can arise from varied attributes that cannot be controlled in experiments, such as differences in the cell effector functions, trafficking and longevity (Cavrois et al. (2017)). Viruses often infect these different subpopulations at different rates. If a virus infects different subpopulation at different rates, but does not alter the marker expressions for any of the subpopulations, then the distribution of the overall viral sample will still be different from the uninfected samples. In these situations the difference in distribution between the infected and the uninfected samples is not due to *viral remodeling* but due to *preferential infection* (for a detailed biological explanation, see Figures 2A and 2B of Cavrois et al., 2017) of the uninfected subpopulations by the virus.

Figure 2 presents two scenarios that may arise when the cloud of infected and uninfected cells are analyzed with respect to a single marker A. In this toy example, Panel 1 in Figure 2 shows that the uninfected T cells arise from three subpopulations with varying expression levels for marker A which may reflect their inherent heterogeneity with respect to cell longevity. The scenario of *preferential infection* is depicted in Panel 2 where the HIV preferentially infects the T cell subpopulation that has a lower expression level for marker A amongst the uninfected cells. Moreover, the virus does not alter the expression levels of these infected cells when compared to Panel 1. In Panel 3, which represents *HIV remodeling*, the virus targets those uninfected cells that have low to medium expression for marker A amongst the uninfected cells and alters their original expression levels upon infection. The distinct pink and yellowish shade of the infected cells in panel 3 depicts their phenotypic change associated with infection. Here, we have described the phenomenon of viral remodeling only for the HIV. However, remodeling analysis is widely conducted across virology for understanding mechanism of other pathogens also. For correct scientific understanding of the viral mechanism, it is extremely important to accurately distinguish the instances of viral remodeling

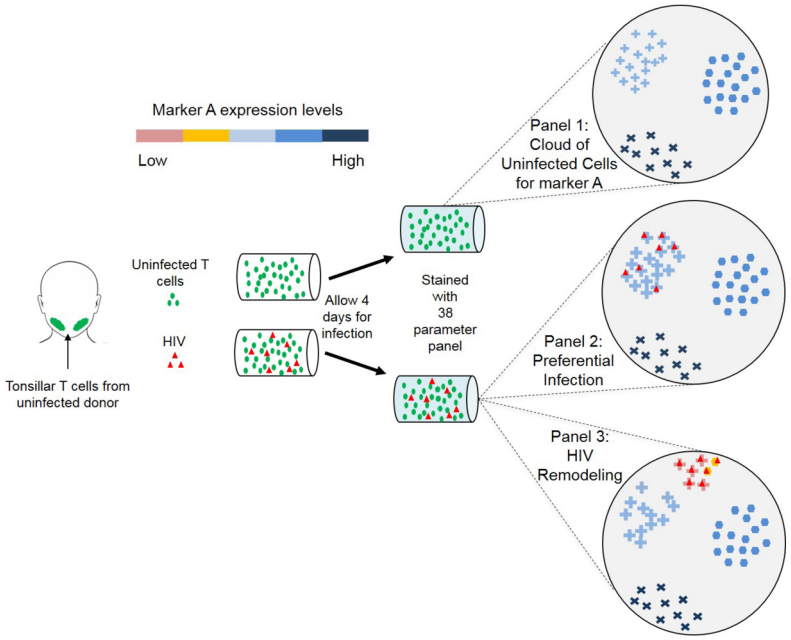

FIG. 2. *Schematic representation of HIV remodeling of T cells with respect to a single marker A. Panel* 1 *shows that the uninfected T cells arise from three subpopulations with varying expression levels for marker A. Panel* 2 *depicts preferential infection where the HIV preferentially infects the T cell subpopulation that has a lower expression level for marker A amongst the uninfected cells and the infection does not alter the expression levels of the T cells when compared to Panel* 1. *Panel* 3 *represents HIV remodeling where the HIV targets those uninfected cells that have low to medium expression for marker A amongst the uninfected cells and alters their original expression levels upon infection, which is represented by the distinct pink and yellowish shade of the infected cells.*

from mere preferential infection. However, popular single-cell based segmentation and classification algorithms (Amir et al. (2013), Bruggner et al. (2014), Linderman et al. (2012), Qiu (2012)) lack a rigorous statistical hypothesis testing framework for conducting two-sample inference and can greatly suffer in testing problems, particularly if there is high imbalance in the sizes of the uninfected (control) and infected (case) samples which is often the situation in virology.

1.3. *Testing procedures in existing literature and statistical challenges.* The statistical framework for testing remodeling falls under the realm of nonparametric two-sample testing. For univariate data, nonparametric two-sample tests like the Kolmogorov–Smirnov test, the Wilcoxon rank-sum test and the Wald–Wolfowitz runs test are extremely popular and find a place in every practitioner's toolkit. Multidimensional versions of these widely used tests date back to the randomization tests of Chung and Fraser (1958) and to the generalized Kolmogorov–Smirnov test of Bickel (1968). Friedman and Rafsky (1979) proposed the first computationally efficient nonparametric two-sample test which applies to high-dimensional data. The Friedman–Rafsky edge-count test, which can be viewed as a generalization of the univariate runs test, computes the Euclidean minimal spanning tree (MST)[1] of the pooled sample and rejects the null if the number of edges with endpoints in different samples is small. Many variants of the edge-count test, based on nearest-neighbor distances and geometric graphs, have been proposed over the years by Hall and Tajvidi (2002), Henze (1984),

[1]Given a finite set $S \subset \mathbb{R}^d$, the *minimum spanning tree* (MST) of S is a connected graph with vertex-set S and no cycles, which has the minimum weight, where the weight of a graph is the sum of the distances of its edges.

Rosenbaum (2005), Schilling (1986), Weiss (1960). Recently, Chen and Friedman (2017) suggested novel modifications of the edge-count test for high-dimensional and object data, and Chen, Chen and Su (2018) proposed new and powerful tests to deal with the issue of sample-size imbalance. Asymptotic properties of two-sample tests based on geometric graphs can be studied in the general framework described in Bhattacharya (2019). Other popular two-sample tests include the test of Baringhaus and Franz (2004), the energy distance test of Aslan and Zech (2005) and the kernel based test using maximum mean discrepancy of Gretton et al. (2007). More recently, Chen, Dou and Qiao (2013) address the problem of sample-size imbalances in the two-sample problem by constructing an ensemble subsampling scheme for the nearest-neighbor tests Henze (1984), Schilling (1986). Very recently, Deb and Sen (2019) and Ghosal and Sen (2019) proposed distribution-free two-sample tests based on the concept of multivariate ranks, defined using optimal transport. Methods based on nearest neighbor distances have been also used extensively in other nonparametric statistical problems, such as density estimation Mack (1983), Mack and Rosenblatt (1979), nonparametric clustering Heckel and Bölcskei (2015), classification Cannings, Berrett and Samworth (2020), Cover and Hart (1967), Gadat, Klein and Marteau (2016), Samworth (2012), entropy and other functional estimation Berrett and Samworth (2019a), Berrett, Samworth and Yuan (2019), Kozachenko and Leonenko (1987) and testing problems, such as testing for normality Vasicek (1976), testing for uniformity Cressie (1976) and independence testing Berrett and Samworth (2019b), Goria et al. (2005).

One of the main challenges for devising a statistically correct test to detect viral remodeling from preferential infection is that the virus may infect different subpopulations at different rates. In Section 2 we show that even in very large sample sizes direct application of existing nonparametric two-sample tests can lead to erroneous inference. We expound this phenomenon by exhibiting explicit scenarios of preferential infection and remodeling where traditional tests fail in a simple setting of $d = 2$ markers. In Figure 3 the green triangles correspond to a sample of uninfected (UI) cells that arise from three different subpopulations while the red dots reflect the infected (VI) cells. The leftmost panel presents a setting where the virus has infected all the three cellular subpopulations and the overlap of the UI and VI cells indicate no remodeling. The middle panel presents a scenario where the cells have undergone remodeling under the influence of the virus, as is evident through a shift in the location of the VI cells. The rightmost panel reflects no remodeling but preferential infection. The different `g-tests` (Chen, Chen and Su (2018), Chen and Friedman (2017), Friedman and Rafsky (1979)), the cross-match test (Rosenbaum (2005)) and the energy test (Aslan and Zech (2005)) reject the null hypothesis of no remodeling in all the three cases and in each

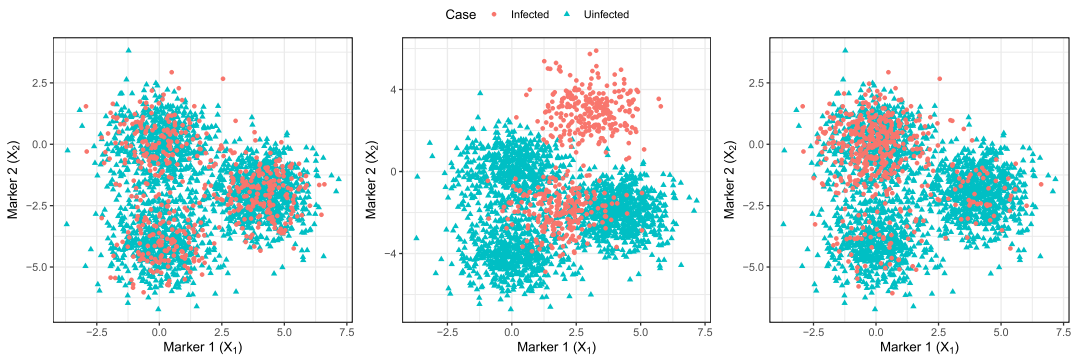

FIG. 3. *Schematic representation of viral remodeling of infected cells versus preferential viral infection with respect to $d = 2$ markers, X_1 and X_2. From left to right, we have (a) no remodeling, (b) remodeling and (c) no remodeling but preferential infection. Uninfected cells are in green whereas virus infected cells are in red.*

of the 100 simulation replications (see Table 3 in Section 3.1). This is not surprising because these tests are designed to test the simple null hypothesis of equality of the two distributions.

Due to the presence of subpopulation level heterogeneity, the problem of testing for remodeling warrants testing a composite null hypothesis. To this end, note that, under preferential infection, the two samples arise from the mixture distribution with identical component distributions but with different mixing weights. This is the case for the rightmost subplot in Figure 3. In this paper we formulate the problem of testing for preferential infection versus remodeling as a composite two-sample hypothesis with mixture distributions and develop a new nearest-neighbor based test that can consistently and efficiently detect the differences between the two samples.

1.4. *The* TRUH *testing framework*: *Novel attributes and our contributions*. In this article we propose a novel procedure for *Testing Remodeling under Heterogeneity* (TRUH) that effectively incorporates the underlying heterogeneity and imbalance in the samples and provides a conservative test for the composite null hypothesis that the two samples arise from the same mixture distribution but may differ with respect to the mixing weights. We summarize its key attributes below:

- The TRUH statistic is based on a nearest-neighbor approach (Cover and Hart (1967), Devroye, Györfi and Lugosi (1996)) that first relies on identifying for every infected cell a predictive precursor cell which is the phenotypically closest cell in the uninfected population. It then measures the relative dissimilarities between the infected cells and their predictive precursors and the predictive precursors to their most phenotypically similar uninfected cells. A large relative dissimilarity between the infected cells and their predictive precursors indicates surface protein regulation or remodeling by the virus, while a small relative dissimilarity provides evidence for preferential infection or no remodeling.
- We describe an efficient bootstrap based approach for calibrating the TRUH test statistic and evaluate its performance in finite-sample simulations. We then use this method to test for viral remodeling in tonsillar T cells under different types of HIV infection, corroborating the efficacy of our proposed procedure.
- We provide an extensive theoretical understanding of the large sample characteristics of our proposed test statistics. We establish the L_2-limit of our proposed statistic using asymptotic properties of functionals of random geometric graphs Penrose and Yukich (2003). The limit can be expressed in terms of the densities of the uninfected and infected populations and dimension dependent constants obtained from nearest-neighbor distances defined on a homogeneous Poisson process. Using these properties, we can select a cut-off for the TRUH statistic that is asymptotically consistent against biologically-relevant location alternatives. Traditional nonparametric tests enjoy these consistency properties in homogeneous populations but not under heterogeneity. We show that, using a nearest-neighbor based approach, this inefficiency of existing nonparametric tests in heterogeneous data can be mitigated.

The rest of the paper is organized as follows: In Section 2 we formulate the problem of testing for remodeling in single-cell virology as a heterogeneous two-sample problem, describe the TRUH framework and show how it can be calibrated using the bootstrap. Numerical experiments demonstrating the nonasymptotic performance of our testing procedure are given in Section 3. In Section 4 we use TRUH for studying remodeling in tonsillar T cells under different types of HIV infection. The asymptotic properties of the test statistic are discussed in Section 5. We conclude the paper in Section 6 with a discussion. The technical details and proofs of the theoretical results are given in the Supplementary Material (Banerjee, Bhattacharya and Mukherjee (2020)).

2. Statistical framework and the proposed TRUH statistic. In this section we formulate the problem of testing for remodeling in single-cell virology as a heterogeneous two-sample problem (Section 2.1), introduce the TRUH statistic (Section 2.2) and discuss how to calibrate it using the bootstrap (Section 2.3).

2.1. *The heterogeneous two-sample problem.* In our virology example the baseline constitutes the m uninfected cells. For each cell, $i \in \{1, \dots m\}$, we denote by U_i a d-dimensional vector of cellular characteristics typically measuring expressions corresponding to different genes or proteins. Denote the uninfected/baseline population by $\boldsymbol{U}_m = \{U_1, \dots, U_m\}$. Let F_0 be the cumulative distribution function (cdf) of the baseline population with the heterogeneity in the population being reflected by K different subgroups, each having unimodal distributions with distinct modes and cdfs $F_1, \dots, F_K$ and mixing proportions $w_1, \dots, w_K$, such that

$$F_0 = \sum_{a=1}^{K} w_a F_a, \quad \text{where } w_a \in (0, 1) \text{ and } \sum_{a=1}^{K} w_a = 1. \tag{2.1}$$

Note that the number of components K, the mixing distributions $F_1, \dots, F_K$ and the mixing weights $w_1, \dots, w_K$ are fixed (nonrandom) attributes which are unknown. Also, as $F_1, \dots, F_K$ are cdfs from unimodal distributions with distinct modes, F_0 is well defined with a unique specification. In addition to the uninfected population, we observe n i.i.d. infected observations $\boldsymbol{V}_n = \{V_1, \dots, V_n\}$ from a distribution function G in $\mathbb{R}^d$. Note that the infected and uninfected samples $\boldsymbol{U}_m$ and $\boldsymbol{V}_n$ are collected from separate experiments and are independent of each other.

Simple vs. composite null. In single-cell virology when an uninfected population is exposed to a pathogen, the virus may infect the different subpopulations at different rates. Therefore, even if the virus does not cause any change in the cellular characteristics, the virus infected sample might have different representations of the uninfected subpopulations than the uninfected mixing proportions $\{w_1, \dots, w_K\}$. As such, it is quite possible that a few of the uninfected subpopulations are completely absent in the viral population which, biologically, implies that the virus preferentially targets few cellular subpopulations. Thus, if the virus does not induce any change in the cellular characteristics, then the distribution of the infected population G lies in a class of distributions $\mathcal{F}(F_0)$ that contains any convex combination of $\{F_1, \dots, F_K\}$, including the boundaries, that is,

$$\mathcal{F}(F_0) = \left\{ Q = \sum_{a=1}^{K} \lambda_a F_a : \lambda_1, \lambda_2, \dots, \lambda_K \in [0, 1] \text{ and } \sum_{a=1}^{K} \lambda_a = 1 \right\}. \tag{2.2}$$

Note that the uninfected cdf F_0 is a particular member of the class $\mathcal{F}(F_0)$. If the virus induces changes in the cellular characteristics, then the viral population distribution would contain at least one nontrivial subpopulation with distribution substantially different from $\{F_1, F_2, \dots, F_K\}$ or their linear combinations. Thus, the test for viral remodeling is tantamount to testing the following composite null hypothesis:

$$H_0 : G \in \mathcal{F}(F_0) \quad \text{versus} \quad H_A : G \notin \mathcal{F}(F_0). \tag{2.3}$$

If the null hypothesis is accepted, we say the virus exhibits *preferential infection*, otherwise we say the virus exhibits *remodeling* (see Figure 6 below), and the hypothesis testing problem (2.3) will be referred to as the problem of *testing remodeling under heterogeneity* (TRUH). Later on, to facilitate proofs of the theoretical properties of our proposed method, we will assume that the baseline cdfs $F_1, \dots, F_K$ have unimodal densities $f_1, \dots, f_K$ (with respect to Lebesgue measure). In this case the baseline uninfected population will have density $f_0 = \sum_{a=1}^{K} w_a f_a$, and the set of distributions in (2.2) can be represented in terms of the densities $f_1, \dots, f_K$ and will be denoted by $\mathcal{F}(f_0)$.

Inefficiency of existing tests. Traditional nonparametric graph-based two-sample tests, such as the edge-count (EC) test of Friedman and Rafsky (1979) or the crossmatch (CM) test of Rosenbaum (2005), are tailored for the null hypothesis $H_0 : F_0 = G$, that is, testing whether the distributions of the uninfected samples $\boldsymbol{U}_m$ and the infected samples $\boldsymbol{V}_n$ are the same. However, not surprisingly, direct application of these tests to the composite hypothesis testing problem, described in (2.3) above, gives nonconservative procedures. To see this, consider the EC test. Recall that the EC test is based on the statistic $\mathcal{R}(\boldsymbol{U}_m, \boldsymbol{V}_n)$ which counts the number of edges in the minimal spanning tree (MST) of the pooled sample $\{U_1, \ldots, U_m, V_1, \ldots, V_n\}$ that connect points from different samples. Then, the EC test rejects the null hypothesis of $F_0 = G$ for small values of $\mathcal{R}(\boldsymbol{U}_m, \boldsymbol{V}_n)$. The cut-off for $\mathcal{R}(\boldsymbol{U}_m, \boldsymbol{V}_n)$ can be chosen based on the asymptotic distribution $\mathcal{R}(\boldsymbol{U}_m, \boldsymbol{V}_n)$ under $F_0 = G$, which was derived by Henze and Penrose (1999) in the usual limiting regime where $m, n \to \infty$ and $n/m \to \rho \in (0, \infty)$. In particular, it follows from Theorem 1 of Henze and Penrose (1999) that

$$\lim_{m,n\to\infty} \mathbb{P}_{F_0=G}\big(\mathcal{R}(\boldsymbol{U}_m, \boldsymbol{V}_n) < C_{m,n}(\alpha)\big) = \alpha, \tag{2.4}$$

with $C_{m,n}(\alpha) = \frac{2mn}{m+n} - z_{1-\alpha}\sigma_d\sqrt{m+n}$, where $z_{1-\alpha}$ is the αth quantile of the standard normal distribution, $\sigma_d^2 = \rho(4\rho + (1-\rho)^2\delta_d)/(1+\rho)^4$ and δ_d is a constant depending only on dimension d. More precisely, δ_d is the variance of the degree of the origin $\mathbf{0} \in \mathbb{R}^d$ in the minimal spanning tree built on a homogeneous Poisson process of rate 1 in $\mathbb{R}^d$ with the origin added to it. Note that (2.4) shows that the test with rejection region $\{\mathcal{R}(\boldsymbol{U}_m, \boldsymbol{V}_n) < C_{m,n}(\alpha)\}$ is asymptotically level α for the null hypothesis of $F_0 = G$.

The following proposition shows that direct application of the EC test, as described above, will not be conservative for testing the hypothesis (2.3) of viral remodeling. In fact, for cases of preferential infection but no remodeling the EC test will produce undesired false discoveries:

PROPOSITION 1. *Fix $\alpha \in (0, 1/2)$. Then, for F_0 as in (2.1) and for any $G \in \mathcal{F}(F_0) \setminus \{F_0\}$ in the usual limiting regime,*

$$\lim_{m,n\to\infty} \mathbb{P}\big(\mathcal{R}(\boldsymbol{U}_m, \boldsymbol{V}_n) < C_{m,n}(\alpha)\big) = 1,$$

with $\boldsymbol{U}_m = \{U_1, \ldots, U_m\}$ i.i.d. from f_0 and $\boldsymbol{V}_n = \{V_1, \ldots, V_m\}$ i.i.d. from g, where f_0 and g are the densities (with respect to the Lebesgue measure) of F_0 and G, respectively.

The proof of the above result is given in the Supplementary Material (Section A of Banerjee, Bhattacharya and Mukherjee (2020)). This shows that, for any level α, the EC test will be inconsistent, as it would reject with certainty all cases of preferential infection but no remodeling. This phenomenon is demonstrated in Figure 4 through a simple univariate simulation experiment. Here, we consider $m = 1000$, $n = 50$ and $d = 1$. The true distribution of the uninfected and infected subpopulations are Gaussian mixtures. We consider two cases:

- *Case A*: Here, F_0 and G are equal-weighted mixtures of three Gaussians, with each subpopulation in G having a different mean from those in F_0, that is, $F_0(u) = \frac{1}{3}\sum_{a=0}^{2} \Phi(u - 4a)$ and $G(u) = \frac{1}{3}\sum_{a=0}^{2} \Phi(u - 4a - 2)$.[2] This is a clear case of viral remodeling.
- *Case B*: Here, $F_0 = \frac{1}{3}\sum_{a=0}^{2} \Phi(u - 10a)$ and $G = \frac{1}{2}\sum_{a=0}^{1} \Phi(u - 20a)$. In this case there is preferential infection but no remodeling, that is, $G \in \mathcal{F}(F_0)$ with the middle population in F_0 being resistant to viral infection.

[2]Throughout, $\Phi(\cdot)$ and $\phi(\cdot)$ will denote the standard normal distribution function and density function, respectively.

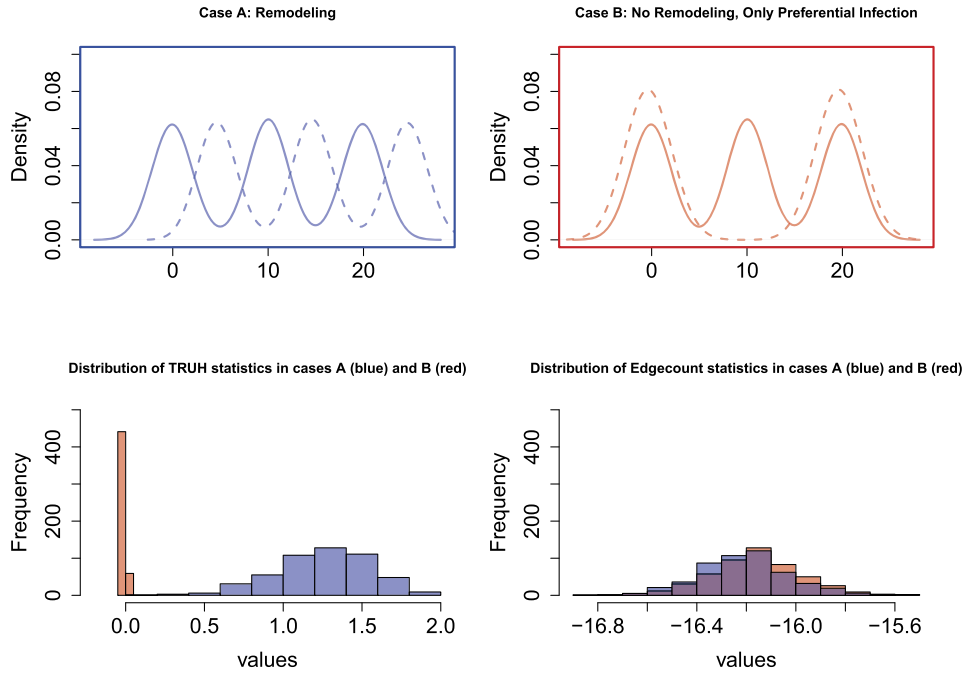

FIG. 4. *Simulation example showing the performance of edge-count test statistic versus the* TRUH *statistic. In the top row we describe the density of the true uninfected* F_0 *(in continuous line) and the density of the infected* G *(in dotted line) for the two cases. In both cases,* F_0 *and* G *are mixtures of normal distributions. In the first case, all the three equiprobable subpopulations in* F_0 *have undergone a discernible location change in* G*. In case B,* F_0 *again has three equiprobable subgroups, while* G *has two of those three subgroups. Thus, while case A signifies viral remodeling, there is no remodeling but only preferential infection in Case B. In the bottom row we have the histogram of the values of the* TRUH *statistic in the left (defined below in* (2.7)*) and the edge-count statistic in the right, respectively, under the two cases.*

Any test for the hypothesis (2.3) should ideally reject Case A and fail to reject Case B. However, Figure 4 shows that the histogram of EC test statistic values across 500 replications under cases A and B have a significant overlap. Table 1 shows the rejection rate (proportion of false discoveries) in Case B and power (proportion of true discoveries) in case A, as the level of the test is varied. From the table it is evident that there does not exists any choice of a critical value such that the rejection rate of the EC test in Case B is commendable, as it rejects all cases of preferential infection presented under Case B. On the other hand, our proposed test statistic (TRUH), described in the following section, entertains possibilities where both the rejection rate and the power attain the desired limit.

TABLE 1

The rejection rate and the power of the edgecount and TRUH *test statistics across* 500 *repetitions of the simulation setting of Figure* 4

	Level	0.01	0.05	0.10	0.20
Power in Case A	edgecount	1.000	1.000	1.000	1.000
	TRUH	1.000	1.000	1.000	1.000
Rejection rate in Case B	edgecount	1.000	1.000	1.000	1.000
	TRUH	0.000	0.000	0.000	0.038

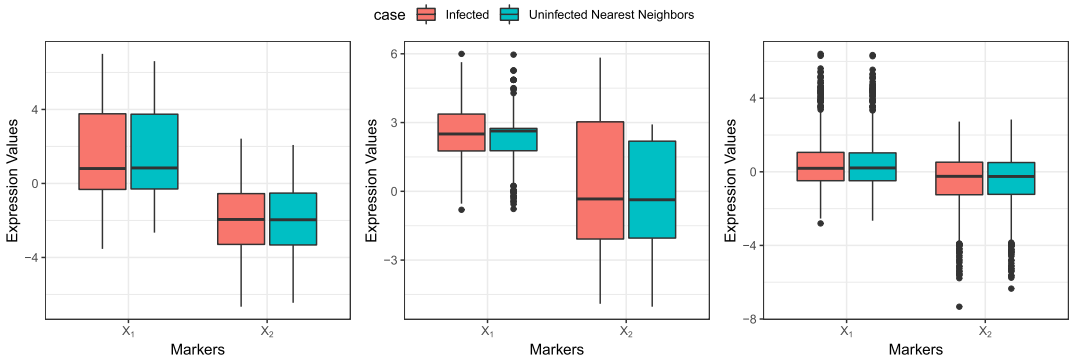

FIG. 5. *Boxplots of the coordinates of* $N_{\boldsymbol{V}_n,\boldsymbol{U}_m} = \{N(V_i, \boldsymbol{U}_m) : 1 \le i \le n\}$ *in green, adjacent to the boxplots of the coordinates of the corresponding infected cells* $\boldsymbol{V}_n$ *in red for each of the scenarios discussed under Figure 3. Recall that, from left to right, we have* (*a*) *no remodeling*, (*b*) *remodeling and* (*c*) *no remodeling but preferential infection.*

2.2. *Proposed test statistic*: TRUH. In this section we describe a nearest-neighbor based statistic for testing the hypothesis of remodeling. To this end, recall that $\boldsymbol{U}_m = \{U_1, \ldots, U_m\}$ is the uninfected sample and $\boldsymbol{V}_n = \{V_1, \ldots, V_n\}$ is the infected sample. Now, for each infected sample $V_i \in \boldsymbol{V}_n$, let

$$D_i = \min_{1 \le j \le m} \|V_i - U_j\|, \tag{2.5}$$

the Euclidean distance of V_i to its nearest point in the uninfected sample $\boldsymbol{U}_m$. The point in $\boldsymbol{U}_m$, which attains this minimum, will be denoted by $N(V_i, \boldsymbol{U}_m)$[3] and constitutes a key point in $\mathbb{R}^d$ for measuring the relative phenotypic difference between the infected cells and their closest uninfected counterparts. In Figure 5 we show the boxplots of the coordinates of $N_{\boldsymbol{V}_n,\boldsymbol{U}_m} = \{N(V_i, \boldsymbol{U}_m) : 1 \le i \le n\}$ in green, for each of the scenarios discussed under Figure 3. Recall from Figure 3 that we have, from left to right, (a) no remodeling, (b) remodeling and (c) no remodeling but preferential infection. We note that, for scenarios (a) and (c), the distributions of $N_{\boldsymbol{V}_n,\boldsymbol{U}_m}$ and $\boldsymbol{V}_n$ appear to overlap. However, in the case of remodeling (scenario (b) in the center plot), there is a clear difference between the two distributions for both the markers. The TRUH statistic captures this phenomenon and deals with the presence of heterogeneous groups (which can make the density within the uninfected sample $\boldsymbol{U}_m$ to vary greatly) by comparing D_i with a feature of the local density of $\boldsymbol{U}_m$ at $N(V_i, \boldsymbol{U}_m)$. For that purpose, define, for each infected observation,

$$C_i = \min_{1 \le j \le m : U_j \ne N(V_i, \boldsymbol{U}_m)} \|N(V_i, \boldsymbol{U}_m) - U_j\| \tag{2.6}$$

which is the distance of $N(V_i, \boldsymbol{U}_m)$ to its nearest neighbor in $\boldsymbol{U}_m$. Our proposed test statistic for testing (2.3), hereafter referred to as the TRUH statistic, is

$$T_{m,n} = \frac{1}{n^{1-\frac{1}{d}}} \left| \sum_{i=1}^{n} (D_i - C_i) \right| = n^{\frac{1}{d}} |\bar{D}_{m,n} - \bar{C}_{m,n}|, \tag{2.7}$$

[3]Given a finite set S and any point $x \in \mathbb{R}^d$, denote by $N(x, S) = \arg\min_{y \in S} \|x - y\|$, that is, the nearest neighbor of x in the set S. If there is a tie, that is, $N(x, S)$ has multiple elements, then we choose a random element from them and set that to $N(x, S)$. However, if the underlying distribution of the data has a continuous density, then there are no ties with probability 1.

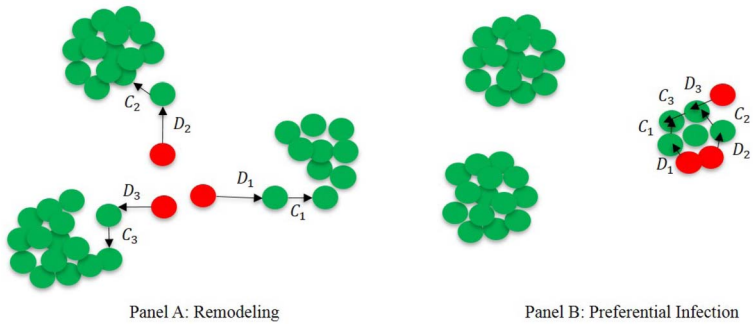

FIG. 6. *Panel A represents the scenario of remodeling, while Panel B exhibits Preferential Infection. Uninfected cells are in green, while infected cells are in red. The gaps are larger in case of remodeling, as infected cells are phenotypically different than their uninfected counterparts.*

where $\bar{D}_{m,n} = \frac{1}{n}\sum_{i=1}^{n} D_i$ and $\bar{C}_{m,n} = \frac{1}{n}\sum_{i=1}^{n} C_i$. Note that the TRUH statistic above is an aggregated measure of how far apart each viral cell is from the uninfected sample compared to the local distance between uninfected sample points in its vicinity. Consider, for example, panel A in Figure 6 that represents a schematic for remodeling, while panel B depicts preferential infection. Here, the three infected cells (in red) in Panel A are phenotypically different than their uninfected counterparts, and thus the average gap $|\bar{D}_{m,n} - \bar{C}_{m,n}|$ in Panel A, averaged over the three infected cells, is relatively larger than what is observed under preferential infection in Panel B. Therefore, we develop a test to reject the null hypothesis of no remodeling for large values of $T_{m,n}$. The cut-off for $T_{m,n}$ can be chosen based on a bootstrap calibration (Section 2.3) or using the asymptotic limit of $T_{m,n}$ (Section 5). Note that, since the nearest neighbor of a point in a cloud of n random points in $\mathbb{R}^d$ typically lies within a ball of radius $n^{-\frac{1}{d}}$ centered at that point, the TRUH statistic is scaled by $n^{1-\frac{1}{d}}$ which makes $T_{m,n}$ bounded in probability.

One of the interesting properties of the quantity $T_{m,n}$ is that it only involves enumeration of distance based features for the viral sample, unlike classical graph-based two-sample tests (Friedman and Rafsky (1979), Rosenbaum (2005)) which are built using the interpoint distances of the pooled sample. As a consequence, the TRUH test statistic is not symmetric in its usage of the uninfected and infected samples, even when the sample sizes are equal and the two samples were actually generated from the same population distribution. This asymmetric sample usage of TRUH helps in tackling possibly different heterogeneity levels in the two samples. Finally, note that, even though the quantities D_i and C_i are defined above using the Euclidean distance, they can be easily generalized to any arbitrary distance function, and the statistic $T_{m,n}$ can potentially be used in non-Euclidean data spaces, such as graph data or functional data, as well.

2.3. *Bootstrap-based calibration for* TRUH. In this section we present a bootstrap-based procedure to determine the cut-off $t_{m,n,\alpha}$ for a level α test using $T_{m,n}$. To this end, recall that $\mathcal{F}(F_0)$ contains any convex combination of the baseline distribution functions $\{F_1, \ldots, F_K\}$. Therefore, the proposed bootstrap procedure relies on the following two steps: (i) random sampling of the mixing proportions a large number of times, and (ii) for each such sampled mixing proportion, surrogate samples from $\mathcal{F}(F_0)$ are constructed to generate a pseudo null distribution which is used to estimate the level α cut-off. The maximum of all the level α cut-offs so obtained, one for each sampled mixing proportion, is then used to calibrate the TRUH statistic.

Our algorithm leverages the fact that, in our virology example, the number m of uninfected samples is much larger than the size n of the infected samples. Therefore, we can use the prediction strength approach of Tibshirani and Walther (2005) on the uninfected samples to obtain an estimate $\hat{K}$ of the unknown number of heterogeneous subgroups K. We then use this value of $\hat{K}$ to estimate the class memberships of the baseline samples $\boldsymbol{U}_m$ using a $\hat{K}$-means algorithm. For $1 \le a \le \hat{K}$, denote by $\hat{J}_a \subseteq \{1, 2, \ldots, m\}$ the subset of indices which belong to class a in the output of the $\hat{K}$-means algorithm. Let $\boldsymbol{U}_{\hat{J}_a} = \{U_i : i \in \hat{J}_a\}$ be the subset of the baseline samples estimated to be in the a^{th} class by the $\hat{K}$-means algorithm. Note that $\boldsymbol{U}_m = \{\boldsymbol{U}_{\hat{J}_a} : a = 1, 2, \ldots, \hat{K}\}$ and $\sum_{a=1}^{\hat{K}} m_a = m$, where $m_a = |\hat{J}_a|$.

Now, for each $b_1 = 1, \ldots, B_1$, denote by $(\lambda_1^{(b_1)}, \ldots, \lambda_{\hat{K}}^{(b_1)})$ a random sample from the $\hat{K}$-dimensional simplex $\mathcal{S}_{\hat{K}} = \{(z_1, \ldots, z_{\hat{K}}) \in \mathbb{R}^{\hat{K}} : z_a \in [0, 1], \text{ for } 1 \le a \le \hat{K}, \text{ and } \sum_{a=1}^{\hat{K}} z_a = 1\}$. Given the mixing weights $\{\lambda_1^{(b_1)}, \ldots, \lambda_{\hat{K}}^{(b_1)}\}$, we construct B_2 surrogate-infected samples from $\mathcal{F}(F_0)$ as follows: for each $b_2 = 1, \ldots, B_2$ and for $1 \le a \le \hat{K}$, randomly sample $\lceil n\lambda_a^{(b_1)} \rceil$ elements without replacement from $\boldsymbol{U}_{\hat{J}_a}$. Denote the chosen elements by

$$\mathcal{V}_a^{(b_2)} = \{U_1^{(b_2)}, \ldots, U_{\lceil n\lambda_a^{(b_1)} \rceil}^{(b_2)}\},$$

and set the remaining $m_a - \lceil n\lambda_a^{(b_1)} \rceil$ elements in $\boldsymbol{U}_{\hat{J}_a}$ as the residual baseline sample $\mathcal{U}_a^{(b_2)}$ in class a. Now, combining the samples over the $\hat{K}$ classes, we get the surrogate infected sample as $\boldsymbol{V}_n^{(b_2)} = \{\mathcal{V}_a^{(b_2)} : a = 1, \ldots, \hat{K}\}$ and the corresponding baseline sample as $\boldsymbol{U}_{\tilde{m}}^{(b_2)} = \{\mathcal{U}_a^{(b_2)} : a = 1, \ldots, \hat{K}\}$, where

$$\tilde{m} = \sum_{a=1}^{\hat{K}} (m_a - \lceil n\lambda_a^{(b_1)} \rceil).$$

Note that, under the null hypothesis of no remodeling ($G \in \mathcal{F}(F_0)$), the bootstrapped samples in the b_2^{th} round, $\boldsymbol{U}_{\tilde{m}}^{(b_2)}$ and $\boldsymbol{V}_n^{(b_2)}$ (which are surrogates for $\boldsymbol{U}_m$ and $\boldsymbol{V}_n$, respectively) can be used to compute the statistic

$$T_{\tilde{m},n}^{(b_2)} = n^{\frac{1}{d}} |\tau_{fc} \cdot \bar{D}_{\tilde{m},n} - \bar{C}_{\tilde{m},n}|. \tag{2.8}$$

For b_1 fixed, $T_{\tilde{m},n}^{(b_2)}$ is the surrogate of the TRUH statistic in the b_2^{th} bootstrap round. Observe that compared to (2.7), we have introduced a tuning parameter τ_{fc} in (2.8) above. We define it as the fold change (fc) hyperparameter and will consider values of $\tau_{fc} \ge 1$. Biologically relevant remodeling corresponds to significant fold change increase or decrease in the magnitude of cellular expressions between the infected and the uninfected cells. As we test the global null hypothesis of no change in any of the concerned genes, alternative hypothesis of remodeling with meager fold changes, if accepted, will only lead to biologically uninteresting discoveries. For discovering virologically interesting alternatives, it is natural to set τ_{fc} slightly larger than 1. (Note that $\tau_{fc} = 1$ corresponds to the bootstrapped version of the TRUH statistic in (2.7).) In the simulation experiments presented later in Section 3, we set $\tau_{fc} = 1$ whereas in Section 4 τ_{fc} is fixed at 1.1 as we study a real-world virology dataset.

The bootstrap procedure described above is summarized in Algorithm 1. The computational complexity of Algorithm 1 is driven by the following two steps: (i) the computation of the estimated number of clusters $\hat{K}$, and (ii) the computation of the TRUH test statistic over the bootstrap samples. While the calculations in step (ii) can be distributed across the $B_1 B_2$ bootstrap samples and n infected samples, the computational cost of estimating $T_{\tilde{m},n}^{(b)}$ for a fixed b is $O(md)$ which is the cost of running the 1-nearest neighbor algorithm twice for

Algorithm 1: Bootstrap cut-off for a level α test using $T_{m,n}$

Input: The parameters n, τ_{fc}, and α. The baseline sample $\boldsymbol{U}_m$, and the estimates $\hat{K}$ and $\{\hat{J}_a : a = 1, \ldots, \hat{K}\}$ from the K-means algorithm.

Output: The bootstrapped level α cutoff $t_{m,n,\alpha}$.

for $b_1 = 1, \ldots, B_1$ **do**

STEP 1: Random sample $\{\lambda_1^{(b_1)}, \ldots, \lambda_{\hat{K}}^{(b_1)}\}$ from the $\hat{K}$-dimensional simplex;

for $b_2 = 1, \ldots, B_2$ **do**

for $a = 1, \ldots, \hat{K}$ **do**

if $\lceil n\lambda_a^{(b_1)} \rceil \leq m_a$ **then**

STEP 2: Draw a simple random sample $\mathcal{V}_a^{(b_2)} = \{U_1^{(b_2)}, \ldots, U_{\lceil n\lambda_a^{(b_1)} \rceil}^{(b_2)}\}$ without replacement from $\boldsymbol{U}_{\hat{J}_a}$;

STEP 3: $\mathcal{U}_a^{(b_2)} = \boldsymbol{U}_{\hat{J}_a} \setminus \mathcal{V}_a^{(b_2)}$ the baseline residual sample in class a;

else

Stop: Go to STEP 1;

Surrogate Case sample: $\boldsymbol{V}_n^{(b_2)} = \{\mathcal{V}_a^{(b_2)} : a = 1, \ldots, \hat{K}\}$;

Baseline sample: $\boldsymbol{U}_{\tilde{m}}^{(b_2)} = \{\mathcal{U}_a^{(b_2)} : a = 1, \ldots, \hat{K}\}$;

STEP 4: Calculate $T_{\tilde{m},n}^{(b_2)} = n^{\frac{1}{d}} |\tau_{fc}\, \bar{D}_{\tilde{m},n} - \bar{C}_{\tilde{m},n}|$;

STEP 5: Return $t_{m,n,\alpha}^{(b_1)} = \min\{T_{\tilde{m},n}^{(b_2)} : \frac{1}{B_2} \sum_{r=1}^{B_2} \mathbf{1}\{T_{\tilde{m},n}^{(r)} \geq T_{\tilde{m},n}^{(b_2)}\} \leq \alpha\}$.

STEP 6: Return $t_{m,n,\alpha} = \max\{t_{m,n,\alpha}^{(b_1)} : 1 \leq b_1 \leq B_1\}$.

each of the n infected samples. To estimate K, we use prediction strength along with a K-means algorithm where the target number of clusters and the maximum number of iterations, over which the K-means algorithm runs before stopping are both fixed, and thus has $O(md)$ complexity. Therefore, the overall computational complexity of Algorithm 1 is $O(md)$. For the numerical experiments and real data analysis of Sections 3 and 4, we set $B_2 = 200$ and implement a version of Algorithm 1 which samples the mixing proportions $\{\lambda_1, \ldots, \lambda_{\hat{K}}\}$ only from the corners of the $\hat{K}$ dimensional simplex $\mathcal{S}_{\hat{K}}$ as follows: we set $B_1 = \hat{K}$ and for $b_1 = 1, \ldots, B_1$, and $a = 1, \ldots, \hat{K}$, we take $\lambda_a^{(b_1)} = 1$ if $b_1 = a$ and 0 otherwise. This sampling scheme ensures that the mechanism for generating the mixing proportions places most weight on the corners of $\mathcal{S}_{\hat{K}}$.

3. Numerical experiments. In this section we evaluate the numerical performance of the `TRUH` procedure across a wide range of simulation experiments. We consider the following six competing testing procedures that use different methodologies to conduct a nonparametric two-sample test: (i) Energy test (`Energy`) of Aslan and Zech (2005) available from the R package `energy`, (ii) Cross-Match test (`Crossmatch`) of Rosenbaum (2005) available from the R package `crossmatch`, (iii) edgecount test (`E Count`) of Friedman and Rafsky (1979), (iv) Generalized edgecount test (`GE Count`) of Chen and Friedman (2017), (v) Weighted edgecount test (`WE Count`) of Chen, Chen and Su (2018) and (vi) the Max Type edgecount test (`MTE Count`) of Zhang and Chen (2017). The aforementioned four edge count-based tests are available from the R package `gtests`. We note that the preceding six testing procedures are not designed to test the composite null hypothesis of equation (2.3) and rely on a simple null hypothesis $H_0 : F_0 = G$ for inference. Nevertheless, the simulation

experiments presented in this section highlight the incorrect inference that may result when traditional two-sample tests are used for testing the composite null hypothesis of equation (2.3).

To assess the performance of the competing testing procedures, we simulate $\boldsymbol{U}_m$ and $\boldsymbol{V}_n$ from F_0 and G, the cdf of the baseline and the infected population, respectively, and for each testing procedure, we measure the proportion of rejections across 100 repetitions of the composite null hypothesis test, described in (2.3) at 5% level of significance. For `TRUH`, we use Algorithm 1 with fold change constant $\tau_{fc} = 1$, $B_2 = 200$ and sample the mixing proportions only from the corners of $\hat{K}$ dimensional simplex $\mathcal{S}_{\hat{K}}$, as described in Section 2.3. The R code that reproduces our simulation results is available in the Supplementary Material (Banerjee, Bhattacharya and Mukherjee (2020)) and online at https://github.com/trambakbanerjee/TRUH_paper. The `TRUH` R package is available at https://github.com/trambakbanerjee/TRUH.

3.1. *Experiment* 1. In the setup of Experiment 1, we consider testing $H_0 : G \in \mathcal{F}(F_0)$vs.$H_A : G \notin \mathcal{F}(F_0)$, when F_0 is the cdf of a d dimensional Gaussian mixture distribution with three components,

$$F_0 = 0.3\mathcal{N}_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + 0.3\mathcal{N}_d(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + 0.4\mathcal{N}_d(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3),$$

where $\boldsymbol{\mu}_1 = \mathbf{0}_d$, $\boldsymbol{\mu}_2 = -3\mathbf{1}_d$, $\boldsymbol{\mu}_3 = -\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}_K$, for $K = 1, 2, 3$, are d dimensional positive definite matrices with eigenvalues randomly generated from the interval $[1, 10]$. To simulate $\boldsymbol{V}_n$ from G, we consider two scenarios as follows:

- Scenario I: Here, $G = 0.1\mathcal{N}_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + 0.1\mathcal{N}_d(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + 0.8\mathcal{N}_d(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$. In this case, G has all the subpopulations present in F_0 but at different proportions. Thus, $G \in \mathcal{F}(F_0)$, and the correct inference here is no remodeling.
- Scenario II: This setting presents a scenario where $G \notin \mathcal{F}(F_0)$ and the composite null H_0 is not true. Here, we consider $G = 0.5\mathcal{N}_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + 0.5\mathcal{N}_d(\boldsymbol{\mu}_4, \boldsymbol{\Sigma}_4)$, where $\boldsymbol{\Sigma}_4$ is a d dimensional positive definite matrix generated independently of $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$, $\boldsymbol{\Sigma}_3$, and $\boldsymbol{\mu}_4 = 4\boldsymbol{\epsilon}_d$, where $\boldsymbol{\epsilon}_d$ is a vector of d independent Rademacher random variables.

For Scenario I, Table 2 reports the rejection rates for 100 repetitions of the test for varying d, m, n when the parameters $\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i 1 \leq i \leq 4\}$ are held fixed across these repetitions. We see that `TRUH` returns the smallest rejection rate. The other six tests all have very high rejection rates as they fail to account for the composite nature of the null hypothesis. The rejection rate for `TRUH` is below the prespecified 0.05 level establishing that it is a conservative test across all the regimes considered in the table. In Scenario II, however, we find that all the

TABLE 2

Rejection rates at 5% level of significance: *Experiment* 1 *and Scenario I wherein* $H_0 : G \in \mathcal{F}(F_0)$ *is true*

	$m = 500, n = 50$			$m = 2000, n = 200$		
Method	$d = 5$	$d = 15$	$d = 30$	$d = 5$	$d = 15$	$d = 30$
`Energy`	1.000	1.000	1.000	1.000	1.000	1.000
`Crossmatch`	0.220	0.150	0.145	0.460	0.410	0.340
`E Count`	0.185	0.115	0.055	0.400	0.335	0.195
`GE Count`	0.170	0.185	0.225	0.510	0.540	0.605
`WE Count`	0.300	0.295	0.360	0.655	0.745	0.735
`MTE Count`	0.230	0.230	0.290	0.605	0.665	0.665
`TRUH`	0.02	0.015	0.015	0.01	0.02	0.01

TABLE 3
Rejection rates at 5% level of significance: Simulation experiment corresponding to Figure 3

	$m = 2000, n = 500, d = 2$		
Method	Left panel: no remodeling ($G \in \mathcal{F}(F_0)$)	Center panel: remodeling ($G \notin \mathcal{F}(F_0)$)	Right panel: preferential infection ($G \in \mathcal{F}(F_0)$)
`Energy`	0.030	1.000	1.000
`Crossmatch`	0.030	1.000	1.000
`E Count`	0.010	1.000	1.000
`GE Count`	0.000	1.000	1.000
`WE Count`	0.060	1.000	1.000
`MTE Count`	0.030	1.000	1.000
`TRUH`	0.000	0.980	0.000

tests correctly identify $G \notin \mathcal{F}(F_0)$ in all the regimes and across all replications. This shows that all the tests exhibit perfect rejection rates in this scenario. These two scenarios under Experiment 1 demonstrate that for testing the composite null hypothesis of equation (2.3), direct application of traditional two-sample tests, such as those considered here, is no longer conservative as these tests rely on a simple null hypothesis for inference. `TRUH`, on the other hand, is adept at detecting $H_0 : G \in \mathcal{F}(F_0)$ and powerful against departures from H_0.

In Table 3 we present the results of the simulation exercise that correspond to the three scenarios described in Figure 3. The two dimensional uninfected marker expressions (X_1, X_2) are randomly sampled from $F_0 = w_1\mathcal{N}_2(\boldsymbol{\mu}_1, \boldsymbol{I}_2) + w_2\mathcal{N}_2(\boldsymbol{\mu}_2, \boldsymbol{I}_2) + w_3\mathcal{N}_2(\boldsymbol{\mu}_3, \boldsymbol{I}_2)$, where $\boldsymbol{\mu}_1 = \mathbf{0}$, $\boldsymbol{\mu}_2 = (0, -4)$, $\boldsymbol{\mu}_3 = (4, -2)$ and the sample size is $m = 2000$. The mixing weights are given by $(w_1, w_2, w_3) = (0.3, 0.3, 0.4)$. For the panel on the left of Figure 3, infected marker expressions arise from F_0 but with sample size $n = 500$, while for the center panel the infected marker expressions represent a random sample of size n from $G = 0.5\mathcal{N}_2(\boldsymbol{\mu}_4, \boldsymbol{I}_2) + 0.5\mathcal{N}_2(\boldsymbol{\mu}_5, \boldsymbol{I}_2)$, where $\boldsymbol{\mu}_4 = 0.25\boldsymbol{\mu}_2 + 0.5\boldsymbol{\mu}_3$ and $\boldsymbol{\mu}_5 = (3/4)\boldsymbol{\mu}_2 + (9/8)\boldsymbol{\mu}_3$. Clearly, in this case $G \notin \mathcal{F}(F_0)$. For the right-most panel, infected marker expressions are again a random sample of size n from $\mathcal{F}(F_0)$ with mixing weights given by the vector $(w_1, w_2, w_3) = (0.8, 0.1, 0.1)$. Under this setting the three `g-tests` (Chen, Chen and Su (2018), Chen and Friedman (2017), Friedman and Rafsky (1979)), the cross-match test of Rosenbaum (2005) and the energy test of Aslan and Zech (2005) infer $G \notin \mathcal{F}(F_0)$ in all of the 100 repetitions of the experiment thus suggesting their inability to tackle subpopulation level heterogeneity.

3.2. *Experiment* 2. For Experiment 2 we consider a more complex setup wherein F_0 is the cdf of a d dimensional mixture distribution which is not necessarily Gaussian. Here,

$$F_0 = 0.5\,\mathrm{Gam}_d(\mathrm{shape} = 5\mathbf{1}_d, \mathrm{rate} = \mathbf{1}_d, \boldsymbol{\Sigma}_1) + 0.5\,\mathrm{Exp}_d(\mathrm{rate} = \mathbf{1}_d, \boldsymbol{\Sigma}_2),$$

where Gam_d and Exp_d are d dimensional Gamma and exponential distributions. For generating correlated Gamma and exponential variables, we use the Gaussian copula approach based function from the R-package `lcmix` (Dvorkin (2012), Song (2000)). We consider tapering matrices with positive and negative autocorrelations: $(\boldsymbol{\Sigma}_1)_{ij} = 0.7^{|i-j|}$ and $(\boldsymbol{\Sigma}_2)_{ij} = -0.9^{|i-j|}$ for $1 \le i, j \le d$. For simulating $\boldsymbol{V}_n$ from G, we consider the following two scenarios:

- Scenario I: Here, $G = \mathrm{Exp}_d(\mathrm{rate} = \mathbf{1}_d, \boldsymbol{\Sigma}_2)$. In this case, G arises from only one of the components of F_0, that is, $G \in \mathcal{F}(F_0)$.

TABLE 4
Rejection rates at 5% level of significance: Experiment 2 and Scenario I wherein $H_0 : G \in \mathcal{F}(F_0)$ *is true*

	$m = 500, n = 50$			$m = 2000, n = 200$		
Method	$d = 5$	$d = 15$	$d = 30$	$d = 5$	$d = 15$	$d = 30$
`Energy`	1.000	1.000	1.000	1.000	1.000	1.000
`Crossmatch`	0.460	0.440	0.390	0.800	0.850	0.760
`E Count`	0.290	0.190	0.280	0.720	0.690	0.560
`GE Count`	0.400	0.430	0.390	0.900	0.920	0.900
`WE Count`	0.560	0.590	0.600	0.970	0.960	0.940
`MTE Count`	0.460	0.510	0.440	0.930	0.950	0.910
`TRUH`	0.000	0.000	0.000	0.000	0.000	0.000

- Scenario II: Here, $G = 0.1\,\mathrm{Gam}_d(\text{shape} = 10\mathbf{1}_d, \text{rate} = 0.5\mathbf{1}_d, \boldsymbol{\Sigma}_1) + 0.9\,\mathrm{Exp}_d(\text{rate} = \mathbf{1}_d, \boldsymbol{\Sigma}_2)$. In this setting, $G \notin \mathcal{F}(F_0)$ and the composite null H_0 is not true. When the ratio n/m is small, this scenario presents a difficult setting for detecting departures from H_0 as majority of the case samples from $\boldsymbol{V}_n$ will arise from $\mathrm{Exp}_d(\text{rate} = \mathbf{1}_d, \boldsymbol{\Sigma}_2)$ and the tests will rely on only a small fraction of samples from $\mathrm{Gam}_d(\text{shape} = 10\mathbf{1}_d, \text{rate} = 0.5\mathbf{1}_d, \boldsymbol{\Sigma}_1)$ to reject the null hypothesis.

Table 4 reports the rejection rates, for 100 repetitions, of the different tests in Scenario I. Note that `TRUH` correctly identifies that $G \in \mathcal{F}(F_0)$ while the remaining tests overwhelmingly support $G \notin \mathcal{F}(F_0)$, especially when m is large, demonstrating their lack of conservatism in testing the composite null hypothesis of the form (2.3). The results for Scenario II (Table 5) are reported for $n/m = 0.02$, where, with the exception of `Energy` test, all the other competing tests demonstrate small rejection rates for $m = 500$. Substantial improvement in the rejection rates is evident when $m = 2000$. However, for both these cases, $m = 500$ and $m = 2000$, the `Energy` test followed by `TRUH` exhibit the largest rejection rates. Although `Energy` test rejects H_0 in almost all of the testing instances in Scenario II, its performance in Scenario I (Table 4) reveals that it can be severely nonconservative when testing under a composite null hypothesis $H_0 : G \in \mathcal{F}(F_0)$.

3.3. *Experiment* 3. For Experiment 3 we introduce zero inflation in both the baseline and case samples to mimic the scenario that is often encountered in virology studies wherein some of the markers exhibit only a small probability of expressing themselves. We let $\boldsymbol{p} = (p_1, \ldots, p_d)$ denote the d dimensional vector of point masses at 0 across dimensions and

TABLE 5
Rejection rates at 5% level of significance: Experiment 2 and Scenario II wherein $H_0 : G \in \mathcal{F}(F_0)$ *is false*

	$m = 500, n = 10$			$m = 2000, n = 40$		
Method	$d = 5$	$d = 15$	$d = 30$	$d = 5$	$d = 15$	$d = 30$
`Energy`	0.930	0.960	1.000	1.000	1.000	1.000
`Crossmatch`	0.400	0.350	0.470	0.600	0.720	0.720
`E Count`	0.180	0.120	0.130	0.340	0.310	0.200
`GE Count`	0.310	0.230	0.160	0.800	0.790	0.770
`WE Count`	0.510	0.490	0.460	0.800	0.790	0.790
`MTE Count`	0.390	0.430	0.380	0.800	0.780	0.770
`TRUH`	0.580	0.580	0.580	0.880	0.940	0.960

consider

$$F_0 = 0.5F_1 + 0.5F_2,$$

where $F_1 = \boldsymbol{p}\delta_{\{\mathbf{0}\}} + (\mathbf{1}_d - \boldsymbol{p})\,\mathrm{Gam}_d(\text{shape} = 5\mathbf{1}_d, \text{rate} = \mathbf{1}_d, \boldsymbol{\Sigma}_1)$ and $F_2 = \boldsymbol{p}\delta_{\{\mathbf{0}\}} + (\mathbf{1}_d - \boldsymbol{p})\,\mathrm{Exp}_d(\text{rate} = \mathbf{1}_d, \boldsymbol{\Sigma}_2)$. In the above representation, $\boldsymbol{p}$ regulates the differential zero inflation across the d dimensions. For the purposes of this experiment, we chose the first $0.8d$ coordinates of $\boldsymbol{p}$ independently from $\mathrm{Unif}(0.5, 0.6)$, and the remaining $0.2d$ coordinates are set to 0. Thus, the zero inflation is encountered only in the first $0.8d$ coordinates of F_0. To simulate the baseline sample $\boldsymbol{U}_m$ from F_0, we use the R-package `lcmix` with $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ as described in Experiment 2 (Section 3.2). For simulating $\boldsymbol{V}_n$ from G, we consider the following two scenarios:

- Scenario I: Let $G = \boldsymbol{p}\delta_{\{\mathbf{0}\}} + (\mathbf{1}_d - \boldsymbol{p})\,\mathrm{Exp}_d(\text{rate} = \mathbf{1}_d, \boldsymbol{\Sigma}_2)$. Here, G arises from only one of the components of F_0, that is, $G \in \mathcal{F}(F_0)$.
- Scenario II: Here, we let $G = 0.5G_1 + 0.5G_2$, where $G_1 = \boldsymbol{q}\delta_{\{\mathbf{0}\}} + (\mathbf{1}_d - \boldsymbol{q})\,\mathrm{Gam}_d(\text{shape} = 5\mathbf{1}_d, \text{rate} = 0.5\mathbf{1}_d, \boldsymbol{\Sigma}_1)$ and $G_2 = \boldsymbol{q}\delta_{\{\mathbf{0}\}} + (\mathbf{1}_d - \boldsymbol{q})\,\mathrm{Exp}_d(\text{rate} = \mathbf{1}_d, \boldsymbol{\Sigma}_2)$, and we set the first $0.8d$ coordinates of $\boldsymbol{q}$ to 0.3 and the remaining $0.2d$ coordinates to 0. Note that in this setting, apart from the difference in the rate parameter of the Gamma distribution, we also have differential zero inflation across G and F_0, as $\boldsymbol{q} \neq \boldsymbol{p}$. Thus, $G \notin \mathcal{F}(F_0)$ and the composite null H_0 is not true. Moreover, when n is small, this scenario presents a challenging setting for detecting departures from H_0, as the tests will have to rely on both the differences in the rate parameter and differential zero expression between $\boldsymbol{U}_m$, $\boldsymbol{V}_n$ to reject the null hypothesis.

Tables 6 and 7 report the rejection rates for 100 repetitions of the test when $\boldsymbol{p}$ is held fixed across these repetitions. For Scenario I (Table 6), we see that `TRUH`, unlike the other six tests, does not excessively reject the null hypothesis and is the only conservative test. In Scenario II (Table 7), when $n = 10$, the `TRUH` and `Energy` tests dominate all the remaining tests and reject H_0 in more than 80% of the testing instances. However, when $n = 40$, all tests are competitive, with the exception of the `Crossmatch` for $d < 30$. Overall, across the above two zero-inflated scenarios, `TRUH` is both conservative and powerful against departures from the null hypothesis $H_0 : G \in \mathcal{F}(F_0)$.

4. Remodeling analysis of HIV-infected T cells. In this section we analyze the data collected in Cavrois et al. (2017). It contains protein expressions of uninfected and HIV-infected CD4 (which is a protein found on the surface of immune cells) positive tonsillar T cells. We show that existing two-sample tests,which rely on a simple null hypothesis for

TABLE 6

Rejection rates at 5% level of significance: Experiment 3 and Scenario I wherein $H_0 : G \in \mathcal{F}(F_0)$ is true

	$m = 500, n = 50$			$m = 2000, n = 200$		
Method	$d = 5$	$d = 15$	$d = 30$	$d = 5$	$d = 15$	$d = 30$
`Energy`	1.000	1.000	1.000	1.000	1.000	1.000
`Crossmatch`	0.340	0.330	0.240	0.730	0.800	0.670
`E Count`	0.200	0.120	0.100	0.670	0.440	0.340
`GE Count`	0.300	0.290	0.330	0.870	0.860	0.870
`WE Count`	0.510	0.460	0.540	0.970	0.920	0.930
`MTE Count`	0.400	0.350	0.460	0.890	0.910	0.920
`TRUH`	0.000	0.000	0.000	0.000	0.000	0.000

TABLE 7
Rejection rates at 5% level of significance: Experiment 3 and Scenario II wherein $H_0 : G \in \mathcal{F}(F_0)$ is false

	$m = 500, n = 10$			$m = 2000, n = 40$		
Method	$d = 5$	$d = 15$	$d = 30$	$d = 5$	$d = 15$	$d = 30$
`Energy`	0.850	0.920	0.940	1.000	1.000	1.000
`Crossmatch`	0.460	0.410	0.590	0.820	0.730	0.970
`E Count`	0.410	0.520	0.730	0.890	0.990	1.000
`GE Count`	0.410	0.480	0.730	0.920	0.960	1.000
`WE Count`	0.550	0.580	0.780	0.920	0.920	1.000
`MTE Count`	0.590	0.570	0.810	0.900	0.960	1.000
`TRUH`	0.760	0.940	0.980	0.970	1.000	1.000

inference, may lead to biologically incorrect inference when testing the composite null hypothesis of equation (2.3). Our proposed `TRUH` hypothesis testing framework, on the other hand, is proficient at detecting $H_0 : G \in \mathcal{F}(F_0)$ and powerful against departures from H_0.

As discussed in Section 1.1, the goal in Cavrois et al. (2017) was to conduct a mass cytometric assessment of subsets of CD4+ T cells that support HIV entry and viral infection in humans using two variants of the HIV virus: `Nef rich HIV` and `Nef deficient HIV`. It is known in the immunology literature that `Nef-rich` cells are more prone to viral remodeling (Basmaciogullari and Pizzato (2014)). The data set we analyze here contains uninfected and infected data from two different sets of experiments. Both the experiments have four replications based on tonsillar T cells from four healthy donors. In Experiment I, the infection was done by `Nef-rich` HIV, whereas in Experiment II the infection was done by `Nef-deficient` HIV. We expect remodeling, if any, in the infected cells to be higher in Experiment I than in Experiment II compared to their respective baseline uninfected populations.

The cells in the data were phenotyped in a 38 parameter CyToF (Bendall et al. (2014)) panel after allowing four days for infection. The panel used *three* markers to classify the cells as uninfected or infected which leaves $d = 35$ of the original 38 markers for our analyses. For donor r, let $\boldsymbol{U}_{m,r} = \{U_{1,r}, \ldots, U_{m,r}\}$ denote the uninfected sample where each $U_{j,r}$ is a d dimensional vector of arcsin transformed marker expression values with cdf F_0. We assume that the heterogeneity in the uninfected population is captured by K heterogeneous cellular subgroups with each having unimodal probability distribution functions with cdfs $F_1, F_2, \ldots, F_K$ and mixing proportions $w_1, w_2, \ldots, w_K$, such that F_0 is of the form represented in equation (2.1). We observe the virus infected sample $\boldsymbol{V}_{n,r} = \{V_{1,r}, \ldots, V_{n,r}\}$ consisting of n i.i.d. d-dimensional arcsin transformed observations from G and the goal is to test $H_0 : G \in \mathcal{F}(F_0)$ vs. $H_A : G \notin \mathcal{F}(F_0)$, where $\mathcal{F}(F_0)$ is the convex hull of $\{F_1, \ldots, F_K\}$ as defined in equation (2.2). Note that rejection of the null hypothesis would indicate that the distribution of the marker expressions under infection is different from F_0 and any convex combination of its components, thus providing evidence in favor of remodeling. Virologists study remodeling in virus infected cells in reference to the expressions of bystander cells. In panels of cells subjected to infection by the virus, not all of the cells get infected. Bystanders are those cells which are not directly infected by the virus but are neighbors of virus infected cells. In these experiments it was seen that, when τ_{fc} is set to 1, then even bystander cells exhibit remodeling in some experiments. However, when τ_{fc} is set to 1.1, there is no remodeling in the bystander population in any experiments. Thus, to detect biologically relevant cases of remodeling and avoid discovering benign instances, we use $\tau_{fc} = 1.1$ throughout this section to obtain the bootstrapped null distribution of the `TRUH` statistic.

Among the 35 markers considered here, it is known that the expressions of the four markers CD4, CCR5, CD28 and CD62L are changed due to HIV infection and these four markers play a significant role in HIV induced remodeling (Garcia and Miller (1991), Matheson et al. (2015), Michel et al. (2005), Ross, Oran and Cullen (1999), Swigut, Shohdy and Skowronski (2001), Vassena et al. (2015)). Consider two testing problems: (A) in which we test the hypothesis for all 35 markers, and (B) in which we test the hypothesis of viral remodeling on 31 markers leaving aside the four markers which are known to be remodeled by HIV. Thus, here we have four different cases on which we conduct the tests of viral remodeling, viz.:

- CASE 1 corresponds to Experiment I A where we test viral remodeling on Nef-rich infected cells based on all 35 markers, including the four which are known to be remodeled.
- CASE 2 corresponds to Experiment I B where we test viral remodeling on Nef-rich infected cells based on 31 markers which are known to be mainly invariant under HIV infection.
- CASE 3 corresponds to Experiment II A where we test viral remodeling on Nef-deficient infected cells based on all 35 markers, including the four which are known to be remodeled.
- CASE 4 corresponds to Experiment II B where we test viral remodeling on Nef-deficient infected cells based on 31 markers which are known to be mainly invariant under HIV infection.

In all of the four cases, we have four replications corresponding to four donors. It has been established through validation experiments in Cavrois et al. (2017) that there is no remodeling but only preferential infection in cases 2 and 4 whereas cases 1 and 3 exhibit remodeling with the intensity of remodeling being much higher in the former than the later. Biologically, it corresponds to the fact that there is Nef-independent remodeling, but the intensity of remodeling is higher in presence of Nef. Also, remodeling in cellular expressions is confined to the four markers CD4, CCR5, CD28 and CD62L in the set of markers considered in the study. Figures 7 and 8 present t-SNE plots (van der Maaten and Hinton (2008)) of the data where the d dimensional uninfected and infected cellular expression levels are projected to a two-dimensional space for each of the four donors across the four cases. While these plots exhibit the underlying heterogeneity in the uninfected sample and the sample size imbalance, instances of remodeling are also visible in cases 1 and 3 (Figure 7) wherein a relatively large fraction of the infected cells in red occupy a distinct position in the two-dimensional space with no overlap with their uninfected counterparts.

For conducting statistical hypothesis tests for the above four cases, along with our proposed TRUH procedure, we also use the siz other competing tests statistics described in Section 3 which are the Energy test (Aslan and Zech (2005)), CrossMatch (Rosenbaum (2005)), E Count(Friedman and Rafsky (1979)), GE Count (Chen and Friedman (2017)), WE Count (Chen, Chen and Su (2018)) and MTE Count (Zhang and Chen (2017)). As discussed in Section 3, these six testing procedures are not designed to test the composite null hypothesis of equation (2.3) and rely on a simple null hypothesis $H_0 : F_0 = G$ for inference. In this section we highlight the biologically incorrect inference that may result when these tests are used for testing the composite null hypothesis of no remodeling.

Figure 9 presents the values of the TRUH statistic and the 2.5th, 50th, 97.5th percentiles of the associated null distribution. From the plots it is evident that, at 5% level, our proposed procedure correctly captures the biological phenomena of remodeling or no remodeling across the four cases. The other six tests fail to correctly detect the phenomena in some of the four cases due to heterogeneity in the data. Next, we describe the results in further detail. In Tables 8 and 9 we report the p-values of the seven competing tests statistics for testing remodeling under HIV infection in Nef-rich environment. In Table 8 all seven tests reject

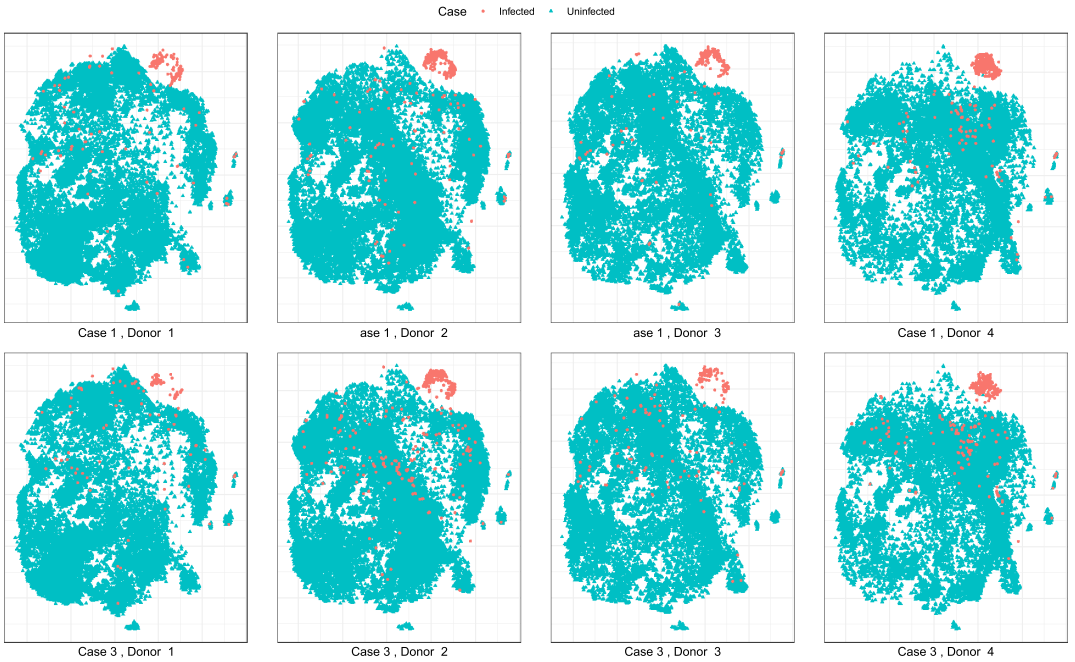

FIG. 7. *This is a t-SNE plot* (*van der Maaten and Hinton* (2008)) *of the data for Cases* 1 *and* 3 *where the* $d = 35$ *dimensional uninfected and infected cellular expression levels are projected to a two-dimensional space for each of the four donors.*

the null hypothesis of no remodeling, thus verifying that CD4+ T cells exhibit remodeling under the influence of `Nef` rich HIV infection. In Table 9, however, we present the p-values of the tests when the four cell surface markers, `CD4`, `CCR5`, `CD28` and `CD62L`, known

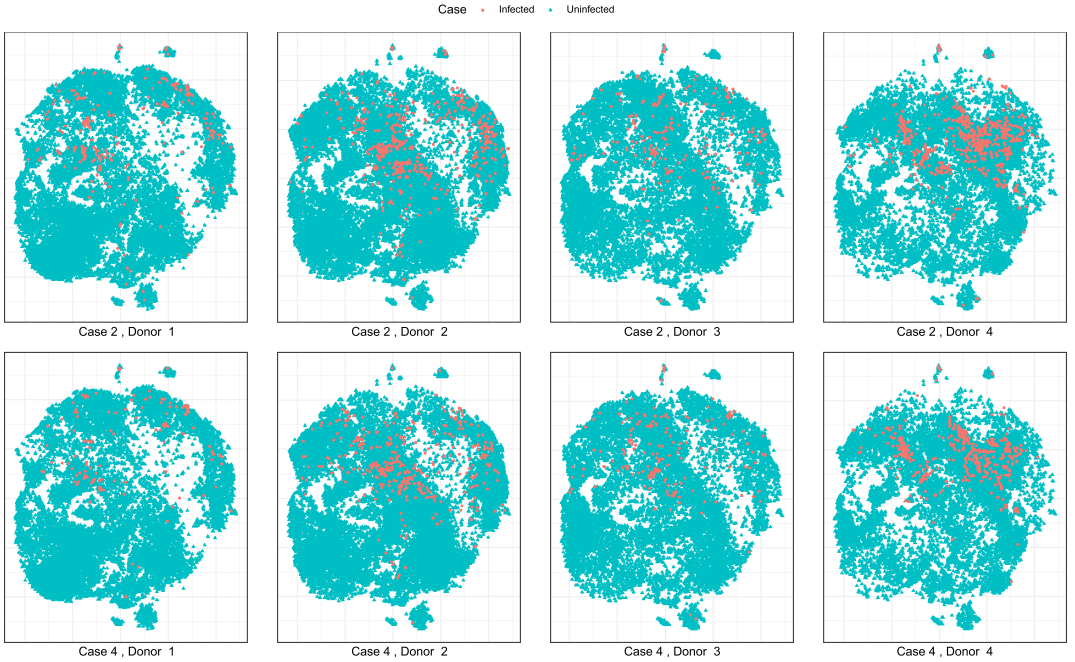

FIG. 8. *This is a t-SNE plot of the data for Cases* 2 *and* 4 *where the* $d = 31$ *dimensional uninfected and infected cellular expression levels are projected to a two-dimensional space for each of the four donors.*

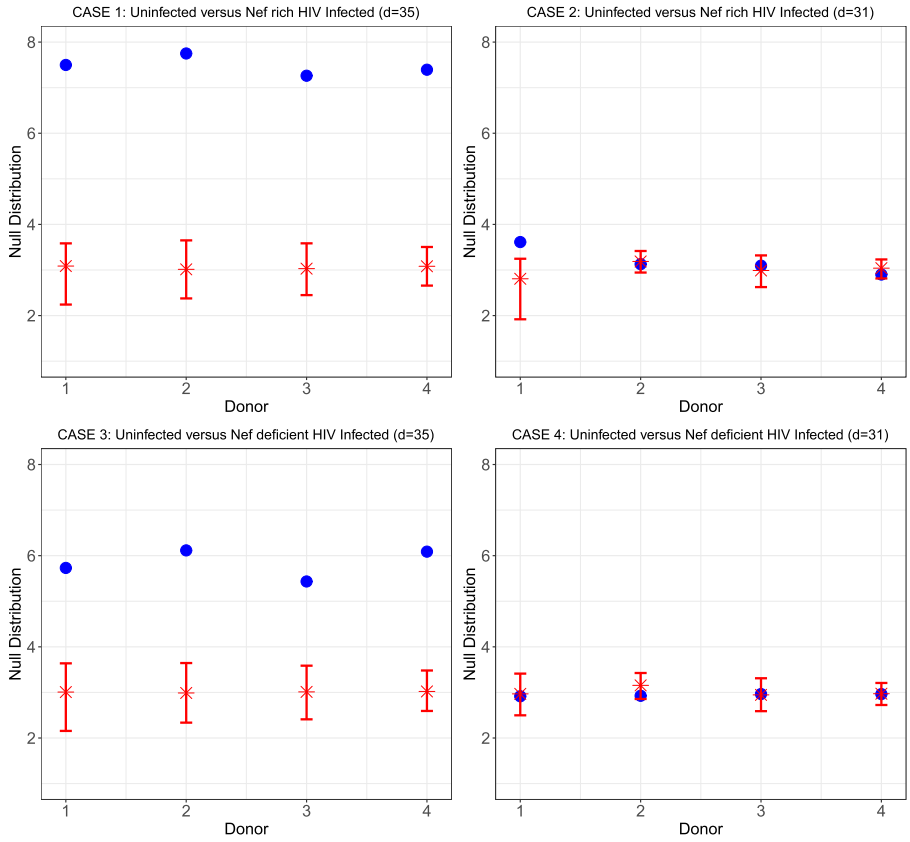

FIG. 9. *Null distribution of the* TRUH *statistic under cases* 1–4. *The blue dots are magnitudes of* TRUH *statistic for each donor under the four cases while the red bars indicate the* 2.5*th*, 50*th and* 97.5*th percentiles of the bootstrapped null distribution obtained from algorithm* 1 *with* $\tau_{fc} = 1.1$.

to be down regulated by Nef, were removed from our analysis ($d = 31$). Other than donor 1, TRUH indicates no remodeling in this scenario for the remaining three donors which is expected given the mechanism of remodeling that Nef pursues by down-regulating CD4, CCR5, CD28 and CD62L (Swigut, Shohdy and Skowronski (2001)). The absence of these four cell markers from the uninfected and infected samples reduces the phenotypic gap between these samples as measured through their surface markers. The top row in Figure 9 shows that, while the null distribution shifts down from CASE 1 (left plot) to CASE 2 (right

TABLE 8
p-values in CASE 1: *Uninfected vs.* Nef-rich *HIV Infected for entire* 35 *markers*

Tests	Donor 1 $m = 24{,}984, n = 245$	Donor 2 $m = 31{,}552, n = 521$	Donor 3 $m = 17{,}704, n = 211$	Donor 4 $m = 22{,}830, n = 660$
Energy	<0.001	<0.001	<0.001	<0.001
CrossMatch	0.005	0.005	0.005	0.005
E Count	<0.001	<0.001	<0.001	<0.001
GE Count	<0.001	<0.001	<0.001	<0.001
WE Count	<0.001	<0.001	<0.001	<0.001
MTE Count	<0.001	<0.001	<0.001	<0.001
TRUH	<0.001	<0.001	<0.001	<0.001

TABLE 9
p-values in CASE 2: *Uninfected vs.* `Nef-rich` *HIV Infected for* 31 *invariant markers*

Tests	Donor 1 $m = 24{,}984, n = 245$	Donor 2 $m = 31{,}552, n = 521$	Donor 3 $m = 17{,}704, n = 211$	Donor 4 $m = 22{,}830, n = 660$
Energy	<0.001	<0.001	<0.001	<0.001
CrossMatch	0.005	0.005	0.005	0.005
E Count	<0.001	<0.001	<0.001	<0.001
GE Count	<0.001	<0.001	<0.001	<0.001
WE Count	<0.001	<0.001	<0.001	<0.001
MTE Count	<0.001	<0.001	<0.001	<0.001
TRUH	<0.001	0.67	0.274	0.914

plot) across all four donors, the drop in the magnitude of the TRUH statistic is far more substantial when the four surface markers are excluded. The remaining six test statistics appear to be insensitive to these subtle changes in the uninfected and infected samples across the two scenarios and continue to detect remodeling in Case 2 which is actually no remodeling but preferential infection. This demonstrates their inability to handle heterogeneity in the data that TRUH tackles via the composite null testing framework of equations (2.1)–(2.3).

In Tables 10 and 11, we present the p-values of the seven test statistics for testing the null hypothesis H_0 of no remodeling when the HIV-infected sample lacks the critical Nef gene (see Construction and validation of reporter viruses in Supplemental Experimental Procedures of Cavrois et al. (2017) for details around the generation of Nef-deficient HIV-infected cells). We see that TRUH rejects the null hypothesis of no remodeling in CASE 3 (Table 10) while it fails to do so in CASE 4 (Table 11), thus corroborating the biological phenomena that: (a) Nef independent remodeling is prevalent in HIV-infected cells and, (b) even in the absence of Nef, the down regulation of the four surface markers by other mechanisms contributes to remodeling. The bottom row in Figure 9 presents the values of the TRUH statistic and the 2.5th, 50th, 97.5th percentiles of the associated null distribution. Similar observations from the top row continue to hold for Cases 3 and 4 in the bottom row of Figure 9, wherein the drop in the magnitude of TRUH statistic is far more significant when the four surface markers are excluded. Moreover, from Figure 9 we see that, for every donor the TRUH statistic obeys, a rank ordering across the scenarios which is of the form $\texttt{TRUH}_1 > \texttt{TRUH}_3 > \texttt{TRUH}_2 > \texttt{TRUH}_4$ where $\texttt{TRUH}_s$ is the magnitude of the TRUH statistic under cases $s = 1, \ldots, 4$. This is not accidental for the relative strength of remodeling is known to be highest under the influence of Nef-rich HIV infection and more so when

TABLE 10
p-values in CASE 3: *Uninfected vs.* `Nef-deficient` *HIV Infected for the entire* 35 *markers*

Tests	Donor 1 $m = 24{,}984, n = 129$	Donor 2 $m = 31{,}552, n = 382$	Donor 3 $m = 17{,}704, n = 174$	Donor 4 $m = 22{,}830, n = 440$
Energy	<0.001	<0.001	<0.001	<0.001
CrossMatch	0.005	0.005	0.005	0.005
E Count	<0.001	<0.001	<0.001	<0.001
GE Count	<0.001	<0.001	<0.001	<0.001
WE Count	<0.001	<0.001	<0.001	<0.001
MTE Count	<0.001	<0.001	<0.001	<0.001
TRUH	<0.001	<0.001	<0.001	<0.001

TABLE 11
p-values in CASE 4: *Uninfected vs.* `Nef-deficient` *HIV Infected for* 31 *invariant markers*

Tests	Donor 1 $m = 24{,}984, n = 129$	Donor 2 $m = 31{,}552, n = 382$	Donor 3 $m = 17{,}704, n = 174$	Donor 4 $m = 22{,}830, n = 440$
`Energy`	<0.001	<0.001	<0.001	<0.001
`CrossMatch`	0.005	0.005	0.005	0.005
`E Count`	<0.001	<0.001	<0.001	<0.001
`GE Count`	<0.001	<0.001	<0.001	<0.001
`WE Count`	<0.001	<0.001	<0.001	<0.001
`MTE Count`	<0.001	<0.001	<0.001	<0.001
`TRUH`	0.58	0.94	0.464	0.524

`Nef` down-regulates the four cell surface markers, `CD4, CCR5, CD28` and `CD62L`. As was seen in Cases 1 and 2, the remaining six tests continue to side in favor of remodeling in both Cases 3 and 4, thus reflecting their relative lack of conservatism in detecting remodeling under our composite null testing framework.

The remodeling analysis of the HIV-infected T Cells reveals that our proposed testing procedure, `TRUH`, conforms to the biologically validated phenomenon of remodeling of human tonsillar T cells under both `Nef-rich` (Case 1) and `Nef-deficient` (Case 3) HIV infection. However, unlike traditional tests that continue to infer remodeling in Cases 2 and 4, `TRUH` detects preferential infection and concludes that phenotypic differences between the HIV-infected and uninfected T cells are primarily driven by variations in the expression levels of `CD4, CCR5, CD28` and `CD62L` across the uninfected and infected cells. Moreover, through Cases 1 and 2, `TRUH` corroborates the findings in Chaudhuri et al. (2007), Michel et al. (2005), Swigut, Shohdy and Skowronski (2001), Vassena et al. (2015) that HIV remodeling of the T cells is driven by `Nef` dependent down-regulation of `CD4, CCR5, CD28, CD62L,` while through Cases 3 and 4 `TRUH` reveals `Nef` independent remodeling of T cells, as evidenced in Cavrois et al. (2017).

5. Optimality properties of the `TRUH` statistic. In this section we derive the L_2-limit of the proposed test statistic $T_{m,n}$ in the usual limiting regime where the sample sizes $m, n \to \infty$, such that $n/m \to \rho > 0$. This can be used to choose a cut-off and construct a test based on $T_{m,n}$ and show asymptotic consistency for biologically relevant location alternatives.

Recall that the uninfected and infected samples are denoted as

$$\boldsymbol{U}_m = \{U_1, \ldots, U_m\} \quad \text{and} \quad \boldsymbol{V}_n = \{V_1, \ldots, V_n\}, \tag{5.1}$$

which are i.i.d. samples from two unknown densities f_0 and g in $\mathbb{R}^d$, respectively. To derive the limit of $T_{m,n}$, we need certain integrability/moment assumptions on f_0 and g.

ASSUMPTION 1. The densities f_0 and g have a common support $S \subseteq \mathbb{R}^d$ and satisfy either one of the following two assumptions, depending on the dimension:

1. For $d \leq 2$, the support S is compact (with a nonempty interior) and f_0 and g are bounded away from zero on S.

2. For $d \geq 3$, f_0 and g satisfy the following conditions: $\int_S f_0(y)^{1-\frac{1}{d}}\,dy < \infty$, $\int_S f_0(y)^{-\frac{1}{d}} g(y)\,dy < \infty$, and $\int_S |y|^r f_0(y)\,dy < \infty$, $\int_S |y|^r g(y)\,dy < \infty$, for some $r > d/(d-2)$.

To describe the limit of $T_{m,n}$, we need a few definitions: For $\lambda > 0$, denote by $\mathcal{P}_\lambda$ the homogeneous Poisson process of intensity λ in $\mathbb{R}^d$, and $\mathcal{P}_\lambda^x = \mathcal{P}_\lambda \cup \{x\}$ for $x \in \mathbb{R}^d$. Now, define the following two quantities:

$$\zeta_1(\mathbf{0}, \mathcal{P}_1) = \inf_{b \in \mathcal{P}_1} \|b\| \quad \text{and} \quad \zeta_2(\mathbf{0}, \mathcal{P}_1) = \inf_{b \in \mathcal{P}_1 \setminus N(\mathbf{0}, \mathcal{P}_1)} \|N(\mathbf{0}, \mathcal{P}_1) - b\|, \tag{5.2}$$

that is, the distance from the origin $\mathbf{0}$ in $\mathbb{R}^d$ to its nearest neighbor in the Poisson process $\mathcal{P}_1$ and the distance of this point to its neighbor in $\mathcal{P}_1$, respectively.

THEOREM 1. *Let $T_{m,n}$ be as in (2.7). Then, for f_0 and g as in Assumption 1 above, as $m, n \to \infty$ such that $n/m \to \rho$,*

$$T_{m,n} \xrightarrow{L_2} \varphi(f_0, g, \rho) = \rho^{\frac{1}{d}} \Delta_d \int \frac{g(y)}{f_0(y)^{\frac{1}{d}}} \, \mathrm{d}y, \tag{5.3}$$

with $\Delta_d = (\zeta_2 - \zeta_1)$, where:

- *$\zeta_1 = \mathbb{E}\zeta_1(\mathbf{0}, \mathcal{P}_1)$, the expected distance from the origin in $\mathbf{0} \in \mathbb{R}^d$ to its nearest neighbor in $\mathcal{P}_1$, and*
- *$\zeta_2 = \mathbb{E}\zeta_2(\mathbf{0}, \mathcal{P}_1)$, the excepted distance between the nearest neighbor of the origin in $\mathcal{P}_1$ to its nearest neighbor in $\mathcal{P}_1$.*

The above theorem gives the L_2-limit of the test statistic for general distributions f_0 and g. The proof of the theorem, which is given in the Supplementary Material (Section B of Banerjee, Bhattacharya and Mukherjee (2020)), uses the machinery of geometric stabilization, introduced by Penrose and Yukich (2003), which obtains the asymptotics of nearest neighbor based functionals in terms of functionals defined on a homogeneous Poisson process. Before we discuss how the result in Theorem 1 can be used to construct a test based on $T_{m,n}$ for the hypothesis (2.2), we discuss some properties and the consequences of the limit in (5.3):

- Note that the finiteness of the limit in (5.3) is ensured by Assumption 1. For $d \geq 3$, the moment conditions in Assumption 1 are required to establish the L_2 convergence in (5.3). This assumption can be relaxed to $\int_S |y|^r f_0(y) \, \mathrm{d}y < \infty$ and $\int_S |y|^r g(y) \, \mathrm{d}y < \infty$, for some $r > d/(d-1)$, if we are only interested in L_1 convergence (by combining the proof of Theorem 1 with that of Penrose and Yukich (2003), Proposition 3.2). However, this still does not apply for $d = 1$, where it is necessary to assume the compactness of the support, in order to ensure that the limit in (5.3) is finite. This is a well-known constraint which arises in a large family of random geometric graphs, while dealing with the asymptotics of edge lengths (see, e.g., Penrose and Yukich (2003), Theorem 1.1, and the references therein). Even though the compactness assumption technically rules out some natural distributions, from a practical standpoint, there is no real concern because one can approximate the univariate density by truncating it to a large interval on which the above result applies. Incidentally, there has been recent work on relaxing the compactness and density bounded below assumptions in the related problems of nearest-neighbor classification Cannings, Berrett and Samworth (2020), Gadat, Klein and Marteau (2016) and entropy estimation Berrett, Samworth and Yuan (2019) which could provide useful insights on how to relax these assumptions from Theorem 1, and what are the effects of tail behavior on the heterogeneity testing problem.
- Note that ζ_1 and ζ_2 are both constants, which depend only on the dimension d. In fact, ζ_1 has a closed form expression which can be easily derived. To this end, denote by V_d and S_d the volume and the surface area of the unit ball in $\mathbb{R}^d$, respectively. It is easy to verify that

$S_d = dV_d$. Moreover, for $r > 0$ and $x \in \mathbb{R}^d$, denote by $B(x, r)$ the ball of radius r centered at $x \in \mathbb{R}^d$. Then, using the observation that a point b is the nearest neighbor of the origin, if are there no points of the Poisson process $\mathcal{P}_1$ in the ball $B(0, \|b\|)$, it follows that

$$\zeta_1 = \mathbb{E}(\zeta_1(\mathbf{0}, \mathcal{P}_1) = \int \|b\| \mathbb{P}\big(b = N\big(\mathbf{0}, \mathcal{P}_1^{0,b}\big)\big)\, \mathrm{d}b = S_d \int_0^\infty t^d e^{-V_d t^d}\, \mathrm{d}t,$$

which, by the change of variable $x = V_d t^d$, equals

$$\left(\frac{1}{V_d}\right)^{\frac{1}{d}} \int_0^\infty x^{\frac{1}{d}} e^{-x}\, \mathrm{d}x = \left(\frac{1}{V_d}\right)^{\frac{1}{d}} \Gamma\left(\frac{d+1}{d}\right), \tag{5.4}$$

where $\Gamma(\cdot)$ denotes the Gamma function.

Theorem 1 shows that, for K fixed densities $f_1, \ldots, f_K$, and $f_0 = \sum_{a=1}^K w_a f_a$,

$$\begin{aligned} \sup_{g \in \mathcal{F}(f_0)} \varphi(f_0, g, \rho) &= \rho^{\frac{1}{d}} \Delta_d \sup_{\lambda_1, \lambda_2, \ldots, \lambda_K} \sum_{a=1}^K \lambda_a \int \frac{f_a(y)}{\big(\sum_{b=1}^K w_b f_b(y)\big)^{\frac{1}{d}}}\, \mathrm{d}y \\ &= \rho^{\frac{1}{d}} \Delta_d \max_{1 \le a \le K} \left\{ \int \frac{\lambda_a f_a(y)}{\big(\sum_{b=1}^K w_b f_b(y)\big)^{\frac{1}{d}}}\, \mathrm{d}y \right\}, \end{aligned} \tag{5.5}$$

where the last step uses the fact that $\lambda_a \in [0, 1]$, for $1 \le a \le K$ and $\sum_{a=1}^K \lambda_K = 1$. Note that the RHS above is unknown, because the densities $f_1, \ldots, f_K$ and the weights $w_1, \ldots, w_K$ as well as the number K of mixture components, are all unknown. However, if we can consistently estimate the RHS of (5.5), then the test, which rejects H_0 in (2.3) when $T_{m,n}$ is greater than the estimated value of (5.5), would have zero asymptotic Type I error and would be powerful whenever g has some separation from the set $\mathcal{F}(f_0)$ (recall definition in (2.2)).

The approach described above is, in general, infeasible because nonparametric estimation of mixture parameters in multivariate problems, especially when the number K is unknown, can often be difficult. In the following we show how in location families, one can obtain a slightly weaker upper bound on $\varphi(f_0, g, \rho)$, which is free of the unknown parameters, that can be used to construct a valid and powerful test for the remodeling hypothesis (2.3). To this end, consider $\{p(y|\theta) = p(y - \theta) : \theta \in \Theta\}$ a family of densities indexed by the parameter space $\Theta \subseteq \mathbb{R}^d$, where $p : \mathbb{R}^d \to \mathbb{R}_{\ge 0}$ such that $\int_{\mathbb{R}^d} p(y)\, \mathrm{d}y = 1$. Throughout, we assume that the densities in the family satisfy Assumption 1. Suppose the baseline samples $U_1, U_2, \ldots, U_m$ are i.i.d. from the density $f_0(\cdot) = \sum_{a=1}^K w_a p(\cdot|\theta_a)$, where $\theta_1, \ldots, \theta_K \in \Theta$ are fixed (but unknown), and there exists a known constant $L > 0$ such that $w_a \ge L$, for all $1 \le a \le K$. If the infected samples $V_1, V_2, \ldots, V_n$ are i.i.d. from a density g in $\mathbb{R}^d$, then the hypothesis of remodeling (2.2), in this parametric setting, becomes,

$$H_0 : g \in \mathcal{F}(\boldsymbol{\theta}) \quad \text{versus} \quad H_A : g \notin \mathcal{F}(\boldsymbol{\theta}), \tag{5.6}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ and $\mathcal{F}(\boldsymbol{\theta})$ is defined as follows:

$$\mathcal{F}(\boldsymbol{\theta}) = \left\{ q(\cdot) = \sum_{a=1}^K \lambda_a p(\cdot|\theta_a) : \lambda_a \in [0, 1], \text{for } 1 \le a \le K, \text{ and } \sum_{a=1}^K \lambda_a = 1 \right\}$$

is the collection of K-mixtures of $p(\cdot|\theta_1), p(\cdot|\theta_2), \ldots, p(\cdot|\theta_K)$. Note that under the null H_0, $g(\cdot) = \sum_{a=1}^K \lambda_a p(\cdot|\theta_a)$, for some $\lambda_1, \lambda_2, \ldots, \lambda_K \in [0, 1]$, such that $\sum_{a=1}^K \lambda_a = 1$. Then, using $\sum_{a=1}^K w_a p(y|\theta_a) > w_b p(y|\theta_b) \ge L p(y|\theta_b)$, for all $b \in \{1, 2, \ldots, K\}$,

$$\begin{aligned} \varphi(f_0, g, \rho) &= \rho^{\frac{1}{d}} \Delta_d \sum_{a=1}^K \lambda_a \int \frac{p(y|\theta_a)}{\big(\sum_{b=1}^K w_b p(y|\theta_b)\big)^{\frac{1}{d}}}\, \mathrm{d}y \\ &< \frac{\rho^{\frac{1}{d}} \Delta_d}{L^{\frac{1}{d}}} \int p(z)^{1 - \frac{1}{d}}\, \mathrm{d}z = \gamma, \end{aligned} \tag{5.7}$$

where the last step follows by the change of variable $z = y - \theta_a$. Note that the constant γ depends on L (the lower bound on the mixing weights of the baseline population), the dimension d and the base function p, defining the location family (which is assumed to be known), but not on the unknown means $(\theta_1, \theta_2, \ldots, \theta_K)$, the unknown weights $(w_1, w_2, \ldots, w_K)$ or the number of components, and hence can be directly calculated. This implies that the test, which rejects when $T_{m,n} > \gamma$, would have zero asymptotic Type I error and would also be powerful whenever g has some separation from the set of possible null distributions $\mathcal{F}(f_0)$, as explained below.

The corollary below shows how the bound in (5.7) can be used to construct a test based on $T_{m,n}$, which is powerful for mixtures of radially symmetric distributions, such as Gaussian mixtures and t-mixtures, among others. Hereafter, we assume $p(y) = r(\|y\|)$ is radially symmetric, where $r : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a uniformly continuous function, such that $\int_{\mathbb{R}^d} r(\|y\|)\,\mathrm{d}y = 1$. (Recall, $\|y\|$ denotes the Euclidean norm of $y \in \mathbb{R}^d$.)

COROLLARY 1. *For the testing problem* (5.6) *in the family* $\{p(y|\theta) = r(\|y - \theta\|) : \theta \in \Theta\}$, *the following hold*:

- *For any* $g \in \mathcal{F}(\boldsymbol{\theta})$, *with* γ *as defined in* (5.7), *we have*

$$\lim_{m,n\to\infty} \mathbb{P}_{f_0,g}(T_{m,n} > \gamma) = 0. \tag{5.8}$$

- *There exists* $\varepsilon(\gamma) > 0$ *such that*

$$\lim_{m,n\to\infty} \mathbb{P}_{f_0,g}(T_{m,n} > \gamma) = 1, \tag{5.9}$$

for any $g(y) = \sum_{a=1}^{K} \bar{\lambda}_a p(y|\theta'_a)$ *with* $\min_{1\leq a,b\leq K} \|\theta'_a - \theta_b\| \mathbf{1}\{\bar{\lambda}_a > 0\} \geq \varepsilon(\gamma)$.

The proof of the corollary is given in the Supplementary Material (Section C of Banerjee, Bhattacharya and Mukherjee (2020)). Note that the condition on $g(y)$ in (5.9) quantifies a natural notion of separation between g and the set $\mathcal{F}(\boldsymbol{\theta})$ by assuming that at least one of the mixture means of g is ε-far (in L_2-distance) from all the unknown null means of the baseline density. Explicit bounds on the separation $\varepsilon(\gamma)$ can be obtained from the proof of Corollary 1, based on the tail decay of the base density p (details given in the Supplementary Material, Section C of Banerjee, Bhattacharya and Mukherjee (2020)).

6. Discussion. We propose a novel nearest-neighbor based two-sample test for detecting changes between the baseline and the case samples, in the presence of heterogeneity, as is often the case in single-cell virology. For integrative analysis involving datasets collected from different experiments with varying external conditions, batch-effect corrections are needed before applying our methodology. Our testing procedure is specially designed for mass cytometry based techniques (Bendall et al. (2011), Giesen et al. (2014)) which produces moderate dimensional ($d \sim 50$) cellular characteristics. In the future it will be interesting to extend our methodology for dealing with single-cell RNA-seq based techniques (Huang et al. (2018), Hwang, Lee and Bang (2018), Jaitin et al. (2014), Schiffman et al. (2017)) which can produce highly multivariate phenotypes ($d \sim 10^4$). A possible approach can be based on random projections of the d dimensional cellular characteristics to a lower dimensional space and then using our testing procedure on the reduced data. Also, it will be interesting to develop efficient testing procedures where the underlying population contains heterogeneous subpopulations with highly varying sizes, including some very rare subpopulations. Finally, extending our hypothesis testing framework to distinguish between depletion and enrichment in remodeled cells will be important.

Acknowledgments. We are grateful to Ann Arvin, Nadia Roan, Adrish Sen, Nandini Sen and Nancy Zhang for numerous stimulating discussions. We thank the Editor, the Associate Editor and three anonymous referees for constructive suggestions that greatly improved the paper.

The third author was partially supported by NSF Grant DMS-1811866.

SUPPLEMENTARY MATERIAL

Supplement: A nearest-neighbor based nonparametric test for viral remodeling in heterogeneous single-cell proteomic data (DOI: 10.1214/20-AOAS1362SUPPA; .pdf). This supplement provides the proofs of the theoretical results and additional numerical experiments.

Supplement: Source code for "A nearest-neighbor based nonparametric test for viral remodeling in heterogeneous single-cell proteomic data" (DOI: 10.1214/20-AOAS1362SUPPB; .zip). This supplement holds the R source code that reproduces the results in Section 3.

REFERENCES

AMIR, E.-A. D., DAVIS, K. L., TADMOR, M. D., SIMONDS, E. F., LEVINE, J. H., BENDALL, S. C., SHENFELD, D. K., KRISHNASWAMY, S., NOLAN, G. P. et al. (2013). Visne enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31** 545–552.

ASLAN, B. and ZECH, G. (2005). New test for the multivariate two-sample problem based on the concept of minimum energy. *J. Stat. Comput. Simul.* **75** 109–119. MR2117010 https://doi.org/10.1080/00949650410001661440

BANERJEE, T., BHATTACHARYA, B. B. and MUKHERJEE, G. (2020). Supplement to "A nearest-neighbor based nonparametric test for viral remodeling in heterogeneous single-cell proteomic data." https://doi.org/10.1214/20-AOAS1362SUPPA, https://doi.org/10.1214/20-AOAS1362SUPPB

BARINGHAUS, L. and FRANZ, C. (2004). On a new multivariate two-sample test. *J. Multivariate Anal.* **88** 190–206. MR2021870 https://doi.org/10.1016/S0047-259X(03)00079-4

BASMACIOGULLARI, S. and PIZZATO, M. (2014). The activity of nef on hiv-1 infectivity. *Front. Microbiol.* **5** 232.

BENDALL, S. C., SIMONDS, E. F., QIU, P., AMIR, E.-A. D., KRUTZIK, P. O., FINCK, R., BRUGGNER, R. V., MELAMED, R., TREJO, A. et al. (2011). Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332** 687–696.

BENDALL, S. C., DAVIS, K. L., AMIR, E.-A. D., TADMOR, M. D., SIMONDS, E. F., CHEN, T. J., SHENFELD, D. K., NOLAN, G. P. and PE'ER, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell* **157** 714–725.

BERRETT, T. B. and SAMWORTH, R. J. (2019a). Efficient two-sample functional estimation and the super-oracle phenomenon. arXiv preprint. Available at arXiv:1904.09347.

BERRETT, T. B. and SAMWORTH, R. J. (2019b). Nonparametric independence testing via mutual information. *Biometrika* **106** 547–566. MR3992389 https://doi.org/10.1093/biomet/asz024

BERRETT, T. B., SAMWORTH, R. J. and YUAN, M. (2019). Efficient multivariate entropy estimation via k-nearest neighbour distances. *Ann. Statist.* **47** 288–318. MR3909934 https://doi.org/10.1214/18-AOS1688

BHATTACHARYA, B. B. (2019). A general asymptotic framework for distribution-free graph-based two-sample tests. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 575–602. MR3961499

BICKEL, P. J. (1968). A distribution free version of the Smirnov two sample test in the p-variate case. *Ann. Math. Stat.* **40** 1–23. MR0256519 https://doi.org/10.1214/aoms/1177697800

BRUGGNER, R. V., BODENMILLER, B., DILL, D. L., TIBSHIRANI, R. J. and NOLAN, G. P. (2014). Automated identification of stratifying signatures in cellular subpopulations. *Proc. Natl. Acad. Sci. USA* **111** E2770–E2777.

CANNINGS, T. I., BERRETT, T. B. and SAMWORTH, R. J. (2020). Local nearest neighbour classification with applications to semi-supervised learning. *Ann. Statist.* **48** 1789–1814. MR4124344 https://doi.org/10.1214/19-AOS1868

CAVROIS, M., BANERJEE, T., MUKHERJEE, G., RAMAN, N., HUSSIEN, R., RODRIGUEZ, B. A., VASQUEZ, J., SPITZER, M. H., LAZARUS, N. H. et al. (2017). Mass cytometric analysis of hiv entry, replication, and remodeling in tissue cd4+ t cells. *Cell Rep.* **20** 984–998.

CHAUDHURI, R., LINDWASSER, O. W., SMITH, W. J., HURLEY, J. H. and BONIFACINO, J. S. (2007). Down-regulation of cd4 by human immunodeficiency virus type 1 nef is dependent on clathrin and involves direct interaction of nef with the ap2 clathrin adaptor. *J. Virol*. **81** 3877–3890.

CHEN, H., CHEN, X. and SU, Y. (2018). A weighted edge-count two-sample test for multivariate and object data. *J. Amer. Statist. Assoc*. **113** 1146–1155. MR3862346 https://doi.org/10.1080/01621459.2017.1307757

CHEN, L., DOU, W. W. and QIAO, Z. (2013). Ensemble subsampling for imbalanced multivariate two-sample tests. *J. Amer. Statist. Assoc*. **108** 1308–1323. MR3174710 https://doi.org/10.1080/01621459.2013.800763

CHEN, H. and FRIEDMAN, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *J. Amer. Statist. Assoc*. **112** 397–409. MR3646580 https://doi.org/10.1080/01621459.2016.1147356

CHUNG, J. H. and FRASER, D. A. (1958). Randomization tests for a multivariate two-sample problem. *J. Amer. Statist. Assoc*. **53** 729–735.

COVER, T. and HART, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13** 21–27.

CRESSIE, N. (1976). On the logarithms of high-order spacings. *Biometrika* **63** 343–355. MR0428583 https://doi.org/10.1093/biomet/63.2.343

DEB, N. and SEN, B. (2019). Multivariate rank-based distribution-free nonparametric testing using measure transportation. arXiv preprint. Available at arXiv:1909.08733.

DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition*. *Applications of Mathematics* (*New York*) **31**. Springer, New York. MR1383093 https://doi.org/10.1007/978-1-4612-0711-5

DVORKIN, D. (2012). lcmix: Layered and chained mixture models. R package version 0.3/r5.

FRIEDMAN, J. H. and RAFSKY, L. C. (1979). Multivariate generalizations of the Wald–Wolfowitz and Smirnov two-sample tests. *Ann. Statist*. **7** 697–717. MR0532236

GADAT, S., KLEIN, T. and MARTEAU, C. (2016). Classification in general finite dimensional spaces with the k-nearest neighbor rule. *Ann. Statist*. **44** 982–1009. MR3485951 https://doi.org/10.1214/15-AOS1395

GARCIA, J. V. and MILLER, A. D. (1991). Serine phosphorylation-independent downregulation of cell-surface cd4 by nef. *Nature* **350** 508.

GHOSAL, P. and SEN, B. (2019). Multivariate ranks and quantiles using optimal transportation and applications to goodness-of-fit testing. arXiv preprint. Available at arXiv:1905.05340.

GIESEN, C., WANG, H. A., SCHAPIRO, D., ZIVANOVIC, N., JACOBS, A., HATTENDORF, B., SCHÜFFLER, P. J., GROLIMUND, D., BUHMANN, J. M. et al. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods* **11** 417.

GORIA, M. N., LEONENKO, N. N., MERGEL, V. V. and NOVI INVERARDI, P. L. (2005). A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *J. Nonparametr. Stat*. **17** 277–297. MR2129834 https://doi.org/10.1080/104852504200026815

GRETTON, A., BORGWARDT, K. M., RASCH, M., SCHÖLKOPF, B. and SMOLA, A. J. (2007). A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems* 513–520.

HALL, P. and TAJVIDI, N. (2002). Permutation tests for equality of distributions in high-dimensional settings. *Biometrika* **89** 359–374. MR1913964 https://doi.org/10.1093/biomet/89.2.359

HECKEL, R. and BÖLCSKEI, H. (2015). Robust subspace clustering via thresholding. *IEEE Trans. Inf. Theory* **61** 6320–6342. MR3418967 https://doi.org/10.1109/TIT.2015.2472520

HENZE, N. (1984). Über die Anzahl von Zufallspunkten mit typ-gleichem nächsten Nachbarn und einen multivariaten Zwei–Stichproben-Test. *Metrika* **31** 259–273. MR0773815 https://doi.org/10.1007/BF01915210

HENZE, N. and PENROSE, M. D. (1999). On the multivariate runs test. *Ann. Statist*. **27** 290–298. MR1701112 https://doi.org/10.1214/aos/1018031112

HOLMES, S. and HUBER, W. (2018). *Modern Statistics for Modern Biology*. Cambridge Univ. Press, Cambridge.

HUANG, M., WANG, J., TORRE, E., DUECK, H., SHAFFER, S., BONASIO, R., MURRAY, J. I., RAJ, A., LI, M. et al. (2018). Saver: Gene expression recovery for single-cell rna sequencing. *Nat. Methods* **15** 539.

HWANG, B., LEE, J. H. and BANG, D. (2018). Single-cell rna sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med*. **50** 1–14.

JAITIN, D. A., KENIGSBERG, E., KEREN-SHAUL, H., ELEFANT, N., PAUL, F., ZARETSKY, I., MILDNER, A., COHEN, N., JUNG, S. et al. (2014). Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science* **343** 776–779.

JIA, C., HU, Y., KELLY, D., KIM, J., LI, M. and ZHANG, N. R. (2017). Accounting for technical noise in differential expression analysis of single-cell rna sequencing data. *Nucleic Acids Res*. **45** 10978–10988.

JIANG, H., SOHN, L. L., HUANG, H. and CHEN, L. (2018). Single cell clustering based on cell-pair differentiability correlation and variance analysis. *Bioinformatics* **34** 3684–3694.

KOZACHENKO, L. F. and LEONENKO, N. N. (1987). A statistical estimate for the entropy of a random vector. *Problemy Peredachi Informatsii* **23** 9–16. MR0908626

LINDERMAN, M. D., BJORNSON, Z., SIMONDS, E. F., QIU, P., BRUGGNER, R. V., SHEODE, K., MENG, T. H., PLEVRITIS, S. K. and NOLAN, G. P. (2012). Cytospade: High-performance analysis and visualization of high-dimensional cytometry data. *Bioinformatics* **28** 2400–2401.

MACK, Y. P. (1983). Rate of strong uniform convergence of k-NN density estimates. *J. Statist. Plann. Inference* **8** 185–192. MR0720150 https://doi.org/10.1016/0378-3758(83)90037-X

MACK, Y. P. and ROSENBLATT, M. (1979). Multivariate k-nearest neighbor density estimates. *J. Multivariate Anal.* **9** 1–15. MR0530638 https://doi.org/10.1016/0047-259X(79)90065-4

MATHESON, N. J., SUMNER, J., WALS, K., RAPITEANU, R., WEEKES, M. P., VIGAN, R., WEINELT, J., SCHINDLER, M., ANTROBUS, R. et al. (2015). Cell surface proteomic map of hiv infection reveals antagonism of amino acid metabolism by vpu and nef. *Cell Host Microbe* **18** 409–423.

MICHEL, N., ALLESPACH, I., VENZKE, S., FACKLER, O. T. and KEPPLER, O. T. (2005). The nef protein of human immunodeficiency virus establishes superinfection immunity by a dual strategy to downregulate cell-surface ccr5 and cd4. *Curr. Biol.* **15** 714–723.

PENROSE, M. D. and YUKICH, J. E. (2003). Weak laws of large numbers in geometric probability. *Ann. Appl. Probab.* **13** 277–303. MR1952000 https://doi.org/10.1214/aoap/1042765669

QIU, P. (2012). Inferring phenotypic properties from single-cell characteristics. *PLoS ONE* **7** e37038.

ROSENBAUM, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 515–530. MR2168202 https://doi.org/10.1111/j.1467-9868.2005.00513.x

ROSS, T. M., ORAN, A. E. and CULLEN, B. R. (1999). Inhibition of hiv-1 progeny virion release by cell-surface cd4 is relieved by expression of the viral nef protein. *Curr. Biol.* **9** 613–621.

SAMWORTH, R. J. (2012). Optimal weighted nearest neighbour classifiers. *Ann. Statist.* **40** 2733–2763. MR3097618 https://doi.org/10.1214/12-AOS1049

SCHIFFMAN, C., LIN, C., SHI, F., CHEN, L., SOHN, L. and HUANG, H. (2017). Sideseq: A cell similarity measure defined by shared identified differentially expressed genes for single-cell rna sequencing data. *Stat. Biosci.* **9** 200–216.

SCHILLING, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.* **81** 799–806. MR0860514

SEN, N., MUKHERJEE, G. and ARVIN, A. M. (2015). Single cell mass cytometry reveals remodeling of human t cell phenotypes by varicella zoster virus. *Methods* **90** 85–94.

SEN, A., ROTHENBERG, M. E., MUKHERJEE, G., FENG, N., KALISKY, T., NAIR, N., JOHNSTONE, I. M., CLARKE, M. F. and GREENBERG, H. B. (2012). Innate immune response to homologous rotavirus infection in the small intestinal villous epithelium at single-cell resolution. *Proc. Natl. Acad. Sci. USA* **109** 20667–20672.

SEN, N., MUKHERJEE, G., SEN, A., BENDALL, S. C., SUNG, P., NOLAN, G. P. and ARVIN, A. M. (2014). Single-cell mass cytometry analysis of human tonsil t cell remodeling by varicella zoster virus. *Cell Rep.* **8** 633–645.

SHI, F. and HUANG, H. (2017). Identifying cell subpopulations and their genetic drivers from single-cell RNA-Seq data using a biclustering approach. *J. Comput. Biol.* **24** 663–674. MR3671106 https://doi.org/10.1089/cmb.2017.0049

SONG, P. X.-K. (2000). Multivariate dispersion models generated from Gaussian copula. *Scand. J. Stat.* **27** 305–320. MR1777506 https://doi.org/10.1111/1467-9469.00191

SWIGUT, T., SHOHDY, N. and SKOWRONSKI, J. (2001). Mechanism for down-regulation of cd28 by nef. *EMBO J.* **20** 1593–1604.

TIBSHIRANI, R. and WALTHER, G. (2005). Cluster validation by prediction strength. *J. Comput. Graph. Statist.* **14** 511–528. MR2170199 https://doi.org/10.1198/106186005X59243

VAN DER MAATEN, L. and HINTON, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9** 2579–2605.

VASICEK, O. (1976). A test for normality based on sample entropy. *J. Roy. Statist. Soc. Ser. B* **38** 54–59. MR0420958

VASSENA, L., GIULIANI, E., KOPPENSTEINER, H., BOLDUAN, S., SCHINDLER, M. and DORIA, M. (2015). Hiv-1 nef and vpu interfere with l-selectin (cd62l) cell surface expression to inhibit adhesion and signaling in infected cd4+ t lymphocytes. *J. Virol.* **JVI–00611**.

WANG, J., HUANG, M., TORRE, E., DUECK, H., SHAFFER, S., MURRAY, J., RAJ, A., LI, M. and ZHANG, N. R. (2018). Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc. Natl. Acad. Sci. USA* **115** E6437–E6446. MR3831463 https://doi.org/10.1073/pnas.1721085115

WEISS, L. (1960). Two-sample tests for multivariate distributions. *Ann. Math. Stat.* **31** 159–164. MR0119305 https://doi.org/10.1214/aoms/1177705995

ZHANG, J. and CHEN, H. (2017). Graph-based two-sample tests for discrete data. arXiv preprint. Available at arXiv:1711.04349.