Supplementary Information to "A Simple and Flexible Test of Sample Exchangeability with Applications to Statistical Genomics"

Alan J. Aw[1,2,3], Jeffrey P. Spence[4], Yun S. Song[1,2,3]
[1] Department of Statistics, University of California, Berkeley
[2] Center for Computational Biology, University of California, Berkeley
[3] Computer Science Division, University of California, Berkeley
[4] Department of Genetics, Stanford University

We provide details supplementing parts of the Main Text and present proofs of results reported in the Main Text.

A Connection to Hypergeometric Distributions and Exponential Families

Our work is related to the theory of sampling from hypergeometric distributions that arise from exponential families conditioned on a sufficient statistic (Diaconis and Sturmfels, 1998). In this theory, we think of sampling as being performed on *counts* of each unique length P binary vector. This perspective will allow us to precisely characterize $H_{\boldsymbol{c}}$, the resampling distribution of $\mathbf{X}$ introduced very recently above. Let $\{0,1\}^P$ denote the collection all length P binary vectors. For $\mathbf{x} \in \{0,1\}^P$, let $\kappa(\mathbf{x}) = \#\{\mathbf{x}_i = \mathbf{x}\}$ be the number of observations $\mathbf{x}_i$ of $\mathbf{X}$ equal to $\mathbf{x}$. Observe that because the ordering of the observations that make up the resampled dataset $\mathbf{X}^*$ does not matter for our computation of $V^* = V(\mathbf{X}^*)$, therefore sampling from the permutation resampling distribution on $\mathbf{X}$ (i.e., uniformly sampling from $\mathcal{Y}_{\boldsymbol{c}}$) is equivalent to sampling from the following collection of *count statistic vectors* $\kappa^* : \{0,1\}^P \to \mathbb{Z}_0^+$,

$$\mathcal{F}_{\boldsymbol{c}} = \left\{ \kappa : \sum_{\mathbf{x}\in\{0,1\}^P} \kappa(\mathbf{x}) \cdot \mathbf{x} = \boldsymbol{c}, \sum_{\mathbf{x}\in\{0,1\}^P} \kappa(\mathbf{x}) = N \right\}.$$

Hence, we may characterize $H_{\boldsymbol{c}}$ as a distribution on count statistic vectors, where each resampled array $\mathbf{X}^*$ has a unique corresponding count statistic vector κ^* (but not the other way round).

To make the connection with exponential families, suppose N length P binary vectors, $\mathbf{x}_1, \ldots, \mathbf{x}_N$, are sampled *with replacement* from $\{0,1\}^P$, where the distribution over $\{0,1\}^P$ is parametrized by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P) \in (0,1)^P$. For a binary vector $\mathbf{x} = (x_1, \ldots, x_P)^T$, let the probability of picking $\mathbf{x}$ be $\pi_{\mathbf{x}} = \prod_{j=1}^P (\theta_j x_j + (1-\theta_j)(1-x_j))$. Under such a sampling scheme, the corresponding count statistic distribution is a multinomial distribution,

$$G_{\boldsymbol{\theta}}(\kappa) = \binom{N}{(\kappa(\mathbf{x}) : \mathbf{x} \in \{0,1\}^P)} \prod_{\mathbf{x}\in\{0,1\}^P} \pi_{\mathbf{x}}^{\kappa(\mathbf{x})}.$$

The product term above is some combination $\prod_{j=1}^P \theta_j^{\omega_j}(1-\theta_j)^{\zeta_j}$. After stacking $\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T$ row-wise to obtain a $N \times P$ array $\mathbf{X}$, notice the indices ω_j and ζ_j are simply

$$\begin{aligned} \omega_j &= \#\{\mathbf{x}_i \text{ contains 1 in column } j\}, \\ \zeta_j &= \#\{\mathbf{x}_i \text{ contains 0 in column } j\} = N - \omega_j. \end{aligned}$$

Let the random vector $\boldsymbol{C} = \boldsymbol{C}(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \sum_{i=1}^N \mathbf{x}_i = \sum_{\mathbf{x}\in\{0,1\}^P} \kappa(\mathbf{x}) \cdot \mathbf{x}$ summarize the number of 1's appearing in each column. By relating $\boldsymbol{C}$ to the quantities ω_j and ζ_j, we may write

$$G_{\boldsymbol{\theta}}(\kappa) = \binom{N}{(\kappa(\mathbf{x}) : \mathbf{x} \in \{0,1\}^P)} \exp\left[\boldsymbol{C}^T \mathrm{logit}(\boldsymbol{\theta}) + N \sum_{j=1}^P \log(1-\theta_j)\right], \tag{S1}$$

where $\mathrm{logit}(\boldsymbol{v})$ denotes the vector whose jth component is the one-to-one logit function $t \mapsto \log[t/(1-t)]$ applied to the jth component of $\boldsymbol{v}$. This is a P-parameter exponential family with sufficient statistic $\boldsymbol{C}$. More importantly, conditioned on $\boldsymbol{C} = \boldsymbol{c}$, where $\boldsymbol{c}$ are the column sums of the observed dataset, we obtain the resampling distribution of $\mathbf{X}$. Thus, $H_{\boldsymbol{c}}$ is the exponential family distribution $G_{\boldsymbol{\theta}}(\kappa)$ conditioned on observing sufficient statistic $\boldsymbol{C}$. It moreover has a closed form probability mass function, given by the hypergeometric density

$$H_{\boldsymbol{c}}(\kappa) = \frac{N!}{|\mathcal{Y}_{\boldsymbol{c}}|} \left(\prod_{\mathbf{x}\in\{0,1\}^P} \kappa(\mathbf{x})!\right)^{-1}. \tag{S2}$$

While the mass function of $H_{\boldsymbol{c}}$ is exact, the mass function and distribution F_{perm} of V^* are not. To see why, we simplify the expressions in eqs. (4) and (5). This simplification is summarized in Proposition S1 below.

Proposition S1 (Simplification Equations)**.** *Recall that for $\mathbf{x} \in \{0,1\}^P$, we let $\kappa(\mathbf{x}) = \#\{\mathbf{x}_i = \mathbf{x}\}$ be the number of observations $\mathbf{x}_i$ of $\mathbf{X}$ equal to $\mathbf{x}$. This produces a vector κ computed from $\mathbf{X}$, the count statistic vector. Let $\boldsymbol{c}$ denote the column sums of $\mathbf{X}$, κ denote its associated count statistic vector, and $\Delta_H = (d_H^2(\mathbf{x},\mathbf{y}))_{\mathbf{x},\mathbf{y}\in\{0,1\}^P}$ denote the $2^P \times 2^P$ matrix of squared Hamming distances between each pair of length P binary vectors. Then, the expressions in* (4) *and* (5) *are equal to the following two expressions.*

$$V(\mathbf{X}) = \frac{1}{P} \cdot \frac{2N}{N-1} \left\langle \frac{\kappa}{N}, \Delta_H \frac{\kappa}{N} \right\rangle - \mu^2, \tag{S3}$$

$$\mu = \frac{2N}{N-1} \sum_{p=1}^P \frac{c_p}{N}\left(1 - \frac{c_p}{N}\right). \tag{S4}$$

Proof of Proposition S1. First, we verify (S4). Observe

$$\begin{aligned}
\mu &= \frac{1}{\binom{N}{2}} \sum_{i<j} \sum_{p=1}^P \mathbb{1}(x_{ip} \neq x_{jp}) \\
&= \frac{1}{\binom{N}{2}} \sum_{p=1}^P \sum_{i<j} \mathbb{1}(x_{ip} \neq x_{jp}) \\
&= \frac{1}{\binom{N}{2}} \sum_{p=1}^P c_p(N - c_p),
\end{aligned}$$

where the last equality follows from counting the number of pairs $\{i,j\}$ in the pth column for which $\mathbb{1}(x_{ip} \neq x_{jp}) = 1$.

Next, we verify (S3). Observe

$$\begin{aligned} V(\mathbf{X}) &= \frac{1}{\binom{N}{2}} \sum_{i<j} d_H^2(\mathbf{x}_i, \mathbf{x}_j) - \left[\frac{2N}{N-1} \sum_{p=1}^{P} \frac{c_p}{N} \left(1 - \frac{c_p}{N}\right) \right]^2 \\ &= \frac{1}{\binom{N}{2}} \sum_{\{\mathbf{x},\mathbf{y}\}:\mathbf{x},\mathbf{y}\in\{0,1\}^P} d_H^2(\mathbf{x}, \mathbf{y}) \kappa(\mathbf{x}) \kappa(\mathbf{y}) - \left[\frac{2N}{N-1} \sum_{p=1}^{P} \frac{c_p}{N} \left(1 - \frac{c_p}{N}\right) \right]^2 \\ &= \frac{2N}{N-1} \left\langle \frac{\kappa}{N}, \Delta_H \frac{\kappa}{N} \right\rangle - \left[\frac{2N}{N-1} \sum_{p=1}^{P} \frac{c_p}{N} \left(1 - \frac{c_p}{N}\right) \right]^2, \end{aligned}$$

where the last equality follows from the definition of Δ_H. □

From Proposition S1, we see that μ depends only on $\boldsymbol{c}$ and is universally bounded in N for any fixed P. Thus, V is dominated by a quadratic form in κ, and so F_{perm} is the image of a hypergeometric distribution under a quadratic map. Because (S3) does not admit a straightforward inverse map, direct computation of F_{perm} is not possible. We thus rely on Algorithm 1 in practice to estimate p-values.

B Permutation Invariance and Valid Testing

Let $\pi : \mathcal{Y}_{\boldsymbol{c}} \to \mathcal{Y}_{\boldsymbol{c}}$ be a transformation obtained from permuting observations within each column feature, across the N observations; see Subsection 2.1 of the Main Text. The set of all such transformations, $\mathscr{G} = \mathfrak{S}_N \times \cdots \times \mathfrak{S}_N$, is the P-fold direct product of the symmetric group of order $N!$. Under this notation, we see that any resampled array $\mathbf{X}^* \in \mathcal{Y}_{\boldsymbol{c}}$ can be written as $\pi(\mathbf{X})$ for some (possibly more than one choice of) $\pi \in \mathscr{G}$. We write $\pi(\mathbf{X})$ as $\pi\mathbf{X}$ to enhance readability. We also recall that $V : \mathcal{Y}_{\boldsymbol{c}} \to \mathbb{R}$ is our test statistic.

The exchangeability null hypothesis can be expressed as the following null hypothesis of permutation invariance:

$$H_p : (\forall \pi \in \mathscr{G})(\mathbf{X} \overset{d}{=} \pi\mathbf{X}). \tag{S1}$$

Let $\mathscr{G} = \{g_1, \ldots, g_{\#\mathscr{G}}\}$. A consequence of H_p is the following permutation invariance of the joint distribution of V:

$$(V(g_1\mathbf{X}), \ldots, V(g_{\#\mathscr{G}}\mathbf{X})) \overset{d}{=} (V(g_1\pi\mathbf{X}), \ldots, V(g_{\#\mathscr{G}}\pi\mathbf{X})),$$

for all $\pi \in \mathscr{G}$.

The permutation test we describe in the Main Text rejects (S1) when $V(\mathbf{X}) > V^{(k)}(\mathbf{X})$, where

$$V^{(1)}(\mathbf{X}) \leqslant \ldots \leqslant V^{(\#\mathscr{G})}(\mathbf{X})$$

are the sorted test statistics, and $k = \lceil (1-\alpha)\#\mathscr{G} \rceil$ with $\alpha \in [0, 1)$. This test is exact, in that for any choice of α and $k = k(\alpha)$, under H_p, $\mathbb{P}(V(\mathbf{X}) > V^{(k)}(\mathbf{X})) \leqslant \alpha$. Our *naïve implementation* of the test, however, resamples a large number R of permutations $\pi \in \mathscr{G}$, *with replacement*, and then *estimates* the quantity $\mathbb{P}(V(\mathbf{X}) > V^{(k)}(\mathbf{X}))$ by the fraction of resampled arrays whose test statistic value exceeds V_{obs}. This procedure produces an unbiased estimate of the true probability, but suffers from an inflated Type I Error for very stringent choices of α. Indeed, for a given choice of R, we see that the resampling algorithm cannot produce outputs lying between 0 and $1/R$, and moreover, there is a non-zero probability $q \neq \alpha$ of the output being 0. Holding R fixed, it is apparent that for any choice of $0 < \alpha < 1/R$, the algorithm will either reject the null (with probability equal to q) or not reject the null (with probability equal to $1 - q$). This leads to an anti-conservative test especially for $\alpha \ll q$.

To ensure the Type I Error is controlled, we *include the identity permutation, id, and evaluate the fraction of resampled arrays whose test statistic value is as extreme as or exceeds* V_{obs}. Intuitively, this works by making sure that the estimated p-value is always positive. Below, we show that in fact, under H_p this simple fix leads to a test that provably controls the Type I Error rate. That is, we show that our procedure meets the following criterion: for any user-specified α and for any R, when uniformly drawing R permutations from $\mathscr{G}$ with replacement, the probability of rejecting H_p, when H_p is true, is at most α.

Theorem S2 (Valid Testing)**.** *Let* $\mathscr{G}'$ *be the set* $\{id, \pi_1, \ldots, \pi_R\}$*, where id is the identity permutation and* $\pi_1, \ldots, \pi_R$ *are elements drawn uniformly at random, with replacement, from* $\mathscr{G}$*. Write* $\pi_0 = id$*, and let* $V^{(0)}(\mathbf{X}, \mathscr{G}') \leqslant \ldots \leqslant V^{(R)}(\mathbf{X}, \mathscr{G}')$ *be the ordered test statistics* $V(\pi_i \mathbf{X})$*,* $0 \leqslant i \leqslant R$*. Let* $\alpha \in [0, 1)$ *and* $k = \lceil (1-\alpha)(R+1) \rceil$*. Reject* H_p *when* $V(\mathbf{X}) > V^{(k)}(\mathbf{X}, \mathscr{G})$*. Then, under* H_p *the rejection probability is at most* α*.*

Proof of Theorem S2. Our proof mirrors that of Theorem 2 of Hemerik and Goeman (2018), although we take care to mention how various group-theoretic and probabilistic assumptions are used throughout. Since $\mathscr{G}'$ contains id, by the group structure of $\mathscr{G}$ it holds that for $0 \leqslant j \leqslant R$ the distributions of $\mathscr{G}'\pi_i^{-1} := \{\pi_j \pi_i^{-1} : j = 0, \ldots, R\}$ and $\mathscr{G}'$ are identical. Let i be uniformly distributed on $\{0, 1, \ldots, R\}$ and write $\tau = \pi_i$. Then

$$\begin{aligned} \mathbb{P}(V(\mathbf{X}) > V^{(k)}(\mathbf{X}, \mathscr{G}')) &= \mathbb{P}(V(\mathbf{X}) > V^{(k)}(\mathbf{X}, \mathscr{G}'\tau^{-1})) \\ &= \mathbb{P}(V(\tau\mathbf{X}) > V^{(k)}(\tau\mathbf{X}, \mathscr{G}'\tau^{-1})), \end{aligned}$$

where the first equality follows from $\mathscr{G}'\tau^{-1} \overset{d}{=} \mathscr{G}'$ and the second equality follows from H_p and the uniform randomness of τ. Since $(\mathscr{G}'\tau^{-1})(\tau\mathbf{X}) = \mathscr{G}'(\tau^{-1}\tau\mathbf{X})$, the last expression equals $\mathbb{P}(V(\tau\mathbf{X}) > V^{(k)}(\tau^{-1}\tau\mathbf{X}, \mathscr{G}') = \mathbb{P}(V(\tau\mathbf{X}) > V^{(k)}(\mathbf{X}, \mathscr{G}'))$. We have thus shown that

$$\mathbb{P}(V(\mathbf{X}) > V^{(k)}(\mathbf{X}, \mathscr{G}')) = \mathbb{P}(V(\tau\mathbf{X}) > V^{(k)}(\mathbf{X}, \mathscr{G}')), \tag{S2}$$

for $\tau = \pi_i$ picked uniformly at random.

Finally, observe that for any $\mathscr{G}'$ chosen,

$$\sum_{i=0}^{R} \mathbb{1}(V(\pi_i\mathbf{X}) > V^{(k)}(\mathbf{X}, \mathscr{G}')) \leqslant \alpha(R+1).$$

Taking expectations and using eq. (S2), we obtain

$$\begin{aligned} (R+1) \cdot \mathbb{P}(V(\mathbf{X}) > V^{(k)}(\mathbf{X}, \mathscr{G}')) &= \mathbb{E}\left(\sum_{i=0}^{R} \mathbb{1}(V(\pi_i\mathbf{X}) > V^{(k)}(\mathbf{X}, \mathscr{G}')) \right) \\ &\leqslant \alpha(R+1), \end{aligned}$$

which completes the proof. □

C Details of Asymptotic Results

C.1 Large P Asymptotics

In Theorem 2.2 of the Main Text, we report that the permutation-induced random variable $V^{(N,P)*}$ has an asymptotic distribution that is a convolution of two chi-square random variables. Moreover, the convolution weights, a_1^N and a_2^N, reportedly depend on the column sums of the dataset $\mathbf{X}$. Here, we explicitly relate a_1^N and a_2^N to the column sums $\boldsymbol{c} = (c_1, \ldots, c_P)$ of $\mathbf{X}$.

Define the quantities

$$\begin{aligned}
\alpha^{N,P} &= \frac{1}{P}\sum_{p=1}^{P}\frac{c_p(N-c_p)}{\binom{N}{2}}\left[1-\frac{c_p(N-c_p)}{\binom{N}{2}}\right],\\
\beta^{N,P} &= \frac{1}{P}\sum_{p=1}^{P}\frac{c_p(N-c_p)}{\binom{N}{2}}\left[\frac{1}{2}-\frac{c_p(N-c_p)}{\binom{N}{2}}\right],\\
\gamma^{N,P} &= \frac{1}{P}\sum_{p=1}^{P}\frac{c_p(N-c_p)}{\binom{N}{2}}\left[\frac{(c_p-1)(N-c_p-1)}{\binom{N-2}{2}}-\frac{c_p(N-c_p)}{\binom{N}{2}}\right],
\end{aligned}$$

and let

$$\begin{aligned}
a_1^{N,P} &= \alpha^{N,P}+(N-4)\beta^{N,P}-(N-3)\gamma^{N,P},\\
a_2^{N,P} &= \alpha^{N,P}-2\beta^{N,P}+\gamma^{N,P}.
\end{aligned}$$

Then, the quantities a_1^N and a_2^N are defined by $\lim_{P\to\infty} a_i^{N,P} = a_i^N$ for $i = 1, 2$. In our software implementation, we compute these quantities for samples $\mathbf{X}$ that have reasonably large number of independent features, P, and use them as the convolution weights reported in Theorem 2.2 of the Main Text.

C.2 Large N and Large P Asymptotics

In Subsection 2.2 of our Main Text we report that the large N and large P asymptotic distribution of $V^{(N,P)*}$ is Gaussian. Here, we provide formal details on how to approximate the null distribution of $V^{(N,P)*}$ when both N and P are large. The theorem below says that $V^{(N,P)*}$ is roughly normally distributed, with mean and variance determined by the column sums of the dataset.

Theorem S3 (Large-P, large-N Limit)**.** *With the random variable $V^{(N,P)*}$ and the quantities $\alpha^{N,P}, \beta^{N,P}, \gamma^{N,P}$, $a_1^{N,P}, a_2^{N,P}$ as defined in Supplementary Material C.1, let $\lim_{N,P\to\infty}\alpha^{N,P} = \alpha$ and define*

$$\begin{aligned}
\tau_N &= \lim_{P\to\infty}\frac{2(N-1)(a_1^{N,P})^2+2\left[\binom{N-1}{2}-1\right](a_2^{N,P})^2}{\binom{N}{2}^2}\\
&= \frac{2(N-1)(a_1^{N})^2+2\left[\binom{N-1}{2}-1\right](a_2^{N})^2}{\binom{N}{2}^2}.
\end{aligned}$$

Then $\tau_N^{-1/2}\left(V^{(N,P)*}-\alpha\right)\xrightarrow{d}\mathcal{N}(0,1)$ as $P\to\infty$ and $N\to\infty$.

Consequently, for N and P large, $V^{(N,P)*}$ is approximately distributed as $\mathcal{N}(\alpha, \tau_N)$.

C.3 Large B and Large P Asymptotics

In Subsection 4.1 of our Main Text we report that in the case of partitionable dependent features, the test statistic $V^{(N,B,P)*}$ has a large B and large P asymptotic distribution under the exchangeable null. Here, we provide formal details on how to approximate this null distribution.

Theorem S4 (Large-P and Large-B Approximation of Block Permutation Null)**.** *Let $V^{(N,B,P)*}$ be the random variable with the block permutation null distribution of $V^{(N,P)}$, where the blocks have delimiters $1 \leqslant P_1 < \cdots < P_B = P$. For each block $b = 1, \ldots, B$, let $d^{(b)}(\mathbf{x}, \mathbf{x}')$ denote the partial Hamming distance of binary vectors $\mathbf{x}$ and $\mathbf{x}'$, that is, the Hamming distance computed along that*

block only. (Note that under this definition the Hamming distance $d_H(\mathbf{x}, \mathbf{x}') = \sum_{b=1}^{B} d^{(b)}(\mathbf{x}, \mathbf{x}')$*.) Define the quantities*

$$\begin{aligned}
\alpha^{N,B,P} &= \frac{1}{P}\sum_{b=1}^{B}\left(\overline{d^{(b)}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})^2} - \left[\overline{d^{(b)}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})}\right]^2\right), \\
\beta^{N,B,P} &= \frac{1}{P}\sum_{b=1}^{B}\left(\overline{d^{(b)}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) d^{(b)}(\mathbf{x}_{i_1}, \mathbf{x}_{i_3})} - \left[\overline{d^{(b)}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})}\right]^2\right), \\
\gamma^{N,B,P} &= \frac{1}{P}\sum_{b=1}^{B}\left(\overline{d^{(b)}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) d^{(b)}(\mathbf{x}_{i_3}, \mathbf{x}_{i_4})} - \left[\overline{d^{(b)}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})}\right]^2\right),
\end{aligned}$$

where the indices i_1, i_2, i_3 *and* i_4 *are distinct and the overline notation denotes taking the average over all pairs, triples, or quadruples of observations. Further, define the quantities*

$$\begin{aligned}
b_1^{N,B,P} &= \alpha^{N,B,P} + (N-4)\beta^{N,B,P} - (N-3)\gamma^{N,B,P}, \\
b_2^{N,B,P} &= \alpha^{N,B,P} - 2\beta^{N,B,P} + \gamma^{N,B,P}.
\end{aligned}$$

Letting $\lim_{B,P\to\infty} b_i^{N,B,P} = b_i^N$ *for* $i = 1, 2$ *and assuming the limits exist, define the random variable*

$$V^{(N,\infty,\infty)} = \frac{b_1^N \chi^2_{N-1} + b_2^N \chi^2_{\binom{N-1}{2}-1}}{\binom{N}{2}}.$$

Then, $V^{(N,B,P)*} \xrightarrow{d} V^{(N,\infty,\infty)}$ *as* $B, P \to \infty$*. In other words, for* B *and* P *large,* $V^{(N,B,P)*} \overset{d}{=} V^{(N,\infty,\infty)}$ *approximately.*

D Null Models for Evaluating Large B and Large P Approximation

In Subsection 4.1 of our Main Text, we report an asymptotic result and mention that we conducted a simulation study to evaluate its accuracy in practice (as measured by FPR control). Here, we describe the models used for simulation.

We simulate data under two different generative models that produce samples with partitionable features: (1) concatenation of a binarized autoregressive time series; and (2) concatenation of a coalescent model commonly used as a generative model for population-genetic datasets. Briefly, for Model (1) we simulate an AR(1) process (i.e., $x_t = \rho x_{t-1} + \varepsilon_t$) with parameter $\rho = 0.5$ before applying a "normal quantile"-based binarization, whereas for Model (2) we simulate haplotypes under the standard coalescent with recombination. In both cases, sequences are simulated in blocks, with concatenation of the blocks to form the final observed unit (simulation code is provided under `block_simulation` directory of Supplementary Material zip file). Whereas Model (1) produces blocks of the same size, Model (2) does not. In both cases we fix $B = 50$ and consider varying sample sizes $N \in \{10, 50, 100, 500, 1000\}$, and perform Monte Carlo estimation of the FPR at significance threshold $\alpha = 0.05$ by simulating $200,000$ replications for each model and running the approximate test.

The results are plotted in Figure S18, and we summarize them in our Main Text.

E A General Scalable Exchangeability Test Requiring Only Pairwise Distance Data

Suppose that we are given a $N \times P$ dataset $\mathbf{X}$ containing P partitionable features, with the features grouped into B blocks. Each feature can be real- or complex-valued, or even be objects lying in a metric space ($B = P$ corresponds to the scenario where the P univariate features are independent, and so each block consists of a single feature.) For each pair of observations $\mathbf{x}, \mathbf{x}'$, let $\{d^{(b)}(\mathbf{x}, \mathbf{x}') : b = 1, \ldots, B\}$ be the collection of B distances where each distance is computed on one

of the blocks of features. In case the B blocks of features come from B underlying metric spaces $\{(\Omega_b, d^{(b)}) : b = 1, \dots B\}$, then all that is needed are the distances computed on the observed data objects. Note $d^{(b)}(\cdot,\cdot)$ and $d^{(b')}(\cdot,\cdot)$ need not be the same distance function for distinct blocks b and b'. In practice, these distance functions are chosen based on the user application, especially when different groups of features come from distinct data modalities. For concreteness we list two examples with two blocks ($b = 1, b' = 2$).

- We could have $d^{(b)}(\mathbf{x}, \mathbf{x}') = \frac{1}{2}\|\mathbf{x}_{1:P_1} - \mathbf{x}'_{1:P_1}\|_1$ and $d^{(b')}(\mathbf{x}, \mathbf{x}') = \max(\mathbf{x}_{(P_1+1):P_2} - \mathbf{x}'_{(P_1+1):P_2})$, where $\mathbf{x}_{k:\ell}$ denotes the subvector of $\mathbf{x}$ obtained by keeping components k up to ℓ of the original.
- Suppose each sample $\mathbf{x} = (\omega_b, \omega_{b'})$ lies in the product of metric spaces $(\Omega_b, d^{(b)}) \otimes (\Omega_{b'}, d^{(b')})$. Here the space $(\Omega_b, d^{(b)})$ could be a space of phylogenetic trees equipped with some tree metric (e.g., Billera-Holmes-Vogtmann treespace with the BHV metric (Billera et al., 2001)), while $(\Omega_{b'}, d^{(b')})$ could be a space of compactly supported probability distributions equipped with the Wasserstein metric.

Instead of V defined by (4), we now let $d_g(\mathbf{x}, \mathbf{x}') = \sum_{b=1}^{B} d^{(b)}(\mathbf{x}, \mathbf{x}')$, and define

$$V_g(\mathbf{X}) = \frac{1}{P\binom{N}{2}} \sum_{i<j} [d_g(\mathbf{x}_i, \mathbf{x}_j) - \mu_g]^2, \tag{S1}$$

with

$$\mu_g = \frac{1}{\binom{N}{2}} \sum_{i<j} d_g(\mathbf{x}_i, \mathbf{x}_j). \tag{S2}$$

With these quantities, the general test is then permuting the blocks independently across the observations and computing the proportion of resampled V_g values larger than or equal to the observed value. This procedure is formalized as Algorithm 2 in Supplementary Material F.

Similar to Theorem S4 in Supplementary Material C.3, a "large B and large P" chi-square approximation to the block permutation null distribution can be obtained for the statistic V_g^*.

F Algorithms

Algorithm 1 Computation of p-value from data array (block version)

Input: Individual-by-feature array $\underset{N\times P}{\mathbf{X}}$, resampling number R, block delimiters $P_1, \dots, P_B$, type of p-value approximation (*unbiased* or *valid*)
Record $\boldsymbol{c} = \boldsymbol{c}(\mathbf{X})$, μ and $V_{\text{obs}} = V(\mathbf{X})$
Set $r = 0$, $\mathcal{V}^* = \emptyset$
while $r < R$ **do**
    Generate resampled array $\mathbf{X}^*$ from block permutation null
    Compute $V^* = V(\mathbf{X}^*)$
    $\mathcal{V}^* \leftarrow \mathcal{V}^* \cup \{V^*\}$
    $r \leftarrow r + 1$
end while
if type is *unbiased* **then**
    Output: $p = \frac{1}{R} \cdot \#[V^* > V_{\text{obs}}]$
else
    Output: $p = \frac{1}{R+1} \cdot (\#[V^* \geqslant V_{\text{obs}}] + 1)$
end if

Algorithm 2 Computation of p-value from data array (general version)

Input: Individual-by-feature array $\underset{N\times P}{\mathbf{X}}$, resampling number R, type of p-value approximation (*unbiased* or *valid*)
Record μ_g and $V_{\text{obs}} = V_g(\mathbf{X})$ (see main text above)
Set $r = 0$, $\mathcal{V}^* = \emptyset$
while $r < R$ **do**
  Generate resampled array $\mathbf{X}^*$ from block permutation null
  Compute $V_g^* = V_g(\mathbf{X}^*)$
  $\mathcal{V}^* \leftarrow \mathcal{V}^* \cup \{V_g^*\}$
  $r \leftarrow r + 1$
end while
if type is *unbiased* **then**
  Output: $p = \frac{1}{R} \cdot \#[V^* > V_{\text{obs}}]$
else
  Output: $p = \frac{1}{R+1} \cdot (\#[V^* \geqslant V_{\text{obs}}] + 1)$
end if

G Proofs of Main Results and Propositions

Throughout, we append a vector or a variable with an asterisk (e.g., x becomes x^*) to denote their random version induced by permuting the entries of each column of the original dataset $\mathbf{X}$. When articulating a mathematical statement requiring no reasoning about randomness (e.g., an equality between two algebraic expressions), we typically drop the asterisk to make the distinction.

Proof of Theorem 2.2. We describe our proof strategy before dotting the "i"s and crossing the "t"s.
Steps Outlining Proof

(1) Write $V^{(N,P)*}$ as the squared ℓ_2 norm of a random vector, $\vec{M}^*$, with $\vec{M}^*$ itself being a mean of independent zero-mean random vectors.
(2) Apply the Central Limit Theorem to $\vec{M}^*$. Together with Step 1, this implies $V^{(N,P)^*}$ is approximately weighted chi-square distributed as $P \to \infty$.
(3) Because the covariance matrix of $\vec{M}^*$ is non-diagonal and singular (conditioning on the sufficient statistic decrements the degrees of freedom by one), apply an orthogonal transformation to $\vec{M}^*$ to quantify the weights of the chi-square distribution.

Step 1: For $1 \leqslant i < j \leqslant N$, let $M_{ij} = \frac{1}{P}(d_H(\mathbf{x}_i, \mathbf{x}_j) - \mu)$ be the centered normalized Hamming distance between observations $\mathbf{x}_i$ and $\mathbf{x}_j$, where μ is the average Hamming distance defined in (5). By permuting the entries of each column of the original dataset $\mathbf{X}$, we obtain random variables $\mathbf{x}_i^*, \mathbf{x}_j^*$ and M_{ij}^*. Note that μ is permutation-invariant, being a deterministic function of the column sum vector $\boldsymbol{c} = (c_1, \ldots, c_P)$ (as verified in Proposition S1). Define

$$\vec{M}^* := \left(M_{ij}^* : \{i, j\} \in \binom{[N]}{2} \right), \tag{S1}$$

a length $\binom{N}{2}$ random vector whose entries are the random variables M_{ij}^*. Then $V^{(N,P)*} = \frac{P}{\binom{N}{2}} \|\vec{M}^*\|_2^2$, verifying the squared ℓ_2 norm assertion.

Next, we verify that $\vec{M}^*$ in (S1) is the mean of P independent zero-mean random vectors. For each feature $p \in [P]$ let $\mu_p = [c_p(N - c_p)]/\binom{N}{2}$, so that $\mu = \mu_1 + \ldots + \mu_P$ (as verified in the proof

of Proposition S1). Define

$$\mathbb{R}^{\binom{N}{2}} \ni \boldsymbol{v}_p = \begin{pmatrix} (x_{1p} - x_{2p})^2 \\ (x_{1p} - x_{3p})^2 \\ \vdots \\ (x_{1p} - x_{Np})^2 \\ \vdots \\ (x_{(N-1)p} - x_{Np})^2 \end{pmatrix} - \mu_p \cdot \vec{1},$$

which is the centered vector whose entries are the distances between pairs of observations. Notice that for any pair of distinct features p and p', the random vectors $\boldsymbol{v}_p^*$ and $\boldsymbol{v}_{p'}^*$ are independent. Moreover, for each feature p and pair of observations i and j, the random variable $(x_{ip}^* - x_{jp}^*)^2$ is marginally distributed as Bernoulli with success probability μ_p, so that $\mathbb{E}[\boldsymbol{v}_p^*] = \mathbf{0}$. Finally, the Hamming distance satisfying $d_H(\mathbf{x}_i, \mathbf{x}_j) = \sum_{p=1}^P (x_{ip} - x_{jp})^2$, it follows that $\sum_{p=1}^P \boldsymbol{v}_p = P\vec{M}$. This shows $\vec{M}^*$ is the mean of P independent zero-mean random vectors.

Step 2: Let $\Sigma_p = \mathbb{E}[\boldsymbol{v}_p^*(\boldsymbol{v}_p^*)^T]$ be the covariance matrix of $\boldsymbol{v}_p^*$, and define $\Sigma = \frac{1}{P}(\Sigma_1 + \ldots + \Sigma_P)$. Note that Σ has dimension $\binom{N}{2} \times \binom{N}{2}$. We claim that

$$\sqrt{P} \cdot \vec{M}^* \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma). \tag{S2}$$

By the Cramér-Wold theorem, as long as we can verify that any non-zero linear combination of $\vec{M}^*$ converges to the corresponding univariate normal distribution, then the convergence in (S2) is true. Thus, let $\vec{t} \in \mathbb{R}^{\binom{N}{2}} \setminus \{\mathbf{0}\}$. We must show that

$$\langle \vec{t}, \vec{M} \rangle \xrightarrow{d} N\left(\mathbf{0}, \frac{1}{P}\langle \vec{t}, \Sigma \vec{t} \rangle\right). \tag{S3}$$

The rest of this Step will be to verify (S3).

First, we compute Σ. Recall that $\Sigma = \frac{1}{P}(\Sigma_1 + \ldots + \Sigma_P)$. To compute a covariance matrix Σ_p, we must compute covariances $\mathrm{Cov}\left((x_{ip}^* - x_{jp}^*)^2, (x_{kp}^* - x_{\ell p}^*)^2\right)$ for pairs of 2-subsets $\{i, j\}$ and $\{k, \ell\}$. Fortunately, these covariances can be computed by splitting into three cases, with each case requiring a combinatorial argument to arrive at the covariance quantity.

- (Case I: $\{i, j\} = \{k, \ell\}$) Then,

$$\mathrm{Cov}\left((x_{ip}^* - x_{jp}^*)^2, (x_{kp}^* - x_{\ell p}^*)^2\right) = \mathrm{Var}((x_{ip}^* - x_{jp}^*)^2) = \frac{c_p(N - c_p)}{\binom{N}{2}}\left[1 - \frac{c_p(N - c_p)}{\binom{N}{2}}\right].$$

- (Case II: $|\{i, j\} \cap \{k, \ell\}| = 1$) Then,

$$\mathrm{Cov}\left((x_{ip}^* - x_{jp}^*)^2, (x_{kp}^* - x_{\ell p}^*)^2\right) = \frac{c_p(N - c_p)}{N(N-1)} - \left[\frac{c_p(N - c_p)}{\binom{N}{2}}\right]^2.$$

- (Case III: $|\{i, j\} \cap \{k, \ell\}| = 0$) Then,

$$\mathrm{Cov}\left((x_{ip}^* - x_{jp}^*)^2, (x_{kp}^* - x_{\ell p}^*)^2\right) = \frac{4c_p(N - c_p)(c_p - 1)(N - c_p - 1)}{N(N-1)(N-2)(N-3)} - \left[\frac{c_p(N - c_p)}{\binom{N}{2}}\right]^2.$$

Thus, (abusing notation $ij = \{i, j\}$ and) letting ij and $k\ell$ each run over all 2-subsets of $[N]$,

$$[\Sigma_p]_{ij,k\ell} = \begin{cases} \frac{c_p(N-c_p)}{\binom{N}{2}}\left[1 - \frac{c_p(N-c_p)}{\binom{N}{2}}\right] & \text{if } ij = k\ell \\ \frac{c_p(N-c_p)}{N(N-1)} - \left[\frac{c_p(N-c_p)}{\binom{N}{2}}\right]^2 & \text{if } |ij \cap k\ell| = 1 \\ \frac{4c_p(N-c_p)(c_p-1)(N-c_p-1)}{N(N-1)(N-2)(N-3)} - \left[\frac{c_p(N-c_p)}{\binom{N}{2}}\right]^2 & \text{if } |ij \cap k\ell| = 0 \end{cases}$$

By summing the matrices Σ_p, the calculations above imply that the entries of Σ satisfy

$$[\Sigma]_{ij,k\ell} = \begin{cases} \frac{1}{P}\sum_{p=1}^{P}\left(\frac{c_p(N-c_p)}{\binom{N}{2}}\left[1 - \frac{c_p(N-c_p)}{\binom{N}{2}}\right]\right) = \alpha^{N,P} & \text{if } ij = k\ell \\ \frac{1}{P}\sum_{p=1}^{P}\left(\frac{c_p(N-c_p)}{N(N-1)} - \left[\frac{c_p(N-c_p)}{\binom{N}{2}}\right]^2\right) = \beta^{N,P} & \text{if } |ij \cap k\ell| = 1 \\ \frac{1}{P}\sum_{p=1}^{P}\left(\frac{4c_p(N-c_p)(c_p-1)(N-c_p-1)}{N(N-1)(N-2)(N-3)} - \left[\frac{c_p(N-c_p)}{\binom{N}{2}}\right]^2\right) = \gamma^{N,P} & \text{if } |ij \cap k\ell| = 0 \end{cases}$$

Now that Σ is computed, we verify (S3) by checking that the Lyapunov condition holds. Here, we require two facts that will be proved in Step 3.

(A) The matrix Σ has eigenvalues

$$\begin{cases} 0, & \text{with multiplicity } 1 \\ \alpha^{N,P} + (N-4)\beta^{N,P} - (N-3)\gamma^{N,P}, & \text{with multiplicity } N-1 \\ \alpha^{N,P} - 2\beta^{N,P} + \gamma^{N,P}, & \text{with multiplicity } \binom{N-1}{2} - 1 \end{cases}$$

(B) The eigenspaces associated with each eigenvalue are as follows.
- For eigenvalue 0, $\mathcal{S}_1 = \text{span}(\{\vec{1}\})$.
- For eigenvalue $\alpha + (N-4)\beta - (N-3)\gamma$, $\mathcal{S}_2 = \text{span}\left(\{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^N\}\right)$, where $\boldsymbol{x}^n$ are defined in Theorem 2.2.
- For eigenvalue $\alpha - 2\beta + \gamma$, $\mathcal{S}_3 = \text{span}\left(\{\boldsymbol{w}^{12}, \ldots, \boldsymbol{w}^{N-1,N}\}\right)$, where $\boldsymbol{w}^{ij}$ are defined in Theorem 2.2.

When $\vec{t} = \lambda\vec{1}$ in (S3), observe that $\langle \vec{M}^*, \vec{1}\rangle = 0$, a consequence of each component of $\vec{M}^*$ having mean zero. Together with Fact (B) ensuring that $\langle \vec{t}, \Sigma\vec{t}\rangle = 0$ whenever $\vec{t} \in \mathcal{S}_1$, (S3) holds trivially. Thus, we assume for the rest of our argument that $\vec{t} \notin \mathcal{S}_1$.

For $\vec{t}$ picked, let its projection onto $\mathcal{S}_1$ be $\vec{t}_\circ$ and its projection onto the orthocomplement (i.e., the space $\mathcal{S}_2 \oplus \mathcal{S}_3$) be $\vec{t}_\perp$. Note that $\vec{t} = \vec{t}_\circ + \vec{t}_\perp$. From Step 1, we have the identity

$$\langle \vec{t}, \vec{M}\rangle = \frac{1}{P}\left(\langle \vec{t}, \boldsymbol{v}_1\rangle + \ldots + \langle \vec{t}, \boldsymbol{v}_P\rangle\right).$$

To show the Lyapunov condition holds, we must verify that there exists $\delta > 0$ such that

$$\lim_{P\to\infty} \frac{1}{s_P^{2+\delta}} \sum_{p=1}^{P} \mathbb{E}\left[\left|\langle \vec{t}, \boldsymbol{v}_p^*\rangle - \mathbb{E}[\langle \vec{t}, \boldsymbol{v}_{p'}^*\rangle]\right|^{2+\delta}\right] = 0, \tag{S4}$$

where

$$s_P := \sqrt{\sum_{p=1}^{P} \text{Var}\left(\langle \vec{t}, \boldsymbol{v}_p\rangle\right)}.$$

As it turns out, essentially the boundedness of the sum of entries of $\boldsymbol{v}_p^*$ ($= c_p$, a constant in P) guarantees that (S4) holds for *any* $\delta > 0$.[1] We check this carefully below.

From earlier computations we know that $s_P = \sqrt{P \cdot \langle \vec{t}, \Sigma \vec{t} \rangle}$. For brevity, introduce shorthand notation for the two non-zero eigenvalues:

$$\begin{aligned} \lambda_1 &= \alpha^{N,P} + (N-4)\beta^{N,P} - (N-3)\gamma^{N,P} \\ \lambda_2 &= \alpha^{N,P} - 2\beta^{N,P} + \gamma^{N,P}. \end{aligned}$$

Further algebra and use of Fact (B) imply that $s_P = \sqrt{P \cdot (\lambda_1 \|\vec{t}_{\perp,2}\|_2^2 + \lambda_2 \|\vec{t}_{\perp,3}\|_2^2)}$, where $\vec{t}_\perp = \vec{t}_{\perp,2} + \vec{t}_{\perp,3}$ is a further projection of $\vec{t}_\perp$ on the remaining two subspaces $\mathcal{S}_2$ and $\mathcal{S}_3$. Note at least one of the vectors $\vec{t}_{\perp,2}$ and $\vec{t}_{\perp,3}$ is non-zero, by assumption.

Now, let's bound each summand $\mathbb{E}\left[|\langle \vec{t}, \boldsymbol{v}_p^* \rangle - \mathbb{E}[\langle \vec{t}, \boldsymbol{v}_p^* \rangle]|^{2+\delta}\right]$ of the numerator of (S4). First, observe that the random vector $\boldsymbol{v}_p^*$ always contains $c_p(N-c_p)$ ones and $\binom{N}{2} - c_p(N-c_p)$ zeros. Denoting by the random variable W_p^* the linear combination, that is $W_p^* := \langle \vec{t}, \boldsymbol{v}_p^* \rangle = \sum_{\{i,j\} \in \binom{[N]}{2}} t_{ij} (\boldsymbol{v}_p^{ij})^*$, then W_p^* is supported on the interval with endpoints

$$\begin{aligned} \tau_{\min}^p &:= \min \left\{ \sum_{\{i,j\} \in \mathcal{A}} t_{ij} : \mathcal{A} \text{ ranges over all subsets of } \binom{[N]}{2} \text{ of size } c_p(N-c_p) \right\} \\ \tau_{\max}^p &:= \max \left\{ \sum_{\{i,j\} \in \mathcal{A}} t_{ij} : \mathcal{A} \text{ ranges over all subsets of } \binom{[N]}{2} \text{ of size } c_p(N-c_p) \right\} \end{aligned}$$

and moreover $\mathbb{E}[W_p^*] = 0$ since $\mathbb{E}[\boldsymbol{v}_p^*] = \mathbf{0}$.

Set $\hat{\tau}_p = |\tau_{\max}^p| \vee |\tau_{\min}^p|$. Then observe that $\hat{\tau}_p \leqslant \|\vec{t}\|_1 \leqslant \sqrt{\binom{N}{2}} \|\vec{t}\|_2$. Moreover, $\frac{W_p^*}{\hat{\tau}_p}$ is bounded between -1 and 1. Recalling the definition of W_p^* and rearranging terms, we obtain the upper bound

$$\mathbb{E}\left[|\langle \vec{t}, \boldsymbol{v}_p^* \rangle - \mathbb{E}[\langle \vec{t}, \boldsymbol{v}_p^* \rangle]|^{2+\delta}\right] \leqslant (\hat{\tau}_p)^\delta \mathrm{Var}(\langle \vec{t}, \boldsymbol{v}_p^* \rangle).$$

Now define $\hat{\tau} = \max_{p \in [P]} \hat{\tau}_p$. Observe that $\hat{\tau} \leqslant \|\vec{t}\|_1 \leqslant \sqrt{\binom{N}{2}} \|\vec{t}\|_2$, which is finite. Moreover, upon replacing $\hat{\tau}_p$ with $\hat{\tau}$ in the upper bound above and summing over p, we obtain the following upper bound on the numerator of (S4):

$$\frac{1}{s_P^{2+\delta}} \sum_{p=1}^P \mathbb{E}\left[|\langle \vec{t}, \boldsymbol{v}_p^* \rangle - \mathbb{E}[\langle \vec{t}, \boldsymbol{v}_p^* \rangle]|^{2+\delta}\right] \leqslant \left(\frac{\hat{\tau}}{s_P}\right)^\delta.$$

To finish the verification of Lyapunov's condition, it suffices to show that the ratio $s_P/\hat{\tau}$ increases in P. We compute this using earlier working:

$$\begin{aligned} \frac{s_P}{\hat{\tau}} &\geqslant \sqrt{\frac{P \cdot \left(\lambda_1 \|\vec{t}_{\perp,2}\|_2^2 + \lambda_2 \|\vec{t}_{\perp,3}\|_2^2\right)}{\binom{N}{2} \|\vec{t}\|_2^2}} \\ &= \sqrt{P} \cdot \frac{1}{\sqrt{\binom{N}{2}}} \cdot \sqrt{\frac{\lambda_1 \|\vec{t}_{\perp,2}\|_2^2 + \lambda_2 \|\vec{t}_{\perp,3}\|_2^2}{\|\vec{t}_\circ\|_2^2 + \|\vec{t}_{\perp,2}\|_2^2 + \|\vec{t}_{\perp,3}\|_2^2}}. \end{aligned}$$

[1] This boundedness also implies that the weaker Lindeberg condition holds, which is also sufficient for verifying (S3).

From the final expression on RHS above it is clear that $s_P/\hat{\tau}$ grows at least like $\sqrt{P}$, which completes the verification of (S4).

Step 3: We have just shown that (S2) holds. If Σ were diagonal, then the continuous mapping theorem applied to

$$f : \mathbb{R}^{\binom{N}{2}} \to \mathbb{R}, f(\mathbf{z}) = \frac{1}{\binom{N}{2}} \|\mathbf{z}\|_2^2 \tag{S5}$$

would imply that $V^{(N,P)*} = f(\sqrt{P}\vec{M}^*)$ is approximately a weighted sum of chi-square random variables with one degree of freedom, where the weights are the diagonal entries of Σ. Because Σ is not actually diagonal, we have to diagonalize Σ by performing an extra eigendecomposition step.

To this end, we shall state and prove a general lemma concerning "combinatorial matrices," by which we mean matrices whose dimensions are indexed by subsets and whose entries are determined by intersection properties of these subsets.[2]

Lemma S5 (Eigendecomposition of combinatorial matrices)**.** *Let Σ be a $\binom{N}{2} \times \binom{N}{2}$ matrix, whose dimensions are indexed by the 2-subsets $\{i, j\}$ of $[N]$. Let the entries of Σ be exactly one of three quantities — a, b and c — with the positions of a, b and c determined by the intersection of the 2-subset indices, as described by the equation below:*

$$[\Sigma]_{ij,k\ell} = \begin{cases} a & \text{if } ij = k\ell \\ b & \text{if } |ij \cap k\ell| = 1 \\ c & \text{if } |ij \cap k\ell| = 0 \end{cases}$$

Then, Σ has the the following eigenvalues,

$$a + (2N-4)b + \binom{N-2}{2}c$$

$$a + (N-4)b - (N-3)c$$

$$a - 2b + c$$

with multiplicities 1, $(N-1)$ and $\binom{N-1}{2} - 1$ respectively. The eigenspaces and eigenvectors can be summarized as follows:

- *For eigenvalue $a + (2N-4)b + \binom{N-2}{2}c$, the eigenspace is span$(\{\mathbf{1}\})$, which has dimension 1.*
- *For eigenvalue $a + (N-4)b - (N-3)c$, the eigenspace is span$(\{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^N\})$, where*

$$\boldsymbol{x}_{ij}^n = \begin{cases} 1 & \text{if } n \in \{i, j\} \\ \frac{-2}{N-2} & \text{if } n \notin \{i, j\} \end{cases},$$

  and moreover $\boldsymbol{x}^1 + \ldots + \boldsymbol{x}^N = \mathbf{0}$.
- *For eigenvalue $a - 2b + c$, the eigenspace is span $\left(\left\{\boldsymbol{w}^{ij} : ij \in \binom{[N]}{2}\right\}\right)$, where*

$$\boldsymbol{w}_{k\ell}^{ij} = \begin{cases} 1 & \text{if } \{i, j\} = \{k, \ell\} \\ \frac{-1}{N-2} & \text{if } |\{i, j\} \cap \{k, \ell\}| = 1 \\ \frac{1}{\binom{N-2}{2}} & \text{if } |\{i, j\} \cap \{k, \ell\}| = 0 \end{cases},$$

  and moreover, for any fixed i, $\sum_{j \neq i} \boldsymbol{w}^{ij} = \mathbf{0}$.

[2]Such matrices, and their higher-order tensor analogues, arise naturally in non-parametric statistics, under the guise of terms involved in computing the second and higher moments of a random vector having a uniform distribution over all permutations.

Remark that the covariance matrix Σ in our problem is a special case of Lemma S5, with $a = \alpha^{N,P}, b = \beta^{N,P}, c = \gamma^{N,P}$. It is a healthy exercise in algebra to verify that $a+(2N-4)b+\binom{N-2}{2}c = 0$ for our problem, and to check that Facts (A) and (B) are special cases of this general lemma.

Proof of Lemma S5. We shall construct the eigenvectors explicitly. Like Σ, the components of these eigenvectors are indexed by 2-subsets, which will allow us to reason about them combinatorially.

First, consider the vector of all 1s, $\mathbf{1}$. For a row in Σ indexed by the 2-subset $\{i, j\}$, we can freely replace i with any $k \in [N] \setminus \{j\}$ and vice-versa to obtain a 2-subset that overlaps $\{i, j\}$ by one, so there are exactly $2(N-2)$ such subsets. We can obtain a subset that does not overlap $\{i, j\}$ by choosing two indices from $[N] \setminus \{i, j\}$ so there are $\binom{N-2}{2}$ such subsets. Therefore, the row sum of Σ corresponding to pair $\{i, j\}$ is $a + (2N-4)b + \binom{N-2}{2}c$. Because $\{i, j\}$ was arbitrary, this is the row sum for each row. Therefore

$$\Sigma\mathbf{1} = \left(a + (2N-4)b + \binom{N-2}{2}c\right)\mathbf{1}.$$

We now consider a vector $\boldsymbol{x}$ constructed as follows:

$$\boldsymbol{x}_{ij} := \begin{cases} 1 & \text{if } 1 \in \{i, j\} \\ \eta & \text{if } 1 \notin \{i, j\} \end{cases},$$

with

$$\eta := \frac{-2b - (N-3)c}{(N-2)b + \binom{N-2}{2}c} = \frac{-2}{N-2}.$$

That is, for every index corresponding to a subset containing 1 the vector's entry is 1, and for all indices that do not contain 1 the entry is η. We now have two cases to consider. First, consider a row of Σ that corresponds to a pair that contains 1, and call this row Σ_{inc}.

$$\Sigma_{\text{inc}}^T\boldsymbol{x} = a + (N-2)b + (N-2)b\eta + \binom{N-2}{2}c\eta,$$

which follows because this entry contains 1 and so a has coefficient 1. Then, there are $(N-2)$ pairs that overlap this set that also contain 1, hence the term $(N-2)b$. There are also $(N-2)$ pairs that overlap this set but do not contain 1 resulting in the term $(N-2)b\eta$. Finally there are $\binom{N-2}{2}$ pairs that do not contain 1 and also do not overlap this set, corresponding to the $\binom{N-2}{2}c\eta$ term. Rearranging terms, we obtain

$$\begin{aligned}\Sigma_{\text{inc}}^T\boldsymbol{x} &= a + (N-2)b + \left[(N-2)b + \binom{N-2}{2}c\right]\eta \\ &= a + (N-2)b - 2b - (N-3)c \\ &= a + (N-4)b - (N-3)c.\end{aligned}$$

Now, consider a row of Σ that corresponds to a pair that does not contain 1, and call this row Σ_{exc}.

$$\Sigma_{\text{exc}}^T\boldsymbol{x} = a\alpha + 2b + 2(N-3)b\eta + (N-3)c + \binom{N-3}{2}c\eta.$$

The first term is because the set under consideration does not contain 1. There are then exactly two sets that overlap the present set that also contain 1, resulting in $2b$. Meanwhile, there are $2(N-3)$ sets that overlap the present set but do not contain 1 giving $2(N-3)b\eta$. The term $(N-3)c$ comes from the $N-3$ sets that contain 1 but do not overlap the present set. Finally, there are $\binom{N-3}{2}$ sets that do not overlap the present set and also do not contain 1 resulting in $\binom{N-3}{2}c\eta$. We can now

rearrange and use the definition of η to see

$$\begin{aligned}\Sigma_{\text{exc}}^T \boldsymbol{x} &= \left(a + 2(N-3)b + \binom{N-3}{2}c\right)\eta + 2b + (N-3)c \\ &= \left(a + 2(N-3)b + \binom{N-3}{2}c\right)\eta - \left((N-2)b + \binom{N-2}{2}c\right)\eta \\ &= (a + (N-4)b - (N-3)c)\,\eta.\end{aligned}$$

Therefore $\boldsymbol{x}$ is an eigenvector with eigenvalue $a + (N-4)b - (N-3)c$.

Because treating 1 as being "special" in the construction of $\boldsymbol{x}$ was arbitrary we can repeat this process N times to obtain N eigenvectors, $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^N$. We now check that these N eigenvectors span a space of dimension $N-1$. Observe that

$$\boldsymbol{x}^1 + \cdots + \boldsymbol{x}^N = \mathbf{0},$$

since each component ij of this sum of vectors is exactly $2 + (N-2)\eta = 2 - 2 = 0$. On the other hand, suppose that $\lambda_1, \ldots, \lambda_n$ are scalars such that $\sum_{k=1}^N \lambda_k \boldsymbol{x}^k = \mathbf{0}$. By considering each component of this sum, we obtain linear equations involving each 2-subset $\{i, j\}$ of $[N]$:

$$\lambda_i + \lambda_j = \frac{2}{N-2} \sum_{k \in [N] \setminus \{i,j\}} \lambda_k.$$

By adding $2/(N-2) \cdot (\lambda_i + \lambda_j)$ to each side of the equation and multiplying by $(N-2)$, we recover $N(\lambda_i + \lambda_j) = N(\lambda_{i'} + \lambda_{j'})$, from which it is easy to see that $\lambda_1 = \ldots = \lambda_N$. This shows that $\dim(\text{span}(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^N)) = N - 1$, as desired.

For the final eigenvalue, we now consider 2 "special" indices. Without loss of generality take these special indices to be 1 and 2. We will separately consider subsets that overlap zero times, once, or twice of $\{1, 2\}$ and allow each to have independent values. Concretely, consider the vector $\boldsymbol{w}$ whose components are given by

$$\boldsymbol{w}_{ij} := \begin{cases} 1 & \text{if } \{1,2\} \cap \{i,j\} = 2 \\ \zeta & \text{if } \{1,2\} \cap \{i,j\} = 1 \\ \vartheta & \text{if } \{1,2\} \cap \{i,j\} = 0 \end{cases},$$

with

$$\begin{aligned}\zeta &:= \frac{-1}{N-2} \\ \vartheta &:= \frac{1}{\binom{N-2}{2}}.\end{aligned}$$

Now, as before, we consider the corresponding three types of rows of Σ. First, consider the row that corresponds to the focal pair $\{1, 2\}$, which we will call Σ_{focal}. We see

$$\Sigma_{\text{focal}}^T \boldsymbol{w} = a + (2N-4)b\zeta + \binom{N-2}{2}c\vartheta,$$

which follows easily because for the focal pair, the sets that overlap exactly, partially, or not at all are exactly those that we used to construct $\boldsymbol{w}$. Substituting the definitions of ζ and ϑ we get

$$\Sigma_{\text{focal}}^T \boldsymbol{w} = a - 2b + c.$$

Consider a row whose indexing subset partially overlaps $\{1, 2\}$, and call it Σ_{partial}.

$$\Sigma_{\text{partial}}^T \boldsymbol{w} = \zeta a + b + (N-2)\zeta b + (N-3)\vartheta b + (n-3)\zeta c + \binom{n-3}{2}\vartheta c.$$

By construction $\{1,2\}$ partially overlaps the present subset and vice-versa giving us $\zeta a + b$. There are $N-2$ subsets that partially overlap $\{1,2\}$ and the present subset resulting in $(N-2)\zeta b$. The term $(N-3)\vartheta b$ comes from the $N-3$ pairs that do not overlap $\{1,2\}$ but do partially overlap the present subset. Conversely, there are $(N-3)$ subsets that do not overlap the present subset but do overlap $\{1,2\}$ giving $(N-3)\zeta c$. Finally, there are $\binom{N-3}{2}$ subsets that do not overlap the present pair of $\{1,2\}$ resulting in $\binom{N-3}{2}\vartheta c$. We again collect terms and apply the definitions of ζ and ϑ to see

$$\begin{aligned}\Sigma^T_{\text{partial}}\boldsymbol{w} &= \Big(a + (N-2)b + (N-3)c\Big)\zeta + b + (N-3)\vartheta b + \binom{N-3}{2}\vartheta c \\ &= \Big(a + (N-2)b + (N-3)c\Big)\zeta - (N-2)b\zeta - \frac{(N-2)(N-3)}{\binom{N-2}{2}}b\zeta - \frac{(N-2)\binom{N-3}{2}}{\binom{N-2}{2}}c\zeta \\ &= (a - 2b + c)\zeta\end{aligned}$$

Lastly, consider a row whose indexing subset does not overlap $\{1,2\}$ and call it Σ_{non}.

$$\Sigma^T_{\text{non}}\boldsymbol{w} = \vartheta a + \ c + 4\zeta b + (2N-8)\vartheta b + (2N-8)\zeta c + \binom{N-4}{2}\vartheta c,$$

where we get ϑa and c because the present subset does not overlap $\{1,2\}$ and vice-versa. There are exactly 4 subsets that partially overlap $\{1,2\}$ and the present subset, giving $4\zeta b$. Meanwhile, there are $(2N-8)$ sets that overlap the present subset but do not overlap $\{1,2\}$ and vice-versa giving $(2N-8)\vartheta b$ and $(2N-8)\zeta c$. Lastly, there are $\binom{N-4}{2}$ subsets that do not overlap the present subset or $\{1,2\}$ yielding $\binom{N-4}{2}\vartheta c$. Rearranging we see

$$\begin{aligned}\Sigma^T_{\text{non}}\boldsymbol{w} &= \left(a + (2N-8)b + \binom{N-4}{2}c\right)\vartheta + c + 4\zeta b + (2N-8)\zeta c \\ &= \left(a + (2N-8)b + \binom{N-4}{2}c\right)\vartheta + \binom{N-2}{2}c\vartheta - \frac{4\binom{N-2}{2}}{N-2}b\vartheta - \frac{(2N-8)\binom{N-2}{2}}{N-2}c\vartheta \\ &= (a - 2b + c)\,\vartheta.\end{aligned}$$

Therefore $\boldsymbol{w}$ is an eigenvector with eigenvalue $a - 2b + c$.

We again can repeat this process choosing different "special" pairs of indices. This results in $\binom{N}{2}$ eigenvectors, $\boldsymbol{w}^{12}, \ldots, \boldsymbol{w}^{N-1,N}$. It is easy to see that there are N redundant eigenvectors (e.g., consider all pairs that contain 1 — the eigenvector corresponding to the last pair is a linear combination of the first $N-2$). By mirroring the analysis in the case of the second eigenvalue above, we see that these $\binom{N}{2}$ eigenvectors span a space of dimension $\binom{N}{2} - N = \binom{N-1}{2} - 1$. □

Lemma S5 implies the following eigendecomposition for our covariance matrix Σ. Let

$$\Sigma = U\Lambda U^T, \tag{S6}$$

with

$$\Lambda = \operatorname{diag}\left(\underbrace{\alpha^{N,P} + (N-4)\beta^{N,P} - (N-3)\gamma^{N,P}}_{(N-1)\text{ times}}, \underbrace{\alpha^{N,P} - 2\beta^{N,P} + \gamma^{N,P}}_{\binom{N-1}{2}-1\text{ times}}, 0\right),$$

and the orthogonal matrix $U = \begin{bmatrix} | & \cdots & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_{\binom{N}{2}} \\ | & \cdots & | \end{bmatrix}$ satisfying

- $\mathbf{u}_1, \ldots, \mathbf{u}_{N-1}$ obtained by performing Gram-Schmidt on the set $\{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^N\}$ (should obtain $N-1$ orthogonal vectors from $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^{N-1}$).

- $\mathbf{u}_N, \ldots, \mathbf{u}_{\binom{N}{2}-1}$ obtained by performing Gram-Schmidt on the set $\left\{\boldsymbol{w}^{ij} : ij \in \binom{[N]}{2}\right\}$.
- $\mathbf{u}_{\binom{N}{2}} = \frac{1}{\sqrt{\binom{N}{2}}}\mathbf{1}$.

By (S2), $\vec{Y}^* = U^T\vec{M}^*$ satisfies

$$\sqrt{P} \cdot \vec{Y}^* \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Lambda).$$

Because the map defined by (S5) is invariant to orthogonal transformations ($\|\vec{Y}\|_2^2 = \vec{M}^T U U^T \vec{M} = \|\vec{M}\|_2^2$), we obtain by the continuous mapping theorem

$$V^{(N,P)*} = f(\sqrt{P}\vec{M}^*) = f(\sqrt{P}\vec{Y}^*) \xrightarrow{d} V^{(N,\infty)},$$

which concludes the proof. □

Proof of Theorem S3. The proof of Theorem 2.2 showed that when P is large, $V^{(N,P)*}$ is approximately a weighted sum of $\binom{N}{2}-1$ chi-square random variables. Concretely, let $\lambda_1 = \ldots = \lambda_{N-1} = \sigma_1^2$ and $\lambda_N = \ldots = \lambda_{\binom{N}{2}-1} = \sigma_2^2$, where

$$\sigma_1^2 = a_1^N, \quad \sigma_2^2 = a_2^N$$

with the quantities a_1^N and a_2^N defined in Theorem 2.2. Then,

$$V^{(N,P)*} \overset{d}{=} \frac{1}{\binom{N}{2}}\left(Y_1^2 + \ldots + Y^2_{\binom{N}{2}-1}\right)$$

holds approximately as long as P is large, where $Y_n \overset{\text{ind}}{\sim} N(0, \lambda_n)$. Thus, we can immediately apply the Central Limit Theorem (in N) to the sequence $\{Y_n^2 : n = 1, \ldots, \binom{N}{2}\}$ and obtain the conclusion. All that remains is to check that the mean and variance of $V^{(N,\infty)}$ converge to the quantities stated in Theorem S3. For the variance, there is nothing to check. For the mean, notice that

$$\begin{aligned}
\lim_{N\to\infty} \mathbb{E}[V^{(N,\infty)}] &= \lim_{N\to\infty}\left[\frac{a_1^N(N-1) + a_2^N\left(\binom{N-1}{2} - 1\right)}{\binom{N}{2}}\right] \\
&= \lim_{N\to\infty}\left[\left(1 - \frac{1}{\binom{N}{2}}\right)\alpha^N + \left(\frac{2}{\binom{N}{2}} - \frac{4}{N}\right)\beta^N - \left(1 + \frac{1}{\binom{N}{2}} - \frac{4}{N}\right)\gamma^N\right] \\
&= \alpha,
\end{aligned}$$

where the last equality follows from $\lim_{N,P\to\infty}\gamma^{N,P} = 0$ (a healthy exercise) and the fact that α^N, β^N and γ^N are uniformly bounded in N.

Finally, as a technical point, because $a_1^{N,P}$ and $a_2^{N,P}$ are uniformly bounded in both N and P, we have $\lim_{N,P\to\infty} a_i^{N,P} = \lim_{N\to\infty}\lim_{P\to\infty} a_i^{N,P} = \lim_{N\to\infty} a_i^N$ for $i = 1, 2$, assuming these limits exist. Similar reasoning justifies $\alpha = \lim_{N,P\to\infty}\alpha^{N,P} = \lim_{N\to\infty}(\lim_{P\to\infty}\alpha^{N,P}) = \lim_{N\to\infty}\alpha^N$ and $0 = \lim_{N,P\to\infty}\gamma^{N,P} = \lim_{N\to\infty}(\lim_{P\to\infty}\gamma^{N,P}) = \lim_{N\to\infty}\gamma^N$, which are equalities implicitly invoked in the previous paragraph. □

Proof of Theorem S4. Our proof strategy is identical to the proof of Theorem 2.2. However, unlike the latter, where there are P independent features, here we have B independent blocks. Concretely,

define

$$\mathbb{R}^{\binom{N}{2}} \ni \boldsymbol{v}_b = \begin{pmatrix} d^{(b)}(\mathbf{x}_1, \mathbf{x}_2) \\ d^{(b)}(\mathbf{x}_1, \mathbf{x}_3) \\ \vdots \\ d^{(b)}(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots \\ d^{(b)}(\mathbf{x}_{N-1}, \mathbf{x}_N) \end{pmatrix} - \overline{d^{(b)}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})} \cdot \vec{1},$$

which is the vector of partial Hamming distances associated with block b. Let $\Sigma_b = \mathbb{E}[\boldsymbol{v}_b^*(\boldsymbol{v}_b^*)^T]$. Then, by mirroring Step 2 in the proof of Theorem 2.2, it holds that

$$\sqrt{B}\left(\frac{1}{B}\sum_{b=1}^{B} \boldsymbol{v}_b^*\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma), \tag{S7}$$

where $\Sigma := \frac{1}{B}(\Sigma_1 + \ldots + \Sigma_B)$.

Now, since $\vec{M}$ defined in Step 1 of the proof of Theorem 2.2 satisfies $\vec{M} = \frac{B}{P}(\frac{1}{B}\sum_{b=1}^{B} \boldsymbol{v}_b)$ and $V^{(N,P)} = V^{(N,B,P)} = \frac{P}{\binom{N}{2}}\|\vec{M}\|_2^2$,

$$V^{(N,B,P)*} = \frac{P}{\binom{N}{2}}\left\|\frac{B}{P}\left(\frac{1}{B}\sum_{b=1}^{B} \boldsymbol{v}_b^*\right)\right\|_2^2 = \frac{1}{\binom{N}{2}}\left\|\sqrt{\frac{B}{P}} \cdot \sqrt{B}\left(\frac{1}{B}\sum_{b=1}^{B} \boldsymbol{v}_b^*\right)\right\|_2^2.$$

Applying (S7) to the expression on the RHS above, we see that, similar to Step 3 in the proof of Theorem 2.2, if $\Sigma' = \frac{B}{P} \cdot \Sigma$ were diagonal, then the continuous mapping theorem would imply that $V^{(N,B,P)*}$ is approximately a weighted sum of chi-square random variables, where the weights are the diagonal entries of Σ'.

To finish the proof, we apply Lemma S5 to the matrix Σ'. It is a healthy exercise to reason that $\alpha^{N,B,P}, \beta^{N,B,P}$ and $\gamma^{N,B,P}$ are the analogues of $\alpha^{N,P}, \beta^{N,P}$ and $\gamma^{N,P}$ in Theorem 2.2, and that the argument of Step 3 there carries over mutatis mutandis to here. □

Proof of Theorem 5.1. We will rely on the following Berry-Esséen bound for sums of independent random vectors, due to Raič (2019). Suppose that $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_P$ is a collection of independent but not necessarily identically distributed zero mean D-dimensional random vectors, and assume that $\sum_{p=1}^{P} \mathrm{Var}(\boldsymbol{v}_p) = I_D$. Define the random vector $\boldsymbol{w} = \sum_{p=1}^{P} \boldsymbol{v}_p$, and denote the standard D-variate Gaussian law by $\mathcal{N}(\mathbf{0}, I_D)\{\cdot\}$, so that for any measurable set $\mathscr{A} \subseteq \mathbb{R}^D$, $\mathcal{N}(\mathbf{0}, I_D)\{\mathscr{A}\} = \int_{\mathscr{A}} (2\pi)^{-D/2} \exp\left(-\frac{1}{2}\|\mathbf{x}\|_2^2\right) d\mathbf{x}$. Then, for any measurable *convex* set $\mathscr{A} \subseteq \mathbb{R}^D$,

$$|\mathbb{P}(\boldsymbol{w} \in \mathscr{A}) - \mathcal{N}(\mathbf{0}, I_D)\{\mathscr{A}\}| \leqslant (42D^{1/4} + 16)\sum_{p=1}^{P} \mathbb{E}\|\boldsymbol{v}_p\|_2^3. \tag{S8}$$

To apply the bound above, we let (a modification of) the random vectors $\boldsymbol{v}_1^*, \ldots, \boldsymbol{v}_P^*$, as defined in the proof of Theorem 2.2, play the role of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_P$. As in the proof of Theorem 2.2, we first describe our proof strategy before presenting the details.

Steps Outlining Proof

(1) Denoting $\mathrm{Var}(\boldsymbol{v}_p^*) = \Sigma_p$ and $\vec{M} = \frac{1}{P}(\boldsymbol{v}_1 + \ldots + \boldsymbol{v}_P)$ just like we did in the proof of Theorem 2.2, we saw in that proof that the covariance matrix $\Sigma = \frac{1}{P}(\Sigma_1 + \ldots + \Sigma_P)$ of $\sqrt{P} \cdot \vec{M}^*$ has eigenvalue 0 with multiplicity 1. We perform a truncation so that the resulting collection of independent random vectors has an invertible covariance matrix allowing the Berry-Esséen bound to be applied.

(2) Using the identity $V^{(N,P)*} = \frac{P}{\binom{N}{2}} \| \frac{1}{P} \sum_{p=1}^{P} \boldsymbol{v}_p^* \|_2^2$ (see Step 1 of proof of Theorem 2.2), we relate the multi-dimensional Berry-Esséen bound to the total variation bound for the random variable $V^{(N,P)*}$.

Step 1: Recall the orthogonal matrix U (see (S6)) from the proof of Theorem 2.2, whose columns are the eigenvectors of Σ and moreover satisfies $\Sigma = U \Lambda U^T$. Define $\vec{Y}_p^* = \frac{1}{\sqrt{P}} U^T \boldsymbol{v}_p^*$ for $p = 1, \ldots, P$. (Note that U and Λ depend only on $\boldsymbol{c}$, which is fixed.).

By the invariance of the test statistic to orthogonal transformations of $\vec{M}$, we see that $V^{(N,P)*} = \frac{1}{\binom{N}{2}} \| \vec{Y}^* \|_2^2$, where $\vec{Y}^* = \vec{Y}_1^* + \ldots + \vec{Y}_P^*$ and satisfies $\mathrm{Cov}(\vec{Y}^*) = \Lambda$. Moreover, because $\mathbb{E}[\vec{Y}^*] = \sqrt{P} \cdot U^T \mathbb{E}[\vec{M}^*] = \mathbf{0}$, the last component of $\vec{Y}^*$ is 0.

Define $\vec{W}_p = \vec{Y}_p[:-1]$, which is $\vec{Y}_p$ without its last component. Observe that $\{\vec{W}_p^* : p = 1, \ldots, P\}$ is a collection of independent random variables, and moreover $V^{(N,P)*} = \frac{1}{\binom{N}{2}} \| \vec{W}_1^* + \ldots + \vec{W}_P^* \|_2^2$, because the last component of $\vec{Y}^*$ is 0. Letting $\vec{W} = \sum_{p=1}^{P} \vec{W}_p$, we have $\mathrm{Cov}(\vec{W}^*) = \Lambda'$, where $\Lambda' = \mathrm{diag}(\lambda_1, \ldots, \lambda_1, \lambda_2, \ldots, \lambda_2)$ is the invertible diagonal matrix obtained from excluding the last row and last column of Λ. (Note that λ_1 and λ_2 depend on $\alpha^{N,P}, \beta^{N,P}$ and $\gamma^{N,P}$, which in turn depend only on $\boldsymbol{c}$.)

Now, let $D = \binom{N}{2} - 1$. Then, $\{(\Lambda')^{-1/2} \vec{W}_p^* : p = 1, \ldots, P\}$ is a collection of $\mathbb{R}^D$-valued independent random vectors, each with mean $\mathbf{0}$. Letting $\boldsymbol{w}$ denote their sum, we see that $\mathrm{Cov}(\boldsymbol{w}) = (\Lambda')^{-1/2} \Lambda' [(\Lambda')^{-1/2}]^T = I_D$. Thus, (S8) holds with $(\Lambda')^{-1/2} \vec{W}_p^*$ in place of $\boldsymbol{v}_p$ in the upper bound expression.

Step 2: Given the collection $\{(\Lambda')^{-1/2} \vec{W}_p^* : p = 1, \ldots, P\}$ satisfying the Berry-Esséen bound, we now verify that the multidimensional bound translates into the desired rate of convergence. To this end, we first establish a further upper bound on the multidimensional bound that depends only on the dimensionality of the problem. Let us bound each term $\mathbb{E} \| (\Lambda')^{-1/2} \vec{W}_p^* \|_2^3$ that appears on the RHS of (S8). Observe

$$\begin{aligned} \| (\Lambda')^{-1/2} \vec{W}_p \|_2^3 &= (\vec{W}_p (\Lambda')^{-1} \vec{W}_p)^{3/2} \\ &\leqslant (\lambda_1 \wedge \lambda_2)^{-3/2} \| \vec{W}_p \|_2^3 \\ &\leqslant (\lambda_1 \wedge \lambda_2)^{-3/2} \| \vec{Y}_p \|_2^3 \\ &\leqslant P^{-3/2} (\lambda_1 \wedge \lambda_2)^{-3/2} \| \boldsymbol{v}_p \|_2^3. \end{aligned}$$

Let us bound the RHS expression above. By noticing that $\boldsymbol{v}_p$ is a vector that contains $c_p(N - c_p)$ copies of $1 - \frac{c_p(N-c_p)}{\binom{N}{2}}$ and $\binom{N}{2} - c_p(N - c_p)$ copies of $\frac{-c_p(N-c_p)}{\binom{N}{2}}$, it is a healthy exercise to show that $\| \boldsymbol{v}_p \|_2^3 \leqslant \frac{1}{4} \binom{N}{2}^{3/2} = \frac{1}{4} (D+1)^{3/2}$. Combining these inequalities, we obtain $\mathbb{E} \| (\Lambda')^{-1/2} \vec{W}_p^* \|_2^3 \leqslant \frac{1}{4} \left(\frac{D+1}{P \cdot \lambda_{\min}} \right)^{3/2}$, where $\lambda_{\min} = \lambda_1 \wedge \lambda_2$. Therefore, we have

$$|\mathbb{P}(\boldsymbol{w} \in \mathscr{A}) - \mathcal{N}(\mathbf{0}, I_D)\{\mathscr{A}\}| \leqslant (42 D^{1/4} + 16) \cdot P \cdot \frac{1}{4} \left(\frac{D+1}{P \cdot \lambda_{\min}} \right)^{3/2} \leqslant C \cdot P^{-1/2} \cdot D^{7/4}, \tag{S9}$$

where C is a constant depending only on $\boldsymbol{c}$ and $\mathscr{A}$ is any measurable convex set.

Given that we now have an upper bound in (S9) that depends only on the dimensionality of the problem, we shall use it to derive our desired total variation bound. Observe that for any convex set $\mathscr{A} \subseteq \mathbb{R}^D$, $\boldsymbol{w} \in \mathscr{A}$ if and only if $\vec{W}^* \in (\Lambda')^{1/2} \mathscr{A}$, where $(\Lambda')^{1/2} \mathscr{A} = \{(\Lambda')^{1/2} \boldsymbol{y} : \boldsymbol{y} \in \mathscr{A}\}$, because the map $\mathscr{A} \mapsto (\Lambda')^{-1/2} \mathscr{A}$ is invertible. This map also preserving measurability and convexity, we obtain $|\mathbb{P}(\vec{W}^* \in \mathscr{A}) - \mathcal{N}(\mathbf{0}, \Lambda')\{\mathscr{A}\}| \leqslant C \cdot P^{-1/2} \cdot D^{7/4}$ for any convex and measurable $\mathscr{A}$. Now, recall

that $V^{(N,P)*} = \frac{1}{\binom{N}{2}}\|\vec{W}^*\|_2^2$. Consider the function $f : \mathbb{R}^D \to \mathbb{R}_0^+$ given by $f(\mathbf{z}) = \frac{1}{\binom{N}{2}}\|\mathbf{z}\|_2^2$. Observe that $V^{(N,\infty)} \stackrel{d}{=} f(\mathbf{z})$, where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \Lambda')$. Define, for any $t \geqslant 0$, $\mathscr{A}(t) = \{\mathbf{z} \in \mathbb{R}^D : f(\mathbf{z}) \leqslant t\}$. Then $\mathscr{A}(t)$ is convex (a healthy exercise). Moreover $\mathbb{P}(\vec{W}^* \in \mathscr{A}(t)) = \mathbb{P}(V^{(N,P)*} \leqslant t)$ (the CDF of the permutation null) and $\mathcal{N}(\mathbf{0}, \Lambda')\{\mathscr{A}(t)\} = \mathbb{P}(V^{(N,\infty)} \leqslant t)$ (the CDF of the large P distribution). These observations together imply that for any $t \geqslant 0$,

$$|\mathbb{P}(V^{(N,P)*} \leqslant t) - \mathbb{P}(V^{(N,\infty)} \leqslant t)| \leqslant \frac{C \cdot D^{7/4}}{\sqrt{P}},$$

which is the desired total variation bound (after absorbing $D^{7/4} = [\binom{N}{2} - 1]^{7/4}$ into the definition of the constant C). □

H Simulation Details

H.1 Type I Error Control Study

In Subsection 3.1 we describe how we simulate data from null models to investigate FPR control. Here, we summarize the results of this simulation study.

Based on the setup described in Subsection 3.1, we perform Monte Carlo sampling to estimate the FPR of our test, and report approximate 95% confidence intervals using the point estimate of the FPR.

We find that our test is exact, meaning that it keeps the Type I error rate at the desired significance level, with occasional small fluctuations due to finite sampling. In the case of $P = 100$ features per sampled row, as the middle row of Figure S2 shows, FPR estimates of our test, together with their 95% confidence intervals, include the nominal significance threshold, with slight deviations occurring mostly when the sample size N is small. (The slight deviations are an artifact of the asymptotic test used when $P = 100$.) On the other hand, the FPR of the TW test is markedly different from the nominal significance threshold, except when N is large. When we simulated datasets with $P = 10$ and $P = 1000$ features, corresponding to small and big datasets, we find that our test remains exact, whereas TW suffers from worse FPR control than seen in the $P = 100$ case, with the FPR estimates and their 95% confidence intervals not including the nominal significance threshold a majority of the time. In particular, when $P = 1000$ and $N \leqslant 30$, the Type I error rate is markedly *higher* than the nominal significance threshold (see the last row of Figure S2). This phenomenon reflects the impact of the scaling factor N/P and the sizes of P and N on the accuracy of the asymptotic TW distribution, an issue that does not surface for our finite-sample and non-parametric approach.

H.2 Statistical Power Study

In Subsection 3.2 of the Main Text we describe a framework comprising seven scenarios and their pairings (Table 1), in order to investigate the statistical power of exchangeability tests such as ours. Here, we provide parametrization details for our seven scenarios and summarize the results of the simulation study.

First, the parametrization details.

(1) *Number of observations.* We generally set $N \in \{10, 50, 100, 500, 1000\}$ to cover a range of small, moderate and large sample sizes; see Scenario 3 for exceptions.
(2) *Closeness of population features or parameters.* We control closeness by adjusting the hyperparameter ε of our hierarchical model ($\varepsilon \in \{0.05, 0.1, 0.15, 0.2\}$).
(3) *Multiple populations.* We consider $K \in \{2, 3, 4, 5\}$. For $K = 3$ and $K = 4$, to ensure both the number of observations drawn from each population is the same and the total number of observations is close to the $K = 2$ default case, we set $N \in \{12, 60, 120, 600, 1200\}$.

(4) *Sparsity of discerning features.* We consider $f \in \{0.1, \ldots, 0.9\}$, and let fP features be discerning between our K populations while the remaining $(1-f)P$ features be non-discerning features that are identically distributed across all K populations in reality but included when performing the statistical test.

(5) *Uneven sampling.* Sample evenness depends on the sampling design when the dataset is not obtained from clustering. On the other hand, uneven representation of populations in a dataset is typical of clusters obtained from partitioning an originally larger dataset, as we expect good clustering algorithms to recover (approximately) homogeneous communities. We consider this scenario for $K = 2$ distinct populations from which observations were drawn to make up the sample, and perform draws with ratios $r \in \{9/1, 8/2, \ldots, 2/8, 1/9\}$ to make up the sample. For example, to form a size $N = 100$ sample with draw ratio $9:1$, we draw 90 observations from Population 1 and 10 observations from Population 2.

(6) *Different sources of heterogeneity.* For $K = 2$ distinct populations, we consider two sources of heterogeneity, which impact the row sums of the overall sample: (i) overall differences in frequencies across all markers; (ii) differences in frequencies across all markers, despite the average marker frequencies for each population being roughly equal. Concretely, in category (i), we draw marker frequencies from the P features for Population 1 and Population 2 as described in Step 2 of the generative process reported in Subsection 3.2. In category (ii), we draw $P/2$ marker frequencies for Population 1 and for Population 2 from the uniform distribution described in Step 2, and then append these marker frequency vectors to obtain two different length P marker frequency vectors for the two populations. For example, with $P = 10$, starting from $(0.42, 0.435, 0.44, 0.422, 0.421)$ and $(0.575, 0.58, 0.572, 0.6, 0.61)$, we get $(0.42, 0.435, 0.44, 0.422, 0.421, 0.575, 0.58, 0.572, 0.6, 0.61)$ and $(0.575, 0.58, 0.572, 0.6, 0.61, 0.42, 0.435, 0.44, 0.422, 0.421)$ as the two final marker frequency vectors.

To visualize the difference between the two categories, consider the two datasets below, where the first two individuals belong to one population and the last two individuals to another distinct population. Both datasets have differences in frequencies across all markers, but the subsamples forming the right dataset have the same average marker frequencies $((1+1+1+0+0+0)/6 = 0.5)$.

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \quad ; \quad \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

Category (i) Category (ii)

Different sources of heterogeneity arise, for instance, often in biological datasets, where technical variation rather than true biological signal accounts for Category (i), which means data pre-processing is required to reduce overall differences before testing for differences that manifest in terms of Category (ii).

(7) *Column flipping.* For binary or binarizable markers, where the binarization provides an interpretation of '1' and '0' for the resulting binary array, this refers to erroneous or opposite binarization, which could arise from errors executed earlier in the data processing pipeline. We perform randomized column flipping after simulating the dataset to further simulate erroneous binarization, and do this alongside a simulation without erroneous binarization to analyze the impact of erroneous binarization on statistical power. We should expect no impact of this procedure on our test, given it is invariant to the "direction" of binarization.

An application of the multiplication principle shows that the number of sets of simulations for power estimation is

$$\underbrace{5 \times 4}_{\text{scenarios 1-2}} \times \underbrace{(4+9+9+1+1)}_{\text{scenarios 3-7}} \times \underbrace{3}_{\text{value of } P} \times \underbrace{2}_{\text{test choice}} = 2880.$$

We find that our test demonstrates robustness to uneven sampling (see Figures S5, S8 and S11). For example, for a small number of features $P = 10$ (Figure S5), when populations are sufficiently far apart, statistical power on an unevenly sampled dataset is no lower than 50% of the maximum power estimated for an evenly sampled dataset. Moreover, when the sample size is large enough, this ratio increases to 80%. In comparison, the TW test sees a drop in power from 1.00 to less than 0.4 for a large sample size $N = 500$.

We also find that the power of our test never falls below the nominal significance level α at which statistical power β is estimated. This is true across all sample sizes N, numbers of features P, and non-null scenarios considered. In comparison, the power of the TW test falls below α, typically when P or N is small, or when the dataset is unevenly sampled, or when the populations from which observations are drawn are very close to each other. (See Figures S3-S11.)

Figure S12 illustrates our two main findings in the case of uneven sampling, where violin plots of statistical powers are compared between our test and the TW test across varying numbers of features P and degrees of evenness. Degree of evenness is measured by the binary Shannon entropy of the empirical frequencies of each sampled population included in the dataset (see Scenario 4 of Table 1); a higher quantity means more evenness. Figure S12 shows that the power of our test always lies to the right of the nominal significance level and also stochastically dominates the power of the TW test in extremely unevenly sampled datasets.

H.3 Area under the receiver-operating curve (AUROC)

In Section 3.3 of the Main Text we report that 4752 AUROCs are computed. We show how we arrive at this number. First, we restrict pairs of sample sizes and feature dimensionality $(N, P) \in \{10, 50, 100\} \times \{10, 100, 1000\}$, since we may only find null and non-null models that generate datasets sharing such dimensionalities. Second, of any such pair, there are 4 non-null scenarios corresponding to Multiple Populations (Scenario 3), Different Sources of Heterogeneity (Scenario 6), and Column Flipping versus Normal (Scenario 7); 9 non-null scenarios corresponding to Sparsity of Discerning Features (Scenario 4); and another 9 non-null scenarios corresponding to Uneven Sampling (Scenario 5). Third, of any such pair, there are 3 null scenarios corresponding to low frequencies ("sparse"), varying frequencies ("regular") and high frequencies ("dense"). Finally, there are 4 choices of hyperparameter ε controlling the Closeness of Population Features (Scenario 2), and 2 choices of tests to evaluate AUROCs for.

An application of the multiplication principle shows that the number of AUROCs is

$$\underbrace{3 \times 3}_{(N,P) \text{ pairs}} \times \underbrace{(4+9+9)}_{\text{non-nulls}} \times \underbrace{3}_{\text{nulls}} \times \underbrace{4}_{\text{closeness hyperparameter}} \times \underbrace{2}_{\text{test choice}} = 4752.$$

References

Billera, L. J. et al. (2001) Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, **27**, 733–767.

Diaconis, P. and Sturmfels, B. (1998) Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics*, **26**, 363–397.

Hemerik, J. and Goeman, J. (2018) Exact testing with random permutations. *Test*, **27**, 811–825.

Raič, M. (2019) A multivariate Berry–Esséen theorem with explicit constants. *Bernoulli*, **25**, 2824–2853.

Supplementary Figures

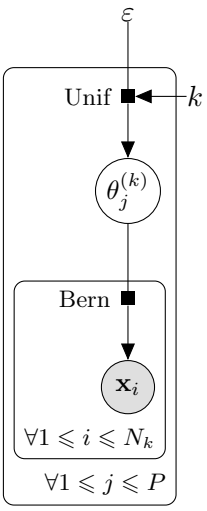

FIGURE S1. Plate diagram for our generative process. Specifically, the generative mechanism for observations drawn from an arbitrary population k ($1 \leqslant k \leqslant K$) is shown. Note that the endpoints of the uniform distribution (i.e., the quantities a_k and b_k of the supporting interval $[a_k, b_k]$ from which $\theta_j^{(k)}$ is uniformly drawn) depend on k.

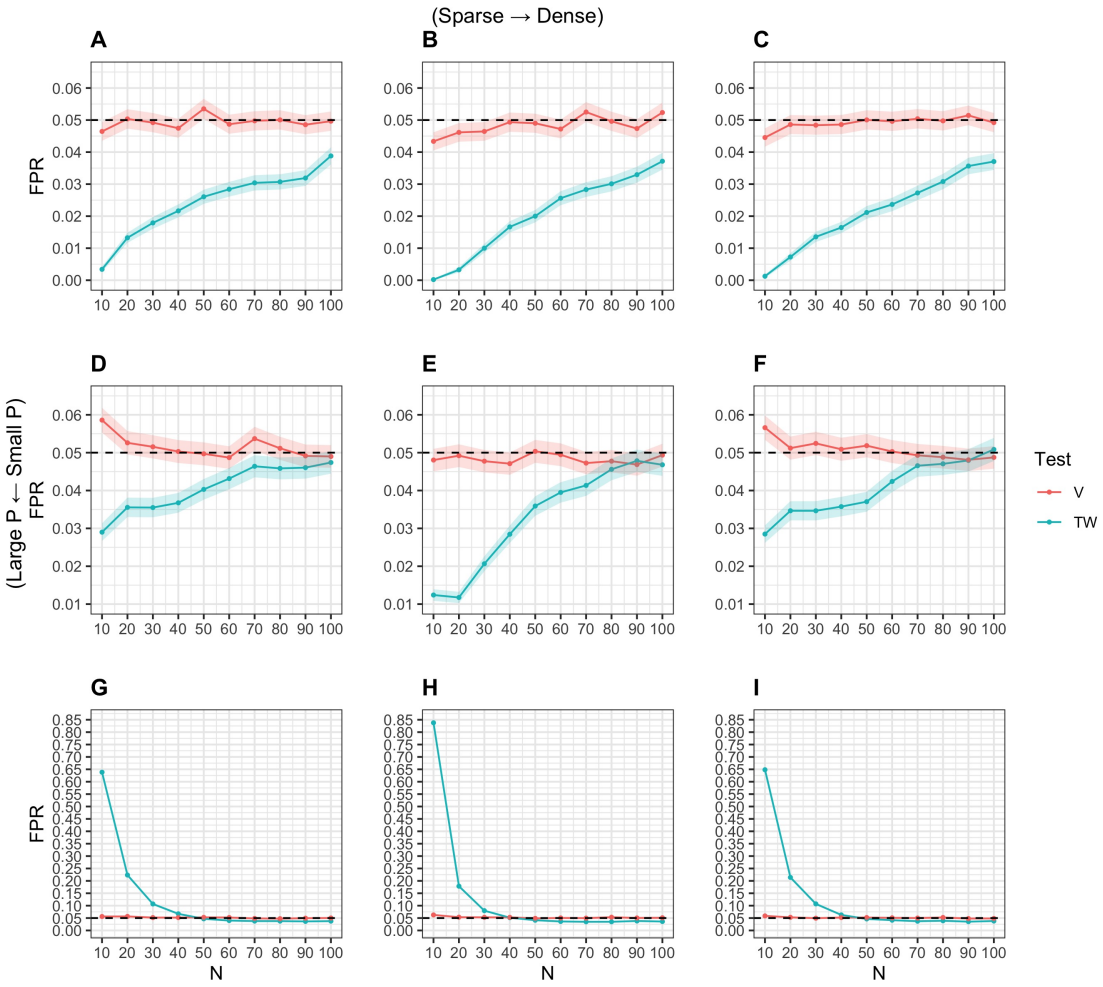

FIGURE S2. FPR of our test versus TW at significance threshold $\alpha = 0.05$, plotted across both numbers of features P (10, 100 or 1000) and marker frequency scenarios. The solid lines connect Monte Carlo estimates of the FPR at each value of N, with red corresponding to our test and blue corresponding to TW.

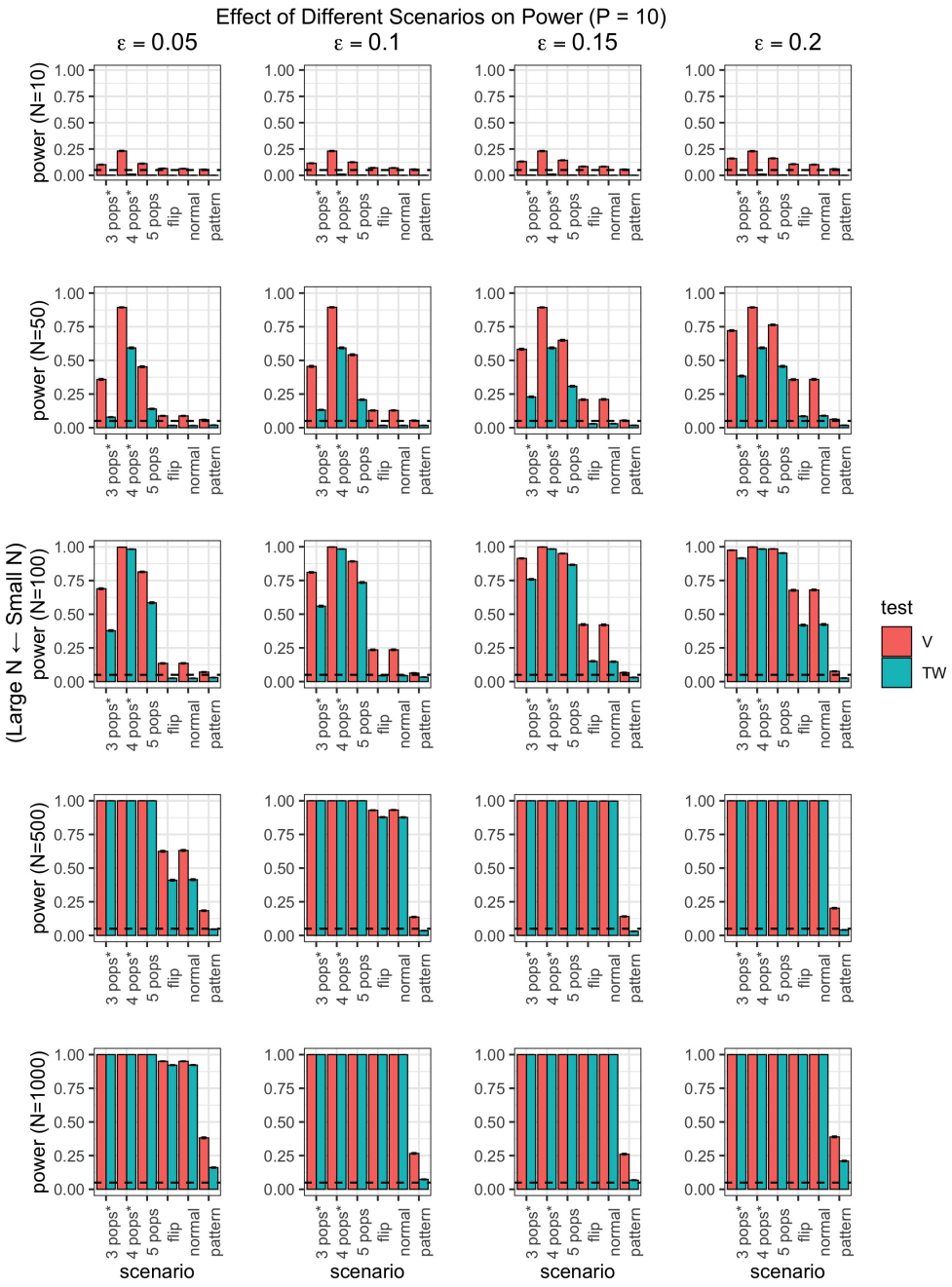

FIGURE S3. The impact of various scenarios on the statistical power of V and TW tests, with $P = 10$ features considered.

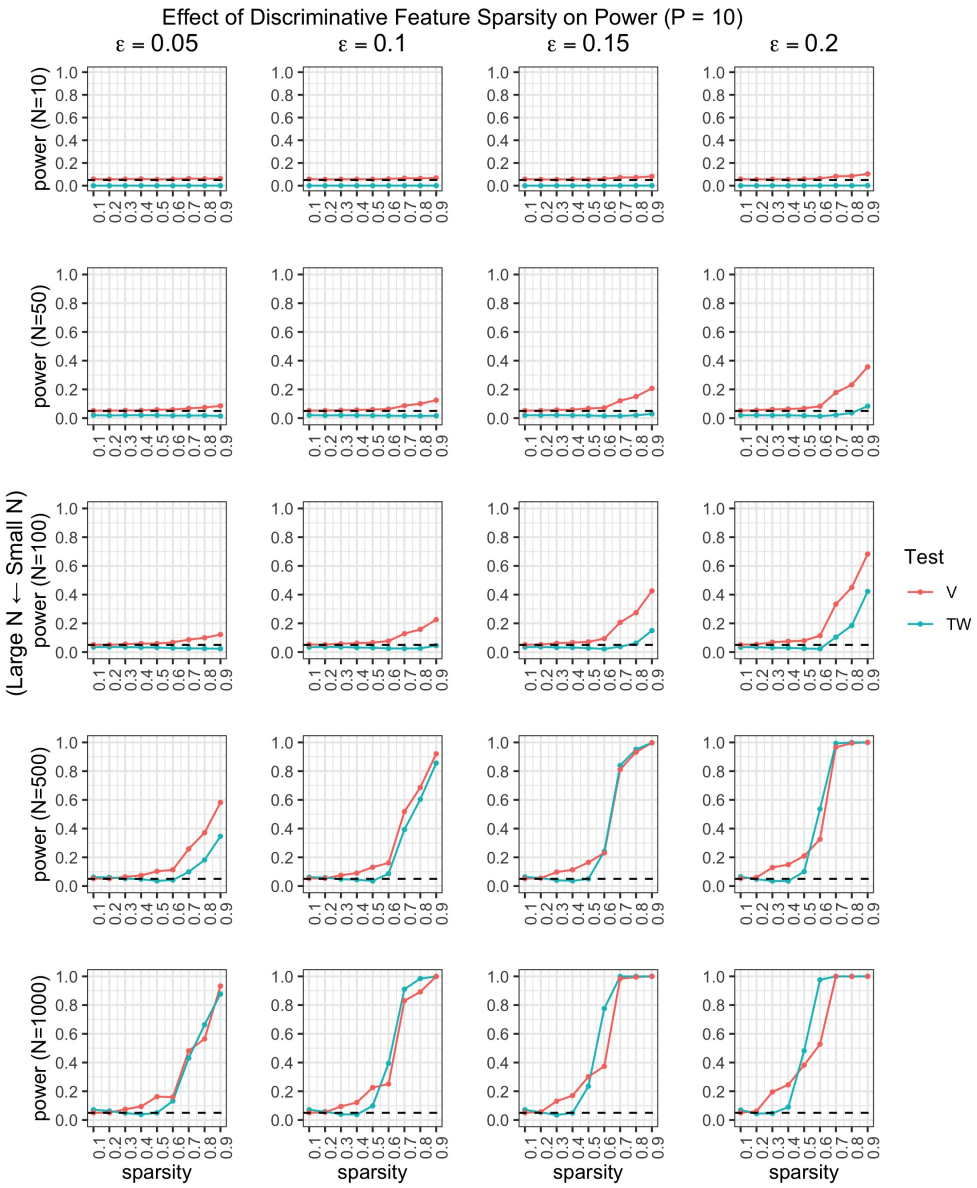

FIGURE S4. The impact of the sparsity of discerning features on the statistical power of V and TW tests, with $P = 10$ features considered.

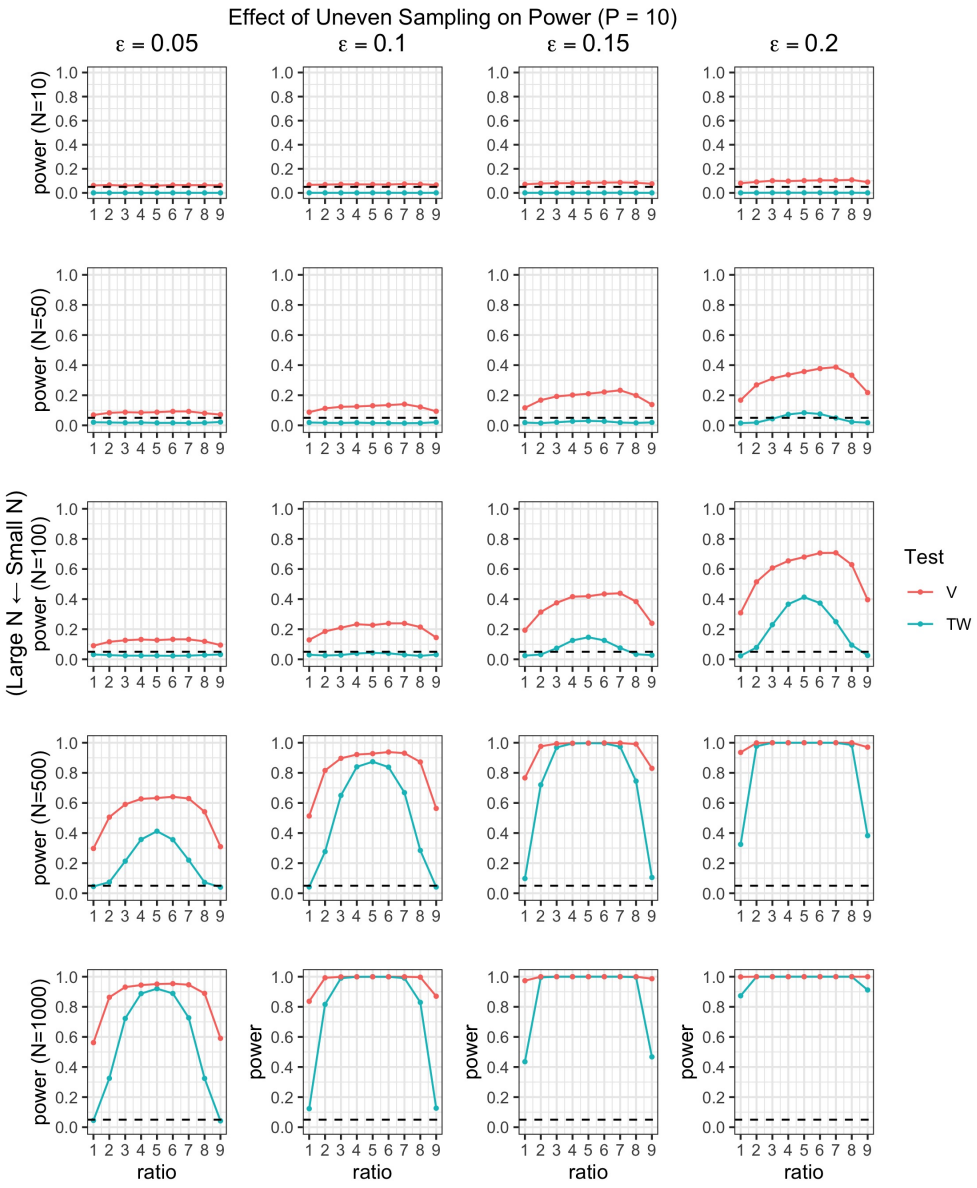

FIGURE S5. The impact of uneven sampling on the statistical power of V and TW tests, with $P = 10$ features considered.

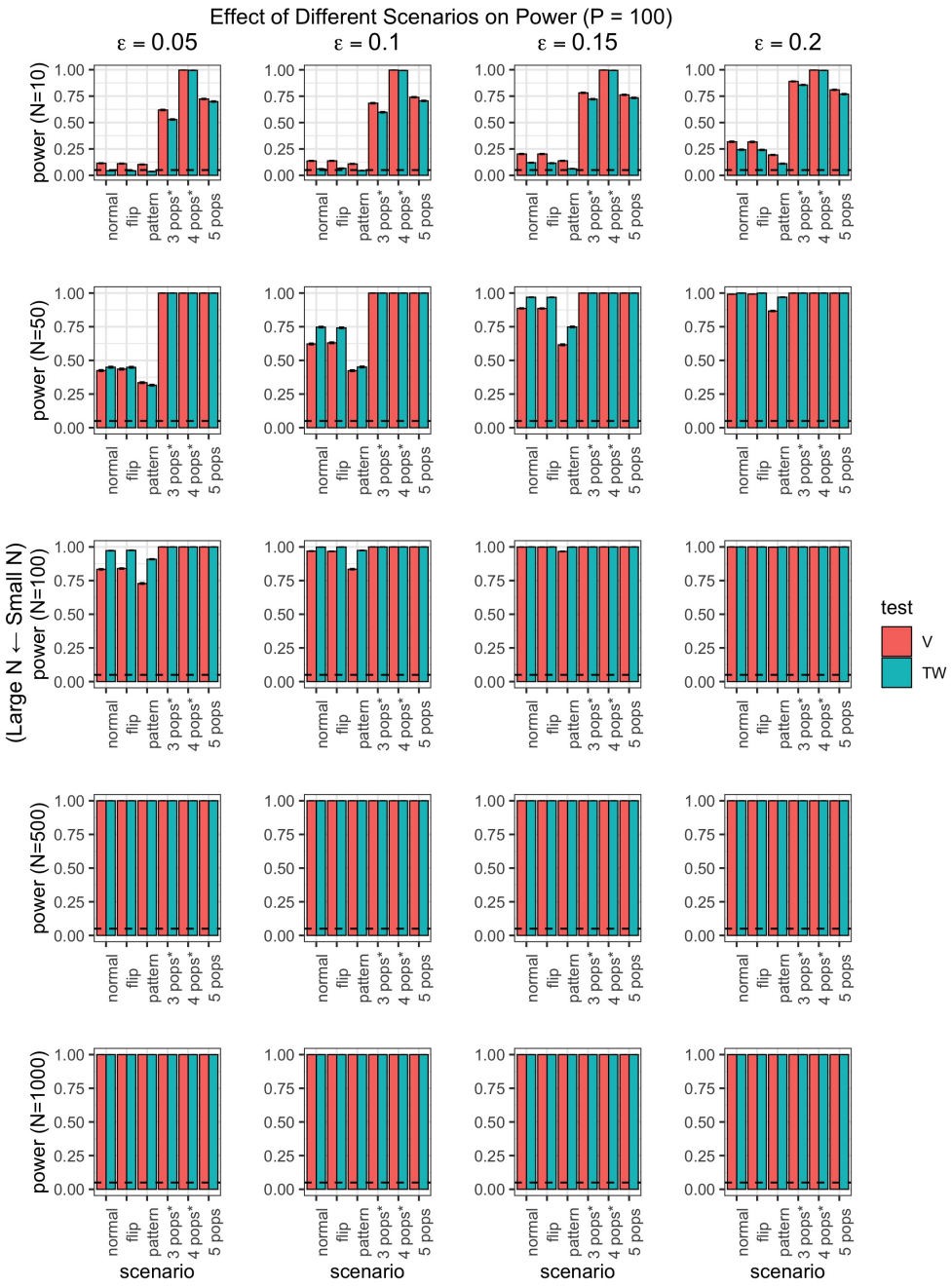

FIGURE S6. The impact of various scenarios on the statistical power of V and TW tests, with $P = 100$ features considered.

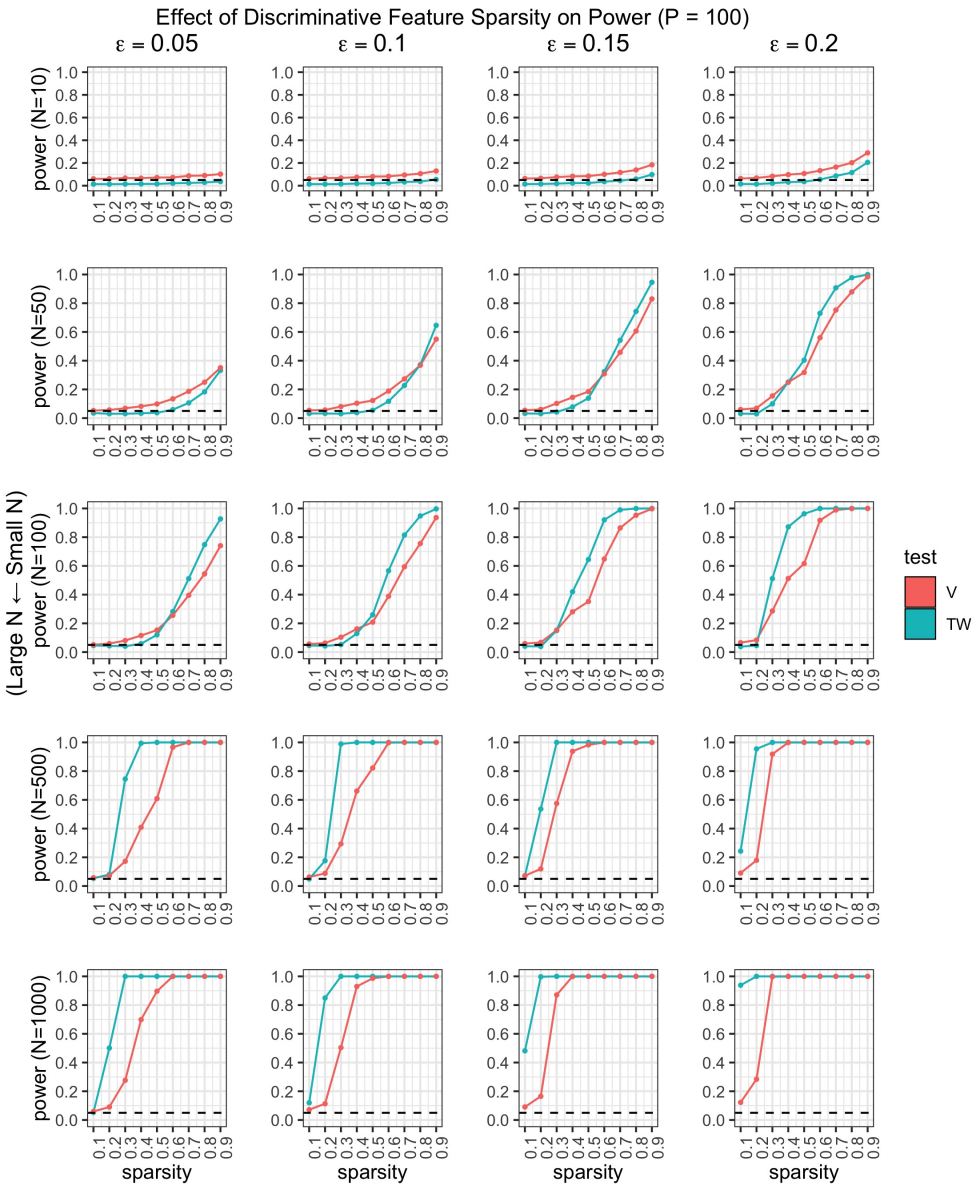

FIGURE S7. The impact of the sparsity of discerning features on the statistical power of V and TW tests, with $P = 100$ features considered.

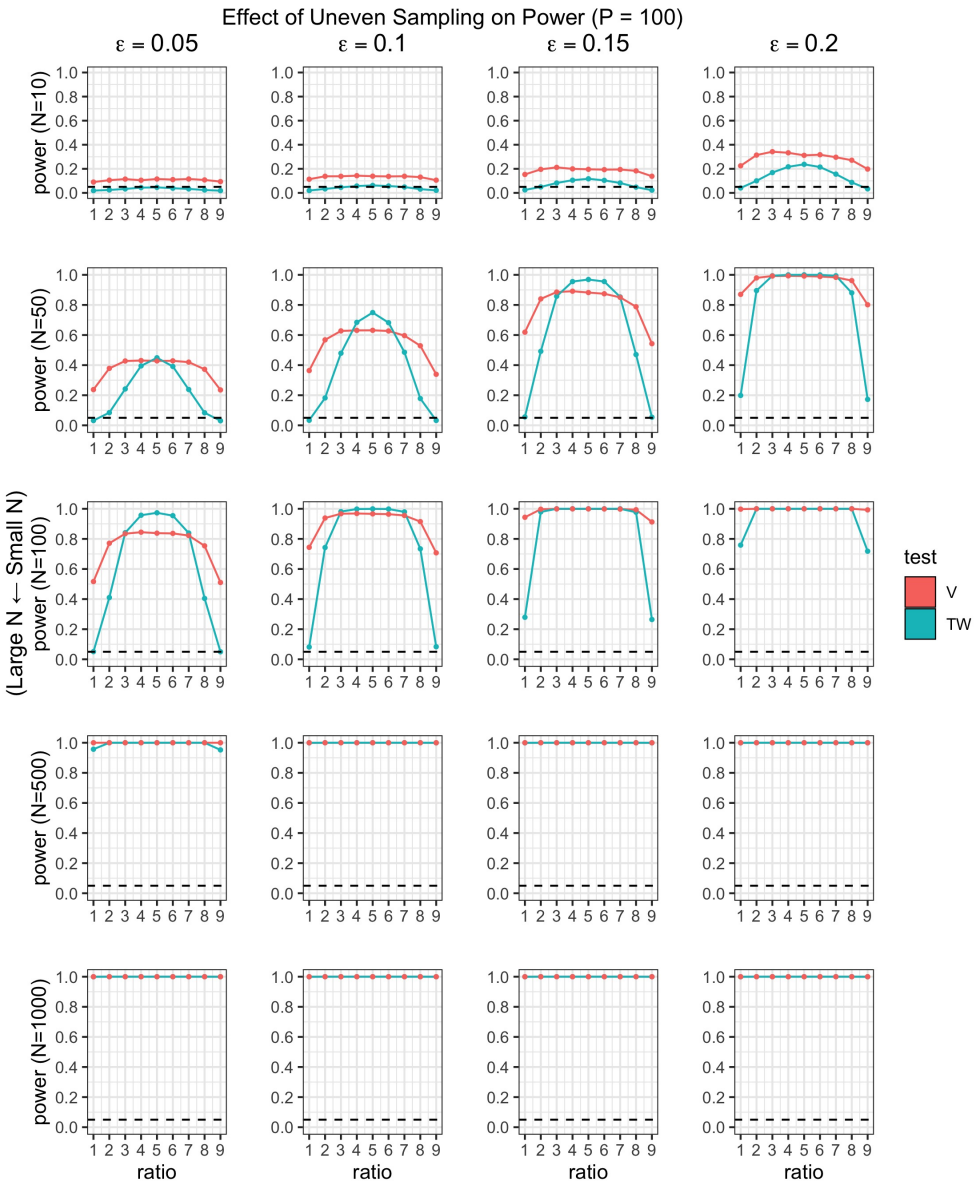

FIGURE S8. The impact of uneven sampling on the statistical power of V and TW tests, with $P = 100$ features considered.

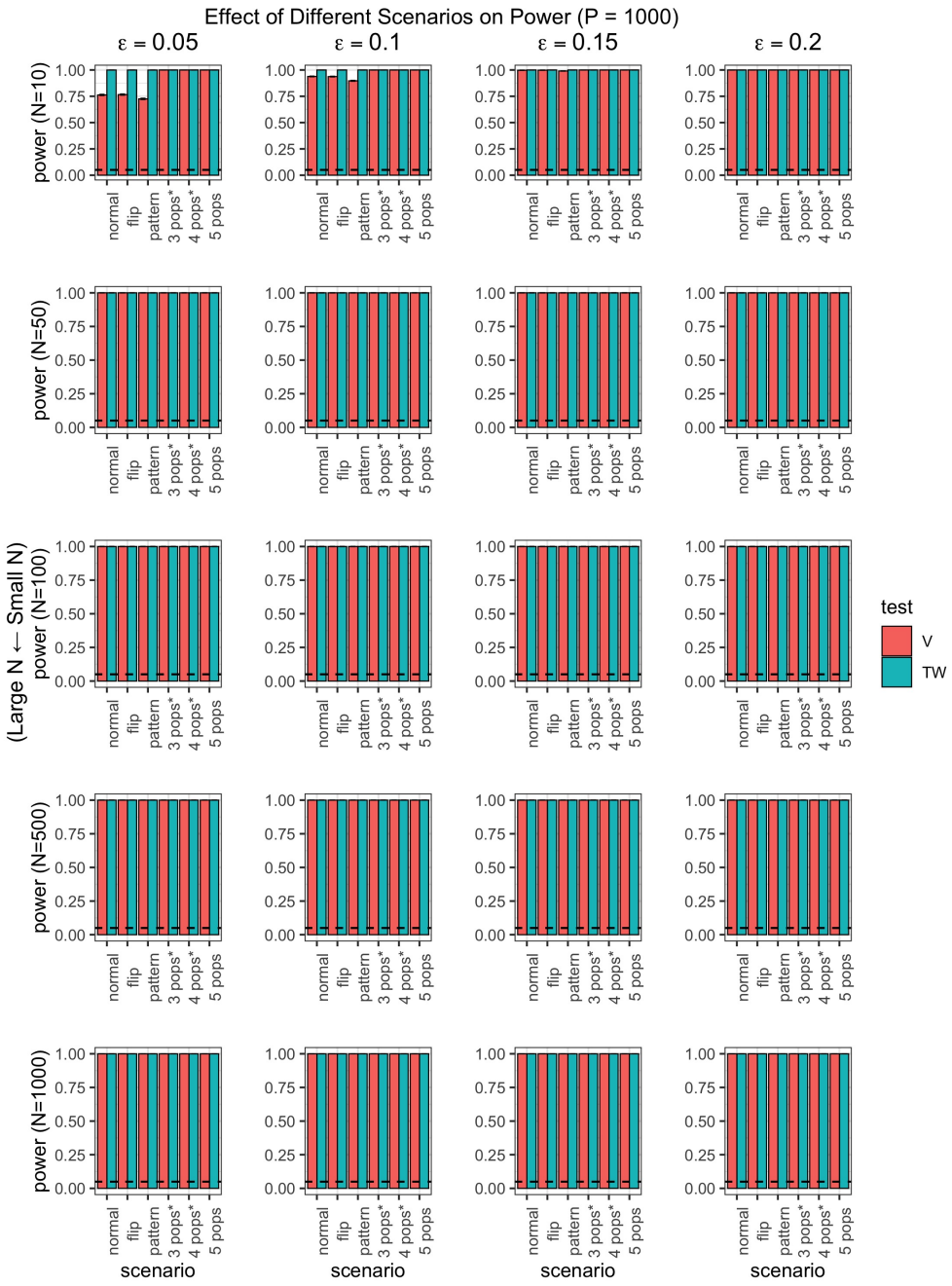

FIGURE S9. The impact of various scenarios on the statistical power of V and TW tests, with $P = 1000$ features considered.

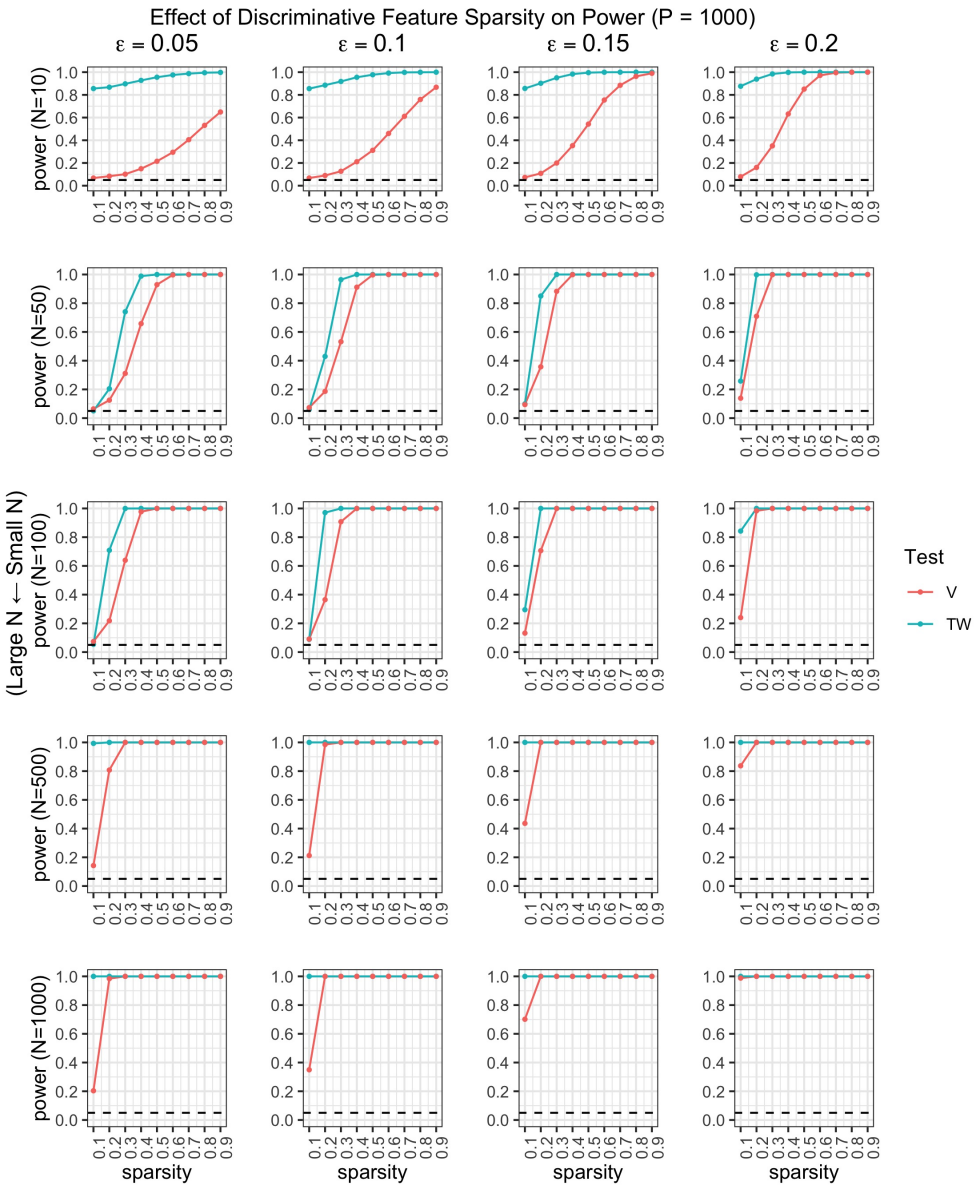

FIGURE S10. The impact of the sparsity of discerning features on the statistical power of V and TW tests, with $P = 1000$ features considered.

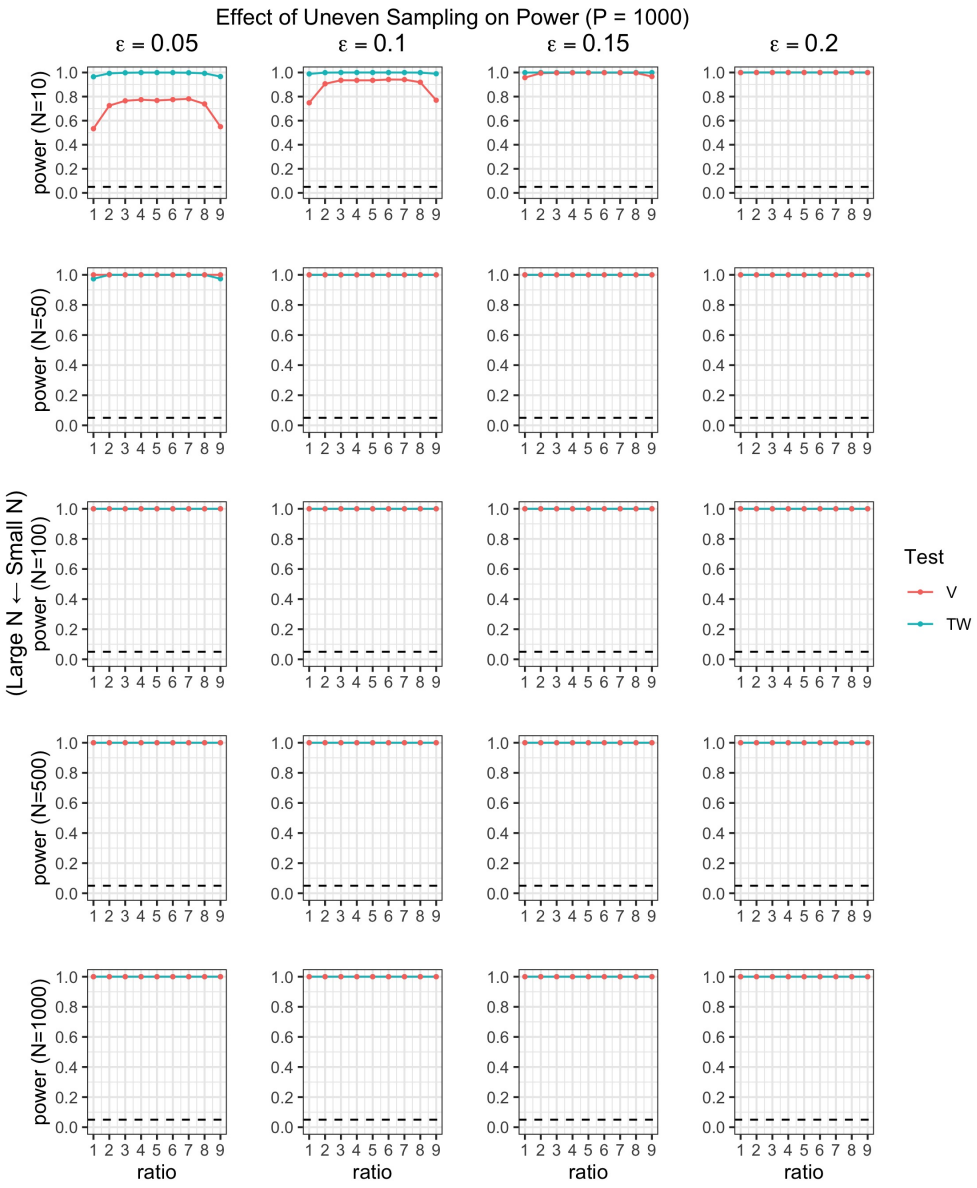

FIGURE S11. The impact of uneven sampling on the statistical power of V and TW tests, with $P = 1000$ features considered.

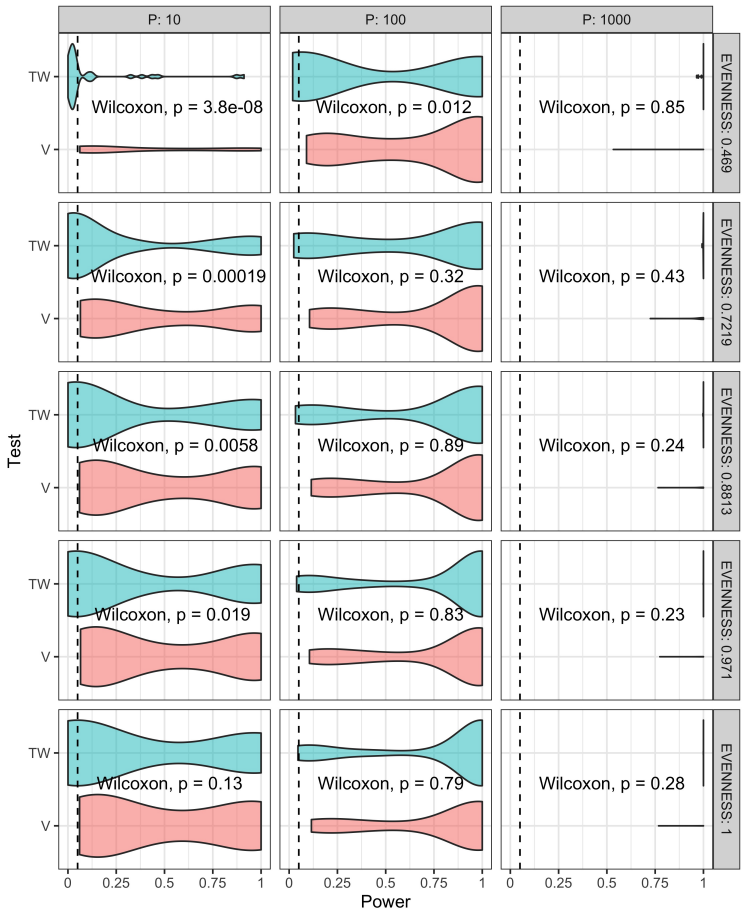

FIGURE S12. Comparison of statistical power between our test and the TW test, across all simulations involving unevenly sampled datasets. Violin plots show kernel density estimates of power. Dashed black line has x-intercept 0.05, which is the nominal significance level α at which statistical powers are computed. For each (UNEVENNESS, P) setting we report Mann-Whitney-Wilcoxon test p-values comparing the two distribution of powers.

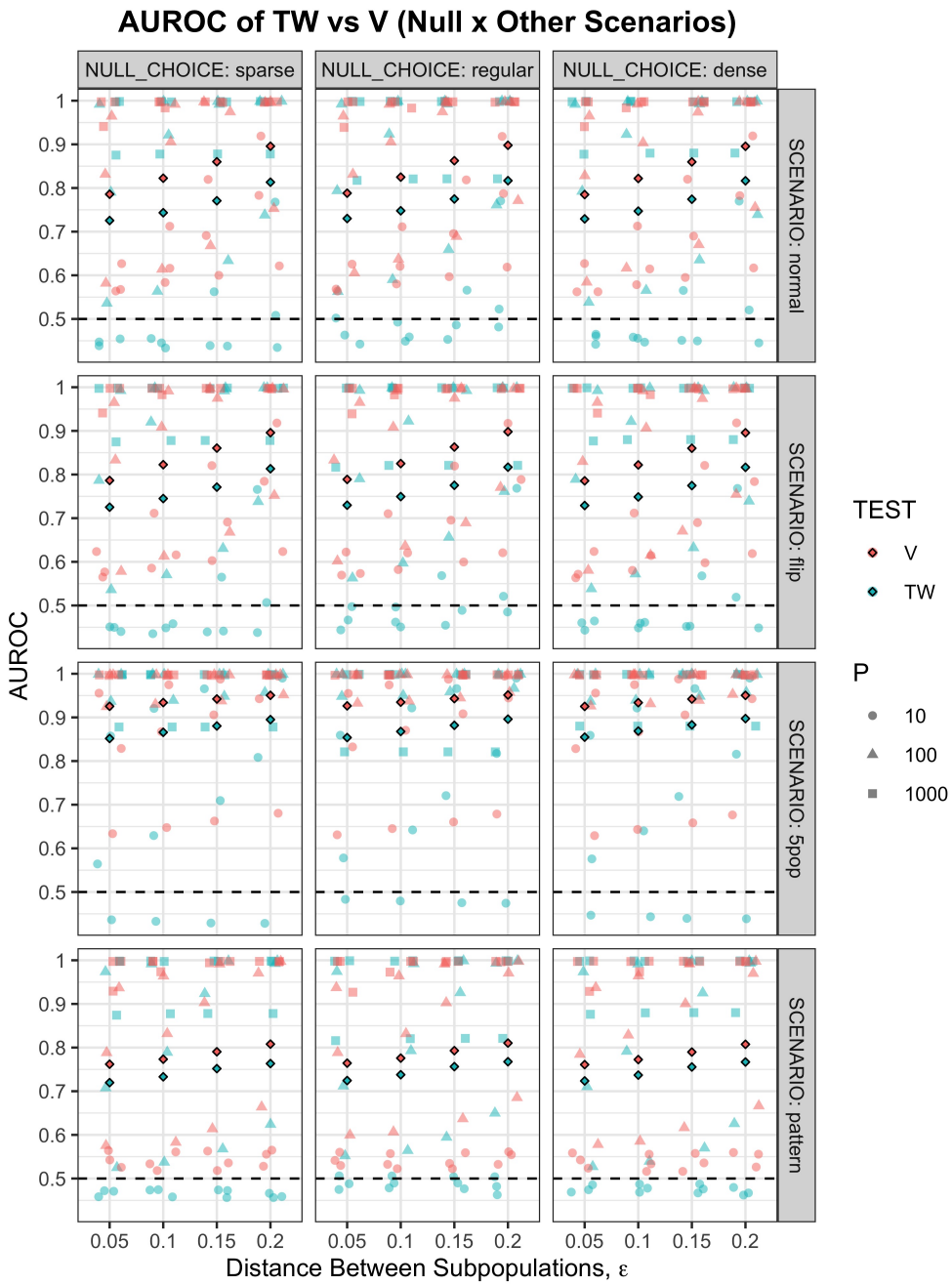

FIGURE S13. AUROC of each classifier based on pairing a null model from our FPR control simulations with a non-null model from our power estimation simulations covering Scenarios (3), (6) and (7) on top of the "normal" scenario. Each AUROC is represented by a point, coloured by the test used, and shaped according to the number of features P involved. Coloured diamonds show the average AUROC across all pairs (N, P) considered for the corresponding test in the particular configuration of (SCENARIO, ε).

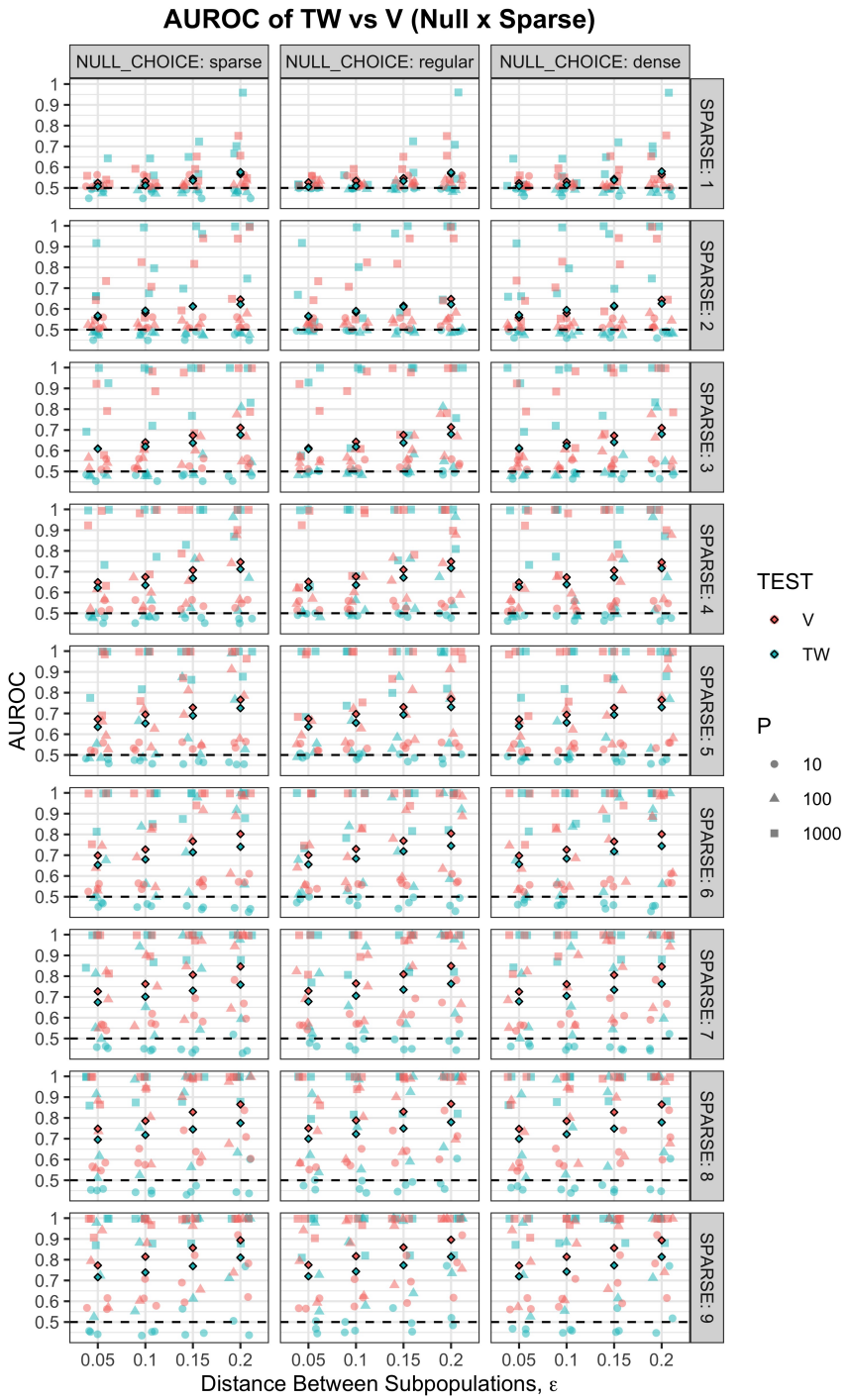

FIGURE S14. AUROC of each classifier based on pairing a null model from our FPR control simulations with a non-null model from our power estimation simulations covering Scenario 4 (sparsity of discerning features). Each AUROC is represented by a point, coloured by the test used, and shaped according to the number of features P involved. Coloured diamonds show the average AUROC across all pairs (N, P) considered for the corresponding test in the particular configuration of (SPARSE, ε).

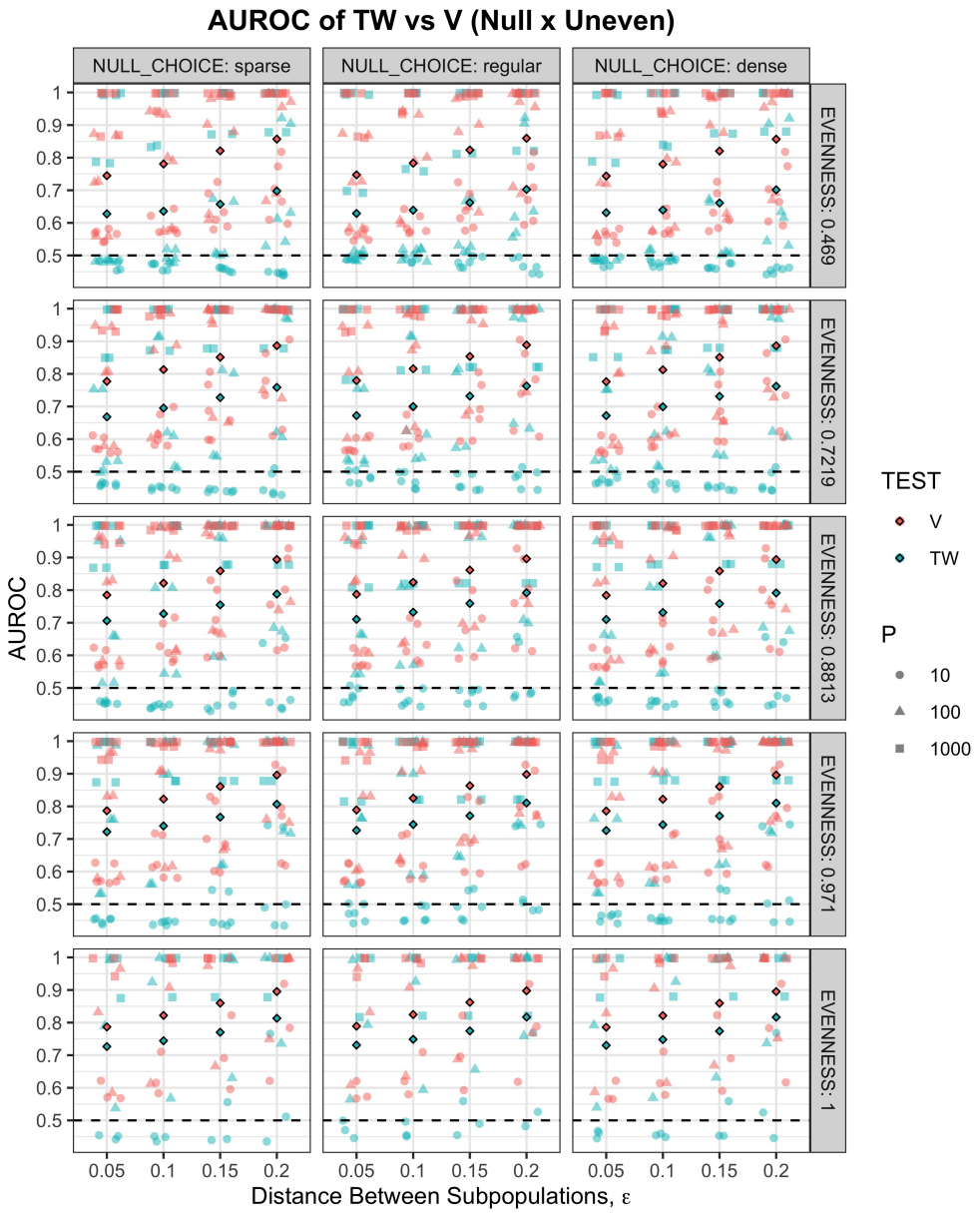

FIGURE S15. AUROC of each classifier based on pairing a null model from our FPR control simulations with a non-null model from our power estimation simulations covering Scenario 5 (uneven sampling). Each AUROC is represented by a point, coloured by the test used, and shaped according to the number of features P involved. Coloured diamonds show the average AUROC across all pairs (N, P) considered for the corresponding test in the particular configuration of (EVENNESS, ε).

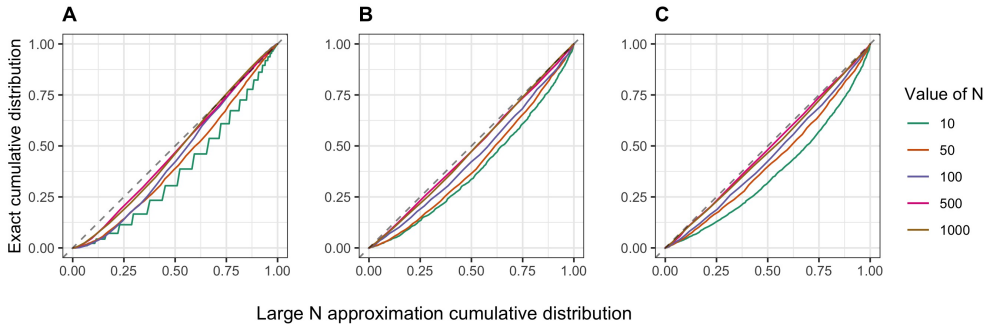

FIGURE S16. Probability-probability plots of the permutation null distribution, F_{perm}, against the parametric bootstrap. **A.** $P = 10$. **B.** $P = 100$. **C.** $P = 1000$.

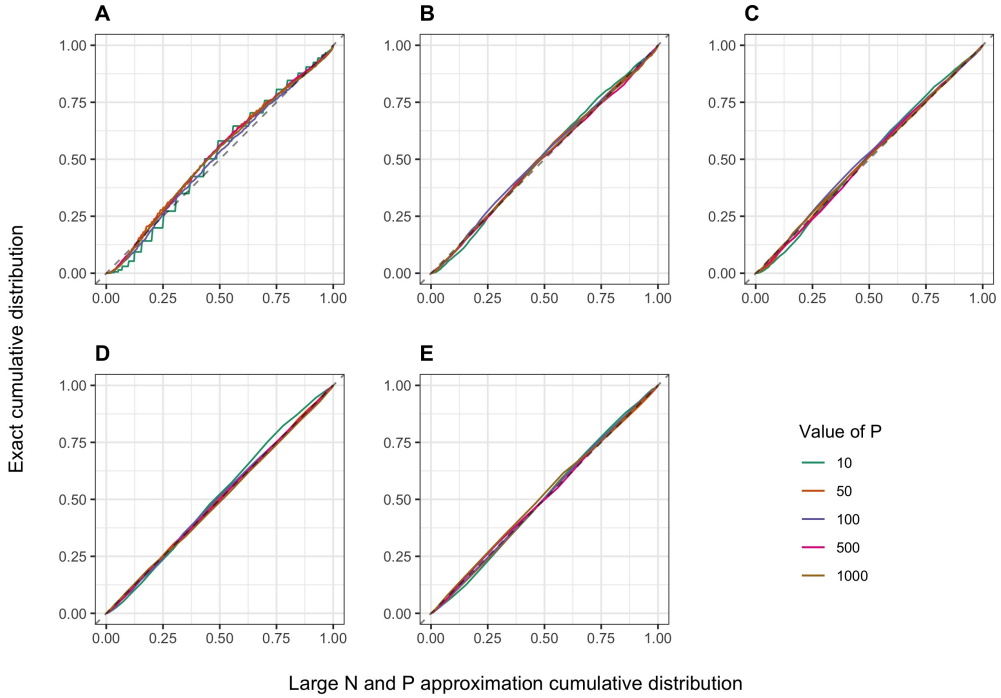

FIGURE S17. Probability-probability plots of the permutation null distribution, F_{perm}, against the large N, large P approximation. **A.** $N = 10$. **B.** $N = 50$. **C.** $N = 100$. **D.** $N = 500$. **E.** $N = 1000$.

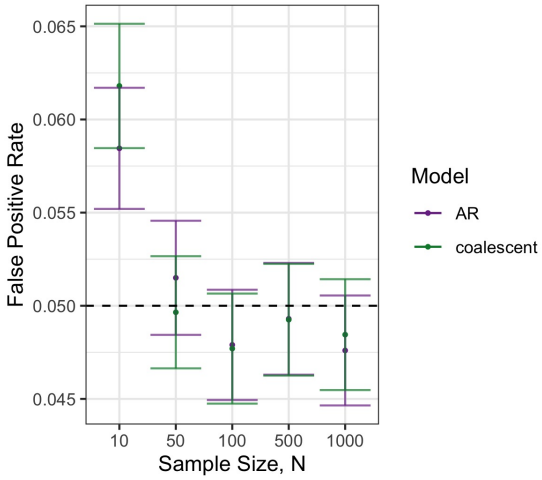

FIGURE S18. Monte Carlo estimates of FPR of the approximation to the block permutation null distribution on datasets simulated under an autoregressive model (AR) and a population-genetic coalescent model. Error bars denote 95% confidence intervals.