Supplementary Appendix to: Hypothesis testing at the extremes: fast and robust association for high-throughput data

Yi-Hui Zhou* and Fred A. Wright

Bioinformatics Research Center and Departments of Statistics and Biological Sciences
North Carolina State University
Broughton Hall
2601 Stinson Dr.
Raleigh, NC 27695

Keywords and phrases: exact testing, density approximation, permutation.

1. Appendix A

We assume exchangeability, but in many applications we expect that the elements of (say) $\mathbf{y}$ are in fact independent. However, the weaker exchangeability requirement is useful to clarify that various forms of pre-treating the data, such as normalization techniques, do not invalidate permutation testing. The same comment applies to normalization of $\mathbf{X}$, provided that $\mathbf{X}$ and $\mathbf{y}$ are normalized separately. Conditioned on the observed $\mathbf{x}$ and $\mathbf{y}$, under H_0 each of the $n!$ permuted versions of $\mathbf{y}$ is equally likely to have arisen in relation to the elements of $\mathbf{x}$, and these permutations are the population upon which inference is based. We use Π to denote a random permutation, drawn uniformly from these possibilities. The permutation in turn produces the random statistic $r(\mathbf{x}, \mathbf{y}_\Pi)$, upon which an exact p-value P_Π is based.

A primary advantage to exact testing is the distribution-free property. If X and Y have continuous densities, r_Π will assume $n!$ unique values, and a p-value (e.g., p_{left}) will be rank-uniform, assuming of the values $\{1/n!, 2/n!, ...1\}$ each with probability $1/n!$. This property ensures approximate validity, $Pr(P_\Pi \leq \alpha | data) \approx \alpha$, and that the p-value is also valid unconditionally, so $Pr(P_\Pi \leq \alpha) \approx \alpha$. Exact testing methods are familiar to many practitioners, but are often discussed in the more limited context of rank methods or discrete data. Thus we make a few additional remarks to clarify our treatment here.

- Exchangeability of *either* X or Y is sufficient. One practical consequence is that the response vector may be fixed by design, and need not be thought of as "random."
- The number of unique $r(\mathbf{x}, \mathbf{y}_\Pi)$ outcomes may be much smaller than $n!$, depending on the choice of statistic, or tied values in $\mathbf{x}$ or $\mathbf{y}$. As an ex-

treme example, consider binary $\mathbf{x}$ and $\mathbf{y}$, and Fisher's exact test of the corresponding 2×2 table. Fisher's p-value is typically computed using summations of hypergeometric outcome probabilities, thus avoiding enumeration of all $n!$ possibilities. However, complete $n!$ enumeration would produce the same p-value.

- A large literature has considered the conservativeness of exact testing (e.g., Agresti and Coull [2]), due to tied values in $\mathbf{x}$ and $\mathbf{y}$ producing tied values in $r(\mathbf{x}, \mathbf{y}_\Pi)$. Although such conservativeness is an important consideration for small sample sizes, for large sample sizes this phenomenon is of lesser importance.
- Even a slight skew in $r(\mathbf{x}, \mathbf{y}_\Pi)$ can, for extreme values of the statistic, produce a marked departure in p-values computed using permutation vs. standard parametric approaches. This matter has received little attention, which we attribute primarily to the historical focus on $\alpha = 0.05$, for which the differences between permutation and parametric analysis are often minimal.

2. Appendix B: citations and derivations for the permutationally equivalent property

Adapting the definition in Pesarin and Salmaso [10] (pg. 48) to our setting, statistics r and s are defined to be *permutationally equivalent* for $\mathbf{x}$ and $\mathbf{y}$ if for all pairs of permutations π and π', $r(\mathbf{x}, \mathbf{y}_\pi) \leq r(\mathbf{x}, \mathbf{y}_{\pi'})$ if and only if $s(\mathbf{x}, \mathbf{y}_\pi) \leq s(\mathbf{x}, \mathbf{y}_{\pi'})$.

The permutationally equivalent property for r has been variously presented for the t statistic for linear regression and equal-variance two-sample t-testing (including our two-sample problem) in Gatti et al. [5]. The one-to-one relationship between r and the contingency table linear trend statistic for two-way tables, which includes the Cochran-Armitage z-statistic and the 2×2 chi-square statistic, is detailed in Stokes and Koch [12] P.99, and also see Andres [3]. These relationships also hold unconditionally, i.e. they are not restricted to fixed $\mathbf{x}$ and $\mathbf{y}$.

The common rank-based procedures include the Spearman correlation coefficient, which is well-known to be identical to the Pearson sample correlation (i.e. r), computed on $rank(\mathbf{x})$ and $rank(\mathbf{y})$. Similarly, the Wilcoxon rank-sum statistic is equivalent to the two-sample mean difference of ranks, as the total sum of ranks is fixed. Thus the permutationally equivalent property follows directly from the definition of r, by computing the correlation on new variables $\mathbf{x}' = rank(\mathbf{x})$, $\mathbf{y}' = rank(\mathbf{y})$.

The relationships described below depend on observed moments of $\mathbf{x}$ or $\mathbf{y}$, and may not hold unconditionally for random X and Y. However, under permutation these moments remain fixed, and thus do not violate the permutationally equivalent property. Directional p-values for Fisher's exact test for 2×2 tables are determined entirely by the cell count in an arbitrarily chosen cell, after conditioning on the margins, and it is evident that the p-values are one-to-one with

the cell count. Suppose without loss of generality that $\mathbf{x}$ and $\mathbf{y}$ are represented by binary $\{0,1\}$ values, and we focus on the cell for which $\mathbf{x}=1, \mathbf{y}=1$. Then that cell count is $r = \sum x_j y_j$, proving the relation.

For the remaining "standard" statistics, such as those based on likelihood ratios, the statistic measures departure from the null in either direction, and may not be one-to-one with r^2. The one-to-one relationship with r_Π is claimed only for separate consideration of permutations with $sign(r) < 0$ and permutations with $sign(r) > 0$. For a generalized linear model, we define μ_j=$E[Y_j]$, with link function $g(\mu_j)$=$\eta_j = \beta_0 + \beta_1 x_j$. For notational convenience we define $z_{j0} = 1$ and $z_{j1} = x_j$, so $\eta_j = \sum_g \beta_g z_{jg}$, for $g \in \{0,1\}$. For any standard link function (McCullagh and Nelder [9] P30), μ_j is monotone with η_j. For the link functions of most common interest, including identity, log, logit, and probit models, the relationship is strictly increasing, and the arguments below pertain to this situation. Decreasing canonical link functions apply to exponential, gamma, and inverse Gaussian data, for which the r, $\hat{\beta}_1$ relationship in the arguments below simply needs to be reversed.

Without loss of generality, we assume that $\mathbf{x}$ and $\mathbf{y}$ have been scaled so that the correlation function $r(\mathbf{x},\mathbf{y}) = \sum_j x_j y_j$. We begin by claiming that, if a unique maximum likelihood estimate exists, then the m.l.e. $\hat{\beta}_1$ and $r(\mathbf{x},\mathbf{y})$ have the same sign. Using notation from Agresti [1], the score function is

$$\frac{\partial L_j}{\partial \beta_g} = \frac{(y_j - \mu_j) z_{jg}}{var(Y_j)} \frac{\partial \mu_j}{\partial \eta_j} \tag{2.1}$$

$$\tag{2.2}$$

with known $var(Y_j) = \frac{\partial \mu_j}{\partial \eta_j} a(\phi)$ and dispersion parameter $a(\phi)$, so that $\frac{\partial L_j}{\partial \beta_g} = (y_j - \mu_j) z_{jg} / a(\phi)$. Careful examination of the score function shows that for nonzero x_j it is always decreasing in β_1, regardless of β_0. This fact may be seen by separate consideration of the four combinations of $\{\mu_j < 0, \mu_j > 0\} \times \{x_j < 0, x_j > 0\}$. Now suppose that $r(\mathbf{x},\mathbf{y}) > 0$. The score function at $\beta_1 = 0$ is equal to $\sum_{j=1}^{n} \frac{(y_j-\mu_j)x_j}{var(Y_j)} \frac{\partial \mu_j}{\partial \eta_j} \propto r(\mathbf{x},\mathbf{y}) > 0$. Therefore, the solution to the score equation must be $\hat{\beta}_1 > 0$. If $r(\mathbf{x},\mathbf{y}) < 0$, the same reasoning implies $\hat{\beta}_1 < 0$, and so $r(\mathbf{x},\mathbf{y})$ and $\hat{\beta}_1$ must share the same sign.

For a particular permutation π, suppose there exists another permutation π' such that $r(\mathbf{x},\mathbf{y}_{\pi'}) > r(\mathbf{x},\mathbf{y}_\pi) > 0$. Using $\hat{\beta}_{g,\pi}$ to denote the maximum likelihood estimates for permutation π, by the argument above,

$$L(\widehat{\beta_0}_\pi, \widehat{\beta_1}_\pi | \mathbf{x}, \mathbf{y}_{\pi'}) > L(\widehat{\beta_0}_\pi, \widehat{\beta_1}_\pi | \mathbf{x}, \mathbf{y}_\pi).$$

By the definition of maximum likelihood and the strict inequality above, it follows that

$$L(\widehat{\beta_0}_{\pi'}, \widehat{\beta_1}_{\pi'} | \mathbf{x}, \mathbf{y}_{\pi'}) > L(\widehat{\beta_0}_\pi, \widehat{\beta_1}_\pi | \mathbf{x}, \mathbf{y}_\pi).$$

Therefore the maximum loglikelihood is a monotone increasing function of the Pearson correlation coefficient over the positive range, and the same argument implies it increases with decreasing negative correlation. The null loglikelihood

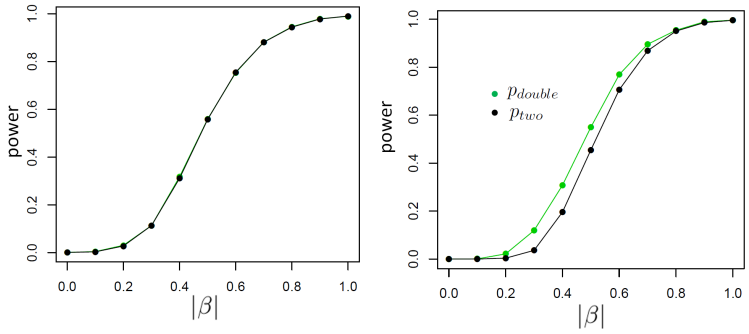

FIG 1. *Left panel: with no skew in r_Π, p_{two} and p_{double} have the same power (black is overlaid over green). Right panel: when r_Π is skewed, p_{double} has a power advantage.*

does not vary over permutations, so the same conclusion applies to the maximum loglikelihood ratio statistic. Note that this argument applies to the loglikelihood ratio, and not necessarily to other statistics. In particular, the logistic Wald statistic can have aberrant behavior for extreme departures from the null Hauck and Donner [6].

3. Appendix C: Power of p_{two} vs. p_{double}.

Figure 1 shows the power for an illustrative model, with $Y = \beta X + \epsilon_Y$, $n = 50$, and significance level $\alpha = 10^{-5}$. 1000 simulations were performed, and 10^6 permutations performed for each simulation to obtain the two types of p-values. Two scenarios are shown: (i) $X \sim N(0,1)$ and $\epsilon_Y \sim exp(1)$ (exponential with mean 1, left panel), and (ii) $X \sim exp(1)$, $\epsilon_Y \sim exp(1)$ (right panel), with the power each $|\beta|$ value averaged over the power for the corresponding positive and negative β. Skew in r_Π requires that both $\mathbf{x}$ and $\mathbf{y}$ be skewed (described in more detail below), and the random variable X is skewed only for scenario (ii). Accordingly, p_{double} and p_{two} are essentially identical in the left panel, while in the right panel, skew in r_Π provides an advantage to p_{double}.

The MCC method approximates the distribution of r_Π using a smooth density $f(r)$ (cdf F), and denote $r_\eta = F^{-1}(\eta)$, $\eta \in (0,1)$. For fixed $\alpha < 0.5$, define r_+ as the value such that $\int_{-\infty}^{-r_+} f(r)dr + \int_{r_+}^{\infty} f(r)dr = \alpha$, and we will denote $r_- = -r_+$. We assume $r_- < r_{\alpha/2}$, which further implies $r_+ < r_{1-\alpha/2}$. This assumption is essentially without loss of generality, as all of the following arguments can be applied by reversing corresponding inequalities, and in the instance of equality the two types of p-value will have equal power. It follows that $F(r_{\alpha/2}) - F(r_-) = F(r_{1-\alpha/2}) - F(r_+)$. The upper panel of Supplementary Figure 2 illustrates these critical values for a right-skewed null distribution and $\alpha = 0.05$. The red boundaries r_-, r_+ are equidistant from the mean, corresponding to rejection thresholds for p_{two}, while the green boundaries $r_{\alpha/2}, r_{1-\alpha/2}$ correspond to p_{double}.

We approximate the alternative density as taking the form $g(r) = \frac{1}{2}f(r-\delta)+\frac{1}{2}f(r+\delta)$ over the appropriate support, where δ determines the power. We define the following *ordering conditions* for fixed δ:

$$F(r_{\alpha/2}-\delta)-F(r_- -\delta) > F(r_{1-\alpha/2}-\delta)-F(r_+-\delta), \quad (3.1)$$
$$F(r_{\alpha/2}+\delta)-F(r_- +\delta) > F(r_{1-\alpha/2}+\delta)-F(r_+ +\delta). \quad (3.2)$$

Given the ordering conditions, it is simple to show that $G(r_{\alpha/2})-G(r_-) > G(r_{1-\alpha/2})-G(r_+)$. The left-hand side corresponds to region 1 in the lower panel of Supplementary Figure 1, which is larger than the right-hand side (region 2). Region 1 is included in the rejection region of p_{double}, while region 2 is included in the rejection region of p_{double}, and the remaining shaded regions are common to both rejection rules. Thus, the ordering conditions imply that the power of p_{double} is greater than that of p_{two}. Applying all of the above arguments when $r_- > r_{\alpha/2}$, and reversing the inequalities, again showing that the power of p_{double} is greater than that of p_{two}.

Ordering conditions for the beta approximation

Whether the ordering conditions hold depends on the precise form of f and on δ. Simulations such as shown in the main text show that for "typical" skewed densities and α values used in practice, the ordering conditions hold across a wide range of δ, and p_{double} clearly has greater power than p_{two}. Using the approximating beta density proposed here for MCC, for r_Π with beta parameters $\alpha < \beta$ (right-skewed), we can obtain a result for extreme α and δ near zero.

For our purposes it is more convenient to work directly with the original beta random variable, rather than the rescaled version which is used to approximate r_Π. We will denote the random variable B, with realized value b, and null pdf and cdf $h(b)$ and $H(b)$. Recall that $r = \frac{b-E(B)}{\sqrt{var(B)(n-1)}}$, $b = r\sqrt{var(B)(n-1)}+E(B)$, and we will use b_-, etc., to refer to the values correspondingly mapped from r_-, etc. Here $E(B) = \frac{\alpha_1}{\alpha_1+\alpha_2} < \frac{1}{2}$. p_{two} experiences an asymmetry for sufficiently small α, as the right tail extends further from the mean than the left tail. Specifically, when $\alpha \in (0, 1-H\big(2E(B)\big))$, $b_- = 0 < b_{\alpha/2}$. Using continuity and unimodal properties of the beta density implies that $h(b_-) < h(b_{\alpha/2})$ for $\alpha \in R$, where $R = (0, \alpha')$ for some positive $\alpha' \geq 1 - H\big(2E(B)\big)$. A first order Taylor expansion is $H(b+\delta) = H(b)+\delta h(b)+o(\delta)$, and we examine the second ordering condition applied to the beta random variable, by computing

$$\begin{aligned} &H(b_{\alpha/2}+\delta)-H(b_-+\delta)-\big\{H(b_{1-\alpha/2}+\delta)-H(b_++\delta)\big\} \\ &= \big\{H(b_{\alpha/2})-H(b_-)-(H(b_{1-\alpha/2})-H(b_+))\big\} \\ &+\delta\big\{h(b_{\alpha/2})-h(b_-)-(h(b_{1-\alpha/2})-h(b_+))\big\}+o(\delta) \\ &= 0+\delta\big\{h(b_{\alpha/2})-h(b_-)-(h(b_{1-\alpha/2})-h(b_+))\big\}+o(\delta). \end{aligned} \quad (1)$$

We have $h(b_{\alpha/2}) > h(b_-)$ and $h(b_{1-\alpha/2}) < h(b_+)$, and thus the term in braces in (1) is positive, implying that the second ordering condition holds for sufficiently small positive δ. Finally, the same argument can be applied if $\alpha_1 > \alpha_2$ (left-skewed), also showing a local power advantage for positive δ near zero.

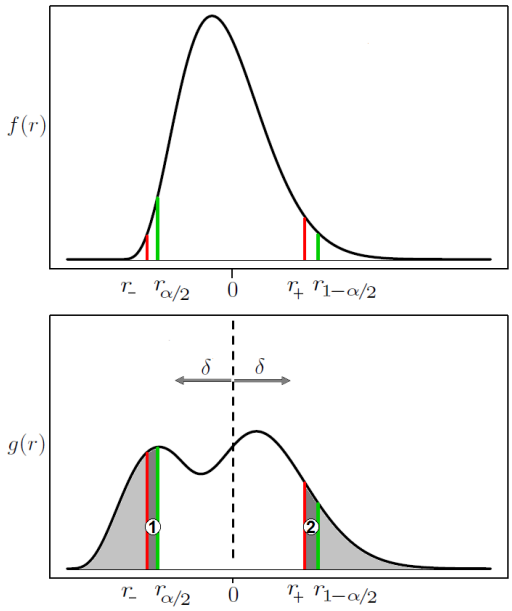

FIG 2. *Illustration of the beta density approximation with* $\alpha_1 = 5, \alpha_2 = 20$. *The region 1 rejected by* p_{double} *is larger than the region 2 rejected by* p_{two}.

4. Appendix D: Kurtosis of r_{Π}

We scale both $\mathbf{x}$ and $\mathbf{y}$ so that $\sum_{j=1}^n x_j = 0$, $\sum_{j=1}^n x_j^2 = 1$ and $\sum_{j=1}^n y_j = 0$, $\sum_{j=1}^n y_j^2 = 1$.

$$\begin{aligned} r &= \frac{(\sum x_j y_j)/(n-1) - \overline{\mathbf{xy}}}{s_x s_y} \\ &= \frac{\frac{\sum x_j y_j}{n-1}}{\sqrt{\frac{\sum_j x_j^2}{n-1}\frac{\sum_j y_j^2}{n-1}}} \\ &= \sum_j x_j y_j \end{aligned} \tag{4.1}$$

We have

$$\begin{aligned} (\sum x_i y_i)^4 &= \sum x_i^4 y_i^4 + 4\sum_i \sum_{j\neq i} x_i^3 x_j y_i^3 y_j + 6\sum_i \sum_{j\neq i} x_i^2 x_j^2 y_i^2 y_j^2 \\ &+ 12\sum_i \sum_{j\neq i} \sum_{l \not\subseteq (i,j)} x_i^2 x_j x_l y_i^2 y_j y_l \\ &+ \sum_i \sum_{j\neq i} \sum_{k \not\subseteq (i,j)} \sum_{m \not\subseteq (i,j,k)} x_i x_j x_k x_m y_i y_j y_k y_m \end{aligned} \tag{4.2}$$

We have the kurtosis of $\mathbf{x}$ (denoted k_x and treating the vector $\mathbf{x}$ as a "population"), $k_x = \frac{\frac{\sum_j x_j^4}{n}}{(\frac{\sum_j x_j^2}{n})^2} - 3 = n\sum_j x_j^4 - 3$, so we have $\sum_j x_j^4 = \frac{k_x+3}{n}$.

$$\begin{aligned}
E[x_j x_k^3] &= \frac{1}{n(n-1)}\sum x_j(\eta - x_j^3) \\
&= -\frac{1}{n(n-1)}\sum x_j^4 \\
&= -\frac{1}{n^2(n-1)}(k_x+3) \\
E[x_j^3 x_k] &= \sum_j\sum_{k\neq j} x_j^3 x_k p(x_j x_k) \\
&= \frac{1}{n(n-1)}\sum_j x_j^3(-x_j) \\
&= -\frac{1}{n(n-1)}\sum_j x_j^4 \\
&= -\frac{1}{n^2(n-1)}(k_x+3) \\
E[x_j^2 x_k^2] &= \sum_j\sum_{k\neq j} x_j^2 x_k^2 p(x_j x_k) \\
&= \frac{1}{n-1}\sum_j x_j^2(1-x_j^2) \\
&= \frac{1}{n-1} - \frac{1}{n^2(n-1)}(k_x+3) \\
E[x_j^2 x_k x_l] &= \frac{1}{n(n-1)(n-2)}\sum_j\sum_{k\neq j}\sum_{l\not\subset(j,k)} x_j^2 x_k x_l \\
&= \frac{1}{n(n-1)(n-2)}\sum_j x_j^2\sum_{k\neq j} x_k \sum_{l\not\subset(j,k)}(-x_j - x_k) \\
&= -\frac{1}{(n-1)(n-2)}\sum_j x_j^3\sum_{k\neq j} x_k - \frac{1}{n(n-1)(n-2)}\sum_j x_j^2\sum_{k\neq j} x_k^2 \\
&= \frac{2}{n^2(n-1)(n-2)}(k_x+3) - \frac{1}{n(n-1)(n-2)} \qquad (4.3)
\end{aligned}$$

$$
\begin{aligned}
E[x_j x_k x_l x_m] &= \frac{1}{n(n-1)(n-2)(n-3)} \sum_j \sum_{k \neq j} \sum_{l \not\subset (j,k)} \sum_{m \not\subset (l,k,j)} x_j x_k x_l x_m \\
&= \frac{1}{n(n-1)(n-2)(n-3)} \sum_j \sum_{k \neq j} \sum_{l \not\subset (j,k)} \sum_{m \not\subset (l,k,j)} x_j x_k x_l - x_j - x_k - x_l \\
&= \frac{1}{n(n-1)(n-2)(n-3)} \sum_j \sum_{k \neq j} \sum_{l \not\subset (j,k)} \{-x_j^2 x_k x_l - x_j x_k^2 x_l - x_j x_k x_l^2\} \\
&= \frac{1}{n(n-1)(n-2)(n-3)} \{-3(\frac{2}{n}(k_x+3)-1)\} \\
&= -\frac{6}{n(n-1)(n-2)(n-3)}(k_x+3) + \frac{1}{n(n-1)(n-2)(n-3)} \qquad (4.4)
\end{aligned}
$$

$$
\begin{aligned}
E[(\sum x_i y_i)^4] &= E[\sum x_i^4 y_i^4] + 4E[\sum_i \sum_{j \neq i} x_i^3 x_j y_i^3 y_j] + 6E[\sum_i \sum_{j \neq i} x_i^2 x_j^2 y_i^2 y_j^2] \\
&+ 12E[\sum_i \sum_{j \neq i} \sum_{l \not\subset (i,j)} x_i^2 x_j x_l y_i^2 y_j y_l] + E[\sum_i \sum_{j \neq i} \sum_{k \not\subset (i,j)} \sum_{m \not\subset (i,j,k)} x_i x_j x_k x_m y_i y_j y_k y_m] \\
&= nE[x_i^4]E[y_i^4] + 4n(n-1)E[x_i^3 x_j]E[y_i^3 y_j] + 6n(n-1)E[x_i^2 x_j^2]E[y_i^2 y_j^2] \\
&+ 12n(n-1)(n-2)E[x_i^2 x_j x_l]E[y_i^2 y_j y_l] \\
&+ n(n-1)(n-2)(n-3)E[x_i x_j x_k x_m]E[y_i y_j y_k y_m]
\end{aligned}
$$

The kurtosis of r_Π is proportional to $E[(\sum x_i y_i)^4]$, and therefore can be written in terms of the linear combinations of the kurtosis of $\mathbf{x}$ (k_x), kurtosis of $\mathbf{y}$ (k_y), and $k_x k_y$.

5. Appendix E: The beta parameters α and β, in terms of skewness and kurtosis of r_Π

We use k to denote the kurtosis of r_Π, and s to denote its skewness. For parameters α and β of the beta density can be expressed analytically in terms of the kurtosis and skewness of the distribution. The skewness and kurtosis of a beta density are given in Chapter 21 of Johnson et al. [7]. Solving for α and β, the

inverse relationship is

$$\begin{aligned}
\alpha &= (3k + 36s\frac{-1}{-k^2s^2 + 32k^2 - 84ks^2 + 96k + 36s^4 - 180s^2}^{1/2} \\
&- 18s^3\frac{-1}{-k^2s^2 + 32k^2 - 84ks^2 + 96k + 36s^4 - 180s^2}^{1/2} - 3s^2 \\
&+ 3k^2s\frac{-1}{-k^2s^2 + 32k^2 - 84ks^2 + 96k + 36s^4 - 180s^2}^{1/2} \\
&- 3ks^3\frac{-1}{-k^2s^2 + 32k^2 - 84ks^2 + 96k + 36s^4 - 180s^2}^{1/2} \\
&+ 24ks\frac{-1}{-k^2s^2 + 32k^2 - 84ks^2 + 96k + 36s^4 - 180s^2}^{1/2} + 6)/(2k - 3s^2) \\
&- \frac{-6s^2 + 6k + 12}{2k - 3s^2}
\end{aligned} \tag{5.1}$$

If the above solution provides $\alpha < 0$, then we instead use

$$\begin{aligned}
\alpha &= (3k - 36s\frac{-1}{-k^2s^2 + 32k^2 - 84ks^2 + 96k + 36s^4 - 180s^2}^{1/2} \\
&+ 18s^3\frac{-1}{-k^2s^2 + 32k^2 - 84ks^2 + 96k + 36s^4 - 180s^2}^{1/2} - 3s^2 \\
&- 3k^2s\frac{-1}{-k^2s^2 + 32k^2 - 84ks^2 + 96k + 36s^4 - 180s^2}^{1/2} \\
&+ 3ks^3\frac{-1}{-k^2s^2 + 32k^2 - 84ks^2 + 96k + 36s^4 - 180s^2}^{1/2} \\
&- 24ks\frac{-1}{-k^2s^2 + 32k^2 - 84ks^2 + 96k + 36s^4 - 180s^2}^{1/2} + 6)/(2k - 3s^2) \\
&- \frac{-6s^2 + 6k + 12}{2k - 3s^2}
\end{aligned} \tag{5.2}$$

Similarly,

$$\begin{aligned}\beta &= -(3k + 36s\frac{-1}{-k^2s^2 + 32k^2 - 84ks^2 + 96k + 36s^4 - 180s^2}^{1/2} \\ &- 18s^3\frac{-1}{-k^2s^2 + 32k^2 - 84ks^2 + 96k + 36s^4 - 180s^2}^{1/2} \\ &- 3s^2 + 3k^2s\frac{-1}{-k^2s^2 + 32k^2 - 84ks^2 + 96k + 36s^4 - 180s^2}^{1/2} \\ &- 3ks^3\frac{-1}{-k^2s^2 + 32k^2 - 84ks^2 + 96k + 36s^4 - 180s^2}^{1/2} \\ &+ 24ks\frac{-1}{-k^2s^2 + 32k^2 - 84ks^2 + 96k + 36s^4 - 180s^2}^{1/2} + 6)/(2k - 3s^2),\end{aligned}$$

and if the above solution is $\beta < 0$, then

$$\begin{aligned}\beta &= -(3k - 36s\frac{-1}{-k^2s^2 + 32k^2 - 84ks^2 + 96k + 36s^4 - 180s^2}^{1/2} \\ &+ 18s^3\frac{-1}{-k^2s^2 + 32k^2 - 84ks^2 + 96k + 36s^4 - 180s^2}^{1/2} \\ &- 3s^2 - 3k^2s\frac{-1}{-k^2s^2 + 32k^2 - 84ks^2 + 96k + 36s^4 - 180s^2}^{1/2} \\ &+ 3ks^3\frac{-1}{-k^2s^2 + 32k^2 - 84ks^2 + 96k + 36s^4 - 180s^2}^{1/2} \\ &- 24ks\frac{-1}{-k^2s^2 + 32k^2 - 84ks^2 + 96k + 36s^4 - 180s^2}^{1/2} + 6)/(2k - 3s^2)\end{aligned}$$

If both $\mathbf{x}$ and $\mathbf{y}$ are highly skewed, occasionally there are no real solutions for α and β, and we instead use a shifted gamma density as an approximation Zhou et al. [13], although in our experience this is a rare occurrence. For example, the approach was not needed for any of 7129 genes in the example in the next section. In this instance, the parameters of a standard gamma density are chosen to match k and s, and then the entire density is shifted to have a mean of zero. The shifted gamma density is then used as an approximation to r_Π, providing left- and right-tailed p-values as above.

6. Appendix F: Small sample perfomance with ties

We consider the case with 2×2 tables as producing the highest number of tied r_Π values, because X and Y are both binary. The Fisher exact p-value can be computed directly and depends on one cell in the table, so exhaustive "permutation" is feasible. Supplementary Figures 3-5 shows the comparisons of MCC vs. p-values from Fisher's exact test. There is reasonable agreement

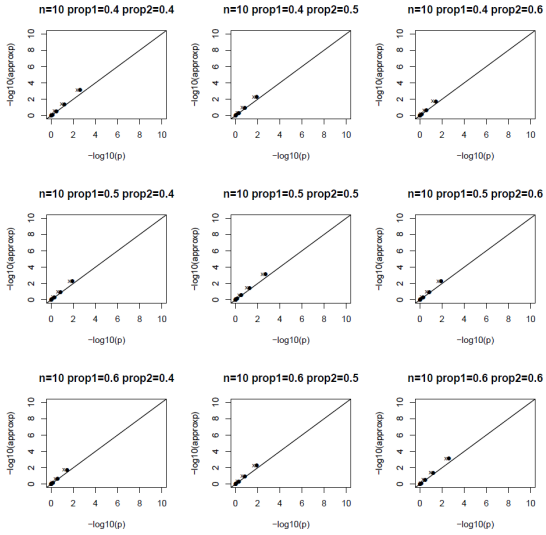

FIG 3. *Results for Fisher's exact comparisons, sample size=10. In each plot, n is the total sample size, and each* 2×2 *table represents a comparison of binary* $X \in \{0, 1\}$ *to* $Y \in \{0, 1\}$. *prop*1 *is the proportion with* $x = 1$, *and prop*2 *the proportion with* $y = 1$. *The x-axis shows the exact right-tail p-value (black dots are the mid p-value, grey x symbols are the conservative exact p-values based on the* 2×2 *odds ratio,* $P_{H_0}(OR \geq OR_{obs})$.

even for $n = 20$, but even for $n = 30$ and skewed X and Y (unbalanced table margins), the extreme MCC p-values can differ noticeably anticonservatively from the true values. We emphasize that 2×2 tables represent an extreme instance for which MCC would not actually be needed.

Another example involving ties is shown using the Wilcoxon rank sum test (X ranks, Y binary). Here again we compute exact p-values by enumerating all possible combinations for the two-group comparisons of n_1 vs. n_2. Supplementary Figure 6 shows the results, up to $n_1 = n_2 = 12$, which has 2.7 million possible combinations, not considering tied observations. Here we chose to introduce tied X values all with identical rank=1, which we reasoned would present a worse-case scenario (as an extreme value), and under combination will produce ties in r_Π due to a large number of equivalent assignments of the tiesd values to groups 1 and 2. For total n as large as 24, we see some departure of MCC from the exact p-value for p-values less than 10^{-4}, and for such small sample sizes recommend using MCC only in settings in which p-values of that magnitude are adequate. The role of ties is mainly to prodice a slight "waviness" in the p-values, due to multiple modes in the tail r_Π distribution.

In summary, MCC in many ways performs adequately for small sample sizes in the range 20-30, as long as p-values in the range $10^{-4} - 10^{-5}$ are adequate. Otherwise, larger sample sizes may be needed.

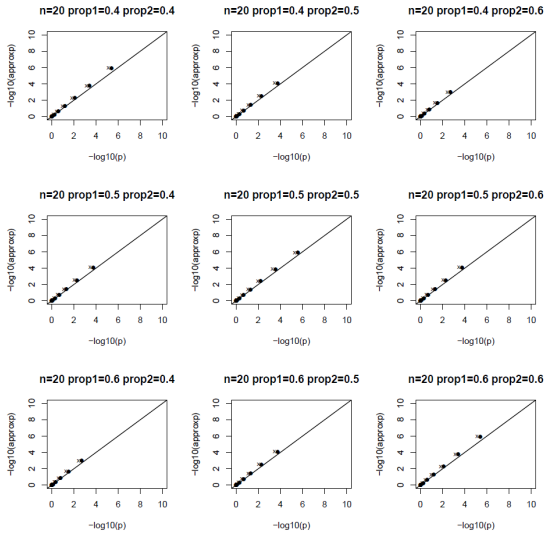

FIG 4. *Results for Fisher's exact comparisons, sample size=20. In each plot, n is the total sample size, and each* 2×2 *table represents a comparison of binary* $X \in \{0, 1\}$ *to* $Y \in \{0, 1\}$*. prop*1 *is the proportion with* $x = 1$*, and prop*2 *the proportion with* $y = 1$*. The x-axis shows the exact right-tail p-value (black dots are the mid p-value, grey x symbols are the conservative exact p-values based on the* 2×2 *odds ratio,* $P_{H_0}(OR \geq OR_{obs})$*.*

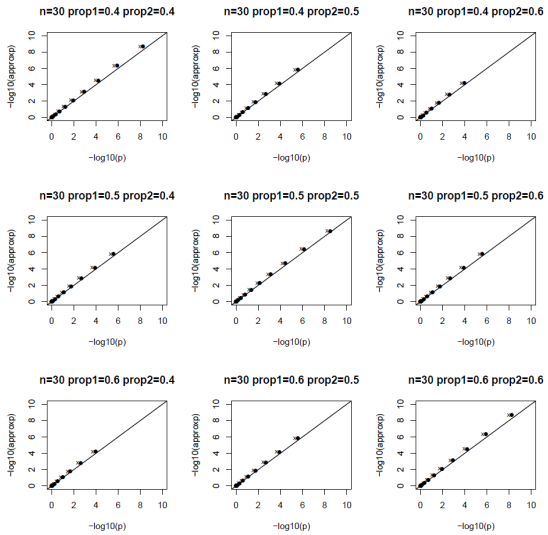

FIG 5. *Results for Fisher's exact comparisons, sample size=30. In each plot, n is the total sample size, and each* 2×2 *table represents a comparison of binary* $X \in \{0, 1\}$ *to* $Y \in \{0, 1\}$*. prop*1 *is the proportion with* $x = 1$*, and prop*2 *the proportion with* $y = 1$*. The x-axis shows the exact right-tail p-value (black dots are the mid p-value, grey x symbols are the conservative exact p-values based on the* 2×2 *odds ratio,* $P_{H_0}(OR \geq OR_{obs})$*.*

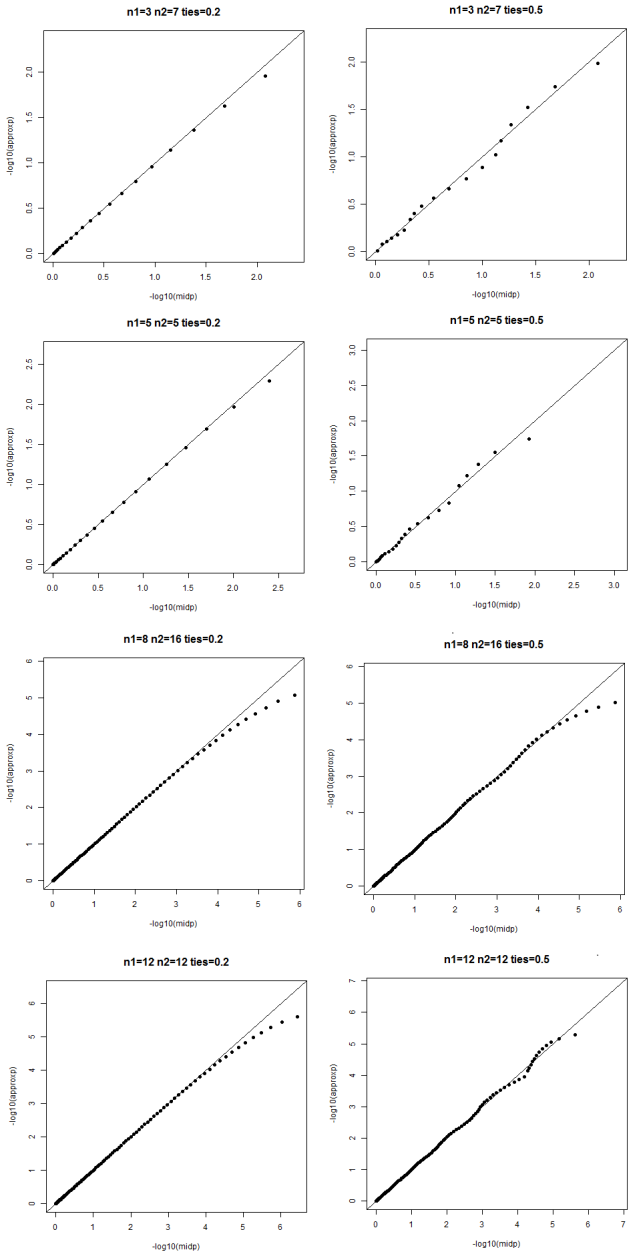

FIG 6. *Results for Wilcoxon rank-sum comparisons, small sample sizes and a high proportion of ties raning from 0.2 to 0.5. In each plot, n_1 and n_2 represent the number of samples in group 1 and group 2, and "ties" is the proportion of all samples with rank=1. The x-axis provides the exact mid p-values (although here the ordinary eact p-values are nearly identical), and the y-axis shows the MCC approximating p-value. The fit is accurate to p-values near 10^{-4}, and often lower, for these small samples regardless of the tied proportion, and further improves for larger sample sizes, as shown in the simulation results in the main manuscript.*

7. Appendix G: Saddlepoint comparisons

A well-studied example facilitates comparison to competing methods of permutation approximation. The two-sample dataset ($n = 16$) analyzing the effect of drugs on pain was described in Lehmann (Lehmann [8]) was further analyzed by Robinson [11] and Booth and Butler [4] to demonstrate saddlepoint methods as an alternative to permutation. The published results include tests and confidence intervals using exact methods (originally based on 100,000 permutations, improved here to 10^8 permutations), saddlepoint and Edgeworth expansions, to which we add our MCC results (Supplementary Table 1). The MCC approach is highly accurate, performing as well or better than the competing approximations, and is easier to implement.

TABLE 1
Hours of pain relief due to drugs, treatment A vs. treatment B. Data originally from Lehmann (1975, p. 37)

A: 6.8,3.1,5.8,4.5,3.3,4.7,4.2,4.9;
B: 4.4,2.5,2.8,2.1,6.6,0.0,4.8,2.3.

pval/Sign. Level	$Exact^a$	$Exact^b$	$Skovgaard1^c$	
	0.102	0.101	0.097	0.089
0.991	$(-1, 3.97)$	$(-1.03, 3.98)$		$(-0.96, 3.88)$
0.975	$(0.62, 3.53)$	$(-0.62, 3.57)$		$(-0.57, 3.5)$
0.95	$(-0.3, 3.26)$	$(-0.32, 3.26)$	$(-0.30, 3.26)$	$(-0.27, 3.22)$

pval/Sign. Level	$Robinson2^a$	$Edgeworth^a$	MCC	MCC_1
	0.101	0.098	0.101	0.098
0.991	$(-1.04, 3.95)$	$(-0.93, 3.87)$	$(-1.03, 3.98)$	$(-1.03, 3.98)$
0.975	$(-0.64, 3.56)$	$(-0.57, 3.51)$	$(-0.61, 3.56)$	$(-0.61, 3.56)$
0.95	$(-0.33, 3.28)$	$(-0.29, 3.23)$	$(-0.31, 3.26)$	$(-0.31, 3.26)$

a From Robinson(1982);
b Our more refined exact value based on 10^8 permutations;
c The Skovgaard p-value and confidence interval repeated from Booth and Butler [4].

As another example, Supplementary Table 2 show the results for the second two-sample dataset, originally from Lehmann (Lehmann [8]) and further analyzed by Robinson [11] to demonstrate saddlepoint methods as an alternative to permutation. The published results include tests and confidence intervals using exact methods (originally based on 100,000 permutations), saddlepoint and Edgeworth expansions, to which we add more precise exact calculations and our MCC results. The MCC approach is highly accurate, performing as well or better than the competing approximations.

8. Appendix H: Computational complexity

For the 36 scenarios decribed in the main text, m ranged from 1024 to 262,144, and n ranged from 512 to 4096. Elapsed time in seconds was computed using the

TABLE 2
Robinson table 2

Data for Robinson Table2:
Effect of analgesia for two classes. Data originally from Lehmann (1975, p 92)
Class I: 17.9,13.3,10.6,7.6,5.7,5.6,5.4,3.3,3.1,0.9
Class II: 7.7,5.0,1.7,0.0,-3.0,-3.1,-10.5.

Sign. Level	*Exact*[a]	*Exact*[b]	*Skovgaard1*[c]	*Robinson1*[a]
	0.012	0.011	0.011	0.010
0.990	$(-0.10, 16.13)$	$(-0.14, 15.97)$		$(0.06, 15.96)$
0.975	$(0.92, 14.68)$	$(0.96, 14.61)$		$(1.18, 14.51)$
0.95	$(1.88, 13.52)$	$(1.88, 13.52)$		$(2.07, 13.46)$
	Robinson2[a]	*Edgeworth*[a]	MCC	MCC_1
	0.011	0.014	0.011	0.096
	$(-0.15, 16.19)$	$(-0.02, 15.76)$	$(-0.16, 15.39)$	$(-0.16, 15.40)$
	$(0.98, 14.72)$	$(1.07, 14.44)$	$(0.96, 14.29)$	$(0.96, 14.31)$
	$(1.86, 13.64)$	$(1.95, 13.45)$	$(1.88, 13.40)$	$(1.88, 13.41)$

[a] From Robinson(1982); [b] Our more refined exact value based on 10^8 permutations; [c] The Skovgaard p-value and 95% CI as repeated in Booth and Butler [4].

system.time function in *R* for the Xeon 2.65 GHz processor. A simple regression model fit to the model $time = \beta mn + \epsilon$, with the resulting fits shown as lines in Supplementary Figure 7. The time is approximately linear on the log scale, as expected. For large m and n, the model fits well, although variations from the fit occur for smaller values.

9. Appendix I: Derivation of the MCC_1 terms

For a given permutation π, denote the first chosen $\mathbf{y}$ value as $y_{\pi[1]}$. Then

$$\begin{aligned} r_\pi &= \sum_j x_j y_{\pi[j]} \\ &= x_1 y_{\pi[1]} + \sum_{j=2}^{n} x_j y_{\pi[j]} \end{aligned}$$

Using "$-$" to represent the removal of an element, we have

$$\begin{aligned} r_{-\pi[1]} &= corr(x_{-1}, y_{-\pi[1]}) \\ &= \frac{(\sum_{j=2}^n x_j y_{\pi[j]})(n-1) - \sum_{j=2}^n x_j \sum_{j=2}^n y_{\pi[j]}}{\sqrt{n\sum_{j=2}^n x_j^2 - (\sum_{j=2}^n x_j)^2}\sqrt{n\sum_{j=2}^n y_{\pi[j]}^2 - (\sum_{j=2}^n y_{\pi[j]})^2}} \\ &= \frac{(\sum_{j=2}^n x_j y_{\pi[j]})(n-1) - (-x_1)(-y_{\pi[1]})}{\sqrt{n(1-x_1^2) - (-x_1)^2}\sqrt{n(1-y_{\pi[1]}^2) - (-y_{\pi[1]})^2}} \end{aligned}$$

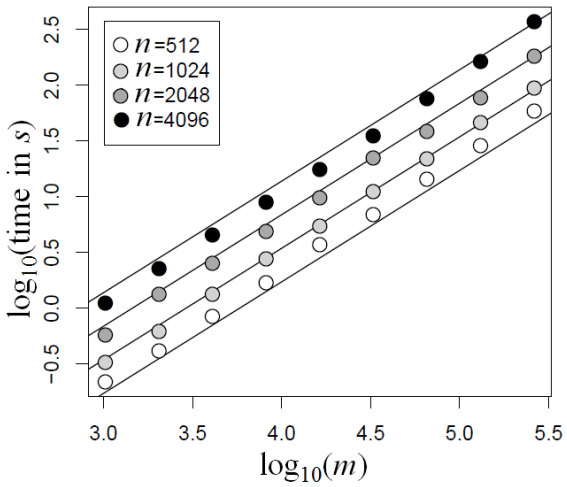

FIG 7. *Elapsed time in seconds to run MCC vs. m for various values of n (axes on* $\log_{10}$ *scale). Lines indicate fits from regression modeling.*

Rearranging terms provides

$$r_\pi = x_1 y_{\pi[1]} + b_{0,\pi[1]} + b_{1,\pi[1]} r_{-\pi[1]},$$

where

$$b_{0,\pi[1]} = (\sqrt{n(1-x_1^2) - (-x_1)^2}\sqrt{n(1-y_{\pi[1]}^2) - (-y_{\pi[1]})^2})(-x_1)(-y_{\pi[1]}),$$

$$b_{1,\pi[1]} = (\sqrt{n(1-x_1^2) - (-x_1)^2}\sqrt{n(1-y_{\pi[1]}^2) - (-y_{\pi[1]})^2})(n-1).$$

10. Appendix J: Example from the Golub data

We further illustrate the concepts with an example from the highly cited Golub ALL/AML expression dataset (R/ Bioconductor *golubEsets*, data signed square root transformed). The data represent expression of $m = 7129$ genes for $n = 38$ samples, 27 of which are from patients with Acute Lymphoblastoid Leukemia, and 11 from patients with Acute Myeloid Leukemia. Differential expression analysis of AML/ALL status by equal-variance t-tests reveals that the gene *LCTS4* shows the most evidence, with $p = 9.6 \times 10^{-11}$. Here $\mathbf{x}$ is the gene's expression, and $\mathbf{y}$ is a $\{0, 1\}$ indicator vector for AML status. Other parametric tests provide wildly differing evidence (8 orders of magnitude), with $p = 3.0 \times 10^{-6}$ for an unequal variance t-test, and $p = 8.6 \times 10^{-3}$ for a logistic regression of ALL/AML status on the gene's expression. These differences result in different qualitative conclusions − for example, a Bonferroni multiple-test correction applied to the

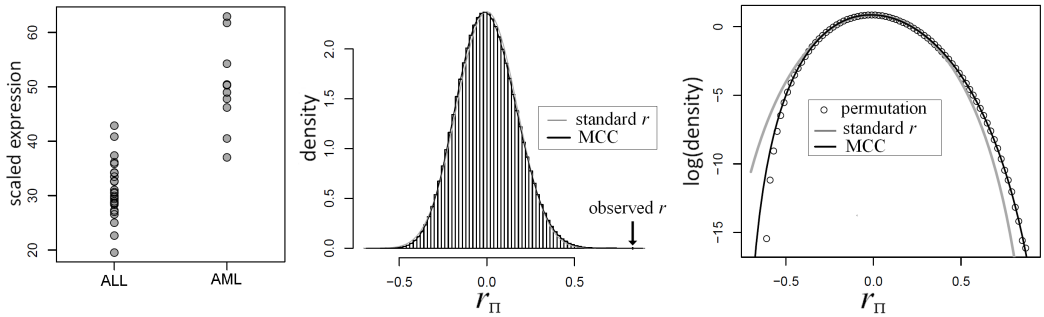

FIG 8. *Expression of LCTS4 vs. AML/ALL status, Golub dataset. The data (left panel) show a large mean difference between the leukemia types. Middle panel: the permuted correlation coefficient between* $\mathbf{x}$ *and* $\mathbf{y}$*, with the overlaid standard and MCC density approximations appearing almost identical. Right panel: the histogram values on the log scale better highlight the contrast between parametric analysis (two-sample t) and permutation, which is closely matched by the MCC approximation.*

equal variance t results in $p_{Bonferroni} = 6.8 \times 10^{-7}$, while for logistic regression is $p_{Bonferroni} = 1$.

Supplementary Figure 8 shows the results for this gene. For much of the range, the histogram is closely approximated by the standard r density, the difference between standard r and MCC is almost imperceptible. However, the standard density fails in the extremes, as can be seen on the log scale in the figure, while the MCC approach continues to work well in the extremes. For this gene, we performed 2×10^9 permutations, determining $p_{double} = 2.3 \times 10^8$ and $p_{two} = 1.1 \times 10^8$, while MCC gives estimates of $p_{double} = 2.1 \times 10^8$ and $p_{two} = 1.1 \times 10^8$.

We note that the standard density approximation is intended for unconditional inference, i.e. not conditioning on the observed $\mathbf{x}$ and $\mathbf{y}$. Thus it is in some sense unfair to expect a close correspondence to the permutation distribution, which is inherently conditional on the data. However, as we show below, if the densities of X and Y are skewed, standard parametric p-values tend to be inaccurate *on average*, in a manner that is largely reflected in comparisons such as shown in Figure 8.

11. Appendix K: Derivation of the moments of stratified $\mathbf{A}=\sum \boldsymbol{A_k}$

Here we propose an approximation to exact testing in which permutation is performed within each stratum level for covariate $\mathbf{z}$.

Suppose we have K strata for $\mathbf{z}$. Let J_k denote the n_k samples belonging to stratum k. Also, define $A_k = \sum_{j \in J_k} x_j y_j$ and $A = \sum_j x_j y_j = \sum_k A_k$.

Without loss of generality, we can center the y values within each stratum, so that $\sum_{j \in I} y_j = 0$. This implies $E_\Pi(A) = 0$, which simplifies analysis. We need to solve for the first four moments of A under permutation, i.e., $E_\Pi(A^2)$,

$E_\Pi(A^3)$, $E_\Pi(A^4)$, and note that the $\{A_k\}$ are independent of each other. Thus we must obtain $E_\Pi(A_k^2)$, $E_\Pi(A_k^3)$, $E_\Pi(A_k^4)$ for each k. After obtaining these moments, we can find $E_\Pi(A^2)$, etc., following standard rules for independent random variables. Similar to Appendix C, ultimately we need only compute moments for $\mathbf{x}$ and $\mathbf{y}$ separately. However, here $\mathbf{x}$ and $\mathbf{y}$ are not scaled, and so we re-derive the needed quantities.

$$
\begin{aligned}
(\sum x_i)^4 &= \sum x_i^4 + 4\sum_i\sum_{j\neq i} x_i^3 x_j + 6\sum_i\sum_{j\neq i} x_i^2 x_j^2 && (11.1)\\
&+ 12\sum_i\sum_{j\neq i}\sum_{k\neq i,k\neq j} x_i^2 x_j x_k + \sum_i\sum_{j\neq i}\sum_{k\neq i,j}\sum_{l\neq i,j,k} x_i x_j x_k x_l && (11.2)\\
(\sum x_i)^3 &= \sum x_i^3 + 3\sum_i\sum_{j\neq i} x_i^2 x_j + \sum_i\sum_{j\neq i}\sum_{k\neq i,k\neq j} x_i x_j x_k\\
\sum_i\sum_{j\neq i} x_i^3 x_j &= \sum_i\sum_j x_i^3 x_j - \sum_i x_i^4 && (11.3)\\
&= (\sum x_i^3)(\sum x_j) - \sum x_j^4\\
\sum_i\sum_{j\neq i} x_i^2 x_j^2 &= \sum_i\sum_j x_i^2 x_j^2 - \sum x_i^4\\
&= (\sum x_i^2)(\sum x_j^2) - \sum x_i^4 && (11.4)
\end{aligned}
$$

$$
\begin{aligned}
E(x_i^3) &= \sum x_i^3/n\\
\sum_i\sum_{j\neq i}\sum_{k\neq i,j} x_i^2 x_j x_k &= \sum_i\sum_j\sum_k x_i^2 x_j x_k - \sum x_i^4 - (11.3) - 2(11.4)\\
&= \sum x_i^2(\sum x_j)^2 - \sum x_i^4 - (11.3) - 2(11.4)
\end{aligned}
$$

$$
\begin{aligned}
E(x_i^2 x_j | i \neq j) &= \frac{\sum_i \sum_{j \neq i} x_i^2 x_j}{n(n-1)} \\
&= \frac{\sum_i \sum_j x_i^2 x_j - \sum_i x_i^3}{n(n-1)} \\
&= \frac{\sum_i x_i^2 (\sum_j x_j) - \sum_i x_i^3}{n(n-1)} \\
&= \frac{\sum_i x_i^2 n\overline{x} - \sum_i x_i^3}{n(n-1)} \\
E(x_i x_j x_k | i \neq j, i \neq k, k \neq j) &= \frac{\sum_i \sum_{j \neq i} \sum_{k \neq i, k \neq j} x_i x_j x_k}{n(n-1)(n-2)} \\
&= \frac{\sum_i \sum_j \sum_k x_i x_j x_k - 3 \sum_i \sum_{j \neq i} x_i^2 x_j - \sum_i x_i^3}{n(n-1)(n-2)} \\
&= \frac{\sum_i x_i (\sum_j x_j (\sum_k x_k)) - 3(\sum_i \sum_j x_i^2 x_j) - \sum_i x_i^3}{n(n-1)(n-2)} \\
&= \frac{n^3 \overline{x}^3 - 3(\sum_i x_i^2) n\overline{x} + 2 \sum_i x_i^3}{n(n-1)(n-2)}
\end{aligned}
$$

$$
\begin{aligned}
E(x_i^2 x_j x_k | i \neq j, i \neq k, k \neq j) &= \frac{\sum_i \sum_{j \neq i} \sum_{k \neq i, k \neq j} x_i^2 x_j x_k}{n(n-1)(n-2)} \\
&= \frac{1}{n(n-1)(n-2)} \{ \sum x_i^2 (\sum x_j)^2 - \sum x_i^4 - (11.3) - 2(11.4) \} \\
&= \frac{1}{n(n-1)(n-2)} \{ \sum x_i^2 (\sum x_j)^2 - 2 \sum x_i^3 \sum x_j \\
&- (\sum x_i^2)^2 + 2 \sum x_i^4 \}
\end{aligned}
$$

$$
\begin{aligned}
(\sum xy)^4 &= \sum x^4 y^4 + \sum_i \sum_{j \neq i} x_i^3 x_j y_i^3 y_j \\
&+ \sum_i \sum_{j \neq i} x_i^2 x_j^2 y_i^2 y_j^2 + \sum_i \sum_{j \neq i} \sum_{k \neq i, k \neq j} x_i^2 x_j x_k y_i^2 y_j y_k \\
&+ \sum_i \sum_j \sum_k \sum_l x_i x_j x_k x_l y_i y_j y_k y_l
\end{aligned}
$$

$$
\begin{aligned}
E_{\Pi}(A_k^4) &= E[(\sum x_i y_i)^4] \\
&= \sum E(x_i^4)E(y_i^4) + \sum_i \sum_{j \neq i} E(x_i^3 x_j)E(y_i^3 y_j) \\
&+ \sum_i \sum_{j \neq i} E(x_i^2 x_j^2)E(y_i^2 y_j^2) + \sum_i \sum_{j \neq i} \sum_{k \neq i, k \neq j} E(x_i^2 x_j x_k)E(y_i^2 y_j y_k) \\
&+ \sum_i \sum_j \sum_k \sum_l E(x_i x_j x_k x_l)E(y_i y_j y_k y_l)
\end{aligned}
$$

References

[1] A. Agresti. *Categorical data analysis*, volume 359. John Wiley & Sons, 2002.

[2] A. Agresti and B. A. Coull. Approximate is Better than "Exact for Interval Estimation of Binomial Proportions. *The American Statistician*, 52(2)(190-203), 1998.

[3] M. Andres. Is fisher's exact test very conservative. *Computational Statistics and Data Analysis*, 19:579–591, 1995.

[4] J. G. Booth and R. W. Butler. Randomization distributions and saddlepoint approximations in generalized linear models. *Biometrika*, 77-4: 787–96, 1990.

[5] D. M. Gatti, A. A. Shabalin, T. Lam, F. A. Wright, I. Rusyn, and A. B. Nobel. FastMap: fast eQTL mapping in homozygous populations. *Bioinformatics*, 25:4:482–489, 2009.

[6] W. W. Hauck and A. Donner. Wald's Test as Applied to Hypotheses in Logit Analysis. *Journal of the American Statistical Association*, 72:851–853, 1977.

[7] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions, vol. 2 of wiley series in probability and mathematical statistics: Applied probability and statistics*. Wiley, New York,, 1995.

[8] E. L. Lehmann. Nonparametrics: Statistical Methods Based on Ranks. *San Francisco: Holden-Day*, 1975.

[9] P. McCullagh and J. A. Nelder. *Generalized Linear Model*. Chapman and Hall, 1983.

[10] F. Pesarin and L. Salmaso. *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons, 2010.

[11] J. Robinson. Saddlepoint Approximations for Permutation Tests and Confidence Intervals. *Journal of the Royal Statistical Society*, 44(1)(91-101), 1982.

[12] D. C. S. Stokes, M. E. and G. G. Koch. Categorical Data Analysis Using the SAS System. *SAS Institute Inc*, 2000.

[13] Y.-H. Zhou, W. T. Barry, and F. A. Wright. Empirical pathway analysis, without permutation. *Biostatistics*, 2013.

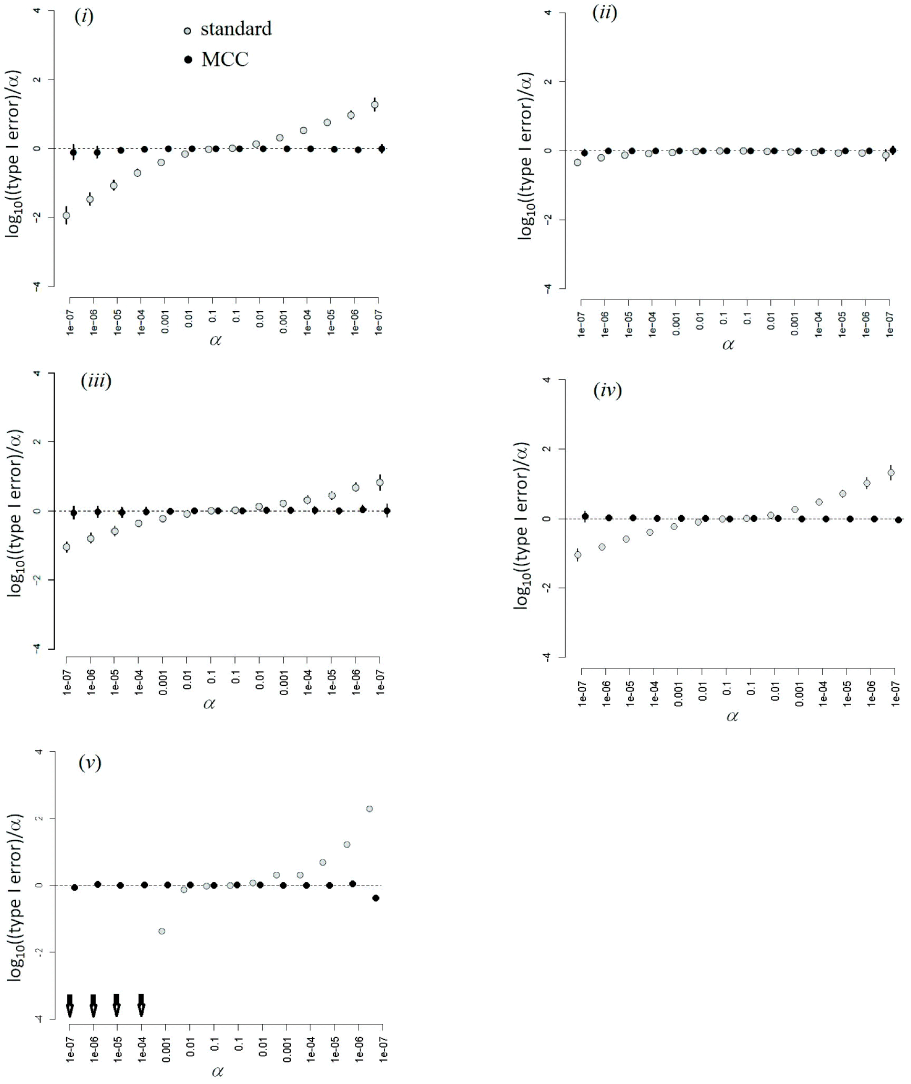

FIG 9. *Simulation scenarios (i)-(vi), sample size n=1000. See legend from main text Figure 5 for details.*

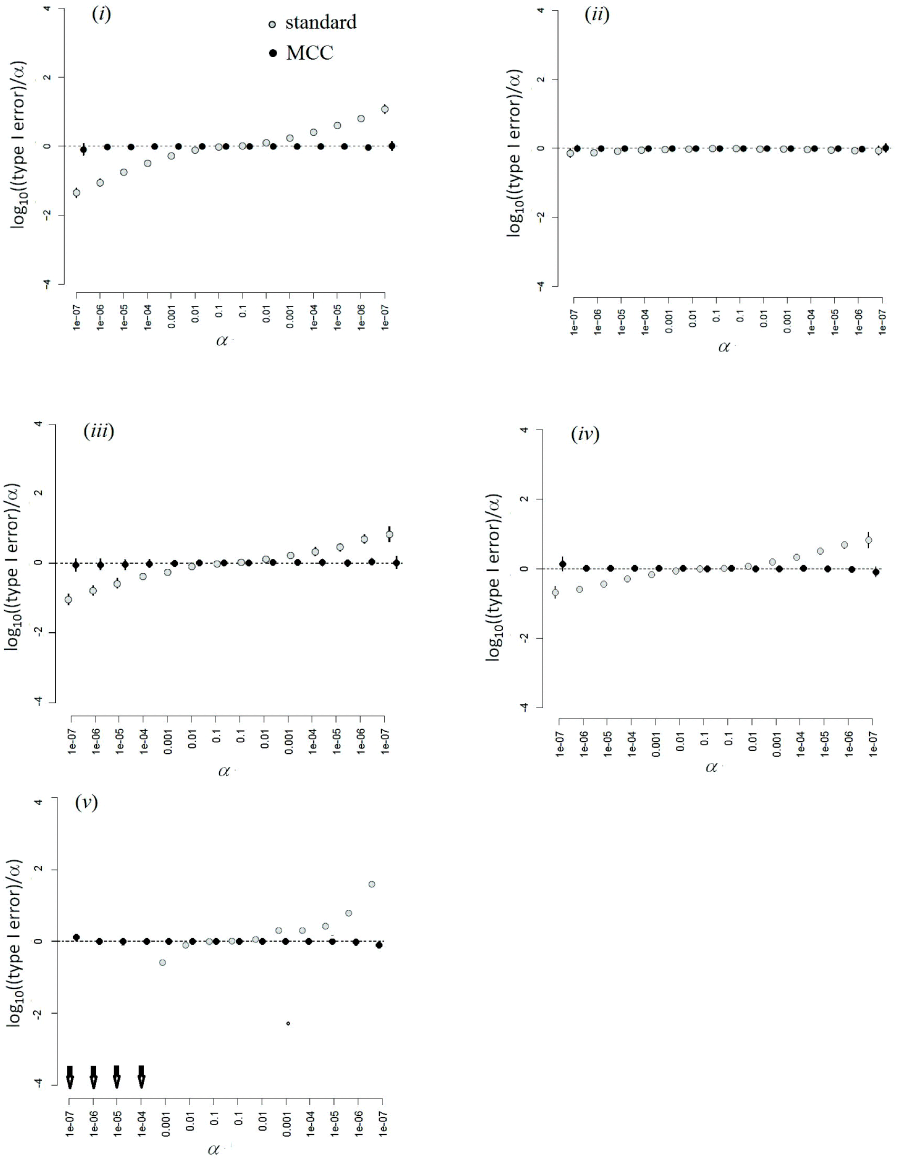

FIG 10. *Scenario (i)-(vi), sample size n=2000. See legend from main text Figure 5 for details*