Biostatistics (2015), **16**, 3, *pp.* 611–625
doi:10.1093/biostatistics/kxv007
Advance Access publication on March 18, 2015

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

YI-HUI ZHOU*

Bioinformatics Research Center, Department of Biological Sciences, North Carolina State University, Raleigh, NC 27695, USA

yihui_zhou@ncsu.edu

FRED A. WRIGHT

Bioinformatics Research Center, Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA

Summary

A number of biomedical problems require performing many hypothesis tests, with an attendant need to apply stringent thresholds. Often the data take the form of a series of predictor vectors, each of which must be compared with a single response vector, perhaps with nuisance covariates. Parametric tests of association are often used, but can result in inaccurate type I error at the extreme thresholds, even for large sample sizes. Furthermore, standard two-sided testing can reduce power compared with the doubled p-value, due to asymmetry in the null distribution. Exact (permutation) testing is attractive, but can be computationally intensive and cumbersome. We present an approximation to exact association tests of trend that is accurate and fast enough for standard use in high-throughput settings, and can easily provide standard two-sided or doubled p-values. The approach is shown to be equivalent under permutation to likelihood ratio tests for the most commonly used generalized linear models (GLMs). For linear regression, covariates are handled by working with covariate-residualized responses and predictors. For GLMs, stratified covariates can be handled in a manner similar to exact conditional testing. Simulations and examples illustrate the wide applicability of the approach. The accompanying *mcc* package is available on CRAN http://cran.r-project.org/web/packages/mcc/index.html.

Keywords: Density approximation; Exact testing; Permutation.

1. Introduction

High-dimensional datasets are now common in a variety of biomedical applications, arising from genomics or other high-throughput platforms. A standard question is whether a clinical or experimental variable (hereafter called the *response*) is related to any of a potentially large number of *predictors*. We use $\mathbf{y}$ to denote the response vector of length n (random vector Y, observed elements y_j), and $\mathbf{X}$ to denote the $m \times n$ matrix of predictors. Standard analysis often begins by testing for association of $\mathbf{y}$ vs. each row $\mathbf{x}_{i.}$ of $\mathbf{X}$, i.e. computing a statistic $r_i = r(\mathbf{x}_{i.}, \mathbf{y})$ for each hypothesis i. The most common corrections for multiple

*To whom correspondence should be addressed.

testing, such as Benjamini–Hochberg false discovery rate control, require only individual p-values for the m test statistics. Thus, at the level of a single hypothesis, the role of m is to determine the stringency of multiple testing. For modern genomic datasets, m can reach 1 million or more. For some datasets, standard parametric p-values may be highly inaccurate at these extremes, even for sample sizes $n > 1000$.

Although the basic problem described here is familiar, current techniques often fail for extreme statistics, or are not designed for arbitrary data types. The researcher often resorts to parametric testing, even when the model is not considered quite appropriate, or may rely on central limit properties without a clear understanding of the limitations for finite samples. In genomics problems, such as single nucleotide polymorphism (SNP) association testing involving contingency tables, the researcher may employ a hybrid approach in which most SNPs are tested parametrically, but those producing low cell counts are subjected to exact testing. Such two-step testing can be computationally intensive and cumbersome, and provides no guidance for situations in which the data are continuous or are mixtures of discrete and continuous observations. Our goal in this paper is to introduce a general trend testing procedure that is fast, provides accurate p-values simultaneously for all m hypotheses, and is largely distribution-free.

2. Exact testing and a summary of the approach

Exact testing is an attractive alternative to parametric testing, in which inference is performed on the observed $\mathbf{y}$ and $\mathbf{x}_{i\cdot}$. In this discussion, i is arbitrary, and we suppress the subscript. We use $\pi = 1, \ldots, n!$ to denote an index corresponding to each of the possible permutations, used as a subscript to represent reordering of a vector, with elements denoted by $\pi[1], \ldots, \pi[n]$. We use Π to denote a random permutation, producing the random statistic $r(\mathbf{x}, \mathbf{y}_{\Pi})$.

The null hypothesis H_0 holds that the distributions generating $\mathbf{x}$ and $\mathbf{y}$ are independent, and we use X, Y to refer to the respective random variables. We assume that at least one of the distributions is exchangeable, so that the joint probability distribution of (say) the response is $P_Y(y_1, y_2, \ldots, y_n) = P_Y(y_{\pi[1]}, y_{\pi[2]}, \ldots, y_{\pi[n]})$ for each π (Good, 2005, p. 268). Appendix A (see supplementary material available at *Biostatistics* online) contains additional remarks on the assumptions underlying exact testing and perspectives for our specific context. The vectors $\mathbf{x}$ and $\mathbf{y}$ are fixed and observed, but the standard parametric tests rely on distributional assumptions for X and Y. Thus, we will informally refer to the observed vectors as "discrete" or "continuous" according to the population assumptions, although the observed vectors are always discrete.

Throughout this paper, we use the statistic $r(\mathbf{x}, \mathbf{y}) = \sum_j x_j y_j$, which is sensitive to linear trend association. For discussion and plotting purposes, it is often convenient to center and scale $\mathbf{x}$ and $\mathbf{y}$ so that r is the Pearson correlation. As we show in Appendix B (see supplementary material available at *Biostatistics* online), most trend statistics of interest, including contingency table trend tests, t-tests, linear regression, and generalized linear model (GLM) likelihood ratios, are permutationally equivalent to r.

Here we introduce the *moment-corrected correlation* (MCC) method of testing. The basic idea is as follows. Using moments of the observed $\mathbf{x}$ and $\mathbf{y}$, we obtain the first four exact permutation moments of r_{Π}. We then apply a density approximation to the distribution, performed for the rows of matrix $\mathbf{X}$ simultaneously to obtain p-values for all m hypotheses. MCC is "robust" in the sense that exact permutation moments are used, with two extra moments beyond the two moments that are used in, e.g. a normal approximations underlying standard parametric statistics.

3. A motivating example

We illustrate the concepts with an example from the genome-wide scan of Wright *and others* (2011), reporting association of ∼570 000 SNPs with lung function in 1978 cystic fibrosis patients with the most common form of the disease. A significant association was reported on chromosome 11p, in the region between the genes *EHF* and *APIP*. The original analysis analyzed the quantitative phenotype vs. genotype as a predictor in a linear regression model, with additional covariates including sex and several genotype principal components, which can equivalently be analyzed by computing the correlation of covariate-corrected phenotype vs. covariate-corrected genotypes (see Section 5). To illustrate the effects of using skewed phenotype y, we further dichotomized the phenotype to consider a hypothetical follow-up regional search for associations to a binary indicator for extreme phenotype ($y = 1$ if the lung phenotype is above the 10th percentile, $y = 0$ otherwise). With a highly skewed phenotype, these data are also emblematic of highly unbalanced case–control data, as might occur when abundant public data are used as controls (Mukherjee *and others*, 2011).

We performed logistic regression for phenotype vs. genotype (covariate-corrected) for 3117 SNPs in a 1.5 Mb region containing the genes, and applied Benjamini–Hochberg q-value adjustment for the

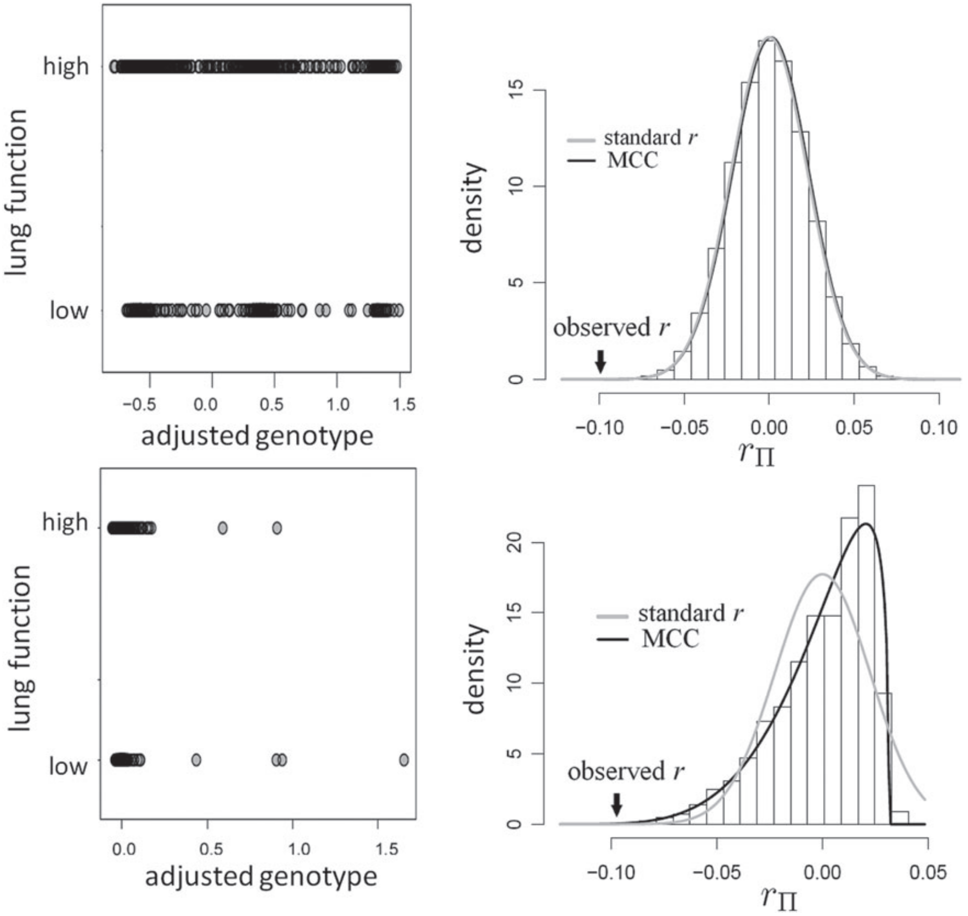

Fig. 1. MCC for genotype association testing. Upper left: data for SNP rs2956073. Although SNP genotypes were initially coded as 0, 1, 2, after covariate adjustment they appear as shown. Upper right: histogram of r_Π, with standard r and MCC fitted densities. Lower left: SNP rs180784621, with a low minor allele frequency producing considerable skew in the adjusted genotypes. Lower right: histogram of r_Π shows that MCC fits much better than standard r.

region. Two SNPs met regional significance at $q < 0.01$, rs2956073 (logistic Wald $p = 7.9 \times 10^{-6}$), and rs180784621 ($p = 1.8 \times 10^{-5}$). The sample size of $n = 1978$ would seem more than sufficient for analysis using large sample approximations. However, histograms of the genotype–phenotype correlation coefficients (Figure 1) for 10^8 permutations for each SNP raises potential concerns for "standard" analysis of the second SNP (lower panels). Here the correlation distribution r_Π is strongly left-skewed, suggesting potential inaccuracy in p-values based on standard parametric approaches. Direct permutation, as shown in the figure, provides accurate p-values, but is computationally intensive, especially when performed for the entire matrix $\mathbf{X}$.

Overlaid on the histograms (Figure 1) in gray is the "standard r" density $f(r) = B(\frac{1}{2}, \frac{1}{2}(n-2))^{-1}(1 - r^2)^{(n-4)/2}$, $r \in (-1, 1)$ where $B()$ is the beta function. This density is the unconditional distribution of r under H_0 if either X or Y is normally distributed (Lehmann and Romano, 2005), and tests based on it are equivalent to t-testing based on simple linear regression or the two-sample equal-variance t, and similar to a Wald statistic from logistic regression.

The example provides a preview of the advantage of using MCC. For the top right panel, the histogram is closely approximated by the standard r density, as well as by MCC (black curve). However, for the lower right panel, MCC is much more accurate than standard r in approximating the histogram, with dramatic differences in the extreme tails. The reason for the improvement is that MCC uses the first four exact moments of r_Π to provide a density fit. When the distribution of r_Π is skewed, more than one type of p-value might reasonably be used. Typical choices include p-values based on either extremity of $|r_\Pi|$, or by doubling the smaller of the two "tail" regions (Kulinskaya, 2008, see below). For the first SNP, these two p-values (based on extremity or tail-doubling) are nearly identical, but can be very different when the distribution of r_Π is skewed, as in the lower panels. Thus, in addition to accuracy of p-values, we must also consider the relative power obtained by the choice of p-value.

4. Trend statistics and p-values

4.1 r_Π *and trend statistics are permutationally equivalent*

Over permutations, r is one-to-one with most standard trend statistics, which are described in terms of distributional assumptions for X and Y. A list of such standard statistics is given below, and Appendix B (see supplementary material available at *Biostatistics* online) provides citations and derivations for permutational equivalence. Standard parametric tests/statistics include simple linear regression (X arbitrary, Y continuous), and the two-sample problem as a special case (X binary, Y continuous). For the latter we do not distinguish between equal-variance and unequal-variance testing, working directly with mean differences in the two samples under permutation. Categorical comparisons include the contingency table linear trend statistic (X ordinal, Y ordinal) (Stokes and Koch, 2000), which includes the Cochran–Armitage statistic (X ordinal, Y binary) and the χ^2 and Fisher's exact tests for 2×2 tables. If X or Y represent ranked values, the standard statistics include the Wilcoxon rank sum (X binary, Y ranked values), and the Spearman rank correlation (X ranked, Y ranked). Other statistics with the property include likelihood ratios or deviances for common two-variable GLMs, when the permutations have been partitioned according to sign(r). These GLMs include logistic and probit (X binary or continuous, Y binary), Poisson (X continuous or discrete, Y integer), and common overdispersion models.

For the standard statistics, it is thus sufficient to work directly with r_Π for testing against the null. Assuming that the investigator is performing permutation testing, there is no need to be concerned over differences among the statistics, or to perform computationally expensive maximum likelihood fitting, because the statistics are equivalent. Finally, we note that the use of correlation makes it obvious that the roles of $\mathbf{x}$ and $\mathbf{y}$ are interchangeable.

4.2 *p-values*

The observed r_{obs} can be compared with r_Π to obtain a two-sided p-value, $p_{\text{two}} = \Pr(|r_\Pi| \geqslant |r_{\text{obs}}|)$. Alternatively, we might obtain left and right-tail p-values $p_{\text{left}} = \Pr(r_\Pi \leqslant r_{\text{obs}})$, $p_{\text{right}} = \Pr(r_\Pi \geqslant r_{\text{obs}})$, with "directional" $p_{\text{directional}} = \min(p_{\text{left}}, p_{\text{right}})$. The directional p-value is not a true p-value, as it uses the data to choose the favorable direction. However, simply doubling it produces a proper p-value, $p_{\text{double}} = 2 \times p_{\text{directional}}$. For skewed r_Π, p_{double} often has a power advantage over p_{two}, provided the investigator maintains equipoise in prior belief of positive vs. negative correlation between X and Y. The intuition behind the increased power of p_{double} comes from the fact that for a skewed r_Π, doubling the smaller of the two tail regions is typically smaller than the sum of the two tail regions used by p_{two}. Appendix C (see supplementary material available at *Biostatistics* online) proves the increased power for local departures from the null for a specific class of skewed densities. The historical use and properties of doubled p-values, as well as alternative constructions, are described in Kulinskaya (2008). The MCC approach described below is accurate for both p_{two} and p_{double}, but we primarily focus on p_{double}, and thus compare MCC and standard parametric tests in terms of accuracy of $p_{\text{directional}}$, except where noted.

5. Density fitting, computation, and an improvement

MCC can be used for a large variety of linear and GLMs and for categorical tests of trend. A simple extension to MCC is also proposed to improve accuracy in the presence of modest outliers. Finally, we describe approaches to handle covariates. Several well-studied examples from the literature, not necessarily high throughput, are used to illustrate. The mean and variance of correlation r_Π over the $n!$ exhaustive permutations are always 0 and $1/(n-1)$, respectively (Pitman, 1937). The exact skewness and kurtosis, however, depend on the moments of $\mathbf{y}$ and $\mathbf{x}$ (and therefore vary with i) and are derived in Pitman (1937) in terms of Fisher k-statistics. In Appendix D (see supplementary material available at *Biostatistics* online), we illustrate key steps in the computations of the kurtosis of r_Π using more familiar expressions. The key to the speed of MCC is the fact that the moments can be computed for all rows of $\mathbf{X}$, and therefore r_Π for each i, using a single set of matrix operations. The entire MCC procedure can be expressed algorithmically as shown below.

Algorithm 1 Compute p-values for moment-corrected correlation

```
Compute moments for $\mathbf{y}$ and all rows of $\mathbf{X}$. These and remaining steps are performed simultaneously for all $i \in \{1, \ldots, m\}$.
Compute moments for $r_{\Pi,i}$ (e.g., Appendix D).
Calculate $\alpha_i$ and $\beta_i$ as the parameters for the beta density having the same skewness and kurtosis as $r_{\Pi,i}$ (Appendix E).
For the beta mean $\mu_i = \alpha_i/(\alpha_i + \beta_i)$ and variance $\sigma_i^2 = \alpha_i\beta_i/\big((\alpha_i + \beta_i)^2(\alpha_i + \beta_i + 1)\big)$, calculate $u_i = \mu_i + \sqrt{n-1}\ \sigma_i\ r_i$. Under $H_0$ the beta density approximation for $u$ is $f_{\alpha_i,\beta_i}(u) = \big(1/B(\alpha_i, \beta_i)\big)u^{\alpha_i - 1}(1-u)^{\beta_i - 1}$ where $B()$ is the beta function, and corresponding cdf $F_{\alpha_i,\beta_i}$.
Compute $\hat{p}_{two,i} = F_{\alpha_i,\beta_i}(\mu_i - |u_i - \mu_i|) + 1 - F_{\alpha_i,\beta_i}(\mu_i + |u_i - \mu_i|)$, $\hat{p}_{left,i} = F_{\alpha_i,\beta_i}(u_i)$, $p_{right,i} = 1 - F_{\alpha_i,\beta_i}(u_i)$, and $\hat{p}_{double,i} = 2\min(\hat{p}_{left,i}, \hat{p}_{right,i})$.
```

If n is very small, or there are numerous tied values in $\mathbf{x}$ and $\mathbf{y}$, the accuracy of the density approximation will be slightly affected by tied instances in r_Π, and the approximation is often closer to the mid p-value, e.g. $\hat{p}_{\text{right}} \approx \Pr(r_\Pi > r_{\text{obs}}) + \frac{1}{2}P(r_\Pi = r_{\text{obs}})$. To examine the effects of tied r_Π values, in Appendix F (see supplementary material available at *Biostatistics* online) we considered the worst-case scenario of using

MCC for the 2 × 2 Fisher exact test for small sample sizes, and for the Wilcoxon rank-sum test with a high proportion of tied observations.

A proposed alternative to direct permutation is to use saddlepoint approximations (Robinson, 1982; Booth and Butler, 1990), which have been examined in considerable detail for a few relatively small datasets. In Appendix G (see supplementary material available at *Biostatistics* online), we illustrate the analysis of two datasets from Lehmann (1975). The datasets show that MCC is at least as accurate as saddlepoint approximations, and far easier to implement. The examples also illustrate that MCC can be used to obtain exact confidence intervals for simple linear models. For the model $Y = \beta_0 + \beta_1 X + \epsilon_Y$, where the ϵ values are assumed drawn independent and identically distributed from an arbitrary density, MCC can be used to provide approximations to exact confidence intervals for β_1, by inverting the test using the MCC p-values for comparing $\mathbf{x}$ to $\mathbf{y} - \beta_1\mathbf{x}$ (the value of β_0 is immaterial in the correlation).

5.1 *Computational cost*

MCC requires several matrix operations performed on $\mathbf{X}$, involving computing element-wise powers (up to 4) followed by row summations, which are $O(mn)$ operations. Other operations are of lower order, so the overall order is $O(mn)$. To empirically demonstrate, we ran the *R* scripts using simulated data with $m = 2^a$, with $a \in \{10, 11, \ldots, 18\}$ (i.e. m ranging from 1024 to 262 144), and $n = 2^b$, with $b \in \{9, \ldots, 12\}$ (i.e. n ranging from 512 to 4096). The $9 \times 4 = 36$ scenarios were analyzed using a Xeon 2.65 GHz processor, and the largest scenario ($m = 262\,144,\ n = 4096$) took 376 s. Computation for a genome-wide association scan with m=1 million markers and $n = 1000$ individuals takes a similar time (≈6 min). Appendix H (see supplementary material available at *Biostatistics* online) shows the timing for all 36 scenarios, and the results of a model fit to the elapsed time. We note that computation of the observed r for all m features is itself an $O(mn)$ computation.

5.2 *A one-step improvement to MCC*

Extreme values in either $\mathbf{x}$ or $\mathbf{y}$ present a challenge for MCC, especially in smaller datasets, as these values have high influence and can even produce a multimodal r_Π distribution. Extensions of MCC using higher moments is possible, but cumbersome. A more direct approach is to condition on an influential observation, which we call the referent sample. Below, without loss of generality we can consider the referent sample to be sample 1. We have

$$r_\pi = \sum_j x_j y_{\pi_{[j]}} = x_1 y_{\pi[1]} + \sum_{j=2}^{n} x_j y_{\pi[i]} = x_1 y_{\pi[1]} + b_{0,\pi[1]} + b_{1,\pi[1]} r_{-\pi[1]},$$

where $r_{-\pi[1]}$ is the random correlation between the $\mathbf{x}$ and $\mathbf{y}$ vectors after removal of the x_1 and $y_{\pi[1]}$ elements (Appendix I of supplementary material available at *Biostatistics* online), and $b_{0,\pi[1]}$, $b_{1,\pi[1]}$ are normalization constants. The n possible $y_{\pi[1]}$ values each generate $(n-1)!$ values of $r_{-\pi[1]}$. We denote the beta density approximation applied to each of the n possibilities as $f(r|x_1, y_{\pi[1]})$, finally obtaining the approximation $g(r) = (1/n)\sum_{\pi[1]=1}^{n} f(r|x_1, y_{\pi[1]})$. We refer to this one-step approximation as MCC$_1$. The motivation behind MCC$_1$ is that the most extreme values of r_Π must contain pairings of extreme $\mathbf{x}$ and $\mathbf{y}$ elements, and so the benefit is often seen in the tail regions.

In order to avoid arbitrariness in the choice of "extreme" value, we can also consider each of the n observations in turn as the referent sample and average over the result (which we call MCC$_{1,\text{all}}$). Applying MCC$_{1,\text{all}}$ adds an additional factor n^2 in computation compared with MCC, and thus in practice we apply it only to features for which the MCC p-value is many orders of magnitude smaller than the standard parametric p-value.

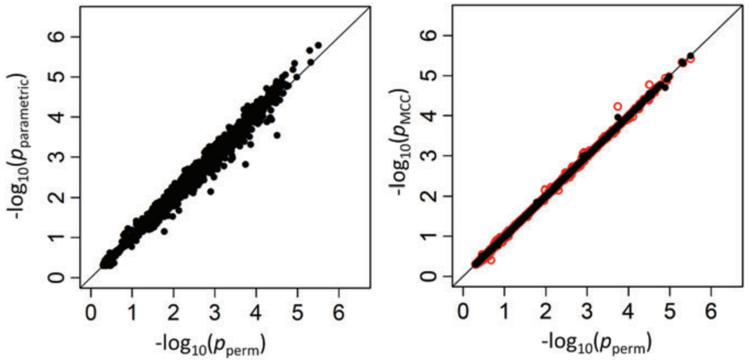

Fig. 2. Performance of MCC for the breast cancer survival data Left panel: directional p-values using a two-sample t test and standard p-values (y-axis) vs. a large number of permutations (x-axis). Right panel: p-values using MCC vs. permutations (red), and using MCC_1 (black).

5.3 *Examples*

As a high-throughput example, we use a breast cancer gene expression dataset, consisting of 236 samples on the Affy U133A expression array, with a disease survival quantitative phenotype (Miller *and others*, 2005). Figure 2 (left panel) shows the results of comparing directional p-values based on the t-statistic from standard linear regression to those of actual permutation. The permutation was conducted in two stages, with 10^6 permutations for each gene in stage 1, and for any gene with a permutation $p < 0.05$ in stage 1, another 10^8 permutations were performed. The right panel shows the analogous results for MCC (red, analyzed in 1 sec for all genes) and MCC_1 (black, analyzed in 1 min). Here for MCC_1 the sample with the most outlying survival phenotype value (judged by absolute deviation from the median) was used as the referent sample. Clearly, both versions of MCC considerably outperform regression in the sense of matching permutation p-values, and here MCC_1 provides a modest improvement over MCC.

Another example, in which both $\mathbf{x}$ and $\mathbf{y}$ are discrete, is given by the dataset published by Takei *and others* (2009), which describes association of Alzheimer disease with several SNPs in the *APOE* region. Although only a few SNPs were investigated, the approaches are identical to those used in genome scans involving up to millions of SNPs. The published analyses used the Cochran–Armitage trend statistic, which is compared with a standard normal. Exact p-values are feasible to compute in this instance. In these data, the case–control ratios are close enough to a 1:1 ratio that the trend statistic performs well, as do most other methods (see Figure 3). An exception is the Wald logistic p-value, which is the default logistic regression approach in genetic analysis tools such as PLINK (Purcell *and others*, 2007), and can depart noticeably from the exact result for the most extreme SNPs. The figure shows two-sided p-values, but the pattern for directional p-values is similar. For modern genomic analyses with over 1 million markers, computing logistic regression likelihood ratios can be time-consuming, as are exact analyses. Moreover, exact methods are not available (except via permutation) for imputed markers, which assume fractional “dosage” values Li *and others* (2010), while MCC is still applicable.

A more detailed examination of r_Π for a significant gene in an expression study is shown in Appendix J (see supplementary material available at *Biostatistics* online), focusing on the behavior in tail regions.

5.4 *Covariate control by residualization or stratification*

Although association testing of two variables is simple, it has wide application for screening purposes. This utility can be further extended to accommodate covariates when a regression model for Y is appropriate.

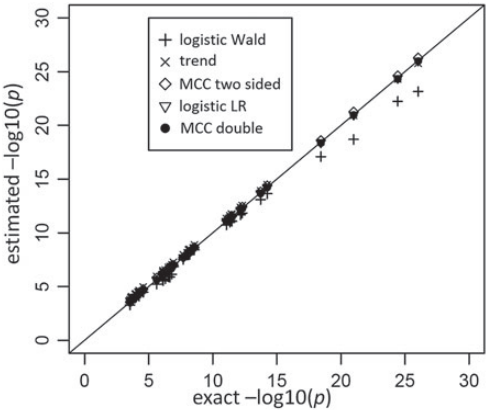

Fig. 3. Results for the analysis of 35 SNPs in the *APOE* region vs. late-onset Alzheimer disease in Japanese, from Takei *and others* (2009).

Suppose $Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon_Y$, where Z is a vector (or matrix) of covariates, β_2 a covariate coefficient (or vector of coefficients), and the ϵ_Y values are drawn independently from an arbitrary density. For standard multiple linear regression, the coefficient estimate $\hat{\beta}_1$ can equivalently be computed using (partial) correlation coefficient between Y and X, after each has been separately corrected/residualized for Z using linear regression (Frisch and Waugh , 1933). Let $\mathbf{y}_z$ denote the residuals after linear regression of Y on Z, and $\mathbf{x}_z$ after linear regression of X on Z. A straightforward testing approach is to use permutation or MCC to compare $\mathbf{y}_z$ to $\mathbf{x}_z$. The residualized quantities $\mathbf{x}_z$ and $\mathbf{y}_z$ are technically no longer exchangeable, even under the null $\beta_1 = 0$, due to error in the estimation of regression coefficients. However, the residualization-permutation approach has considerable empirical support (Kennedy and Cade, 1996), and for large sample sizes and few covariates, the impact of coefficient estimation error becomes negligible, especially in comparison to the inaccuracies produced by reliance on standard parametric p-values. To evaluate the effectiveness of residualized covariate control, for a fixed dataset we can compare the distribution of the true $r(\epsilon_{\mathbf{x}}, \epsilon_{\mathbf{y},\Pi})$ to that of $r(\mathbf{x}_z, \mathbf{y}_{z,\Pi})$, where $\mathbf{y}_{z,\pi}$ denotes the π-permutation of $\mathbf{y}_z$. An example of this kind of covariate control is shown in later simulations.

For GLMs under permutation, covariate control is not as straightforward, as there are no precisely analogous results to the partial correlations described above (or even quantities such as $\epsilon_{\mathbf{y}}$). We consider a discrete covariate vector $\mathbf{z} \in (1, \ldots, K)$ and define J_k as the indexes for the observations assuming the kth covariate value, i.e. $J_k = \{j : \mathbf{z} = k\}$. Denoting the within-stratum sum $A_k = \sum_{j \in J_k} x_j y_j$, we have $A = \sum_{j=1}^{n} x_j y_j = \sum_{k=1}^{K} A_k$. The moments of A are described in Appendix K (see supplementary material available at *Biostatistics* online). For this subsection, we use different notation (A instead of r) because, in the stratified setting, there is no algebraic advantage to rescaling $\mathbf{x}$ and $\mathbf{y}$ to be equivalent to the Pearson correlation. However, A is used and interpreted essentially in the same manner as r. The key to stratified covariate control is to perform permutation between $\mathbf{x}$ and $\mathbf{y}$ *within* strata, so there are $\Pi_{k=1}^{K}(n_k!)$ total permutations. We note that this stratified approach is similar to the principle underlying exact conditional logistic regression (Cox and Snell, 1989; Corcoran *and others*, 2001). The moments of each A_k under permutation are obtained using the same approach described earlier for r_Π, and because the strata are permuted independently, the moments for stratified A_Π are straightforward. We note that stratification does not change the computational complexity. For the 36 scenarios described in the earlier timing subsection, stratification by a 32-level covariate in fact reduced the computational time approximately 22% when averaged over the scenarios, due to some savings in lower-order computation.

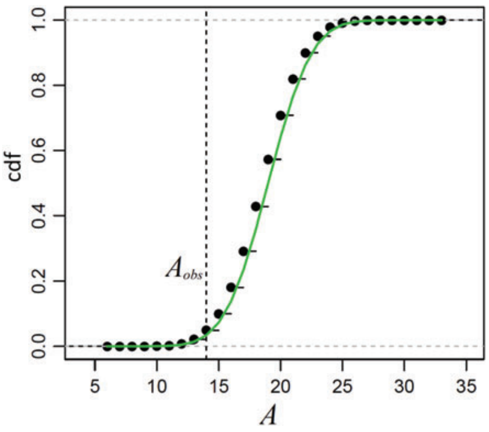

Fig. 4. The distribution of A for the endometrial cancer data of Breslow and Day (1980), with gall bladder disease as a predictor and matched case–control pairs. The empirical cdf is based on 10^7 stratified permutations, while the green curve is based on the MCC fit.

Figure 4 shows the result of applying MCC to the data from Breslow and Day (1980) on binary outcome data for endometrial cancer for 63 matched pairs, with gall bladder disease as the predictor and the matched pairs used to form covariate strata. This is an extreme instance with 63 strata. The figure shows the close fit of MCC to the permutation distribution, although due to discrete outcomes on the integers, a continuity correction is necessary for accuracy. For $A_{\text{observed}} = 14$, the doubled p-value is obtained by computing MCC after applying a 0.5 offset, resulting in $p_{\text{double}} = 0.1007$. The exact p-value obtained from 10^7 permutations is 0.0996.

6. Additional simulated datasets

We now consider additional simulations involve discrete outcomes or covariates, using "$\sim$" to signify the distribution from which values are drawn. We perform 10^9 permutations, for each of $n = 500$, 1000, 2000, performed for 10 simulations. The relatively large sample sizes are intended to match large-scale omics datasets, where large sample sizes are necessary to achieve stringent significance thresholds.

(i) *Two-sample mixed discrete/continuous*: we consider X drawn as a mixture of 50% zeros and the remainder drawn from a χ^2_1 density, $Y \sim \text{Binom}(1, 0.2)$. One "standard" approach is the two-sample unequal-variance t-test, although some investigators might be uncomfortable doing so in the presence of a large number of zero values, and permutation might be preferred.

(ii) *Ranks of mixed discrete/continuous*: we consider an initial X' drawn as a mixture with $X' = 0$ with probability 0.2, $X' = 3.0$ with probability 0.1, and the remainder drawn from a χ^2_1 density, $Y \sim \text{Binom}(1, 0.2)$. Then for observed $\mathbf{x}'$, we use the ranks $\mathbf{x} = \text{rank}(\mathbf{x}')$. The standard approach is the two-sample Wilcoxon rank-sum test, but due to the large number of ties, the standard distributional approximation for the Wilcoxon may not be accurate.

(iii) *Case/control*: $X \sim \text{Binom}(2, 0.1)$, $Y \sim \text{Binom}(1, 0.2)$, which mimics the outcome of an unbalanced case–control study with $\mathbf{y}$ as an indicator for case status, and $\mathbf{x}$ a discrete covariate such as SNP genotype. Standard approaches are the Cochran–Armitage trend test (shown here) or logistic regression.

(iv) *Continuous with continuous covariates*: To illustrate the effect of continuous covariate control, we simulated $\epsilon_X \sim \exp(1)$, $\epsilon_Y \sim \exp(1)$, with true models $Y = Z_1 + \epsilon_Y$, $X = 2Z_1 + \epsilon_X$. The

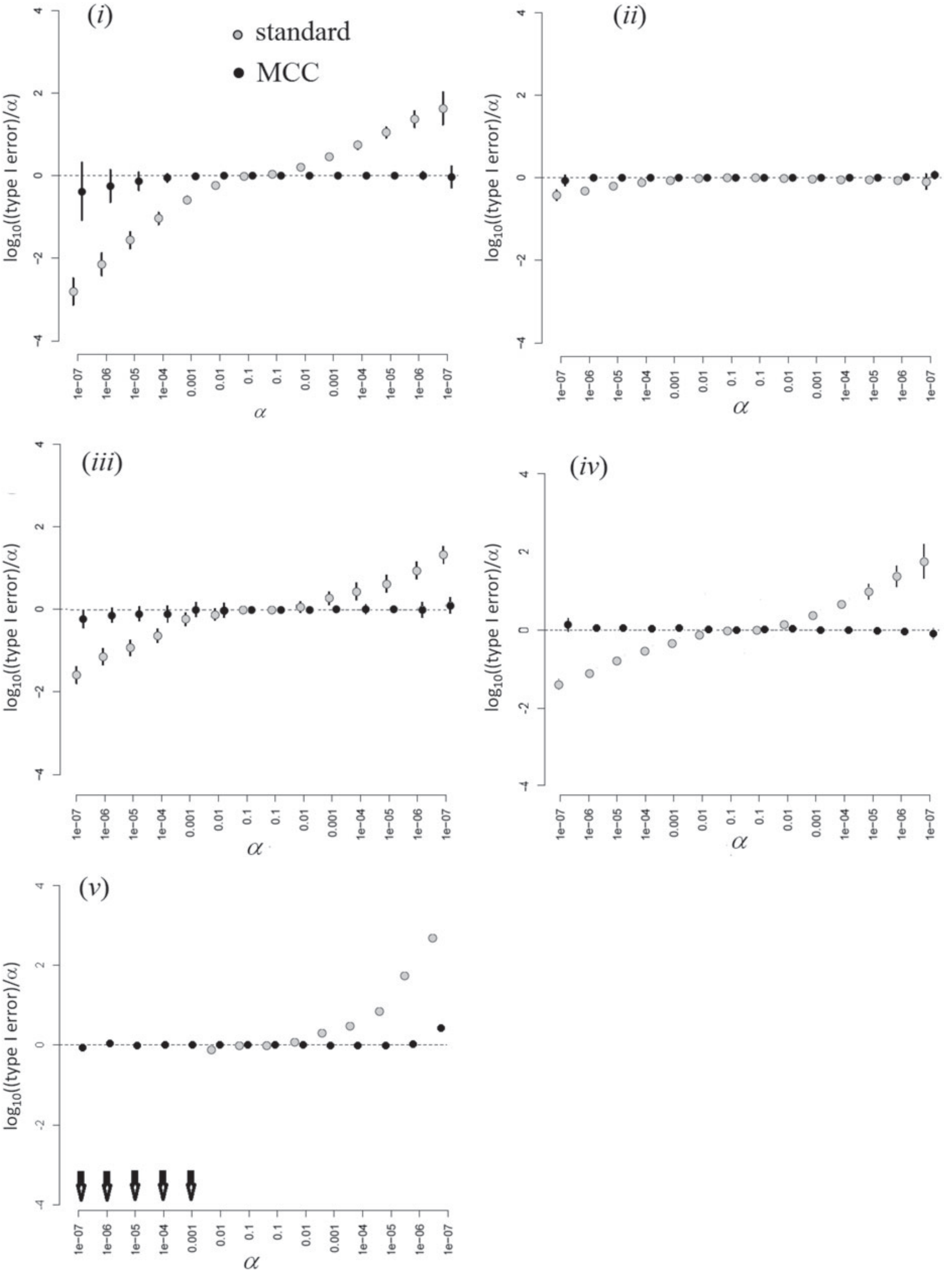

Fig. 5. Simulations with n = 500, simulation scenarios (i)–(v). Each x-axis is the false positive rate for a single tail of the r_{Π} distribution, with the correct threshold determined by 10^9 permutations (values on the left are for the left tail, values on the right for the right tail). The plotted points are the actual false positive rates for these thresholds, expressed as a $\log_{10}$ ratio compared with intended, via one-sided standard or MCC p-values. Arrows in panel (v) show outcomes for which the logistic regression likelihood ratio statistic did not converge. Error bars represents $+/-1$ SD for the 10 different simulations per scenario. Standard p-values are often incorrect by more than 2 orders of magnitude, and substantial inaccuracy persists for $n = 1000$ and $n = 2000$ (Figures 9 and 10 of supplementary material available at *Biostatistics* online).

covariates $Z_1 \sim N(0, 1)$ and $Z_2 \sim \exp(1)$ were fitted to the data, although only Z_1 was correlated with X and Y. The standard approach is linear regression. Here the α thresholds were determined using true realized errors $\epsilon_{\mathbf{x}}, \epsilon_{\mathbf{y}}$, and thus the performance of MCC reflects the merits of both the method and the residualization strategy.

(v) *Discrete with a stratified covariate*: We first simulated covariate $Z \sim \text{Binom}(1, 0.5)$, and then $X \sim \text{Binom}(2, 0.02 + 0.16\,Z)$, $Y \sim \text{Binom}(1, 0.04 + 0.32\,Z)$. Marginally, this is similar to (iii), except that X and Y have removable correlation induced by Z. The standard approach is logistic regression, with the effect of Z modeled as an additive covariate, which is correct under H_0. To determine α thresholds, the covariate was acknowledged by performing stratified permutation of Y vs. X under stratification, and MCC also used the stratified approach.

Figure 5, and Figures 9 and 10 of supplementary material available at *Biostatistics* online show the performance of directional p under the various scenarios. Performance is described in terms of $\log_{10}$((true type I error)/α), where the true type I error is the probability that $p_{\text{directional}} \leqslant \alpha$ for each of the 10 simulations, and the values are shown as mean $+/-$ 1 standard deviation. For scenarios (i), (iii),

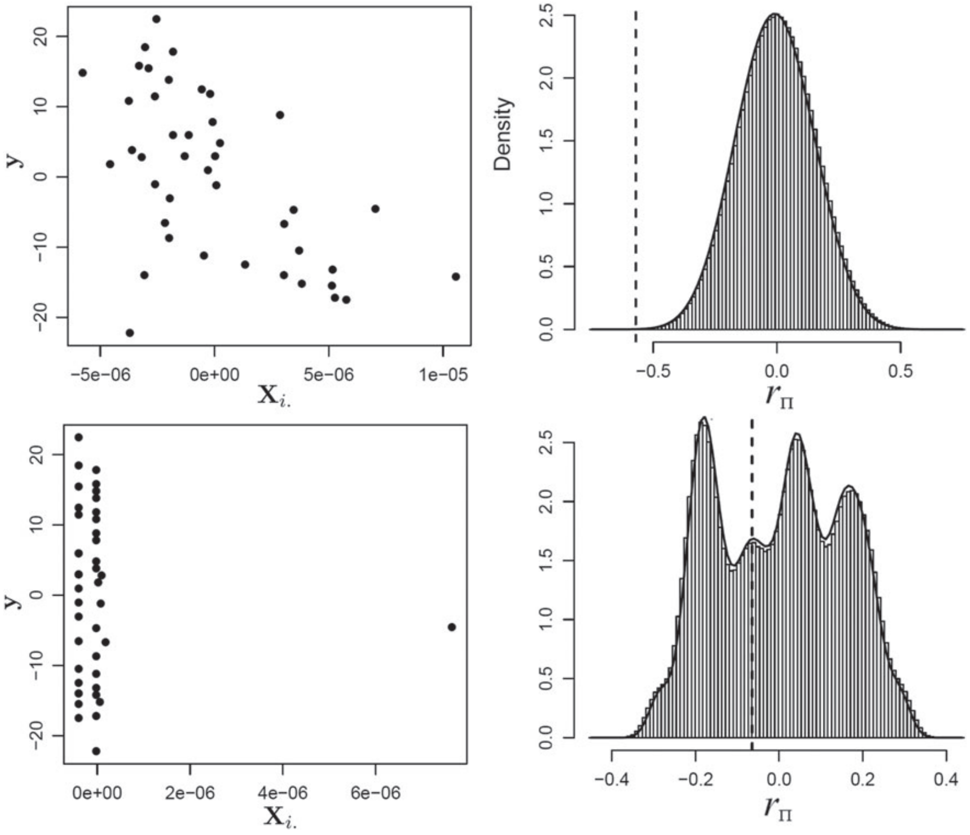

Fig. 6. Normalized RNA-Seq data vs. etoposide IC_{50}. Residualized $\mathbf{y}$ vs. $\mathbf{x}_{i.}$ and null permutation histograms for the gene *TEAD4* (upper panels) and *AGT* (lower panels). The fitted $\text{MCC}_{1,\text{all}}$ densities are overlaid on the histograms, and the observed r_{obs} shown as a dashed line.

(iv), and (v), both X and Y are skewed, and the standard approaches are highly anticonservative in the right tail and conservative in the left tail (see Figure 5). In fact, for scenario (v), the standard left directional p-values are often unable to achieve sufficiently small values in order to be rejected. The performance of standard approaches is particularly poor for $n = 500$, but the performance remains poor even for $n = 2000$ (Figure 10 of supplementary material available at *Biostatistics* online). MCC is much more accurate, down to $\alpha = 10^{-7}$. The standard approach for scenario (ii) is only modestly conservative in the left tail, which we attribute to the use of ranks, although due to ties some skew remains.

In summary, the standard approaches often have difficulty with type I error control, if both X and Y are skewed. However, MCC is well-behaved across all the scenarios. If the direction of skew were reversed for either X or Y, the conservativeness would appear on the right.

6.1 *An RNA-Seq example*

As a final example, incorporating several of the aspects described above, we consider the RNA-Seq expression data of Montgomery *and others* (2010) from $n = 42$ HapMap CEU cell lines, with ranked IC_{50} values from exposure to etoposide (Huang *and others*, 2007) used as a response $\mathbf{y}$. For these samples, $m = 30\,009$ genes which vary across the samples were used. We applied the residualization approach as described earlier, with sex as a stratified covariate. The RNA-Seq data were originally based on integer counts, which were then normalized as described in Zhou *and others* (2011) and covariate-residualized. We applied $\text{MCC}_{1,\text{all}}$ to the data for all features, requiring 25 min on the desktop PC used earlier for timing comparisons.

Figure 6 (top panels) shows the results for the most significant gene as determined by MCC, although not genome-wide significant (empirical $p_{\text{double}} = 7.4 \times 10^{-5}$ based on 10^8 permutations, $\text{MCC}_{1,\text{all}}$ $p_{\text{double}} = 9.5 \times 10^{-5}$). The lower panels show an example gene that is not significant, but for which the distribution is highly multimodal, due to the presence of extreme count values in $\mathbf{X}_i$. Nonetheless, $\text{MCC}_{1,\text{all}}$ can effectively fit the density, due to its successive conditioning strategy.

7. Discussion

We have described a coherent and fast approach to perform trend testing of a single vector vs. all rows of a matrix, which is a canonical testing problem arising in genomics and other high-throughput applications. As implemented in the *mcc* R package, the investigator need only provide $\mathbf{X}$ and $\mathbf{y}$, and possibly strata, and p_{double} and p_{two} will be automatically computed.

We emphasize that the idea of approximating permutation distributions is not new. In addition to saddlepoint approaches as described (Robinson, 1982; Booth and Butler, 1990), approaches using moment approximations for density fits include (Zhou *and others*, 2009; Zhou *and others*, 2013). However, these approaches have not fully exploited the simplicity of the score statistic and the attendant extreme speed of computation achieved here. We also note that our p-values are not adjusted for multiple comparisons, and thus are most immediately useful for methods such as Bonferroni or false discovery control. However, another important aspect of our approach is that, by ensuring greater uniformity of null p-values, each tested feature is placed on the same scale. Thus, as the computation for MCC is of the same order as computing the statistic r itself, MCC might be subjected to family-wise (across all features) permutation, or importance sampling (Kimmel and Shamir, 2006).

Our approach largely eliminates the need to be concerned over the appropriate choice of trend statistic, or whether parametric testing can be justified for the data at hand. In specific settings, such as genotype association testing, concern over the minor allele frequencies often leads investigators to perform exact testing for a subset of markers. We clarify here that the primary difficulty arises when both $\mathbf{x}$ and $\mathbf{y}$ are skewed, but the effects of the fourth moments may also be noticeable for extreme testing thresholds. For standard case–control studies with samples accrued in a 1:1 ratio, skewness may not be severe. However, for the analysis of binary secondary traits, the case:control ratio may depart from 1:1, and thus **y** may be highly skewed. In addition, the expense of sequence-based genotyping has increased interest in using shared or common sets of controls, which could then be much larger than the number of cases.

A possible alternative approach is to simply transform **x** and/or **y** (e.g. to match quantiles of a normal density) so that standard approximations fit well. Although this approach may provide correct type I error, it may also distort the interpretability of a meaningful trait or phenotype. In addition, for discrete data, such as those used in case–control genetic association studies, no such transformation may be feasible. We also note that it is rare for such transformations to be considered prior to fitting GLMs, and thus our methodology remains highly relevant.

We note that the standard density approximation is intended for unconditional inference, i.e. not conditioning on the observed **x** and **y**. Thus, it might be considered in some sense unfair to expect a close correspondence to the permutation distribution, which is inherently conditional on the data. However, the results in Figures 5, 9, and 10 of supplementary material available at *Biostatistics* online are highly consistent across independent simulations, showing that if the densities of X and Y are skewed, standard parametric p-values tend to be inaccurate *on average*. Thus, we can recommend MCC as generally preferred over standard trend testing for high-throughput datasets.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

We thank Dr Alan Agresti for pointing out the relevance of the Hauk and Donner 1977 paper described in Appendix (see supplementary material available at *Biostatistics* online). We gratefully acknowledge the CF patients, the Cystic Fibrosis Foundation, the UNC Genetic Modifier Study, and the Canadian Consortium for Cystic Fibrosis Genetic Studies, funded in part by Cystic Fibrosis Canada and by Genome Canada through the Ontario Genomics Institute per research agreement 2004-OGI-3-05, with the Ontario Research Fund-Research Excellence Program. *Conflict of Interest*: None declared.

Funding

This work was supported in part by the Gillings Statistical Genomics Innovation Lab, EPA RD83382501, NCI P01CA142538, NIEHS P30ES010126, P42ES005948, HL068890, MH101819 and DMS-1127914.

References

Breslow, N. E. and Day, N. E. (1980). Statistical methods in cancer research. Vol. 1. The analysis of case-control studies. *IARC Scientific Publications* **1**, 5–338.

Booth, J. G. and Butler, R. W. (1990). Randomization distributions and saddlepoint approximations in generalized linear models. *Biometrika* **77**(4), 787–796.

Corcoran, C., Mehta, C., Patel, N. and Senchaudhuri, P. (2001). Computational tools for exact conditional logistic regression. *Statistics in Medicine* **20**(17–18), 2723–2739.

Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data*. Boca Raton: Chapman and Hall.

Frisch, R. and Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica* **1**, 387–401.

Good, P. I. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Berlin: Springer.

Huang, S. T., Duan, S., Bleibel, W. K., Kistner, E. O., Zhang, W., Clark, T. A., Chen, T. X., Schweitzer, A. C., Blume, J. E., Cox, N. J. *and others* (2007). A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. *Proceedings of the National Academy of Sciences* **104**(23), 9758–9763.

Kennedy, P. E. and Cade, B. S. (1996). Randomization tests for multiple regression. *Communications in Statistics—Simulation and Computation* **25**(4), 923–936.

Kimmel, Gad and Shamir, Ron. (2006). A fast method for computing high-significance disease association in large population-based studies. *The American Journal of Human Genetics* **79**(3), 481–492.

Kulinskaya, E. (2008). On two-sided P-values for nonsymmetric distributions. Arxiv (0810:2124).

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Berlin: Springer.

Li, Y., Willer, C. J., Ding, J., Scheet, P. and Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *American Journal of Human Genetics* **34**(8), 816–834.

Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E. T. *and others* (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences* **102**(38), 13550–13555.

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R. and Dermitzakis, E. T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**(7289), 773–777.

Mukherjee, S., Simon, J., Bayuga, S., Ludwig, E., Yoo, S., Orlow, I., Viale, A., Offit, K., Kurtz, R. C, Olson, S. H. *and others* (2011). Including additional controls from public databases improves the power of a genome-wide association study. *Human Heredity* **72**(1), 21–34.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any populations. II. The correlation coefficient test. *Supplement to the Journal of the Royal Statistical Society* **4**, 225–232.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, J. D., Maller, S. P., de Bakker, P. I., Daly, M. J. and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**(3), 559–575.

Robinson, J. (1982). Saddlepoint approximations for permutation tests and confidence intervals. *Journal of the Royal Statistical Society* **44**(1), 91–101.

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

Takei, N., Miyashita, A., Tsukie, T., Arai, H., Asada, T., Imagawa, M., Shoji, M., Higuchi, S., Urakami, K., Kimura, H. *and others* (2009). Genetic association study on in and around the APOE in late-onset Alzheimer disease in Japanese. *Genomics* **93**(5), 441–448.

Wright, F. A, Strug, L. J., Doshi, V. K., Commander, C. W., Blackman, S. M., Sun, L., Berthiaume, Y., Cutler, D., Cojocaru, A. and Collaco, J. M. *and others* (2011). Genome-wide association and linkage identify modifier loci of lung disease severity in cystic fibrosis at 11p13 and 20q13. 2. *Nature Genetics* **43**(6), 539–546.

Zhou, Y.-H., Mayhew, G., Sun, Z., Xu, X., Zou, F. and Wright, F. A. (2013). Space-time clustering and the permutation moments of quadratic forms. *Statistics* **2**(1), 292–302.

Zhou, C., Wang, H. J. and Wang, Y. M. (2009). Efficient moments-based permutation tests. *Advances in Neural Information Processing Systems*, pp. 2277–2285.

Zhou, Y. H., Xia, K. and Wright, F. A. (2011). A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics* **27**(19), 2672–2678.

[*Received September 17, 2014; revised February 16, 2015; accepted for publication February 16, 2015*]