Fast and Accurate P-value Estimation by Fitting a Density Adaptive Distribution for Improved Computational Efficiency

ABSTRACT:

Estimating the p-value of test statistics with unknown null-distributions is a required computation in many hypothesis testing analyses. The empirical permutation method for p-value estimation is popular due to its simplicity and ability to provide accurate estimates under various null-distributions. However, its key weakness is its large computational cost which makes it unfeasible for large-scale genetic testing. Other distribution fitting techniques are preferred for smaller sample sizes of null-statistics; however, they either require a priori distribution information for fitting the null-statistics or a large sample size for fitting a subset of extreme value statistics. We propose a method that fits a density adaptive distribution (DAD) of multiple known distributions that is both precise and computationally efficient. Compared to the three existing methods under a wide range of test scenarios, the proposed DAD-fitting method exhibits superior performance in reducing p-value bias, controlling type-I error rate, and providing precise estimations under various null-distributions. Using Genetic Analysis Workshop 17 data for gene-based association analysis, the proposed method can accurately estimate p-values 10 times quicker than the empirical permutation method. With this development, new discoveries of gene correlations and better natural disaster prevention can be best achieved through the superior DAD-fitting method for p-value estimation.

Index

I. Introduction

Advances in high throughput technology have enabled the generation of abundant genomic data, which promises to unravel the genetic architecture of complex traits. The ability to map out the exact effects of each genetic mutation, paired with CRISPR genome-editing technology, could revolutionize the entire medical field. Families would no longer have to worry about miscarriages, birth defects, or the potential for diseases like cystic fibrosis or childhood cancers, because this technology would be able to both predict and correct the presence of these health issues [Prescott and Wilkie 2007]. However, there are statistical challenges to accurately and efficiently determining the genetic relevance of the available data. For example, for many hypothesis-testing applications, the distribution of a test statistic is either unknown or hard to approximate due to the complexities of its test statistic and data compositions. In such situations, the empirical permutation method, which estimates the p-value from the empirical cumulative distribution function (CDF) of a set of permutation-based null statistics, has become a popular approach due to its computational simplicity and intrinsic efficiency under various null distributions. A key weakness of this empirical method is its large computational cost in estimating the small p-values, because each additional digit of p-value precision requires ten times the sample size of null statistics [Knijnenburg et al., 2009]. For instance, the asymptotic distributions are unavailable for many test statistics used in gene-based rare variant association studies, especially in those adaptive methods designed for detecting the varied association signals [Derkach et al., 2013; Liu et al., 2013; Sha et al., 2013]. It is crucial to be able to obtain precise small p-values in the contemporary genetic association studies since the simultaneous testing of a large number of gene sets or pathways requires multiple testing corrections to

prevent excessive false positives. For example, in order to test the correlations between a susceptible single-nucleotide-polymorphism (SNP) and specific phenotype traits, we need to scan 3 billion SNPs across the whole human genome, which would take an incredible amount of computing time. The corresponding type-1 error rate threshold for this process in its current form is $1.7 \times 10^{-11} \approx (\frac{0.05}{3.0 \times 10^9})$, and the corresponding p value must be extremely small. The computational time of applying the currently favored empirical permutation method to estimating p- values of thousands and even millions of hypothesis tests has thereby become impractical in the setting of increasingly large scale genomic studies.

Alternative p-value fitting methods and shortcomings

To obtain p-values using a relatively small batch of null statistics while simultaneously reducing computational costs, several different distribution fitting methods have been previously created - each with significant shortcomings. One such method is the Gumbel-fitting method, which fits a set of Gumbel-like space-time scan statistics to a Gumbel distribution [Abrams et al., 2010]. However, the applicability of this approach to test statistics of other distributions remains unclear and needs further investigation. Another method is the GPD-fitting method, which fits a subset of extreme-value statistics to a generalized Pareto distribution (GPD) [Knijnenburg et al., 2009].

The latter method yields accurate and dependable estimates when the sample size is sufficiently large (5000 or more null statistics), but suffers from large bias and poor calculability when the sample size is small (500 or fewer null statistics). Several application-dependent procedures have also been developed, such as a p-value estimation process effective only for testing the scan statistics in linkage analysis [Shi et al., 2007] and a numerical-integration method suited solely to testing the correlated normal statistics in association analysis [Conneely

and Boehnke 2007]. These specific techniques achieve computational efficiency at the cost of application generality. Therefore, an improved method for p-value estimation is needed that remains accurate with a small sample size of null statistics, is computationally fast, and is application independent.

To meet this need, we propose to estimate the p-value by fitting a density adaptive distribution (DAD), which can uniquely continue to provide useful results even with a small to moderate number of null statistics. We evaluated the proposed method with extensive simulation studies and applied it to the Genetic Analysis Workshop 17 (GAW17) data to compare its performance with that of three existing p-value estimation methods, when applicable.

II. Methods

P-value estimation by fitting a density adaptive distribution

Let x_0 be a positive test statistic with unknown null distribution, and $X = (x_1, \cdots, x_n)$ be a set of permutation-based statistics generated under the null hypothesis. We propose to fit the null statistics X to a DAD of five known distributions: Chi-square (χ^2), Gamma, Log-normal, Weibull, and Gumbel [Krishnamoorthy 2006], with an additional adjustment from its empirical distribution (referred to hereafter as the DAD-fitting method). The probability density function of a DAD is as follows:

$$f(x|\,\Theta, \mathrm{W}) = \sum_{j=1}^{5} w_j f_j(x|a_j, b_j) \qquad (1)$$

where $\left(f_j\left(x\middle|\, a_{j,}\, b_j\right), j \in (1, \cdots, 5)\right)$ are the probability density functions of five known distributions:

$$Y_1 = \chi^2(a_1, b_1) \qquad Y_2 = Gamma(a_2, b_2) \qquad Y_3 = Lognormal(a_3, b_3)$$

$$Y_4 = Weibull(a_4, b_4) \qquad Y_5 = Gumbel(a_5, b_5)$$

respectively, $\Theta = (a_j,\ b_j,\ j \in (1, \cdots, 5)$ are the unknown distribution parameters, and $W = (w_j, j \in (1, \cdots, 5)|$ each $w_j \geq 0$ and $\sum_{j=1}^{5} w_j = 1)$ are the unknown density weights which satisfy the unit sum condition. The logic of equation (1) relies on an intuitive distribution assumption that the unknown distribution of a test statistic is either close to one of the known distributions or lying somewhere between those distributions [Cordy and Thomas 1997; Laird 1978]. These five distributions are significant mainly because of their varying distribution patterns in relation to the first four moments, and partially due to their popularity as asymptotic distributions of many known statistics. Particularly, many common tests applied in bioinformatics are based on the test statistics that are asymptotically distributed or related to these five known distributions [Breslow and Clayton 1993; Davies 1980; Lin et al., 2013; Tzeng et al., 2011; Wu et al., 2011]. The computations of the DAD-fitting method are described below as steps 1, 2 and 3.

Step 1: Parameter Estimation. We estimated the distribution parameters Θ in relation to the null statistics *X.* The four pairs of $(a_j, b_j, j \in (1, 2, 3 \text{ and } 5)) \in \Theta$ of the χ^2, Gamma, Log–Normal and Gumbel distributions are estimated via the standard moment methods [Krishnamoorthy 2006], and the parameters $(a_4, b_4) \in \Theta$ of the Weibull distribution are estimated using a moment-based correlation and coefficient of variation method [Teimouri and Gupta 2013]. The computations for deriving the five pairs of distribution parameters $\widehat{\Theta} = (\hat{a}_j,\ \hat{b}_j,\ j \in (1, \cdots, 5))$ in Table 1 at the end of the report.

Step 2: Weight Optimization. The weight parameters *W* are optimized by maximum likelihood analysis given the null statistics *X* and estimated parameters $\widehat{\Theta}$. By using the function

of (1), the log like-likelihood of weight parameters W can be expressed as:

$$L(W|\, X, \widehat{\Theta}) = \sum_{i=1}^{5} log(\sum_{j=1}^{5} w_j, \hat{f}_{ij}) \tag{2}$$

where each density value $\hat{f}_{ij} = f_j(x_i|\, \hat{a}_j,\, \hat{b}_j)$ is calculated in relation to a null statistic x_i and a pair of estimated distribution parameters $(\hat{a}_j,\, \hat{b}_j)$. To maximize the log likelihood of (2) under the unit sum condition of W, each move of one parameter w_j for a positive or negative step size $\Delta_{j(k)}$ at an iteration k (i.e., $w_{j(k)} = w_{j(k-1)+}\ \Delta_{j(k)}$) is accompanied with a proportional update of the remaining parameters (i.e., $w_{j'(k)} = w_{j'(k-1)}\,[1-\Delta_{j(k)}/(1-w_{j(k-1)})]$ for each $(j' \neq j)$. We employ a derivative-free optimization procedure of the Hooke and Jeeves algorithm [Hooke and Jeeves 1961; Kelley 1999] to obtain $\widehat{W} = (\widehat{w}_j,\ j \in (1, \cdots, 5))$ as the maximum likelihood estimates of W.

Step 3: P-value Estimation. The p-value of a test statistic x_0 is estimated from the fitted CDF of DAD obtained in step 2 and adjusted from the empirical CDF of X when applicable. To do this, let x_{min} , x_{max} , and x_{med} be the minimum, maximum, and median of X, respectively. The right tail p-value of a test statistic x_0 by the DAD-fitting method is estimated by:

$$\hat{p}_{dad} = \begin{cases} 1 - c_f, & if\ x_0 \leq\ x_{min}\ or\ x_0 \geq\ x_{max} \\ 1\ -\ [w_e c_e + (1 - w_e) c_f\,], & if\ x_{min} <\ x_0\ <\ x_{max} \end{cases} \tag{3}$$

where $c_f = \sum_{j=1}^{5} \widehat{w}_j\ \Pr(Y_j \leq x_0 |\, \hat{a}_j\,, \hat{b}_j\,)$ is the fitted CDF of DAD evaluated at $\widehat{\Theta}$, $\widehat{W}$ and x_0, $c_e = \{\sum_{i=1}^{n} I(x_0 \geq x_i) + 0.5[I(x_0 < x_{min}) - I(x_0 \geq x_{max})]\}/n$ is the empirical CDF of X at x_0 and $w_e = I(x_0 < x_{med})\left[\frac{x_0 - x_{min}}{x_{med} - x_{min}}\right] + I(x_0 \geq x_{med})[\frac{x_{max} - x_0}{x_{max} - x_{med}}]$ is an adjustment weight representing the relative importance of c_e versus c_f. The $\hat{p}_{pdad}$ value at the second line of equation (3) represents an adjustment to the fitted CDF from its empirical CDF when the

observed test statistic x_0 is within the range of null statistics (x_{min}, x_{max}).

Bias of estimated p – value

To evaluate the accuracy of p-value estimation, we define the bias statistic of an estimated p-value $\hat{p}$ relative to its true p-value $p^{(t)}$ by a scaled log difference as:

$$bias(\hat{p}|\, p^t) = \left|\frac{\log(\hat{p}/p^t)}{\log[\min(p^t,0.9)]}\right| \quad (4)$$

where the true p-value $p^{(t)}$ is either calculated from the CDF of the null distribution of estimated using a very large number of null statistics, depending on the knowledge and type of the true null distributions. The denominator of $log\ [min\ (p^{(t)},\ 0.9)\ ]$ in equation (4) is a scaled factor for ensuring the comparable bias values for all $p^{(t)} \in (0,1)$ and avoiding the infinite bias value when $p^{(t)}$ is close to 1.

Comparison methods

To evaluate the performance of the proposed DAD-fitting method, we conducted extensive simulation studies and real data analysis by comparing the performance of the DAD-fitting method with that of the three existing methods of empirical permutation [Pesarin and Salmaso 2012], Gumbel-fitting [Abrams et al., 2010], and GPD-fitting [Knijnenburg et al., 2009], when applicable. The compared performance endpoints include the estimated p-value and its bias from true p-value, along with the type-I error rate and its bias from nominal significance level.

Results Comparison of p-value accuracy

For a small-scale comparison of p-value accuracy using four scenarios (two sample sizes: 100 and 200; two null distributions: $\chi^2(6,3)$ and a weighted sum of two χ^2 distributions) and six selected true p-values (0.50, 0.05, 0.01, 10^{-3}, 10^{-4}, and 10^{-5}), the DAD-fitting method

achieved the smallest p-value biases in 67% (16/24) of cases tested, whereas the other three methods reached the smallest biases in 25% or fewer cases (Table 2, right panel). Featuring the average bias among the six true p-values as a summary statistic, the DAD-fitting method consistently showed the smallest bias values among all four methods in all four scenarios A to D (Table 2, the rows for average bias). Of the remaining three methods, the Gumbel-fitting method performed best in three scenarios A, B and C, and the GDP-fitting method performed best only in scenario D. For a more extensive comparison using 30 scenarios (five sample sizes: 100, 200, 500, 1000, and 2000; six null distributions: $\chi^2(6,3)$, Gamma(3.3750, 2.6667), Lognormal(2.0675, 0.5094), Weibull(1.9120, 10.1441), Gumbel(6.7952, 3.8197), and a weighted sum of two χ^2 distributions) and six true p-values (0.50, 0.05, 0.01, 10^3, 10^4, and 10^5), we reported the summary statistics of p-value biases for a total of 6000 replicated analyses (Table 3, lower panel.)

Comparison of type-I error rate

For type-I error rate evaluation, we considered the Gumbel-fitting and DAD-fitting methods by testing three small sample sizes (100, 200, and 500) and six null distributions $\chi^2(6,3)$ Gamma(3.3750, 2.6667), Lognormal(2.0675, 0.5094), Weibull(1.9120, 10.1441), Gumbel(6.7952, 3.8197), and a weighted sum of two randomly selected distributions. We did not consider the results of empirical permutation and GPD-fitting methods because of their poor bias performance under small sample-size settings. For the Gumbel-fitting method, the type-I error rate of estimated p-values was accurate at the significance level of 0.05 (Table 4, upper left panel) and its bias increased noticeably as the significance level decreased from 0.01 to 0.0001. For the DAD-fitting method, the type-I error rate was accurate at all four significance levels (Table 4, upper right panel), although its bias showed a modest increase as the significance level

decreased. On average across the six null distributions and as the significance level decreased from 0.05 to 0.0001, the mean biases of type-I error rate by the DAD-fitting and Gumbel-fitting methods showed a modest increase from 0.01-0.03 to 0.02-0.12 and a very large increase from 0.01-0.02 to 0.90-0.91, respectively (Table 4, lower panel).

Application to the GAW17 mini-exome sequence data

In an application to real data analysis, we utilized a GAW17 dataset that consists of 697 unrelated individuals' genotypes and phenotypes [Almasy et al., 2011]. The single nucleotide polymorphism (SNP) genotypes of the 697 individuals were obtained from a mini-exome sequence data provided by the 1000 Genomes Project Consortium (www.1000genomes.org). The Q1 phenotype used was a simulated quantitative trait influenced by nine causal genes. For an analysis of rare and uncommon variants with minor allele frequency (MAF) below 0.05, each of the nine causal genes consisted of 1-10 causal variants and 0-22 neutral variants (Table 5, column 2). To compare the performance of the empirical permutation method using 105 null statistics with that of the DAD-fitting method using 200 null statistics, we tested the SNP set associations between Q1 phenotype and nine casual genes in the replicate 1 dataset using the linear and quadratic tests as described by Derkach et al. (2013). For the FLT1 gene with a very strong association affect, the DAD fitting method identified highly significant p-values of $7.7x10^{-12}$ and $5.3x10^{-13}$ by the linear and quadratic tests, respectively. For the remaining eight genes, the p-values from the empirical permutation and DAD-fitting methods were very similar. In addition, for the GAW17 dataset with 697 individuals and a gene size of 1 to 32 variants, the computational time of p-value estimation by the DAD-fitting analysis using 200 null statistics was approximately 100 times faster than that by an empirical permutation analysis using 105 null statistics (Table 5; the ratios of total computational time in second between the empirical and

DAD-fitting methods were 44.0 / 0.5 ≈ 88 and 74.2 / 0.7 ≈ 106 when using the linear and quadratic tests, respectively). Analogously, if we ran 1 million tests using the same computational complexities (a dataset of 697 individuals, a gene size of 1-32 variants, and a test statistic similar to the linear and quadratic tests as in the above example), then the computational time of p-value estimation would be 18.5 hours when referencing the DAD fitting method with 200 null statistics, or 1824.1 hours when referencing the empirical permutation method with 10^5 null statistics.

III. Discussion

We have shown that the DAD-fitting method is effective in estimating p-values under a wide range of null distributions and exhibits superior performance in reducing p-value bias and controlling type-I error rate as compared to the three existing methods - empirical permutation, Gumbel-fitting, and GPD-fitting. Unlike the fixed distribution-fitting methods which may utilize a variety of uniform distributions [Cordy and Thomas 1997], a three-parameter Weibull distribution [Cousineau 2009] or a four-parameter generalized Lambda distribution [Fournier et al., 2007], our DAD-fitting method jointly fits five distributions from differing distribution families, making it more adaptive to the different unknown distributions. The most important benefit of the DAD-fitting method is its ability to accurately estimate small p-values using a modest number of null statistics, thus reducing computational time dramatically. As shown in Table 3, the p-value biases obtained from the DAD-fitting method using 200 null statistics are approximately equivalent to those acquired from the empirical permutation method using 2000 null statistics. In practice, the DAD-fitting method can estimate any small or large p-values via a two-stage procedure. We suggest first applying the DAD-fitting method by inputting a small number of 100 null statistics. An estimated p-value of 10^{-3} or larger can be accepted within an

upper-limit bias of 0.1 (Table 2, last column). Otherwise, if the estimated p-value is less than 10^{-3} , we can recalculate the p-value estimation by applying the DAD-fitting method and inputting a large number of 1000 null statistics. Alternatively, for more estimation accuracy at an additional computational cost, we can run the above average two-stage using 200 and 2000 null statistics. Respectively, one limitation of the current DAD-fitting method is its applicability to test only positive statistics. We can achieve an extension of this method to statistics taking both positive and negative values by fitting a DAD composed of the normal, Cauchy, logistic, and beta distributions, with each representing a different distribution pattern. Also, for specific hypothesis testing, we ensure improved p-value accuracy by fitting a DAD with other density functions in addition to the five pre-selected functions utilized in equation (1). Another consideration of the DAD-fitting method is whether we can better achieve the weight optimization of equation (2) under a balanced consideration of distribution fitting performance and computational speed. We use the Hooke and Jeeves algorithm for maximum likelihood analysis - mainly for its computational simplicity and speed - and we have no reason to expect that we can obtain the global optima of density weights in all situations. Potentially, to improve the weight optimization analysis, we could utilize a more intensive search approach, such as a simulated annealing algorithm [Press et al., 2002], for better performance at another cost of additional computations. In summary, for p-value estimations under a wide range of scenarios, the DAD-fitting method exhibits superior performance in reducing p-value bias, controlling type-I error rate, and providing precise estimations under various null distributions when compared to that of the empirical permutation, Gumbel-fitting, and GPD-fitting methods. For many hypothesis testing analyses with unknown null distributions (such as the SNP set association testing by the frequency weighted sum test [Madsen and Browning 2009], adaptive sum test

[Han and Pan 2010], variable threshold test [rice et al., 2010], C-alpha test [Ionita-Laza et al., 2011], nonlinear genotype aggregation test [Zhu et al., 2012], adaptive weighting test [Sha et al., 2013], and Fisher's combination test [Derkach et al., 2013]), the replacement of the empirical permutation method with the DAD-fitting method can greatly reduce the computational cost and thus facilitate large scale genomic studies. The DAD-fitting method has been implemented in a pDAD software and is available at https://github.com/siemensPDAD/pdad_programs [Competition Entrant 2017].

Applicability and the future

Efficient p-value testing, as we achieve through our standardized PDAD method, serves to increase the accessibility of statistic-dependent research worldwide. Any experiment performed or study conducted that requires hypotheses testing must administer a p-value test to conduct meaningful analysis of collected data, meaning our methods serve to benefit the large majority of research projects. Our standardized PDAD method is accurate, time efficient, and requires a smaller sample size than both Gumbel and GPD methods; this method hereby allows researchers to spend less time running p-value tests and more time contributing to scientific advancement. In addition, the reduced computational time needed to carry out p-value testing when one uses the PDAD method allows both individuals and smaller institutions to analyze data just as effectively as any larger institution, regardless of equipment available. In fact, our analysis method is so efficient and accessible that it can be performed entirely on a personal laptop computer, therefore providing the opportunity for advanced data analysis to be conducted without concern for excessive use of computational time or funds.

P-value testing is also highly applicable in natural disaster predictions; earthquake prediction and prevention efforts already widely use the Gumbel method [Izumiya and Ojima 2013], but serve to benefit in accuracy from usage of the PDAD method. In our increasingly unstable environment, natural disasters continue to occur more frequently and cause more severe and widespread destruction. Hurricane Harvey alone has caused more than 180 billion dollars in damages to the State of, and is expected to take over 10 years of intensive cleanup to restore the Houston metropolitan area to its former glory. Thousands of families have been displaced, and many do not have financial ability to rebuild their houses and their livelihoods. The path of Harvey was well predicted in terms of the tools widely available today, but there is currently no reliable way to ensure accurate predictions of damage for storms such as Harvey. With PDAD p-value testing, researchers would be able to conduct an enormous quantity of p-value estimations, meaning the exact costs and effects a hurricane would have on a metropolitan city could be predicted with significantly higher accuracy at significantly lower costs, allowing better preparation overall for future natural disasters. The lower processor requirements to run the DAD-fitting method means that it can be processed on a run-of-the-mill consumer laptop, eliminating the need for supercomputers or high-performance clusters. Researchers in impoverished countries can now get the same, accurate p-value tests without access to expensive computer equipment. The possibilities are truly endless with efficient p-value testing, and this statistical tool has the ability to create more accurate and useful scientific results across every discipline of science, which could serve to enhance the future of scientific discovery.

IV. Tables

Table 1. Parameter estimation for five known distributions

Distribution	Parameter	Estimation method[a]	Comments
Chi-square	$a_1 > 0$, degree of freedom $b_1 > 0$, non-centrality	$\hat{a}_1 = 2m - v/2$ $\hat{b}_1 = v/2 - m$	The estimation of (a_1, b_1) is valid for $v/4 < m \le v/2$. Otherwise, ignore this estimation and force $w_1 \equiv 0$ in equation (1).
Gamma	$a_2 > 0$, shape $b_2 > 0$, scale	$\hat{a}_2 = m^2/v$ $\hat{b}_2 = v/m$	The estimations of $(a_j, b_j, 2 \le j \le 5)$ are each valid for any $m > 0$ and $v > 0$.
Log-normal	$a_3 \in (-\infty, \infty)$, log scaled mean $b_3 > 0$, log scaled variance	$\hat{a}_3 = \log(m) - \hat{b}_3/2$ $\hat{b}_3 = \log(v/m^2 + 1)$	The log-normal distribution has a heavier right tail than the other four distributions.
Weibull	$a_4 > 0$, shape $b_4 > 0$, scale	$\hat{a}_4 = \dfrac{-\log(2)}{\log\left(1 - \dfrac{rc}{\sqrt{3}}\sqrt{\dfrac{n+1}{n-1}}\right)}$ $\hat{b}_4 = m/\Gamma(1 + 1/\hat{a}_4)$	The estimation of (a_4, b_4) is based on a moment-based method by Teimouri and Gupta (2013). $\Gamma(y)$ is the Gamma function value at $y > 0$.
Gamma	$a_5 \in (-\infty, \infty)$, location $b_5 > 0$, scale	$\hat{a}_5 = m - \gamma\hat{b}_5$ $\hat{b}_5 = \sqrt{6v}/\pi$	The circle-diameter ratio constant $\pi \approx 3.1415926535$, and the Euler-Mascheroni constant $\gamma \approx 0.5772156649$.

[a]The estimation of distribution parameters are based on a set of null statistics $(x_1, x_2, \cdots, x_n)$. The four intermediate statistics used in the analysis are the mean $m = \sum x_i/n$, variance $v = \sum (x_i - m)^2/(n-1)$, coefficient of variation $c = \sqrt{v}/m$, and the Pearson correlation coefficient of $r = \sum (x_i - m)(x_i^* - m^*)/\sqrt{\sum (x_i - m)^2 \sum (x_i^* - m^*)^2}$, where $(x_1^*, x_2^*, \cdots, x_n^*)$ are the ascending ranks of $(x_1, x_2, \cdots, x_n)$, and $m^* = \sum x_i^*/n$.

Table 2. Estimated p-value and bias for two sample sizes and two null distributions

True p-value	Statistic	Estimated p-value[a]				P-value bias[b]			
		Empirical	Gumbel	GPD	DAD	Empirical	Gumbel	GPD	DAD
Scenario A: Using $N = 100$ null statistics from a χ^2 distribution of $\chi^2(6, 3)$									
0.50	8.166	6.00×10^{-1}	5.79×10^{-1}	6.00×10^{-1}	5.97×10^{-1}	0.262	**0.212**	0.262	0.257
0.05	18.236	3.58×10^{-2}	3.79×10^{-2}	4.37×10^{-2}	3.67×10^{-2}	0.112	0.093	**0.045**	0.103
0.01	23.850	1.84×10^{-2}	1.12×10^{-2}	8.11×10^{-3}	9.93×10^{-3}	0.133	0.024	0.045	**0.002**
10^{-3}	31.178	5.00×10^{-3}	1.47×10^{-3}	9.71×10^{-5}	7.32×10^{-4}	0.233	0.056	0.338	**0.045**
10^{-4}	38.042	5.00×10^{-3}	2.20×10^{-4}	5.00×10^{-3}	5.92×10^{-5}	0.425	0.085	0.425	**0.057**
10^{-5}	44.616	5.00×10^{-3}	3.55×10^{-5}	5.00×10^{-3}	4.82×10^{-6}	0.540	0.110	0.540	**0.063**
Average bias:						0.284	0.097	0.276	**0.088**
Scenario B: Using $N = 100$ null statistics from a weighted sum of two χ^2 distributions of $0.75X + 0.25Y$, where X and Y independently follow $\chi^2(6, 3)$ (abbreviated as the WS2C distribution)									
0.50	10.084	4.71×10^{-1}	4.93×10^{-1}	4.71×10^{-1}	4.71×10^{-1}	0.087	**0.021**	0.087	0.085
0.05	18.507	6.34×10^{-2}	4.67×10^{-2}	4.83×10^{-2}	5.14×10^{-2}	0.080	0.023	0.012	**0.009**
0.01	22.943	9.46×10^{-3}	1.17×10^{-2}	4.99×10^{-3}	1.06×10^{-2}	0.012	0.035	0.151	**0.012**
10^{-3}	28.700	5.00×10^{-3}	1.92×10^{-3}	5.00×10^{-3}	1.66×10^{-3}	0.233	0.094	0.233	**0.073**
10^{-4}	34.080	5.00×10^{-3}	3.53×10^{-4}	5.00×10^{-3}	2.91×10^{-4}	0.425	0.137	0.425	**0.116**
10^{-5}	39.142	5.00×10^{-3}	7.16×10^{-5}	5.00×10^{-3}	5.66×10^{-5}	0.540	0.171	0.540	**0.151**
Average bias:						0.230	0.080	0.241	**0.074**
Scenario C: Using $N = 2000$ null statistics from the same $\chi^2(6, 3)$ distribution as in scenario A									
0.50	8.166	5.00×10^{-1}	4.94×10^{-1}	5.00×10^{-1}	5.00×10^{-1}	**0.001**	0.018	**0.001**	**0.001**
0.05	18.236	5.45×10^{-2}	5.47×10^{-2}	5.45×10^{-2}	5.53×10^{-2}	**0.029**	**0.030**	0.029	0.034
0.01	23.850	7.56×10^{-3}	1.10×10^{-2}	7.56×10^{-3}	8.49×10^{-3}	0.061	**0.021**	0.061	0.036
10^{-3}	31.178	7.81×10^{-4}	1.62×10^{-3}	6.35×10^{-4}	8.79×10^{-4}	0.036	0.070	0.066	**0.019**
10^{-4}	38.042	2.50×10^{-4}	2.67×10^{-4}	1.90×10^{-5}	9.24×10^{-5}	0.099	0.107	0.180	**0.009**
10^{-5}	44.616	2.50×10^{-4}	4.75×10^{-5}	9.11×10^{-8}	9.85×10^{-6}	0.280	0.135	0.408	**0.001**
Average bias:						0.084	0.064	0.124	**0.017**
Scenario D: Using $N = 2000$ null statistics from the same WS2C distribution as in scenario B									
0.50	9.936	5.18×10^{-1}	5.19×10^{-1}	5.18×10^{-1}	5.18×10^{-1}	**0.051**	0.055	**0.051**	**0.051**
0.05	18.507	4.50×10^{-2}	5.07×10^{-2}	4.50×10^{-2}	4.62×10^{-2}	0.036	**0.005**	0.036	0.026
0.01	24.139	1.25×10^{-2}	1.11×10^{-2}	1.25×10^{-2}	1.23×10^{-2}	0.049	**0.024**	0.049	0.044
10^{-3}	33.165	1.16×10^{-3}	7.87×10^{-4}	1.41×10^{-3}	1.08×10^{-3}	0.022	0.035	0.050	**0.011**
10^{-4}	43.667	4.02×10^{-4}	3.58×10^{-5}	1.34×10^{-4}	1.03×10^{-4}	0.151	0.112	0.032	**0.003**
10^{-5}	56.537	2.50×10^{-4}	8.10×10^{-7}	9.69×10^{-6}	4.08×10^{-6}	0.280	0.218	**0.003**	0.078
Average bias:						0.098	0.075	0.037	**0.028**

[a]The four methods for p-value estimation are labeled as "Empirical" for the empirical permutation method, "Gumbel" for the Gumbel-fitting method, "GPD" for the GPD-fitting method, and "DAD" for the proposed DAD-fitting method. [b]The smallest bias among the four methods are highlighted in bold. The average bias among all six true p-values is reported as a summary statistic for the overall bias under each scenario.

Table 3. P-value bias for five sample sizes and six null distributions

Sample size	Null distribution[a]	P-value bias[b]			
		Empirical	Gumbel	GPD	DAD
Mean ± SD of p-value biases among six true p-values					
100	χ^2	0.286 ± 0.177	**0.098 ± 0.066**	0.302 ± 0.529	0.166 ± 0.217
100	Gm	0.286 ± 0.178	**0.087 ± 0.064**	0.286 ± 0.484	0.158 ± 0.186
100	Ln	0.289 ± 0.176	0.259 ± 0.194	0.312 ± 0.560	**0.163 ± 0.214**
100	Wb	0.288 ± 0.175	0.196 ± 0.107	0.281 ± 0.245	**0.141 ± 0.100**
100	Gb	0.289 ± 0.176	**0.090 ± 0.069**	0.290 ± 0.401	0.174 ± 0.220
100	WS2C	0.286 ± 0.178	**0.139 ± 0.097**	0.273 ± 0.272	0.148 ± 0.147
200	χ^2	0.238 ± 0.160	**0.092 ± 0.061**	0.240 ± 0.393	0.118 ± 0.104
200	Gm	0.234 ± 0.164	**0.072 ± 0.052**	0.276 ± 0.615	0.114 ± 0.086
200	Ln	0.238 ± 0.161	0.248 ± 0.178	0.312 ± 0.729	**0.112 ± 0.107**
200	Wb	0.237 ± 0.162	0.194 ± 0.106	0.236 ± 0.256	**0.113 ± 0.086**
200	Gb	0.232 ± 0.164	**0.057 ± 0.045**	0.237 ± 0.307	0.100 ± 0.096
200	WS2C	0.238 ± 0.161	0.129 ± 0.091	0.251 ± 0.385	**0.119 ± 0.111**
500	χ^2	0.168 ± 0.149	0.083 ± 0.051	0.208 ± 0.478	**0.072 ± 0.058**
500	Gm	0.167 ± 0.148	**0.063 ± 0.041**	0.180 ± 0.425	0.078 ± 0.059
500	Ln	0.166 ± 0.147	0.223 ± 0.152	0.210 ± 0.757	**0.061 ± 0.047**
500	Wb	0.168 ± 0.149	0.190 ± 0.104	0.206 ± 0.438	**0.076 ± 0.062**
500	Gb	0.166 ± 0.147	**0.037 ± 0.028**	0.185 ± 0.441	0.068 ± 0.055
500	WS2C	0.164 ± 0.148	0.119 ± 0.083	0.192 ± 0.248	**0.081 ± 0.068**
1000	χ^2	0.138 ± 0.117	0.079 ± 0.050	0.166 ± 0.454	**0.057 ± 0.049**
1000	Gm	0.139 ± 0.117	0.059 ± 0.038	0.197 ± 0.702	**0.058 ± 0.049**
1000	Ln	0.139 ± 0.117	0.216 ± 0.143	0.124 ± 0.197	**0.044 ± 0.037**
1000	Wb	0.138 ± 0.117	0.192 ± 0.105	0.153 ± 0.229	**0.062 ± 0.054**
1000	Gb	0.137 ± 0.117	**0.027 ± 0.021**	0.129 ± 0.196	0.048 ± 0.040
1000	WS2C	0.138 ± 0.117	0.114 ± 0.080	0.154 ± 0.364	**0.066 ± 0.055**
2000	χ^2	0.106 ± 0.098	0.076 ± 0.049	0.119 ± 0.224	**0.043 ± 0.041**
2000	Gm	0.107 ± 0.098	0.055 ± 0.036	0.119 ± 0.327	**0.045 ± 0.040**
2000	Ln	0.109 ± 0.098	0.219 ± 0.140	0.109 ± 0.328	**0.031 ± 0.026**
2000	Wb	0.108 ± 0.097	0.190 ± 0.104	0.115 ± 0.169	**0.041 ± 0.039**
2000	Gb	0.108 ± 0.098	**0.019 ± 0.015**	0.116 ± 0.331	0.036 ± 0.030
2000	WS2C	0.107 ± 0.098	0.117 ± 0.082	0.138 ± 0.527	**0.056 ± 0.054**
Mean ± SD of p-value biases among six null distributions and six true p-values					
100	All six	0.287 ± 0.177	**0.145 ± 0.126**	0.291 ± 0.432	0.158 ± 0.186
200	All six	0.236 ± 0.162	0.132 ± 0.121	0.259 ± 0.479	**0.113 ± 0.099**
500	All six	0.166 ± 0.148	0.119 ± 0.110	0.197 ± 0.488	**0.073 ± 0.059**
1000	All six	0.138 ± 0.117	0.115 ± 0.109	0.154 ± 0.401	**0.056 ± 0.048**
2000	All six	0.108 ± 0.098	0.113 ± 0.109	0.119 ± 0.337	**0.042 ± 0.040**

[a]The null distributions are labeled as "χ^2" for $\chi^2(6, 3)$, "Gm" for Gamma(3.3750, 2.6667), "Ln" for Lognormal(2.0675, 0.5094), "Wb" for Weibull(1.9120, 10.1441), and "Gb" for Gumbel(6.7952, 3.8197). The distribution of "WS2C" refers to a weighted sum of two χ^2 distributions of $0.75X + 0.25Y$, where X and Y follow two independent distributions of $\chi^2(6, 3)$. "All six" refers to the mean ± SD results in all six null distributions. Of note, the "χ^2" and "WS2C" distributions in this table are the same as those in Table 2. [b]See footnote a in Table 2 for the meaning of "Empirical", "Gumbel", "GPD", and "DAD". Each mean ± SD of p-value bias in the upper panel is estimated from 1000 replicated analyses among the six true p-values of 0.50, 0.05, 0.01, 10^{-3}, 10^{-4}, and 10^{-5} under a specified sample size and null distribution. Each mean ± SD in the lower panel is estimated from 6000 replicated analyses among the six null distributions and six true p-values under a specified sample size. The smallest p-value bias among the four methods are highlighted in bold.

Table 4. Type I error rate and bias for three small sample sizes and six null distributions

Sample size	Null distribution[a]	Gumbel-fitting method at significance level 0.05	0.01	0.001	0.0001	DAD-fitting method at significance level 0.05	0.01	0.001	0.0001
Type I error rate[b]									
100	χ^2	5.4×10^{-2}	9.8×10^{-3}	7.2×10^{-4}	4.6×10^{-5}	5.5×10^{-2}	1.2×10^{-2}	1.6×10^{-3}	2.9×10^{-4}
100	Gm	5.5×10^{-2}	1.1×10^{-2}	9.1×10^{-4}	7.4×10^{-5}	5.5×10^{-2}	1.2×10^{-2}	1.6×10^{-3}	2.9×10^{-4}
100	Ln	5.6×10^{-2}	1.6×10^{-2}	3.1×10^{-3}	7.3×10^{-4}	5.5×10^{-2}	1.3×10^{-2}	2.2×10^{-3}	4.5×10^{-4}
100	Wb	5.1×10^{-2}	5.6×10^{-3}	1.2×10^{-4}	0	5.2×10^{-2}	9.2×10^{-3}	9.2×10^{-4}	1.3×10^{-4}
100	Gb	5.3×10^{-2}	1.2×10^{-2}	1.4×10^{-3}	1.8×10^{-4}	5.4×10^{-2}	1.3×10^{-2}	2.2×10^{-3}	4.3×10^{-4}
100	WS2D	5.1×10^{-2}	8.7×10^{-3}	7.6×10^{-4}	1.0×10^{-4}	5.4×10^{-2}	1.2×10^{-2}	1.8×10^{-3}	3.5×10^{-4}
200	χ^2	5.3×10^{-2}	9.1×10^{-3}	6.1×10^{-4}	3.1×10^{-5}	5.3×10^{-2}	1.1×10^{-2}	1.2×10^{-3}	1.8×10^{-4}
200	Gm	5.3×10^{-2}	9.9×10^{-3}	7.8×10^{-4}	6.3×10^{-5}	5.3×10^{-2}	1.1×10^{-2}	1.3×10^{-3}	1.8×10^{-4}
200	Ln	5.4×10^{-2}	1.5×10^{-2}	2.8×10^{-3}	6.5×10^{-4}	5.3×10^{-2}	1.2×10^{-2}	1.6×10^{-3}	2.4×10^{-4}
200	Wb	4.9×10^{-2}	5.1×10^{-3}	9.0×10^{-5}	0	5.1×10^{-2}	9.3×10^{-3}	7.9×10^{-4}	8.6×10^{-5}
200	Gb	5.2×10^{-2}	1.1×10^{-2}	1.2×10^{-3}	1.4×10^{-4}	5.3×10^{-2}	1.2×10^{-2}	1.6×10^{-3}	2.2×10^{-4}
200	WS2D	4.9×10^{-2}	8.0×10^{-3}	6.6×10^{-4}	8.7×10^{-5}	5.3×10^{-2}	1.1×10^{-2}	1.3×10^{-3}	2.0×10^{-4}
500	χ^2	5.2×10^{-2}	8.7×10^{-3}	5.6×10^{-4}	2.1×10^{-5}	5.2×10^{-2}	1.1×10^{-2}	1.0×10^{-3}	9.5×10^{-5}
500	Gm	5.3×10^{-2}	9.5×10^{-3}	7.1×10^{-4}	4.6×10^{-5}	5.1×10^{-2}	1.1×10^{-2}	1.0×10^{-3}	1.1×10^{-4}
500	Ln	5.2×10^{-2}	1.4×10^{-2}	2.6×10^{-3}	5.8×10^{-4}	5.1×10^{-2}	1.1×10^{-2}	1.2×10^{-3}	1.4×10^{-4}
500	Wb	4.9×10^{-2}	4.8×10^{-3}	8.0×10^{-5}	0	5.1×10^{-2}	9.6×10^{-3}	7.7×10^{-4}	7.3×10^{-5}
500	Gb	5.1×10^{-2}	1.0×10^{-2}	1.1×10^{-3}	1.2×10^{-4}	5.1×10^{-2}	1.1×10^{-2}	1.2×10^{-3}	1.4×10^{-4}
500	WS2D	4.8×10^{-2}	7.7×10^{-3}	5.8×10^{-4}	8.1×10^{-5}	5.1×10^{-2}	1.1×10^{-2}	1.0×10^{-3}	1.2×10^{-4}
Bias of type I error rate[c]									
100	All six	0.02 ± 0.01	0.05 ± 0.05	0.10 ± 0.11	0.90 ± 2.01	0.03 ± 0.01	0.04 ± 0.02	0.08 ± 0.04	0.12 ± 0.05
200	All six	0.01 ± 0.01	0.05 ± 0.05	0.12 ± 0.12	0.91 ± 2.01	0.02 ± 0.01	0.03 ± 0.01	0.05 ± 0.02	0.07 ± 0.03
500	All six	0.01 ± 0.00	0.06 ± 0.06	0.12 ± 0.13	0.91 ± 2.00	0.01 ± 0.00	0.02 ± 0.01	0.02 ± 0.02	0.02 ± 0.01

[a]The null distributions are labeled as "χ^2" for $\chi^2(6, 3)$, "Gm" for Gamma(3.3750, 2.6667), "Ln" for Lognormal(2.0675, 0.5094), "Wb" for Weibull(1.9120, 10.1441), and "Gb" for Gumbel(6.7952, 3.8197). The null distribution of "WS2D" is a weighted sum of two distributions of $0.75X + 0.25Y$, where X and Y follow two independent distributions randomly chosen from the five distributions of χ^2, Gm, Ln, Wb, and Gb. "All six" refers to the mean ± SD results in all six null distributions, which represents an overall type I error rate under different null distributions. [b]Each type I error rate at each significance level is estimated from 10^6 replicated analyses under each null distribution, respectively. [c]Each reported statistic is the mean ± SD bias of type I error rate for all six null distributions combined. The bias of type I error rate under a null distribution is defined as $\text{bias}(\alpha|\beta) = |\log[\max(\alpha, \beta^6)/\beta]/\log(\beta)|$, where α is the type I error rate, β is a corresponding nominal significance level, and β^6 is an ad hoc lower limit of type I error rate to avoid the infinite bias value in the case where α is equal or very close to 0 (as occurred with the Gumbel-fitting method for the null Weibull distribution and significance level 0.0001).

Table 5. Comparison of the empirical permutation and DAD-fitting methods for estimated p-values using the linear and quadratic tests[a]

Gene name	No. of variants[b]	P-value of linear test		P-value of quadratic test	
		Emp.	DAD	Emp.	DAD
ELAVL4	8, 2	0.1219	0.0959	0.0060	0.0037
ARNT	17, 5	0.0008	0.0002	0.1432	0.1296
KDR	15, 9	$< 5.0\times10^{-6}$	7.3×10^{-6}	2.0×10^{-4}	5.6×10^{-4}
VEGFC	1, 1	0.0007	0.0003	0.0007	0.0003
FLT4	10, 2	0.1338	0.1258	0.5866	0.5714
VEGFA	6, 1	0.0020	0.0017	0.0287	0.0258
FLT1	32, 10	$< 5.0\times10^{-6}$	7.7×10^{-12}	$< 5.0\times10^{-6}$	5.3×10^{-13}
HIF1A	8, 4	0.0206	0.0184	0.1399	0.1755
HIF3A	17, 3	0.6555	0.6632	0.1867	0.1951
Total time in second		44.0	0.5	74.2	0.7

[a]The analyzed data is the GAW17 replicate 1 dataset that consists of 697 unrelated individuals and their Q1 phenotype and genotypes for nine causal genes. The estimated p-values calculated by the empirical permutation and DAD-fitting methods (labeled as "Emp." and "DAD") are based on 10^5 and 200 null statistics, respectively. The linear and quadratic tests for testing the SNP set association are based on the methods described in Derkach et al. (2013). [b]The reported numbers are the number of all variants and number of causal variants whose MAFs are less than 0.05 and their locations are within each analyzed gene.

References

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061-1073.

Abrams AM, Kleinman K, Kulldorff M. 2010. Gumbel based p-value approximations for spatial scan statistics. *Int. J. Health Geographics* 9:61.

Almasy L, Dyer T, Peralta J, Kent JJ, Charlesworth J, Curran J, Blangero J. 2011. Genetic Analysis Workshop 17 mini-exome simulation. *BMC Proc.* 5:S2.

Breslow KN, Clayton DG. 1993. Approximate inference in generalized linear mixed models. *J. Am. Stat.Assoc.* 88:9-25.

Competition Entrant 2017, pDAD Github repository. https://github.com/siemensPDAD/pdad_programs. Username: siemensPDAD Password:siemens1234

Conneely KN, Boehnke M. 2007. So many correlated tests, so little time! Rapid adjustment of p values for multiple correlated tests. *Am. J. Hum. Genet.* 81:1158-1168.

Cordy CB, Thomas DR. 1997. Deconvolution of a distribution function. *J. Am. Stat. Assoc.* 92:1459- 1465.

Cousineau D. 2009. Fitting the three-parameter weibull distribution: review and evaluation of existing and new methods. *Dielectrics and Electrical Insulation* 16:281-288.

Davies RB. 1980. The distribution of a linear combination of chi-square random variables. *Applied Statistics* 29:323-333.

Derkach A, Lawless JF, Sun L. 2013. Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genet Epidemiol* 37:110-121.

Fournier B, Rupin N, Bigerelle M, Najjar D, Iost A, Wilcox R. 2007. Estimating the

parameters of a generalized lambda distribution. *Comput. Stat. Data Anal.* 51:2813-2835.

Han F, Pan W. 2010. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum. Hered.* 70:42-54.

Hooke R, Jeeves TA. 1961. Direct search solution of numerical and statistical problems. *J. Asso. Comp.Machinery* 8:212-229.

Ionita-Laza I, Buxbaum JD, Laird NM, Lange C. 2011. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet.* 7:e1001289.

Izumiya, T., Uchiyama, S., & Ojima, Y. (2013). Estimaton Of Earthquake Occurrence Probability Using The Gutenberg-Richter Relation. *Journal of Japan Society of Civil Engineers, Ser. B3 (Ocean Engineering), 69*(2). doi:10.2208/jscejoe.69.i_431

Kelley CT. 1999. Iterative methods for optimization. *Soc. Indust. Appl. Math.* 18.

Knijnenburg TA, Wessels LF, Reinders MJ, Shmulevich I. 2009. Fewer permutations, more accurate P- values. *Bioinformatics* 25:i161-168.

Krishnamoorthy K. 2006. *Handbook of statistical distributions with applications*. New York: Chapman and Hall/CRC.

Laird N. 1978. Nonparametric Maximum Likelihood Estimation of a Mixing Distribution. *Am. J. Hum. Genet.* 73:805-811.

Lin X, Lee S, Christiani DC, Lin X. 2013. Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics* 14:667-681.

Liu K, Fast S, Zawistowski M, Tintle NL. 2013. A geometric framework for evaluating rare variant tests of association. *Genet. Epidemiol.* 37:345-357.

Madsen BE, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 5:e1000384.

Pesarin F, Salmaso L. 2012. A review and some new results on permutation testing for multivariate problems. *Stat. Comput.* 22:639-646.

Prescott, K. R., & Wilkie, A. O. M. (2007). Genetic aspects of birth defects: new understandings of old problems. *Archives of Disease in Childhood. Fetal and Neonatal Edition*, *92*(4), F308–F314.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 2002. *Numerical Recipes in C++*. New York: Cambridge University Press.

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. 2010. Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* 86:832-838.

Sha Q, Wang S, Zhang S. 2013. Adaptive clustering and adaptive weighting methods to detect disease associated rare variants. *Eur. J. Hum. Genet.* 21:332-337.

Shi J, Siegmund D, Yakir B. 2007. Importance sampling for estimating p values in linkage analysis. *J. Am. Stat. Assoc.* 102:929-937.

Teimouri M, Gupta AK. 2013. On the three-parameter Weibull distribution shape parameter estimation. *J. Data Sci.* 11:403-414.

Tzeng JY, Zhang DW, Pongpanich M, Smith C, McCarthy MI, Sale MM, Worrall BB, Hsu FC, Thomas DC, Sullivan PF. 2011. Studying gene and gene-environment effects of uncommon and common variants on continuous traits: A marker-set approach using gene-trait similarity regression. *Am. J. Hum. Genet.* 89:277-288.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for

sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89:82-93.

Zhu H, Li L, Zhou H. 2012. Nonlinear dimension reduction with Wright-Fisher kernel for genotype aggregation and association mapping. *Bioinformatics* 28:i375-i381.