PLOS ONE

RESEARCH ARTICLE

An Optimal Bahadur-Efficient Method in Detection of Sparse Signals with Applications to Pathway Analysis in Sequencing Association Studies

Hongying Dai[1,2]*, Guodong Wu[3], Michael Wu[4], Degui Zhi[5]

1 Health Services and Outcomes Research, Children's Mercy Hospital, Kansas City, MO, United States of America, **2** Department of Biomedical & Health Informatics, University of Missouri-Kansas City, Kansas City, MO, United States of America, **3** Lovelace Respiratory Research Institute, Albuquerque, New Mexico, United States of America, **4** Biostatistics and Biomathematics Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, United States of America, **5** Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, United States of America

* hdai@cmh.edu

Citation: Dai H, Wu G, Wu M, Zhi D (2016) An Optimal Bahadur-Efficient Method in Detection of Sparse Signals with Applications to Pathway Analysis in Sequencing Association Studies. PLoS ONE 11(7): e0152667. doi:10.1371/journal.pone.0152667

Editor: Zhi Wei, New Jersey Institute of Technology, UNITED STATES

Received: November 12, 2015

Accepted: March 17, 2016

Published: July 5, 2016

Data Availability Statement: Case study data are available from http://csg.sph.umich.edu//abecasis/public/lipids2013/.

Funding: This work is supported by NHGRI of the National Institutes of Health under award number R01HG008115 (D.Z.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Next-generation sequencing data pose a severe curse of dimensionality, complicating traditional "single marker—single trait" analysis. We propose a two-stage combined p-value method for pathway analysis. The first stage is at the gene level, where we integrate effects within a gene using the Sequence Kernel Association Test (SKAT). The second stage is at the pathway level, where we perform a correlated Lancaster procedure to detect joint effects from multiple genes within a pathway. We show that the Lancaster procedure is optimal in Bahadur efficiency among all combined p-value methods. The Bahadur efficiency, $\lim_{\varepsilon \to 0} N^{(2)}/N^{(1)} = \phi_{12}(\theta)$, compares sample sizes among different statistical tests when signals become sparse in sequencing data, i.e. $\varepsilon \to 0$. The optimal Bahadur efficiency ensures that the Lancaster procedure asymptotically requires a minimal sample size to detect sparse signals ($P_{N^{(i)}} < \varepsilon \to 0$). The Lancaster procedure can also be applied to meta-analysis. Extensive empirical assessments of exome sequencing data show that the proposed method outperforms Gene Set Enrichment Analysis (GSEA). We applied the competitive Lancaster procedure to meta-analysis data generated by the Global Lipids Genetics Consortium to identify pathways significantly associated with high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, triglycerides, and total cholesterol.

Introduction

Next-generation sequencing (NGS) technology has opened a new era for studying genetic associations with complex diseases. Yet, although whole genome searching has become easier and less costly to perform, our ability to critically evaluate such high throughput data has not

improved substantially. Sequencing data often contain millions of genetic variants. However, testing millions of markers using the "single marker—single trait" analysis often loses power after the multiple-testing adjustment. Genome-wide significance requires strict Bonferroni correction with p-value $< 2.5\times10^{-6}$ for a total of 20,000 gene-based statistical tests. To maintain statistical power of detecting rare variants, a theoretical sample size of n>10,000 may be required for sequencing data [1].

These dimensional challenges motivate us to aggregate effects from multiple genes using pathway analysis. Genetic pathways comprise molecular entities that interact with each other to regulate specific cell functions, metabolic processes, biosynthesis, and embryonic developments. For non-Mendelian diseases and complex traits, multiple genetic risk factors may function together in the pathway. As a result, signals may not be significant in the "single marker-single trait" analysis, but many such values from related genes might provide valuable information regarding gene function and regulation. The pathway information can be extracted from bioinformatic resources, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [2], the PANTHER classification system for protein sequence data [3], and Reactome database for human pathway data [4].

We propose a two-stage combined p-value method for pathway (gene set) analysis of NGS data. The first stage is at the gene level, where we integrate effects from rare and common variants within a gene. The goal of the first stage analysis is to generate a p-value that summarizes an overall effect within a gene. The second stage is at the pathway level, where we aggregate p-values among all genes in a pathway.

An exome sequencing simulation study was conducted to compare the SKAT-Lancaster procedure and Gene Set Enrichment Analysis (GSEA) [5]. We applied the competitive Lancaster procedure to meta-analysis data generated by the Global Lipids Genetics Consortium.

Methods

Two-Stage Pathway Analysis for Sequencing Data

There is a different nature of effects between gene and pathway. At the gene level, we are interested in identifying *rare* genetic variants from high throughput data. At the pathway level, genes with similar functions work together to fulfill biological tasks. Thus, we are interested in detecting *small and common* effects among genes. The proposed "SKAT-Lancaster" procedure provides a two-stage framework in order to (1) reduce the dimension of genetic variants, (2) combine effects from multiple genes, and (3) take genetic correlation architecture into account.

Stage I—Gene Level Testing. In the first stage, we suggest integrating effects from rare variants within the i^{th} gene using the Sequence Kernel Association Test (SKAT) [6]. Several tests have been proposed to analyze rare variants at the gene level, including burden tests and the C-alpha test. We choose SKAT because SKAT has been proven to be a locally most powerful score test [7].

Let G_{ij} be the j^{th} variant of the i^{th} gene. Let $\beta_i = (\beta_{i1}, \cdots, \beta_{ij}, \cdots)$ be the effects from markers in the i^{th} gene. Generate a p-value, P_i for the i^{th} hypothesis testing $H_{0i} : \beta_i = \vec{0}$ vs. $H_{ai} : \beta_i \neq \vec{0}$ in the i^{th} gene. A $\rightarrow$ is added to denote the zero vector. SKAT is a locally most powerful score test on the variance component of a regression model $Y = \alpha^t X + \beta_i^t G_i + \varepsilon$, where Y is a phenotype, α is a vector of fixed effects from covariates X, and ε is an error term. To increase the power, SKAT tests H_{0i}: $\beta_i = 0$ by treating β_{ij} as a random variable with mean zero and variance $w_{ij}\tau_i$, where τ_i is a common variance component and w_{ij} is a pre-specified weight for variant G_{ij}. As a result,$H_{0i} : \beta_i = \vec{0}$ is equivalent to H_{0i}: $\tau_i = 0$. The variance component score statistic is $Q = (Y - \hat{\mu})^t G_i W_i G_i^t (Y - \hat{\mu})$, where $\hat{\mu} = \hat{\alpha}^t X$ is the predicted mean of Y under H_{0i}, and

$W_i = diag(w_{i1}, \cdots)$ are the weights of the variants. Under the null hypothesis, Q follows a mixture of chi-square distributions [6].

Common variants, population stratification, and other covariates can also be included as fixed effects in the model. The goal of the first stage analysis is to generate a p-value that summarizes the overall effect for each gene.

Stage II—Pathway Level P-value Combination. The second stage is at the pathway level, where we perform the modified Lancaster procedure to combine effects from multiple genes within a pathway. We choose the Lancaster procedure because it is optimal in Bahadur efficiency among all weighted combined p-value methods. The original Lancaster procedure is based on the independent p-value assumption. However, genetic data are highly correlated and ignoring the correlation structure will severely inflate the Type I error rate. Thus we need a modification of the Lancaster procedure to take the complex correlation structure among genetic variants into account [8].

Consider m sequences of test statistics, $\{T_{n_i}^{(i)}\},\ i = 1,\ 2,\ \cdots,\ m$ and the corresponding significance levels,$\{P_{n_i}^{(i)}\}$, where n_i is the sample size for the i^{th} test statistic. Let the Lancaster statistic $T_n^{Lancaster} = \sum_{i=1}^{m} F_i^{-1}(1 - P_{n_i}^{(i)})$, where $P_{n_i}^{(i)}$ F_i^{-1} is the inverse cumulative distribution function (CDF) of $\chi^2_{w_i}$ with $w_i > 0$ for $i = 1, 2, \cdots, m$. When p-values are correlated, $T_n^{Lancaster}$ does not follow $\chi^2_{\sum_{i=1}^{m} w_i}$. The null distribution of $T_n^{Lancaster}$ does not have an explicit analytical form. To address this issue, we approximate the $T_n^{Lancaster}$ statistic with a scaled chi-square distribution. Let $T_n^{Lancaster} \approx c\chi^2_\nu$ where $c > 0$ is a scalar and $\nu > 0$ is the degrees of freedom for the approximate chi-square distribution. Under H_0: $\theta \in \Theta_0$, we have

$$E(T_n^{Lancaster}) = E\left(\sum_{i=1}^{m} F_i^{-1}(1 - P_{n_i}^{(i)})\right) = \sum_{i=1}^{m} w_i$$

and

$$\mathrm{var}(T_n^{Lancaster}) = \sum_{i=1}^{m} \mathrm{var}(F_i^{-1}(1 - P_{n_i}^{(i)})) + 2\sum_{i<j} \mathrm{cov}(F_i^{-1}(1 - P_{n_i}^{(i)}), F_j^{-1}(1 - P_{n_j}^{(j)})) = 2\sum_{i=1}^{n} w_i + 2\sum_{i<j} \rho_{ij},$$

where $\rho_{ij} = \mathrm{cov}(F_i^{-1}(1 - P_{n_i}^{(i)}), F_j^{-1}(1 - P_{n_j}^{(j)}))$ takes the correlation among p-values into account. We use the Satterthwaite method to match the mean and variance of $T_n^{Lancaster}$ and $c\chi^2_\nu$, and solve the equations to derive c and ν. Thus we have $T_n^{Lancaster} \approx c\chi^2_\nu$, where $c = 0.5\mathrm{var}(T_n^{Lancaster})/E(T_n^{Lancaster})$ and $\nu = 2[E(T_n^{Lancaster})]^2/\mathrm{var}(T_n^{Lancaster})$.

As genetic variants have very complex correlation architecture, there is no analytical form for the exact correlated Lancaster procedure. The Satterthwaite approximation is an effective approach to summarize the distribution of the exact correlated Lancaster procedure. Q-Q plots from simulated data suggest a good match between the approximated $T_n^{Lancaster}$ and exact $T_n^{Lancaster}$, with a very slight deviation in the tail part. By introducing the correlation structure, the Satterthwaite approximation can significantly reduce the Type I error among correlated p-values.

Self Contained vs. Competitive Lancaster Procedure

The main difference between competitive and self-contained tests lies in the formulation of the null hypothesis [9]. Let μ_i stand for the effect size from the i^{th} pathway. The null hypothesis for

the self-contained test of the i^{th} pathway is $H_{0,\ \text{self-contained}}$: $\mu_i = 0$. Thus, the correlated Lancaster procedure can be considered as a self-contained test.

The null hypothesis in the competitive test is $H_{0,\ \text{competitive}}$: $\mu_1 = \mu_2 = \cdots = \mu_i = \cdots$ The competitive Lancaster test can be carried out using permutation testing:

Step 1: Let P_i be the p-value from the Lancaster procedure in the i^{th} real pathway.

Step 2: Create L, say 100000, permutated pathways by shuffling genes among pathways. The permutated pathway sizes should resemble the real pathway sizes. Let P^l be the p-value from the Lancaster procedure in the l^{th} permutated pathway for $l = 1, 2, \cdots, L$.

Step 3: The p-value of the competitive Lancaster procedure is $\sum_{l=1}^{L} I\{P_i \geq P^l\}/L$ for the i^{th} real pathway, where $I\{.\}$ is an indicator function.

Meta-analysis in Sequencing Association Studies

Due to cost, the rarity of diseases involved, and high dimensionality of variants, sequencing association studies are often underpowered to detect modest genetic effects. Meta-analysis can be used to address this issue by analyzing data across studies. Meta-analysis uses study-specific summary statistics, allowing investigators to combine information across studies when individual-level data cannot be shared.

The Lancaster procedure is independent from the SKAT test. One can directly apply the Lancaster procedure to meta-analysis, as we demonstrated in our analysis of the Global Lipids Genetics Consortium data. In this work, we choose SKAT to pair with the Lancaster procedure in order to detect rare variants in exome sequencing data. For other types of sequencing data, we suggest replacing SKAT with other statistical tests, such as FaST-LMM [10] or GEMMA [11], at the gene level and then applying the Lancaster procedure to combine multiple effects at the pathway level.

Lancaster Procedure Is Optimal in Bahadur Efficiency

Several weighted combined p-value methods have been developed. See [12] for a comprehensive review. Since high throughput sequencing data pose a severe challenge in retaining the statistical power for small sample sizes in detection of sparse signals, it is critical to theoretically assess the efficiency among the weighted combined p-value methods. Let P_i, $(i = 1, 2, \cdots, m)$ be p-values from m hypothesis tests. Littell and Folks [13, 14] showed that Fisher's method of combining independent tests ($T^{Fisher} = -2\sum_{i=1}^{m} \ln(P_i)$) is asymptotically Bahadur efficient. However, Fisher's method does not allow a weight function when combining p-values.

The weight function can be used to integrate multiple-source omics data from varying sequencing platforms. For instance, one can apply weight functions to integrate microarray data and CHIP-TIE data to identify the protein involved in transcription. In this case, weight functions can be considered as prior information to ensure the binding calling is a real signal instead of an artifact. As [15] pointed out, carefully chosen weights can generally improve power for a combination of p-values.

There is no uniformly most powerful method of combining p-values. The Bahadur efficiency is an important way to compare sample sizes required by two statistics in detection of sparse signals ($\varepsilon \to 0$).

The Notation of Bahadur Relative Efficiency. Consider a hypothesis test for H_0: $\theta \in \Theta_0$ vs. H_a: $\theta \in \Theta - \Theta_0$. Bahadur efficiency offers an asymptotic relative comparison between two

competing test statistics. Under H_a, a test statistic whose significance level converges to zero at a faster rate is considered more Bahadur efficient.

Let T_n be a real valued test statistic depending on an independent sample, $x_1, x_2, \cdots, x_n$ for $n = 1, 2, \cdots$ Assume for all $\theta \in \Theta_0$, T_n follows the same null CDF F_0. Let t be the value attained by T_n, then the significance level of T_n is $P_n = 1 - F_0(t)$. Suppose that $-2\ln P_n / n$ converges to $c(\theta)$ with probability 1, i.e.,

$$\Pr(\lim_{n\to\infty} -(2/n)\ln P_n = c(\theta)) = 1$$

for some $c(\theta) > 0$ under H_a: $\theta \in \Theta - \Theta_0$. The value $c(\theta)$ is dependent on θ under the alternative hypothesis and $c(\theta)$ is called the Bahadur efficiency slope of T_n when $n \to \infty$. Consider two competing sequences of test statistics, $\{T_n^{(1)}\}$ and $\{T_n^{(2)}\}$, with the Bahadur efficiency slopes $c_1(\theta)$ and $c_2(\theta)$, respectively. The ratio

$$\phi_{12}(\theta) = c_1(\theta)/c_2(\theta)$$

is the Bahadur efficiency of $\{T_n^{(1)}\}$ relative to $\{T_n^{(2)}\}$. Let $N^{(i)}$ be the minimal sample size satisfying $P_{N^{(i)}} < \varepsilon$ for the i^{th} test. Bahadur [16] shows that

$$\lim_{\varepsilon\to 0} N^{(2)}/N^{(1)} = \phi_{12}(\theta)$$

with probability 1 under H_a: $\theta \in \Theta - \Theta_0$, which indicates that the Bahadur efficiency ratio $\phi_{12}(\theta)$ gives the limiting ratio of sample sizes required by the two statistics to attain an equally small significance level. As a result, $\{T_n^{(1)}\}$ is deemed superior to, i.e. more Bahadur efficient than, $\{T_n^{(2)}\}$ if $\phi_{12}(\theta) \geq 1$ under H_a: $\theta \in \Theta - \Theta_0$.

Bahadur Efficiency for Lancaster Procedure, Weighted Z-test, and Good's test. Consider m sequences of test statistics, $\{T_{n_i}^{(i)}\},\ i = 1,\ 2,\ \cdots,\ m$ and the corresponding significance levels,$\{P_{n_i}^{(i)}\}$, where n_i is the sample size for the i^{th} test statistic. Assume that for each $i = 1, 2, \cdots, m$, the sequence $\{T_{n_i}^{(i)}\}$ has a Bahadur efficiency slope $c_i(\theta)$. That is, $\Pr\left(\lim_{n_i\to\infty} -(2/n_i)\ln P_{n_i}^{(i)} = c_i(\theta)\right) = 1$ for some $c_i(\theta) > 0$ under H_a: $\theta \in \Theta - \Theta_0$. Assume also that the sample sizes $n_1, \cdots, n_m$ have an average sample size $n = (n_1 + \cdots + n_m) / m$ and $\lim_{n\to\infty} n_i/n = \lambda_i$ for $i = 1, 2, \cdots, m$. Then we have $\Pr\left(\lim_{n\to\infty} -(2/n)\ln P_{n_i}^{(i)} = \lambda_i c_i(\theta)\right) = 1$. For each n, it is desired to combine the m statistics $T_{n_1}^{(1)},\ \cdots,\ T_{n_m}^{(m)}$ into an overall test statistic T_n for testing H_0: $\theta \in \Theta_0$ vs. H_a: $\theta \in \Theta - \Theta_0$.

We first derive the Bahadur efficiency for the Lancaster test. Let f_i, F_i and F_i^{-1} be the PDF, CDF, and inverse CDF for $\chi^2_{w_i}$, with $w_i > 0$ for $i = 1, 2, \cdots, m$.

[Theorem 1] Assume m independent test statistics $T_{n_1}^{(1)},\ \cdots,\ T_{n_m}^{(m)}$ have significance levels $P_{n_1}^{(1)},\ \cdots,\ P_{n_m}^{(m)}$ respectively. The Lancaster statistic, $\sum_{i=1}^{m} F_i^{-1}(1 - P_{n_i}^{(i)})$, has the Bahadur efficiency slope $c_{Lancaster}(\theta) = \sum_{i=1}^{m} \lambda_i c_i(\theta)$ under H_a: $\theta \in \Theta - \Theta_0$.

[13] derived the Bahadur efficiency slope for the regular (unweighted) Z-test, $\sum_{i=1}^{m} \Phi^{-1}(1 - P_{n_i}^{(i)})/\sqrt{m}$. Here we generalize their findings to the weighted Z-test.

[**Theorem 2**] Assume m independent test statistics $T_{n_1}^{(1)}, \cdots, T_{n_m}^{(m)}$ have significance levels $P_{n_1}^{(1)}, \cdots, P_{n_m}^{(m)}$ respectively. The weighted Z-test, $\sum_i w_i \Phi^{-1}(1-P_{n_i}^{(i)})/\sqrt{\sum_i w_i}$, has the Bahadur efficiency slope $c_{Weighted\,z}(\theta) = \left(\sum_{i=1}^{m} w_i \sqrt{\lambda_i c_i(\theta)}/\sqrt{\sum_{i=1}^{m} w_i^2}\right)^2$ under H_a: $\theta \in \Theta - \Theta_0$.

When $w_i = 1$ for all $i = 1, 2, \cdots, m$, the weighted Z-test is reduced to the regular z statistic and the Bahadur efficiency slope is $c_{regular\,z}(\theta) = \left(\sum_{i=1}^{m} \sqrt{\lambda_i c_i(\theta)}/\sqrt{m}\right)^2$. This finding is in agreement with the Bahadur efficiency finding for the regular Z-test in [13].

The Lancaster test statistic is superior to the weighted Z-test and regular Z-test in terms of Bahadur efficiency. Using the induction method, we show that the Bahadur relative efficiency $\phi_{12} = c_{Lancaster}(\theta)/\, c_{Weighted\,z}(\theta) \geq 1$ and $\phi_{12} = c_{Lancaster}(\theta)/\, c_{regular\,z}(\theta) \geq 1$ for all $\theta \in \Theta - \Theta_0$. The fact that $\lim_{\varepsilon \to 0} N^{(2)}/N^{(1)} = \phi_{12}(\theta)$ indicates that the Lancaster procedure will require smaller sample sizes as compared to the weighted Z-test to achieve the same significance level.

[17] suggested a weighted Fisher's method. Let $Q = \prod_{i=1}^{m} (P_{n_i}^{(i)})^{w_i}$, so $-\ln(Q) = -\sum_{i=1}^{m} w_i \ln(P_{n_i}^{(i)})$. When the weights are unequal, the null CDF of Q is given by $\Pr(Q < q) = \sum_{i=1}^{m} \Lambda_i q^{1/w_i}$, $q \in [0,\ 1]$, where $\Lambda_i = (w_i)^{m-1}/\prod_{j \neq i}(w_i - w_j)$. Below we will derive the Bahadur efficiency for Good's test.

[**Theorem 3**] Assume m independent test statistics $T_{n_1}^{(1)}, \cdots, T_{n_m}^{(m)}$ have significance levels $P_{n_1}^{(1)}, \cdots, P_{n_m}^{(m)}$ respectively. Good's test statistic, $-\ln(Q) = -\sum_{i=1}^{m} w_i \ln(P_{n_i}^{(i)})$, has the Bahadur efficiency slope $c_{Good}(\theta) = \sum_{i=1}^{m} w_i \lambda_i c_i(\theta)/\max_i(w_i)$ under H_a: $\theta \in \Theta - \Theta_0$.

The maximal weight, $\max_i(w_i)$, has a strong impact on the Bahadur efficiency in Good's test. Only the individual test(s) assigned with the maximal weight reserves its Bahadur efficiency in Good's test. That is, if $w_i = \max_i(w_i)$, then $w_i \lambda_i c_i(\theta)/\max_i(w_i) = \max_i(w_i)\lambda_i c_i(\theta)/\max_i(w_i) = \lambda_i c_i(\theta)$. Other individual tests will relatively lose more Bahadur efficiency as the maximal weight gets larger. That is, if $w_i < \max_i(w_i)$, then $w_i \lambda_i c_i(\theta)/\max_i(w_i) < \lambda_i c_i(\theta)$.

The Lancaster procedure is superior to Good's test in terms of the Bahadur efficiency, i.e. $\phi_{12} = c_{Lancaster}(\theta)/\, c_{Good}(\theta) \geq 1$ for all $\theta \in \Theta - \Theta_0$ and $\lim_{\varepsilon \to 0} N^{(2)}/N^{(1)} = \phi_{12}(\theta)$. For large-scale tests, which often occur in next-generation sequencing data, the Lancaster procedure will require relatively smaller sample sizes as compared to Good's test, i.e., $N^{Lancaster} \leq N^{Good}$ when the significance level goes to 0, which represents sparse signaling in high throughput data.

Lancaster Procedure Has the Optimal Bahadur Efficiency. We can further prove that the Lancaster procedure reaches the upper bounds of Bahadur efficiency among all non-decreasing T_n. Thus the Lancaster procedure has the optimal Bahadur efficiency compared to all other combination methods under mild conditions.

[**Proposition 1**] Let T_n be any function of m independent test statistics $T_{n_1}^{(1)}, \cdots, T_{n_m}^{(m)}$. Let $c_{any}(\theta) > 0$ be the Bahadur efficiency slope of T_n. Assume T_n is non-decreasing in a way that $t_1 \leq t_1^*, \cdots, t_m \leq t_m^* \Rightarrow T_n(t_1, \cdots, t_m) \leq T_n(t_1^*, \cdots, t_m^*)$. Then the Lancaster statistics have the optimal Bahadur efficiency, with $c_{Lancaster}(\theta) \geq c_{any}(\theta)$ for all $\theta \in \Theta - \Theta_0$.

The Lancaster procedure and Fisher's test both have the optimal Bahadur efficiency among all non-decreasing combined tests. Since the Lancaster procedure can incorporate weight

functions for auxiliary information in modeling and testing, the Lancaster procedure has more flexibility and it can be considered as the optimal generalized Fisher's method. The non-decreasing condition, $t_1 \le t_1^*, \ \cdots, t_m \le t_m^* \ \Rightarrow \ T_n(t_1, \ \cdots, t_m) \le T_n(t_1^*, \ \cdots, t_m^*)$, is easy to meet in practice.

Comparing Bahadur Efficiency for Correlated Data. It is critical to assess Bahadur efficiency for correlated data as it will shed light on the impact of correlation structures on the asymptotic convergence rate of significance levels and will further impact the sample sizes required for the experiments. This is a challenging topic since the distributions of combined test statistics under complex correlation structures have no closed analytical forms. To address this issue, we give an approximate Bahadur efficiency using the techniques described in the Methods Section. Below are some interesting results.

[**Proposition 2**] When m statistics $T_{n_1}^{(1)}, \ \cdots, \ T_{n_m}^{(m)}$ are correlated, under H_a: $\theta \in \Theta - \Theta_0$:

- the Lancaster statistic, has an approximate Bahadur efficiency slope

$$c_{Lancaster}^{Correlated}(\theta) \approx \frac{\sum_i w_i}{\sum_i w_i + 2\sum_{i<j}\rho_{ij}} \sum_{i=1}^{m} \lambda_i c_i(\theta),$$

  where

$$\rho_{ij} = \mathrm{cov}(F_i^{-1}(1 - P_{n_i}^{(i)}), F_j^{-1}(1 - P_{n_j}^{(j)}));$$

- the Good's test statistic has an approximate Bahadur efficiency slope

$$c_{Good}^{Correlated}(\theta) \approx \frac{\sum_i w_i}{\sum_i w_i^2 + 2\sum_{i<j}\tilde{\rho}_{ij}} \sum_{i=1}^{m} w_i \lambda_i c_i(\theta),$$

  where

$$\tilde{\rho}_{ij} = \mathrm{cov}(\ln(P_{n_i}^{(i)}), \ln(P_{n_j}^{(j)}));$$

- the Fisher's statistic has an approximate Bahadur efficiency slope

$$c_{Fisher}^{Correlated}(\theta) \approx \frac{m}{m + 2\sum_{i<j}\tilde{\rho}_{ij}} \sum_{i=1}^{m} \lambda_i c_i(\theta).$$

Simulation Study

We conducted an extensive simulation study to assess the type I error and power of the SKA-T-Lancaster procedure. We further compared our proposed method to Gene Set Enrichment Analysis (GSEA) [5]. The empirical assessment was based on rigorous simulation algorithms for sequencing-based genome-wide association studies [18].

The simulation was conducted using the whole exome sequencing genotype data from the 1000 Genomes Project Phase 1 study (n = 822 individuals). After filtration, 40,918 biallellic protein-changing coding variants in selected pathways were mapped to KEGG and Biocarta pathways. To avoid testing over- or under- sized pathways, we selected pathways containing 10 to 100 genes. This yielded 353 pathways with 3304 genes for our simulation study.

We applied a genome-wide additive model to evaluate pathway-testing methods using realistic genetic architectures. Let $Y_i = X_i\beta + \varepsilon_i$, where Y_i is a continuous trait, the vector X_i is the whole exome sequencing genotype for the i^{th} subject and $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ is random noise. The vector β contains genetic effect regression coefficients corresponding to genotype variants. In simulation, the j^{th} variant is causal if $|\beta_j| > 0$; pathways and genes are causal if they harbor causal variants. We adopted a stochastic hierarchical effect model $\beta_{pgv} = C_p \times C_g d_g \times C_{gv} d_{gv} \times e_{pgv}$ to distribute the total genetic variance into pathways, genes, and individual variants [18]. Within a central causal pathway, we first randomly selected 50% of the genes to be associated with the trait. Then we randomly selected 70% of the variants in causal genes to be associated with the trait. We randomly assigned 80% (20%) of causal genes to be detrimental (protective). For variants within causal genes, 80% were detrimental and 20% were protective. We set the whole-genome heritability $h_2 = Var(X\beta)/(Var(X\beta) + \sigma^2) = 20\%$. This resembles heritability in real data, which often ranges between 20% and 30%. We used Bonferroni correction to control Family-Wise Error Rate (FWER) and set the genome-wide significance level at $\alpha = 0.05/353 = 1.4164E\text{-}4$ for testing 353 pathways. We performed principal component analysis and included the top 3 principal components as covariates in regression analyses to adjust for the population stratification.

In the SKAT-Lancaster procedure, we first performed SKAT to test overall effects on the gene level. Then we considered 4 weight functions in the Lancaster procedure when combining p-values among genes in a pathway:

- Gene size weight $= 2\sqrt{\tilde{n}/n_i}$, where n_i is the number of SNPs in the i^{th} gene and $\tilde{n} = median(n_i)$ is the median gene size. This weight can remove bias when testing overly small or overly generalized pathways.
- AIC weight, BIC weight: these weight functions calculate the degrees of variations summarized by the gene level multi-SNP regression.
- Uniform weight = 2.

We considered 3 simulation scenarios (Table 1). In Scenario 1, we assessed the global null hypothesis type I error by setting all genetic effect coefficients as zero, i.e. $\beta = \vec{0}$. Any pathways or genes reaching the significance level were considered as false positives. The results in Table 2 indicate that the SKAT-Lancaster procedure has well-controlled type I error rates (~10E-4). We further investigated the Q-Q plot by comparing observed p-values versus expected p-values (Fig 1). The type I error inflation factor (λ) is the ratio between the area under the curve and the area under the diagonal reference line. Fig 1 indicates that SKAT-Lancaster procedure with 4 weight functions has no inflation of the global null hypothesis type I error rate ($\lambda < 1$).

In Scenarios 2 and 3, we assessed the stringent power and lenient power when randomly generating one central causal pathway in each simulation (Table 1). The stringent power calculates the percentage of times the central causal pathway is found significant. Due to the correlation among pathways, pathways that share causal genes with the central causal pathway are overlapping causal pathways. The lenient power calculates the percentage of times (central and overlapping) causal pathways are found significant.

The results in Table 2 indicate that the SKAT-Lancaster procedure outperformed GSEA. In Scenario 2, the SKAT-Lancaster procedure with 4 weight functions had lenient power ranging between 0.826 and 0.884, while GSEA had lenient power of 0.373. In Scenario 3, the SKAT-Lancaster procedure with 4 weight functions had lenient power ranging between 0.543 and 0.645, while GSEA had lenient power of 0.505. We randomly assign the causal variances in Scenarios 2 and 3, the SKAT-Lancaster procedure with uniform weight had the best detection power.

Table 1. Simulation Scenarios and parameters*.

Simulation Scenarios 1
• Include 353 pathways, 3304 genes.
• Phenotype is normally distributed.
• Assume heritability is 20%.
• $\beta = \vec{0}$.
• No pathways, genes or variations are associated with the trait.
• Significance level is 0.05/353. All significant results are considered as type 1 errors.
Simulation Scenario 2
• Include 353 pathways and 3304 genes.
• Phenotype is normally distributed.
• Randomly assign one central causal pathway. Within the central causal pathway, randomly assign 50% causal genes. Randomly assign 70% causal variants in associated genes.
• Randomly assigned 80% (20%) of causal genes to be detrimental (protective). For variants within the causal genes, 80% are detrimental and 20% are protective.
• Associated variants' effect size ~ $\log_{10}(MAF)$.
• Significance level is 0.05/353.
Simulation Scenarios 3
• Include 353 pathways and 3304 genes.
• Phenotype is normally distributed.
• Assume heritability is 20%.
• Randomly assign one central causal pathway. Within the central causal pathway, randomly assign 50% causal genes. Randomly assign 70% causal variants in associated genes.
• Randomly assigned 80% (20%) of causal genes to be detrimental (protective). For variants within the causal genes, 80% are detrimental and 20% are protective.
• Associated variants' effect size ~$1/\sqrt{MAF * (1 - MAF)}$.
• Significance level is 0.05/353.

* Covariates: top 3 principal components for population stratification are included as covariates in all three simulation scenarios.

doi:10.1371/journal.pone.0152667.t001

Regarding the computing time, the self-contained Lancaster procedure compares a test statistic to an asymptotic distribution, thus it does not require intensive computation. The competitive Lancaster procedure is based on permutation and it has similar computation efficiency as compared with GSEA.

Case Study: Lipid Meta-Analysis

We illustrate our method using meta-analysis data generated by the Global Lipids Genetics Consortium. To identify new loci and validate existing loci associated with lipids, [19] we analyzed the levels of low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol, triglycerides (TG) and total cholesterol (TC) of 196,475 individuals from 60 studies. A total of 1,048,161 Single Nucleotide Polymorphisms (SNPs) were genotyped using the genome-wide association study (GWAS) arrays and Metabochip arrays. These variants were selected from promising loci associated with lipid and coronary artery disease, based on findings from previous GWAS studies and the 1000 Genome Project. Subjects taking lipid-lowering

Table 2. Comparison of type I error and power among competing methods.

Simulation Scenario 1			
		Type_1 error	Inflation factor
Test	*Weight function*	(10E-4)	λ
SKAT- Lancaster	*Uniform*	1.1615	0.9921
SKAT- Lancaster	*Gene size*	1.3598	0.9852
SKAT- Lancaster	*AIC*	0.9632	0.9477
SKAT- Lancaster	*BIC*	1.1331	0.9770
GSEA		12.0000	1.2390
Simulation Scenario 2			
Test	*Weight function*	Stringent Power	Lenient Power
SKAT- Lancaster	*Uniform*	0.870	0.884
SKAT- Lancaster	*Gene size*	0.810	0.836
SKAT- Lancaster	*AIC*	0.832	0.854
SKAT- Lancaster	*BIC*	0.809	0.826
GSEA		0.279	0.373
Simulation Scenario 3			
Test	*Weight function*	Stringent Power	Lenient Power
SKAT- Lancaster	*Uniform*	0.610	0.645
SKAT- Lancaster	*Gene size*	0.509	0.543
SKAT- Lancaster	*AIC*	0.585	0.628
SKAT- Lancaster	*BIC*	0.540	0.558
GSEA		0.468	0.505

doi:10.1371/journal.pone.0152667.t002

medications were excluded in the meta-analysis. The additive effect of each SNP on blood lipid levels after adjusting for age and sex was analyzed and p-value was generated for each SNP and each lipid variable. Genomic control values for the initial meta-analyses were 1.10–1.15, indicating that population stratification had only a minor impact on the results [20].

The SKAT-Lancaster procedure can only be applied to original data. Remarkably, as the Lancaster procedure is independent from the SKAT test, it can be applied to secondary data analysis. To identify pathways that are more significant than others, we performed the competitive Lancaster procedure. In the competitive test, we performed 100,000 times of permutations and ensured that the permutated pathways preserved the size and characteristics of original pathways. Our simulation study showed that the competitive Lancaster procedure had well-controlled type I error rates to prevent false discoveries.

Before comparing the proposed method to Fisher's method [21] and weighted Z-test [22], we considered 4 weight functions for the Lancaster procedure:

- $w_1 = 2\sqrt{\tilde{n}/n_i}$, where n_i is the number of SNPs in the i^{th} gene and $\tilde{n} = median(n_i)$ is the median gene size. This is a weight adjusted by gene size to remove the bias from large genes.
- $w_2 = 4\sqrt{MAF(1-MAF)}$, where MAF stands for minor allele frequency. Common variants receive higher weights.
- $w_3 = 1/\sqrt{MAF(1-MAF)}$. Rare variants receive higher weights.
- Uniform weight: $w_4 = 2$.

Pathway analysis was performed using the gene ontology (GO) gene sets from http://www.broadinstitute.org/gsea/index.jsp. A total of 1454 pathways were analyzed and multiple testing

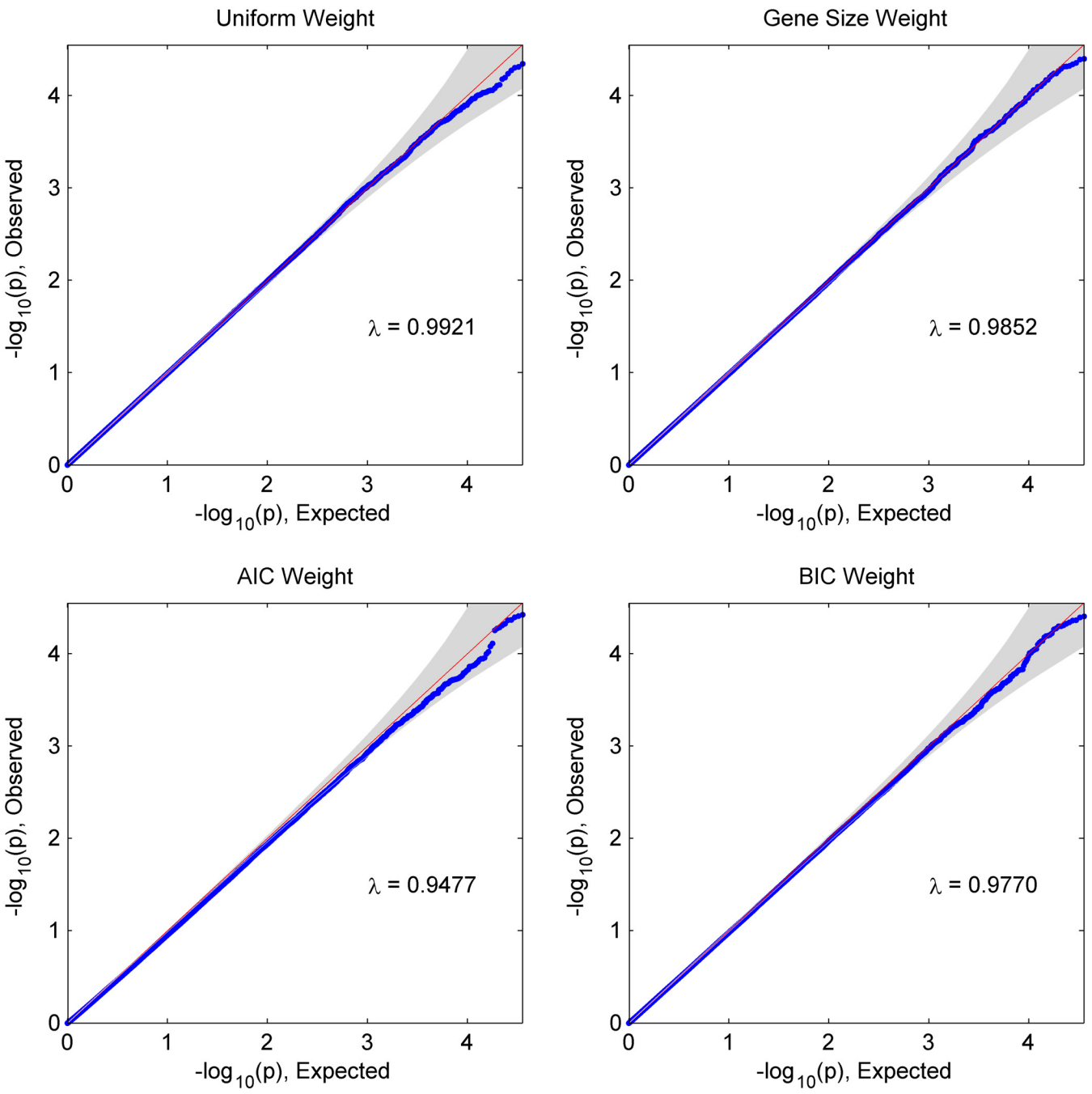

Fig 1. Q-Q plots investigating global null hypothesis type-I errors for the SKAT-Lancaster procedure under Simulation Scenario 1 (*λ* is the inflation factor for the Type I error rate). The type I error inflation factor (*λ*) is the ratio between the area under the curve and the area under the diagonal reference line.

doi:10.1371/journal.pone.0152667.g001

was adjusted by False Discovery Rate (FDR) [23]. The numbers of significant pathways are summarized in Fig 2 As shown in Table 3, the Lancaster procedure outperformed Fisher's method and weighted Z-test by identifying more significant pathways. When the Lancaster procedure was assigned with uniform weights (w_4), it performed equivalently to Fisher's method. The weighted Z-test is not optimal in Bahadur efficiency, so it identified fewer

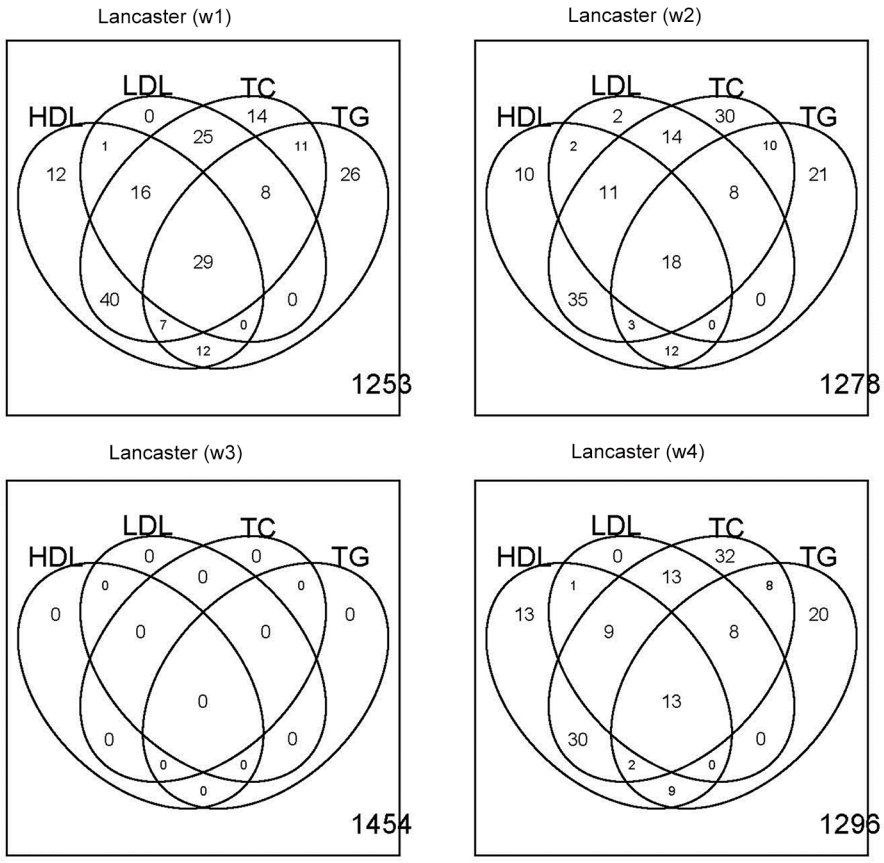

Fig 2. Venn Diagrams for Significant Pathways (FDR < 0.05).

doi:10.1371/journal.pone.0152667.g002

pathways than the Lancaster procedure and Fisher's method. Weight functions w_1 and w_2 outperformed w_3 and w_4, indicating that removing gene size bias and assigning higher weights to common variants can improve power of the Lancaster procedure.

We compared our pathway findings with findings from the MAGENTA analysis in [19] (Table 4). The Lancaster procedure (w_1) showed that the "enzyme binding" pathway is significantly associated with HDL (FDR$<10^{-5}$), which agrees with the finding from [19] (FDR = 0.038). The "enzyme binding" pathway contained 178 genes interacting selectively and non-covalently with any enzyme. The Lancaster procedure (w_1, w_2, w_4) showed that the "lipid

Table 3. Number of Significant pathways.

FDR < 0.05	HDL	LDL	TC	TG
Lancaster (w1)	117	79	150	93
Lancaster (w2)	91	55	129	72
Lancaster (w3)	0	0	0	0
Lancaster (w4)	77	44	115	60
Fisher	77	44	115	60
Weighted Z-test (w1)	2	1	3	2
Weighted Z-test (w2)	5	0	5	4
Weighted Z-test (w3)	0	0	0	0
Weighted Z-test (w4)	4	0	4	6

doi:10.1371/journal.pone.0152667.t003

Table 4. Comparison of pathway analysis p-values.

Pathway Name	enzyme binding	lipid transport	lipoprotein metabolic process
GO Accession	GO:0019899	GO:0006869	GO:0042157
Gene Ontology	molecular function	biological process	biological process
Description	Interacting selectively with any enzyme	The directed movement of lipids into, out of, within or between cells. Lipids are compounds soluble in an organic solvent but not, or sparingly, in an aqueous solvent.	The chemical reactions and pathways involving any conjugated, water-soluble protein in which the nonprotein moiety consists of a lipid or lipids.
number of genes	178	28	33
number of SNPs	12089	1058	769
(Willer 2013)*	0.038	0.0016	0.00017
Lancaster (w1)*	$<10^{-5}$	$<10^{-5}$	$<10^{-5}$
Lancaster (w2) *	0.80	$<10^{-5}$	$<10^{-5}$
Lancaster (w3)*	0.98	0.44	0.44
Lancaster (w4)*	0.82	$<10^{-5}$	$<10^{-5}$
Fisher*	0.82	$<10^{-5}$	$<10^{-5}$
Weighted z (w1)*	0.90	0.36	0.58
Weighted z (w2)*	0.94	0.21	0.35
Weighted z (w3)*	0.99	0.65	0.55
Weighted z (w4)*	0.94	0.23	0.36

* FDR adjusted p-values

doi:10.1371/journal.pone.0152667.t004

transport" pathway is significantly associated with LDL (FDR adjusted p-value $<10^{-5}$), which agrees with the finding from [19] (FDR = 0.0016). The "lipid transport" pathway contains 28 genes involving directed movement of lipids into, out of, within, or between cells. Lipids are compounds soluble in an organic solvent but not, or sparingly, in an aqueous solvent. The Lancaster procedure (w_1, w_2, w_4)found that the "lipoprotein metabolic process" pathway is significantly associated with LDL (FDR adjusted p-value $<10^{-5}$), which agrees with the finding from [19] (FDR = 0.00017). The "lipoprotein metabolic process" pathway contains 33 genes involving the chemical reactions. The pathway also involves any conjugated, water-soluble protein in which the non-protein moiety consists of a lipid or lipids.

Discussion and Conclusions

The proposed two-stage approach is a powerful tool to integrate information in pathway analysis of sequencing association studies. The first stage is the gene-based testing, where effects from rare variants within a gene are summarized into one p-value using the SKAT test. In the second stage, p-values from multiple genes are combined for pathway analysis and meta-analysis using the correlated Lancaster procedure. In this work, we prove that the Lancaster procedure is optimal in Bahadur efficiency among all combined p-value methods.

We assess the Bahadur efficiency among weighted combined p-value methods and further prove that the Lancaster procedure is optimal in Bahadur efficiency under very mild conditions. There has been a lack of theatrical comparison among combined p-value methods. Several simulation studies have compared weighted combined p-value methods [15, 22, 24]. With more than 400 citations in the literature, these studies have been a subject of intense interest to the research community heated discussions in the research community, but yield controversial results in different simulation scenarios. Thus, we fill the gap by comparing the Bahadur efficiency among methods.

The Bahadur efficiency is a critical measure of performance of statistical testing [25] [26]. In [25], Bahadur efficiency has been applied for sensitivity analyses in observation studies. The Bahadur efficiency, $\lim_{\varepsilon\to 0} N^{(2)}/N^{(1)} = \phi_{12}(\theta)$, compares sample sizes among different statistical tests when signals become sparse in sequencing data, i.e. $\varepsilon \to 0$. As the number of genetic variants scanned by the sequencing technology increases from thousands to millions, signals that are associated with phenotypes become sparse, requiring a more stringent statistical significance level to detect sparse signals, i.e. $(P_{N^{(i)}} < \varepsilon \to 0)$. The optimal Bahadur efficiency ensures that the Lancaster procedure asymptotically requires a minimal sample size to detect sparse signals.

Among combined p-value methods, the Lancaster procedure can be considered as the generalized Fisher's method with a weight function. Weight functions, when used appropriately, can generally increase the power of combined p-value methods [27–29].

Evaluating Bahadur efficiency for high-throughput genetic data is critical since there is no combined p-value method that that is uniformly the most powerful. Bahadur efficiency calculates the limiting ratio of sample sizes required by two statistics to attain an equally small significance level. The optimal Bahadur efficiency indicates that the Lancaster procedure asymptotically requires a minimal sample size to attain the significance level.

Data and Software

R package 'CombinePValue' has been created for the proposed Lancaster procedure. Case study data are available from http://csg.sph.umich.edu//abecasis/public/lipids2013/. Source codes for simulation analyses can be provided upon contacting Dr Guodong Wu.

Appendix

Lemma 1 [16, 30] is needed to derive the Bahadur efficiency.

[Lemma 1] If the following two conditions are met,

(Condition 1) there exists a function $b(\theta)$, $0 < b(\theta) < \infty$, such that $T_n/\sqrt{n} \to b(\theta)$ with probability 1 under H_a: $\theta \in \Theta - \Theta_0$;

(Condition 2) there exists a function $f(t)$, $0 < f(t) < \infty$, which is continuous in some open set containing the range of $b(\theta)$ such that for each t in the open set $-n^{-1}\ln[1 - F_0(\sqrt{n}t)] \to f(t)$, then the Bahadur efficiency slope of $\{T_n\}$ is $c(\theta) = 2f(b(\theta))$.

Proof of Theorem 1: Since the equivalent tests have the same Bahadur efficiency, we can consider $T_n^{Lancaster} = \sqrt{\sum_{i=1}^{m} Z_{n_i}^{(i)}}$, where $Z_{n_i}^{(i)} = F_i^{-1}(1 - P_{n_i}^{(i)})$. Under H_a: $\theta \in \Theta_0$, $P_{n_i}^{(i)} \sim Uniform(0,\ 1)$ and $Z_{n_i}^{(i)} \sim \chi^2_{w_i}$. According to Theorem 2.1 by [31], we have $-2\ln P_{n_i}^{(i)} = -2\ln(1 - F_i(Z_{n_i}^{(i)})) = -2\ln(f_i(Z_{n_i}^{(i)})) + o(1) = Z_{n_i}^{(i)}(1 + o(1))$ as $Z_{n_i}^{(i)} \to \infty$. So $n^{-1}Z_{n_i}^{(i)}(1 + o(1)) = -(2/n)\log P_{n_i}^{(i)} \to \lambda_i c_i(\theta)$ with probability 1 under H_a: $\theta \in \Theta - \Theta_0$. It follows that

$$T_n^{Lancaster}/\sqrt{n} = \sqrt{\textstyle\sum_{i=1}^{m} Z_{n_i}^{(i)}/n} \to \sqrt{\textstyle\sum_{i=1}^{m} \lambda_i c_i(\theta)}. \quad (1)$$

Now, for $\theta \in \Theta_0$, $T_n^{Lancaster}$ is distributed as the square root of $\chi^2_{\sum_i w_i}$ with the CDF F and PDF f. According to Theorem 2.1 by [31], we have,

$$-n^{-1}\ln\left[1 - F(\sqrt{n}t)\right] = -n^{-1}\ln\left[f(\sqrt{n}t)\right] = 0.5t^2(1 + o(1)) \to 0.5t^2. \quad (2)$$

Plug the results from Eqs (1) and (2) to Lemma 1. Then the Bahahur efficiency slope for the Lancaster statistic is $c_{Lancaster}(\theta) = \sum_{i=1}^{m} \lambda_i c_i(\theta)$ under H_a: $\theta \in \Theta - \Theta_0$.

Proof of Theorem 2: Rewrite the weighted z statistic as $T_n^{Weighted\,z} = \sum_{i=1}^{m}(w_i Z_{n_i}^{(i)})/\sqrt{\sum_{i=1}^{m} w_i^2}$, where $Z_{n_i}^{(i)} = \Phi^{-1}(1 - P_{n_i}^{(i)})$. Under H_0: $\theta \in \Theta_0$, $P_{n_i}^{(i)} \sim Uniform(0,\ 1)$ and $Z_{n_i}^{(i)} \sim N(0,\ 1)$. According to Theorem 2.1 by [31], we have $-2\ln P_{n_i}^{(i)} = -2\ln(1 - \Phi(Z_{n_i}^{(i)})) = -2\ln(f(Z_{n_i}^{(i)})) + o(1) = [Z_{n_i}^{(i)}]^2(1 + o(1))$ as $Z_{n_i}^{(i)} \to \infty$. So $n^{-1}[Z_{n_i}^{(i)}]^2(1 + o(1)) = -2n^{-1}\log P_{n_i}^{(i)} \to \lambda_i c_i(\theta)$ with probability 1 under H_a: $\theta \in \Theta - \Theta_0$. It follows that

$$T_n^{Weighted\,z}/\sqrt{n} = \sum_{i=1}^{m}(w_i Z_{n_i}^{(i)})/\sqrt{n\sum_{i=1}^{m} w_i^2} \to \sum_{i=1}^{m} w_i\sqrt{\lambda_i c_i(\theta)}/\sqrt{\sum_{i=1}^{m} w_i^2}. \quad (3)$$

Now, for $\theta \in \Theta_0$, $T_n^{Weighted\,z} \sim N(0,\ 1)$. Thus,

$$-n^{-1}\ln[1 - \Phi(\sqrt{n}t)] = 0.5t^2(1 + o(1)) \to 0.5t^2. \quad (4)$$

Eqs (3) and (4) and Lemma 1 imply that the Bahadur efficiency slope of $\{T_n^{Weighted\,z}\}$ is $c_{Weighted\,z}(\theta) = \left(\sum_{i=1}^{m} w_i\sqrt{\lambda_i c_i(\theta)}/\sqrt{\sum_{i=1}^{m} w_i^2}\right)^2$.

Proof of Theorem 3: It is equivalent to consider $T_n^{Good} = \sqrt{-\sum_{i=1}^{m} w_i \ln(P_{n_i}^{(i)})}$. Then

$$T_n^{Good}/\sqrt{n} = \sqrt{-n^{-1}\sum_{i=1}^{m} w_i \ln(P_{n_i}^{(i)})} \to \sqrt{0.5\sum_{i=1}^{m} w_i\lambda_i c_i(\theta)}$$

with probability 1 under H_a: $\theta \in \Theta - \Theta_0$. Direction calculation shows that the PDF of $\sqrt{-\ln(Q)}$ is $f(x) = 2x\sum_{i=1}^{m}\frac{\Lambda_i}{w_i}\exp(-x^2/w_i)$ under H_0. Let I be the index corresponds to $\max_i(w_i) = w_I$. One can construct the upper and lower bounds of $f(x)$,

$$2x\frac{\Lambda_I}{w_I}\exp(-x^2/w_I) \le f(x) \le 2xm\frac{\Lambda_I}{w_I}\exp(-x^2/w_I)$$

when x is greater than a certain finite number. This implies that

$$-n^{-1}\ln[1 - F(\sqrt{n}t)] = -n^{-1}\ln f(\sqrt{n}t) + o(1) \to t^2/\max(w_i).$$

By Lemma 1, the Bahadur efficiency slope for Good's test is $c_{Good}(\theta) = \sum_{i=1}^{m} w_i\lambda_i c_i(\theta)/\max_i(w_i)$ for $\theta \in \Theta - \Theta_0$.

Proof of Proposition 1: Let P_n be the significance level of T_n and let $t^{(1)}, \cdots, t^{(m)}$ be the observed values of $T_{n_1}^{(1)},\ \cdots,\ T_{n_m}^{(m)}$. For any non-decreasing T_n, we have

$$P_n \ge \Pr(T_{n_1}^{(1)} > t^{(1)},\ \cdots,\ T_{n_m}^{(m)} > t^{(m)}) \ge \prod_{i=1}^{m}\Pr(T_{n_i}^{(i)} > t^{(i)}).$$

Therefore, $c_{Lancaster}(\theta) = \sum_{i=1}^{m} \lambda_i c_i(\theta) = -n^{-1} \sum_{i=1}^{m} \ln(\Pr(T_{n_i}^{(i)} > t^{(i)})) \geq -n^{-1} \ln(P_n) = c_{any}(\theta)$ for all $\theta \in \Theta - \Theta_0$.

Proof of Proposition 2: We give the proof to the Lancaster statistic. Note that any correlation structure among p-values has no impact to the first condition of Lemma 1. As a result, one can repeat the derivation for Theorem 1 to get

$$T_n^{Lancaster}/\sqrt{n} = \sqrt{\sum_{i=1}^{m} Z_{n_i}^{(i)}/n} \to \sqrt{\sum_{i=1}^{m} \lambda_i c_i(\theta)} \quad (5)$$

under H_a: $\theta \in \Theta - \Theta_0$.

When $P_{n_1}^{(1)}, \cdots, P_{n_m}^{(m)}$ are correlated, the null distribution of $T_n^{Lancaster} = \sum_{i=1}^{m} F_i^{-1}(1 - P_{n_i}^{(i)})$ no longer follows $\chi^2_{\sum_i w_i}$ distribution. One can approximate it by a scaled chi-square distribution such as $T_n^{Lancaster} \approx c\chi_\nu^2$. By matching expectation and variance between two sides, one can solve for c and ν. Under H_0: $\theta \in \Theta_0$, direct calculations show that

$$-n^{-1}\ln[1 - F(\sqrt{n}t)] \approx -n^{-1}\ln[1 - \tilde{F}(\sqrt{n}t)] = -n^{-1}\ln[\tilde{f}(\sqrt{n}t)] = \frac{0.5t^2 \sum_i w_i}{\sum_i w_i + 2\sum_{i<j} \rho_{ij}}(1 + o(1)) \quad (6)$$

where F is the CDF of $T_{Lancaster}^{Correlated}$ and $\tilde{F}$ ($\tilde{f}$) are the CDF (PDF) of $c^{-1}\chi_\nu^2$.

Plug the results from Eqs (5) and (6) to Lemma 1. Then the Bahahur efficiency slope for the Lancaster statistic is $c_{Lancaster}^{Correlated}(\theta) \approx \frac{\sum_i w_i}{\sum_i w_i + 2\sum_{i<j} \rho_{ij}} \sum_{i=1}^{m} \lambda_i c_i(\theta)$ under H_a: $\theta \in \Theta - \Theta_0$.

Acknowledgments

There are no competing interests to this work. We thank three reviewers for their constructive comments, which helped us improve the manuscript. This work is supported by NHGRI of the National Institutes of Health under award number R01HG008115 (D.Z.). We also thank David Williams and the Medical Writing Center at Children's Mercy Hospital for their comments, support, and assistance with the manuscript, which significantly enhanced readability of our work.

Author Contributions

Conceived and designed the experiments: HD. Performed the experiments: HD GW. Analyzed the data: HD. Contributed reagents/materials/analysis tools: MW DZ. Wrote the paper: HD.

References

1. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012; 337 (6090):64–9. Epub 2012/05/19. doi: 10.1126/science.1219240 science.1219240 [pii]. PMID: 22604720; PubMed Central PMCID: PMC3708544.
2. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004; 32(Database issue):D277–80. Epub 2003/12/19. doi: 10.1093/nar/gkh063 32/suppl_1/D277 [pii]. PMID: 14681412; PubMed Central PMCID: PMC308797.
3. Nikolsky Y, Bryant J. Protein networks and pathway analysis. Preface. Methods Mol Biol. 2009; 563:v–vii. Epub 2009/09/19. PMID: 19760825.
4. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, et al. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res. 2009; 37(Database issue):D619–22.

Epub 2008/11/05. doi: 10.1093/nar/gkn863 gkn863 [pii]. PMID: 18981052; PubMed Central PMCID: PMC2686536.

5. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet. 2007; 81(6):1278–83. Epub 2007/10/30. S0002929707637756 [pii] doi: 10.1086/522374 PMID: 17966091; PubMed Central PMCID: PMC2276352.
6. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011; 89(1):82–93. Epub 2011/07/09. doi: 10.1016/j.ajhg.2011.05.029 S0002-9297(11)00222-9 [pii]. PMID: 21737059; PubMed Central PMCID: PMC3135811.
7. Lin X. Variance component testing in generalised linear models with random effects. Biometrika. 1997; 84:309–26.
8. Dai H, Leeder JS, Cui Y. A modified generalized Fisher method for combining probabilities from dependent tests. Front Genet. 2014; 5:32. Epub 2014/03/07. doi: 10.3389/fgene.2014.00032 PMID: 24600471; PubMed Central PMCID: PMC3929847.
9. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nat Rev Genet. 2010; 11(12):843–54. Epub 2010/11/19. doi: 10.1038/nrg2884 nrg2884 [pii]. PMID: 21085203.
10. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Methods. 2011; 8(10):833–5. Epub 2011/09/06. doi: 10.1038/nmeth.1681 nmeth.1681 [pii]. PMID: 21892150.
11. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods. 2014; 11(4):407–9. Epub 2014/02/18. doi: 10.1038/nmeth.2848 nmeth.2848 [pii]. PMID: 24531419.
12. Dai H, Charnigo R, Srivastava T, Talebizadeh Z, Ye S. Integrating P-values for Genetic and Genomic Data Analysis. Journal of Biometrics and Biostatistics. 2012; 3(7):e117
13. Littell RC, Folks JL. Asymptotic optimality of Fisher's method of combining independent tests. Journal of American Statistical Association. 1971; 66:802–6.
14. Littell RC, Folks JL. Asymptotic optimality of Fisher's method of combining independent tests. II. Journal of American Statistical Association. 1973; 68:193–4.
15. Zaykin DV. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. J Evol Biol. 2011; 24(8):1836–41. Epub 2011/05/25. doi: 10.1111/j.1420-9101.2011.02297.x PMID: 21605215; PubMed Central PMCID: PMC3135688.
16. Bahadur RR. Rates of convergence of estimates and test statistics. Annals of Mathematical Statistics. 1967; 38:303–24.
17. Good IJ. On the weighted combination of significance tests. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 1955; 17:264–5.
18. Wu G, Zhi D. Pathway-based approaches for sequencing-based genome-wide association studies. Genet Epidemiol. 2013; 37(5):478–94. Epub 2013/05/08. doi: 10.1002/gepi.21728 PMID: 23650134; PubMed Central PMCID: PMC3856324.
19. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013; 45(11):1274–83. Epub 2013/10/08. doi: 10.1038/ng.2797 ng.2797 [pii]. PMID: 24097068; PubMed Central PMCID: PMC3838666.
20. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999; 55(4):997–1004. Epub 2001/04/21. PMID: 11315092.
21. Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, et al. Gene and pathway-based second-wave analysis of genome-wide association studies. Eur J Hum Genet. 2010; 18(1):111–7. Epub 2009/07/09. doi: 10.1038/ejhg.2009.115 ejhg2009115 [pii]. PMID: 19584899; PubMed Central PMCID: PMC2987176.
22. Whitlock MC. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J Evol Biol. 2005; 18(5):1368–73. Epub 2005/09/02. JEB917 [pii] doi: 10.1111/j.1420-9101.2005.00917.x PMID: 16135132.
23. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc Ser B. 1995; 57:289–833.
24. Chen Z. Is the weighted z-test the best method for combining probabilities from independent tests? J Evol Biol. 2011; 24(4):926–30. Epub 2011/03/16. doi: 10.1111/j.1420-9101.2010.02226.x PMID: 21401770.
25. Rosenbaum PR. Bahadur Efficiency of Sensitivity Analyses in Observational Studies. Journal of the American Statistical Association. 2015; 110(509).

26. He X, Shao Q-M. Bahadur efficiency and robustness of studentized score tests. Annals of the Institute of Statistical Mathematics 1996; 48(2):295–314.

27. Dai H, Charnigo R. D_CDF Test of Negative Log Transformed P-values with Application to Genetic Pathway Analysis. Statistics and Its Interface. 2014.

28. Genovese CR, Roeder K, Wasserman L. False discovery control with p-value weighting. Biometrika. 2006; 93:509–24.

29. Benjamini Y, Hochberg Y. Multiple hypothesis testing with weights. Scandinavian Journal of Statistics. 1997; 24:407–17.

30. Savage IR. Nonparametric Statistics: A Personal Review. Sankhya. 1969; 31:107–44.

31. Killeen TJ, Hettmansperger TP, Sievers GL. An elementary theorem on the probability of large deviations. Annals of Mathematical Statistics. 1972; 43:181–92.