Accuracy Evaluation of the Unified *P*-Value from Combining Correlated *P*-Values

Gelio Alves, Yi-Kuo Yu*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America

Abstract

Meta-analysis methods that combine *P*-values into a single unified *P*-value are frequently employed to improve confidence in hypothesis testing. An assumption made by most meta-analysis methods is that the *P*-values to be combined are independent, which may not always be true. To investigate the accuracy of the unified *P*-value from combining correlated *P*-values, we have evaluated a family of statistical methods that combine: independent, weighted independent, correlated, and weighted correlated *P*-values. Statistical accuracy evaluation by combining simulated correlated *P*-values showed that correlation among *P*-values can have a significant effect on the accuracy of the combined *P*-value obtained. Among the statistical methods evaluated those that weight *P*-values compute more accurate combined *P*-values than those that do not. Also, statistical methods that utilize the correlation information have the best performance, producing significantly more accurate combined *P*-values. In our study we have demonstrated that statistical methods that combine *P*-values based on the assumption of independence can produce inaccurate *P*-values when combining correlated *P*-values, even when the *P*-values are only weakly correlated. Therefore, to prevent from drawing false conclusions during hypothesis testing, our study advises caution be used when interpreting the *P*-value obtained from combining *P*-values of unknown correlation. However, when the correlation information is available, the weighting-capable statistical method, first introduced by Brown and recently modified by Hou, seems to perform the best amongst the methods investigated.

Citation: Alves G, Yu Y-K (2014) Accuracy Evaluation of the Unified *P*-Value from Combining Correlated *P*-Values. PLoS ONE 9(3): e91225. doi:10.1371/journal.pone.0091225

Editor: Frank Emmert-Streib, Queen's University Belfast, United Kingdom

Received December 3, 2013; **Accepted** February 9, 2014; **Published** March 24, 2014

Funding: This work was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health/Department of Health and Human Services. Funding for Open Access publication charges for this article was provided by the National Institutes of Health. 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.

* E-mail: yyu@ncbi.nlm.nih.gov

Introduction

Meta-analysis methods that combine *P*-values into a single unified *P*-value are commonly used to rank or score a list of hypotheses [1]. For each hypothesis tested, the *P*-values to be combined are often acquired from studying different features associated with the hypothesis or from using different data analysis methods (DAM) to analyze a chosen feature. Either approaches conducted to test the same list of hypotheses assign an overall *P*-value to each hypothesis tested. These *P*-values are then usually sorted, with the most significant result ranking first in the list. Given that different features may not be completely independent and that different DAMs may share protocols and use similar information, it is likely that the *P*-values obtained for a hypothesis are correlated.

Most *P*-value combining methods assume that the *P*-values to be combined are independent or weakly correlated [2,3]. When the unified *P*-value is computed by combining correlated *P*-values, without properly taking into account the correlation, there can be notable effects in the significance assignment of the hypothesis tested. As the *P*-values to be combined are possibly correlated, it is important to investigate the effect that correlation has on the unified *P*-value. The current study is designed to evaluate the accuracy of the unified *P*-value computed by combining (positively) correlated *P*-values using some commonly applied statistical methods. By *P*-value accuracy, we mean how well on average does reported *P*-value agree with the one-sided cumulative distribution function of the random variable (associated with the null hypotheses tested) at the critical region. In other words, accurate *P*-value means that when one controls type-I error rate at a level α, the type-I error rate is really controlled at the level α. To keep this paper focused, we will not provide a lengthy introduction. For methods that we will evaluate, more details are provided in the Methods sections. For others, we will only provide the readers with appropriate references.

Several studies have been performed to evaluate methods that combine independent *P*-values [4–10]. For example, Rosenthal has evaluated nine methods for combining *P*-values and has summarized advantages, limitations and applications for each method [4]. Loughin [5] has also conducted a systematic comparison of methods for combining *P*-values and recommended practitioners to choose a method based on the structure and expectation for the problem being studied. Recently, Whitlock [6] has showed that the weighted Z-method has more power and precision than Fisher's test. In other studies, Chen [8] as well as Chen and Nadarajah [9], have shown that either the generalized Fisher method due to Lancaster or a special case of Lancaster's test outperform the weighted Z-method, while Zaykin [10] has shown that the weighted Z-method has similar power to Lancaster's method when the weights are selected to be the square roots of sample sizes.

As for combining correlated *P*-values, only few studies have been conducted to evaluate the accuracy of the unified *P*-value computed by existing statistical methods [11,12]. Evidently, more comprehensive investigations that incorporate different methods, encompass a wide range of correlation strength, and have a large number of simulations can further our understanding on the effect of correlation has on computing a unified *P*-value. To advance towards this direction, we systematically investigate a family of statistical methods for combining *P*-values. Because we are interested in combining *P*-values obtained from the right-tailed tests, we have limited our study to methods that combine *P*-values based on the normal distribution (e.g. Stouffer's method) and on the Chi-square distribution (e.g. Fisher's method), the general purpose method and the right-tail method recommended by Loughin [5]. The two aforementioned methods, aside from being frequently used to combine *P*-values, are useful and important to study for the following reason. Both methods mentioned have variations that weight *P*-values while computing the combined *P*-value: Lipták, Good and Bhoj methods [13–15], and variations that take into account the correlation among *P*-values: Hartung and Hou methods [16,17]. In addition, all methods mentioned above either have closed-form formulas, i.e., distribution functions, or approximation formulas that can provide the unified *P*-value with minimum computation cost.

In summary, our study presents an accuracy evaluation of the unified *P*-value obtained from statistical methods designed to combine independent, weighted independent, correlated, and weighted correlated *P*-values. We have evaluated the accuracy of the unified *P*-value from combining positively correlated *P*-value vectors with correlation among *P*-value vectors in the range [0,1]. Our results show that methods designed to combine independent *P*-values but with the capability of assigning weights to *P*-values perform better than methods that combine independent *P*-values without weights. Also methods that take into account the correlation between *P*-values perform significantly better than methods designed to combine independent *P*-values. Based on this study, the method first introduced by Brown [18] to combine correlated *P*-values and later adapted to include weights by Hou [17] is the best performing one amongst the methods investigated.

Methods

The main task of combining *P*-values is described below. Given a list of hypotheses $\mathcal{H}=\{H_1,H_2,H_3,\cdots,H_{\mathcal{K}}\}$, let each hypothesis have m *P*-values associated with it. These $m\times\mathcal{K}$ *P*-values can be organized as m *P*-value vectors, $\mathrm{P}_1,\mathrm{P}_2,\ldots,\mathrm{P}_m$, each having $\mathcal{K}$ components. Each *P*-value vector may result from analyzing one out of m different features of every hypothesis or may be from analyzing a single feature using one of the m different DAMs. The m *P*-values associated with hypothesis H_i are $\{P_1(i),P_2(i),\ldots,P_m(i)\}$. Given those values, one needs to combine them to form a single unified *P*-value. This scenario can occur in many applications. As an example, when different studies are performed to test a set of genetic loci for allelic imbalance [19], the number of genetic regions tested will correspond to the number of hypotheses $\mathcal{K}$ and each region will carry with them m *P*-values, one from each of the m studies. To fairly rank these possible $\mathcal{K}$ regions, for each region one would need a unified *P*-value resulting from combining the m *P*-values associated with it. For database search based peptide identification using mass spectrometry, it is possible to analyze the data using multiple analysis methods. Here for each experimental spectrum, the number of hypotheses tested $\mathcal{K}$ equals the number of scored peptides in the database and each peptide receives a *P*-value from each of the m analysis methods. To fairly rank the candidate peptides, it is again natural to combine the m *P*-values associated with each scored peptide [3] to reach a unified *P*-value. In the sequence homology detection where multiple motifs are used as a query to a sequence database, it is often needed to combine the *P*-values, each from one of the m motifs, to assign the statistical significance to a sequence in the sequence database [2]. In this case, $\mathcal{K}$ is the number of sequences in the database, while m is the number of motifs used as the query.

To make the notation uniform, we will use F_s and F_s^{-1} to represent the cumulative distribution and inverse cumulative distribution. When the subscript $s=n,\chi,\gamma$, F_s represents respectively the cumulative Normal, Chi-squared, and Gamma distributions. All the parameters of these distributions will be shown as arguments enclosed by a pair of parentheses following the symbol F.

Combining Independent *P*-values

We begin this subsection with a brief introduction of Stouffer's (Z-transform test) and Fisher's (Chi-square test) methods. Generalizations of both methods to combine weighted *P*-values are also described.

Method 1. The combined Z-transform test was first used by Stouffer *et al.* [20] and later generalized to include weights by Lipták [13]. Under the null hypothesis, the *P*-values are uniformly distributed between [0,1]. Given a list of *P*-values $(p_{(H_i,1)},p_{(H_i,2)},\cdots,p_{(H_i,m)})$ associated with a given H_i, one transforms the *P*-values to a new variable $(x_{H_i,j})$ by a simple transformation

$$x_{H_i,j}=F_n^{-1}(1-p_{(H_i,j)}),\qquad 1\leq j\leq m\ ,$$

where F_n^{-1} stands for the inverse of the cumulative normal distribution. For the Z-transform test the distribution function used is the standard Normal (Gaussian) distribution with probability density function given by

$$\mathrm{pdf}_n(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

with parameters $\mu=0$ and $\sigma=1$.

Stouffer's way to combine the above *P*-values is by defining a new variable

$$\tau'=\frac{\sum_{j=1}^{m}x_{(H_i,j)}}{\sqrt{m}},$$

which is also Gaussian distributed with *P*-value given by the formula

$$P(\tau'\geq\tau)=\int_{\tau}^{\infty}e^{-\frac{t^2}{2}}\frac{dt}{\sqrt{2\pi}}\qquad(1)$$

A generalization of the above equation that assigns weights (w_j) to the variable $x_{H_i,j}$ is know as the weighted Z-transform test [13]

$$\tau' = \frac{\sum_{j=1}^{m} w_j x_{(H_i,j)}}{\sqrt{\sum_{j=1}^{m} w_j^2}}.$$

The variable of the weighted Z-transform τ' also follows Normal distribution, and the formula for the *P*-value is also given by eq. (1).

Method 2. Fisher's method [21] is one of the most used method to combine independent *P*-values. The combined Fisher *P*-value is obtained through the following variable:

$$\tau' = -2\sum_{j=1}^{m} \ln\left(p_{(H_i,j)}\right),$$

which follows a Chi-squared distribution $\mathrm{pdf}_{\chi}(\tau;2m)$ with $2m$ degrees of freedom. Computing the unified *P*-value using the Chi-squared distribution is not the most efficient approach because of the significant computational cost in calculating the cumulative distribution F_{χ}. A more efficient way to obtain the unified *P*-value has been proposed [2,3], where the unified *P*-value of $\tilde{\tau}' \equiv e^{-\tau'/2}$ has a closed form given by

$$P(\tilde{\tau}' \le \tilde{\tau}) = \tilde{\tau} \sum_{v=0}^{m-1} \frac{[-\ln(\tilde{\tau})]^v}{v!}, \tag{2}$$

or in terms of the τ' variable

$$P(\tau' \ge \tau) = e^{-\tau/2} \sum_{v=0}^{m-1} \frac{\tau^v}{2^v\, v!}\ . \tag{3}$$

Note that as τ' increases $\tilde{\tau}'$ decreases and vice versa.

Fisher's method does not assign weights to the *P*-values to be combined. However, when information is available regarding how *P*-values were obtained, it might be beneficial to weight *P*-values. Lancaster *et al.* [22] addresses this issue by replacing the random variable $-2\ln(p_{(H_i,j)})$ with $F_{\chi}^{-1}(1-p_{(H_i,j)};d_j)$, a variable following a Chi-squared distribution with d_j degrees of freedom not necessarily equal to two.

In Lancaster's procedure, summarized below, one can exploit the equivalence between the Chi-squared distribution $\mathrm{pdf}_{\chi}(x;d)$ and the gamma distribution $\mathrm{pdf}_{\gamma}(x;\alpha=\frac{d}{2},\beta=\frac{1}{2})$ to reach a different weighting generalization. For hypothesis H_i, the variable τ' can now be written as

$$\tau' = \sum_{j=1}^{m} F_{\chi}^{-1}(1-p_{(H_i,j)};d_j) = \sum_{j=1}^{m} F_{\gamma}^{-1}(1-p_{(H_i,j)};d_j/2,1/2),$$

which evidently follows a Chi-squared distribution with $\sum_j^m d_j$ degrees of freedom. In the expression above, Fisher's method is recovered by setting $d_j=2$ for all j. Another way to incorporate weights is to keep $d_j=2$ while retaining a general β value. Specifically, one may choose, with w_j being the weight factor, to use the following new variable

$$\tau' = \sum_{j=1}^{m} w_j F_{\gamma}^{-1}(1-p_{(H_i,j)};1,\beta) = -\sum_{j=1}^{m} \frac{w_j}{\beta} \ln\left(p_{(H_i,j)}\right).$$

The *P*-value for τ' can be easily evaluated using the same technique as that in [3] and is given below

$$P(\tau' \ge \tau) = \sum_{j=1}^{m} \frac{e^{-\beta\tau/w_j} w_j^{m} - 1}{\prod_{v=1,v\neq j}^{m} (w_j - w_v)}. \tag{4}$$

Interestingly, with $\beta=1/2$, eq. (4) corresponds to the unified *P*-value of multiplying weighted independent *P*-values obtained earlier by Good [14]. This can be seen by the following observation. Good defined his variable

$$\tilde{\tau}' = \prod_{j=1}^{m} p_{(H_i,j)}^{w_j} = e^{-\tau'/2},$$

and the corresponding *P*-value is given by

$$P(\tilde{\tau}' \le \tilde{\tau}) = \sum_{j=1}^{m} \left[\frac{(\tilde{\tau})^{1/w_j} w_j^{m-1}}{\prod_{v=1,v\neq j}^{m} (w_j - w_v)}\right] \tag{5}$$

When expressed in the variable $\tau \equiv -2\ln\tilde{\tau}$, we easily see that

$$P(\tau' \ge \tau) = \sum_{j=1}^{m} \left[\frac{e^{-\tau/2w_j} w_j^{m-1}}{\prod_{v=1,v\neq j}^{m} (w_j - w_v)}\right],$$

in agreement with eq. (4) when $\beta=1/2$.

A question that arises naturally when using methods such as the weighted Z-transform's test, Good's test, and Lancaster's test is how to obtain the optimal weights (w_j)? This difficult question has been raised and it was suggested that the choice of weights may vary by cases [23]. Existing methods to assign/estimate the weights include, but are not limited to: (1) weight in proportion to the reciprocal of the variance estimated from each study [6], (2) estimate the weights from one's prior belief about a method or feature [24], (3) select weights to stabilize the variance of the combined test statistics [25], and (4) use weights that improve the testing power [26]. Because there is no universal procedure to compute the optimal weights to be used, in this study the weights, when used, were randomly generated and normalized to sum to one (see Table 1).

There are also two apparent problems with Lancaster's eq. (4) and Good's eq. (5). The first problem is that the weights used can't be identical, otherwise singularities can occur [14,15]. Second, if the difference between some of the weights are small, numerical instability can occur [15,17,27]. In order to address the problem of numerical instability associated with identical and almost identical weights, Bhoj [15] suggested an approximation using a linear combination of m gamma density functions (with $\tau' = -2\sum_{j=1}^{m} w_j \ln(p_{(H_i,j)})$)

Table 1. Breakdown of Methods Used to Combine *P*-values Investigated.

Method Name	Ref. number	Eq. number	Acc. weights	Nor. weights	Account for corr.
Fisher	[21]	3	no	none	no
Stouffer	[20]	1	no	none	no
Bhoj	[15]	6	yes	$\sum_{i=1}^{m} w_i = 1$	no
Good	[14]	5	yes	$\sum_{i=1}^{m} w_i = 1$	no
Lipták	[13]	1	yes	$\sum_{i=1}^{m} w_i^2 = 1$	no
Hartung	[16]	9	yes	$\sum_{i=1}^{m} w_i = 1$	yes
Hou	[17]	14	yes	$\sum_{i=1}^{m} w_i = 1$	yes

The first column of the table provides the names of the methods used to combine *P*-values investigated in our study. The second column lists the reference number cited in this paper for the publication (Ref) corresponding to the method used. The third column provides the equation number for the method distribution function used to compute the formula *P*-value. The fourth column indicates if a method equation can accommodate (acc.) weight when combining *P*-value. The fifth column gives the normalization (nor.) procedure used to normalize the weights. Finally, the last column conveys the information about a method's capability to account for correlation (corr.) between *P*-values.

doi:10.1371/journal.pone.0091225.t001

$$P(\tau' \geq \tau) = 1 - \sum_{j=1}^{m} \frac{w_j \gamma\left(\frac{1}{w_j}, \frac{\tau}{2w_j}\right)}{\Gamma\left(\frac{1}{w_j}\right)} = \sum_{j=1}^{m} w_j \left[1 - F_\gamma\left(\frac{\tau}{2w_j}; \frac{1}{w_j}\right)\right] \tag{6}$$

where $\gamma(a,x) = \int_0^x t^{a-1} e^{-t} dt$ is the incomplete gamma function and $\Gamma(a) = \gamma(a,\infty)$ is the gamma function. Although the approximation provided by Bhoj does reduce to Fisher's distribution when the weights are all equal and does not encounter singularities when weights are identical or nearly identical, this approximation does not lead to Good's distribution when the weights are all different. A recent publication [27] has provided an analytical formula that not only is numerically stable when combining *P*-values with nearly degenerate or identical weights but also correctly reproduces Fisher's and Good's results as limiting cases.

Combining Dependent *P*-values

In this subsection we summarize two statistical methods that are generalizations of Stouffer's test (Z-transform test) and Fisher's test (Chi-square test) that attempt to account for the correlation among *P*-values to be combined.

Method 3. Hartung [16] incorporates the correlation among *P*-values via introducing in the Z-transform test (eq. (1)) the correlation-matrix, with elements ρ_{jv} computed from the variable pairs $(x_{H_i,j}, x_{H_i,v})$, and by defining a new variable

$$\tau' = \frac{\sum_{j=1}^{m} w_j x_{H_i,j}}{\sqrt{(1 - \mathrm{E}[\rho]) \sum_{j=1}^{m} w_j^2 + \mathrm{E}[\rho] (\sum_{j=1}^{m} w_j)^2}},$$

where

$$\mathrm{E}[\rho] = 2 \frac{\sum_{j=1}^{m} \sum_{v>j}^{m} \rho_{jv}}{m(m-1)}, \tag{7}$$

and

$$\rho_{jv} = \frac{\sum_{i=1}^{\mathcal{K}} (x_{H_i,j} - \overline{x_{H_\bullet,j}})(x_{H_i,v} - \overline{x_{H_\bullet,v}})}{\mathcal{K}\ \sigma_j\ \sigma_v}, \tag{8}$$

where $\overline{x_{H_\bullet,j}} = \sum_{i=1}^{\mathcal{K}} x_{H_i,j} / \mathcal{K}$ is the average value of $x_{H_i,j}$, $\mathcal{K}$ is the total number of hypotheses tested, and σ_j^2 is the variance of $x_{H_i,j}$.

The *P*-value for τ' is then approximated by the standard Normal distribution

$$P(\tau' \geq \tau) \approx \int_{\tau}^{\infty} e^{-t^2/2} \frac{dt}{\sqrt{2\pi}}, \tag{9}$$

which nevertheless becomes exact in the two extreme limits of $\rho_{jv} = 1\ \forall j,v$ and $\rho_{jv} = 0\ \forall j \neq v$. Although in general the distribution of τ' is only approximately normal, it is arguable that ignoring correlation can cause more damage to the combined *P*-value than the deviations from normality. Applications and extensions of Hartung's idea can also be found in more recent publications [12,28].

Method 4. Following Satterthwaite's procedure [29], there have been some attempts, when combining correlated *P*-values, to obtain approximate unified *P*-value for the Fisher's variable (no weight) [18,30] and for the Good's variable (unequal weights) [17]. The main idea of Satterthwaite's procedure is to equate the first two moments of the uncharacterized distribution to that of a Chi-squared distribution. Brown [18] and Kost *et al.* [30] tried to approximate the distribution of the Fisher's variable

$$\tau' = \sum_{j=1}^{m} -2 \ln (p_{H_i,j}),$$

and Hou [17] the distribution of Good's variable

$$\tau' = \sum_{j=1}^{m} -2 w_j \ln (p_{H_i,j}),$$

to that of a Chi-squared distribution $\mathrm{pdf}_\chi(\tau'/c; f)$, with c being a scale factor to be determined.

The expectation value ($E[\tau']$) the variance ($V[\tau']$) of τ' by formal operation are given respectively by

$$E[\tau'] = E[\sum_{j=1}^{m} -2 w_j \ln (p_{H_i,j})] = 2 \sum_{j=1}^{m} w_j, \text{ and} \tag{10}$$

$$V[\tau'] = V[\sum_{j=1}^{m} -2w_j \ln(p_{H_i,j})]$$

$$= 4\sum_{j=1}^{m} w_j^2 + 2\sum_{j=1}^{m}\sum_{v=1,j<v}^{m} w_j w_v \,\mathrm{cov}(-2\ln(p_{H_i,j}), -2\ln(p_{H_i,v})). \quad (11)$$

On the other hand, the expectation value and variance of τ' using $\mathrm{pdf}_\chi(\tau'/c;f)$ yields

$$E[\tau'] = cf, \text{ and} \quad (12)$$

$$V[\tau'] = 2c^2 f. \quad (13)$$

Equating (10) to (12) and (11) to (13) yields

$$c = \frac{2\sum_{j=1}^{m} w_j^2 + \sum_{j=1}^{m}\sum_{v=1,j<v}^{m} w_j w_v \,\mathrm{cov}(-2\ln(p_{H_i,j}), -2\ln(p_{H_i,v}))}{2\sum_{j=1}^{m} w_j},$$

and

$$f = \frac{4(\sum_{j=1}^{m} w_j)^2}{2\sum_{j=1}^{m} w_j^2 + \sum_{j=1}^{m}\sum_{v=1,j<v}^{m} w_j w_v \,\mathrm{cov}(-2\ln(p_{H_i,j}), -2\ln(p_{H_i,v}))}.$$

The covariance (cov) term used above was first estimated by Brown [18] and recently an improved estimation (through numerically tabulating the covariance as a function of the correlation and then performing polynomial fits) was provided by Kost and McDermott [30]

$$\mathrm{cov}(-2\ln(p_{H_i,j}), -2\ln(p_{H_i,v})) = 3.263\rho_{jv} + 0.710\rho_{jv}^2 + 0.027\rho_{jv}^3,$$

where ρ_{jv} above is the correlation between $\ln(p_{H_i,j})$ and $\ln(p_{H_i,v})$. The *P*-value for τ' is then approximated by that of a Chi-squared distribution

$$P(\tau' \geq \tau) = 1 - F_\chi^{-1}(\tau/c;f). \quad (14)$$

Equation (14) reduces to Fisher's formula eq. (3) when the *P*-values are independent and the weights are all same. However, the above equation does not reduces to Good's formula eq. (5) when the *P*-values are independent and each carries a different weight.

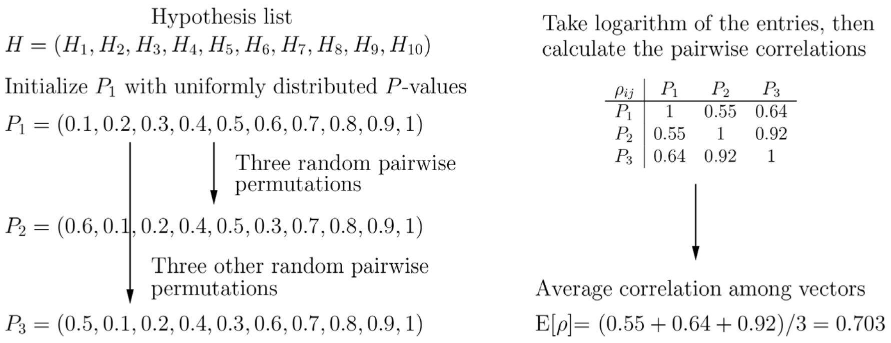
Generating Correlated P-value Vectors

By definition, the *P*-values of null hypotheses should be uniformly distributed between 0 and 1, which is often assumed by methods of combining *P*-values. However, the uniformity of *P*-values, when assigned by available statistical tools to a group of null hypotheses, is often lost. This would handicap the efficacy of methods for combining *P*-values from the start. To eliminate the effect of nonuniform null *P*-values from our evaluation, we enforce the quasi-uniformity of null *P*-values by first constructing a starter *P*-value vector $\tilde{\mathbf{P}}$ of size $\mathcal{K}$ with the *i*th element $\tilde{\mathbf{P}}(i) = i/\mathcal{K}$, for $1 \leq i \leq \mathcal{K}$. (See next paragraph for more details.) This guarantees an even sample of the *P*-values (in the range from $1/\mathcal{K}$ to 1). To achieve correlations of various strengths, we have used *P*-value vectors, each of which is obtained via permuting (pairwise) the elements of a fixed vector, the starter vector with a small perturbation, by a randomly chosen number. The basic idea is that when the number of pairwise permutations is not large, the resulting *P*-value vectors will be correlated to the fixed vector and will be correlated among one another. It is worth pointing out that this approach does not generate correlations with a *prescribed* strength: even with the same number of random pairwise permutations of the vector elements, the correlation between any pair of such *permuted* vectors does not have a fixed strength. We believe this is closer to the real-world scenario than having a fixed correlation strength among the *P*-value vectors. The value of $\mathcal{K}$ should not matter in terms of testing whether a method can provide accurate combined *P*-value. If a small $\mathcal{K}$ is used, however, the combined *P*-value will have a large statistical fluctuation that may reduce the resolution of the comparison. On the other hand, making $\mathcal{K}$ large causes a long computational time. We find that using $\mathcal{K} = 10{,}000$ yields enough separations among methods tested without significantly slowing down the computation.

For each method investigated, we have performed a simulation of 500,000 realizations, each of which was conducted as follows. First, pick a random positive integer r with $1 \leq r \leq \mathcal{K}/2$. Second, generate the first *P*-value vector $\mathbf{P}_1$ by adding a small random perturbation ($\pm\delta$) between 0 and 5×10^{-5} to each vector element of $\tilde{\mathbf{P}}$: $\mathbf{P}_1(i) \leftarrow \tilde{\mathbf{P}}(i) \pm \delta_i$. Evidently, by increasing the upper bound for δ, one will produce **P**-values with larger variations from exactly uniform distribution. In the third step, generate more size-$\mathcal{K}$ vectors $\mathbf{P}_2, \mathbf{P}_3, \cdots, \mathbf{P}_m$ and initialize them to $\mathbf{P}_1$. For each vector generated, its vector elements are pairwise permuted r (chosen at the first step) times. After that using $-\ln \mathbf{P}_j(i)$ in place of $x_{H_i,j}$ the pairwise correlation ρ_{jv} was computed using eq. (8) and the average correlation $\mathrm{E}[\rho]$ among vectors was computed using eq. (7). This work flow is illustrated in Figure 1 with $\delta(i) \rightarrow 0$ for simplicity. The constructed random **P**-value vectors $\mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_m$ were then combined to obtain a unified *P*-value vector (**F**) using the various methods listed in Table 0. Once the unified *P*-value vector (**F**) was calculated, its elements were sorted in increasing order and it was then compared against the rank (**R**) vector, whose element is obtained by dividing the rank of a **F** element by $\mathcal{K}$, i.e., $R(i) = i/\mathcal{K}$ for i ranging from 1 to $\mathcal{K}$. We shall call $\mathbf{R}(i)$, the *i*th element of the rank vector, the *normalized rank* of rank *i*.

Statistical Accuracy Evaluation of the Combined *P*-value (F)

If a method yields a unified *P*-value vector **F** agreeing with **R**, the scatter plot of $\mathbf{F}(i)$ versus $\mathbf{R}(i)$ should produce a straight line with slope one and intercept zero [31]. It is also important to mention that the smallest computed *P*-value is expected to be inversely proportional to the sample size, which for the current case is of the order of 10^{-4}. An example of a logarithmic plot of **F** versus **R** generated from a single iteration of our simulation is shown in Figure 2. Using the textbook definition of *P*-value, the linear slope obtained from the logarithmic plot of **R** versus **F** should be approximately one for methods with accurate statistics. To quantify how well **F** agrees with **R** we use four measures: (1) the average weighted sum of squares error ($AWSE$), (2) the distance (D) between **F** and **R**, (3) the expected rank $\mathrm{E}[R(F_c)]$, and (4)the expected error of $F(i)$. Figure 2 also illustrates what is being computed by the above four measures.


Hypothesis list
$H=(H_1,H_2,H_3,H_4,H_5,H_6,H_7,H_8,H_9,H_{10})$
Initialize P_1 with uniformly distributed P-values
$P_1=(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1)$
Three random pairwise permutations
$P_2=(0.6,0.1,0.2,0.4,0.5,0.3,0.7,0.8,0.9,1)$
Three other random pairwise permutations
$P_3=(0.5,0.1,0.2,0.4,0.3,0.6,0.7,0.8,0.9,1)$
Take logarithm of the entries, then calculate the pairwise correlations
Average correlation among vectors
$\mathrm{E}[\rho]=(0.55+0.64+0.92)/3=0.703$

ρ_{ij}	P_1	P_2	P_3
P_1	1	0.55	0.64
P_2	0.55	1	0.92
P_3	0.64	0.92	1

Figure 1. Example workflow of generating correlated P-values and pairwise correlations. In this example figure, $\mathcal{K}$ is 10, the number of P-value vectors is $m=3$, the number of pairwise permutations $r=3$, and the perturbations $\delta(i)$s are set to zero for clarity and simplicity. The resulting pairwise correlations by using $-\ln P_j(i)$ in place of $x_{H_i j}$ are displayed in a symmetric matrix form.
doi:10.1371/journal.pone.0091225.g001

Average Weighted Sum of Squares Error. We define the average weighted sum of squares error as

$$AWSE=\frac{1}{\mathcal{K}}\sum_{i=1}^{\mathcal{K}}\frac{[\ln(R(i))-\ln(F(i))]^2}{R(i)}. \qquad (15)$$

The weight factor (w), $1/R(i)$, in the above equation was chosen so that each point in the transformed variable domain carries the same contribution to the $AWSE$. By construction, the P-values in the random vector R are uniformly distributed between $[10^{-4},1]$. However, once we make the logarithmic transformation, $y_i=-\ln(R(i))$, we find the new variable y to be exponentially distributed, i.e., $\mathrm{pdf}(y)=e^{-y}$. One may thus introduce $w(y)$, a weight factor making $w(y)\mathrm{pdf}(y)=1$, to compensate the non-uniformity in y. This leads to $w(y_i)=e^{y_i}=1/R(i)$, the weight factor used in eq. (15).

Angular Distance Between F and R. To compute the distance between **F** and **R**, we began by first computing the slope (b) of the logarithmic plot of **R** versus **F** using a weighted least-square regression, which aims to minimize the weighted sum of squares error (WSE)

$$WSE=\sum_{i=1}^{\mathcal{K}}\frac{[\ln(R(i))-a-b\ln(F(i)]^2}{R(i)}.$$

Taking the derivative of the above expression with respect to a and b and setting them equal to zero gives the following equations:

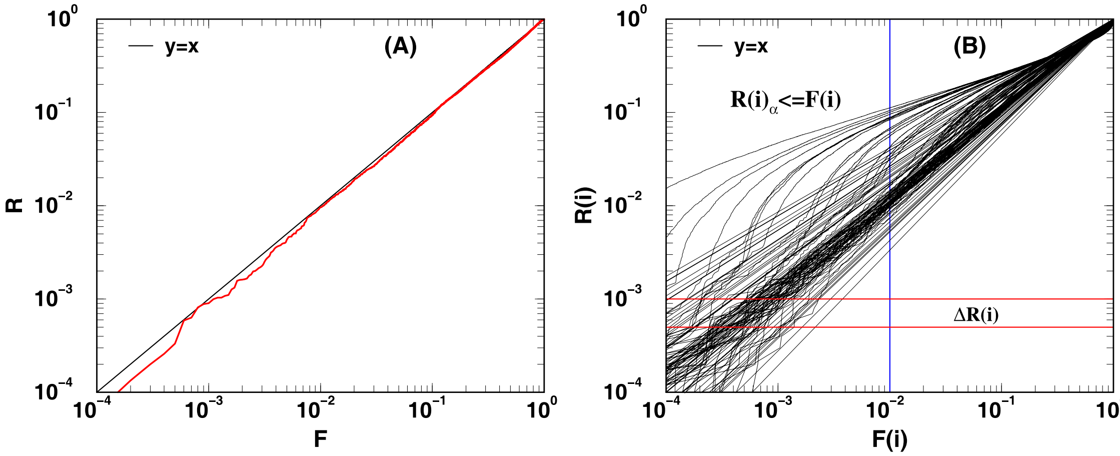


Figure 2. Log-log plot of the unified P-value vector F versus the rank vector R. The curves in panels (A) and (B) were obtained from combining the P-values of four P-value vectors, each of size 10,000, using Stouffer's method. In panel (A), the red circles show the scatter plot of normalized rank versus computed P-value from a randomly picked iteration (realization) of very weak average correlation. It is through curves like the one displayed in panel (A) that enables one to calculate the average sum of squares error using eq. (15) and the distance measure using eq. (16). Panel (B) shows 1000 curves, each of which is obtained from performing the same task as that leads to the curve in (A) but with different average correlation strengths. The lines that go significantly above $y=x$ line are from cases with stronger average correlations. They yield unified P-values that are much exaggerated perhaps due to the fact that the Stouffer's method does not account for correlations. By averaging the normalized rank R along the blue line ($F=F_c$) yields the value $E[R(F_c)$ (see eq. (17)). By shifting the blue line to different F_c values renders the entire $E[R(F_c)]$ versus F_c curve. The red horizontal line illustrates the case when $i=10$ (or normalized rank $R(10)=10^{-3}$). By averaging the $-\ln F$ values along this line, the $E[\ln(\frac{R(i)}{F(i)})]$ value is obtained for $i=10$ by simply adding $\ln(10^{-3})$ to the averaged value (see eq. (18)).
doi:10.1371/journal.pone.0091225.g002

$$\frac{\partial(WSE)}{\partial a}=\sum_i\frac{\ln(R(i))}{R(i)}-b\sum_i\frac{\ln(F(i))}{R(i)}-a\sum_i\frac{1}{R(i)}=0$$

and

$$\frac{\partial(WSE)}{\partial b}=\sum_i\frac{\ln(R(i))\ln(F(i))}{R(i)}$$
$$-b\sum_i\frac{\ln(F(i))\ln(F(i))}{R(i)}-a\sum_i\frac{\ln(F(i))}{R(i)}=0.$$

Solving the above two equations simultaneously for a and b gives

$$a=\frac{\sum_i\frac{\ln(R(i))}{R(i)}}{\sum_i\frac{1}{R(i)}}-b\frac{\sum_i\frac{\ln(F(i))}{R(i)}}{\sum_i\frac{1}{R(i)}}=M_R-bM_F$$

where M_R and M_F are the weighted average of $\ln(R)$ and $\ln(F)$ respectively and

$$b=\frac{\sum_i\frac{1}{R(i)}[\ln(R(i))-M_R][\ln(F(i))-M_F]}{\sum_i\frac{1}{R(i)}[\ln(R(i))-M_R]^2}.$$

From b and a, a normalized vector $\mathbf{V_F}=(\frac{1}{\sqrt{1+b^2}},\frac{b}{\sqrt{1+b^2}})$ was computed using the points $(0,a)$ and $(1,b+a)$ along the regression line. Similarly another normalized vector $\mathbf{V_R}=(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}})$ was obtained using the points (0,0) and (1,1) along the ideal line. Finally, the (angular)distance between the two unit vectors $\mathbf{F}$ and $\mathbf{R}$ was computed

$$D=\sqrt{2(1-\mathbf{V_F}\cdot\mathbf{V_R})}=\sqrt{2(1-\frac{1+b}{\sqrt{2+2b^2}})}. \qquad (16)$$

Methods with accurate statistics are expected to have $b=1$ and $a=0$. Evidently, $b=1$ leads to $D=0$ (see eq. (16)). The independence of the angular distance D on the intercept parameter a implies that D only measures the relative accuracy of the *P*-value, not the absolute accuracy. For example, if $\mathbf{F}(i)=\eta\mathbf{R}(i)$, even when the positive constant η is different from 1, D is still zero.

Expected Rank E[$R(F_c)$]. For iteration $1\leq\alpha\leq(\mathcal{N}=500{,}000)$, we denote by $R_\alpha(F_c)$ the largest normalized rank whose corresponding reported *P*-value is less than or equal to a selected cutoff *P*-value F_c. The expected rank E[$R(F_c)$] is computed by averaging R_α over all realizations and can be written as

$$E[R(F_c)]=\frac{1}{\mathcal{N}}\sum_{\alpha=1}^{\mathcal{N}}R_\alpha(F_c)\ . \qquad (17)$$

In the ideal case of absolute accuracy, $R_\alpha(F_c)=F_c$. In reality, this is hardly the case and that is why we use the expectation value of $R_\alpha(F_c)$ versus F_c as the measure. For methods with accurate statistics a plot of E[$R(F_c)$] versus F_c should trace closely the line $y=x$.

Expected Error of $F(i)$. The expected error of $F(i)$ relative to $R(i)=i/\mathcal{K}$ (for $1\leq i\leq\mathcal{K}$) is defined as

$$E[\ \ln(\frac{R(i)}{F(i)})\ ]=\frac{1}{\mathcal{N}}\sum_{\alpha=1}^{\mathcal{N}}\ln(\frac{R(i)}{F_\alpha(i)}), \qquad (18)$$

and the standard deviation

$$\sigma[\ \ln(\frac{R(i)}{F(i)})\ ]=\sqrt{\frac{1}{\mathcal{N}}\sum_{\alpha=1}^{\mathcal{N}}\{\ \ln(\frac{R(i)}{F_\alpha(i)})-E[\ln(\frac{R(i)}{F_\alpha(i)})]\ \}^2}. \qquad (19)$$

For methods with accurate statistics, plotting $E[\ \ln(\frac{R(i)}{F(i)})\ ]$ versus R should track the line $y=0$ well and have small standard deviations for various $R(i)$.

Results and Discussion

The four measures mentioned in the methods section are used to evaluate the accuracy of the unified *P*-value computed. In Figures 3, 4, 5 and 6, we show the results of combining a list of 12 *P*-values. The layout of each of these figure is identical. For each method considered, our simulation includes a total of $\mathcal{N}=500{,}000$ iterations. At each iteration, we generated $\mathcal{K}$ lists, within which the ith list is obtained by taking the i entry of each of the 12 *P*-value vectors, $(\mathrm{P}_1,\mathrm{P}_2,\cdots,\mathrm{P}_{12})$. By computing the pairwise correlation (see eq. (8)) among the *P*-value vectors, one obtains the average pairwise correlation E[ρ] given by eq. (7). Each iteration, generating a 12-tuples of *P*-value vectors, thus yields an average correlation $E[\rho]$.

For Figures 3, 4, 5, 6, the data points in panels A and B respectively display the expected average sums of square errors (E[$ASWE$]) and expected distances (E[D]) versus E[ρ]. More specifically, every data point plotted with x-axis value $\rho_k=0.025+0.05*(k-1)$ represents an average of 25,000 iterations, each of which has its 12-tuple's average correlation $E[\rho]$ fall in the range of $\rho_k\pm0.025$. For panels C, D, E and F, each data point plotted is computed using all the $\mathcal{N}$ iterations from our simulation. The curves in panel C show the expected number of events with unified *P*-value computed less than or equal to a cutoff value F_c. For methods with accurate statistics, by the definition of *P*-value, a plot of E[$R(F_c)$] versus F_c should follow the line $y=x$. Panels D and E (and F for Figures 3 and 4) display the expected F value together with its standard deviation as a function of R. Similar plots for the combination of 4 and 8 *P*-value vectors can be found in File S1.

Figure 3 displays the results for methods that assume the the *P*-values to be combined are independent: Fisher's (eq. 3), Stouffer's (eq. 1) and Bhoj's (eq. 6) methods. These methods are expected to compute accurate combined *P*-values for E[ρ] ∼ 0, corresponding to the first few data points of panels A and B. The data points in

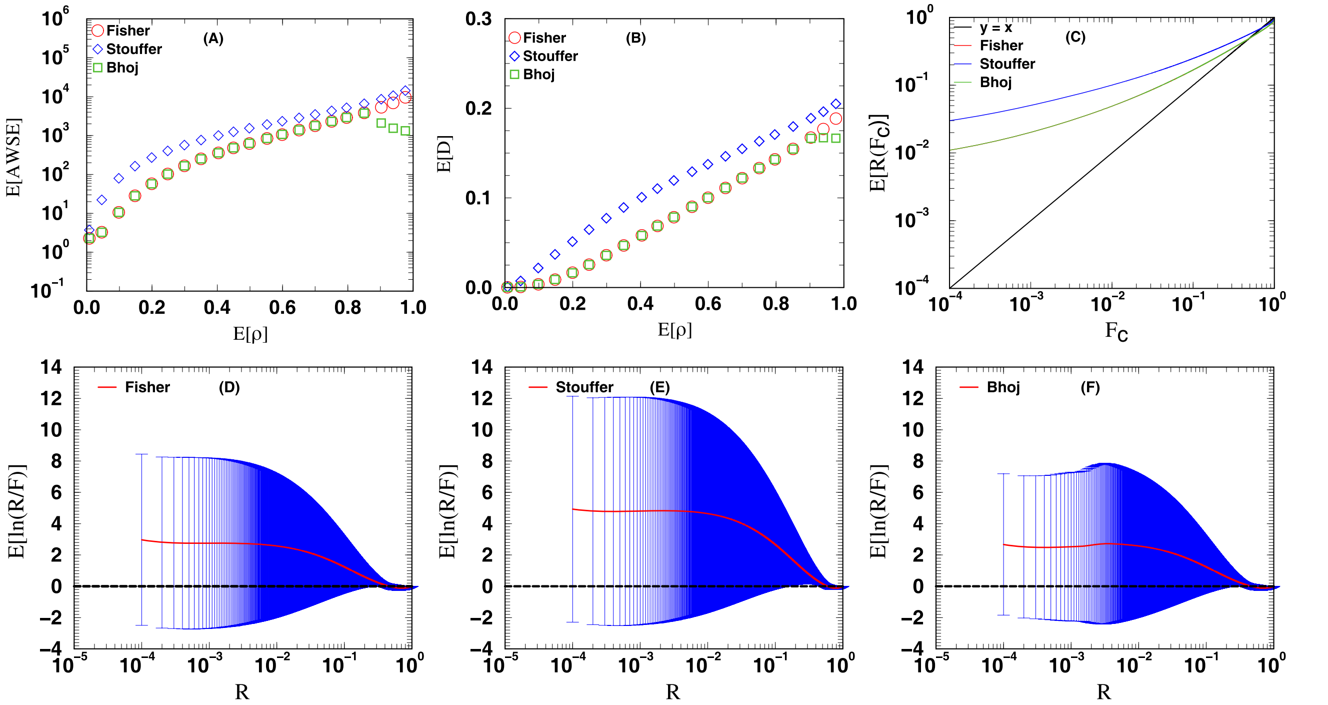


Figure 3. Methods that combine independent P-values: Fisher, Stouffer and Bhoj. The curves plotted above are the curves for the four different measures used to evaluate the accuracy of the computed P-value from combining the P-values of 12 P- value vectors.In panel C, note that the Fisher curve (red) is almost completely covered by the Bhoj curve (green). See text for more details.

doi:10.1371/journal.pone.0091225.g003

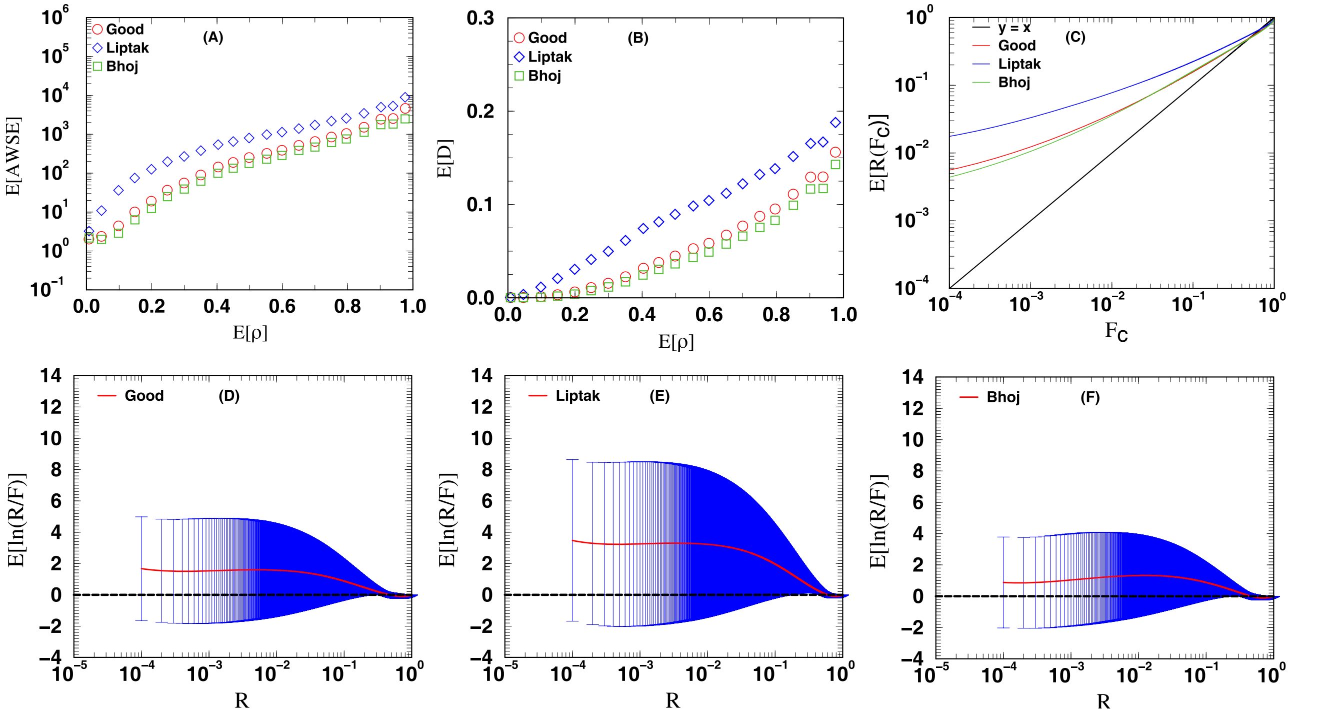


Figure 4. Methods that combine weighted independent P-values: Good, Lipták and Bhoj. The curves plotted above are the curves for the four different measures used to evaluate the accuracy of the computed P-value from combining the P-values of 12 P-value vectors. See text for more details.

doi:10.1371/journal.pone.0091225.g004

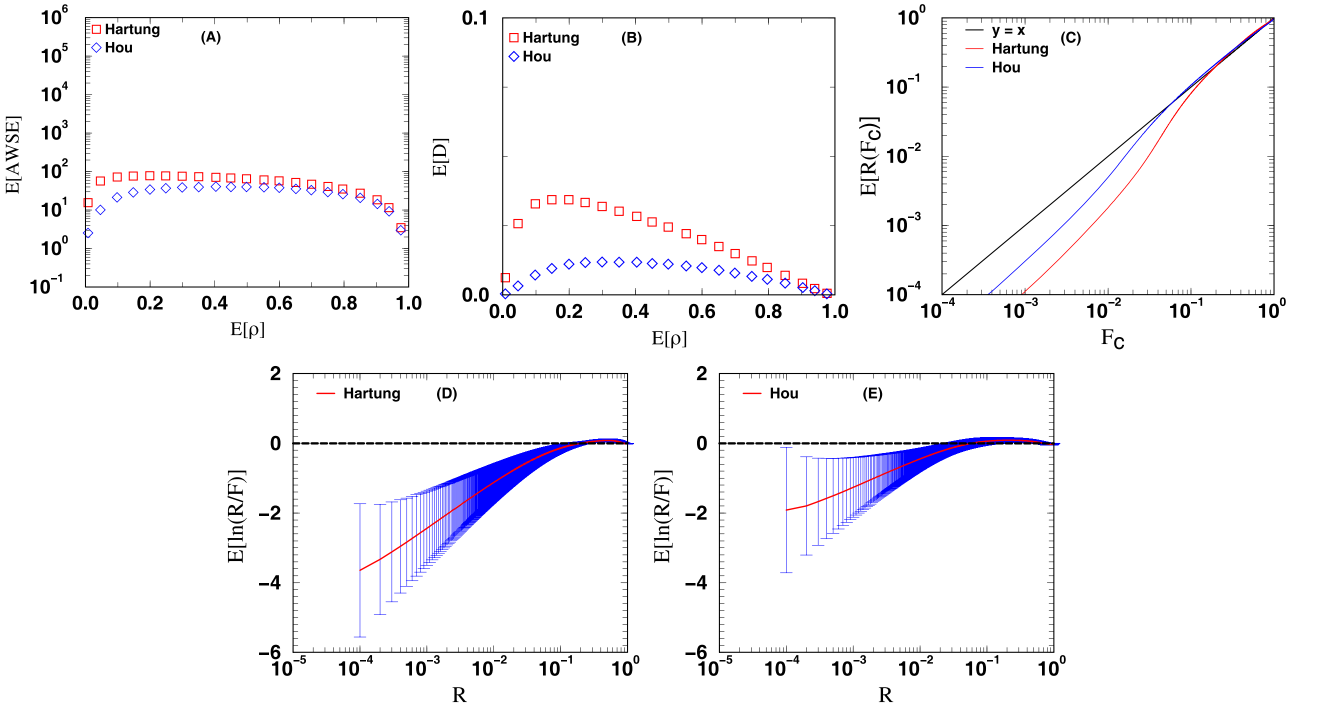


Figure 5. Methods that combine correlated *P*-values: Hartung and Hou. The curves plotted above are the curves for the four different measures used to evaluate the accuracy of the computed *P*-value from combining the *P*-values of 12 *P*-value vectors. See text for more details.
doi:10.1371/journal.pone.0091225.g005

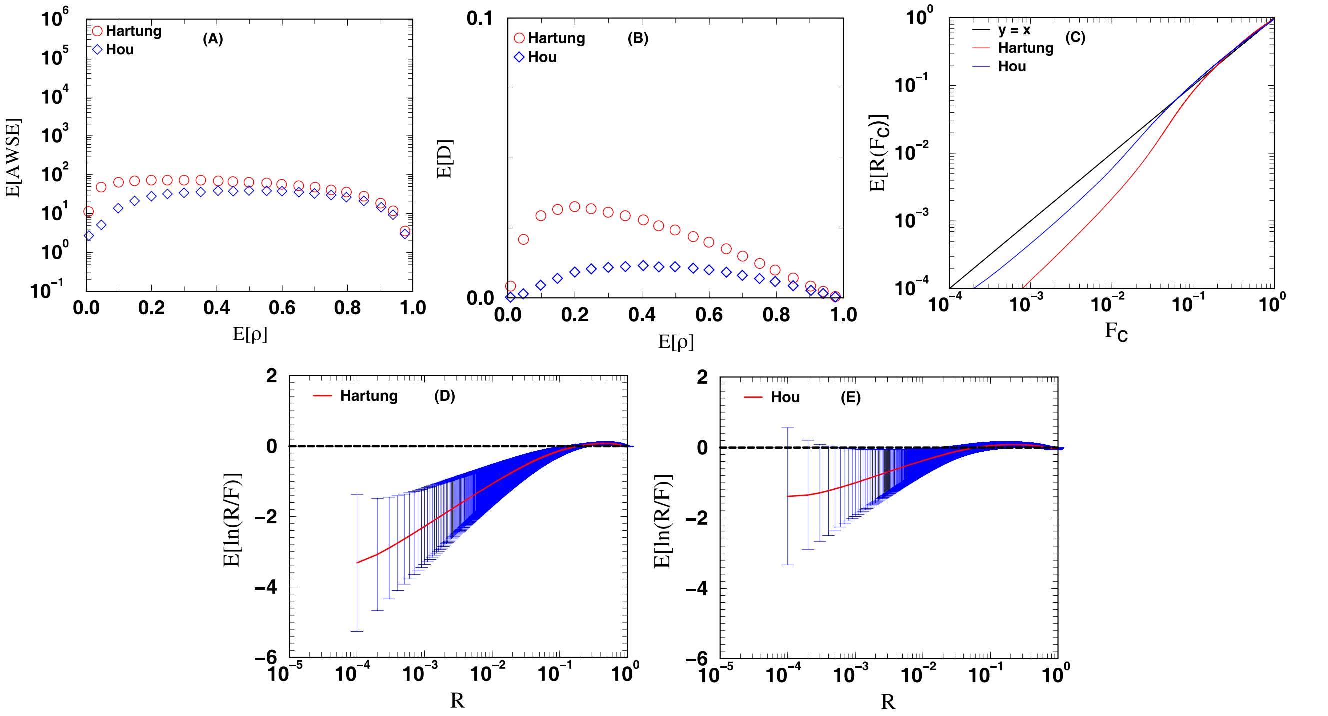


Figure 6. Methods that combine weighted correlated *P*-values: Hartung and Hou. The curves plotted above are the curves for the four different measures used to evaluate the accuracy of the computed *P*-value from combining the *P*-values of 12 *P*-value vectors. See text for more details.
doi:10.1371/journal.pone.0091225.g006

panels A and B show that as E[ρ] increases so does the E[$ASWE$] and E[D], indicating the methods' inadequacy for handling correlation among *P*-values. All three curves in panel C lie above the $y=x$ line, indicating that all three methods exaggerate significance when combining correlated *P*-values. The curves in panels D, E and F show that the average value (red solid curve) of $\ln(F/R)$ can deviate significantly from $y=0$ axis with wild fluctuations (error bars shown in blue). Also, a comparison with the plots obtained from combining 4, 8, and 12 *P*-value vectors indicates that the accuracy of the unified *P*-value decreases as the number of *P*-values combined increases from 4 to 12.

Figure 4 shows the results for methods that combine weighted independent *P*-values: Good's (eq. 5), Lipták's (eq. 1) and Bhoj's (eq. 6) methods. These three methods may be viewed as extensions of the previous three methods with *P*-value weighting enabled. Comparison of the panels of Figure 4 with that of Figure 3 shows noticeable improvement on the accuracy of the combined *P*-values. Although the accuracy has improved by weighting the *P*-values, the computed *P*-value still differs significantly from the expected value. The observed improvement suggests that weighting *P*-values might weaken the effect of correlation by promoting one *P*-value over the rest in the list of *P*-values to be combined. Other studies have also recommended [32,33] weighting *P*-values to improve statistical power. Even though weighting *P*-values is recommended, there exists no consensus on how to determine the optimal weights [6,24-26]. This is why in our simulation we have assigned random weights to the *P*-values to be combined. In principle, the accuracy of the computed *P*-value from the three methods above could be improved by using a different procedure to compute the weights. Such an investigation, although worth pursuing in its own right, is beyond the scope of the current study.

Figure 5 shows the results from using methods designed to combine correlated *P*-values: Hartung's (eq. 9) and Hou's (eq. 14) methods. The curves in Figure 5 when compared with the curves of Figure 3 and 4 show a significant improvement in the accuracy of the combined *P*-value computed. From the curves of Figure 5 Hou's method seems to be the better performing one, it has a smaller expected error and standard deviation when compared with the curves obtained from Hartung's method. As shown in panel C of Fig. 6, Hou's E[$R(F_c)$] vs F_c curve also traces reasonable well the line $y=x$, deviating from it only by a factor of about 4.0 for $F \leq 0.1$.

Finally, in Figure 6 we have the evaluation results of methods that combine weighted correlated *P*-values: Hartung's (eq. 9) and Hou's (eq. 14) methods. When the curves of Figure 6 are compared with that of Figure 5, as before it shows that weighting *P*-values tends to improve the accuracy of the the computed *P*-value The curves also show that Hou's method has a larger improvement in accuracy by using weights in comparison to Hartung's method. As articulated earlier and supported by the observed results, there is a possibility that the accuracy of the combined *P*-value could be further improved by having a statistically and mathematically rigorous procedure that could render the optimal weights to be used.

In a brief summary, methods designed for combining *independent* *P*-values tend to yield exaggerated *P*-values when used to combining correlated *P*-values. On the other hand, most methods designed to handle correlated *P*-values tend to provide conservative estimates for the unified *P*-values. The first case can be understood easily since one is effectively using nearly identical evidences to corroborate one another. For the latter case, however, we can not provide an intuitive interpretation except that it might result from the heuristics those methods employed. Weighting *P*-values seems to weaken the effect of correlation. This can be roughly understood as follows. By weighting each of the m *P*-values, only the *P*-values assigned the highest weights play a role. This increase the likelihood of having the highest weighted *P*-values be nearly independent, thereby reducing the effect of correlations. Not only does it help the methods designed for combining independent *P*-values, it also helps the ones for combining correlated *P*-values as most of these methods are heuristic-based and get more accurate results when the correlation is weaker. Based on these results, when the lists of the *P*-value vectors are complete, it is best to calculate the corresponding pairwise correlations between any two *P*-value vectors, introduce weights, and then assign the final unified statistical significance to each hypothesis.

In real applications, however, one is often faced with incomplete lists of *P*-values. That is, one only has the *P*-values for the highest ranking hypotheses, not for all hypotheses tested. This prevents one from computing the correlations needed for the formalism for combining correlated *P*-values. In this case, *i.e.*, when combining *P*-values of unknown correlation, one should exercise caution. Absent the correlation information, a better option might be to use the smallest of the *P*-values to be combined and then apply the Bonferroni correction by multiplying the smallest *P*-value by m, the number of *P*-values to be combined. This will guarantee a conserved statistics. However, under this approach, one might run into cases where the smallest *P*-values considered is larger than $1/m$, thereby obtaining a corrected *P*-value that is larger than 1. Even if each of the *P*-value lists is complete, there are still scenarios not covered in this paper. For example, it is possible that higher order correlations (such as the three-body or four-body) exist among the *P*-value vectors. We did not consider these cases since we are not aware of any readily available methods designed to deal with such type of higher order correlations.

In conclusion our study recommends that the unified *P*-value obtained from combining *P*-values of unknown correlation should be used with caution to prevent from drawing false conclusions. Results from our study agree with previous investigations [6,8,10], supporting the hypothesis that weighting *P*-values has the potential to improve the accuracy of the combined *P*-value. However, the important issues of choosing the weights to optimize a method's power and estimating the correlation matrix elements among *P*-values from small sample sizes remain challenging [34,35]. Our results also show that when combining independent or weighted independent *P*-values, Bhoj's method produces more accurate *P*-values than other methods tested. In the case when the correlation information is available, among the methods investigated, Hou's method, able to accommodate *P*-value weighting, seems to be the best performing method.

Supporting Information

File S1 This pdf file contains eight figures showing *P*-value accuracy evaluation of methods considered in this manuscript when combining 4 and 8 *P*-value vectors.
(PDF)

Acknowledgments

We thank the administrative group of the National Institutes of Health Biowulf Clusters, where all the computational tasks were carried out. We also thank the National Institutes of Health Fellows Editorial Board for editorial assistance.

Author Contributions

Conceived and designed the experiments: YKY. Performed the experiments: GA. Analyzed the data: GA YKY. Wrote the paper: GA YKY.

References

1. Olkin I (1995) Statistical and theoretical considerations in meta-analysis. J Clin Epidemiol 48: 133–146.
2. Bailey TL, Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches. Bioinformatics 14: 48–54.
3. Alves G, Wu WW, Wang G, Shen RF, Yu YK (2008) Enhancing peptide identification confidence by combining search methods. J Proteome Res 7: 3102–3113.
4. Rosenthal R (1978) Combining Results of Independent studies. Psychological Bulletin 85: 185–193.
5. Loughin TM (2004) A systematic comparison of methods for combining *p*-values from independent tests. Computational Statistics & Data Analysis 47: 467–485.
6. Whitlock MC (2005) Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J Evol Biol 18: 1368–1373.
7. Won S, Morris N, Lu Q, Elston RC (2009) Choosing an optimal method to combine P-values. Stat Med 28: 1537–1553.
8. Chen Z (2011) Is the weighted z-test the best method for combining probabilities from independent tests? J Evol Biol 24: 926–930.
9. Chen Z, Nadarajah S (2014) On the optimally weighted -test for combining probabilities from independent studies. Computational Statistics & Data Analysis 70: 387–394.
10. Zaykin DV (2011) Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. J Evol Biol.
11. Dudbridge F, Koeleman BP (2003) Rank truncated product of P-values, with application to genomewide association scans. Genet Epidemiol 25: 360–366.
12. Demetrescu M, Hassler U, Tarcolea AI (2006) Combining significance of correlated statistics with application to panel data. Oxford Bulletin of Economics and Statistics 68: 647–663.
13. Lipták P (1958) On the combination of independent tests. Magyar Tud Akad Nat Kutato int Kozl 3: 171–197.
14. Good IJ (1955) On the weighted combination of significance tests. Journal of the Royal Statistical Society Series B (Methodological) 17: 264–265.
15. Bhoj DS (1992) On the distribution of the weighted combination of independent probabilities. Statistics & Probability Letters 15: 37–40.
16. Hartung J (1999) A note on combining dependent tests of significance. Biometrical Journal 41: 849–855.
17. Hou CD (2005) A simple approximation for the distribution of the weighted combination of nonindependent or independent probabilities. Statistics & Probability Letters 73: 179–187.
18. Brown MB (1975) A method for combining non-independent, one-sided tests of significance. Biometrics 31: 987–992.
19. Vattathil S, Scheet P (2013) Haplotype-based profiling of subtle allelic imbalance with SNP arrays. Genome Res 23: 152–158.
20. Stouffer S, Suchman E, DeVinney L, Star S, Williams RMJ (1949) The American Soldier, Vol. 1: Adjustment during Army Life. Princeton: Princeton University Press.
21. Fisher RA (1932) Statistical Methods for Research Workers, vol. II. Edinburgh: Oliver and Boyd.
22. Lancaster HD (1961) The combination of probabilities: an application of orthogonal functions. Austr J Statist 3: 20–33.
23. Hedges L, Olkin I (1985) Statistical methods for meta-analysis. New York: Academic Press.
24. Zelen M, Joel LS (1959) The weighted compounding of two independent significance tests. The Annals of Mathematical Statistics 30: pp. 885–895.
25. Pepe MS, Fleming TR (1989) Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. Biometrics 45: 497–507.
26. Loesgen S, Dempfle A, Golla A, Bickeboller H (2001) Weighting schemes in pooled linkage analysis. Genet Epidemiol 21 Suppl 1: S142–147.
27. Alves G, Yu YK (2011) Combining independent, weighted p-values: Achieving computational stability by a systematic expansion with controllable accuracy. PLoS ONE 6: e22647.
28. Delongchamp R, Lee T, Velasco C (2006) A method for computing the overall statistical significance of a treatment effect among a group of genes. BMC Bioinformatics 7 Suppl 2: S11.
29. Satterthwaite FE (1946) An approximate distribution of estimates of variance components. Biometrics Bulletin 2: 110–114.
30. Kost JT, McDermott MP (2002) Combining dependent p-values. Statistics & Probability Letters 60: 183–190.
31. Schweder T, Spjotvoll E (1982) Plots of p-values to evaluate many tests simultaneously. Biometrika 69: 493–502.
32. Genovese CR, Roeder K, Wasserman L (2006) False discovery control with p-value weighting. Biometrika 93: 509–524.
33. Hu JX, Zhao H, Zhou HH (2010) False Discovery Rate Control With Groups. J Am Stat Assoc 105: 1215–1227.
34. Liechty JC, Liechty MW, Muller P (2004) Bayesian correlation estimation. Biometrika 91: 1–14.
35. Peng J, Wang P, Zhou N, Zhu J (2009) Partial correlation estimation by joint sparse regression models. Journal of the American Statistical Association 104: 735–746.