This paper proposes DANUBE, a new unbiased method of statistical meta-analysis that is applied to combine the results of multiple experiments performed for the same biological condition.

By Tin Nguyen, Cristina Mitrea, Student Member IEEE, Rebecca Tagett, and Sorin Draghici, Senior Member IEEE

ABSTRACT | Identifying the pathways and mechanisms that are significantly impacted in a given phenotype is challenging. Issues include patient heterogeneity and noise. Many experiments do not have a large enough sample size to achieve the statistical power necessary to identify significantly impacted pathways. Meta-analysis based on combining p-values from individual experiments has been used to improve power. However, all classical meta-analysis approaches work under the assumption that the p-values produced by experiment-level statistical tests follow a uniform distribution under the null hypothesis. Here, we show that this assumption does not hold for three mainstream pathway analysis methods, and significant bias is likely to affect many, if not all, such meta-analysis studies. We introduce DANUBE, a novel and unbiased approach to combine statistics computed from individual studies. Our framework uses control samples to construct empirical null distributions, from which empirical p-values of individual studies are calculated and combined using either a Central Limit Theorem approach or the additive method. We assess the performance of DANUBE using four different pathway analysis methods. DANUBE is compared to five meta-analysis approaches, as well as with a pathway analysis approach that employs multiple datasets (MetaPath). The 25 approaches have been tested on 16 different datasets related to two human diseases, Alzheimer’s disease (7 datasets) and acute myeloid leukemia (9 datasets). We demonstrate that DANUBE overcomes bias in order to consistently identify relevant pathways. We also show how the framework improves results in more general cases, compared to classical meta-analysis performed with common experiment-level statistical tests such as Wilcoxon and t-test.

KEYWORDS | Acute myeloid leukemia; alzheimer’s disease; empirical distribution; meta-analysis; pathway analysis; p-values

I. INTRODUCTION

The proliferation of high-throughput genomics technologies has resulted in an abundance of data, for many different biomedical conditions. Large public repositories such as Gene Expression Omnibus [1], [2]. The Cancer
The major drawback of the available p-value-based meta-analysis frameworks is that they work under the assumption that the p-values provided by the individual statistical tests follow a uniform distribution under the null hypothesis. Previous reports describe nonuniform distributions of p-values under the null as due to specific factors such as improper normalization, cross-hybridization, poorly characterized variance, and heteroskedasticity in microarray data analysis [33], [34], or even due to properties of some more general distributions [35]. Here, we show that this assumption also does not hold in the realm of pathway analysis methods, severely compromising the reliability of the results. In addition to strong statistical assumptions, the current methods for combining p-values are sensitive to outliers. For example, using Fisher’s method, a p-value of zero in one individual case will result in a combined p-value of zero regardless of the other p-values. The same is true for the minP and maxP statistics, where outliers greatly influence the combined p-value.

Here, we propose DANUBE (Data-driven meta-ANalysis using UnBiased Empirical distributions), a new meta-analysis framework that can combine the p-values of multiple studies in a better way. Our contribution is twofold. First, we use empirical null distributions to calculate p-values for individual studies. This approach learns from the data under the null hypothesis and compensates for any bias potentially introduced by an individual pathway analysis method. Second, we combine the individual p-values using a method based on the Central Limit Theorem. This is less sensitive to outliers and provides more reliable results. Our simulation experiments demonstrate that both type-I and type-II errors of DANUBE are better than those of classical meta-analysis approaches using both parametric and nonparametric tests.

We apply DANUBE in the context of pathway analysis using 16 public gene expression datasets from two biological conditions and four different pathway analysis methods. Gene Set Enrichment Analysis (GSEA) [36] and Gene Set Analysis (GSA) [37] are Functional Class Scoring methods [36]–[39], Down-weighting of Overlapping Genes (PADOG) [38] is an enrichment method [40]–[42], and Signaling Pathway Impact Analysis (SPIA) [43], [44] is a topology-aware method [43], [45]. These pathway analysis methods are applied on the human signaling pathways from KEGG [19], [20].

We show that with the exception of GSEA, each of the other three methods GSA, SPIA, and PADOG have different biases, leading to nonuniform distributions of p-values under the null hypothesis. Not surprisingly, when combining p-values using classical methods such as Fisher’s or the additive method, each of the three pathway analysis methods (GSA, SPIA, and PADOG) yields a very different list of significantly impacted pathways. We then apply the DANUBE framework using the empirical distributions characteristic to each of these methods. The DANUBE results yield much more consistent lists of significant pathways that are also pertinent to the phenotypes.
II. BACKGROUND

We first recapitulate the classical methods of combining p-values, such as Fisher’s method [24] and the additive method [25]–[27]. We then demonstrate the shortcomings of existing approaches in pathway analysis.

A. Fisher’s Method

Fisher’s method [24] is one of the most widely used methods for combining independent p-values. Considering a set of \( m \) independent significance tests, the resulting p-values \( P_1, P_2, \ldots, P_m \) are independent and uniformly distributed on the interval \([0, 1] \) under the null hypothesis. Denoting \( X_i = -2 \ln P_i \) (\( i \in \{1, 2, \ldots, m\} \)) as new random variables, the cumulative distribution function of \( X_i \) can be calculated as follows:

\[
F_i(x) = \Pr(X_i \leq x) = \Pr(-2 \ln P_i \leq x) = \Pr\left(P_i \geq e^{-x}\right) = \int_{e^{-x}}^{\infty} f(p)dp = 1 - e^{-x/2}.
\]

The above function is the cumulative distribution function of a chi-squared distribution with two degrees of freedom \( (\chi^2_2) \). Since the sum of chi-squared random variables is also a chi-squared random variable, \(-2 \sum_{i=1}^{m} \ln(P_i)\) follows a chi-squared distribution with \( 2m \) degrees of freedom \( (\chi^2_{2m}) \). In summary, the log product of \( m \) independent p-values follows a chi-squared distribution with \( 2m \) degrees of freedom:

\[
X = -2 \sum_{i=1}^{m} \ln(P_i) \sim \chi^2_{2m}. \quad (1)
\]

We note that if one of the individual p-values approaches zero, which is often the case for empirical p-values, then the combined p-value approaches zero as well, regardless of other individual p-values. For example, if \( P_1 \to 0 \), then \( X \to \infty \) and therefore, \( \Pr(X) \to 0 \) regardless of \( P_2, P_3, \ldots, P_m \). Therefore, we see that Fisher’s method is sensitive to outliers.

In practice, most pathway analysis methods use some kind of permutation or bootstrap approach to construct an empirical distribution of a statistic under the null. For example, the empirical null distribution of the \( t \) statistic is \( \xi = \{t_1, t_2, \ldots, t_N\} \). The empirical p-value calculated from such a distribution is the fraction of the statistics’ values in the \( N \) random trials performed that are more extreme than the observed one. Many times, there are no occurrences of values more extreme than the observed one, yielding an empirical p-value of zero. In this situation, the combined p-value calculated using Fisher’s method will be zero, even if all other p-values are equal to one. It is important to note that this phenomenon occurs because many methods choose to round the reported empirical p-value down to zero (when in fact, the real p-value is somewhere in the interval \( [0, 1/N] \)), and not because of the mathematical formulation of Fisher’s method.

B. Additive Method

The additive method proposes an alternative approach that uses the sum of p-values instead of the log product. Consider \( m \) random variables \( P_1, P_2, \ldots, P_m \) that are independent and uniformly distributed on the interval \([0, 1] \). Denoting \( X = \sum_{i=1}^{m} P_i \) as a new random variable, then \( X \) follows the Irwin–Hall distribution [26], [27]. The cumulative distribution function of \( X \) can be calculated as follows:

\[
F(x) = \frac{1}{2} + \frac{1}{2m!} \sum_{i=0}^{m} (-1)^i \binom{m}{i} (x - i)^m \text{sgn}(x - i). \quad (2)
\]

Using the above cumulative distribution function, we can calculate the probability of observing the sum \( X = \sum_{i=1}^{m} P_i \). We note that the concept of the additive method was also presented in [25] with a slightly different formulation and proof than in [26] and [27]. However, they are equivalent and can be transformed into one another.

The additive method is not as sensitive to extremely small individual p-values as Fisher’s method. However, both methods assume the uniformity of the p-values under the null hypothesis. We will show that this assumption does not hold for three mainstream pathway analysis methods. The inherent bias of these pathway analysis methods is most likely to affect the classical meta-analysis in most cases, and thus lead to systematic bias in identifying significant pathways.

C. Pitfalls of the Existing Approaches

Null distributions are used to model populations so that statistical tests can determine whether an observation is unlikely to occur by chance. The p-values produced by a sound statistical test must be uniformly distributed in the interval \([0, 1]\) when the null hypothesis is true [33]–[35], [46]. For example, the p-values that result from comparing two groups using a t-test should be distributed uniformly if the data are normally distributed [35]. When the assumptions of statistical models do not hold, the resulting p-values are not uniformly distributed under the null hypothesis. We will demonstrate this fact using gene expression data and pathway analysis.

Using only the control samples from seven publicly available Alzheimer’s datasets (\( N = 74 \)), we simulate 40 000 datasets as follows. We randomly label 37 as “control” samples and the remaining 37 as “disease” samples. We repeat this procedure 10 000 times to generate different groups of 37 control and 37 disease samples. To
make the simulation more general, we also create 10,000 datasets consisting of 10 control and 10 disease samples, 10,000 datasets consisting of 10 control and 20 disease samples, and 10,000 datasets consisting of 20 control and 10 disease samples. We then calculate the p-values of the KEGG (version 65) human signaling pathways (extracted as graph objects by the R package ROntoTools1.2.0 [44] version 1.2.0) using the following methods: GSEA [36], GSA [37], SPIA [43], [44], and PADOG [38].

Fig. 1 displays the empirical null distributions of p-values using GSA, SPIA, and PADOG. The horizontal axes represent p-values, while the vertical axes represent p-value densities. Blue panels (A0–A6) show p-value distributions from GSA, while purple (B0–B6) and green (C0–C6) panels show p-value distributions from SPIA and PADOG, respectively. For each method, the larger panel (A0, B0, and C0) shows the cumulative p-values from all KEGG signaling pathways. The small panels, six per method, display extreme examples of nonuniform p-value distributions for specific pathways. For each method, we show three distributions severely biased towards zero (e.g., A1–A3), and three distributions severely biased towards one (e.g., A4–A6).

These results show that, contrary to generally accepted beliefs, the p-values are not uniformly distributed for three out of the four methods considered. Therefore, one should expect a very strong and systematic bias in identifying significant pathways for each of these methods. Pathways that have p-values biased towards zero will often be falsely identified as significant (false positives). Likewise, pathways that have p-values biased towards one are likely to rarely meet the significance requirements, even when they are truly implicated in the given phenotype (false negatives). Systematic bias, due to nonuniformity of p-value distributions, results in failure of the statistical methods to correctly identify the biological pathways implicated in the condition, and also leads to inconsistent and incorrect results. For example, all three of the zero-biased GSA pathways shown in Fig. 1—Prostate cancer (A1), Adherens junction (A2), and Pathways in cancer (A3)—are reported as statistically significant by GSA (see Section IV), even though these data were collected in an experiment comparing Alzheimer’s disease patients versus healthy subjects, an experiment that has nothing to do with cancer.

In this section, we introduce the DANUBE framework and its application in the context of pathway analysis.

III. METHODS

In this section, we describe the DANUBE framework.

A. DANUBE Framework

We propose a new framework for meta-analysis that makes no assumptions on the data and is therefore expected to perform much better than any of the classical methods when the individual p-values are not distributed uniformly, as we have shown that it is the case for the pathway analysis methods. Fig. 2 displays a flowchart comparison between classical meta-analysis and DANUBE. Both approaches take m independent studies as input. The pipeline marked by blue arrows (I–II) shows the classical meta-analysis, and the one marked by black arrows (I–IV) is DANUBE.

The classical approach first calculates a p-value for each study using a parametric or nonparametric test, then it combines the individual p-values into one. The main limitation of the classical approach is that it relies on the assumption of uniformity of the p-values under the null hypothesis, which often does not hold true. As shown in Fig. 1, this assumption is not true for real transcriptomics data and KEGG pathways.

In the DANUBE framework, instead of modeling the data under a specific assumption, we construct empirical distributions and use them to calculate empirical p-values.
Following the black arrows (1–4) in Fig. 2, we initially calculate the values \( t_1, t_2, \ldots, t_m \) of the discriminating statistic for the \( m \) studies in step (1). For example, instead of using a statistical test to directly calculate the p-values, we could calculate the means of the data samples over the \( m \) studies. In step (2), we construct the empirical null distribution \( \xi_T \) for the chosen statistic. In step (3), we calculate the empirical p-values \( e_{p_1}, e_{p_2}, \ldots, e_{p_m} \) for the \( m \) studies with respect to the empirical null distribution \( \xi_T \). For all \( i \in \{1, 2, \ldots, m\} \), \( e_p \) is calculated as the number...
of elements in $\xi_T$ more extreme than $t_i$, divided by the total number of elements in $\xi_T$. We will prove that the resulting empirical p-values are uniformly distributed under the null hypothesis.

Lemma 1: Let $T$ be a random variable with the empirical distribution $\xi_T$ and the cumulative distribution function $F_T(T)$. We define the new random variable $X$ as follows:

$$X = \frac{\left\{x : x \in \xi_T \wedge x \leq t_i\right\}}{|\xi_T|}.$$  (3)

where the numerator represents the number of elements of $\xi_T$ that are smaller than or equal to $T$. If $\xi_T$ consists of enough data points to be considered as continuous, then $X$ is uniformly distributed on the interval $[0, 1]$.

Proof: Denote $F_T(T)$ as the cumulative distribution function of $T$. For any value $t \in \xi_T$, $F_T(t)$ can be calculated as follows:

$$F_T(t) = \frac{\left\{x : x \in \xi_T \wedge x \leq t\right\}}{|\xi_T|}.$$  (4)

We can see that $X = F_T(T)$. In addition, $F_T(t)$ is a strictly increasing function for all values $t \in \xi_T$. Let $F_X(X)$ be the

Fig. 2. DANUBE framework for meta-analysis. The blue arrows (I and II) show the classical meta-analysis pipeline, while black arrows (1–4) show the pipeline of DANUBE. The first step (I) of the classical approach is to perform a parametric or nonparametric test for each study. This step provides individual p-values which are independent and identically distributed (i.i.d.), but not necessarily uniformly distributed under the null, as shown in Fig. 1. The second step (II) of the classical approach is to use a classical method, such as Fisher’s, to combine the individual p-values, relying heavily on the assumption of uniformity under the null. In step (I) of DANUBE, we choose the discriminating statistic and calculate the values of this statistic in each study $(t_1, t_2, \ldots, t_m)$. In step (2), we generate the empirical distribution $\xi_T$ of the discriminating statistic under the null hypothesis. In step (3), we calculate the probability of observing $t_1, t_2, \ldots, t_m$ using $\xi_T$. In step (4), we combine the $m$ empirical p-values using either the additive method or the Central Limit Theorem (CLT).
cumulative distribution function of X, we have the following formula:

\[ F_X(x) = \Pr(X \leq x) = \Pr(F_T(T) \leq F_T(t)) = \Pr(T \leq t) = F_T(t) = x. \]  

(5)

We note that \( F_X(x) = x \) is the cumulative distribution function of the continuous uniform distribution on \([0, 1]\). Therefore, if we have enough data for \( F_T(T) \) to be considered continuous, then X will be a uniformly distributed random variable.

In step (4), we combine the empirical p-values using either the additive method or the Central Limit Theorem (CLT). According to Lemma 1, the resulting p-values after step (3) are now truly uniformly distributed under the null hypothesis and thus can be combined using the additive method as described in (2). However, the additive method can be computationally intensive when \( m \) is large. For this reason, we use the CLT to approximate the combined p-value [47]. The uniform distribution has mean and variance of \( 1/2 \) and \( 1/12 \), respectively. According to the CLT, the average of \( m \) independent and identically distributed (i.i.d.) variables (with large \( m \)) follows a normal distribution with mean \( \mu = 1/2 \) and variance \( \sigma^2 = 1/(12m) \). By default, we use this to approximate the combined p-value when \( m \geq 20 \). We note that the additive method of combining p-values in our framework may be substituted by any other method of combining p-values.

**B. Application of DANUBE in Pathway Analysis**

Here, we present the application of DANUBE in the context of pathway analysis (Fig. 3). Let us consider a method \( M \), which can be GSEA, GSA, SPIA, or PADOG, or any other method that outputs a p-value for each pathway in the pathway database. We treat this p-value as the discriminating statistic. In step (1), we calculate the p-values of the pathways using the method \( M \). A pathway \( i \) will have \( m \) p-values \( \{p_{i1}, p_{i2}, \ldots, p_{im}\} \) for the \( m \) studies. The \( m \) p-values for a pathway are i.i.d. However, these p-values are not necessarily uniformly distributed under the null hypothesis (see Fig. 1). Therefore, combining these p-values will lead to systematic bias in identifying significant pathways as shown in Section II-C and as will be further illustrated in Section IV. Instead of combining these p-values, we treat them as observed values of the discriminating statistic.

To calculate the probability of observing such values, we need to construct the empirical distribution under the null hypothesis as described in steps (2)–(5) above. In step (2), we take all of the control samples from the \( m \) studies to create a set of control samples as shown in (C) in Fig. 3. In step (3), we generate the \( k \) synthetic datasets by random sampling from the pool of control samples. For example, for a simulation, we choose two groups of samples from the pool and label them as controls and diseases. In our case study using the Alzheimer’s dataset, as described in Section II-C, we generated 10,000 simulations of 10 control and 10 disease samples, 10,000 simulations of 10 control and 20 disease samples, 10,000 of 20 control and 10 disease samples, and 10,000 of 37 control and 37 disease samples, for a total of 40,000 simulations.

After generating \( k \) simulations from the control samples, we proceed to calculate the p-values for each pathway and each simulation using the same method \( M \). For a pathway \( i \), we have a set of p-values \( sp_{i1}, sp_{i2}, \ldots, sp_{ik} \). Since all of these p-values are calculated from the real control samples (i.e., healthy people), they can be considered as p-values under the null hypothesis. These p-values will be used to construct the empirical distribution \( \xi_i \) in step (5). In summary, steps (2)–(5) produce an empirical distribution for each pathway, resulting in a total of \( n \) empirical distributions for \( n \) pathways. These distributions will be used to calculate the empirical p-values of the measurements done in step (1).

After steps (1)–(5), for a pathway \( i \), we have \( m \) p-values \( p_{i1}, p_{i2}, \ldots, p_{im} \) and an empirical distribution \( \xi_i \). Using the formula described in (2), we calculate the empirical p-values \( ep_{i1}, ep_{i2}, \ldots, ep_{im} \). According to Lemma 1, these empirical p-values are independent and uniformly distributed under the null hypothesis. In step (7), we combine these empirical p-values using the additive method to have a single p-value pDANUBE, for pathway \( i \).

**IV. RESULTS AND VALIDATION**

In this section, we illustrate the limitations of combining p-values using classical meta-analysis approaches, and show that DANUBE overcomes these limitations. Section IV-A and B compare the classical approaches to DANUBE for the specific application domain of pathway analysis. Section IV-C and D compare the classical meta-analysis approaches to DANUBE in the general case, applicable to any meta-analysis.

For the pathway analysis applications on which we focus in this paper, we compare DANUBE to five other classical meta-analysis methods: Stouffer’s, Z-method, Brown’s, Fisher’s, and the additive method [14], [24], [48], [49], each of them combined with each of the four pathway analysis methods (GSEA, GSA, SPIA, and PADOG). We also compare these methods to a standalone meta-analysis method, MetaPath. In total, we analyze the results of 25 approaches: six meta-analyses combined with four pathway analysis methods, plus MetaPath [11], [50]. Each of these methods is tested on two diseases—one is Alzheimer’s disease with seven, and the other is acute myeloid leukemia (AML) with nine
Fig. 3. DANUBE’s application in pathway analysis. The input is m studies (datasets), and a pathway database, such as KEGG. Each dataset has a certain number of control and disease samples. Step (1): Perform pathway analysis using a method M (e.g., GSA, SPIA, or PADOG). For each pathway, the resulting m p-values are i.i.d. However, these p-values are not uniformly distributed under the null hypothesis (see Fig. 1), and therefore combining them would result in systematic bias. Step (2): Pool the control samples from the m datasets to produce a large set of control samples. Step (3): Generate k simulated datasets by randomly sampling from the pool. Since the “disease” and “control” samples in each of the simulated datasets were chosen only from the control samples of the original m studies, the resulting p-values are calculated under the null hypothesis. Step (4): Perform pathway analysis on the simulated data. Step (5): Collect the empirical distributions for each pathway. Step (6): Calculate the empirical p-values for each pathway using the additive method. Step (7): Combine the m empirical p-values obtained for each pathway using either the additive method or the Central Limit Theorem.
datasets. These conditions were selected for two reasons. First, there is a pathway in KEGG for each of the diseases. We refer to this as the target pathway and use it to validate the methods. Second, there are multiple experiments available in the public domain for both of these diseases.

A. Pathway Analysis Applications: Alzheimer’s Disease

The Alzheimer’s datasets we use in our data analysis are GSE28146 (hippocampus) and GSE5281 [six different tissues: entorhinal cortex (EC), hippocampus (HIP), medial temporal gyrus (MTG), posterior cingulate (PC), superior frontal gyrus (SFG), and primary visual cortex (VCX)]. The four pathway analysis methods, GSEA, GSA, SPIA, and PADOG, were used to process the expression data in each study and output a p-value for each study and for each pathway. Details of all datasets are provided in Supplementary Section 3.

The rankings and FDR-corrected p-values of the target pathway Alzheimer’s disease for the seven Alzheimer’s datasets are displayed in Fig. 4. The graphs demonstrate that the adjusted p-values and rankings of the target pathway vary substantially between the four methods for a given study, and from one study to the next. Furthermore, both GSA and PADOG report the target pathway Alzheimer’s disease as not significant in all seven studies.

We combine the four pathway analysis methods with six meta-analyses: Stouffer’s, Z-method, Brown’s, Fisher’s, the additive method, and DANUBE. Using a pathway analysis method M, each pathway has seven p-values—one per study. These seven p-values are combined using each of the six meta analysis methods Therefore, each pathway analysis method produces six lists of pathways. Each list has 150 pathways ranked according to the combined p-values. We then adjusted the combined p-values for multiple comparisons in each list using FDR.

In order to run DANUBE, we generated the null distributions from control samples as described in Section III-B. We took the 74 control samples from the seven Alzheimer’s datasets and randomly divided them into “control” and “disease” subgroups. We generated 10,000 simulations of 10 controls and 10 diseases, 10,000 simulations of 10 controls and 20 diseases, 10,000 of 20 controls and 10 diseases, and 10,000 of 37 controls and 37 diseases, for a total of 40,000 simulations. For each pathway analysis method, we constructed 150 empirical distributions for 150 KEGG signaling pathways (totally 600 empirical distributions for the four methods GSEA, GSA, SPIA, and PADOG). We used these empirical distributions to calculate the empirical p-values before applying the additive method to combine the empirical p-values for each pathway, resulting in 150 combined p-values. We then adjusted the combined p-values for multiple comparisons using FDR. Running time is reported in Supplementary Section 5 and Tables S1 and S2.

Table 1 displays the results using GSA combined with the six meta-analysis methods. The horizontal line across each list marks the 1% significance threshold. The pathway highlighted green is the target pathway Alzheimer’s disease. Pathways highlighted in red are examples of false positives. These pathways were expected to be reported as false positives because their null distribution is very skewed towards zero (see Fig. 1 panels A1–A3 and Supplementary Figure S3). These include Adherens junction and several cancer-related pathways, none of which are known to be implicated in Alzheimer’s disease. Stouffer’s method, the additive method, and DANUBE identify the target pathway as significant. DANUBE yields the best ranking.
Both Stouffer’s and the additive method identify the target pathway as significant using GSA, as shown in Table 1. However, the inherent bias of the null distribution brings irrelevant results into the list of significant pathways. For Stouffer’s method, pathways having p-values biased toward zero, such as Prostate cancer, Adherens junction, Pathways in cancer, and Pancreatic cancer, are still among the significant pathways. For the additive method, pathways having p-values biased toward zero, such as Prostate cancer, Adherens junction, and Pathways in cancer, are still among the significant pathways.

Table 2 displays the results using PADOG combined with the six meta-analysis methods. Only DANUBE identifies the target pathway as significant. Z-method and Brown’s method return no significant pathways. For Stouffer’s, Fisher’s, and the additive method, the systematic bias of the pathway analysis method greatly influences the outcome of the meta-analyses. Pathways having p-values biased toward zero, such as Adherens junction and cancer-related pathways (see Fig. 1 panels C1–C3 and Supplementary Figure S5), are among the significant pathways.

Supplementary Table S3 displays the results using SPIA combined with the six meta-analysis methods. The target pathway Alzheimer’s disease is highlighted in green. Pathways highlighted in red are examples of false positives. These pathways were expected to be reported as false positives because their null distributions are very skewed toward zero (see Figure 1 panels A1-A3 and Supplementary Figure S5). These include Adherens junction and several cancer-related pathways, which are not considered to be implicated in Alzheimer’s disease.

Both Stouffer’s and the additive method identify the target pathway as significant using GSA, as shown in Table 1. However, the inherent bias of the null distribution brings irrelevant results into the list of significant pathways. For Stouffer’s method, pathways having p-values biased toward zero, such as Prostate cancer, Adherens junction, Pathways in cancer, and Pancreatic cancer, are still among the significant pathways. For the additive method, pathways having p-values biased toward zero, such as Prostate cancer, Adherens junction, and Pathways in cancer, are still among the significant pathways.

Table 2 displays the results using PADOG combined with the six meta-analysis methods. Only DANUBE identifies the target pathway as significant. Z-method and Brown’s method return no significant pathways. For Stouffer’s, Fisher’s, and the additive method, the systematic bias of the pathway analysis method greatly influences the outcome of the meta-analyses. Pathways having p-values biased toward zero, such as Adherens junction and cancer-related pathways (see Fig. 1 panels C1–C3 and Supplementary Figure S5), are among the significant pathways.

Supplementary Table S3 displays the results using SPIA combined with the six meta-analysis methods. The target pathway Alzheimer’s disease is highlighted in green. Pathways highlighted in red are examples of false positives. These pathways were expected to be reported as false positives because their null distributions are very skewed toward zero (see Figure 1 panels A1-A3 and Supplementary Figure S5). These include Adherens junction and several cancer-related pathways, which are not considered to be implicated in Alzheimer’s disease.
Supplementary Table S4 displays the results using GSEA combined with the six meta-analysis methods. The horizontal line across each list marks the cutoff $FDR = 0.01$. The pathway highlighted green is the target pathway Alzheimer’s disease. The target pathway is significant for all the six meta-analysis methods. Because GSEA is unbiased, the additive method and DANUBE have equivalent results. These two methods have a shorter list of significant pathways and rank the target pathway higher than other methods. In addition, all the four significant pathways, Cardiac muscle contraction, Huntington’s disease, Alzheimer’s disease, and Parkinson’s disease, appear in the lists of significant pathways when we combine DANUBE with GSA, PADOG, and SPIA.

There is no gold standard for assigning true or false values to each of the results, apart from the expectation that a disease under study should impact its namesake pathway. Indeed, the target pathway Alzheimer’s disease is ranked as significant for all of the four pathway analysis methods when combined with DANUBE. The target pathway is also ranked higher when using DANUBE compared to the results of other five meta-analysis methods. In addition, the pathways Parkinson’s disease, Alzheimer’s disease, Cardiac muscle contraction, and Huntington’s disease consistently appear as significant in the results of all the four pathway analysis methods when combined with DANUBE.

Alzheimer’s, Parkinson’s, and Huntington’s diseases are three neurological disorders that have many commonalities including abnormal protein folding, endoplasmic reticulum stress, and ubiquitin mediated breakdown of proteins, leading to programmed cell death. Given that
the pathway Alzheimer's disease is influenced by the mitochondrial compartment, which is strongly implicated in the disease [51]–[54], it is not surprising that other pathways with strong mitochondrial components also garner high rankings. Previous studies [55] have shown the presence of a cross-talk that makes the neurological disease pathways, Alzheimer's disease, Parkinson's disease, and Huntington's disease, along with Cardiac muscle contraction, appear as significant simultaneously, due to their dominant mitochondrial module. Cardiac muscle contraction has a strong mitochondrial component and is highly dependent on calcium signaling, which is also prevalent in Synaptic vesicle cycle, Alzheimer's disease, and Huntington's disease. Ca2+ regulates mitochondrial metabolism, but calcium overload to mitochondria can result in cell damage from reactive oxygen [56].

We also use MetaPath to combine the seven studies. MetaPath is a standalone meta-analysis method, which does not need an external pathway analysis tool. This method performs meta-analysis at both gene (MAPE_G) and pathway levels (MAPE_P), and then combines the results (MAPE_I) to give the final p-value and ranking of pathways. Supplementary Table S5 shows the top seven pathways using MetaPath for the seven Alzheimer's datasets. The target pathway Alzheimer's disease is not significant and is outranked by six other pathways.

B. Pathway Analysis Applications: AML

The AML datasets we use in our data analysis are GSE14924 (CD4 and CD8 T cells), GSE17054 (stem cells), GSE12662 (CD34+ cells, promyelocytes, and neutrophils and PR9 cell line), GSE57194 (CD34+ cells), GSE33223 (peripheral blood, bone marrow), GSE42140 (peripheral blood, bone marrow), GSE8023 (CD34+ cells), and GSE15061 (bone marrow). The rankings and FDR-corrected p-values of the target pathway Acute myeloid leukemia are highlighted in green.

Table 3 Twenty-One Top-Ranked Pathways and FDR-Corrected p-Values Obtained by Combining the GSA p-Values Using Six Meta-Analysis Methods for AML. The Target Pathway Acute myeloid leukemia is Significant for Stouffer’s, the Additive Method, and DANUBE With Rankings 13th, 2nd, and 1st, Respectively

<table>
<thead>
<tr>
<th>Pathway</th>
<th>GSA + Stouffer's method</th>
<th>GSA + Z-method</th>
<th>GSA + Brown's method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Ebb signaling pathway</td>
<td>&lt; 10^{-4}</td>
<td>&lt; 10^{-4}</td>
<td>Ebb signaling pathway</td>
</tr>
<tr>
<td>2 Sulfur relay system</td>
<td>&lt; 10^{-4}</td>
<td>&lt; 10^{-4}</td>
<td>Sulfur relay system</td>
</tr>
<tr>
<td>3 Adrenergic junction</td>
<td>&lt; 10^{-4}</td>
<td>&lt; 10^{-4}</td>
<td>Adrenergic junction</td>
</tr>
<tr>
<td>4 Tight junction</td>
<td>&lt; 10^{-4}</td>
<td>&lt; 10^{-4}</td>
<td>Tight junction</td>
</tr>
<tr>
<td>5 Circadian rhythm</td>
<td>&lt; 10^{-4}</td>
<td>&lt; 10^{-4}</td>
<td>Circadian rhythm</td>
</tr>
<tr>
<td>6 Alcoholism</td>
<td>&lt; 10^{-4}</td>
<td>&lt; 10^{-4}</td>
<td>Alcoholism</td>
</tr>
<tr>
<td>7 Glutathion</td>
<td>&lt; 10^{-4}</td>
<td>&lt; 10^{-4}</td>
<td>Glutathion</td>
</tr>
<tr>
<td>8 Transcriptional regulation in cancer</td>
<td>&lt; 10^{-4}</td>
<td>&lt; 10^{-4}</td>
<td>Transcriptional regulation in cancer</td>
</tr>
<tr>
<td>9 Renal cell carcinoma</td>
<td>&lt; 10^{-4}</td>
<td>&lt; 10^{-4}</td>
<td>Renal cell carcinoma</td>
</tr>
<tr>
<td>10 Glioma</td>
<td>&lt; 10^{-4}</td>
<td>&lt; 10^{-4}</td>
<td>Glioma</td>
</tr>
<tr>
<td>11 Systemic lupus erythematosus</td>
<td>&lt; 10^{-4}</td>
<td>&lt; 10^{-4}</td>
<td>Systemic lupus erythematosus</td>
</tr>
<tr>
<td>12 Non-small cell lung cancer</td>
<td>&lt; 10^{-4}</td>
<td>&lt; 10^{-4}</td>
<td>Non-small cell lung cancer</td>
</tr>
<tr>
<td>13 Acute myeloid leukemia</td>
<td>0.0003</td>
<td>0.0003</td>
<td>Acute myeloid leukemia</td>
</tr>
<tr>
<td>14 VEGF signaling pathway</td>
<td>0.0017</td>
<td>0.1139</td>
<td>VEGF signaling pathway</td>
</tr>
<tr>
<td>15 Endothelial cancer</td>
<td>0.0025</td>
<td>0.1288</td>
<td>Endothelial cancer</td>
</tr>
<tr>
<td>16 mTOR signaling pathway</td>
<td>0.0033</td>
<td>0.1136</td>
<td>mTOR signaling pathway</td>
</tr>
<tr>
<td>17 Chronic myeloid leukemia</td>
<td>0.0083</td>
<td>0.1333</td>
<td>Chronic myeloid leukemia</td>
</tr>
<tr>
<td>18 Proteasome</td>
<td>0.0081</td>
<td>0.1333</td>
<td>Proteasome</td>
</tr>
<tr>
<td>19 Pancreatic cancer</td>
<td>0.0097</td>
<td>0.2037</td>
<td>Pancreatic cancer</td>
</tr>
<tr>
<td>20 HIF-1 signaling pathway</td>
<td>0.3170</td>
<td>0.2394</td>
<td>HIF-1 signaling pathway</td>
</tr>
</tbody>
</table>

The horizontal lines show the 1% significance threshold. The target pathway Acute myeloid leukemia is highlighted in green.
leukemia for the nine AML datasets are displayed in Supplementary Figure S12. The graphs demonstrate that the adjusted p-values and rankings of the target pathway vary substantially between the four methods for a given study, and from one study to the next. Furthermore, the AML pathway was not found to be significant by any method in any dataset.

We combine the four pathway analysis methods with the six meta-analysis methods. Using a pathway analysis method $M$, each pathway has nine p-values—one per study. These nine p-values are combined using each of the six meta-analysis methods. Therefore, each pathway analysis method produces six lists of pathways. Each list has 150 pathways ranked according to the combined p-values. We then adjust the combined p-values for multiple comparisons in each list using FDR.

In order to run DANUBE, we generated the null distributions from control samples as described in Section III-B. We took the 140 control samples of the nine AML datasets and randomly designated "control" and "disease" subgroups. We generated 10,000 simulations of 10 controls and 10 diseases, 10,000 of 30 controls and 50 diseases, 10,000 of 50 controls and 30 diseases, and 10,000 of 70 controls and 70 diseases, for a total of 40,000 simulations. For each pathway analysis method, we constructed 150 empirical distributions for 150 KEGG signaling pathways (totally 600 empirical distributions for the four pathway analysis methods). We then used the empirical distributions to calculate the empirical p-values before applying the additive method to combine the empirical p-values for each pathway, resulting in 150 combined p-values. Finally,

### Table 4: Twenty-Three Top-Ranked Pathways and FDR-Corrected p-Values Obtained by Combining the PADOG p-Values Using Six Meta-Analysis Methods for AML. The Target Pathway Acute myeloid leukemia Is Significant for Stouffer's, Fisher's, the Additive Method, and DANUBE. DANUBE Yields the Best Ranking

<table>
<thead>
<tr>
<th>Pathway + Stouffer's method</th>
<th>Pathway + Z-method</th>
<th>Pathway + Brown's method</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pathway</strong></td>
<td>p-value</td>
<td><strong>Pathway</strong></td>
</tr>
<tr>
<td>1 Non-small cell lung cancer</td>
<td>$&lt; 10^{-4}$</td>
<td>Non-small cell lung cancer</td>
</tr>
<tr>
<td>2 Chronic myeloid leukemia</td>
<td>$&lt; 10^{-4}$</td>
<td>Chronic myeloid leukemia</td>
</tr>
<tr>
<td>3 Glioma</td>
<td>$&lt; 10^{-4}$</td>
<td>Glioma</td>
</tr>
<tr>
<td>4 ErbB signaling pathway</td>
<td>$&lt; 10^{-4}$</td>
<td>ErbB signaling pathway</td>
</tr>
<tr>
<td>5 Colorectal cancer</td>
<td>$&lt; 10^{-4}$</td>
<td>Colorectal cancer</td>
</tr>
<tr>
<td>6 Prostate cancer</td>
<td>$&lt; 10^{-4}$</td>
<td>Prostate cancer</td>
</tr>
<tr>
<td>7 Acute myeloid leukemia</td>
<td>$&lt; 10^{-4}$</td>
<td>Acute myeloid leukemia</td>
</tr>
<tr>
<td>8 VEGF signaling pathway</td>
<td>0.0001</td>
<td>VEGF signaling pathway</td>
</tr>
<tr>
<td>9 Endometrial cancer</td>
<td>0.0001</td>
<td>Endometrial cancer</td>
</tr>
<tr>
<td>10 Pancreatic cancer</td>
<td>0.0001</td>
<td>Pancreatic cancer</td>
</tr>
<tr>
<td>11 Pathways in cancer</td>
<td>0.0001</td>
<td>Pathways in cancer</td>
</tr>
<tr>
<td>12 Transcriptional misregulation in cancer</td>
<td>0.0005</td>
<td>Transcriptional misregulation in cancer</td>
</tr>
<tr>
<td>13 T cell receptor signaling pathway</td>
<td>0.0012</td>
<td>T cell receptor signaling pathway</td>
</tr>
<tr>
<td>14 mTOR signaling pathway</td>
<td>0.0012</td>
<td>mTOR signaling pathway</td>
</tr>
<tr>
<td>15 Circadian rhythm</td>
<td>0.0015</td>
<td>Circadian rhythm</td>
</tr>
<tr>
<td>16 Neurotrophin signaling pathway</td>
<td>0.0021</td>
<td>Neurotrophin signaling pathway</td>
</tr>
<tr>
<td>17 Small cell lung cancer</td>
<td>0.0024</td>
<td>Small cell lung cancer</td>
</tr>
<tr>
<td>18 Retinal cell carcinoma</td>
<td>0.0054</td>
<td>Retinal cell carcinoma</td>
</tr>
<tr>
<td>19 Insulin signaling pathway</td>
<td>0.0063</td>
<td>Insulin signaling pathway</td>
</tr>
<tr>
<td>20 Endostrosis</td>
<td>0.0070</td>
<td>Endostrosis</td>
</tr>
<tr>
<td>21 Adherin junction</td>
<td>0.0079</td>
<td>Adherin junction</td>
</tr>
<tr>
<td>22 Wnt signaling pathway</td>
<td>0.0168</td>
<td>Wnt signaling pathway</td>
</tr>
<tr>
<td>23 Melanoma</td>
<td>0.0185</td>
<td>Melanoma</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pathway + Fisher's method</th>
<th>Pathway + Additive method</th>
<th>Pathway + DANUBE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pathway</strong></td>
<td>p-value</td>
<td><strong>Pathway</strong></td>
</tr>
<tr>
<td>1 Chronic myeloid leukemia</td>
<td>$&lt; 10^{-4}$</td>
<td>Acute myeloid leukemia</td>
</tr>
<tr>
<td>2 Non-small cell lung cancer</td>
<td>$&lt; 10^{-4}$</td>
<td>Chronic myeloid leukemia</td>
</tr>
<tr>
<td>3 Glioma</td>
<td>$&lt; 10^{-4}$</td>
<td>Glioma</td>
</tr>
<tr>
<td>4 ErbB signaling pathway</td>
<td>$&lt; 10^{-4}$</td>
<td>ErbB signaling pathway</td>
</tr>
<tr>
<td>5 Colorectal cancer</td>
<td>$&lt; 10^{-4}$</td>
<td>Colorectal cancer</td>
</tr>
<tr>
<td>6 Prostate cancer</td>
<td>$&lt; 10^{-4}$</td>
<td>Prostate cancer</td>
</tr>
<tr>
<td>7 Acute myeloid leukemia</td>
<td>$&lt; 10^{-4}$</td>
<td>Acute myeloid leukemia</td>
</tr>
<tr>
<td>8 VEGF signaling pathway</td>
<td>0.0001</td>
<td>VEGF signaling pathway</td>
</tr>
<tr>
<td>9 Endometrial cancer</td>
<td>0.0001</td>
<td>Endometrial cancer</td>
</tr>
<tr>
<td>10 Pancreatic cancer</td>
<td>0.0001</td>
<td>Pancreatic cancer</td>
</tr>
<tr>
<td>11 Pathways in cancer</td>
<td>0.0001</td>
<td>Pathways in cancer</td>
</tr>
<tr>
<td>12 Transcriptional misregulation in cancer</td>
<td>0.0005</td>
<td>Transcriptional misregulation in cancer</td>
</tr>
<tr>
<td>13 T cell receptor signaling pathway</td>
<td>0.0012</td>
<td>T cell receptor signaling pathway</td>
</tr>
<tr>
<td>14 mTOR signaling pathway</td>
<td>0.0012</td>
<td>mTOR signaling pathway</td>
</tr>
<tr>
<td>15 Circadian rhythm</td>
<td>0.0015</td>
<td>Circadian rhythm</td>
</tr>
<tr>
<td>16 Neurotrophin signaling pathway</td>
<td>0.0021</td>
<td>Neurotrophin signaling pathway</td>
</tr>
<tr>
<td>17 Small cell lung cancer</td>
<td>0.0024</td>
<td>Small cell lung cancer</td>
</tr>
<tr>
<td>18 Retinal cell carcinoma</td>
<td>0.0054</td>
<td>Retinal cell carcinoma</td>
</tr>
<tr>
<td>19 Insulin signaling pathway</td>
<td>0.0063</td>
<td>Insulin signaling pathway</td>
</tr>
<tr>
<td>20 Endostrosis</td>
<td>0.0070</td>
<td>Endostrosis</td>
</tr>
<tr>
<td>21 Adherin junction</td>
<td>0.0079</td>
<td>Adherin junction</td>
</tr>
<tr>
<td>22 Wnt signaling pathway</td>
<td>0.0168</td>
<td>Wnt signaling pathway</td>
</tr>
<tr>
<td>23 Melanoma</td>
<td>0.0185</td>
<td>Melanoma</td>
</tr>
</tbody>
</table>

The horizontal lines show the 1% significance threshold. The target pathway Acute myeloid leukemia is highlighted in green.
we adjusted the combined p-values for multiple comparisons using FDR.

Table 3 displays the results of GSA combined with the six meta-analysis methods, ordered by the FDR corrected p-values. We place a horizontal line across each list to mark our 1% cutoff. Stouffer’s method, the additive method, and DANUBE identify the target pathway as significant. DANUBE yields the best ranking (ranked 1st), followed by the additive (2nd) and Stouffer’s method (13th). In addition, the target pathway is the only significant pathway in DANUBE’s result.

Table 4 shows the results of PADOG combined with the six meta-analysis methods. The target pathway is significant for the four methods: DANUBE, Stouffer’s, Fisher’s, and the additive method. For DANUBE, Acute myeloid leukemia is ranked 1st compared to 7th using the other three meta-analysis methods. There are no significant pathways using the Z-method and Brown’s method.

Supplementary Table S6 shows the results of SPIA combined with the six meta-analysis methods, ordered by the FDR corrected p-value. Again, the target pathway is significant using Stouffer’s, Fisher’s, and the additive method. For DANUBE, Acute myeloid leukemia is ranked 1st compared to 7th using the other three meta-analysis methods. There are no significant pathways using the Z-method and Brown’s method.

Supplementary Table S7 displays the results of GSEA combined with the six meta-analysis methods. The target pathway Acute myeloid leukemia (AML) as significant in the AML data.

We also use MetaPath to combine the nine acute myeloid leukemia studies. Supplementary Table S8 shows the top five pathways using MetaPath. The target pathway is not significant (p = 0.4) and is outranked by two other pathways.

Table 5 summarizes all the results for the 25 approaches (four pathway analysis methods each combined with one of six meta-analysis approaches, plus MetaPath). On average, DANUBE performs best in terms of ranking, as well as in terms of identifying the target pathway as significant at the 1% cutoff.

We note that for both diseases, DANUBE and the additive methods have the same results when combined with GSEA because GSEA is an unbiased method with uniform distributions of p-values under the null. In addition, the results of the two methods for SPIA are almost equivalent because the distributions of the p-values produced by SPIA under the null are closer to the expected uniform. Notably, DANUBE is more useful in conjunction with methods that have more skewed empirical null distributions.

C. General Case: t-Test and Wilcoxon Test

In this section, we will demonstrate the generality of the problem, beyond pathway analysis applications. In order to do so, we have used the one sample t-test [57], [58] and the one sample Wilcoxon signed-rank test [59]–[61] as illustrative examples of parametric and nonparametric tests. Using simulated null distributions, we show that both the t-test and Wilcoxon tests have systematic bias depending on the shape and the symmetry of the null distribution. When the p-values are biased towards zero, combining multiple studies results in an
increase of type-I error (prevalence of false positives). When the p-values are biased towards one, the test loses power and more evidence is needed to identify true positives.

In Fig. 5, panel (a) displays a simulated null distribution $H_0$ that is not symmetrical and does not follow any standard distribution. Panel (b) displays an alternative distribution $H_1$, which has the same shape as $H_0$, but a slightly smaller median. Panel (c) displays another alternative distribution $H_2$, which has the same shape as $H_0$ but a slightly larger median. Each population has 100,000 elements. The goal here is to investigate the ability of each approach to distinguish between $H_0$ and $H_1$, and between $H_0$ and $H_2$, respectively. This is attempted using both a t-test and a Wilcoxon test.

Denoting $M_0$ and $m_0$ as the mean and median of the null distribution $H_0$, $M_0$ is used as the parameter (mean) for the t-tests where $m_0$ is used as the parameter (median) for Wilcoxon test. To make the analysis more general, the sample size is randomized between 3 and 10 every time we pick a sample. Since DANUBE uses the additive method to combine the p-values, we also use the additive method to combine the p-values of t-test and Wilcoxon test. When the number of studies is larger or equals to 20, the combined p-values are calculated using the Central Limit Theorem as described in Section III.

Panels (d)–(h) show the results using the one sample left-tailed t-test for the mean; panels (i)–(m) show the results using the one sample right-tailed t-test for the mean; panels (n)–(r) show the results using the one sample left-tailed Wilcoxon test for the median; panels (s)–(w) show the results using one sample right-tailed Wilcoxon test for the median.

Panel (d) shows the distribution of p-values for samples drawn from the null distribution $H_0$. To plot this panel, we randomly select 100,000 samples from $H_0$ and then calculate the p-values using the left-tailed t-test. Since the null distribution $H_0$ is not normal, the resulting p-values are not uniformly distributed. Panel (e) displays the distribution of combined p-values for samples drawn from the null distribution $H_0$. To calculate a combined p-value, we randomly pick 10 samples from the null population $H_0$ and then calculate the 10 p-values using the left-tailed t-test. From these 10 p-values, we calculate a combined p-value using the additive method. This procedure is repeated 100,000 times to generate the distribution of the combined p-values under the null hypothesis. Similarly, panel (f) displays the distribution of the combined p-values for samples drawn from the alternative distribution $H_1$.

The red dashed lines in panels (e, f) show the 0.05 cutoff. Since the combined p-values in (e) are calculated under the null hypothesis, values smaller than the cutoff are false positives. Therefore, the blue area to the left of the red dashed line is type-I error of the classical meta-analysis using the left-tailed t-test. Similarly, combined p-values larger than the cutoff in panel (f) are false negatives. The blue area to the right of the red line panel (f) displays type-II error.

The results show that combined p-values will be biased towards zero since p-values of the left-tailed t-test are biased towards zero. To understand the behavior of the meta-analysis, we display type-I and type-II error in panels (g) and (h) with varying numbers of studies to be combined. As the number of studies increases, the meta-analysis becomes more biased, and type-I error increases. For example, when the number of studies reaches 50, the analysis has more than 60% false positives. Paradoxically, increasing the number of studies will make the meta-analysis less useful due to the increase of type-I error.

Panels (i)–(m) display the results of the right-tailed t-test. Panel (i) displays the distribution of p-values for samples drawn from the null distribution $H_0$. Panel (j) displays the combined p-values for samples drawn from the null distribution $H_0$. Panel (k) displays the combined p-values for samples drawn from the alternative distribution $H_2$. Each combined p-value is calculated from 10 individual p-values. The right-tailed t-test is biased towards one, therefore more evidence is required to identify true positives. Compared to the left-tailed t-test, the right-tailed t-test has smaller type-I error but larger type-II error (less power). Therefore, many more studies would be required for this test to identify true positives. Panel (m) shows that for the case of combining 10 studies, the type-II error of the right-tailed t-test is about 0.5, whereas the type-II error of the left-tailed t-test is less than 0.2.

Panels (n)–(r) display the results of meta-analysis using the one sample left-tailed Wilcoxon test for the median. In this example, the left-tailed Wilcoxon test is biased towards one, so more evidence is required to identify true positives. As shown in panel (r), the expected type-II error of the meta-analysis is about 0.6 when combining 10 studies. Interestingly, the behavior of the meta-analysis using the left-tailed Wilcoxon test is similar to that of the right-tailed t-test. In both cases, the meta-analysis needs a large number of studies to identify true positives. Panels (m) and (r) show that type-II error converges to zero as the number of studies increases.

Panels (s)–(w) display the results of meta-analysis using the one sample right-tailed Wilcoxon test for the median. Similar to the t-test, the right-tailed Wilcoxon test is biased towards zero. As shown in panels (g) and (v), type-I error using either of the two tests increases as the number of studies increases.

D. General Case: DANUBE

In this section, we analyze the performance of DANUBE using the same null and alternative distributions that were used for the t-test and Wilcoxon tests. Fig. 6 displays the results using DANUBE. Panels (a)–(c) show the null distribution $H_0$ and two alternative distributions $H_1$.
Fig. 5. Type-I and Type-II errors of the classical meta-analysis using one sample t-test and Wilcoxon signed-rank test. (a) Probability distribution under the null hypothesis $H_0$. (b) Alternative distribution $H_1$ that has the same shape as the null distribution with a slightly smaller median. (c) Another alternative distribution $H_2$ that has the same shape as the null distribution with a slightly larger median. (d)–(h) Results using left-tailed t-tests. (d) Distribution of $p$-values using left-tailed t-test for samples drawn from the null distribution $H_0$. (e) Distribution of combined $p$-values using left-tailed t-test for samples drawn from the null distribution $H_0$. The red dashed line represents the threshold (0.05) below which the null hypothesis will be rejected. The blue area to the left of the red dashed line is type-I error (false positives). (f) Distribution of combined $p$-values using left-tailed t-test for samples drawn from the alternative distribution $H_1$. The blue area to the right of the red dashed line is type-II error (false negatives). (g) Type-I error with varying number of studies. (h) Type-II error with varying number of studies using a left-tailed t-test for samples drawn from the alternative distribution $H_1$. Similarly, (i)–(m) display the results using right-tailed t-test; (n)–(r) display the results of left-tailed Wilcoxon signed-rank test; (s)–(w) display the results of right-tailed Wilcoxon signed-rank test. In this example, the left-tailed t-test and right-tailed Wilcoxon tests are biased towards 0 as shown in (e) and (f). Therefore, an increase in the number of studies makes the combined $p$-values more biased towards 0, causing an increase in type-I error as shown in (g) and (v). On the contrary, the right-tailed t-test and left-tailed Wilcoxon test are biased towards 1. This kind of bias makes the test less powerful. For example, with 10 studies, type-II errors using right-tailed t-test and left-tailed Wilcoxon test are 0.51 and 0.61, respectively.
Fig. 6. **Type-I and type-II errors of DANUBE using mean and median as discriminative statistics.**

(a) Probability distribution under the null hypothesis ($H_0$), which has the same shape as the null distribution but a slightly smaller median. (c) Alternative distribution ($H_2$) that has the same shape as the null distribution but a slightly larger median. (d)–(h) Results of the left-tailed DANUBE using mean; (i)–(m) Results of the right-tailed DANUBE using mean; (n)–(r) Results of left-tailed DANUBE using median; (s)–(w) Results of right-tailed DANUBE using median. (d), (i), (n), and (s) show the p-value distributions for samples drawn from the null. For all four tests, p-values are uniformly distributed under the null hypothesis. Consequently, the combined p-values (using the additive method) are also uniformly distributed under the null hypothesis as shown in (e), (j), (o), and (t). The result is that the type-I error equals the threshold (0.05) regardless of the number of studies combined, as shown in (g), (l), (q), and (v). (h), (m), (r), and (w) show that the type-II error converges quickly to zero. Combining 10 studies, the type-II error of left- and right-tailed DANUBE for the mean are both less than 0.3 compared to 0.51 for the right-tailed t-test. Similarly, using the median, the type-II error of DANUBE is less than 0.2 compared to 0.61 for the left-tailed Wilcoxon test.
and $H_2$. Panels (d)–(h) display the results using left-tailed DANUBE for the mean; panels (i)–(m) display the results using right-tailed DANUBE for the mean; panels (n)–(r) display the results using left-tailed DANUBE for the median; panels (s)–(w) display the results using right-tailed DANUBE for the median.

We randomly select 10,000 samples from the null distribution and use them to construct the empirical distribution of sample means [panels (d)–(m)] and likewise of sample medians [panels (n)–(w)]. For a given empirical distribution, we calculate the probability of observing the discriminating statistic in a study. Panel (d) displays the distribution of empirical $p$-values for samples drawn from the null distribution $H_0$; we see that these are uniformly distributed under the null hypothesis. Panel (e) displays the distribution of combined $p$-values for samples drawn from the null distribution $H_0$. Each combined $p$-value is calculated from 10 individual empirical $p$-values. The blue area to the left of the red dashed line is type-I error. Since the individual $p$-values are uniformly distributed, the combined $p$-values are also uniformly distributed. Consequently, the type-I error of this test is equal to the threshold. Panel (f) displays the distribution of combined $p$-values for samples drawn from the alternative distribution $H_1$. The blue area to the right of the red dashed line is the type-II error.

Panels (g) and (h) display the type-I and type-II errors of DANUBE with varying numbers of combined studies. The graphs show that the type-I error of DANUBE consistently equals the threshold, while type-II error decreases when the number of studies increases. When combining 10 studies, the type-I and type-II errors of the left-tailed DANUBE for the mean are 0.05 and 0.27, respectively, compared to 0.24 and 0.14 for the left-tailed $t$-test. When the number of the studies increases over 30, one can expect DANUBE to give a 0.05 type-I error and an almost zero type-II error.

Similar to the left-tailed test, right-tailed DANUBE on the mean has the expected type-I error and a reasonable type-II error as shown in panels (i) and (m). With 10 studies to be combined, the right-tailed DANUBE's type-I and type-II errors are 0.05 and 0.25, respectively, compared to 0.01 and 0.51 for the right-tailed $t$-test. The results for the mean show that both left- and right-tailed type-I errors are equal to the threshold while the type-II error decreases rapidly. On the contrary, the left- and right-tailed $t$-tests have unpredictable behavior due to the skewness of the null distribution.

Panels (n)–(w) show the results of left- and right-tailed DANUBE for the median. As expected, the type-I error for the median is also equal to the threshold, regardless of the number of studies that are combined. The test is proven to be powerful for both tails with type-II error less than 0.2 for 10 studies. When compared to the left-tailed Wilcoxon test on 10 studies, the DANUBE left-tailed type-II error is 0.17 as opposed to 0.61.

V. CONCLUSION

In this paper, we present a new framework to combine the results of multiple studies in order to gain more statistical power. Our framework first calculates the empirical $p$-values for each study using the empirical distribution of the discriminating statistic. It then combines the empirical $p$-value using either the Central Limit Theorem or the additive method. The new framework makes no statistical assumptions about the data and is therefore usable in many practical cases when no simple model is appropriate. In addition, use of the additive method makes the framework more robust to outliers.

The advantage of the new meta-analysis framework is demonstrated using both simulation and real-world data. In our simulation study, we compare the results of DANUBE to the classical additive method using the one sample $t$-test and Wilcoxon signed-rank test. The skewness and the nonnormality of the simulated null distribution produces systematic bias in classical meta-analysis, either increasing type-I error or decreasing the power of the test. In contrast, the type-I error of DANUBE is equal to the threshold cutoff and type-II error declines quickly when the number of studies increases.

To evaluate the proposed framework for pathway analysis applications, we examine seven Alzheimer’s and nine acute myeloid leukemia datasets using 25 approaches: six meta-analysis methods, Stouffer’s, Z-method, Brown’s, Fisher’s, the additive method, and DANUBE, each of them combined with four representative pathway analysis methods, GSA, SPIA, PADOG, and GSEA, plus an additional independent meta-analysis method MetaPath. The results confirm the advantage of DANUBE over classical meta-analysis to identify pathways relevant to the phenotype.

This work describes an important limitation of current meta-analysis techniques and provides a general statistical approach to increase the power of an analysis method using empirical distributions. With vast databases of biological data being made available, this framework may be powerful because it lets the data speak for itself. The proposed framework is flexible enough to be applicable to various types of studies, including gene-level analysis, pathway analysis, or clinical trials to assess the effect of a therapy in complex diseases.

Acknowledgment

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies. The authors thank D. Diaz for help and useful discussion.
REFERENCES


[16] P. Hall, “The distribution of means for samples of size n drawn from a population in which the variate takes values between 0 and 1, all such values being equally probable,” Biometrika, vol. 19, no. 3–4, pp. 240–244, 1927.


ABOUT THE AUTHORS

**Tin Nguyen** received the B.Sc. and M.Sc. degrees in computer science from Eotvos Lorand University, Budapest, Hungary, in 2004 and 2008, respectively, and is currently pursuing the Ph.D. degree in computer science at Wayne State University, Detroit, MI, USA.

He is a member of the Intelligent Systems and Bioinformatics Laboratory (ISBL), Department of Computer Science, Wayne State University. His research interests include computational and statistical methods for analyzing high-throughput data. His current focus is meta-analysis, multi-omics data integration, and disease subtyping.

**Cristina Mitrea** (Student Member, IEEE) received the Master of Science degree in computer science from Wayne State University, Detroit, MI, USA, in 2012, and is currently pursuing the Ph.D. degree in computer science at Wayne State University.

She is a member of the Intelligent Systems and Bioinformatics Laboratory (ISBL), Department of Computer Science, Wayne State University. Her work is focused on research in data mining techniques applied to bioinformatics and computational biology. The main focus of her research is developing bioinformatics tools for cancer studies. Other interests include network discovery and meta-analysis applied to pathway analysis.

Ms. Mitrea is also a student member of the ACM.

**Rebecca Tagett** received the bachelor’s degree in physics from Emory University, Atlanta, GA, USA, in 1984, the master’s degree in atmospheric sciences from the University of California, Los Angeles (UCLA), CA, USA, in 1987, and the master’s degree in molecular biology from the City University of New York (CUNY), New York, NY, USA, in 1992, and is currently pursuing the Ph.D. degree in computer science at Wayne State University, Detroit, MI, USA.

She has 10 years of R&D experience in industry as a Computational Biologist. She is currently a member of the Intelligent Systems and Bioinformatics Laboratory (ISBL), Department of Computer Science, Wayne State University. Her research focuses on phenotypic prediction using multi-omics. Her interests are functional genomics, scientific writing, bioinformatics and biostatistics.

Ms. Tagett is a member of the International Society for Computational Biology (ISCB).

**Sorin Draghici** (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of St Andrews, St Andrews, U.K., in 1995.

He is the Associate Dean for Innovation and Entrepreneurship and Director of the James and Patricia Anderson Engineering Ventures Institute, College of Engineering, Wayne State University, Detroit, MI, USA. He currently holds the Robert J. Sokol, MD Endowed Chair in Systems Biology, as well as appointments as Full Professor with the Department of Computer Science and the Department of Obstetrics and Gynecology, Wayne State University. He is also the head of the Intelligent Systems and Bioinformatics Laboratory (ISBL) in the Department of Computer Science. His work is focused on research in artificial intelligence, machine learning, and data mining techniques applied to bioinformatics and computational biology. He has published two best-selling books on data analysis of high-throughput genomics data, eight book chapters, and over 160 peer-reviewed journal and conference papers. His research laboratory has a strong track record in developing tools for data analysis of high-throughput data. His laboratory has developed eight analysis tools in this area, tools that have been made available over the Web for more than 10 years to over 17,000 scientists from five continents. He has also coauthored three analysis packages in Bioconductor.


**[57] W. S. Gosset, “The probable error of a mean,” Biometrika, vol. 6, pp. 1–25, 1908.**


