DANUBE: Data-driven meta-ANalysis using UnBiased Empirical distributions - applied to biological pathway analysis

Tin Nguyen, Cristina Mitrea, 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 with 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.**

***Index Terms*—meta-analysis, p-values, empirical distribution, pathway analysis, Alzheimer's disease, acute myeloid leukemia.**

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 Genome Atlas (cancergenome.nih.gov), ArrayExpress [3, 4], and Therapeutically Applicable Research to Generate Effective Treatments (ocg.cancer.gov/programs/target) store thousands of datasets, within which there are independent experimental series with similar patient cohorts and experiment design. Gene expression data, as measured by microarrays, are particularly prevalent in public databases, such that some disease conditions are represented by half a dozen studies or more.

Experiments comparing two phenotypes, such as disease and control, yield lists of genes that are differentially expressed (DE). However, lists of DE genes obtained from similar but independent experiments tend to have little in common, and taken alone, they usually fail to elucidate the underlying biological mechanisms. Effective meta-analysis approaches are needed to unify the biological knowledge spread out over such similar studies with apparently incongruent results.

The goal of the meta-analysis is to combine the results of independent but related studies and provide increased statistical power and robustness compared to individual studies analyzed alone [5, 6]. In spite of the numerous sophisticated tools for meta-analysis, many biological applications still use only Venn diagrams (intersection/union) or vote counting for combining multiple studies [7, 8]. Such approaches are useful for demonstrating consistency when combining a few studies. However, when combining many studies, Venn diagrams are either too conservative (for intersection) or too anti-conservative (for union), while vote counting is statistically inefficient [5, 9, 10]. Regarding microarray data, meta-analysis has been used at both gene level [5, 7, 11–13] and pathway level [11, 14]. Pathway analysis [15–18] was developed to correlate differential gene expression evidence with a-priori defined functional modules, organized into biological pathway databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [19, 20], Reactome [21], Biocarta (www.biocarta.com), or Molecular Signatures Database (MSigDB) [22].

One straightforward and flexible way of integrating diverse studies is to combine the individual p-values provided by each study. Classical meta-analysis methods of combining p-values have been reviewed and compared in [23]. These include Fisher's method based on the chi-squared distribution [24], the additive method [25] using the Irwin-Hall distribution [26, 27], minP [28], and maxP [29].

In an early study, Rhodes and others [13] collected multiple prostate cancer microarray datasets and combined p-values using Fisher's method. Since then, other sophisticated approaches have been proposed including the weighted Fisher's method [30] and the latent variable approach [31, 32].

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 non-uniform distributions of p-values under the null as due to specific factors such as improper normalization, cross-hybridization, poorly characterized variance,

Tin Nguyen, Cristina Mitrea, and Rebecca Tagett are with the Department of Computer Science, Wayne State University, Detroit, MI 48202. E-mail: {tin, cristina, rtt}@wayne.edu.

Sorin Draghici is with the Department of Computer Science and the Department of Obstetrics and Gynecology, Wayne State University, Detroit, MI 48202. E-mail: sorin@wayne.edu

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 which can combine the p-values of multiple studies in a better way. Our contribution is two-fold. 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 non-parametric tests.

We apply DANUBE in the context of pathway analysis using 16 public gene expression datasets from two biological conditions, and 4 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 non-uniform 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, …, 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:

$$F_i(x) = Pr(X_i \leq x) = Pr(-2 \ln P_i \leq x) = Pr(P_i \geq e^{\frac{x}{2}})$$
$$= \int_{e^{-\frac{x}{2}}}^{1} f(p)dp = 1 - e^{-\frac{x}{2}}$$

The above function is the cumulative distribution function of a chi-squared distribution with two degrees of freedom (χ_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 (χ_{2m}^2). In summary, the log product of m independent p-values follows a chi-squared distribution with $2m$ degrees of freedom:

$$X = -2 \sum_{i=1}^{m} \ln(P_i) \sim \chi_{2m}^2 \tag{1}$$

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, …, 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 $\xi_t = \{t_1, t_2, \ldots, t_N\}$. 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, …, 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:

$$F(x) = \frac{1}{2} + \frac{1}{2m!} \sum_{i=0}^{m} (-1)^i \binom{m}{i} (x-i)^m \mathrm{sgn}(x-i) \tag{2}$$

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, 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 7 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].

Figure 1 displays the empirical null distributions of p-values using GSA, SPIA, and PADOG. The horizonal 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, 6 per method, display extreme examples of non-uniform p-value distributions for specific pathways. For each method, we show three distributions severely biased towards zero (eg. A1–A3), and three distributions severely biased towards one (eg. 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 non-uniformity 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 Figure 1: *Prostate cancer* (A1), *Adherens junction* (A2), and *Pathways in cancer* (A3), are reported as statistically significant in the results shown in Table I even though these data were collected in an experiment comparing Alzheimer's disease patients vs. healthy subjects, an experiment that has nothing to do with cancer.

The effect of combining control (i.e. healthy) samples from different experiments is to uniformly distribute all sources of bias among the random groups of samples. If we compare groups of control samples based on experiments, there could be true differences due to batch effects. By pooling them together, we form a population which is considered the reference population. This approach is similar to selecting from a large group of people that may contain different sub-groups (e.g. different ethnicities, gender, race, or living conditions). When we randomly select samples (for the two random groups to be compared) from the reference population, we expect all bias (e.g. ethnic subgroups) to be represented equally in both random groups and therefore, we should see no difference between these random groups, no matter how many distinct ethnic subgroups were present in the population at large. Therefore, the p-values of a test for difference between the two randomly selected groups should be equally probable between zero and one (see Supplementary Section 4 and Figures S10–S11 for more discussion).

We apply this procedure for the popular Gene Set Enrichment Analysis (GSEA) [36] using the exact same $40,000$ datasets simulated from the pool of control samples of Alzheimer's data. The resulting p-value distributions are uniform, as displayed in Supplementary Figure S1, showing not only that our resampled data correctly models the null, but also that GSEA is an unbiased test. This supports the idea that the non-uniformity of the distributions is due to the methods rather than the data. We also plot the top 24 most biased null distributions of GSEA (Figures S2) using the exact same data and exact same random grouping of samples. In each figure, the panels are sorted by the distribution means. The distributions of GSEA (Figures S2, S6) are uniform while those of GSA (Figures S3, S7), SPIA (Figures S4, S8), and PADOG (Figures S5, S9) are biased. Therefore, the bias is indeed due to the methods and not to one specific pathway.

III. Methods

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

A. The 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

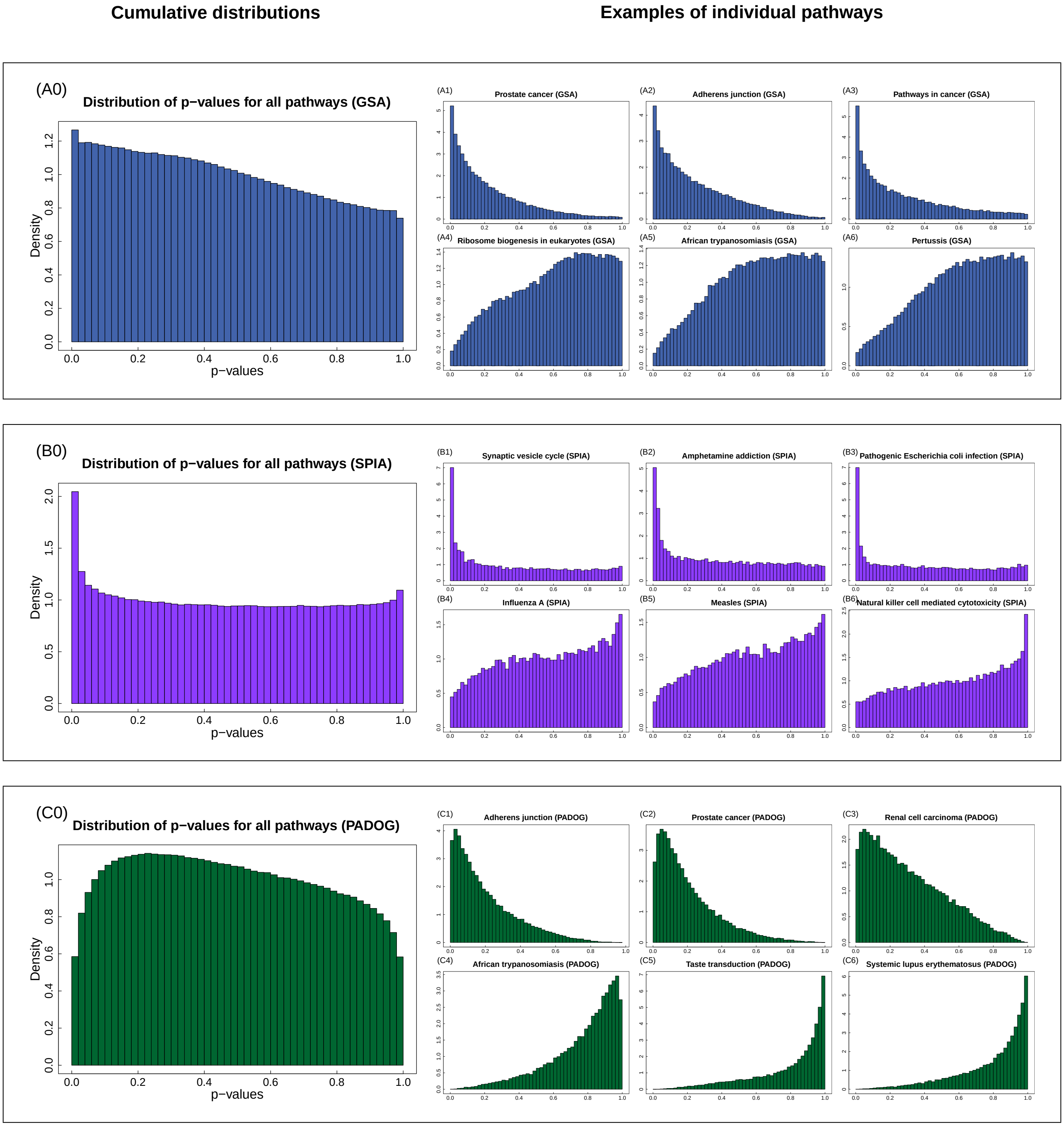


Fig. 1: The empirical null distributions of p-values using: Gene Set Analysis (GSA) - top, Signaling Pathway Impact Analysis (SPIA) - middle, and Down-weighting of Overlapping Genes (PADOG) - bottom. The distributions are generated by re-sampling from 74 control samples obtained from 7 public Alzheimer's datasets. The horizontal axes display the p-values while the vertical axes display the p-value densities. Panels A0-A6 (blue) show the distributions of p-values from GSA; panels B0-B6 (purple) show the distribution of p-values from SPIA; panels C0-C6 (green) show the distribution of p-values from PADOG. The large panels on the left, A0, B0, and C0, display the distributions of p-values cumulated from all KEGG signaling pathways. The smaller panels on the right display the p-value distributions of selected individual pathways, which are extreme cases. For each method, the upper three distributions, for example A1-A3, are biased towards zero and the lower three distributions, for example A4-A6, are biased towards one. Since none of these p-value distributions are uniform, there will be systematic bias in identifying significant pathways using any one of the 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 more likely to be among false negative results even if they may be implicated in the given phenotype.

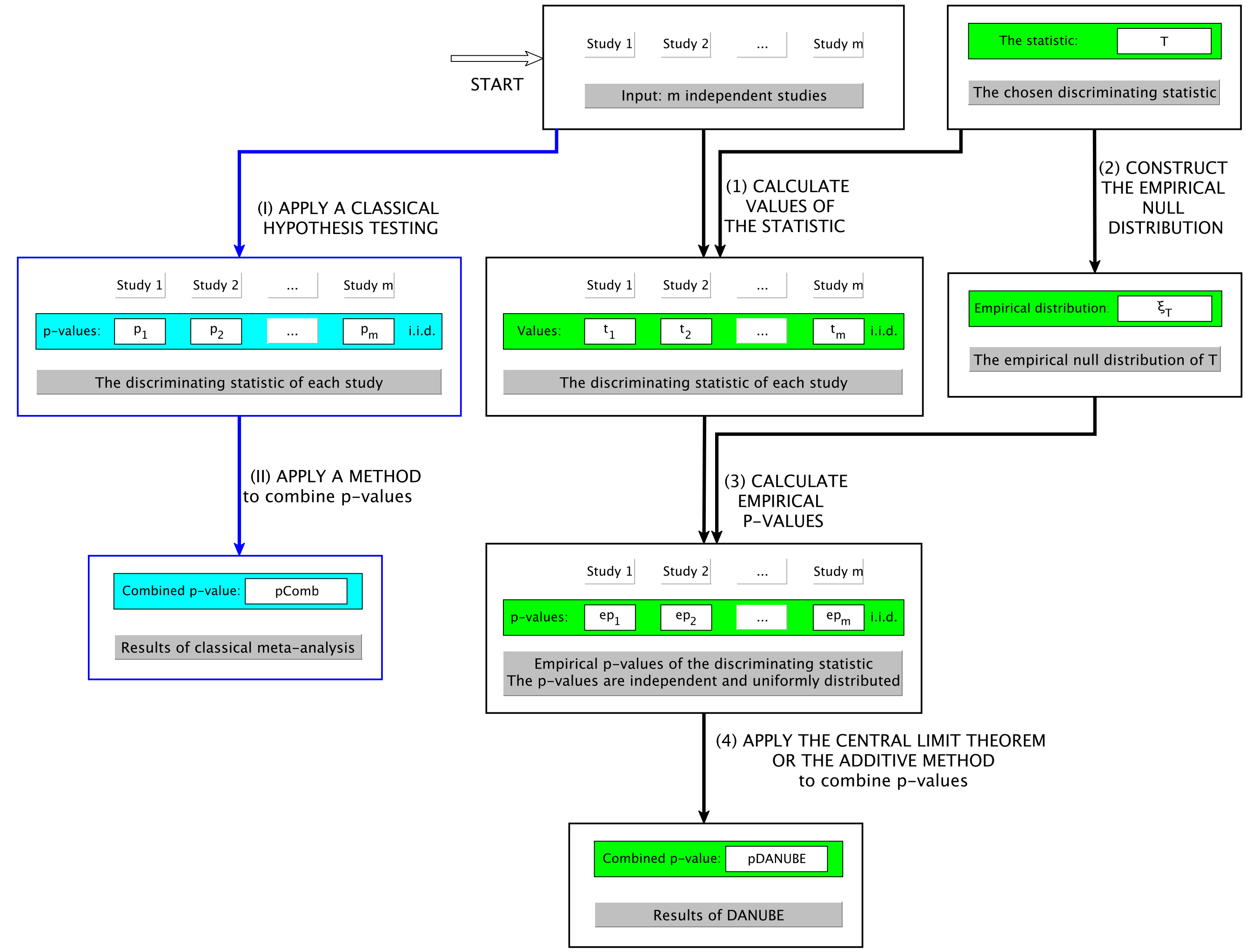

START
Study 1
Study 2
...
Study m
Input: m independent studies
The statistic:
T
The chosen discriminating statistic
(I) APPLY A CLASSICAL HYPOTHESIS TESTING
(1) CALCULATE VALUES OF THE STATISTIC
(2) CONSTRUCT THE EMPIRICAL NULL DISTRIBUTION
p-values:
p_1
p_2
p_m
i.i.d.
The discriminating statistic of each study
Values:
t_1
t_2
t_m
Empirical distribution
ξ_T
The empirical null distribution of T
(II) APPLY A METHOD to combine p-values
(3) CALCULATE EMPIRICAL P-VALUES
Combined p-value:
pComb
Results of classical meta-analysis
ep_1
ep_2
ep_m
Empirical p-values of the discriminating statistic
The p-values are independent and uniformly distributed
(4) APPLY THE CENTRAL LIMIT THEOREM OR THE ADDITIVE METHOD to combine p-values
pDANUBE
Results of DANUBE

Fig. 2: The 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 non-parametric 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 (1) 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 ξ_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 ξ_T. In step (4), we combine the m empirical p-values using either the additive method or the Central Limit Theorem (CLT).

the individual p-values are not distributed uniformly, as we have shown that it is the case for the pathway analysis methods. Figure 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 (1–4) is DANUBE.

The classical approach first calculates a p-value for each study using a parametric or non-parametric 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 Figure 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 Figure 2, we initially calculate the values t_1, t_2, …, 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 ξ_T for the chosen statistic. In step (3), we calculate the empirical p-values ep_1, ep_2, …, ep_m for the m studies with respect to the empirical null distribution ξ_T. For all $i \in \{1, 2, \ldots, m\}$, ep_i is calculated as the number of elements in ξ_T more extreme than t_i, divided by the total number of elements in ξ_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 ξ_T and the cumulative distribution function $F_T(T)$. We define the new random variable X as follows:*

$$X = \frac{|\{x : x \in \xi_T \wedge x \leq T\}|}{|\xi_T|} \tag{3}$$

where the numerator represents the number of elements of ξ_T that are smaller than or equal to T. If ξ_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:

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

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 cumulative distribution function of X, we have the following formula:

$$\begin{aligned} F_X(x) &= Pr(X \leq x) \\ &= Pr(F_T(T) \leq F_T(t)) \\ &= Pr(T \leq t) = F_T(t) = x \end{aligned} \tag{5}$$

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 equation (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 $\frac{1}{2}$ and $\frac{1}{12}$, 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 $\mu = \frac{1}{2}$ and variance $\sigma^2 = \frac{1}{12m}$. 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. The application of DANUBE in pathway analysis

Here we present the application of DANUBE in the context of pathway analysis (Figure 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 independent and identically distributed (i.i.d.). However, these p-values are not necessarily uniformly distributed under the null hypothesis (see Figure 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 Figure 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 datasets, 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 ξ_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 ξ_i. Using the formula described in Equation (2), we calculate the empirical p-values ep_{i1}, ep_{i2}, $\ldots$, ep_{im}. As we showed in the Methods section, 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 $pDANUBE_i$ 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. Sections IV-A and IV-B compare the classical approaches with DANUBE for the specific application domain of pathway analysis. Sections IV-C and IV-D compare the classical meta-analysis approaches with 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 with 5 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 4 pathway analysis methods (GSEA, GSA, SPIA, and PADOG). We also compare these methods with a stand-alone meta-analysis method, MetaPath. In total, we analyze the results of 25 approaches: 6 meta-analyses combined with 4 pathway analysis methods, plus MetaPath [11, 50]. Each of these methods is tested on two diseases, one is Alzheimer's disease with 7 and the other is acute myeloid leukemia (AML) with 9 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.

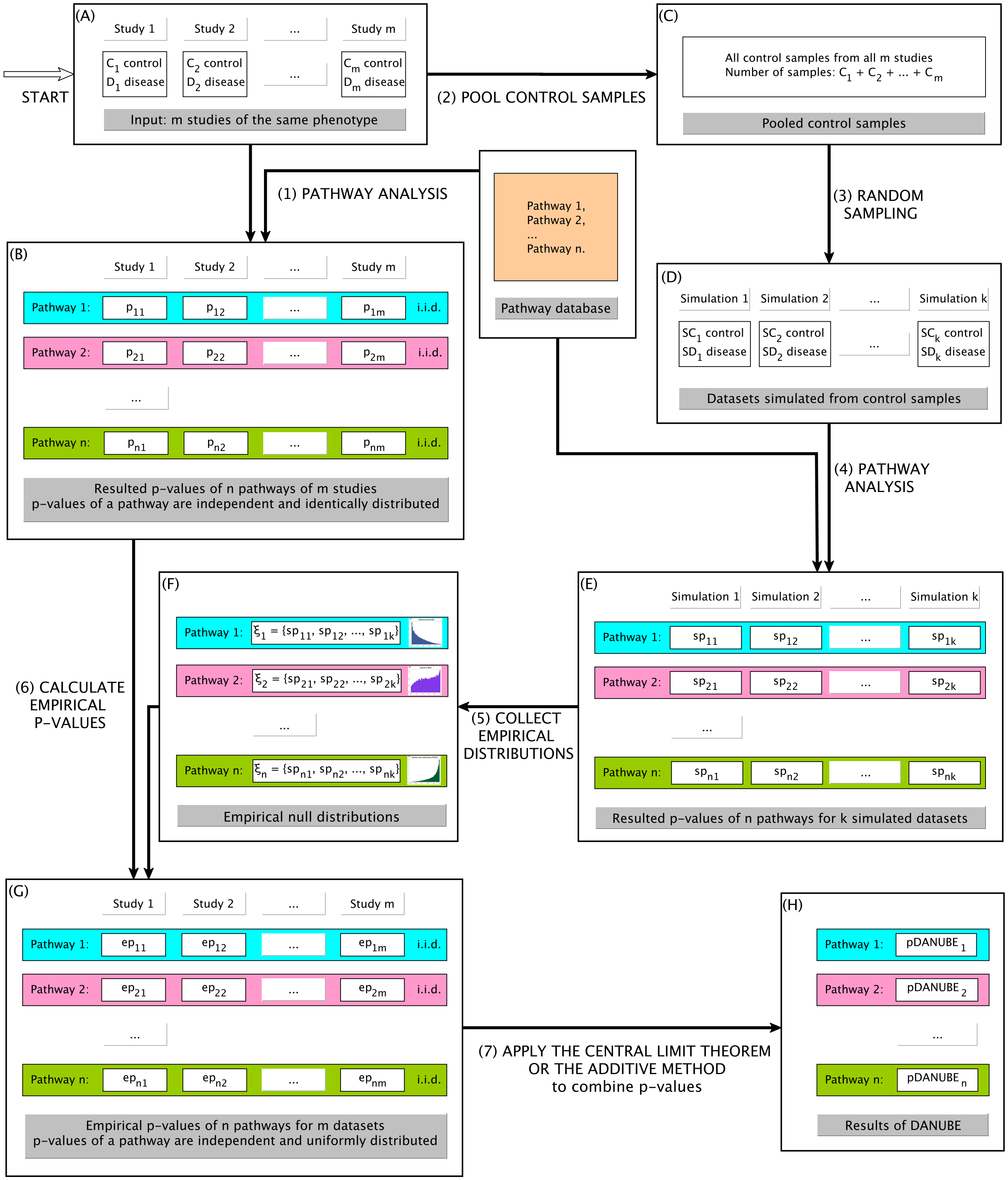

(A)
Study 1
Study 2
...
Study m
C_1 control
D_1 disease
C_2 control
D_2 disease
...
C_m control
D_m disease
Input: m studies of the same phenotype
START
(2) POOL CONTROL SAMPLES
(C)
All control samples from all m studies
Number of samples: C_1 + C_2 + ... + C_m
Pooled control samples
(1) PATHWAY ANALYSIS
Pathway 1,
Pathway 2,
...
Pathway n.
Pathway database
(3) RANDOM SAMPLING
(B)
Study 1
Study 2
...
Study m
Pathway 1: p_{11} p_{12} ... p_{1m} i.i.d.
Pathway 2: p_{21} p_{22} ... p_{2m} i.i.d.
...
Pathway n: p_{n1} p_{n2} ... p_{nm} i.i.d.
Resulted p−values of n pathways of m studies
p−values of a pathway are independent and identically distributed
(D)
Simulation 1
Simulation 2
...
Simulation k
SC_1 control
SD_1 disease
SC_2 control
SD_2 disease
...
SC_k control
SD_k disease
Datasets simulated from control samples
(4) PATHWAY ANALYSIS
(6) CALCULATE EMPIRICAL P−VALUES
(F)
Pathway 1: $\xi_1 = \{sp_{11}, sp_{12}, ..., sp_{1k}\}$
Pathway 2: $\xi_2 = \{sp_{21}, sp_{22}, ..., sp_{2k}\}$
...
Pathway n: $\xi_n = \{sp_{n1}, sp_{n2}, ..., sp_{nk}\}$
Empirical null distributions
(5) COLLECT EMPIRICAL DISTRIBUTIONS
(E)
Simulation 1
Simulation 2
...
Simulation k
Pathway 1: sp_{11} sp_{12} ... sp_{1k}
Pathway 2: sp_{21} sp_{22} ... sp_{2k}
...
Pathway n: sp_{n1} sp_{n2} ... sp_{nk}
Resulted p−values of n pathways for k simulated datasets
(G)
Study 1
Study 2
...
Study m
Pathway 1: ep_{11} ep_{12} ... ep_{1m} i.i.d.
Pathway 2: ep_{21} ep_{22} ... ep_{2m} i.i.d.
...
Pathway n: ep_{n1} ep_{n2} ... ep_{nm} i.i.d.
Empirical p−values of n pathways for m datasets
p−values of a pathway are independent and uniformly distributed
(7) APPLY THE CENTRAL LIMIT THEOREM
OR THE ADDITIVE METHOD
to combine p−values
(H)
Pathway 1: $pDANUBE_1$
Pathway 2: $pDANUBE_2$
...
Pathway n: $pDANUBE_n$
Results of DANUBE

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 (eg. GSA, SPIA, or PADOG). For each pathway, the resulting m p-values are independent and identically distributed (i.i.d.). However, these p-values are not uniformly distributed under the null hypothesis (see Figure 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): build an empirical distribution for each pathway, which consists of k p-values obtained under the null hypothesis. Step (6): calculate an empirical p-value for each p-value obtained from step (1). For example, using the empirical distribution ξ_1, we calculate the empirical p-value ep_{11} as the probability of observing a p-value more extreme than p_1, i.e., $ep_{11} = |\{sp_{1i} \leq p_{11}, i \in [1..k]\}|$. Step (7): combine the m empirical p-values obtained for each pathway using either the additive method or the Central Limit Theorem.

A. Pathway analysis applications: Alzheimer's disease

The Alzheimer's datasets we use in our data analysis are GSE28146 (hippocampus) and GSE5281 (6 different tissues: entorhinal cortex (EC), hippocampus (HIP), medial temporal gyrus (MTG), posterior cingulate (PC), superior frontal gyrus (SFG), and primary visual cortex (VCX)). The 4 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 7 Alzheimer's datasets are displayed in Figure 4. The graphs demonstrate that the adjusted p-values and rankings of the target pathway vary substantially between the 4 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 7 studies.

We combine the 4 pathway analysis methods with 6 meta-analyses: Stouffer's, Z-method, Brown's, Fisher's, the additive method, and DANUBE. Using a pathway analysis method M, each pathway has 7 p-values – one per study. These 7 p-values are combined using each of the 6 meta analysis methods Therefore, each pathway analysis method produces 6 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 7 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 4 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–S2.

Table I displays the results using GSA combined with the 6 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 Figure 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 I. 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 II displays the results using PADOG combined with the 6 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 Figure 1 panels C1–C3 and Supplementary Figure S5) are among the significant pathways.

Supplementary Table S3 displays the results using SPIA combined with the 6 meta-analysis methods. The target pathway is significant and is ranked near the top for all methods. DANUBE yields the shortest list of significant pathways. All the 5 significant pathways, *Parkinson's disease*, *Alzheimer's disease*, *Synaptic vesicle cycle*, *Cardiac muscle contration*, and *Huntington's disease* are also significant when we combine DANUBE with GSA and PADOG.

Supplementary Table S4 displays the results using GSEA combined with the 6 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 6 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 4 significant pathways, *Cardiac muscle contration*, *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 4 pathway analysis methods when combined with DANUBE. The target pathway is also ranked higher when using DANUBE compared to the results of other 5 meta-analysis methods. In addition, the pathways *Parkinson's disease*, *Alzheimer's disease*, *Cardiac muscle constration*, and *Huntington's disease*, consistently appear as significant in the results of all the 4 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*

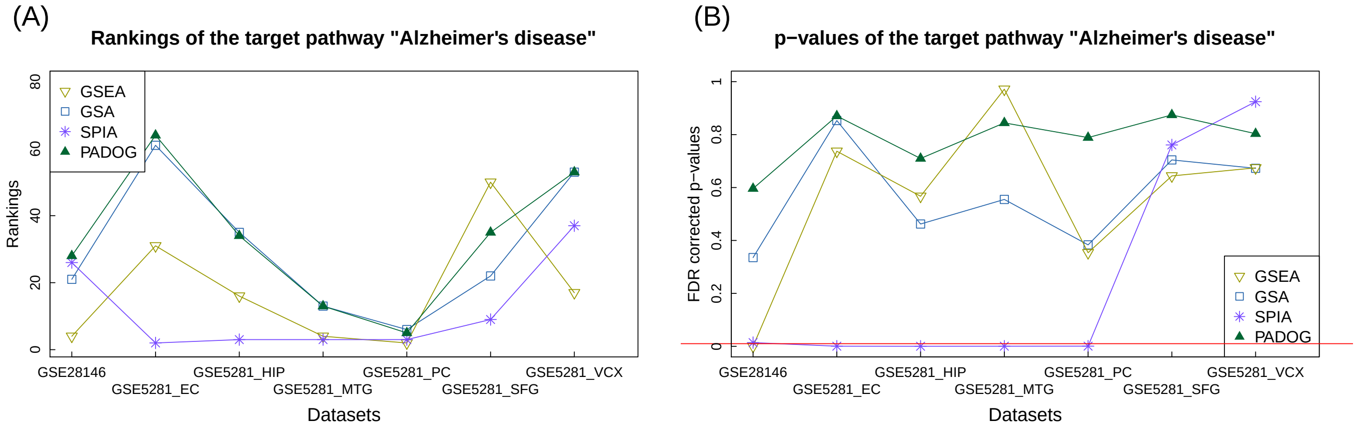


Fig. 4: Ranks (panel A) and p-values (panel B) of the KEGG target pathway, *Alzheimer's disease*, for 7 Alzheimer's datasets, using the pathway analysis methods: Gene Set Enrichment Analysis (GSEA), Gene Set Analysis (GSA), Signaling Pathway Impact Analysis (SPIA), and Down-weighting of Overlapping Genes (PADOG). The horizontal axes show the 7 Alzheimer's datasets. The vertical axis in panel (A) shows the rankings of the target pathway for each dataset using the 4 methods. The vertical axis in panel (B) shows the FDR-corrected p-values of the target pathway. The red horizontal line in (B) shows the threshold 0.01. Note how the rankings and p-values of the target pathway vary greatly across different datasets and methods, making the interpretation of the results very difficult.

TABLE I: The 17 top ranked pathways and FDR-corrected p-values obtained by combining the GSA p-values using 6 meta-analysis methods for Alzheimer's disease. Stouffer's method, the additive method, and DANUBE, identify the target pathway as significant and rank it in positions 11^{th}, 6^{th}, and 2^{nd}, respectively. DANUBE yields the best ranking.

	GSA + Stouffer's method		GSA + Z-method		GSA + Brown's method	
	Pathway	pvalue.fdr	Pathway	pvalue.fdr	Pathway	pvalue.fdr
1	Vasopressin-regulated water reabsorption	$< 10^{-4}$	Vasopressin-regulated water reabsorption	$< 10^{-4}$	Vasopressin-regulated water reabsorption	$< 10^{-4}$
2	Pathogenic Escherichia coli infection	$< 10^{-4}$	Pathogenic Escherichia coli infection	$< 10^{-4}$	Pathogenic Escherichia coli infection	$< 10^{-4}$
3	Prostate cancer	$< 10^{-4}$	Prostate cancer	0.0307	Prostate cancer	0.0418
4	Pathways in cancer	0.0003	Pathways in cancer	0.1352	Adherens junction	0.1722
5	Adherens junction	0.0003	Adherens junction	0.1352	Pathways in cancer	0.1722
6	Hippo signaling pathway	0.0004	Hippo signaling pathway	0.1352	Hippo signaling pathway	0.1765
7	Synaptic vesicle cycle	0.0032	Synaptic vesicle cycle	0.2443	Synaptic vesicle cycle	0.2625
8	Vibrio cholerae infection	0.0032	Vibrio cholerae infection	0.2443	Endocrine and other factor-regulated calcium reabsorption	0.2625
9	Endocrine and other factor-regulated calcium reabsorption	0.0032	Endocrine and other factor-regulated calcium reabsorption	0.2443	Vibrio cholerae infection	0.2625
10	Shigellosis	0.0071	Shigellosis	0.2808	Pancreatic cancer	0.2625
11	Alzheimer's disease	0.0073	Alzheimer's disease	0.2808	Focal adhesion	0.2950
12	Bacterial invasion of epithelial cells	0.0073	Bacterial invasion of epithelial cells	0.2808	Shigellosis	0.3027
13	Pancreatic cancer	0.0095	Pancreatic cancer	0.2808	Bacterial invasion of epithelial cells	0.3034
14	Focal adhesion	0.0112	Focal adhesion	0.2808	Notch signaling pathway	0.3254
15	Parkinson's disease	0.0112	Parkinson's disease	0.2808	Alzheimer's disease	0.3254
16	Huntington's disease	0.0112	Huntington's disease	0.2808	HIF-1 signaling pathway	0.3274
17	Wnt signaling pathway	0.0112	Wnt signaling pathway	0.2808	SNARE interactions in vesicular transport	0.3274

	GSA + Fisher's method		GSA + Additive method		**GSA + DANUBE**	
	Pathway	pvalue.fdr	Pathway	pvalue.fdr	Pathway	pvalue.fdr
1	Vasopressin-regulated water reabsorption	$< 10^{-4}$	Prostate cancer	$< 10^{-4}$	Cardiac muscle contraction	0.0014
2	Pathogenic Escherichia coli infection	$< 10^{-4}$	Pathways in cancer	0.0002	Alzheimer's disease	0.0014
3	Prostate cancer	$< 10^{-4}$	Hippo signaling pathway	0.0005	Huntington's disease	0.0014
4	Adherens junction	0.0019	Adherens junction	0.0015	Parkinson's disease	0.0014
5	Pathways in cancer	0.0023	Endocrine and other factor-regulated calcium reabsorption	0.0042	Hippo signaling pathway	0.0025
6	Hippo signaling pathway	0.0030	Alzheimer's disease	0.0042	Vibrio cholerae infection	0.0047
7	Synaptic vesicle cycle	0.0097	Vibrio cholerae infection	0.0057	Synaptic vesicle cycle	0.0081
8	Vibrio cholerae infection	0.0121	Shigellosis	0.0057	Prostate cancer	0.0112
9	Endocrine and other factor-regulated calcium reabsorption	0.0133	Huntington's disease	0.0057	Vasopressin-regulated water reabsorption	0.0112
10	Pancreatic cancer	0.0133	Bacterial invasion of epithelial cells	0.0057	Epithelial cell signaling in Helicobacter pylori infection	0.0118
11	Focal adhesion	0.0190	Parkinson's disease	0.0057	Systemic lupus erythematosus	0.0150
12	Shigellosis	0.0222	Glioma	0.0057	Amyotrophic lateral sclerosis (ALS)	0.0174
13	Bacterial invasion of epithelial cells	0.0245	Vasopressin-regulated water reabsorption	0.0057	Shigellosis	0.0193
14	Alzheimer's disease	0.0334	Cardiac muscle contraction	0.0057	Endocrine and other factor-regulated calcium reabsorption	0.0193
15	Notch signaling pathway	0.0334	Wnt signaling pathway	0.0057	Phagosome	0.0302
16	SNARE interactions in vesicular transport	0.0465	Synaptic vesicle cycle	0.0057	Lysosome	0.0302
17	Wnt signaling pathway	0.0465	Dorso-ventral axis formation	0.0119	Ribosome biogenesis in eukaryotes	0.0302

The horizontal lines show the 1% significance threshold. 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 S3). These include *Adherens junction* and several cancer-related pathways, which are not considered to be implicated in Alzheimer's disease.

TABLE II: The 20 top ranked pathways and FDR-corrected p-values obtained by combining the PADOG p-values using 6 meta-analysis methods for Alzheimer's disease. Only DANUBE identifies the target pathway *Alzheimer's disease* as significant and ranks it in position 6^{th}.

	PADOG + Stouffer's method		PADOG + Z-method		PADOG + Brown's method	
	Pathway	pvalue.fdr	Pathway	pvalue.fdr	Pathway	pvalue.fdr
1	Adherens junction	$< 10^{-4}$	Adherens junction	0.6725	HIF-1 signaling pathway	0.6495
2	Shigellosis	0.0002	Shigellosis	0.6725	Adherens junction	0.6495
3	Renal cell carcinoma	0.0002	Renal cell carcinoma	0.6725	Gap junction	0.6495
4	Prostate cancer	0.0005	Prostate cancer	0.6725	Long-term potentiation	0.6495
5	Bacterial invasion of epithelial cells	0.0014	Bacterial invasion of epithelial cells	0.6725	Long-term depression	0.6495
6	Long-term depression	0.0036	Long-term depression	0.6725	Endocrine and other factor-regulated calcium reabsorption	0.6495
7	Pathogenic Escherichia coli infection	0.0036	Pathogenic Escherichia coli infection	0.6725	Bacterial invasion of epithelial cells	0.6495
8	Colorectal cancer	0.0036	Colorectal cancer	0.6725	Vibrio cholerae infection	0.6495
9	Gap junction	0.0036	Gap junction	0.6725	Pathogenic Escherichia coli infection	0.6495
10	Glioma	0.0036	Glioma	0.6725	Shigellosis	0.6495
11	Pancreatic cancer	0.0036	Pancreatic cancer	0.6725	Colorectal cancer	0.6495
12	Vibrio cholerae infection	0.0036	Vibrio cholerae infection	0.6725	Renal cell carcinoma	0.6495
13	Endocrine and other factor-regulated calcium reabsorption	0.0043	Endocrine and other factor-regulated calcium reabsorption	0.6725	Pancreatic cancer	0.6495
14	ErbB signaling pathway	0.0053	ErbB signaling pathway	0.6725	Endometrial cancer	0.6495
15	Endometrial cancer	0.0063	Endometrial cancer	0.6725	Glioma	0.6495
16	HIF-1 signaling pathway	0.0063	HIF-1 signaling pathway	0.6725	Prostate cancer	0.6495
17	Neurotrophin signaling pathway	0.0067	Neurotrophin signaling pathway	0.6725	ErbB signaling pathway	0.6533
18	Long-term potentiation	0.0076	Long-term potentiation	0.6725	Neurotrophin signaling pathway	0.6533
19	Synaptic vesicle cycle	0.0160	Synaptic vesicle cycle	0.7324	mRNA surveillance pathway	0.7157
20	VEGF signaling pathway	0.0317	VEGF signaling pathway	0.7324	MAPK signaling pathway	0.7157

	PADOG + Fisher's method		PADOG + Additive method		**PADOG + DANUBE**	
	Pathway	pvalue.fdr	Pathway	pvalue.fdr	Pathway	pvalue.fdr
1	Adherens junction	0.0008	Adherens junction	$< 10^{-4}$	Vibrio cholerae infection	$< 10^{-4}$
2	Shigellosis	0.0022	Renal cell carcinoma	$< 10^{-4}$	Shigellosis	$< 10^{-4}$
3	Renal cell carcinoma	0.0022	Shigellosis	$< 10^{-4}$	Parkinson's disease	0.0007
4	Prostate cancer	0.0049	Prostate cancer	0.0001	Synaptic vesicle cycle	0.0007
5	Bacterial invasion of epithelial cells	0.0065	Long-term depression	0.0006	Gap junction	0.0007
6	Pathogenic Escherichia coli infection	0.0149	Colorectal cancer	0.0009	Alzheimer's disease	0.0007
7	Endocrine and other factor-regulated calcium reabsorption	0.0199	Gap junction	0.0011	Pathogenic Escherichia coli infection	0.0007
8	Glioma	0.0199	ErbB signaling pathway	0.0013	Cardiac muscle contraction	0.0007
9	Pancreatic cancer	0.0199	Bacterial invasion of epithelial cells	0.0013	Epithelial cell signaling in Helicobacter pylori infection	0.0009
10	Long-term depression	0.0199	Vibrio cholerae infection	0.0013	Huntington's disease	0.0013
11	Gap junction	0.0199	Pancreatic cancer	0.0021	Renal cell carcinoma	0.0024
12	Colorectal cancer	0.0199	Glioma	0.0022	Vasopressin-regulated water reabsorption	0.0047
13	Vibrio cholerae infection	0.0199	Neurotrophin signaling pathway	0.0028	VEGF signaling pathway	0.0052
14	Long-term potentiation	0.0226	HIF-1 signaling pathway	0.0037	Endocrine and other factor-regulated calcium reabsorption	0.0072
15	Endometrial cancer	0.0226	Pathogenic Escherichia coli infection	0.0042	Bacterial invasion of epithelial cells	0.0078
16	HIF-1 signaling pathway	0.0257	Endometrial cancer	0.0052	GABAergic synapse	0.0102
17	ErbB signaling pathway	0.0326	VEGF signaling pathway	0.0052	Adherens junction	0.0103
18	Neurotrophin signaling pathway	0.0352	Endocrine and other factor-regulated calcium reabsorption	0.0052	Long-term depression	0.0103
19	Synaptic vesicle cycle	0.0600	Synaptic vesicle cycle	0.0086	Salmonella infection	0.0134
20	Dopaminergic synapse	0.1305	Long-term potentiation	0.0106	Colorectal cancer	0.0198

The horizontal lines show the 1% significance threshold. The target pathway *Alzheimer's disease* is highlighted in green. Pathways highlighted in red are examples of false positives (see Figure 1 panels C1-C3 and Supplementary Figure S5).

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 7 studies. MetaPath is a stand-alone 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 7 pathways using MetaPath for the 7 Alzheimer's datasets. The target pathway *Alzheimer's disease* is not significant and is outranked by 6 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* for the 9 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 4 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 4 pathway analysis methods with the 6 meta-analysis methods. Using a pathway analysis method M, each pathway has 9 p-values – one per study. These 9 p-values are combined using each of the 6 meta-analysis methods Therefore, each pathway analysis method produces 6 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 9 AML datasets, and randomly designated "control" and "disease" subgroups. We generated $10,000$ simulations of 10 controls and 10 diseases, $10,000$ simulations 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 4 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, we adjusted the combined p-values for multiple comparisons using FDR.

Table III displays the results of GSA combined with the 6 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 1^{st}), followed by the additive (2^{nd}) and Stouffer's method (13^{th}). In addition, the target pathway is the only significant pathway in DANUBE's result.

Table IV shows the results of PADOG combined with the 6 meta-analysis methods. The target pathway is significant for the 4 methods: DANUBE, Stouffer's, Fisher's, and the additive method. For DANUBE, *Acute myeloid leukemia* is ranked 1^{st} compared to 7^{th} 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 6 meta-analysis methods, ordered by the FDR corrected p-value. Again, the target pathway is significant using Stouffer's, Fisher's, the additive method, and DANUBE. The additive method and DANUBE have the same list of significant pathways. In addition, both methods place the target pathway higher than the other two methods.

Supplementary Table S7 displays the results of GSEA combined with the 6 meta-analysis methods. The target pathway *Acute myeloid leukemia* is highlighted in green. For all 6 meta-analyses, the target pathway is not significant despite being ranked among the top pathways. Since GSEA has no bias, the additive method and DANUBE yield similar results. In essence, even though it is completely unbiased, GSEA lacks the power to identify the *Acute myeloid leukemia* (AML) as significant in the AML data.

We also use MetaPath to combine the 9 acute myeloid leukemia studies. Supplementary Table S8 shows the top 5 pathways using MetaPath. The target pathway is not significant (p=0.4), and is outranked by 2 other pathways.

Table V summarizes all the results for the 25 approaches (4 pathway analysis methods each combined with one of 6 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 non-parametric 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 Figure 5, panel (a) displays a simulated null distribution H_0 which 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 everytime 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

TABLE III: The 21 top ranked pathways and FDR-corrected p-values obtained by combining the GSA p-values using 6 meta-analysis methods for acute myeloid leukemia (AML). The target pathway *Acute myeloid leukemia* is significant for Stouffer's, the additive method, and DANUBE with rankings 13^{th}, 2^{nd}, and 1^{st}, respectively.

	GSA + Stouffer's method		GSA + Z-method		GSA + Brown's method	
	Pathway	pvalue.fdr	Pathway	pvalue.fdr	Pathway	pvalue.fdr
1	ErbB signaling pathway	$< 10^{-4}$	ErbB signaling pathway	$< 10^{-4}$	ErbB signaling pathway	$< 10^{-4}$
2	Sulfur relay system	$< 10^{-4}$	Sulfur relay system	$< 10^{-4}$	Sulfur relay system	$< 10^{-4}$
3	Adherens junction	$< 10^{-4}$	Adherens junction	$< 10^{-4}$	Adherens junction	$< 10^{-4}$
4	Tight junction	$< 10^{-4}$	Tight junction	$< 10^{-4}$	Tight junction	$< 10^{-4}$
5	Circadian rhythm	$< 10^{-4}$	Circadian rhythm	$< 10^{-4}$	Circadian rhythm	$< 10^{-4}$
6	Alcoholism	$< 10^{-4}$	Alcoholism	$< 10^{-4}$	Alcoholism	$< 10^{-4}$
7	Shigellosis	$< 10^{-4}$	Shigellosis	$< 10^{-4}$	Shigellosis	$< 10^{-4}$
8	Transcriptional misregulation in cancer	$< 10^{-4}$	Transcriptional misregulation in cancer	$< 10^{-4}$	Transcriptional misregulation in cancer	$< 10^{-4}$
9	Renal cell carcinoma	$< 10^{-4}$	Renal cell carcinoma	$< 10^{-4}$	Renal cell carcinoma	$< 10^{-4}$
10	Glioma	$< 10^{-4}$	Glioma	$< 10^{-4}$	Glioma	$< 10^{-4}$
11	Systemic lupus erythematosus	$< 10^{-4}$	Systemic lupus erythematosus	$< 10^{-4}$	Systemic lupus erythematosus	$< 10^{-4}$
12	Non-small cell lung cancer	0.0003	Non-small cell lung cancer	0.0606	Non-small cell lung cancer	0.1250
13	Acute myeloid leukemia	0.0012	Acute myeloid leukemia	0.1011	mTOR signaling pathway	0.2120
14	VEGF signaling pathway	0.0017	VEGF signaling pathway	0.1139	VEGF signaling pathway	0.2120
15	Endometrial cancer	0.0025	Endometrial cancer	0.1298	Pathways in cancer	0.2120
16	Pathways in cancer	0.0029	Pathways in cancer	0.1352	Acute myeloid leukemia	0.2120
17	mTOR signaling pathway	0.0033	mTOR signaling pathway	0.1386	HIF-1 signaling pathway	0.2252
18	Chronic myeloid leukemia	0.0081	Chronic myeloid leukemia	0.1933	Endometrial cancer	0.2252
19	Prostate cancer	0.0081	Prostate cancer	0.1933	Prostate cancer	0.2252
20	Pancreatic cancer	0.0097	Pancreatic cancer	0.2037	Insulin signaling pathway	0.2379
21	HIF-1 signaling pathway	0.0150	HIF-1 signaling pathway	0.2394	Pancreatic cancer	0.2628

	GSA + Fisher's method		GSA + Additive method		**GSA + DANUBE**	
	Pathway	pvalue.fdr	Pathway	pvalue.fdr	Pathway	pvalue.fdr
1	ErbB signaling pathway	$< 10^{-4}$	Non-small cell lung cancer	0.0003	Acute myeloid leukemia	0.0065
2	Sulfur relay system	$< 10^{-4}$	Acute myeloid leukemia	0.0003	Transcriptional misregulation in cancer	0.0231
3	Adherens junction	$< 10^{-4}$	VEGF signaling pathway	0.0005	VEGF signaling pathway	0.0489
4	Tight junction	$< 10^{-4}$	ErbB signaling pathway	0.0005	Alcoholism	0.1161
5	Circadian rhythm	$< 10^{-4}$	Endometrial cancer	0.0008	Non-small cell lung cancer	0.5968
6	Alcoholism	$< 10^{-4}$	Transcriptional misregulation in cancer	0.0020	Bladder cancer	0.5968
7	Shigellosis	$< 10^{-4}$	Chronic myeloid leukemia	0.0038	HIF-1 signaling pathway	0.5968
8	Transcriptional misregulation in cancer	$< 10^{-4}$	mTOR signaling pathway	0.0043	Apoptosis	0.5968
9	Renal cell carcinoma	$< 10^{-4}$	Pathways in cancer	0.0043	mTOR signaling pathway	0.5968
10	Glioma	$< 10^{-4}$	Colorectal cancer	0.0084	Cocaine addiction	0.5968
11	Systemic lupus erythematosus	$< 10^{-4}$	Glioma	0.0108	Autoimmune thyroid disease	0.6141
12	Non-small cell lung cancer	0.0048	Pancreatic cancer	0.0108	Amyotrophic lateral sclerosis (ALS)	0.6458
13	Pathways in cancer	0.0153	Prostate cancer	0.0108	Notch signaling pathway	0.6458
14	Acute myeloid leukemia	0.0181	Small cell lung cancer	0.0177	ErbB signaling pathway	0.6458
15	mTOR signaling pathway	0.0188	Bacterial invasion of epithelial cells	0.0177	HTLV-I infection	0.6458
16	VEGF signaling pathway	0.0188	Adherens junction	0.0184	Natural killer cell mediated cytotoxicity	0.6458
17	Endometrial cancer	0.0243	Renal cell carcinoma	0.0239	Chronic myeloid leukemia	0.6458
18	HIF-1 signaling pathway	0.0252	Melanoma	0.0326	Endocytosis	0.6458
19	Prostate cancer	0.0252	Endocytosis	0.0403	Small cell lung cancer	0.6458
20	Insulin signaling pathway	0.0295	HIF-1 signaling pathway	0.0447	Fc gamma R-mediated phagocytosis	0.6458
21	Pancreatic cancer	0.0378	Circadian rhythm	0.0447	African trypanosomiasis	0.6458

The horizontal lines show the 1% significance threshold. The target pathway *Acute myeloid leukemia* is highlighted in green.

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, 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

TABLE IV: The 23 top ranked pathways and FDR-corrected p-values obtained by combining the PADOG p-values using 6 meta-analysis methods for acute myeloid leukemia (AML). The target pathway *Acute myeloid leukemia* is significant for Stouffer's, Fisher's, the additive method and DANUBE. DANUBE yields the best ranking.

	PADOG + Stouffer's method		PADOG + Z-method		PADOG + Brown's method	
	Pathway	pvalue.fdr	Pathway	pvalue.fdr	Pathway	pvalue.fdr
1	Non-small cell lung cancer	$< 10^{-4}$	Non-small cell lung cancer	0.0705	Chronic myeloid leukemia	0.0412
2	Chronic myeloid leukemia	$< 10^{-4}$	Chronic myeloid leukemia	0.0705	Non-small cell lung cancer	0.0412
3	Glioma	$< 10^{-4}$	Glioma	0.2152	Glioma	0.1240
4	ErbB signaling pathway	$< 10^{-4}$	ErbB signaling pathway	0.2239	ErbB signaling pathway	0.2149
5	Colorectal cancer	$< 10^{-4}$	Colorectal cancer	0.2565	VEGF signaling pathway	0.2806
6	Prostate cancer	$< 10^{-4}$	Prostate cancer	0.2565	Pathways in cancer	0.2806
7	Acute myeloid leukemia	$< 10^{-4}$	Acute myeloid leukemia	0.2565	Colorectal cancer	0.2806
8	VEGF signaling pathway	0.0001	VEGF signaling pathway	0.2565	Pancreatic cancer	0.2806
9	Endometrial cancer	0.0001	Endometrial cancer	0.2565	Prostate cancer	0.2806
10	Pancreatic cancer	0.0001	Pancreatic cancer	0.2565	Acute myeloid leukemia	0.2806
11	Pathways in cancer	0.0001	Pathways in cancer	0.2565	Endometrial cancer	0.3398
12	Transcriptional misregulation in cancer	0.0005	Transcriptional misregulation in cancer	0.3509	mTOR signaling pathway	0.4198
13	T cell receptor signaling pathway	0.0012	T cell receptor signaling pathway	0.4055	T cell receptor signaling pathway	0.4198
14	mTOR signaling pathway	0.0012	mTOR signaling pathway	0.4055	Circadian rhythm	0.4198
15	Circadian rhythm	0.0015	Circadian rhythm	0.4061	Insulin signaling pathway	0.4198
16	Neurotrophin signaling pathway	0.0021	Neurotrophin signaling pathway	0.4184	Transcriptional misregulation in cancer	0.4198
17	Small cell lung cancer	0.0024	Small cell lung cancer	0.4184	Small cell lung cancer	0.4491
18	Renal cell carcinoma	0.0054	Renal cell carcinoma	0.4837	Neurotrophin signaling pathway	0.4568
19	Insulin signaling pathway	0.0063	Insulin signaling pathway	0.4837	mRNA surveillance pathway	0.4695
20	Endocytosis	0.0070	Endocytosis	0.4837	MAPK signaling pathway	0.4695
21	Adherens junction	0.0070	Adherens junction	0.4837	HIF-1 signaling pathway	0.4695
22	Wnt signaling pathway	0.0168	Wnt signaling pathway	0.5674	Endocytosis	0.4695
23	Melanoma	0.0195	Melanoma	0.5674	Wnt signaling pathway	0.4695

	PADOG + Fisher's method		PADOG + Additive method		**PADOG + DANUBE**	
	Pathway	pvalue.fdr	Pathway	pvalue.fdr	Pathway	pvalue.fdr
1	Chronic myeloid leukemia	$< 10^{-4}$	Non-small cell lung cancer	$< 10^{-4}$	Acute myeloid leukemia	$< 10^{-4}$
2	Non-small cell lung cancer	$< 10^{-4}$	Chronic myeloid leukemia	$< 10^{-4}$	VEGF signaling pathway	0.0007
3	Glioma	$< 10^{-4}$	ErbB signaling pathway	$< 10^{-4}$	Non-small cell lung cancer	0.0008
4	ErbB signaling pathway	$< 10^{-4}$	Endometrial cancer	$< 10^{-4}$	T cell receptor signaling pathway	0.0021
5	Colorectal cancer	0.0003	Glioma	$< 10^{-4}$	Colorectal cancer	0.0023
6	Prostate cancer	0.0006	Colorectal cancer	$< 10^{-4}$	Chronic myeloid leukemia	0.0027
7	Acute myeloid leukemia	0.0006	Acute myeloid leukemia	$< 10^{-4}$	Endometrial cancer	0.0057
8	Pancreatic cancer	0.0007	Prostate cancer	$< 10^{-4}$	Transcriptional misregulation in cancer	0.0095
9	VEGF signaling pathway	0.0007	Transcriptional misregulation in cancer	0.0001	Glioma	0.0153
10	Pathways in cancer	0.0009	VEGF signaling pathway	0.0001	mTOR signaling pathway	0.0160
11	Endometrial cancer	0.0021	Pathways in cancer	0.0001	Prostate cancer	0.0203
12	Transcriptional misregulation in cancer	0.0056	Pancreatic cancer	0.0002	Apoptosis	0.0239
13	T cell receptor signaling pathway	0.0080	mTOR signaling pathway	0.0005	ErbB signaling pathway	0.0390
14	mTOR signaling pathway	0.0098	Neurotrophin signaling pathway	0.0005	B cell receptor signaling pathway	0.0464
15	Insulin signaling pathway	0.0098	Renal cell carcinoma	0.0006	Circadian rhythm	0.0521
16	Circadian rhythm	0.0098	T cell receptor signaling pathway	0.0006	Thyroid cancer	0.0844
17	Small cell lung cancer	0.0138	Circadian rhythm	0.0006	Progesterone-mediated oocyte maturation	0.1040
18	Neurotrophin signaling pathway	0.0165	Small cell lung cancer	0.0011	Oocyte meiosis	0.1040
19	Adherens junction	0.0318	Endocytosis	0.0036	Systemic lupus erythematosus	0.1441
20	Endocytosis	0.0356	Adherens junction	0.0052	Neurotrophin signaling pathway	0.1697
21	Renal cell carcinoma	0.0502	Melanoma	0.0072	Shigellosis	0.1697
22	Axon guidance	0.0564	Bacterial invasion of epithelial cells	0.0081	Fc epsilon RI signaling pathway	0.1697
23	Wnt signaling pathway	0.0564	Wnt signaling pathway	0.0128	Pancreatic cancer	0.2083

The horizontal lines show the 1% significance threshold. The target pathway *Acute myeloid leukemia* is highlighted in green.

TABLE V: Ranking and significance of the target pathway for Alzheimer's disease and acute myeloid leukemia (AML). The first and second columns show the disease and the pathway analysis methods. The next 6 columns show the ranking of the target pathways for 6 meta-analysis combined with the 4 pathway analysis methods. Each row shows the result of the 6 meta-analysis methods combined with the same pathway analysis method. Each cell shows the ranking of the target pathways. The Y(es) or N(o) letters next to the ranking denote if the target pathway is significant or not. Cells highlighted in green are those that are significant and have the best rankings in their row. The last column shows the result of MetaPath. For both diseases, and for all the 4 pathway analysis methods, the target pathway is significant and is ranked the highest when using DANUBE. The target pathway is not significant for AML data when the GSEA p-values are combined with any of the 6 meta-analysis methods.

	Meta-analysis Pathway analysis	Stouffer's method	Z-method	Brown's method	Fisher's method	Additive method	**DANUBE**	MetaPath
Alzheimer's	GSEA	4 (Y)	4 (Y)	4 (Y)	4 (Y)	3 (Y)	3 (Y)	7 (N)
	GSA	11 (Y)	11 (N)	15 (N)	14 (N)	6 (Y)	2 (Y)	
	SPIA	2 (Y)	2 (Y)	3 (Y)	3 (Y)	2 (Y)	2 (Y)	
	PADOG	21 (N)	21 (N)	31 (N)	23 (N)	21 (N)	6 (Y)	
AML	GSEA	1 (N)	1 (N)	4 (N)	4 (N)	1 (N)	1 (N)	4 (N)
	GSA	13 (Y)	13 (N)	16 (N)	14 (N)	2 (Y)	1 (Y)	
	SPIA	4 (Y)	4 (N)	6 (N)	6 (Y)	2 (Y)	2 (Y)	
	PADOG	7 (Y)	7 (N)	10 (N)	7 (Y)	7 (Y)	1 (Y)	

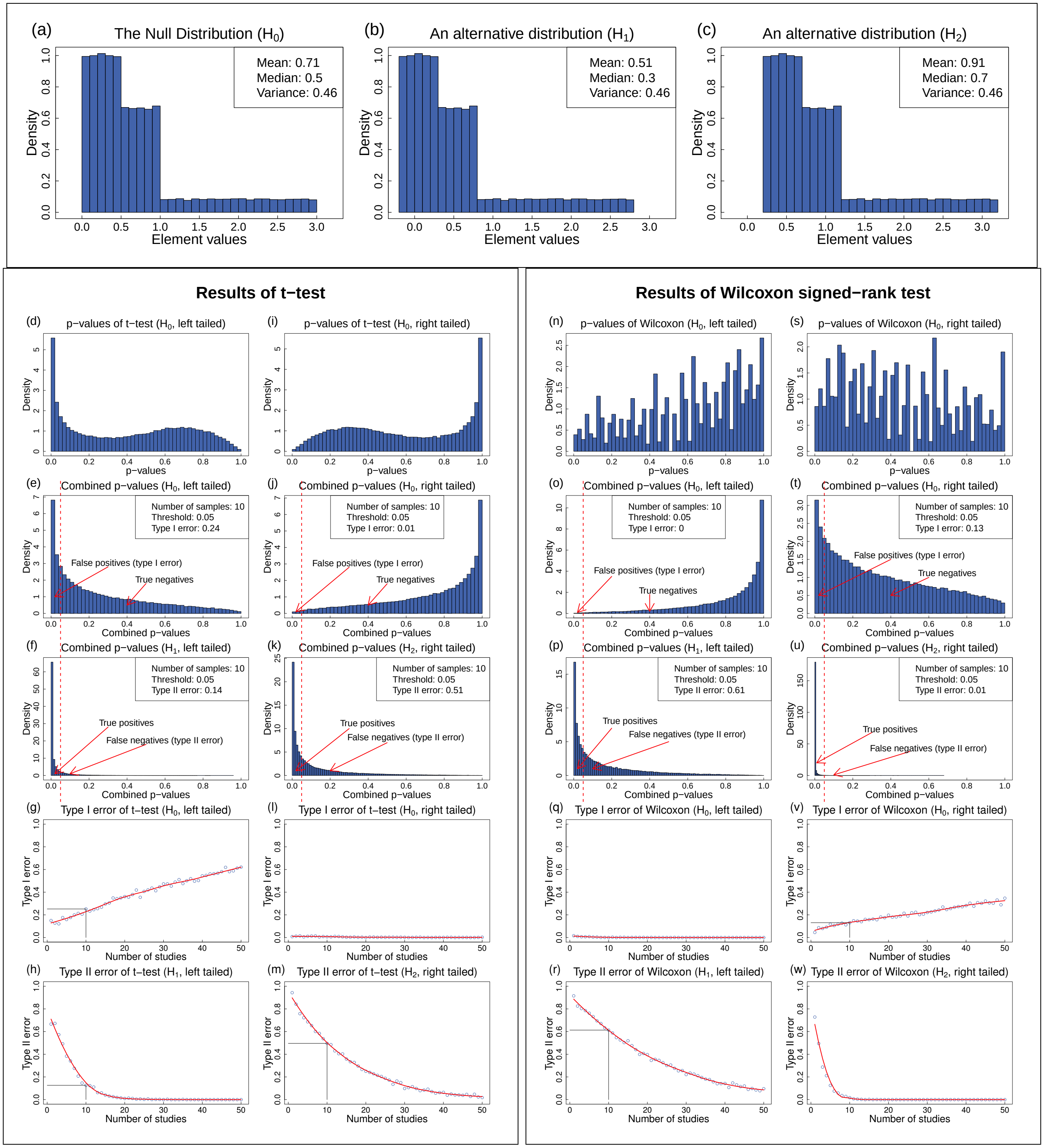


Fig. 5: Type I and Type II errors of the classical meta-analysis using one sample t-test and Wilcoxon signed-ranked test. Panel (a) displays the probability distribution under the null hypothesis H_0. Panel (b) displays an alternative distribution H_1 which has the same shape as the null distribution with a slightly smaller median. Panel (c) displays another alternative distribution H_2 which has the same shape as the null distribution with a slightly larger median. Panels (d–h) display the results using left-tailed t-tests. Panel (d) displays the distribution of p-values using left-tailed t-test for samples drawn from the null distribution H_0. Panel (e) displays the 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). Panel (f) displays the distribution of combined p-values using a 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). Panel (g) displays the type I error with varying number of studies. Panel (h) displays the type II error with varying number of studies using a left-tailed t-test for samples drawn from the alternative distribution H_1. Similarly, panels (i–m) display the results using right-tailed t-test; panels (n–r) display the results of left-tailed Wilcoxon signed-rank test; panels (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,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,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.

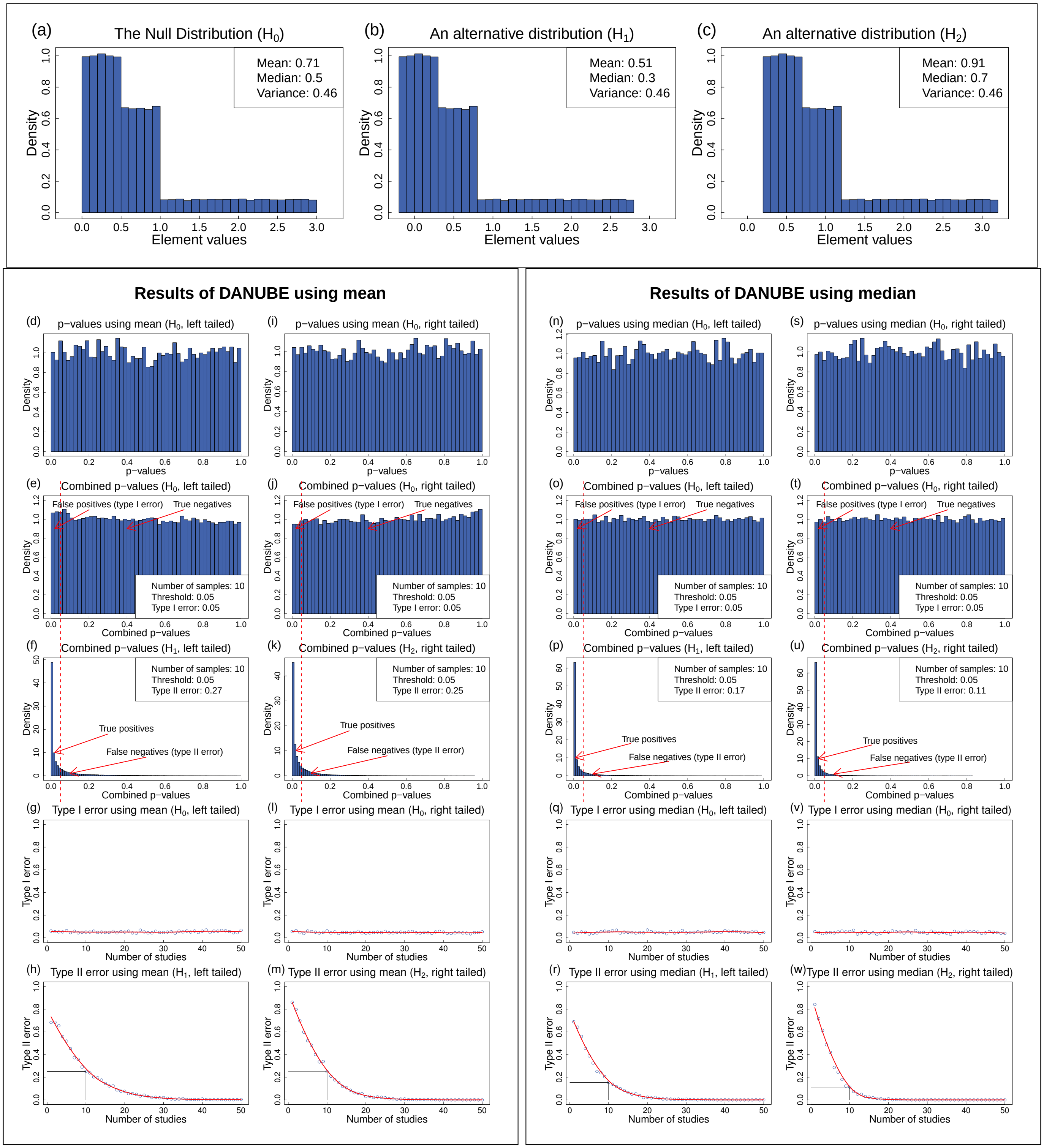


Fig. 6: Type I and type II errors of DANUBE using mean and median as discriminative statistics. Panel (a) displays the probability distribution under the null hypothesis (H_0). Panel (b) displays an alternative distribution (H_1), which has the same shape as the null distribution but a slightly smaller median. Panel (c) displays an alternative distribution (H_2) which has the same shape as the null distribution but a slightly larger median. Panels (d–h) display the results of the left-tailed DANUBE using mean; panels (i–m) display the results of the right-tailed DANUBE using mean; panels (n–r) display the results of left-tailed DANUBE using median; panels (s–w) display the results of right-tailed DANUBE using median. Panels (d, i, n, 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, 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, v). Panels (h, m, r, w) show that the type II error converges quickly to zero. Combining 10 studies, the type II errors 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.

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 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, 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. Figure 6 displays the results using DANUBE. Panels (a, b, c) show the null distribution H_0 and two alternative distributions H_1 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, h) display the type I and type II error 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 (l, 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. Conclusions

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 non-normality 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 7 Alzheimer's and 9 acute myeloid leukemia datasets using 25 approaches: 6 meta-analysis methods, Stouffer's, Z-method, Brown's, Fisher's, the additive method and DANUBE, each of them combined with 4 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

This research was supported in part by the following grants: NIH R01 DK089167, R42 GM087013 and NSF DBI-0965741, and by the Robert J. Sokol Endowment in Systems Biology. 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. We thank Diana Diaz for help and useful discussion.

References

[1] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "NCBI GEO: archive for functional genomics data sets–update," *Nucleic Acids Research*, vol. 41, no. D1, pp. D991–D995, 2013.

[2] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 2002. [Online]. Available: http://nar.oxfordjournals.org/cgi/content/abstract/30/1/207

[3] G. Rustici, N. Kolesnikov, M. Brandizi, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, J. Ison, M. Keays, N. Kurbatova, J. Malone, R. Mani, A. Mupo, R. P. Pereira, E. Pilicheva, J. Rung, A. Sharma, Y. A. Tang, T. Ternent, A. Tikhonov, D. Welter, E. Williams, A. Brazma, H. Parkinson, and U. Sarkans, "ArrayExpress update–trends in database growth and links to data analysis tools," *Nucleic Acids Research*, vol. 41, no. D1, pp. D987–D990, 2013.

[4] A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, A. Oezcimen, P. Rocca-Serra, and S.-A. Sansone, "ArrayExpress–a public repository for microarray gene expression data at the EBI," *Nucleic Acids Research*, vol. 31, no. 1, pp. 68–71, 2003.

[5] G. C. Tseng, D. Ghosh, and E. Feingold, "Comprehensive literature review and statistical considerations for microarray meta-analysis," *Nucleic Acids Research*, vol. 40, no. 9, pp. 3785–3799, 2012.

[6] A. Ramasamy, A. Mondry, C. C. Holmes, and D. G. Altman, "Key issues in conducting a meta-analysis of gene expression microarray datasets," *PLoS Medicine*, vol. 5, no. 9, p. e184, 2008.

[7] T. Manoli, N. Gretz, H.-J. Gröne, M. Kenzelmann, R. Eils, and B. Brors, "Group testing for pathway analysis improves comparability of different microarray datasets," *Bioinformatics*, vol. 22, no. 20, pp. 2500–2506, 2006.

[8] F. Borovecki, L. Lovrecic, J. Zhou, H. Jeong, F. Then, H. Rosas, S. Hersch, P. Hogarth, B. Bouzou, R. Jensen, and D. Krainc, "Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 31, pp. 11 023–11 028, 2005.

[9] L. Friedman, "Why vote-count reviews don't count," *Biological Psychiatry*, vol. 49, no. 2, pp. 161–162, 2001.

[10] L. V. Hedges and I. Olkin, "Vote-counting methods in research synthesis," *Psychological Bulletin*, vol. 88, no. 2, p. 359, 1980.

[11] K. Shen and G. C. Tseng, "Meta-analysis for pathway enrichment analysis when combining multiple genomic studies," *Bioinformatics*, vol. 26, no. 10, pp. 1316–1323, 2010.

[12] S. R. Setlur, T. E. Royce, A. Sboner, J.-M. Mosquera, F. Demichelis, M. D. Hofer, K. D. Mertz, M. Gerstein, and M. A. Rubin, "Integrative microarray analysis of pathways dysregulated in metastatic prostate cancer," *Cancer Research*, vol. 67, no. 21, pp. 10 296–10 303, 2007.

[13] D. R. Rhodes, T. R. Barrette, M. A. Rubin, D. Ghosh, and A. M. Chinnaiyan, "Meta-analysis of microarrays interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer," *Cancer Research*, vol. 62, no. 15, pp. 4427–4433, 2002.

[14] A. Kaever, M. Landesfeind, K. Feussner, B. Morgenstern, I. Feussner, and P. Meinicke, "Meta-analysis of pathway enrichment: combining independent and dependent omics data sets," *PloS One*, vol. 9, no. 2, p. e89297, 2014.

[15] C. Mitrea, Z. Taghavi, B. Bokanizad, S. Hanoudi, R. Tagett, M. Donato, C. Voichiţa, and S. Drăghici, "Methods and approaches in the topology-based analysis of biological pathways," *Frontiers in Physiology*, vol. 4, p. 278, 2013.

[16] P. Khatri, M. Sirota, and A. J. Butte, "Ten years of pathway analysis: current approaches and outstanding challenges," *PLoS Computational Biology*, vol. 8, no. 2, p. e1002375, 2012.

[17] E. Kotelnikova, M. A. Shkrob, M. A. Pyatnitskiy, A. Ferlini, and N. Daraselia, "Novel approach to meta-analysis of microarray datasets reveals muscle remodeling-related drug targets and biomarkers in Duchenne muscular dystrophy," *PLoS Computational Biology*, vol. 8, no. 2, p. e1002365, 2012.

[18] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Research*, vol. 37, no. 1, pp. 1–13, 2009.

[19] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, January 2000.

[20] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 27, no. 1, pp. 29–34, 1999.

[21] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, and P. D'Eustachio, "The Reactome pathway knowledgebase," *Nucleic Acids Research*, vol. 42, no. D1, pp. D472–D477, 2014.

[22] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, 2011.

[23] T. M. Loughin, "A systematic comparison of methods for combining p-values from independent tests," *Computational Statistics & Data Analysis*, vol. 47, no. 3, pp. 467–485, 2004.

[24] R. A. Fisher, *Statistical methods for research workers*. Edinburgh: Oliver & Boyd, 1925.

[25] E. S. Edgington, "An additive method for combining probability values from independent experiments," *The Journal of Psychology*, vol. 80, no. 2, pp. 351–363, 1972.

[26] 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.

[27] J. O. Irwin, "On the frequency distribution of the means of samples from a population having any law of frequency with finite moments, with special reference to Pearson's Type II," *Biometrika*, vol. 19, no. 3-4, pp. 225–239, 1927.

[28] L. H. C. Tippett, *The methods of statistics*. London: Williams & Norgate, 1931.

[29] B. Wilkinson, "A statistical consideration in psychological research." *Psychological Bulletin*, vol. 48, no. 2, p. 156, 1951.

[30] J. Li and G. C. Tseng, "An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies," *The Annals of Applied Statistics*, vol. 5, no. 2A, pp. 994–1019, 2011.

[31] H. Choi, R. Shen, A. M. Chinnaiyan, and D. Ghosh, "A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments," *BMC Bioinformatics*, vol. 8, no. 1, p. 364, 2007.

[32] R. Shen, D. Ghosh, and A. M. Chinnaiyan, "Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data," *BMC Genomics*, vol. 5, no. 1, p. 94, 2004.

[33] S. J. Barton, S. R. Crozier, K. A. Lillycrop, K. M. Godfrey, and H. M. Inskip, "Correction of unexpected distributions of P values from analysis of whole genome arrays by rectifying violation of statistical assumptions," *BMC Genomics*, vol. 14, no. 1, p. 161, 2013.

[34] A. A. Fodor, T. L. Tickle, and C. Richardson, "Towards the uniform distribution of null P values on Affymetrix microarrays," *Genome Biology*, vol. 8, no. 5, p. R69, 2007.

[35] M. Bland, "Do baseline p-values follow a uniform distribution in randomised trials?" *PloS One*, vol. 8, no. 10, p. e76010, 2013.

[36] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceeding of The National Academy of Sciences of the Unites States of America*, vol. 102, no. 43, pp. 15 545–15 550,

2005.

[37] B. Efron and R. Tibshirani, “On testing the significance of sets of genes,” *The Annals of Applied Statistics*, vol. 1, no. 1, pp. 107–129, 2007.

[38] A. L. Tarca, S. Drăghici, G. Bhatti, and R. Romero, “Down-weighting overlapping genes improves gene set analysis,” *BMC Bioinformatics*, vol. 13, no. 1, p. 136, 2012.

[39] V. K. Mootha, C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstråle, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop, “PGC-1$1\alpha$-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes,” *Nature Genetics*, vol. 34, no. 3, pp. 267–273, Jul 2003.

[40] P. Khatri, S. Drăghici, G. C. Ostermeier, and S. A. Krawetz, “Profiling gene expression using Onto-Express,” *Genomics*, vol. 79, no. 2, pp. 266–270, 2002.

[41] S. Drăghici, P. Khatri, R. P. Martins, G. C. Ostermeier, and S. A. Krawetz, “Global functional profiling of gene expression,” *Genomics*, vol. 81, no. 2, pp. 98–104, 2003.

[42] T. Beißbarth and T. P. Speed, “GOstat: find statistically overrepresented Gene Ontologies within a group of genes.” *Bioinformatics*, vol. 20, pp. 1464–1465, June 2004.

[43] A. L. Tarca, S. Drăghici, P. Khatri, S. S. Hassan, P. Mittal, J.-s. Kim, C. J. Kim, J. P. Kusanovic, and R. Romero, “A novel signaling pathway impact analysis,” *Bioinformatics*, vol. 25, no. 1, pp. 75–82, 2009.

[44] C. Voichita and S. Draghici, *ROntoTools: R Onto-Tools suite*, 2013, R package. [Online]. Available: http://www.bioconductor.org

[45] S. Drăghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichiţa, C. Georgescu, and R. Romero, “A systems biology approach for pathway level analysis,” *Genome Research*, vol. 17, no. 10, pp. 1537–1545, 2007.

[46] J. D. Storey and R. Tibshirani, “Statistical significance for genomewide studies,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9440–9445, 2003.

[47] E. S. Edgington, “A normal curve method for combining probability values from independent experiments,” *The Journal of Psychology*, vol. 82, no. 1, pp. 85–89, 1972.

[48] S. Stouffer, E. Suchman, L. DeVinney, S. Star, and J. Williams, RM, *The American Soldier: Adjustment during army life*. Princeton: Princeton University Press, 1949, vol. 1.

[49] M. B. Brown, “A method for combining non-independent, one-sided tests of significance,” *Biometrics*, pp. 987–992, 1975.

[50] X. Wang, D. D. Kang, K. Shen, C. Song, S. Lu, L.-C. Chang, S. G. Liao, Z. Huo, S. Tang, Y. Ding, N. Kaminski, E. Sibille, Y. Lin, J. Li, and G. C. Tseng, “An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection,” *Bioinformatics*, vol. 28, no. 19, pp. 2534–2536, 2012.

[51] R. H. Swerdlow, “Brain aging, Alzheimer’s disease, and mitochondria,” *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1812, no. 12, pp. 1630–1639, 2011.

[52] A. Maruszak and C. Żekanowski, “Mitochondrial dysfunction and Alzheimer’s disease,” *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 35, no. 2, pp. 320–330, 2011.

[53] X. Zhu, G. Perry, M. A. Smith, and X. Wang, “Abnormal mitochondrial dynamics in the pathogenesis of Alzheimer’s disease,” *Journal of Alzheimer’s Disease*, vol. 33, pp. S253–S262, 2013.

[54] H. W. Querfurth and F. M. LaFerla, “Mechanisms of disease,” *New England Journal of Medicine*, vol. 362, no. 4, pp. 329–344, 2010.

[55] M. Donato, Z. Xu, A. Tomoiaga, J. G. Granneman, R. G. MacKenzie, R. Bao, N. G. Than, P. H. Westfall, R. Romero, and S. Drăghici, “Analysis and correction of crosstalk effects in pathway analysis,” *Genome Research*, vol. 23, no. 11, pp. 1885–1893, 2013.

[56] P. S. Brookes, Y. Yoon, J. L. Robotham, M. Anders, and S.-S. Sheu, “Calcium, ATP, and ROS: a mitochondrial love-hate triangle,” *American Journal of Physiology-Cell Physiology*, vol. 287, no. 4, pp. C817–C833, 2004.

[57] W. S. Gosset, “The Probable Error of a Mean,” *Biometrika*, vol. 6, pp. 1–25, 1908.

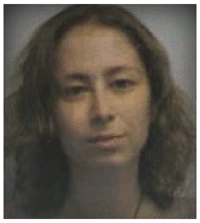
[58] E. Peaeson and H. Haetlet, “Biometrika tables for statisticians,” *Biometrika Trust*, 1976.

[59] F. Wilcoxon, “Individual comparisons by ranking methods,” *Biometrics*, vol. 1, no. 6, pp. 80–83, 1945.

[60] F. Wilcoxon, S. Katti, and R. A. Wilcox, “Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test,” *Selected tables in mathematical statistics*, vol. 1, pp. 171–259, 1970.

[61] M. Hollander, D. A. Wolfe, and E. Chicken, *Nonparametric statistical methods*. John Wiley & Sons, 2013, vol. 751.

Tin Nguyen received the BSc and MSc degrees in computer science from Eotvos Lorand University in Budapest, Hungary. He is currently is a PhD Candidate and a member of the Intelligent Systems and Bioinformatics Laboratory (ISBL) in the Department of Computer Science at Wayne State University, Michigan. 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 is a PhD Candidate and a member of the Intelligent Systems and Bioinformatics Laboratory (ISBL) in the Department of Computer Science at Wayne State University. In 2012, she received the Master of Science in Computer Science from 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. She is also a student member of IEEE and ACM.

Rebecca Tagett has a Bachelors in Physics, a Masters in Molecular Biology, and 10 years R&D experience in industry as a Computational Biologist. A PhD Candidate and a member of the Intelligent Systems and Bioinformatics Laboratory (ISBL) in the Department of Computer Science at Wayne State University, her research focuses on phenotypic prediction using multi-omics. Her interests are Functional Genomics, Scientific Writing, Bioinformatics and Biostatistics. She is a member of the International Society for Computational Biology (ISCB).

Sorin Draghici is the Associate Dean for Innovation and Entrepreneurship, and Director, James and Patricia Anderson Engineering Ventures Institute in the College of Engineering at Wayne State University. He currently holds the Robert J. Sokol, MD Endowed Chair in Systems Biology, as well as appointments as full professor in the Department of Computer Science and the Department of Obstetrics and Gynecology, Wayne State University. Professor Draghici 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, 8 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 8 analysis tools in this area, tools that have been made available over the web for over 10 years to over 11,000 scientists from 5 continents. He has also co-authored 3 analysis packages in Bioconductor. His top 4 papers in this area have over 2,000 total citations, while this entire work gathered over 7,000 citations. During his 17 year appointments as faculty, he was able to attract $8,262,283 as PI and $27,418,291 as co-PI in NIH and NSF grants.