A novel bi-level meta-analysis approach - applied to biological pathway analysis

Supplementary Material

Tin Nguyen[1], Rebecca Tagett[1], Michele Donato[1], Cristina Mitrea[1], and Sorin Draghici[1,2,*]

[1]**Department of Computer Science, Wayne State University, Detroit, Michigan, USA.**
[2]**Department of Obstetrics and Gynecology, Wayne State University, Detroit, Michigan, USA.**

1 Introduction

The supplementary material contains additional details and figures which were not included in the main text due to space limitations.

2 The additive and the add-CLT method

The additive method [6–8, 11, 12] uses the sum of the p-values as the test statistic, instead of the log product. This way of combining p-values is rarely used, but we show it to be robust and powerful in our data analysis. Let us denote the p-values resulting from the m independent significance tests as P_1, P_2, ..., P_m. These p-values are independent and uniformly distributed between zero and one under the null. Denote the sum of these p-values, $X = \sum_{i=1}^{m} P_i$ ($X \in [0, m]$), as the new random variable. X is known to follow the Irwin-Hall distribution [11, 12] with the following probability density function (pdf) $f(x)$, and cumulative distribution function (cdf) $F(x)$ as follows:

$$\begin{aligned} f(x) &= \frac{1}{(m-1)!}\sum_{i=0}^{\lfloor x \rfloor}(-1)^i \binom{m}{i}(x-i)^{m-1} \\ F(x) &= \frac{1}{m!}\sum_{i=0}^{\lfloor x \rfloor}(-1)^i \binom{m}{i}(x-i)^{m} \end{aligned} \tag{1}$$

The above formulae were derived by Lagrange as described in Feller's book [8], but a similar formulation was also reported in [6, 11, 12]. The probability density function of the Irwin-Hall distribution, for different numbers of studies, m, is displayed in the left panel of Figure S1A. For each value of m, X takes values in the interval $[0, m]$ and the distributions are symmetrical with mean $\frac{m}{2}$. Using the above cdf, the additive method calculates the probability of observing the sum of individual p-values.

Unlike Fisher's method, the additive method is not sensitive to extremely small individual p-values. However, we note that the additive method faces a different practical problem. For large values of m, Equation (1) involves some intensive computation due to a sum of combinatorial and division by a factorial, the result of which can lead to an "arithmetic underflow". In other words, the result can be a number smaller than what a computer can actually store in memory.

To demonstrate this, we calculate the values of the cdf when $X = m$ for each value of $m \in [1..100]$. For a given value of m, the variable X takes values in the interval $[0, m]$, and $F(X = m)$ is the area under the curve of the pdf. Since the area under a pdf is 1, $F(X = m)$ should be 1 for each value of m, and therefore the log absolute value of $F(X = m)$ should be 0. However, this does not hold in practice. The right panel in Figure S1A displays the log absolute values of $F(X = m)$ calculated by a 64-bit implementation of R (version 3.1.1, 2014-Jul-10). Due to underflow, the calculation is accurate only up to around $m = 30$, after which the log absolute value of $F(X = m)$ departs linearly from 0. Therefore, the additive method can be inaccurate, not due to the mathematical formulation of the Irwin-Hall distribution, but rather due to computational limitations.

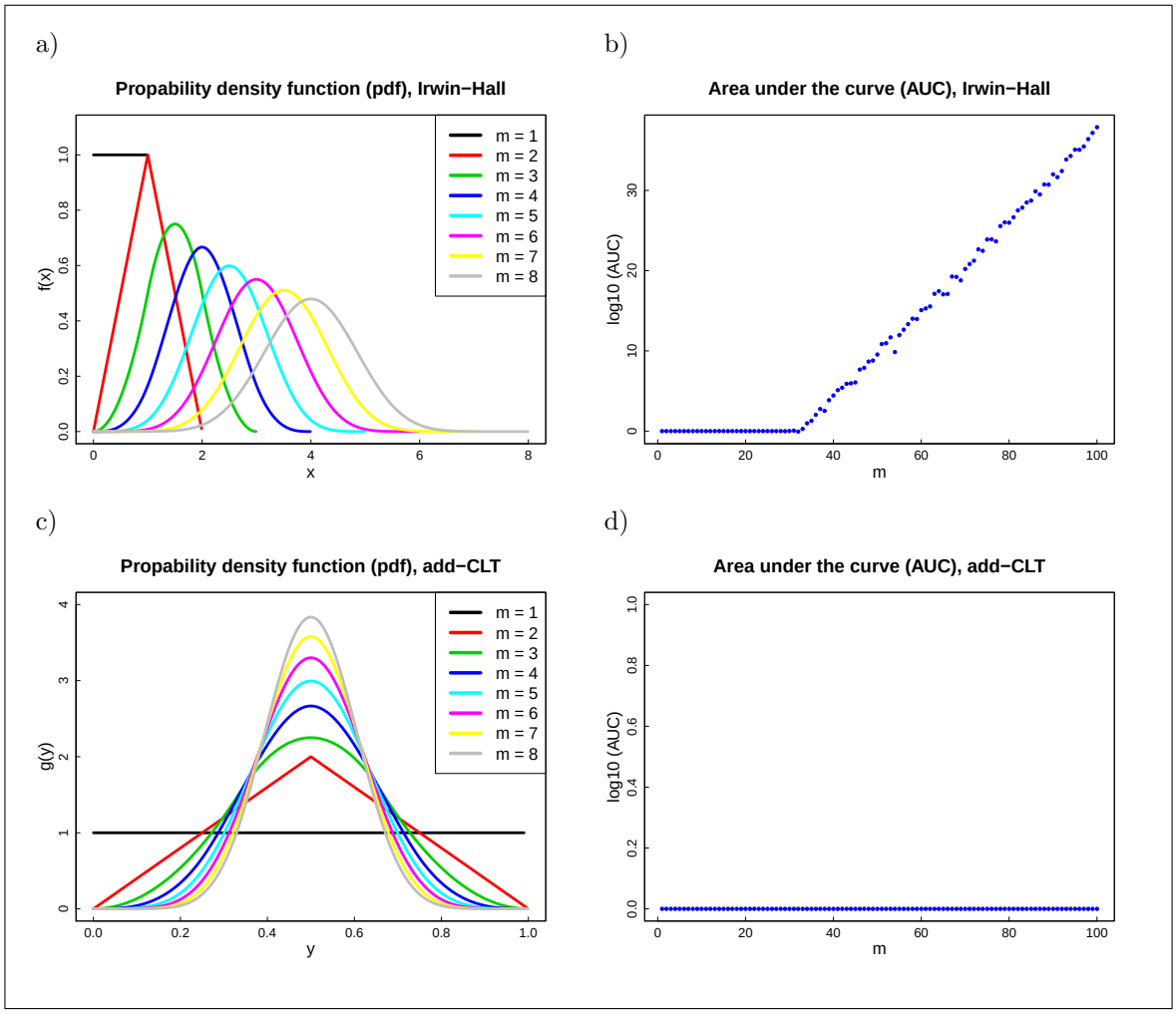

Figure S1: Probability density functions (pdf) and the area under the pdf curve (AUC) of the Irwin-Hall and add-CLT distributions. In (A), the left panel displays the pdf of the Irwin-Hall distribution for different values of m. The horizontal axis displays the values of the random variable $X \in [0, m]$ while the vertical axis displays the density of X. The right panel shows the AUC calculated by a 64-bit implementation of R (version 3.1.1, 2014-Jul-10). The horizontal axis shows increasing values of m while the vertical axis shows the log absolute value of $F(X = m)$, i.e. the area under the entire pdf curve. For each value of m, $F(X = m)$ should be 1 and therefore the log absolute value of $F(X = m)$ should be 0. However, due to the complexity of the Irwin-Hall formula and arithmetic underflow, the calculation is completely unreliable when $m > 30$. Similarly, the left panel in (B) displays the pdf for different values of m. The horizontal axis shows the random variable $Y \in [0, 1]$ while the vertical axis shows the density of Y. For $m < 20$, we use the linear transformation of the Irwin-Hall distribution to calculate the pdf of Y. For $m \geq 20$, we use the normal distribution with mean $\mu = \frac{1}{2}$ and variance $\sigma^2 = \frac{1}{12m}$ to calculate the pdf of Y (Central Limit Theorem). The right panel shows the log absolute value of $G(Y = 1)$, i.e. the area under the entire pdf curve. For each value of m, $G(Y = 1)$ should be 1 and therefore the log absolute value of $G(Y = 1)$ should be 0. The right panel shows that the add-CLT method overcomes the computational problem presented in the classical additive method using Irwin-Hall distribution.

Here we describe an enhancement to the additive method that makes it more reliable for larger values of m. First, we change the random variable from the sum of the p-values to the average of the p-values. Second, when m is large, we replace the additive method with the Central Limit Theorem (CLT). The reason for the modification is that the additive method is accurate for small values of m, while the CLT is more accurate for large values of m. We select $m = 20$ as a conservative cut-off. In other words, we will use the additive method when $m < 20$, and the CLT when $m \geq 20$. To show the validity of using the CLT for large m, we define a random variable $Y = \frac{\sum_{i=1}^{m} P_i}{m}$ ($Y \in [0, 1]$) as the average of p-values. Since $Y = \frac{X}{m}$, we can derive the probability density function (pdf) and the cumulative distribution function (cdf) using a linear transformation of X as follows:

$$g(y) = \frac{m}{(m-1)!} \sum_{i=0}^{\lfloor m \cdot y \rfloor} (-1)^i \binom{m}{i} (m \cdot y - i)^{m-1} \tag{2}$$

$$G(y) = \frac{1}{m!} \sum_{i=0}^{\lfloor m \cdot y \rfloor} (-1)^i \binom{m}{i} (m \cdot y - i)^{m} \tag{3}$$

The variable Y is the mean of m independent and identically distributed (i.i.d.) random variables (the p-values from each individual experiment), that follow a uniform distribution with a mean of $\frac{1}{2}$ and a variance of $\frac{1}{12}$. From the Central Limit Theorem [13], the average of such m i.i.d. variables follows a normal distribution with mean $\mu = \frac{1}{2}$ and variance $\sigma^2 = \frac{1}{12m}$, i.e. $Y \sim \mathcal{N}(\frac{1}{2}, \frac{1}{12m})$ for sufficiently large values of m.

The pdf of Y for different m values and the corresponding AUCs are displayed in Figure S1B. The left panel displays the distribution used in add-CLT method while the right panel displays the area under the pdf curve. We can see that the add-CLT overcomes the computational problem presented in the classical additive method using Irwin-Hall distribution. We refer to our proposed combination of the Irwin-Hall distribution and the Central Limit Theorem as "add-CLT", for "additive-Central Limit Theorem", in order to distinguish it from the classical additive method. As noted above, the transition from the additive method to the Central Limit Theorem takes place at the $m \geq 20$ threshold.

3 Experimental studies using gene expression data

3.1 Type II diabetes data

The diabetes datasets we use in our data analysis were obtained from Gene Expression Omnibus (GEO) with IDs: GSE25462 (skeletal muscle, 10 cases and 15 controls), GSE19420 (skeletal muscle, 10 cases and 12 controls), GSE18732 (skeletal muscle, 45 cases and 47 controls), GSE23343 (liver biopsy, 10 cases and 7 controls), and GSE22309 (skeletal muscle, 30 cases and 40 controls).

The platform for GSE22309 is the Affymetrix Human Genome U95A array while the platform for the other datasets is the Affymetrix Human Genome U133 Plus 2.0 array. Affymetrix *CEL* files containing raw expression data were downloaded from GEO for each dataset and processed using *R* and *Bioconductor 2.13*. Quality control was performed using the *qc* method from the package *simpleaffy 2.38.0* [14]. Pre-processing was performed on individual datasets using the *threestep* function from the package *affyPLM version 1.38.0* [3–5]. The parameters used for the *threestep* function are: robust multi-array analysis (RMA) background adjustment, quantile normalization, and median polish summarization.

We use Gene Set Enrichment Analysis (GSEA) [16] to analyze the 5 datasets individually, before performing the meta-analysis. Figure S2 illustrates the rankings and FDR-corrected p-values of the target pathway *Type II diabetes mellitus* for the 5 datasets. The left panel shows the FDR-corrected p-values while the right panel shows the rankings of the target pathway. The horizontal axes show the 5 diabetes datasets. The vertical axis in the left panel shows the FDR-corrected p-values of the target pathway. The vertical axis in the right panel shows the rankings of the target pathway. For GSEA, the adjusted p-values (black squares) of the target pathway are not significant and greatly vary between 0.5 and 1 while the rankings vary between 4 and 133. This is a clear case in which the correct pathway is missed in every single one of the 5 individual datasets available; such a situation calls for meta-analysis.

In the same figure, the blue triangles show the results of the intra-experiment analysis (combined with GSEA). This method returns similar adjusted p-values (compared to GSEA) for GSE25462 and GSE23343 but offers a significant improvement in p-values for the other three datasets. In addition, the rankings of the target pathway are also more consistent (between 1 and 20). However, for both GSEA and intra-experiment

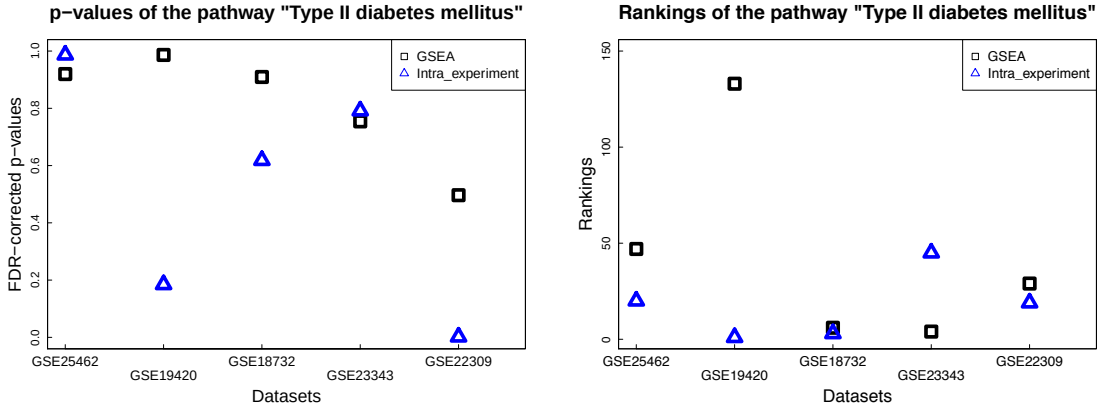

Figure S2: P-values (left panel) and rankings (right panel) of the KEGG target pathway, *Type II diabetes mellitus*, for 5 diabetes datasets, using Gene Set Enrichment Analysis (GSEA) and the intra-experiment analysis. The horizontal axes show the 5 diabetes datasets. The vertical axis in the left panel shows the FDR-corrected p-values of the target pathway. The vertical axis in the right panel shows the rankings of the target pathway. For GSEA, the adjusted p-values (black squares) of the target pathway are not significant and greatly vary between 0.5 and 1 while the rankings vary between 4 and 133. The intra-experiment analysis (blue triangles) returns similar adjusted p-values for GSE25462 and GSE23343 but offers a significant improvement in p-values for the other three datasets. In addition, the rankings of the target pathway are also more consistent (between 1 and 20). However, for both methods, the inconsistency in p-values and rankings across datasets makes biological interpretation quite difficult, showing the need for the second level of analysis.

analysis methods, the inconsistency in p-values and rankings across datasets makes biological interpretation quite difficult. Therefore, an added level of meta-analysis is needed for both methods.

3.2 Acute myeloid leukemia data

The following acute myeloid leukemie (AML) datasets from GEO were used for our analysis: GSE14924 (CD4 T cells, 10 cases and 9 controls, and CD8 T cells, 10 cases and 11 controls), GSE17054 (hematopoietic stem cells, 5 cases and 4 controls), GSE12662 (fractionated bone marrow: CD34+ cells, promyelocytes, neutrophils and the PR9 cell line, 75 cases and 24 controls), GSE57194 (primary CD34+ cells, 6 cases and 6 controls), GSE33223 (peripheral blood mononuclear cells, 20 cases and 10 controls), GSE42140 (peripheral blood mononuclear cells, 26 cases and 5 controls), GSE8023 (CD34+ cells from cord blood, 9 cases and 3 controls), and GSE15061 (bone marrow, 201 cases and 68 controls). The platform for all of the AML datasets is the Affymetrix Human Genome U133 Plus 2.0 array. Affymetrix *CEL* files containing raw expression data were downloaded from GEO for each dataset and processed the same way as the diabetes datasets.

Figure S3 shows the p-values (left panel) and rankings (right panel) of the KEGG target pathway, *Acute myeloid leukemia*, for 9 acute myeloid leukemia (AML) datasets, using Gene Set Enrichment Analysis (GSEA) and the intra-experiment analysis. The horizontal axes show the 9 AML datasets. The vertical axis in the left panel shows the FDR-corrected p-values of the target pathway. The vertical axis in the right panel shows the rankings of the target pathway. For GSEA (black squares), the adjusted p-values of the target pathway are not significant and vary greatly between 0.23 and 1 while the rankings vary between 12 and 114. The intra-experiment analysis (blue triangles) returns a similar adjusted p-values (compared to GSEA) for GSE14924_CD8, GSE17054, GSE57194, and GSE8023 but offers a significant improvement in p-values for the other five datasets. The rankings are also more consistent (between 10 and 58). However, for both methods, the inconsistency in p-values and rankings across datasets makes biological interpretation quite difficult. In both cases, another level of meta-analysis is necessary.

3.3 Alzheimer's disease

The Alzheimer's datasets we use in our data analysis were obtained from Gene Expression Omnibus (GEO) with IDs: GSE1297 (hippocampus, 22 cases and 9 controls), GSE28146 (hippocampus, 22 cases and 8 controls) and GSE5281 (a mixture of entorhinal cortex, hippocampus, medial temporal gyrus, posterior cingulate,

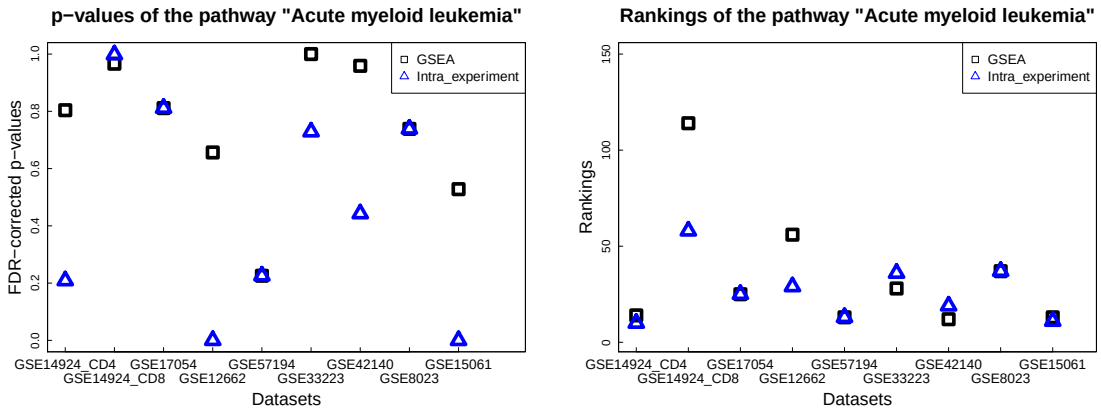

Figure S3: P-values (left panel) and rankings (right panel) of the KEGG target pathway, *Acute myeloid leukemia*, for 9 acute myeloid leukemia (AML) datasets, using Gene Set Enrichment Analysis (GSEA) and the intra-experiment analysis. The horizontal axes show the 9 AML datasets. The vertical axis in the left panel shows the FDR-corrected p-values of the target pathway. The vertical axis in the right panel shows the rankings of the target pathway. For GSEA, the adjusted p-values (black squares) of the target pathway are not significant and vary greatly between 0.23 and 1 while the rankings vary between 12 and 114. The intra-experiment analysis (blue triangles) returns similar adjusted p-values for GSE14924_CD8, GSE17054, GSE57194, and GSE8023 but offers a significant improvement in p-values for the other five datasets. The rankings are also more consistent (between 10 and 58). However, for both methods, the inconsistency in p-values and rankings across datasets makes biological interpretation quite difficult, showing the need for the second level of analysis.

superior frontal gyrus, and primary visual cortex, 87 cases and 74 controls), GSE16759 (parietal lobe cortex, 4 cases and 4 controls), GSE48350 (a mixture of post central gyrus, superior frontal gyrus, hippocampus, and entorhinal cortex, 80 cases and 173 controls), GSE39420 (brain tissues, 14 cases and 7 controls), and GSE4757 (entorhinal cortex, 10 cases and 10 controls). The platform for GSE1297 is the Affymetrix Human Genome U95A array, for GSE39420 is the Affymetrix Human Gene 1.1 ST Array, while the platform for the other datasets is the Affymetrix Human Genome U133 Plus 2.0 array.

4 Simulation studies

The goal of simulation studies is to demonstrate the advantage of the bi-level meta-analysis in a more general context. First, we compare the classical two-sample t-test [10, 15] against a combination of t-test and the intra-experiment analysis (with add-CLT, by default), and show that the latter has a better true positive rate (Figure 3, main text). Second, we investigate the false positive rate of the bi-level meta-analysis. Third, we investigate the effects of group size setting. Finally, we compare add-CLT with Fisher's method, and show that the add-CLT method is more powerful and more robust to bias.

4.1 Comparison between t-test and intra-experiment analysis

Our bi-level meta-analysis framework is comprised of intra- and inter-experiment analysis. In intra-experiment analysis, we split each experiment into smaller studies and then combine the small studies using add-CLT. In inter-experiment analysis, we combine multiple experiments using add-CLT again. Our reasoning for the first stage is that performing a statistical test on a large experiment is not as powerful as splitting it into smaller studies and then combining them using add-CLT. We demonstrate this using the classical two-sample t-test [10, 15].

In the first simulation, we use normal null (control) and alternative (disease) distributions that have the same variance but different means. Analogous to case-control analyses done in biological experiments, we randomly pick a set of samples from the null distribution and a set of samples from the alternative distribution and then compare the two sets. We compare true positive rates (TPR) using two approaches, the two-sample t-test, and a combination of the t-test and our intra-experiment analysis method. For the intra-experiment

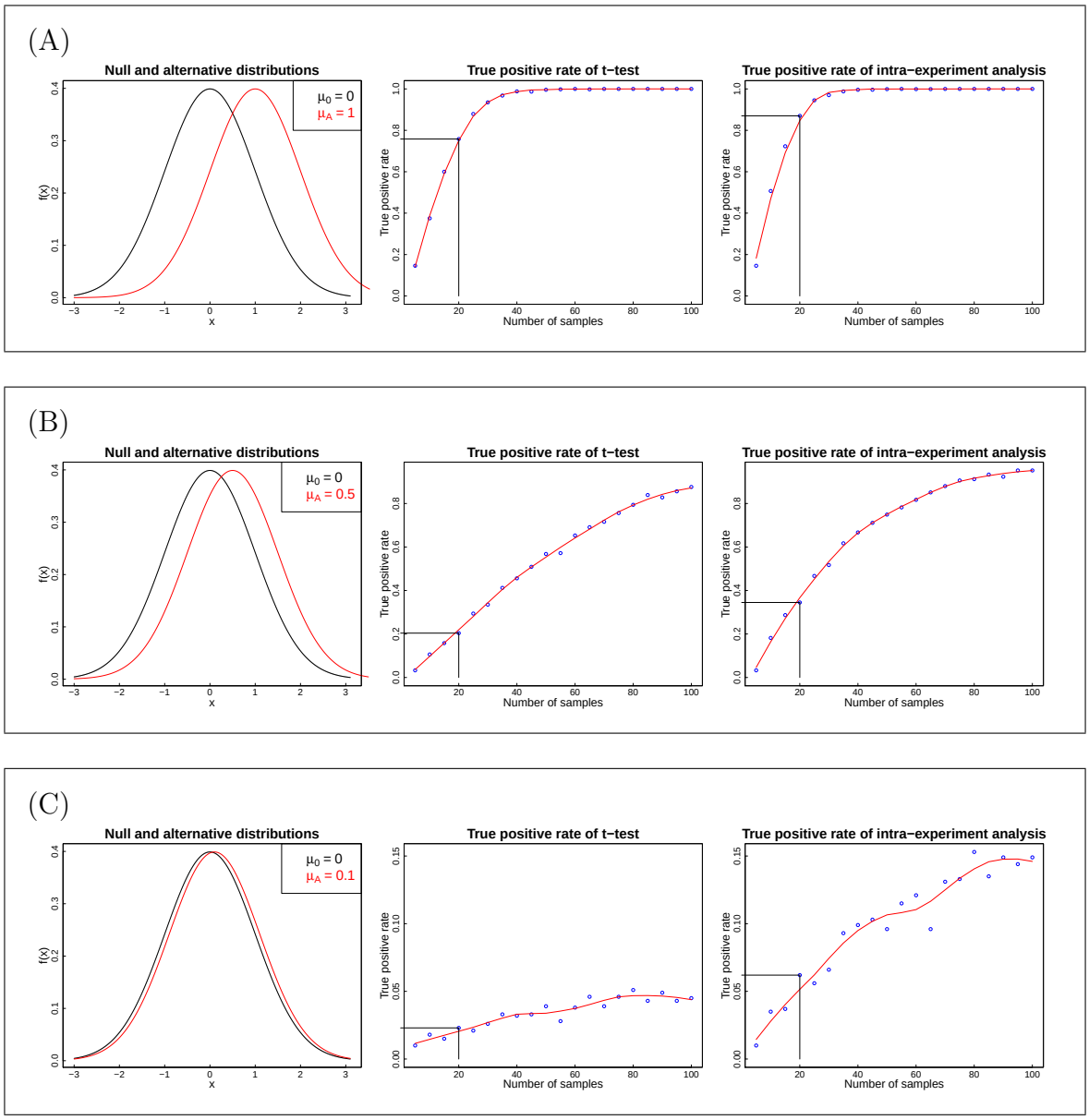

Figure S4: Comparison between the t-test and the intra-experiment analysis method. In each row, the left row panel displays a hypothetical null (control) distribution H_0 (black) and an alternative (disease) distribution H_A (red). The middle and right panels show the true positive rate (TPR) of the t-test and the intra-experiment analysis method as they vary with increasing the number of samples. The horizontal axes of the middle and right panels represent the number of samples while the vertical axes represent the TPR. Given n as the number of samples, we randomly pick n control samples from H_0 and n disease samples from H_A. The first approach is to use a right-tailed t-test to compare the two groups of samples. The second approach is to split the disease group into smaller groups of size 5 and then perform a right-tailed t-test to compare each newly created small disease group against the control group. The resulting p-values are then combined using add-CLT. We repeat the procedure $1,000$ times to get $1,000$ p-values for the t-test and $1,000$ p-values for the intra-experiment analysis method. We then calculate the TPR for each approach as the number of p-values smaller than the threshold 0.01 divided by $1,000$. In (A), when the null and alternative distributions are very different, the intra-experiment analysis method is only slightly better than the t-test in terms of TPR (85% TPR compared to 75% TPR with 20 samples). In (B), when the alternative distribution is closer to the null distribution, the difference in TPR of the two approaches increases (35% TPR compared to 75% TPR with 20 samples). In (C), when the alternative distribution is very close to the null distribution, the difference in TPR between the two approaches is much larger. The TPR of the intra-experiment analysis method is almost three times higher than the TPR of the t-test. In summary, when the two distributions are closer, the intra-experiment analysis method is much more powerful than the t-test and its power increases more rapidly than that of the t-test.

analysis, we split the disease set into smaller sets of size 5 to form multiple small studies. We then perform a right-tailed t-test to compare each newly created small disease set against the full control set. The resulting p-values are then combined using add-CLT. We will show that a combination of the t-test and the intra-experiment analysis method is more powerful than t-test alone (right-tailed as well).

Figure S4 shows comparisons between the t-test and the intra-experiment analysis method for three different alternative distributions. In the left panel of Figure S4A, the null distribution H_0 (black) is a standard normal distribution with mean $\mu_0 = 0$ and variance $\sigma_0^2 = 1$ while the alternative distribution H_A (red) has the same variance but different mean, $\mu_A = 1$. The horizontal axis represents the expression values and the vertical axis represents the density of the expression values. Given n as the number of samples, we pick n control samples from H_0 and n disease samples from H_A. We then use the t-test and the intra-experiment analysis method (combined with t-test) to compare the two sets of samples. Each of the two approaches produces a p-value. We repeat the procedure $1,000$ times to get $1,000$ p-values for the t-test and $1,000$ p-values for the intra-experiment analysis method. We then calculate the true positive rate (TPR) of each method as the number of p-values smaller than the threshold 0.01 divided by $1,000$.

The middle and right panels of Figure S4A display the true positive rate (TPR) of t-test and of the intra-experiment analysis method. The horizontal axes show the number of samples while the vertical axes show the TPR. These panels show that the intra-experiment analysis method has a slightly better TPR than the t-test alone. With 20 samples, the TPR of the intra-experiment analysis method is around 85% whereas the TPR of t-test is close to 75%.

Figure S4B shows the comparison between the two approaches for another alternative distribution with mean $\mu_A = 0.5$. The alternative distribution is closer to the null distribution and thus it is harder for both approaches to identify the true positives. In this case, the intra-experiment analysis method (combined with t-test) has notably higher TPR than the t-test. With 20 samples, the TPR of the intra-experiment analysis method is 35% compared to 20% for the t-test.

Similarly, Figure S4C shows the comparison between the two approaches for another alternative distribution with mean $\mu_A = 0.1$. In this case, the alternative distribution is very close to the null distribution and thus both approaches have much lower TPR than cases (B) and (C). Interestingly, the TPR of the t-test barely increases when the number of samples increases - only from 2% to 5% when the number of samples increases from 20 to 100. Similarly, the TPR of the intra-experiment analysis method is also low but it *increases rapidly* when the number of samples increases. For example, the TPR increases from 6% with 20 samples to 15% with 100 samples. The figure shows that the TPR of the intra-experiment analysis is approximately three times higher that that of t-test.

In summary, the intra-experiment analysis method (combined with the t-test) is more powerful than the t-test alone for comparative analysis. According to the comparison shown in Figure S4, the difference in performance between the two approaches increases when the difference between the null and alternative distributions decreases. Therefore, the intra-experiment analysis approach is especially powerful when there is only a small change in gene expression profile between the two phenotypes compared.

4.2 False positive rate of the bi-level meta-analysis

Since the control samples are used in every test in the intra-experiment meta-analysis, the individual p-values of the split datasets in the intra-experiment analysis are not completely independent. As a result, the combined p-values of the intra-experiment analysis may not be uniformly distributed in certain situations, in particular when the number of control samples is small. However, we will show that this does not have significant impact on the final results of the bi-level meta-analysis.

Given a sound statistical test, the empirical p-values of independent datasets should be uniformly distributed under the null. Let us consider one dataset DS_i, which has an associated empirical p-value p_i (produced by something like a t-test). The intra-experiment analysis produces a set of p-values (one for each split dataset) and then combines these p-values to produce one p-value $\hat{p}_i$. We can consider that $\hat{p}_i$ is a transformation of p_i using the add-CLT function. Under the null hypothesis, the empirical p-values $\{p_i\}$ are symmetrical around 0.5 (because they are uniformly distributed). Since, the add-CLT function is symmetrical around 0.5, one can expect that the transformed p-values $\hat{p}_i$ are also symmetrical around 0.5. This is also demonstrated by the simulations in Figure S5. These p-values are combined again in the inter-experiment analysis using add-CLT. Since add-CLT relies on the mean of these intra-experiment p-values, the distribution of the inter-experiment p-values will become closer to the uniform distribution. As a result, the false positive rate (FPR) of the bi-level meta-analysis will be close to the significance threshold even if the FPR yielded at

the intra-experiment level were to be higher than the threshold.

Please note that the intra-experiment p-values that are potentially biased, are only intermediate values that are not used to draw any conclusions. This phenomenon is illustrated in Figure S5. Panel (a) shows the distribution of the intermediate intra-experiment p-values obtained with the same setting of 20 control and 20 disease samples. Due to the small number of control samples and the lack of independence, this distribution is not perfectly uniform. However, even in this case, the distribution is symmetrical. When these intermediate p-values are combined in the inter-experiment analysis, the distribution of the p-values yielded by the add-CLT method is close to the uniform. Note that this effect of the lack on independence caused by the use of the same control samples is greatly alleviated when the number of control samples increases. Panels (c) and (d) in Figure S5 show that the distributions of the intermediate intra-experiment p-values become uniform as the number of control samples increases.

Figure S6A shows the false positive rate (FPR) of the intra-experiment and inter-experiment analyses using t-test. The left panel displays the hypothetical null distribution H_0 while the middle and right panels displays the FPR of the intra- and inter-experiment analyses, respectively. The middle panel shows that the intermediate FPR of the intra-experiment analysis is notably higher than then threshold 0.01, due to the dependency of the split datasets from a study. However, when multiple studies are combined using add-CLT in the inter-experiment analysis, the final FPR is reduced to the threshold again. The right panel shows that the FPR of the bi-level meta-analysis is consistently at $1 - 2\%$ when the number of studies increases.

In the same figure, we illustrate the true positive rate (TPR) of the bi-level meta-analysis when the null (control) and alternative (disease) distributions are not identical. In Figure S6B, the left panel displays the null distribution (black) H_0 and the alternative distribution (red) H_A. The middle and right panels displays the TPR of the intra-experiment and inter-experiment analyses, respectively. In this case, the TPR is calculated as the number of p-values smaller than the threshold 0.01 divided by $1,000$. The TPR of the bi-level meta-analysis increases linearly with the number of studies. The TPR increases from 10% with 2 studies to almost 50% with 20 studies. On the contrary, the FPR stays at $1 - 2\%$ regardless of the number of studies and the number of samples in each study.

4.3 The effect of group size setting

We would like to have as many split datasets as possible to increase the power of the add-CLT. Many splits would suggest small group sizes. However, we also need the sample size of disease groups to be large enough so each split group has enough biological replicates needed for a meaningful analysis. This would require large group sizes. The default group size of 5 allows each split dataset to have a reasonable sample size, while still allowing for a sufficient number of splits.

However, the choice of the group size does not influence dramatically the results of the approach. Figure S7 shows the true positive rate (TPR) of the intra-experiment analysis using a t-test with different settings of group size. In each row, the left panel displays a hypothetical null distribution H_0 (black) and an alternative distribution H_A (red). The middle panels show the true positive rate (TPR) of the intra-experiment analysis method as they vary with increasing the number of samples. The right panels show the confidence interval of the intra-experiment p-values. Each segment represents the mean and 95% confidence interval of the p-values for a given g (group size) and n (number of samples). The figure shows that the curves of size 6, 7, and 8 are notably lower than the curves of size 3, 4, and 5. This suggests the TPR of the test is likely to decrease if we were to increase the number of group size above 5. However, changing the settings of group size from 3 to 8 makes the TPR and p-value distributions change only slightly. Most importantly, the TPR curves tend to converge when the number of samples increases, showing that the proposed approach is very stable as the number of data points increases. In summary, the TPR of the intra-experiment analysis is stable against the settings of group size in our analysis.

4.4 Comparison between Fisher's method and add-CLT

Here we compare the add-CLT method with Fisher's method, and show that the add-CLT method is more powerful and more robust to bias. Both Fisher's method and add-CLT 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 due to specific factors such as improper normalization, cross-hybridization, poorly characterized variance, and heteroskedasticity in microarray data analysis [1, 9], or even due to properties of some other more general distributions [2]. Therefore, we need to investigate how robust the two meta-analysis methods are, Fisher's method and add-CLT, against

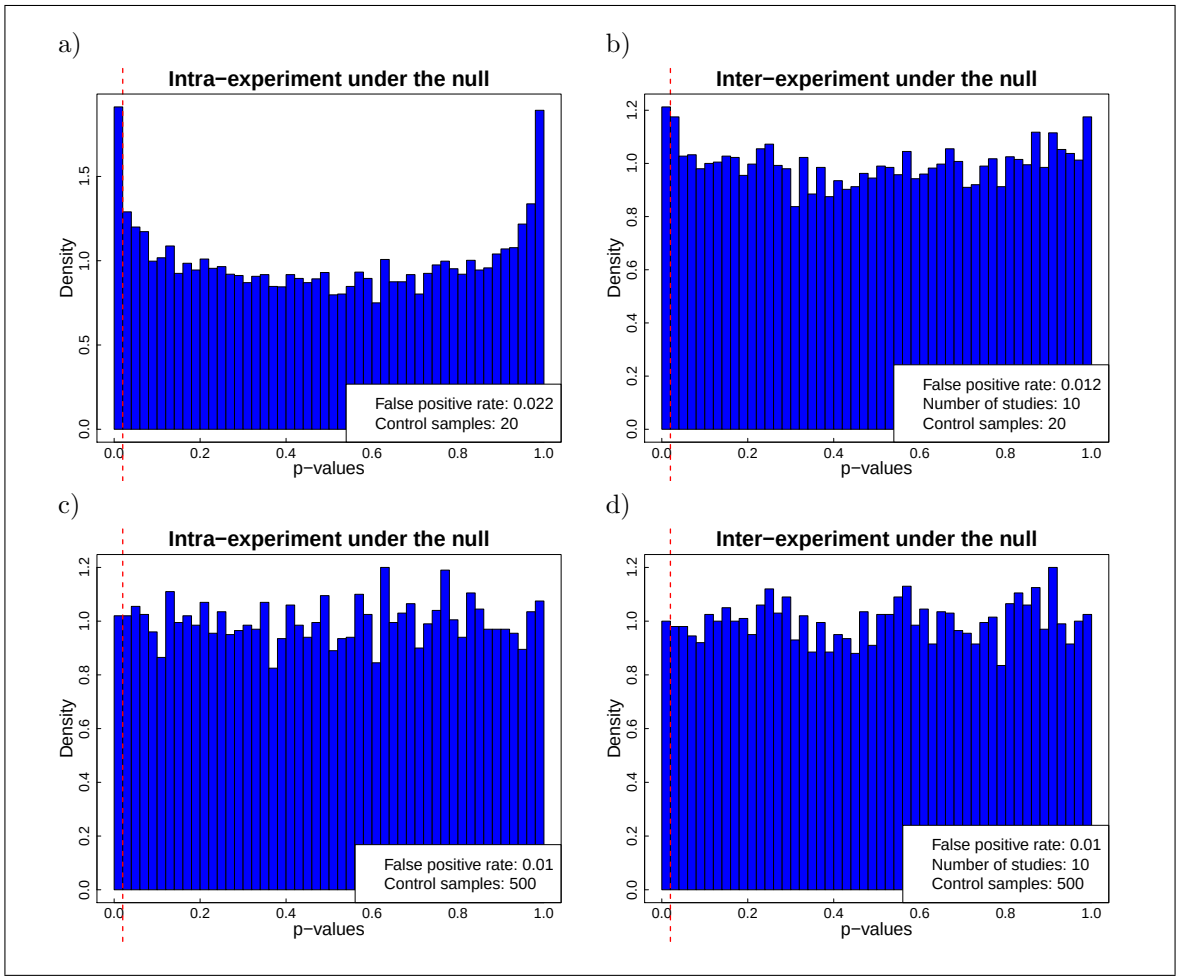

Figure S5: Distributions of p-values under the null for the intra-experiment and inter-experiment analyses. The horizontal axes display the p-values and the vertical axes display the density of the p-values. Panels (a) and (b) display the distributions of the intra-experiment and inter-experiment p-values when the number of control samples is 20. In each study, we randomly pick 20 disease samples and 20 control samples, both from the null distribution, and then calculate intra-experiment p-values. We repeat the procedure $10,000$ times to get the distribution of the intra-experiment p-values. To calculate a meta p-value for the inter-experiment analysis, we combine 10 independent studies using add-CLT. We repeat this procedure $10,000$ times to get the distribution of the inter-experiment p-values. The red dashed line in each panel represents the cutoff 0.01. The p-values to the left of this line are false positives. Panels (c) and (d) are constructed in a similar way, but the number of control samples is set to 500. Panel (a) shows that the distribution of the intra-experiment analysis is not uniform under the null when the number of control samples is small. However, when these intermediate p-values are combined again in the inter-experiment analysis, the distribution of p-values yielded by the add-CLT method is closer to the uniform (panel b). In addition, this non-uniformity is also alleviated when the number of control samples increases as shown in panels (c) and (d). In summary, the two levels of add-CLT keeps the false positive rate close to the threshold 0.01.

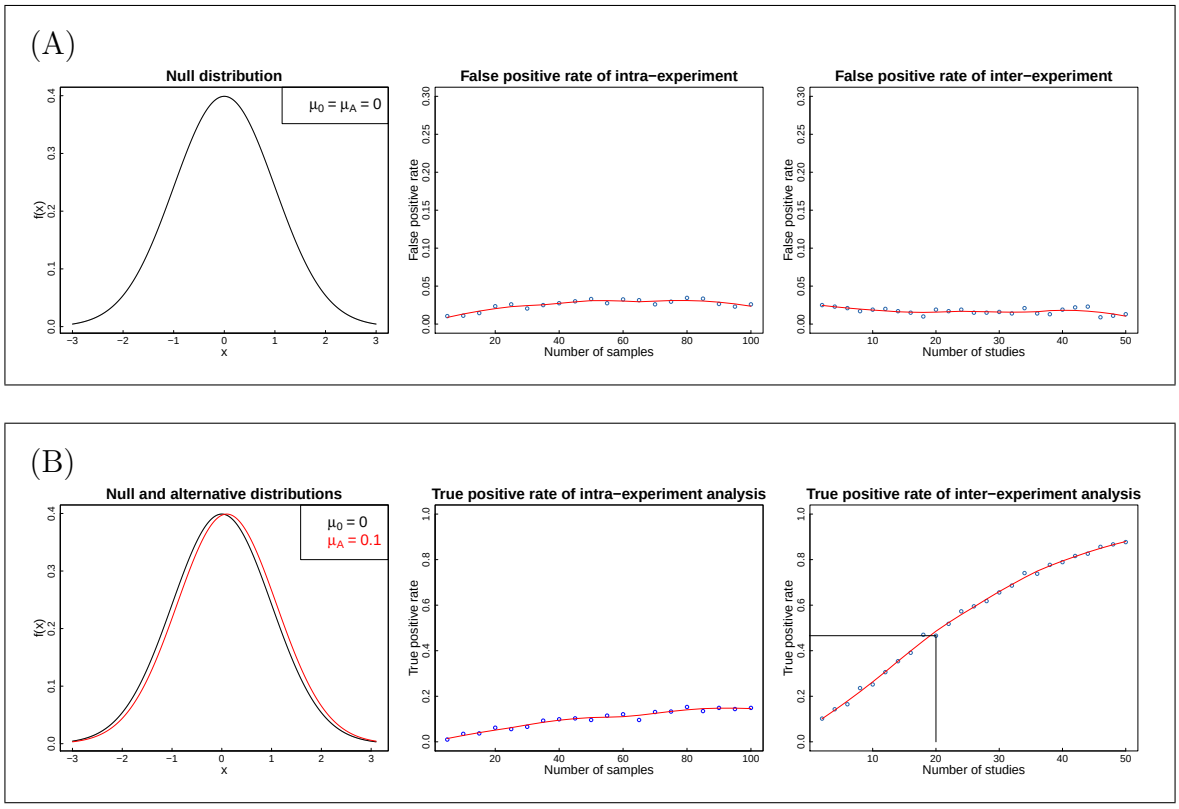

Figure S6: False positive rate (FPR) and true positive rate (TPR) of the bi-level meta-analysis using t-test. In the first row (A), the left panel displays a hypothetical null distribution H_0 while the middle and right panels display the FPR of the intra-experiment and inter-experiment analyses, respectively. Given n as the number of samples for the intra-experiment analysis, we randomly pick n control samples and n disease samples, all from the null distribution H_0. We split the disease group into smaller groups of size 5 and then perform a t-test to compare each newly created small disease group against the control group. The resulting p-values are then combined using add-CLT. We repeat the procedure $1,000$ times to get $1,000$ p-values for the intra-experiment analysis. We then calculate the FPR for the intra-experiment analysis as the number of p-values smaller than the threshold 0.01 divided by $1,000$. Given m as the number of studies in the inter-experiment analysis, we generate m independent datasets, perform intra-experiment analysis for each dataset, and then combine the m resulted p-values using add-CLT. In each generated study, the disease and control groups have the same size, but the number of samples in each group is randomized. We repeat the procedure $1,000$ times to get $1,000$ p-values for the inter-experiment analysis. We then calculate the FPR as the number of p-values smaller than the threshold 0.01 divided by $1,000$. The panels show that the FPR of the inter-experiment analysis is smaller than that of the intra-experiment analysis and is consistently at $1 - 2\%$. In the second row (B), the left panel displays the null distribution (black) H_0 and the alternative distribution (red) H_A. The sampling procedure in (B) is similar to (A) with the exception is that the disease samples are picked from the alternative distribution H_A. In this case, the TPR is calculated as the number of p-values smaller than the threshold 0.01 divided by $1,000$. The middle and right panels display the TPR of the intra- and inter-experiment analyses, respectively. The TPR of the bi-level meta-analysis increases linearly with the number of studies. The TPR increases from 10% with 2 studies to almost 50% with 20 studies. On the contrary, the FPR stays at $1 - 2\%$ regardless of the number of studies and the number of samples in each study.

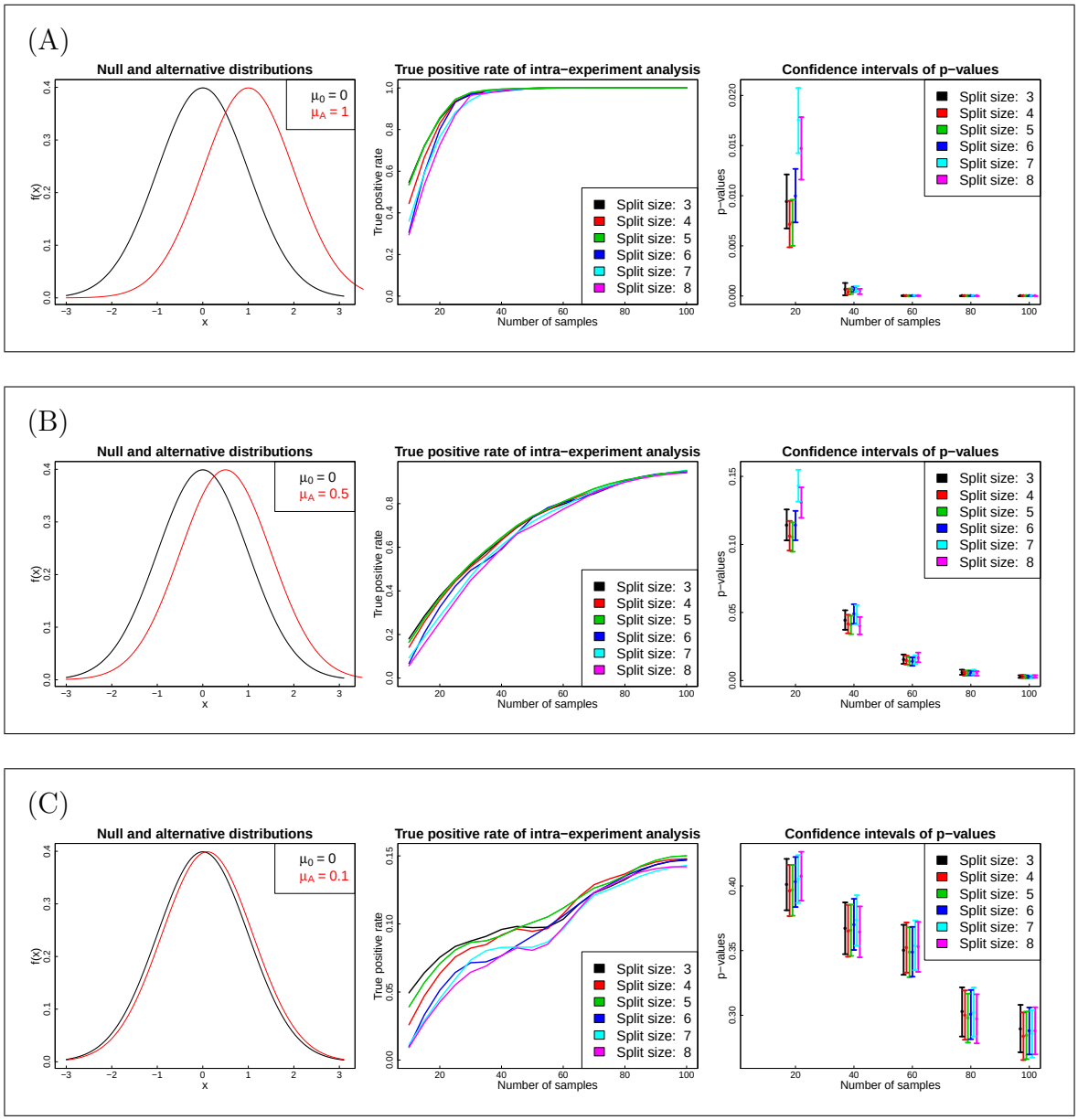

Figure S7: True positive rate (TPR) of the intra-experiment analysis with varying group size. In each row, the left panel displays a hypothetical null distribution H_0 (black) and an alternative distribution H_A (red). The middle panels show the true positive rate (TPR) of the intra-experiment analysis method as they vary with increasing the number of samples. The horizontal axes represent the number of samples while the vertical axes represent the TPR. The right panels show the confidence interval of the intra-experiment p-values. Given n as the number of samples and g as the minimum group size, we randomly pick n control samples from H_0 and n disease samples from H_A. We split the disease group into smaller groups of size g and then perform a right-tailed t-test to compare each newly created small disease group against the control group. The resulting p-values are then combined using add-CLT. We repeat the procedure $1,000$ times to get $1,000$ p-values for the intra-experiment analysis method. We then calculate the TPR as the number of p-values smaller than the threshold 0.01 divided by $1,000$. Each curve in the middle panels shows the TPR for each setting of group size g. In the right panels, each segment represents the mean and 95% confidence interval of the p-values for a given group size g and number of samples n. The middle panels show that the curves for group size 6, 7, and 8 are notably lower than the curves for group size 3, 4, and 5. This suggests the TPR of the test can only decrease when we increase the number of group size over 5. In addition, the TPR and p-values distribution are very similar when g equals to 3, 4, and 5.

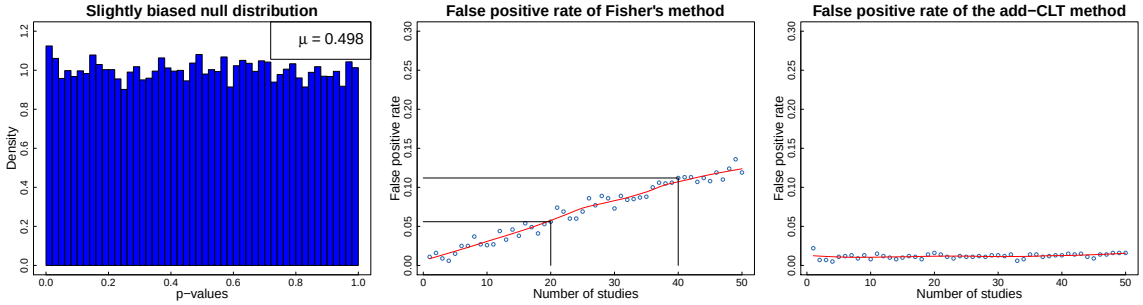

Figure S8: False positive rate (FPR) of Fisher's and the add-CLT method. The method add-CLT is a combination of the additive method and the Central Limit Theorem. The left panel shows a slightly biased distribution of $40,000$ p-values under the null (with mean of 0.498). The horizontal axis displays the p-values while the vertical axis displays the density. The middle and right panels display the FPR using Fisher's method and add-CLT, respectively, varying the number of studies to be combined. The horizontal axes display the number of studies to be combined while the vertical axes display the FPR. Given m as the number of studies, we calculate the FPR of a meta-analysis method as follows. We randomly pick m p-values from the distribution and then combine them to have one combined p-value. We repeat the process 1000 times to have 1000 combined p-values. We then compute the FPR as the number of combined p-values that are smaller than the significance threshold (0.01) divided by 1000. The false positive rate of Fisher's method increases linearly with the number of studies. The FPR of Fisher's method is around 1% with one study, and then reaches 5% and 10% with 20 and 40 studies, respectively. On the contrary, the FPR of add-CLT is constantly near $1-2\%$ regardless of the number of studies to be combined. The figure shows that add-CLT is more robust against weak bias compared to Fisher's method.

weak bias of p-values under the null.

Here we construct a distribution of $40,000$ p-values. This distribution is slightly biased towards zero with distribution mean 0.498. In Figure S8, the left panel shows the distribution of the p-values under the null while the middle and right panels show the false positive rate (FPR) of Fisher's method and add-CLT. For a given value of m as the number of studies, we calculate the FPR of a meta-analysis method as follows. We randomly pick m p-values from the null distribution and then combine them into one combined p-value. We repeat the process $1,000$ times to have $1,000$ combined p-values. We then calculate the FPR of the meta-analysis method for the given value of m as the number of combined p-values that are smaller than the threshold 0.01 divided by $1,000$. We calculate the FPR of both meta-analysis methods for all values of m in the range between 1 and 50.

The middle panel of Figure S8 shows that the FPR of Fisher's method increases linearly with the number of studies. The FPR of Fisher's method is at 1% with one study, and then reaches 5% and 10% with 20 and 40 studies, respectively. This indicates that Fisher's method is sensitive to weak bias under the null. The right panel of Figure S8 shows that the FPR of add-CLT constantly at $1-2\%$ regardless of the number of studies to be combined. In summary, add-CLT is robust against bias under the null.

Theoretically, when there are no disease effects, individual p-values are equally probable between zero and one. Let us now consider the distribution of the p-values for those pathways (or genes) that are truly implicated in the condition over a number of datasets. Some of the truly implicated pathways will have individual p-values that are smaller than the significance threshold, will be identified as relevant and therefore will be true positives. Others, will have p-values higher than the significance threshold, will not be identified as relevant and therefore will be false negatives. Let us denote ξ as the empirical distribution of p-values *only for the truly implicated pathways*. It is reasonable to assume that the mean of this distribution is notably smaller than 0.5.

We simulate three cases of ξ as shown in Figure S9. The left-most panels show the distribution of the p-values. Each distribution consists of $40,000$ p-values. In these panels, the horizontal axes represent the p-values while the vertical axes represent the density of p-values. The middle and right panels show the true positive rate of Fisher's method and add-CLT, varying the number of studies to be combined. Given m as the number of studies to be combined, we calculate the TPR of a meta-analysis method as follows. We randomly sample m p-values from ξ and then combine them into one combined p-value. We repeat the process $1,000$ times to have $1,000$ combined p-values. We then compute the TPR for m as the number of combined p-values that are smaller than 0.01 divided by $1,000$.

In the first case (A), the p-values are highly concentrated near zero and are almost uniformly distributed

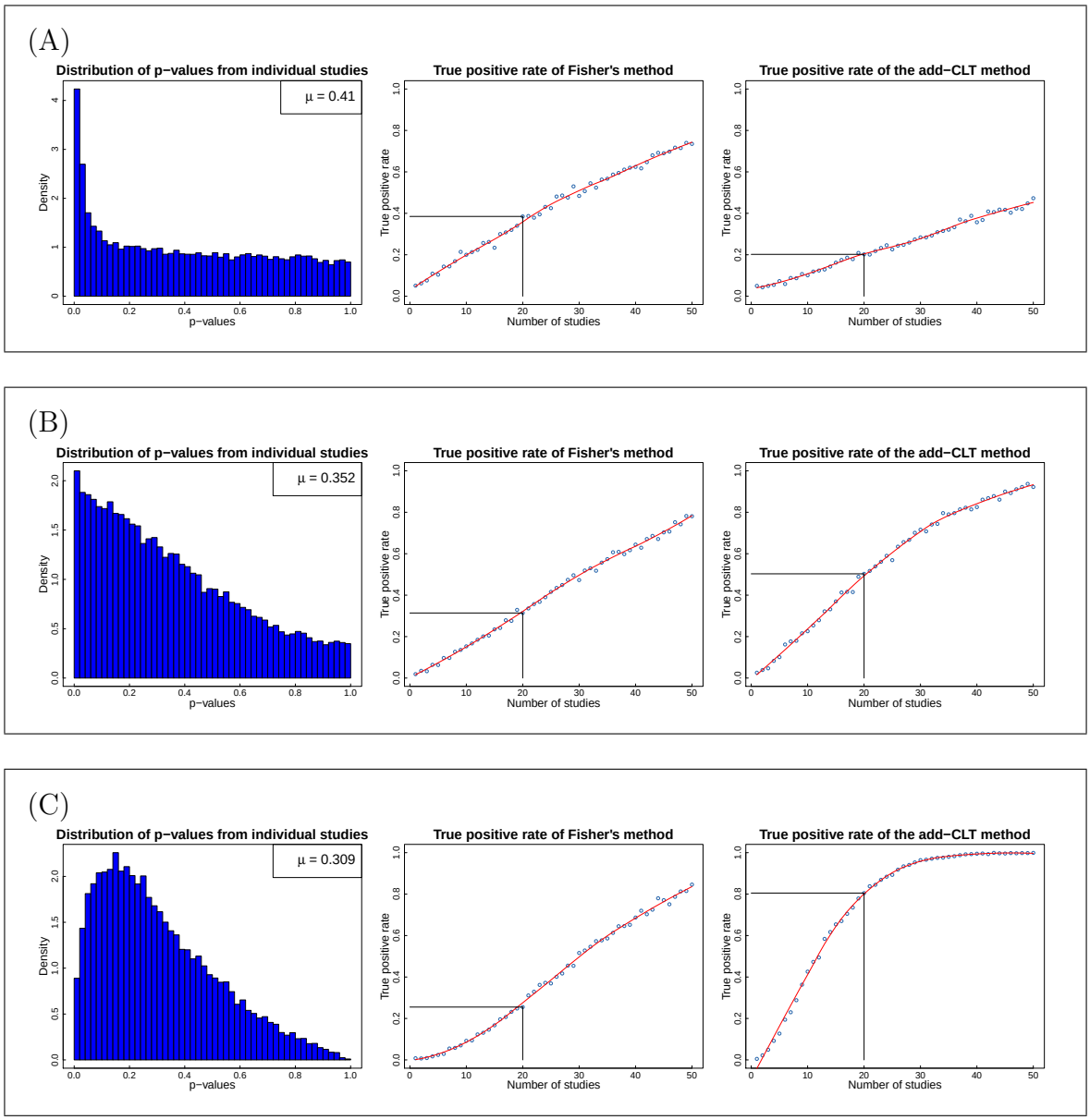

Figure S9: Comparison of true positive rates (TPR) obtained by applying Fisher's and the add-CLT method to three simulated data distributions. The left-most panels in each row show the simulated p-value distributions, ordered by decreasing mean. The horizontal axes of these panels display the p-values while the vertical axes display the densities. The middle and right panels display the TPR using Fisher's method and add-CLT, respectively, with varying numbers of studies to be combined. The horizontal axes display the number of studies to be combined while the vertical axes display the TPR. Given m as the number of studies to be combined, we calculate the TPR of a meta-analysis method as follows. We randomly pick m p-values from the corresponding distribution and then combine them to have one combined p-value. We repeat the process 1000 times to have 1000 combined p-values. We then compute the TPR of the method as the number of combined p-values that are smaller than the threshold 0.01 divided by 1000. In (A), we see that Fisher's method has more power than add-CLT when the p-values are highly concentrated near zero, and are almost uniformly distributed elsewhere. In (B), add-CLT is more powerful than Fisher's method when the density of p-values increases linearly with decreasing of p-values. In (C), add-CLT is much more powerful than Fisher's method when the density peaks at 0.2. Interestingly, Fisher's method loses power when the distribution mean decreases. This shows that Fisher's method is highly sensitive to the shape of the distribution of p-values. On the contrary, the power of add-CLT increases quickly when the distribution mean decreases, regardless of the shape of the distribution. With 20 studies, the TPR of add-CLT increases from 20% to 80% when the distribution mean decreases from 0.41 to 0.309.

in the rest of the interval. In this case, Fisher's method has a higher true positive rate than add-CLT. With 20 studies to be combined, the TPR of Fisher's method is higher - about 40% compared to 20% for add-CLT. The advantage of Fisher's is due to the fact that it favors extremely small p-values. However, this kind of distribution is not likely to represent the p-values of the truly implicated pathways. One would not expect these p-values to be distributed uniformly on most of the interval $[0, 1]$ as shown in Figure S9A, but rather highly concentrated below the significance threshold.

In the second case (B), the density of the p-values linearly increases with the decreasing p-value. In this case, add-CLT has more power than Fisher's method. With 20 studies to be combined, the TPR of add-CLT is 50% while the TPR of Fisher's method is 30%. Although the mean of the distribution in (B) is smaller than that in (A), Fisher's method loses power because the p-values are less concentrated near zero. On the contrary, add-CLT gains power because the distribution mean decreases.

In the third case (C), the density of p-values peaks at 0.2 but the mean of the distribution is even smaller than in (A) and (B). In this case add-CLT is much more powerful than Fisher's method. One likely reason for the shift in power of the two meta-analysis methods is that add-CLT's power greatly improves when the distribution mean decreases. The second reason is that Fisher's method loses power because the p-values are less concentrated near zero. With 20 studies, the TPR of add-CLT is 80% compared to less than 30% TPR of Fisher's method.

The three cases in Figure S9 are sorted with the decreasing order of distribution mean of p-values. One would expect that a meta-analysis method would gain more power when the distribution of p-values shifts towards zero. However, Fisher's method loses power in this scenario due to its sensitivity to extremely small p-values, which are less frequent in this scenario. With 20 studies, the TPR of Fisher's method decreases from 40% in (A) to less than 30% in (C). On the contrary, add-CLT's power greatly increases when the distribution mean decreases. For example, with 20 studies, the TPR of add-CLT increases from 20% to 80% when the distribution mean decreases from 0.41 to 0.309. We conclude that add-CLT has some advantage over Fisher's method in terms of both false positive rate and true positive rate. Therefore, we use add-CLT as the default meta-analysis method to combine p-values in our bi-level meta-analysis framework.

References

[1] Barton, S. J., Crozier, S. R., Lillycrop, K. A., Godfrey, K. M., and Inskip, H. M. (2013). Correction of unexpected distributions of P values from analysis of whole genome arrays by rectifying violation of statistical assumptions. *BMC Genomics*, **14**(1), 161.

[2] Bland, M. (2013). Do baseline p-values follow a uniform distribution in randomised trials? *PloS One*, **8**(10), e76010.

[3] Bolstad, B. M. (2004). *Low-level analysis of high-density oligonucleotide array data: background, normalization and summarization*. Ph.D. thesis, University of California.

[4] Bolstad, B. M., Collin, F., Brettschneider, J., Simpson, K., Cope, L., Irizarry, R., and Speed, T. P. (2005). Quality assessment of Affymetrix GeneChip data. In *Bioinformatics and computational biology solutions using R and bioconductor*, pages 33–47. Springer, New York.

[5] Brettschneider, J., Collin, F., Bolstad, B. M., and Speed, T. P. (2008). Quality assessment for short oligonucleotide microarray data. *Technometrics*, **50**(3).

[6] Edgington, E. S. (1972a). An additive method for combining probability values from independent experiments. *The Journal of Psychology*, **80**(2), 351–363.

[7] Edgington, E. S. (1972b). A normal curve method for combining probability values from independent experiments. *The Journal of Psychology*, **82**(1), 85–89.

[8] Feller, W. (2008). *An introduction to probability theory and its applications*, volume 1. John Wiley & Sons, New York-London-Sydney.

[9] Fodor, A. A., Tickle, T. L., and Richardson, C. (2007). Towards the uniform distribution of null P values on Affymetrix microarrays. *Genome Biology*, **8**(5), R69.

[10] Gosset, W. S. (1908). The Probable Error of a Mean. *Biometrika*, **6**, 1–25.

[11] Hall, P. (1927). 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*, **19**(3-4), 240–244.

[12] Irwin, J. O. (1927). 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*, pages 225–239.

[13] Kallenberg, O. (2002). *Foundations of modern probability*. Springer-Verlag, New York.

[14] Miller, C. J. (2013). *simpleaffy: Very simple high level analysis of Affymetrix data*. R package version 2.38.0.

[15] Peaeson, E. and Haetlet, H. (1976). Biometrika tables for statisticians. *Biometrika Trust*.

[16] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). 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*, **102**(43), 15545–15550.