Bioinformatics, 36(2), 2020, 524–532
doi: 10.1093/bioinformatics/btz589
Advance Access Publication Date: 29 July 2019
Original Paper

Gene expression

P-value evaluation, variability index and biomarker categorization for adaptively weighted Fisher's meta-analysis method in omics applications

Zhiguang Huo[1], Shaowu Tang[2], Yongseok Park[3,*] and George Tseng [3,*]

[1]Department of Biostatistics, University of Florida, Gainesville, FL 32611, USA, [2]Roche Molecular Solutions, Inc., Pleasanton, CA 94588, USA and [3]Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on September 17, 2018; revised on May 25, 2019; editorial decision on July 15, 2019; accepted on July 24, 2019

Abstract

Motivation: Meta-analysis methods have been widely used to combine results from multiple clinical or genomic studies to increase statistical powers and ensure robust and accurate conclusions. The adaptively weighted Fisher's method (AW-Fisher), initially developed for omics applications but applicable for general meta-analysis, is an effective approach to combine *P*-values from K independent studies and to provide better biological interpretability by characterizing which studies contribute to the meta-analysis. Currently, AW-Fisher suffers from the lack of fast *P*-value computation and variability estimate of AW weights. When the number of studies K is large, the $3^K - 1$ possible differential expression pattern categories generated by AW-Fisher can become intractable. In this paper, we develop an importance sampling scheme with spline interpolation to increase the accuracy and speed of the *P*-value calculation. We also apply bootstrapping to construct a variability index for the AW-Fisher weight estimator and a co-membership matrix to categorize (cluster) differentially expressed genes based on their meta-patterns for intuitive biological investigations.

Results: The superior performance of the proposed methods is shown in simulations as well as two real omics meta-analysis applications to demonstrate its insightful biological findings.

Availability and implementation: An R package *AWFisher* (calling C++) is available at Bioconductor and GitHub (https://github.com/Caleb-Huo/AWFisher), and all datasets and programing codes for this paper are available in the Supplementary Material.

Contact: ctseng@pitt.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

High-throughput biological experiments play a key role in deciphering biological mechanisms behind complex diseases. Advanced experimental techniques allow us to obtain high-resolution genomic information with affordable price. Over the years large amount of omics data are accumulated in public databases and repositories: The Cancer Genome Atlas (TCGA) http://cancergenome.nih.gov, Gene Expression Omnibus (GEO) http://www.ncbi.nlm.nih.gov/geo/ and Sequence Read Archive (SRA) http://www.ncbi.nlm.nih.gov/sra/, just to name a few. For a given transcriptomic study from microarray or RNA-seq, many statistical methods have been developed for detecting differentially expressed (DE) genes as candidate biomarkers (Pan, 2002; Soneson and Delorenzi, 2013). The analysis of a single study, however, contains small to moderate sample size (usually $N = 20 \sim 50$), producing unstable and inaccurate results (Domany, 2014; Simon, 2005). Meta-analysis to combine multiple transcriptomic studies has become a common practice to improve statistical power and reproducibility. Readers who are interested may refer to Ramasamy *et al.* (2008) for a practical guideline of microarray meta-analysis, and Tseng *et al.* (2012); Begum *et al.* (2012) for comprehensive reviews of microarray and genome-wide association study meta-analysis.

Among the numerous meta-analysis methods proposed in the literature, combining *P*-values from multiple studies is a simple and flexible solution to combine studies of different experimental design and avoid complexity from batch effect (i.e. systematic non-biological differences between studies because of differences such as sample platforms and experimental protocols; Luo *et al.*, 2010). When combining two studies, it has been demonstrated that such batch effects often cannot be properly removed when by normalization (Guerra and Goldstein, 2009). Meta-analysis methods (e.g. *P*-value combination methods) will lead to increased statistical power by combining the summary statistics (i.e. *P*-values). These summary statistics, representing the strength of association, are usually considered standardized and independent of the batch effect (Gibbons *et al.*, 2018). Under the framework of such *P*-value combination methods, multiple hypothesis testing settings have been considered to address different biological questions. Following the convention of Song and Tseng (2014) (also see Birnbaum, 1954; Li and Tseng, 2011), three major hypothesis settings have been considered in the literature: HS_A targets on the hypothesis testing setting to detect biomarkers that are DE in all cohorts:

$$H_0 : \vec{\theta} \in \bigcap \{\theta_k = 0\}$$
$$H_A : \vec{\theta} \in \bigcap \{\theta_k \neq 0\},$$

where HS_A is the acronym for '*H*ypothesis *S*etting *A*ll' mentioned above, and θ_k is the effect size of study k, $1 \leq k \leq K$). HS_B targets on biomarkers DE in one or more studies:

$$H_0 : \vec{\theta} \in \bigcap \{\theta_k = 0\}$$
$$H_A : \vec{\theta} \in \bigcup \{\theta_k \neq 0\},$$

where '*B*' in HS_B is the counterpart of '*A*' in HS_A. The HS_B method is traditionally also referred to as 'union–intersection test (UIT)' or 'conjunction null hypothesis' in the statistical literature. HS_r targets on biomarkers DE in at least *r* studies:

$$H_0 : \vec{\theta} \in \bigcap \{\theta_k = 0\}$$
$$H_A : \sum \mathbb{I}\{\theta_k \neq 0\} \geq r,$$

where r in HS_r represents the *r*th order statistics (Song and Tseng, 2014), $\mathbb{I}\{\cdot\}$ is an indicator function taking value one if the statement is true and zero otherwise, and r is usually pre-specified with $K/2 \leq r \leq K$).

Biologically HS_A is preferred when the purpose is to find concordant genes across all studies. HS_r can be considered as a robust form of HS_A to seek for concordant genes in majority of studies. HS_B is considered when heterogeneity is expected and we are interested in biomarkers statistically significant in at least one study.

In the literature, HS_B is an UIT (Roy, 1953) and is also called a conjunction or intersection hypothesis (Benjamini and Heller, 2008). Many statistical tests have been developed for this hypothesis setting, including Fisher's method (Fisher, 1992), Stouffer's (Stouffer *et al.*, 1949) method, the minimum *P*-value method (also referred as Tippett's method) (Tippett, 1931) and many others. Fisher's method defines the test statistic as the sum of log-transformed *P*-values: $T^F = -2\sum_{k=1}^{K} \log p_k$, where p_k is the *P*-value from the *k*th study; Stouffer's method uses $T^S = -\frac{1}{\sqrt{K}}\sum_{k=1}^{K} \Phi^{-1}(p_k)$, where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function (CDF) of standard normal distribution. A larger Fisher (or Stouffer) score indicates stronger differential expression evidence. Under the null assumption and assuming independence across studies, the null distribution of Fisher's statistics follows $\chi^2{}_{2K}$ and Stouffer's follows $N(0,\ 1)$. Although Fisher's method has many theoretical advantages (e.g. asymptotic Bahadur optimality under certain restricted Gaussian assumptions; see Littell and Folks, 1971), it has a critical pitfall when heterogeneity is expected across studies. For example, suppose $\vec{p}_1 = (0.001,\ 1,\ 1)$ represents *P*-values of three studies of Gene 1 and $\vec{p}_2 = (0.1,\ 0.1,\ 0.1)$ represents *P*-values of Gene 2. Both genes produce the same Fisher's test statistics and meta-analysis *P*-values ($T^F = 13.8$ and $p^F = 0.032$) but the biological interpretation of the two genes are obviously different. $\vec{p}_1$ indicates strong statistical significance only in the first study, while $\vec{p}_2$ shows marginal statistical significance in all three studies. To characterize study heterogeneity in meta-analysis, Li and Tseng (2011) proposed an adaptively weighted Fisher's method (AW-Fisher) where the Fisher's score is modified as weighted sum and the 0–1 weights can be viewed as latent variables of whether a study contributes DE information to the meta-analysis (details see next paragraph). Aside from additional biological interpretability of AW weights, AW-Fisher also enjoys nice theoretical properties. It has been shown to be admissible (Li and Tseng, 2011) and asymptotic Bahadur optimal under certain Gaussian assumptions (Fang *et al.*, 2019). In addition, Fisher's method is more powerful when all studies are significant and the minimum *P*-value method is more powerful when only one study has small *P*-value. AW-Fisher theoretically takes advantage of both methods on their favored extreme situations (Li and Tseng, 2011). Chang *et al.* (2013) performed a comprehensive comparative study to evaluate 12 popular microarray meta-analysis methods and categorized them into the three complementary hypothesis settings, HS_A, HS_B and HS_r. AW-Fisher was the best performer in the HS_B setting when considering a variety of data and heterogeneity assumptions.

Below we describe the method and rationale for AW-Fisher (Li and Tseng, 2011).

For the convenience of discussion, we focus on two class comparison for detecting DE genes in this paper. However, users can easily extend to multi-class, continuous or survival outcomes using conventional packages for differential expression analysis to generate *P*-values as input of our method. Throughout the manuscript, we use *limma* (Smyth, 2005) for continuous data (e.g. microarray or RNA-seq RPKM data) and *edgeR* (Robinson *et al.*, 2010) for count data (e.g. RNA-seq count data) to calculate *P*-values for each individual study. Define $T(\vec{\mathbf{P}};\vec{\mathbf{w}}) = -2\sum_{k=1}^{K} w_k \log P_k$, where $\vec{\mathbf{w}} = (w_1,\ldots,w_K) \in \{0,\ 1\}^K$ is the AW weight associated with K studies and $\vec{\mathbf{P}} = (P_1,\ldots,P_K) \in (0,\ 1)^K$ is the random variable of input *P*-value vector for *K* studies. Under the null distribution and conditionally on $\vec{\mathbf{w}}$, the significance level obtained by $T(\vec{\mathbf{P}};\vec{\mathbf{w}})$ is $L(T(\vec{\mathbf{P}};\vec{\mathbf{w}})) = 1 - F_{\chi^2_{d(\vec{\mathbf{w}})}}(T(\vec{\mathbf{P}};\vec{\mathbf{w}}))$, where $d(\vec{\mathbf{w}}) = 2\sum_{k=1}^{K} w_k$ and $F_{\chi^2_d}(\cdot)$ is the CDF of χ^2-distribution with degrees of freedom d. Given *P*-value vector $\vec{\mathbf{P}}$, the test statistic of AW-Fisher is defined as

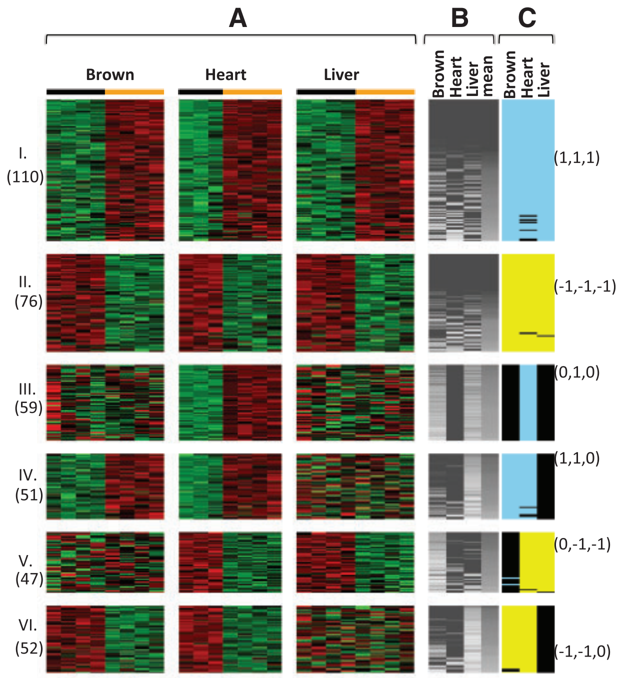

Fig. 1. Six meta-pattern modules of biomarkers from mouse metabolism example. Each gene module (Modules I, II, …, VI) shows a set of detected biomarkers with similar meta-pattern of differential signals. (**A**) Heatmaps of detected genes (on the rows) and samples (on the columns) for each tissue (brown fat, heart, liver), where each tissue represents a study. In the heatmap, red color represents higher expression level, and the green color represents lower expression level. Black color bar on top represents wild-type (control) and orange color bar on top represents VLCAD−/− mice (case). Number of genes is shown on the left under each module number. (**B**) Variability index (genes on the rows and studies on the columns). Variability index is described in Section 2.2. Gray heatmap range from 0 (black) to 1 (white), which is the maximum of the variability index. Genes of each module are sorted based on the mean variability index. (**C**) Signed AW-Fisher weights $\hat{v}_{gk}$ for gene g and study k. Light blue represents $\hat{v}_{gk} = 1$, yellow corresponds to $\hat{v}_{gk} = -1$ and black for $\hat{v}_{gk} = 0$. Representative signed AW-Fisher weights for each module are shown on the right. Note brown represents brown fat tissue. (Color version of this figure is available at *Bioinformatics* online.)

$$s(\vec{\mathbf{P}}) = \min_{\vec{\mathbf{w}}} L(T(\vec{\mathbf{P}}; \vec{\mathbf{w}})). \qquad (1)$$

The optimal weight for $\hat{\mathbf{w}}$ is determined by $\hat{\mathbf{w}} = w(\vec{\mathbf{P}}) = \text{argmin}_{\vec{\mathbf{w}}} L(T(\vec{\mathbf{P}}; \vec{\mathbf{w}}))$. Here we denote by s the mapping from P-value vector to the AW-Fisher test statistic and S is the random variable for the AW-Fisher test statistic which can be obtained by $S = s(\vec{\mathbf{P}})$. We further define signed AW-Fisher weights by:

$$\hat{\mathbf{v}} = (\hat{v}_1, \ldots, \hat{v}_K) = (\hat{w}_1 \cdot \text{sign}(\hat{\theta}_1), \ldots, \hat{w}_K \cdot \text{sign}(\hat{\theta}_K)),$$

where $(\hat{\theta}_1, \ldots, \hat{\theta}_K)$ is the estimate of effect size of each study and $\text{sign}(x) = x/|x|$ if $x \neq 0$ and $\text{sign}(x) = 0$ otherwise. Note that $\hat{v}_k$ can be 0, 1 or −1 for $1 \leq k \leq K$. AW-Fisher is appealing in applications since the AW weight estimate $\hat{\mathbf{w}}$ characterizes which study contributes to the meta-analysis result. In the previous simple example, we have $\hat{\mathbf{w}} = (1,\ 0,\ 0)$ for Gene 1 and $\hat{\mathbf{w}} = (1,\ 1,\ 1)$ for Gene 2, which indicates Gene 1 ($\vec{\mathbf{P}} = (0.001,\ 1,\ 1)$) is a first-study-specific biomarker while Gene 2 ($\vec{\mathbf{P}} = (0.1,\ 0.1,\ 0.1)$) is an all-study-consistent biomarker. Figure 1A shows heatmap of candidate biomarkers declared as DE by the AW-Fisher's method in a mouse metabolism microarray example combining three studies (tissues): brown fat, heart, liver (see detailed description in Section 3.2.1). In each study, VLCAD−/− mutant mice (orange bar on top) were compared to VLCAD+/+ wild-type mice (black bar) and DE analysis was performed using *limma* (Smyth, 2005). Meta-analysis P-values were calculated for each gene using the AW-Fisher method. Benjamini–Hochberg's procedure (Benjamini and Hochberg, 1995) was used to account for multiple comparisons and false discovery rate (FDR) was controlled at 5% level. Among detected biomarkers, some genes are up-regulated DE genes across all tissues (e.g. genes in Module I, $\hat{v} = (1,\ 1,\ 1)$)); many others are tissue specific (e.g. heart-specific biomarkers in Module III, $\hat{v} = (0,\ 1,\ 0)$). If applying Fisher's method, these different gene modules will not be distinguished, which may hinder biologists for further biological investigation and hypothesis generation. Despite the advantages of AW-Fisher in theory and applications, applying AW-Fisher currently encounters three major issues:

1. In the original paper, Li and Tseng (2011) did not derive a closed-form solution for calculating the null distribution of the AW statistics. Instead, a permutation method (permuting case/control labels in each study independently) was suggested. This results in high computing demand, when especially high P-value numerical precision is needed to account for multiple comparisons in omics applications. The search space of all possible weights also becomes high ($2^K - 1$) when K goes large. This limits AW-Fisher in general genomic applications.
2. The original AW weight estimate can generate unexpected discontinuity and is thus not stable. For example, the following two genes were taken from the mouse metabolism example in Figure 1 gene Module II. P-values of the three tissues for probeset 1419484*_a_at* were (0.000391, 0.0962, 0.00211), and P-values for another probeset 1425567*_a_at* were (0.000356, 0.1026, 0.00206). Despite their very similar P-value inputs, 1419484*_a_at* produced an AW weight $\hat{\mathbf{w}} = (1,\ 1,\ 1)$ with P-value 5.64×10^{-5} using AW-Fisher and 1425567*_a_at* produced an AW weight $\hat{\mathbf{w}} = (1,\ 0,\ 1)$ with P-value 5.22×10^{-5}, showing unstable weight estimate of the second study. In other words, the AW weight estimate is a hard classification with no variability estimate and biomarker categorization is thus unstable.
3. Given K studies, the resulting genes could be categorized into $(3^K - 1)$ groups based on their unique AW weight estimate and effect size direction (if separating up-regulation and down-regulation into 1 and −1 weight using $\hat{\mathbf{v}}$; see Fig. 1). This becomes intractable for further biological investigation when K is large. For example, combining $K = 5$ studies produces $3^5 - 1 = 242$ categories of biomarkers.

Our methods aim to solve these issues of the AW-Fisher's method. The performance will be evaluated in both simulations and real data applications.

2 Materials and methods

2.1 Fast computation of AW-Fisher

In this section, we provide solutions to the two computational problems mentioned in Issue 1. We propose a fast algorithm of searching the adaptive weights and an interpolation approach to obtain accurate P-values. In Supplementary Section SI, we also derive closed-form solution for the cases $K = 2$ to benchmark the performance of the proposed method, and $K = 3$ for the purpose of demonstrating difficulties of deriving closed-form solution in general K.

Recall that the AW-Fisher method $T(\vec{\mathbf{P}};\vec{\mathbf{w}}) = -2\sum_{k=1}^{K} w_k \log P_k$ belongs to HS_B, which targets on biomarkers DE in one or more studies $(H_0 : \vec{\theta} \in \bigcap\{\theta_k = 0\}$ versus $H_A : \vec{\theta} \in \bigcup\{\theta_k \neq 0\})$. The search space $\Omega = \{\vec{\mathbf{w}} : \vec{\mathbf{w}} \neq 0, \vec{\mathbf{w}} = (w_1, \ldots, w_K) \in \{0,\ 1\}^K\}$ contains $2^K - 1$ non-zero vectors of weights and searching the whole space Ω to find the AW-Fisher test statistic $s(\vec{\mathbf{P}}) = \min_{\vec{w} \in \Omega} L(T(\vec{\mathbf{P}};\vec{\mathbf{w}}))$ and the adaptive weights $w(\vec{\mathbf{P}}) = \text{argmin}_{\vec{\mathbf{w}} \in \Omega} L(T(\vec{\mathbf{P}};\vec{\mathbf{w}}))$ becomes computationally expensive when K is large. The amount of computation is even more challenging when the AW-Fisher's method is applied to genomic data, where the same procedure is repeated for thousands of genes or even millions of single nucleotide polymorphisms. To overcome this difficulty, we propose a fast algorithm to find $\hat{\mathbf{w}}$ based on the ordered P-values $\{P_{(i)}\}_{i=1}^{K}$ with $P_{(1)} \leq \ldots \leq P_{(K)}$. Specifically, by decomposing Ω into $\Omega = \bigcup_{k=1}^{K} \Omega_k$ with $\Omega_k = \{\vec{\mathbf{w}} : \sum_{j=1}^{K} w_j = k\}$, it can be seen that $s(\vec{\mathbf{P}}) = \min_{\vec{\mathbf{w}} \in \Omega}\{L(T\ (\vec{\mathbf{w}};\vec{\mathbf{P}}))\} = \min_{1 \leq k \leq K} \min_{\vec{\mathbf{w}} \in \Omega_k}\ \{L(T(\vec{\mathbf{w}};\ \vec{\mathbf{P}}))\}$. Given $1 \leq k_0 \leq K$, denote by $\vec{\mathbf{w}}^{k_0} = (w_1{}^{k_0}, \ldots, w_K{}^{k_0})$ the vector of weights such that $-2\sum_{j=1}^{K} w_j{}^{k_0} \log(P_j) = -2\sum_{j=1}^{k_0} \log(P_{(j)})$ (i.e. the Fisher's statistic using the first k_0 smallest P-values). Then it is straightforward to see that the test statistic involving the first k_0 ordered P-values will generate the most significant $L(T(\vec{\mathbf{P}};\vec{\mathbf{w}}))$ in Ω_{k_0}. This implies in Ω_{k_0}, only $\vec{\mathbf{w}}^{k_0}$ has to be considered for further comparison. Therefore, instead of searching the whole space Ω, it is enough to search only K vectors of weights $\{\vec{\mathbf{w}}^1, \ldots, \vec{\mathbf{w}}^K\}$ to find the adaptive weights $\hat{\mathbf{w}}$. The proposed fast algorithm contains two steps: first sorting K P-values [usually with complexity of $\mathcal{O}(K\log(K))$] and then searching K vectors of weights (with complexity of $\mathcal{O}(K)$). Therefore, the fast searching algorithm proposed in this section reduces the computational complexity from $\mathcal{O}(2^K)$ to $\mathcal{O}(K\log(K))$, which can significantly reduce computing time when K is large.

Denote by $\vec{\mathbf{p}}_{\text{obs}}$ the observed P-values from individual studies and $s_{\text{obs}} = s(\vec{\mathbf{p}}_{\text{obs}})$ the observed AW-Fisher statistic. Theoretically, the P-value of the AW-Fisher's method $\mathbb{P}_{H_0}(S \leq s_{\text{obs}})$ can be calculated analytically for any $K \geq 2$. However, the formulae involve the evaluation of a K-fold integral and the integration domain becomes very complicated for $K \geq 3$, which makes the derivation of the closed-form solution tedious and fallible. For illustration, closed-form derivation of $K = 2$ and $K = 3$ are shown in Supplementary Section SI. In Li and Tseng (2011), a permutation test by randomly permuting class labels in each study was proposed. Although this non-parametric approach has its merit of maintaining gene dependency structure, it is computationally demanding and difficult for generating precise small P-value, such as when $P\text{-value} < 10^{-4}$, which is a critical requirement for multiple testing correction on thousands of genes. In this paper, we propose to use importance sampling to obtain an accurate numerical approximation of $\mathbb{P}_{H_0}(S \leq s_{\text{obs}})$. Importance sampling is a method to accurately estimate the expectation of a function with very small value using Monte Carlo sampling. The idea behind importance sampling is to draw samples from a suitable new distribution function rather than the original one of interest and assign a weight to each sample based on the ratio of two density functions.

To evaluate AW-Fisher P-value $\mathbb{P}_{H_0}(S \leq s_{\text{obs}})$ using importance sampling, we propose a beta-distribution density function $f^*(\cdot)$ to draw $\vec{\mathbf{P}}$ instead of natural uniform distribution $f(\cdot)$ so that we can 'over-sample' those small P-values that result in a large S. It holds that

$$\begin{aligned}\mathbb{P}_{H_0}(S \leq s_{\text{obs}}) = \mathbb{E}_{H_0}[\mathbb{I}\{S \leq s_{\text{obs}}\}] &= \int \mathbb{I}\{S \leq s_{\text{obs}}\} f(\vec{\mathbf{P}}) d\vec{\mathbf{P}} \\ &= \int \mathbb{I}\{S \leq s_{\text{obs}}\} \frac{f(\vec{\mathbf{P}})}{f^*(\vec{\mathbf{P}})} f^*\left(\vec{\mathbf{P}}\right) d\vec{\mathbf{P}} \\ &= \mathbb{E}^*[\mathbb{I}\{S \leq s_{\text{obs}}\} \times W(\vec{\mathbf{P}})], \end{aligned} \tag{2}$$

where $f(\cdot)$ is the density of $\vec{\mathbf{P}}$ under the null and $f^*(\cdot)$ is the proposed density function of $\vec{\mathbf{P}}$ for importance sampling. Importance sampling weight $W(\cdot) = f(\cdot)/f^*(\cdot)$, $\mathbb{E}(\cdot)$ and $\mathbb{E}^*(\cdot)$ are the expectation with respect to $f(\cdot)$ and $f^*(\cdot)$ respectively. Therefore, we can obtain expectation from the original measure using a more efficient new one by applying weights for different samples in Monte Carlo method. Under the null hypothesis and independence assumption between different studies, $P_{\underset{\rightarrow}{k}} \sim \text{UNIF}(0,\ 1)$ for all $1 \leq k \leq K$, so the joint distribution of $f(\vec{\mathbf{P}}) = 1$. If we instead use $\text{Beta}(\eta,\ \ 1)$ distribution as the proposed distribution of each study for importance sampling, then $f^*(\vec{\mathbf{P}}) = \eta^K(\prod_{k=1}^{K} P_k)^{\eta-1}$. To implement importance sampling, suppose we simulate $\vec{p}_i = (p_{i1}, \ldots, p_{iK})$, where $p_{ik} \overset{\text{i.i.d.}}{\sim} \text{Beta}(\eta,\ \ 1)$ for $1 \leq i \leq n$ and $1 \leq k \leq K$. Denote by $s_i = s(\vec{p}_i)$. From Equation (2), we calculate estimate of $\mathbb{P}_{H_0}(S \leq s_{\text{obs}})$ by

$$\hat{\mathbb{P}}_{H_0}\left(S \leq s_{\text{obs}}; \eta,\ \ \vec{p}_1, \ldots,\ \ \vec{p}_n\right) = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\mathbb{I}\{s_i \leq s_{\text{obs}}\} \cdot \frac{1}{\eta^K (\prod_{k=1}^{K} p_{ik})^{\eta-1}} \right). \tag{3}$$

Our P-value evaluation procedure (also see the flowchart in Supplementary Fig. S1) has the following steps:

1. For a targeted number of studies K and target P-value c, calculate its AW-Fisher test statistic.
    a. (Identify suitable η for given c_t and K): Note that different η can provide better importance sampling for different range of targeted c_t given K. When $\eta = 1$, the importance sampling method reduces to the naive Monte Carlo method. To identify an appropriate η given c_t and K, we simulate $\vec{q}_i = (q_{i1}, \ldots, q_{iK})$, where $1 \leq i \leq 1000$ and $q_{ik} \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,\ 1)$. Denote by $q_i{}^{1/\eta} = (\vec{q}_{i1}{}^{1/\eta}, \ldots, q_{iK}{}^{1/\eta})$ with element-wise power to $1/\eta$ and $r_i{}^{(\eta)} = s(\vec{q}_i{}^{1/\eta})$. Define $r_0 = \text{median}_{1 \leq i \leq 1000}(r_i{}^{(\eta)})$. Note that since $q_{ik} \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,\ 1)$, $q_{iK}^{1/\eta} \sim \text{Beta}(\eta, 1)$. From Equation (3), we have

$$\begin{aligned}\phi(\eta) &= \hat{\mathbb{P}}_{H_0}(S \leq r_0; \eta,\ \vec{q}_i{}^{1/\eta}, \ldots, \vec{q}_{1000}{}^{1/\eta}) \\ &= \frac{1}{1000} \cdot \sum_{i=1}^{1000} \left(\mathbb{I}\left\{r_i^{(\eta)} \leq r_0\right\} \cdot \frac{1}{\eta^K \left(\prod_{k=1}^{K} q_{iK}{}^{1/\eta}\right)^{\eta-1}} \right). \end{aligned} \tag{4}$$

We choose $\eta(K, c_t)$ as the root of $\phi(\eta) = c_t$, which can be numerically obtained using 'uniroot()' function in R. This choice of η guarantees half of the simulated samples will effectively contribute to the importance sampling calculation for each targeted c_t. Alternatively, one can choose η by minimizing the variance (i.e. find η such that $\text{Var}_{H_0}[\mathbb{I}\{S \leq s_{\text{obs}}\}]$ is minimized). However, for $c_t \geq 0.01$, we set $\eta = 1$ since the gain of importance sampling diminishes.

    b. (Derive corresponding AW-Fisher statistic for targeted P-value c_t): Next, we derive the corresponding AW-Fisher statistic $S_{K,t}$ for a targeted P-value c_t given K. Given K and c_t, we use $\eta(K, c_t)$ (abbreviated as η hereafter) from the previous step to draw $\vec{o}_i = (o_{i1}, \ldots, o_{iK})$, where $1 \leq i \leq 10^7$ and $\vec{o}_i \overset{\text{i.i.d.}}{\sim} \text{Beta}(\eta, 1)$. Denote by $t_i = s(\vec{o}_i)$ the corresponding AW-Fisher statistic of $\vec{o}_i$ and $t_{(1)} \leq t_{(2)} \leq \ldots \leq t_{(10^7)}$ are ordered from $t_1, \ldots, t_{10^7}$. Define

$$
\begin{aligned}
m_i &= \hat{\mathbb{P}}_{H_0}(S \le t_{(i)}; \eta, \vec{o}_1, \ldots, \vec{o}_{10^7}) \\
&= \frac{1}{10^7} \cdot \sum_{j=1}^{10^7} \left(\mathbb{I}\{t_j \le t_{(i)}\} \cdot \frac{1}{\eta^K \left(\prod_{k=1}^{K} o_{jk} \right)^{\eta-1}} \right).
\end{aligned}
\tag{5}
$$

Note that m_i is monotonically increasing with respect to i and $m_1 \approx 0$. There exists i^* such that $m_{i^*} \le c_t < m_{i^*+1}$. The corresponding AW-Fisher statistic $S_{K,t}$ given K and c_t is chosen as $S_{K,t} = t(i^*)$.

2. Specify a grid of targeted $K = 2,\ 3,\ \ldots,\ 100$ and targeted AW-Fisher P-values as $\{c_t, t = 1,\ 2,\ \ldots,\ 198\} = \{1,\ 0.99,\ 0.98,\ 0.97,\ \ldots,\ 0.03,\ 0.02,\ 0.01, 10^{-3}, 10^{-4}, 10^{-5}, \ldots, 10^{-100}\}$.
3. (Interpolation to calculate P-value of a given S_{obs}): From Step 1b, the library of c_t and $S_{K,t}$ ($t = 1, \ldots, 198$ and $K = 2, \ldots, 100$) is established for interpolation. For any given AW-Fisher statistic S_{obs} and K, we apply function *splinefun* in R with *monoH.FC* option using $(\log(S_{K,t}),\ \log(c_t))$, where $t = 1,\ 2, \ldots,\ 198$, to fit a smooth curve and identify the corresponding P-value of S_{obs}. Note that we apply spline on log-scale P-value to avoid numerical overflow.

Remark. *In Step 1a, given K, we simulate* $q_{ik} \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,\ 1)$ *and take the power of* $1/\eta$, *instead of simulating from* $\text{Beta}(\eta, 1)$. *This design guarantees* $\phi(\eta)$ *is a monotone function with respect to* η *by eliminating the uncertainty from sampling* q_{ik} *for each* η.

For any future input P-values, we only need to calculate the AW-Fisher statistics and interpolate the statistics to obtain AW-Fisher P-value by the spline curve fitting. The design of our base library $\{(\log(S_{K,t}), \log(c_t));\ (t = 1, \ldots, 198,\ \text{and}\ \ K = 2, \ldots, 100)\}$ facilitates accurate estimation for the AW-Fisher P-value in the range of $(10^{-100}, 1)$ and K up to 100. Although the computation is demanding to generate the base library, it only runs once before we generate our AW-Fisher R package and will not affect computing for users. In fact, as shown in Section 3.1.1, the new approach is even faster than closed-form solution once the base library in the R package is established.

2.2 Variability index of adaptive weights

As discussed in Issue 2 in the Section 1, the AW weight estimate $\hat{w}_g = (\hat{w}_{g1}, \ldots, \hat{w}_{gK})$ is discontinuous as a function of the input P-values and thus may not be stable. Denote by $U_{gk} = 4 \cdot \text{Var}(\hat{w}_{gk})$ the variability index of the AW weight estimate for gene g in study k, where the normalization factor 4 scales U_{gk} to range within $[0,\ 1]$. The variability index gauges the stability of $\hat{w}_{gk}$, where a smaller variability index indicates a stable AW weight estimate. However, U_{gk} is not easy to evaluate since $\hat{w}_{gk}$ is binary. Here, we propose a bootstrap procedure to calculate an estimate of U_{gk}. The procedure is as follows:

1. Obtain a bootstrap sample and repeat the following procedure B ($b = 1, \ldots, B$) times.
   - Denote by $D_k \in \mathbb{R}^{G \times N_k}$ the data matrix of study k, where G is the total number of genes and N_k is the total number of samples for study k. c_{ki} is the case–control label, where $i \in \{1, \ldots, N_k\}$ is the sample index and $c_{ki} = 0$ or 1, representing sample i belongs to control or case group.
   - Perform bootstrapping by resampling columns with replacements within case and control, respectively. To be specific, we create an empty data matrix ${D_k}^{(b)} \in \mathbb{R}^{G \times N_k}$. Then sample the ith column of ${D_k}^{(b)}$ using jth column of D_k, where $j \in \{j' : c_{kj'} = c_{ki}\}$. This bootstrap procedure is stepped through for $i = 1, \ldots, N_k$ with replacement (allowing ${D_k}^{(b)}$ as identical columns).

     Use bootstrapped data matrix ${D_k}^{(b)}$ to generate an AW weight estimate ${\hat{w}_{gk}}^{(b)}$ and an effect size estimate ${\hat{\theta}_{gk}}^{(b)}$.
   - Calculate the variability index estimate $\hat{U}_{gk}$ of $\hat{w}_{gk}$ for gene g in study k, where $\hat{U}_{gk} = \frac{4}{B} \sum_{b=1}^{B} \left({\hat{w}_{gk}}^{(b)} - \frac{1}{B} \sum_{b'=1}^{B} {\hat{w}_{gk}}^{(b')} \right)^2$.

Here $\hat{U}_{gk}$ ranges from 0 to 1 with $\hat{U}_{gk} = 0$ represents ${\hat{w}_{gk}}^{(b)} = \hat{w}_{gk}$ for all b, which indicates stable estimate of the AW weight since its bootstrap variance is 0. $\hat{U}_{gk} = 1$ represents ${\hat{w}_{gk}}^{(b)} = 0$ for half of $b's$ and ${\hat{w}_{gk}}^{(b)} = 1$ for the other half of $b's$. A large variability index indicates an unstable estimate of the AW weight.

2.3 Resampling-based ensemble clustering for biomarker categorization

In order to categorize detected genes into biomarker groups with similar differential meta-pattern (Issue 3), we extended the bootstrapping procedure in Section 2.2 to obtain a co-membership matrix for all pairs of genes where each element of the co-membership matrix represents a similarity of signed AW weight $\hat{\text{v}}$ of two genes. Specifically, denote by ${\hat{v}_{gk}}^{(b)} = {\hat{w}_{gk}}^{(b)} \cdot \text{sign}({\hat{\theta}_{gk}}^{(b)})$ from Section 2.2. Define co-membership matrix from each bootstrap sample b as $W^{(b)} \in \mathbb{R}^{G \times G}$ with elements ${W_{gg'}}^{(b)} = 1$ if ${\hat{v}_{gk}}^{(b)} = {\hat{v}_{g'k}}^{(b)}$ for all k, and ${W_{gg'}}^{(b)} = 0$ otherwise. The final co-membership matrix is defined as $V = \sum_{b=1}^{B} W^{(b)}/B$. We further applied the tight clustering algorithm (Tseng and Wong, 2005) (*tight.clust* function within R package *tightClust*) using the co-membership matrix V to obtain tight modules. Tight clustering is able to produce tight and stable gene modules without forcing all genes into clusters. Note that for the clustering algorithm, the dissimilarity measure between two genes i and j can be calculated as $d_{ij} = D(V_i, V_j)$, where $V_i \in [0, 1]^G$ is the ith column of the co-membership matrix V and D can be any distance measurement mapping. Throughout this paper, we calculate the dissimilarity measure using the Euclidean distance $d_{ij} = ||V_i - V_j||_2$. The resulting gene modules show unique DE patterns across multiple studies (namely meta-pattern). We perform the biomarker categorization (clustering) procedure only on declared DE genes at certain FDR cutoff. Genes of each resulting module are then sorted by the variability index and visualized by heatmaps. Below we perform simulations to demonstrate the performance of the resampling-based ensemble clustering for biomarker categorization.

Note that based on the co-membership matrix, one can use any clustering algorithm to obtain the DE patterns. We used the tight clustering algorithm throughout this manuscript due to its capability to remove noise genes (genes that are not assigned to any cluster).

3 Results

3.1 Simulation results

3.1.1 Simulation and numerical evaluation for fast AW-Fisher computing

In Section 2.1 we introduced the fast computation for the AW-Fisher P-value via importance sampling and interpolation by spline smoothing. In this section, this interpolation approach will be compared to the original permutation-based approach in Li and Tseng (2011) and Wang *et al.* (2012). The comparisons include evaluation of accuracy and computing speed. In terms of computing speed, our

Table 1. AW-Fisher *P*-value accuracy in terms of rMSE comparing interpolation approach, permutation-based approach and Monte Carlo approach with closed-form solution as benchmark

P-value range	Interpolation	Permutation		Monte Carlo	
		$B = 10^3$	$B = 10^4$	$B = 10^3$	$B = 10^4$
(0.01, 1]	0.0002	0.0031	0.0014	0.0015	0.0002
(0.001, 0.01]	0.0003	0.071	0.035	0.042	0.0047
(0.0001, 0.001]	0.0007	0.31	0.15	0.25	0.037
(1e-10, 0.0001]	0.0006	3.3	2.5	3.3	2.5
(1e-50, 1e-10]	0.0023	16.6	15.7	16.6	15.7
(1e-100, 1e-50]	0.0069	59.1	58.1	59.1	58.1
Time	0.011 s	18.5 min	2.1 h	9.5 min	1.6 h

Note: Two studies (sample size $N_1 = 20$, $N_2 = 20$) are included as input. *B* is the number of permutations/samplings, and the closed-form solution and the interpolation approach do not require any permutation. The range of the resulting AW-Fisher *P*-values is displayed in the first column. The computing time for each method is displayed in the last row. The computing time for the closed-form solution is 0.06 s.

approach applies a new linear sorting algorithm for searching weights and an interpolation for *P*-value calculation. The improvement of linear sorting algorithm is quite obvious: the search space reduces from an exponential order $\mathcal{O}(2^K)$ to almost linear order $\mathcal{O}(K\log(K))$. Below we utilize the closed-form solution for $K = 2$ (details in Supplementary Section SI.1) as the underlying truth to compare the new approach with the existing permutation approach. The linear sorting does not improve computing speed when $K = 2$ and the improvement will mainly come from the interpolation. The simulation setting is described in Supplementary Section SII.

Using the closed-form solution as the underlying truth, we evaluated the performance of the AW-Fisher *P*-value with $K = 2$ studies from the interpolation approach using the permutation approach as the primary baseline. Since the importance sampling method reduces to Monte Carlo method when setting $\eta = 1$ in Equation (3), we also treat the Monte Carlo method as a secondary baseline. To formally evaluate the accuracy from the permutation approach, Monte Carlo approach, or the interpolation approach, we utilized root mean square error (rMSE): $\text{rMSE} = \sqrt{\sum_{g=1}^{G_1} (\alpha_g - \beta_g)^2 / G_1}$, where α_g is the $-\log_{10}$ (AW-Fisher *P*-value) for gene *g*; β_g is the $-\log_{10}$ (AW-Fisher *P*-value) for gene *g* from closed-form solution; G_1 is the total number of genes being evaluated; and the rMSE indicates the accuracy of *P*-value estimates with smaller rMSE for better estimation. The result for (i) $N_1 = 20$, $N_2 = 20$ is shown in Table 1, the results for (ii) $N_1 = 50$, $N_2 = 50$ and (iii) $N_1 = 20$, $N_2 = 50$ are shown in Supplementary Tables S1 and S2. Our proposed interpolation approach is superior to permutation-based approach and the Monte Carlo approach in terms of accuracy. In terms of computing time, we want to de-emphasize the permutation-based approach because in some cases the per study *P*-values will be calculated by permutation tests, where the permutations occur anyway so that calculating the meta-analysis *P*-value through permutation will not incur much additional cost. However, the permutation test is not always available, for example, when sample sizes per study may be too small to perform permutation, or in a meta-analysis context one may only have access to computed *P*-values, but maybe not the raw data. Under these scenarios, Monte Carlo approach is more appropriate to calculate the *P*-value. Regardless using the permutation-based approach or the Monte Carlo approach as baseline, we observe that the interpolation approach is much faster. Note that the interpolation approach is even faster than closed-form solution because the interpolation is only based on spline curve fitting using data in the library and does not implement importance sampling method while the closed-form method requires evaluation of power and logarithmic functions.

To further evaluate whether our proposed method can well control the Type I error rate, we have included QQ plots as well as estimated Type I errors at various levels under the null. From the QQ plots (Supplementary Figs S2, S3 and S4), we can see the genomic inflation factors ranging from 0.999 to 1.001, indicating that the Type I error rate is well controlled. From the Type I error rate estimation table (Supplementary Table S3), we observe that the Type I error rate is correctly estimated around their nominal level, further corroborating the accurate Type I error rate control of the proposed interpolation method.

3.1.2 Simulation results for the variability index

The main simulation setting mimics a transcriptomic study by considering generative process of DE genes and correlation structures between genes. The procedure generally follows Song and Tseng (2014). Note that this simulation also applies to Section 3.1.1. Details of the simulation procedure are in Supplementary Section SII.

To evaluate the performance of the variability index, we considered different combinations of biological variance ($\sigma = 1,\ 1.5,\ 2$) and sample sizes ($N = 20,\ 50,\ 80$). The result in Supplementary Figure S5 shows that when the dataset has smaller sample size or larger biological variation, the variability index becomes larger. Since the variability index gauges the stability of the AW weight estimate, it can be seen that noisy datasets tend to generate large variability index. In addition, we drew bar plot of the variability index with respect to true positive (TP), false positive (FP) and false negative (FN). As shown in Supplementary Figure S6, FN has larger variability index, followed by FP, and lastly TP, under different scenarios of sample sizes and noise level, which is expected since the weight of TPs are likely to be stable, and the weight of FP and FN are on the decision boundary, and thus unstable. The scatter plots of *P*-value and the variability index are shown in Supplementary Figure S7. We observed that with respect to the increase of $-\log_{10}(P - \text{value})$, the variability index goes up, and finally goes down to zero.

3.1.3 Simulation result for biomarker categorization

To evaluate the performance of biomarker categorization, we adopted a simulation procedure similar to Section 3.1.2 and Huo *et al.* (2019). We simulated $K = 4$ studies in total and 50 control subjects and 50 case subjects in each study. Among the $G = 10\,000$ genes, we set 4% as homogeneously concordant DE genes, DE with the same direction in all studies (all positive or all negative). We denote 'homo+' as the homogeneously concordant DE genes with all positive effect sizes and 'homo−' as the homogeneously concordant DE genes with all negative effect sizes. We also set another 4% as study-specific DE genes—differential expressed only in one study. Among them, 1/4 are DE genes only in the first study with positive effect sizes (denoted as 'ssp1+'), 1/4 are DE genes only in the first study with negative effect sizes (denoted as 'ssp1−'), 1/4 are DE genes only in the second study with positive effect sizes (denoted as 'ssp2+'), and the rest 1/4 are DE genes only in the second study with negative effect sizes (denoted as 'ssp2−'). The rest of the genes are

Table 2. Contingency table of 794 detected DE genes with simulation underlying truth (on the columns) and the tight clustering result with 6 target modules (on the rows)

Module	homo−	homo+	ssp1−	ssp1+	ssp2−	ssp2+	Non-DE
1	0	**177**	0	0	0	0	0
2	**184**	0	0	0	0	0	0
3	0	0	0	**74**	0	0	1
4	0	0	**60**	0	0	0	1
5	0	0	0	0	0	**102**	2
6	0	0	0	0	**85**	0	3
0	13	24	19	11	6	5	**27**

Note: 0 represents the scattered gene group. $1\sim6$ represent 6 detected modules. Bolded numbers are genes with correct assignment.

non-DE (denoted as 'non-DE'). The biological variation parameter σ is set to 1 in this simulation.

By applying the AW-Fisher method, we obtained 794 genes based on FDR at 5% under HS$_B$ HS_B. Co-membership of these genes were calculated with $B = 1000$ and used as input for our gene module detection using the tight clustering algorithm. We identified 6 gene modules in these 794 genes. The detected gene modules are tabulated against the true gene modules simulated in Table 2 (Module 0 contains scattered genes not assigned to any of the six modules). The FDR is well controlled at $34/794 = 4.3\%$ while the nominal FDR is 5%. The detected gene modules clearly correspond to the true modules, and most of the non-DE genes were left out as the noises. The meta-pattern, variability index and AW weight estimates of these six modules are shown in Supplementary Figure S8. This simulation study showed that the proposed algorithm can recover the underlying gene meta-pattern.

Based on the co-membership matrix, we can perform any clustering algorithm to obtain the gene meta-pattern. Hence, we also applied *K*-means and hierarchical clustering for comparison. The confusion tables of the resulting gene modules and the underlying true gene modules are shown in Supplementary Tables S4 and S5. In order to benchmark the results, we adopted the adjusted Rand index (Hubert and Arabie, 1985), which is a measure for the similarity between two clustering assignment results (ranges from 1 to −1), with larger number indicating better consistency. Compared to the underlying truth, the adjusted Rand index for tight clustering, *K*-means and hierarchical clustering are 0.83, 0.82 and 0.80, respectively.

3.2 Transcriptomic meta-analysis applications

3.2.1 Mouse metabolism example

Very long-chain acyl-CoA dehydrogenase (VLCAD) deficiency was found to be associated with energy metabolism disorder in children. Two genotypes of the mouse model—wild-type (VLCAD +/+) and VLCAD-deficient (VLCAD −/−)—were studied for four types of tissues (brown fat, liver, heart and skeleton) with 3–4 mice in each genotype group. In order to better demonstrate the biological interpretability of the proposed method, we focused on the first three tissues (brown fat, liver and heart) as our primary analysis. We also analyzed all four tissues in order to demonstrate when the number of studies *K* is large, the biomarker categorization based on the unique AW weight estimate becomes infeasible but our proposed method is much more tractable. Total number of probesets from these three transcriptomic microarray studies is 14 495. Supplementary Table S6a shows details of the study design. Two-sided *P*-values and effect sizes were calculated using *limma* comparing wild-type (VLCAD +/+) versus mutant (VLCAD −/−) mice in each tissue. AW-Fisher meta-analysis *P*-values were obtained and *q*-values were calculated by applying Benjamini–Hochberg procedure. By controlling FDR at 5%, we obtained 967 DE genes. We calculated the variability index and generated gene co-membership matrix using resampling techniques. We further applied the tight clustering algorithm on the co-membership to identify gene modules with unique meta-pattern. In this example, we successfully detected six gene modules with different meta-patterns in Figure 1. For example, the first and second biomarker modules (Gene clusters I and II) are concordant genes that are up-regulated (or down-regulated) in all tissues. The other biomarker modules have study-specific differential patterns. For example, DE genes in Gene module III are up-regulated in heart but not in brown fat or liver. To examine the biological functions of these modules, we performed pathway enrichment analysis for genes in each module using Fisher's exact test. The pathway database was downloaded from Molecular Signatures Database (MSigDB) v5.0 (http://bioinf.wehi.edu.au/software/MSigDB/), where a mouse-version pathway database was created by combining pathways from KEGG, BIOCARTA, REACTOME and GO databases and mapping all the human genes to their orthologs in mouse using the Jackson Laboratory Human and Mouse Orthology Report (http://www.informatics.jax.org/orthology.shtml). Among the six gene modules with distinct meta-patterns, Module I is enriched in enzyme activities (e.g. GO COFACTOR BINDING; $p = 3.85 \times 10^{-4}$); Module II is enriched in pathways for amino acid catabolism (e.g. REACTOME BRANCHED CHAIN AMINO ACID CATABOLISM; $p = 9.31 \times 10^{-5}$); Module III is enriched in defense related pathways (e.g. DEFENSE RESPONSE; $p = 2.11 \times 10^{-6}$); Module IV is enriched in pathways of metabolism of amino acids (e.g. REACTOME METABOLISM OF AMINO ACIDS; $p = 2.36 \times 10^{-3}$); Module V is enriched in stimulus related pathways (e.g. EXTERNAL STIMULUS; $p = 1.33 \times 10^{-3}$); For Module VI, we did not detect any significantly enriched pathways. Interestingly, all of these pathways are known to be related to different aspects of metabolism, which indicates that our method is able to detect homogeneous and heterogeneous gene modules that are biologically meaningful. Such biomarker categorizations may help enhance meta-analysis interpretations and motivate further biological hypotheses. For example, it is intriguing why the defense related genes in Module III are up-regulated only in heart but not in liver and brown fat, and why the stimulus related genes in Module V are down-regulated in heart and liver but not in brown fat. Among the six detected meta-pattern modules, 5 (83%) of them are biologically interpretable (defined as genes in the module are statistically enriched using Fisher's exact test with *P*-value < 0.005 in at least one pathway). In contrast, the biomarker categorization based on the unique AW weight estimate is hard to characterize as there are $3^3 - 1 = 26$ modules. In addition, among these modules only 13 (50%) are significantly enriched in at least one pathway with *P*-value < 0.005.

In order to benchmark the improvement of computational speed and feasibility, we also applied the original permutation approach (Li and Tseng, 2011) with number of permutation $B = 1000$. The entire analysis from raw data only took our proposed method 0.34 s while the permutation approach required 6.57 min.

The variability index helps characterize the instability of the AW weight estimate. Supplementary Table S7 listed 11 genes with similar *P*-values, but their AW weight estimates in the second study (heart) can be different since their AW variabilities are large. For example, *P*-values for probeset 1419484_*a*_*at* were (0.00039, 0.096, 0.0021) and its AW weight estimates were $\hat{w} = (1, 1, 1)$, while *P*-values for probeset 1425567_*a*_*at* were (0.00036, 0.10, 0.0021) and its AW weight estimates were $\hat{w} = (1, 0, 1)$.

The variability index of $\hat{w} = (1,\ 1,\ 1)$ in 1419484*_a_at* is (0, 0.93, 0) and variability index of $\hat{w} = (1,\ 0,\ 1)$ in 1425567*_a_at* is (0, 0.94, 0), showing unstable weight estimate of the second study for both gene probes.

In addition, we drew scatter plots of *P*-value and the variability index in Supplementary Figure S9. Similar to what has been observed in the simulation, with respect to the increase of $-\log_{10}(P - \text{value})$, the variability index goes up, and finally goes down to zero. This is also expected because the weight on the boundary (with moderate *P*-values) tends to be FP and FN, which has larger variability. The extreme significant *P*-values provide strong evidence to be TP, and thus have 0 variability.

Following the same procedure, we further applied the proposed method on all four tissues. 1073 DE genes were detected at FDR 5%. The resulting meta-pattern visualization is shown in Supplementary Figure S10. Among the top 10 detected meta-pattern modules, 8 (80%) of them were biologically interpretable. In contrast, the biomarker categorization based on the original AW weight estimate was intractable, since it produced $3^4 - 1 = 80$ modules. In addition, among these modules, only 26 (33%) were significantly enriched in at least one pathway with *P*-value < 0.005.

3.2.2 HIV transgenic rat RNA-seq data

Li *et al.* (2013) conducted studies to determine gene expression differences between F344 and HIV transgenic rats using RNA-seq (GSE47474 in Gene expression Omnibus database. The HIV transgenic rat model is designed to study learning, memory and vulnerability to drug addiction and other psychiatric disorders to HIV positive patients. Twelve F334 untreated rats and 12 HIV transgenic rats in prefrontal cortex (PFC), hippocampus (HIP) and striatum (STR) regions are sequenced for RNA-seq (see Supplementary Table S6b). Tophat (Trapnell *et al.*, 2009) was applied for alignment (adopted by Li *et al.*, 2013) and the alignment results were converted to RNA-seq count data with 16 821 genes by BEDTools (Quinlan and Hall, 2010). Genes with <100 total counts within any brain region were filtered out and 11 824 genes remained. Potential outliers were removed by checking the sample correlation heatmaps (see Supplementary Fig. S11). R package *edgeR* (Robinson *et al.*, 2010) was adopted to perform differential expression gene detection and two-sided *P*-values were obtained. AW-Fisher meta-analysis *P*-values were evaluated and *q*-values were obtained by applying the Benjamini–Hochberg procedure. By controlling FDR at 30%, we obtained 145 DE genes. We loosen the FDR criteria to 30% since it is well known that the transcriptomic signals in brain are generally weak. We calculated the variability index and performed biomarker categorization by using resampling techniques and the tight clustering algorithm. The result is shown in Supplementary Fig. S12. To examine the biological functions of these modules, we also performed pathway enrichment analysis using the same procedure as in Section 3.2.1. As the results show, Module I is up-regulated in all the three brain regions, and is enriched in pathways related to response to virus (e.g. GO RESPONSE TO VIRUS; $p = 1.59 \times 10^{-3}$); Module II is down-regulated in all the three brain regions, and is enriched in pathways related to rhythmic process (e.g. GO RHYTHMIC PROCESS; $p = 6.23 \times 10^{-4}$); Module III is especially interesting since it is down-regulated in HIP, but up-regulated in PFC and STR. It is enriched in transporter related pathways (e.g. REACTOME INORGANIC CATION ANION SLC TRANSPORTERS; $p = 3.67 \times 10^{-3}$). Since the brain is responsive for the virus invasion, we anticipate that genes responding to virus to be up-regulated, as observed in Module I. The down-regulation of rhythmic process genes in Module II indicates that HIV virus may have caused loss of rhythmic pattern in multiple brain regions. Moreover, because different brain regions have different functions, it is not surprising that transporter related genes (Module III) respond differently to HIV in different brain regions. Among the three detected meta-pattern modules, 3 (100%) of them were biologically interpretable. In contrast, the biomarker categorization based on the original AW weight estimate was difficult to characterize as it generated $3^3 - 1 = 26$ modules. In addition, among these modules, only 13 (50%) were biologically interpretable.

In addition, we drew scatter plots of *P*-value and the variability index in Supplementary Fig. S13. Similar to what has been observed in the simulation and the mouse metabolism data, we observed that with respect to the increase of $-\log_{10}(P - \text{value})$, the variability index goes up, and finally goes down to zero.

In order to benchmark the improvement of computational speed, we also applied the original permutation approach Li and Tseng (2011) with $B = 1000$. It only took our proposed method 18.0 s while it took the permutation approach 4.53 h.

4 Conclusion and discussion

Emerging omics datasets in the public domain have made genome-wide meta-analysis appealing. AW-Fisher has become useful and popular due to its capability to characterize study-specific contributions to the meta-analysis result. In this paper, we proposed novel methods to further generalize the application of AW-Fisher. The contributions of this paper are 3-fold. (i) We developed an AW-Fisher weight variability index. This is essential to determine the stability of AW-Fisher weight estimates. (ii) We proposed a biomarker categorization algorithm via a resampling procedure, which can efficiently obtain gene modules of different meta-analysis differential expression patterns. These meta-patterns can help establish biological hypotheses to quantify homogeneous and heterogeneous DE signals across studies. (iii) The previous version of the AW algorithm relied on permutation analysis to calculate *P*-values, which set a limitation for accuracy and computing speed. We proposed a fast computation and weight searching algorithm for the AW algorithm based on importance sampling, interpolation and a linear searching complexity of the AW weight, which makes the AW-Fisher algorithm more applicable for large-scale genomic applications. Finally, the superior performance of the proposed methods is demonstrated in extensive simulations and two real applications (mouse brain HIV data and mouse metabolism data).

One potential limitation of the AW-Fisher's method is that when sample sizes vary dramatically among different studies, it is possible that a DE gene from a study with larger sample size is more likely to have weight one than a study with smaller sample size, because under the alternative hypothesis, the significance level is also driven by the sample size. However, this is a fundamental issue for all *P*-value combination methods and a potential solution is by effect-size combination methods where effect sizes and sample sizes can be simultaneously considered in the meta-analysis.

Our current AW-Fisher's method has several potential extensions. First, the adaptive weight concept can be extended from Fisher's method to other *P*-value combination meta-analysis methods, such as Stouffer's method. The linear weight searching, importance sampling and spline smoothing can equally be applied in order to efficiently obtain accurate *P*-values (e.g. AW-Stouffer's method). Second, in addition to the AW-Fisher weight variance, another interesting estimator (π_{gk}) is the proportion of bootstrapping resamples of w_{gk} that are 1. π_{gk} is akin to stability selection (Meinshausen and Bühlmann, 2010), which could achieve potential consistent weight selection under certain assumptions.

An R package *AWFisher* (calling C++) is available at Bioconductor and GitHub (https://github.com/Caleb-Huo/AWFisher). All datasets and programing code used to perform all analyses in this paper are available in the Supplementary Material.

Acknowledgements

We also thank the reviewers for their many insightful suggestions to improve this paper.

Funding

Z.H. and G.T. are supported by National Institutes of Health [R01CA190766].

Conflict of Interest: none declared.

References

Begum,F. *et al.* (2012) Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Res.*, **40**, 3777–3784.

Benjamini,Y. and Heller,R. (2008) Screening for partial conjunction hypotheses. *Biometrics*, **64**, 1215–1222.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Methodol.*, **57**, 289–300.

Birnbaum,A. (1954) Combining independent tests of significance. *J. Am. Stat. Assoc.*, **49**, 559–574.

Chang,L.-C. *et al.* (2013) Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC Bioinformatics*, **14**, 368.

Domany,E. (2014) Using high-throughput transcriptomic data for prognosis: a critical overview and perspectives. *Cancer Res.*, **74**, 4612–4621.

Fang,Y. *et al.* (2019) Properties of adaptively weighted fisher's method. *arXiv preprint arXiv:1908.00583*.

Fisher,R.A. (1992) Statistical methods for research workers. In: *Breakthroughs in Statistics*. Springer, New York, NY, pp. 66–70.

Gibbons,S.M. *et al.* (2018) Correcting for batch effects in case–control microbiome studies. *PLoS Comput. Biol.*, **14**, e1006102.

Guerra,R. and Goldstein,D.R. (2009) *Meta-Analysis and Combining Information in Genetics and Genomics*. Chapman and Hall/CRC, New York.

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classif.*, **2**, 193–218.

Huo,Z. *et al.* (2019) Bayesian latent hierarchical model for transcriptomic meta-analysis to detect biomarkers with clustered meta-patterns of differential expression signals. *Ann. Appl. Stat.*, **13**, 340.

Li,J. and Tseng,G. (2011) An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *Ann. Appl. Stat.*, **5**, 994–1019.

Li,M.D. *et al.* (2013) Transcriptome sequencing of gene expression in the brain of the HIV-1 transgenic rat. *PLoS One*, **8**, e59582.

Littell,R. and Folks,J. (1971) Asymptotic optimality of Fisher's method of combining independent tests. *J. Am. Stat. Assoc.*, **66**, 802–806.

Luo,J. *et al.* (2010) A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *Pharmacogenomics J.*, **10**, 278.

Meinshausen,N. and Bühlmann,P. (2010) Stability selection. *J. R. Stat. Soc. Series B Stat. Methodol.*, **72**, 417–473.

Pan,W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **18**, 546–554.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Ramasamy,A. *et al.* (2008) Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.*, **5**, e184.

Robinson,M.D. *et al.* (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Roy,S. (1953) On a heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Stat.*, **24**, 220–238.

Simon,R. (2005) Development and validation of therapeutically relevant multi-gene biomarker classifiers. *J. Natl. Cancer Inst.*, **97**, 866–867.

Smyth,G.K. (2005) Limma: linear models for microarray data. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York, NY, pp. 397–420.

Soneson,C. and Delorenzi,M. (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, **14**, 1.

Song,C. and Tseng,G.C. (2014) Hypothesis setting and order statistic for robust genomic meta-analysis. *Ann. Appl. Stat.*, **8**, 777.

Stouffer,S. *et al.* (1949) *The American Soldier: Adjustment during Army Life*. Princeton University Press, Princeton.

Tippett,L. (1931) *The Methods of Statistics*. Williams Norgate Ltd, London.

Trapnell,C. *et al.* (2009) Tophat: discovering splice junctions with RNA-seq. *Bioinformatics*, **25**, 1105–1111.

Tseng,G. *et al.* (2012) Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.*, **40**, 3785.

Tseng,G.C. and Wong,W.H. (2005) Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. *Biometrics*, **61**, 10–16.

Wang,X. *et al.* (2012) An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. *Bioinformatics*, **28**, 2534–2536.