Submitted to the Annals of Applied Statistics
arXiv: arXiv:0000.0000

BAYESIAN LATENT HIERARCHICAL MODEL FOR TRANSCRIPTOMIC META-ANALYSIS TO DETECT BIOMARKERS WITH CLUSTERED META-PATTERNS OF DIFFERENTIAL EXPRESSION SIGNALS †

By Zhiguang Huo[1], Chi Song[2,*], and George Tseng[1,*]

[1] *University of Pittsburgh*
AND
[2] *The Ohio State University*

Due to rapid development of high-throughput experimental techniques and fast dropping prices, many transcriptomic datasets have been generated and accumulated in the public domain. Meta-analysis combining multiple transcriptomic studies can increase statistical power to detect disease related biomarkers. In this paper, we introduce a Bayesian latent hierarchical model to perform transcriptomic meta-analysis. This method is capable of detecting genes that are differentially expressed (DE) in only a subset of the combined studies, and the latent variables help quantify homogeneous and heterogeneous differential expression signals across studies. A tight clustering algorithm is applied to detected biomarkers to capture differential meta-patterns that are informative to guide further biological investigation. Simulations and three examples including a microarray dataset from metabolism related knockout mice, an RNA-seq dataset from HIV transgenic rats and cross-platforms prostate cancer datasets are used to demonstrate performance of the proposed method.

1. Introduction. With the rapid development of high-throughput experimental techniques and dropping prices, many transcriptomic datasets have been generated and deposited into public databases. In general, each dataset contains small to moderate sample size that brings caution of accuracy and reproducibility of detected biomarkers (Simon et al., 2003; Simon, 2005; Domany, 2014). Meta-analysis combining multiple transcriptomic studies can increase statistical power and provide robust conclusions from various platforms and sample cohorts (Ramasamy et al., 2008). Tseng, Ghosh and Feingold (2012) presented a comprehensive review of methods and applications in the microarray meta-analysis field and categorized ex-

*To whom correspondence should be addressed.
†Supported by NIH RO1CA190766 for Z.H. and G.T.
Keywords and phrases: transcriptomic differential analysis, meta-analysis, Bayesian hierarchical model, Dirichlet process

isting methods into combining p-values, combining effect sizes, direct merging (a.k.a mega-analysis) and combining non-parametric ranks. In general, meta-analysis can be viewed as information combination tools for adjusting batch effects (such as different experimental platform, protocols and bias) across studies to draw a more efficient and accurate conclusion. Since numerous known and unknown factors can aggregate to cause complicated batch effects in genomic studies, data normalization and then directly merging across studies is usually not appealing. Combining summary statistics such as p-values, effect sizes and ranks provides a more attractive alternative to avoid the need to normalize batch effect across studies. By combining p-values, analysts can use meta-data from different microarray or RNA-seq platforms with various forms of test statistics.

Following conventions in Birnbaum (1954) and Li and Tseng (2011), two hypothesis testings have been considered in meta-analysis. In the first setting (namely HS_A), we aim to detect biomarkers that are differentially expressed (DE) in all studies: $H_0 : \vec{\theta} \in \bigcap\{\theta_s = 0\}$ vs $H_A : \vec{\theta} \in \bigcap\{\theta_s \neq 0\}$, where θ_s is the effect size of study s, $1 \leq s \leq S$. Throughout this manuscript, effect size refers to unstandardized effect size (Cooper, Hedges and Valentine, 2009) (difference of group means or unstandardized regression coefficients). In the second setting (HS_B), targeted biomarkers are DE in one or more studies: $H_0 : \vec{\theta} \in \bigcap\{\theta_s = 0\}$ vs $H_A : \vec{\theta} \in \bigcup\{\theta_s \neq 0\}$. In view of overly stringent criterion in HS_A in noisy genomic data and when S is large, Song and Tseng (2014) proposed a robust setting HS_r, requiring that r or more studies are DE: $H_0 : \vec{\theta} \in \bigcap\{\theta_s = 0\}$ vs $H_A : \vec{\theta} \in \sum \mathbb{I}\{\theta_s \neq 0\} \geq r$, where $\mathbb{I}\{\cdot\}$ is an indicator function taking value one if the statement is true and zero otherwise and r is usually pre-estimated with $S/2 \leq r \leq S$. Song and Tseng (2014) also proposed to use the rth-ordered p-value (rOP, $T^{rOP} = p_{(r)}$) to test HS_r. Generally speaking, HS_A and HS_r are of the major biological interests to detect consensus biomarkers in the disease or phenotypic contrast. However, when heterogeneous differential expression signals across studies are expected (e.g. studies come from different tissues or brain regions in the two mouse/rat examples in Section 4), biomarkers detected from HS_B can be of interest. Chang et al. (2013) conducted a comparative study evaluating 12 popular microarray meta-analysis methods targeting on the three complementary hypothesis testings (HS_A, HS_B and HS_r).

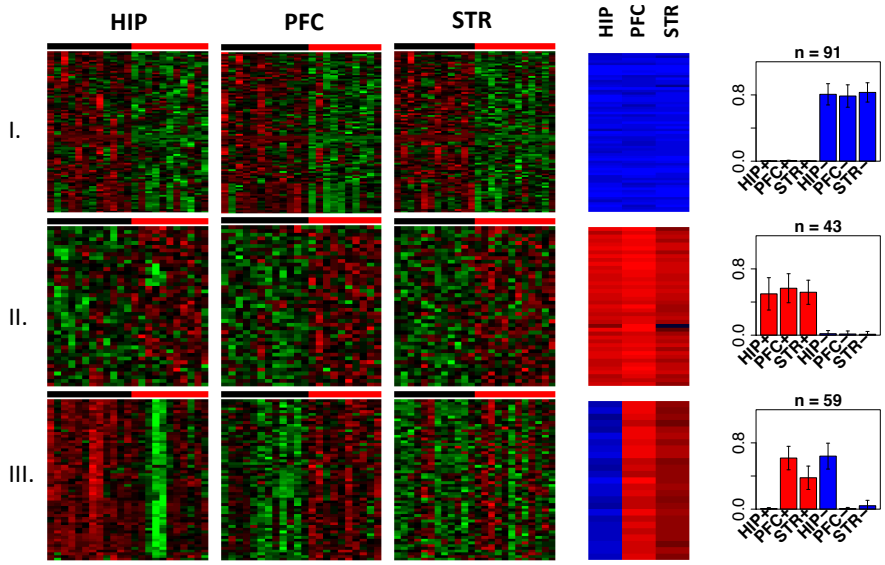
Strictly speaking, HS_B is a sound hypothesis testing, and statistical tests for HS_B are easier to develop. Most popular p-values aggregation methods such as Fisher (Fisher, 1934) and Stouffer (Stouffer et al., 1949) methods aim for this purpose. In the literature, HS_B is also called a conjunction or intersection hypothesis (Benjamini and Heller, 2008). On the other hand,

HS_A is a somewhat defective hypothesis testing. If we apply the maximum p-value test ($T^{maxP} = \max_s p_s$, where p_s is the p-value for study s) for HS_A and reasonably assuming that p-values are independently distributed with UNIF(0,1), the null distribution of T^{maxP} is Beta(S,1) and it is easily seen that the test is anti-conservative when there exist genes DE in partial studies. The problem mainly comes from the non-complementary null and alternative spaces in HS_A (and also HS_r) and the more adequate hypothesis testing should be $\text{HS}_{\bar{A}} \equiv H_0 : \vec{\theta} \in \bigcup\{\theta_s = 0\}$ vs $H_A : \vec{\theta} \in \bigcap\{\theta_s \neq 0\}$ (and $\text{HS}_{\bar{r}} \equiv H_0 : \sum \mathbb{I}\{\theta_s \neq 0\} < r$ vs $H_A : \sum \mathbb{I}\{\theta_s \neq 0\} \geq r$). Benjamini and Heller (2008) proposed a legitimate but conservative test for $\text{HS}_{\bar{A}}$ and $\text{HS}_{\bar{r}}$. In genomic application, the composite null hypotheses of $\text{HS}_{\bar{A}}$ and $\text{HS}_{\bar{r}}$ are complicated by the fact that genes can be differentially expressed in up to "$S-1$" (for $\text{HS}_{\bar{A}}$) or "$r-1$" (for $\text{HS}_{\bar{r}}$) studies with different levels of effect sizes. Under such a scenario, it becomes very difficult to characterize the null distribution for hypothesis test in a frequentist setting. Theoretically it is necessary to borrow differential expression information across genes to significantly improve statistical power. Bayesian hierarchical modeling can provide a convenient solution for this purpose. Efron et al. (2008) and Efron (2009) applied empirical Bayes methods to control false discovery rate (FDR) in single microarray studies. Muralidharan (2010) further extended these works to allow the simultaneous modeling of both empirical null and alternative distribution of the test stiatistics. Zhao, Kang and Yu (2014) incorporated pathway information to select genes using the Bayesian mixture model. Despite these successful applications, a Bayesian method that combines multiple studies and detect DE genes based on various meta-analysis hypothesis settings is yet to be developed. In this paper, we propose a Bayesian latent hierarchical model (BayesMP, named after Bayesian method for meta-patterns) that uses non-parametric Bayesian method to effectively combine information across genes for direct testing for $\text{HS}_{\bar{A}}$ (as well as HS_B and $\text{HS}_{\bar{r}}$) in genome-wide scale. In simulations, we show successful Bayesian false discovery rate control of BayesMP while the original maxP and rOP method using beta null distribution loses FDR control for the $\text{HS}_{\bar{A}}$ and $\text{HS}_{\bar{r}}$ setting.

Traditionally, meta-analysis aims to pool consensus signals to increase statistical power (e.g. by fixed or random effects models). Recently, researchers have recognized existence of heterogeneous signals among cohorts and the importance of their characterization in meta-analysis. For example, Figure 1(a) shows three modules of detected biomarkers from the RNA-seq HIV transgenic rat data using BayesMP and meta-pattern clustering (See Section 4.2 for details). Module I and II are consensus biomarkers that are

all down-regulated or all up-regulated across three brain regions. In contrast, the biomarkers in module III are down-regulated in HIP but up-regulated in PFC and STR. Such kind of biomarkers are somewhat expected biologically because it is well-known that different brain regions are responsible for different functions such as reasoning, recognition, visual inspection and memory/speech. Several approaches such as adaptive weighting (or subset) method (Li and Tseng, 2011; Bhattacharjee et al., 2012) or lasso variable selection (Li et al., 2014) have been proposed for quantifying and inferring such heterogeneity. In the adaptively weighted Fisher's method (AW-Fisher; Li and Tseng 2011), for example, heterogeneity of differential expression signals in each study is categorized by w_{gs} as 0 or 1 weights (1 representing differential expression for gene g in study s and 0 for non-differential expression). Specifically, AW-Fisher considers weighted Fisher's statistics (i.e. $T(\vec{W}_g) = -2\sum_{s=1}^{S} w_{gs} \cdot \log(p_{gs})$, where $\vec{W}_g = (w_{g1}, \ldots, w_{gS})$ is the vector of 0-1 weights reflecting gene-specific heterogeneous contribution of each study and p_{gs} is the p-value of gene g in study s) and adaptively searches the best weight vector for gene g by minimizing the resulting p-value: $\widehat{\vec{W}_g} = \arg\min_{\vec{W}_g} p(T(\vec{W}_g)) = \arg\min_{\vec{W}_g} 1 - F^{-1}_{\chi^2_{2\cdot\sum_{s=1}^{S} w_{gs}}}(T(\vec{W}_g))$, where $F^{-1}_{\chi^2_w}$ is the inverse CDF of chi-squared distribution with degree of freedom w. The 0-1 weight estimated from AW-Fisher helps cluster detected biomarkers by their differential expression meta-patterns but has a disadvantage of hard-thresholding without quantification of variability. When S is large, the number of all $2^S - 1$ possible weight combinations also makes it intractable. In BayesMP, the differential expression indicators naturally come with variability estimates from posterior distribution (see a confidence score to be defined later in Section 2.3). In BayesMP, we also adopt a cosine dissimilarity measure on these posterior distributions and apply tight clustering (Tseng and Wong, 2005) to identify biomarkers of different meta-patterns (e.g. see the three modules of biomarkers in Figure 1). Unsupervised clustering of the expression pattern across studies identifies data-driven modules of biomarkers of different meta-patterns and provides interpretable results for further biological investigation. For example, it is interesting to investigate why biomarkers in module III are down-regulated in HIP but up-regulated in PFC and STR. We note here that our proposed cluster analysis to categorize detected biomarkers by study heterogeneity in meta-analysis is a relatively novel concept. It is different from popular practices of clustering gene for identifying co-expression gene modules or clustering samples for discovering disease subtypes (e.g. Huo et al. (2016)).

The paper is structured as follows. Section 2 establishes the methodology,



(a) Heatmap (b) CS (c) bar plot

FIG 1. *Three meta-pattern modules of biomarkers from HIV transgenic rats example. Each row (Module I, II and III) shows a set of detected biomarkers showing similar meta-pattern of differential signals. 1(a) Heatmaps of detected genes (on the rows) and samples (on the columns) for each brain region (HIP, PFC or STR). Black color bar on top represents F334 rats (control) and red color bar on top represents HIV transgenic rats (case). 1(b) Heatmaps of confidence scores (CS) (genes on the rows and three studies on the columns). Confidence score is described in Section 2.3, which ranges from -1 (blue color for down-regulation) to 1 (red color for up-regulation). 1(c) Bar plots of mean posterior probability (error bar represents standard deviation across all genes in the module) of differential expression indicator in each brain region. Up-regulation is shown by red and down-regulation is shown by blue. Number of genes is shown on top of each barplot.*

estimation and inference of BayesMP. Section 3 evaluates the performance of the proposed method using simulation datasets. Section 4 shows the application to three real examples of multi-tissue microarray studies using metabolism related knockout mice, multi-brain-region RNA-seq studies using HIV transgenic rats, and cross-platform prostate cancer studies. Finally, Section 5 provides final conclusion and discussion.

2. Methods. For the ease of discussion, we focus on detecting DE genes in two-class comparison in this manuscript. The method can be easily extended for studies with numerical or survival outcomes. In a meta-analysis combining S studies with G genes, we denote p_{gs} as the one sided p-value testing for down-regulation for gene g in study s, where $1 \leq g \leq G$ and $1 \leq s \leq S$. These p-values can be calculated from SAM (Tusher, Tibshirani and Chu, 2001) or limma (Smyth, 2005) for microarray studies (or RNA-seq studies with RPKM data), and edgeR (Robinson, McCarthy and Smyth, 2010) or DEseq (Anders and Huber, 2010) for RNA-seq studies with count data. As a result, our model is flexible to mixed studies of different platforms (e.g. microarray or RNA-seq) and study designs (e.g. case-control, numerical outcome or survival outcome). Throughout this manuscript, we use limma and edgeR to obtain the p-values. For modeling convenience, we transform the one-sided p-values into Z-statistics, i.e. $Z_{gs} = \Phi^{-1}(p_{gs})$, where $\Phi^{-1}(\cdot)$ is the inverse cumulative density function (CDF) of standard Gaussian distribution. Z_{gs}'s are the input data for BayesMP.

2.1. *Bayesian hierarchical mixture model.* Denote by θ_{gs} the effect size of gene g in study s and Y_{gs} an indicator variable s.t. $Y_{gs} = 1$ if $\theta_{gs} > 0$ (up-regulation), $Y_{gs} = -1$ if $\theta_{gs} < 0$ (down-regulation) and $Y_{gs} = 0$ if $\theta_{gs} = 0$ (non-DE gene). We assume that the Z-statistics from study s are sampled from a mixture distribution with three mixing components depending on Y_{gs}: $f^{(s)}(Z_{gs}|Y_{gs}) = f_0^{(s)}(Z_{gs})\cdot\mathbb{I}(Y_{gs} = 0)+f_{+1}^{(s)}(Z_{gs})\cdot\mathbb{I}(Y_{gs} = 1)+f_{-1}^{(s)}(Z_{gs})\cdot\mathbb{I}(Y_{gs} = -1)$, where $f^{(s)}(\cdot)$ is the pdf of Z-statistics in study s, and $f_0^{(s)}$, $f_{+1}^{(s)}$ and $f_{-1}^{(s)}$ are the pdfs of the null, positive and negative components in study s.

In most situations, if an appropriate statistical test is adopted, one can expect that $p_{gs} \sim \text{Unif}(0, 1)$, hence reasonably assume $f_0^{(s)} \equiv \text{N}(0, 1)$. In case the p-value distribution is not uniform under null hypothesis, one can also empirically estimate $f_0^{(s)}$ following Efron (2004) . In our simulation and real data applications, we assume $f_0^{(s)} \equiv \text{N}(0, 1)$ because there is no evidence that this assumption is violated and estimating null component is not our major focus. Unlike $f_0^{(s)}$, $f_{\pm1}^{(s)}$ are usually unknown and their estimation is not trivial. To account for the uncertainty of $f_{\pm1}^{(s)}$, we model

them non-parametrically by assuming they are also mixtures of distributions using Dirichlet processes (DPs). DP is a Bayesian non-parametric method and we use DPs to capture the variability and uncertain of the alternative distribution of Z-statistics. DPs are widely discussed and applied in the literature (Neal, 2000; Müller and Quintana, 2004) and density estimation using DPs has also been discussed (Escobar and West, 1995). In our model, when $Y_{gs} \neq 0$, we assume $Z_{gs} \sim \mathrm{N}(\mu_{gs}, 1)$, and μ_{gs} follows distribution G_{s+} or G_{s-} generated from DPs. Specifically, the generative process of Z_{gs} given $Y_{gs} = \pm 1$ is as follow.

$$
\begin{aligned}
&G_{s+} \sim \mathrm{DP}(G_{0+}, \alpha_+) \text{ and } G_{s-} \sim \mathrm{DP}(G_{0-}, \alpha_-).\\
&\mu_{gs} \sim \begin{cases} G_{s+} \text{ if } Y_{gs} = 1,\\ G_{s-} \text{ if } Y_{gs} = -1. \end{cases}\\
&Z_{gs} \sim \mathrm{N}(\mu_{gs}, 1).
\end{aligned}
$$

$\mathrm{DP}(G, \alpha)$ denotes a DP with base distribution G and concentration parameter α, and G_{0+} (G_{0-}) denotes normal density $\mathrm{N}(0, \sigma_0^2)$ left (right) truncated at 0. We find that the selection of σ_0 and $\alpha_\pm$ do not much affect the performance of this model in simulation (see details in Supplementary Table S4). It should be noted that we assume $Z_{gs} \sim \mathrm{N}(\mu_{gs}, 1)$ where the variance is fixed at 1 to ensure that $f_{+1}^{(s)}/f_0^{(s)}$ ($f_{-1}^{(s)}/f_0^{(s)}$) is monotonically increasing (decreasing) while $Z_{gs} > 0$ ($Z_{gs} < 0$), which in turn guarantees the posterior probability of gene g being DE in study s to increase as $|Z_{gs}|$ increases (see Theorem 2.1). In addition, this assumption makes the MCMC simpler and hence speeds up the algorithm.

THEOREM 2.1. *If $f_k(x) \equiv N(\mu_k, \sigma_k^2)$ with $\mu_k > 0$, $\sigma_k^2 \geq 1$ and $1 \leq k \leq K$, and $f_0(x) = N(0, 1)$, then $\sum_{k=1}^K w_k f_k / f_0$ is monotonically increasing when $x \geq 0$, where $w_1, \ldots, w_K > 0$ and $\sum_{k=1}^K w_k = 1$.*

PROOF. $\forall 1 \leq k \leq K$, $f_k(x) = 1/(\sigma_k\sqrt{2\pi})\exp(-(x-\mu_k)^2/(2\sigma_k^2))$, $f_0(x) = 1/\sqrt{2\pi}\exp(-x^2/2)$, So $g_k(x) = f_k(x)/f_0(x) = 1/\sigma_k \exp(-(x-\mu_k)^2/(2\sigma_k^2)) + x^2/2)$ and $\log g_k(x) = -(x-\mu_k)^2/(2\sigma_k^2) + x^2/2 - \log\sigma_k$. Take derivative, $[\log g_k(x)]' = -(x-\mu_k)/\sigma_k^2 + x = (1 - 1/\sigma_k^2)x + \mu_k/\sigma_k^2 > 0$, when $x \geq 0$ (actually $x > \mu_k/(1-\sigma_k^2)$ is enough). Therefore $g_k = f_k/f_0$ is monotonically increasing when $x \geq 0$, and $\sum_{k=1}^K w_k f_k/f_0$ is also monotonically increasing when $x \geq 0$. □

In order to borrow information across studies, we further assume that Y_{gs} is generated depending on (1) the prior probability π_g that gene g is a DE

gene and (2) the conditional probability δ_g for gene g in study s being up-regulated (or $1-\delta_g$ for down-regulated) given gene g is DE. Specifically, we assume $\vec{W}_{gs} \sim \text{Mult}\left(1, (1-\pi_g, \pi_g^+, \pi_g^-)\right)$ and $Y_{gs} = \vec{W}_{gs} \cdot (0, 1, -1)$, where $\pi_g^+ = \pi_g \delta_g$, $\pi_g^- = \pi_g(1-\delta_g)$ and $\cdot$ is the inner product of two vectors. Given $Y_{gs} = y$, Z_{gs} is generated from $f_y^{(s)}(Z)$. The graphical representation of the full generative model is shown in Figure 2.

We assume that each gene g is DE in different studies in the same probability π_g, i.e. $\pi_g = \Pr(Y_{gs} \neq 0)$, and $\pi_g \sim \text{Beta}(\gamma/(G-\gamma), 1)$. The prior parameter γ can be interpreted as the expected number of DE genes, since the expectation of π_g from this prior is γ/G. Because there is no immediate conjugate prior for γ, we choose to obtain γ following Storey (2002) using pooled Z-statistics and the procedure is given in Section 2.5.

For each gene g, define $\delta_g = \Pr(Y_{gs} = 1 | \text{gene g is a DE gene})$. We assume $\delta_g \sim \text{Beta}(\beta, \beta)$. We set $\beta = 1/2$ in this paper, which gives a non-informative prior. Note that this conditional probability provides flexibility for a DE gene to contain conflicting differential expression directions (i.e. up-regulation in one study but down-regulation in another study; e.g. module III in Figure 1)

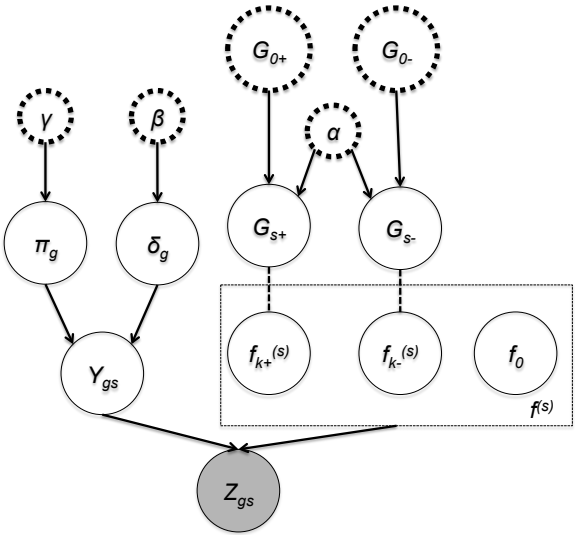


FIG 2. *Graphical representation of Bayesian latent hierarchical model. Shaded nodes are observed variables. Dashed nodes are pre-estimated/fixed parameters. Arrows represent generative process. Dashed lines represent equivalent variables. s is the study index and g is the gene index.*

2.2. *Model fitting.* Since conjugate priors were used in the generative model, we can generate the posterior samples efficiently using MCMC pro-

cedure. In order to update the DP with infinite number of components, we take the alternative view of DP as the Chinese restaurant process. We define $C_{gs} \in \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$ as the auxiliary component variable, s.t. $C_{gs} = 0$ indicates $Y_{gs} = 0$; $C_{gs} = k$ and $k > 0$ indicates that $Y_{gs} = 1$ and μ_{gs} is sampled from the kth component (or kth table in Chinese restaurant process) of G_{s+}; and $C_{gs} = k$ and $k < 0$ indicates that $Y_{gs} = -1$ and μ_{gs} is sampled from the k^{th} component of G_{s-}. The following steps provide details of our MCMC iterations.

1. Update π_g's:

$$\pi_g | Y_{gs} \sim \text{Beta}(\gamma/(G-\gamma) + Y_g^+ + Y_g^-, S - Y_g^+ - Y_g^- + 1),$$

   where $Y_g^+ = \sum_s \mathbb{I}(Y_{gs} = 1)$ and $Y_g^- = \sum_s \mathbb{I}(Y_{gs} = -1)$.
2. Update δ_g's:

$$\delta_g | Y_{gs} \sim \text{Beta}(\beta + Y_g^+, \beta + Y_g^-).$$

3. Update Y_{gs}'s:
   First update C_{gs}'s s.t.

$$\Pr(C_{gs} = k | C_{-g,s}, Z_{gs}, \pi_g^{\pm}) \propto h_k^{(s)}(Z_{gs}|C_{-g,s})(\pi_g^+)^{\mathbb{I}(k>0)}(\pi_g^-)^{\mathbb{I}(k<0)}(1-\pi_g)^{\mathbb{I}(k=0)},$$

   where $C_{-g,s}$ denotes all the C's in study s excluding gene g. Note that $h_k^{(s)}$ can be calculated directly following the convention of Algorithm 3 in Neal (2000). Details of $h_k^{(s)}(Z_{gs}|C_{-g,s})$ are given in Supplementary Section II. Finally, we set $Y_{gs} = \text{sgn}(C_{gs})$, where $\text{sgn}(\cdot)$ is the sign function.

2.3. *Decision space and inference making.* A main benefit of Bayesian modeling is its capability of making inference by statistical decision theory, a generalized framework that covers the traditional hypothesis testing framework as a special case (Berger, 2013). Take $\text{HS}_{\bar{A}}$ from Section 1 as an example. Traditional hypothesis testing considers null hypothesis $H_0 : \vec{\theta}_g \in \Omega^0_{\bar{A}}$ where $\Omega^0_{\bar{A}} = \{\vec{\theta}_g : \sum_{s=1}^S \mathbb{I}(\theta_{gs} \neq 0) < S\}$ and $\vec{\theta}_g = (\theta_{g1}, \cdots, \theta_{gS})$ versus alternative hypothesis $H_A : \vec{\theta}_g \in \Omega^1_{\bar{A}}$ where $\Omega^1_{\bar{A}} = \{\vec{\theta}_g : \sum_{s=1}^S \mathbb{I}(\theta_{gs} \neq 0) = S\}$. When observed data are unlikely to happen (i.e. type I error controlled at 5%) under null hypothesis, we reject the null hypothesis. One notable feature is that traditional hypothesis testing views null and alternative hypothesis spaces differently. The decision of hypothesis testing is only based on the null hypothesis - either reject or accept. The alternative hypothesis plays little role in decision making. In view of decision theory framework, a decision space (a.k.a. action space) is designed as $\mathcal{D}_{\bar{A}} = (\Omega^0_{\bar{A}}, \Omega^1_{\bar{A}})$. The inference

generates a decision function f that maps from the observed data space Z to $\mathcal{D}_{\bar{A}}$ (i.e. $f : Z \to \mathcal{D}_{\bar{A}}$). Under this framework, type I error can be expressed as $\Pr(f(Z) = \Omega^1_{\bar{A}} | \vec{\theta}_g \in \Omega^0_{\bar{A}})$; and type II error is $\Pr(f(Z) = \Omega^0_{\bar{A}} | \vec{\theta}_g \in \Omega^1_{\bar{A}})$. Hypothesis testing is a special case under this framework with adequate type I error control to determine the decision function f. Unlike hypothesis testing, decision theory treats $\Omega^0_{\bar{A}}$ and $\Omega^1_{\bar{A}}$ equally, because in decision theory, the decision is made through cost analysis, which weighs the costs of making wrong decisions in both spaces. One can easily design a realistic loss (cost) function based on the two types of errors to determine their balance and achieve the best decision function. In this paper in order to make fair comparison with classical hypothesis testing, we use posterior probabilities from Bayesian modeling and adopt a false discovery control described by Newton et al. (2004) to determine the decision function.

Denote by $\xi_g = \Pr(\vec{\theta}_g \in \Omega^0_{\bar{A}} | Z) = 1 - \Pr(\vec{\theta}_g \in \Omega^1_{\bar{A}} | Z)$, which is local FDR (Efron and Tibshirani, 2002) by defination. Given a threshold κ, we declare gene g as a DE gene if $\xi_g \leq \kappa$ and the expected number of false discoveries is $\sum_g \xi_g \mathbb{I}(\xi_g \leq \kappa)$. False discovery rate is defined as $\frac{\sum_g \xi_g \mathbb{I}(\xi_g \leq \kappa)}{\sum_g \mathbb{I}(\xi_g \leq \kappa)}$, which is referred to as global FDR (gFDR) in order to distinguish from the local FDR. In simulation and real applications, we compare performance of FDR control from traditional hypothesis testing and gFDR control from Bayesian modeling. Note that, we can consider $\mathcal{D}_B = (\Omega^0_B, \Omega^1_B)$ where $\Omega^0_B = \{\vec{\theta}_g : \sum_{s=1}^S \mathbb{I}(\theta_{gs} \neq 0) = 0\}$ and $\Omega^1_B = \{\vec{\theta}_g : \sum_{s=1}^S \mathbb{I}(\theta_{gs} \neq 0) > 0\}$ to correspond to HS_B and $\mathcal{D}_{\bar{r}} = (\Omega^0_{\bar{r}}, \Omega^1_{\bar{r}})$ where $\Omega^0_{\bar{r}} = \{\vec{\theta}_g : \sum_{s=1}^S \mathbb{I}(\theta_{gs} \neq 0) < r\}$ and $\Omega^1_{\bar{r}} = \{\vec{\theta}_g : \sum_{s=1}^S \mathbb{I}(\theta_{gs} \neq 0) \geq r\}$ to correspond to $\text{HS}_{\bar{r}}$.

Finally for a declared DE gene, we are given the posterior probability of whether gene g in study s is a non-DE gene ($\Pr(Y_{gs} = 0|Z)$), an up-regulated gene ($\Pr(Y_{gs} = 1|Z)$) or a down-regulated gene ($\Pr(Y_{gs} = -1|Z)$). We propose a gene and study specific confidence score $V_{gs} = \Pr(Y_{gs} = 1|Z) - \Pr(Y_{gs} = -1|Z)$, which ranges between -1 and 1. We are confident that gene g is up-regulated in study s if V_{gs} is close to one, and vise versa when V_{gs} is close to minus one. See Figure 1(b) for an example.

2.4. *Biomarker clustering for meta-patterns of homogenous and heterogenous differential signals.* Several recently developed meta-analysis methods (Li and Tseng, 2011; Bhattacharjee et al., 2012; Li et al., 2014) provide modeling of homogeneous and heterogeneous differential signals. The 0-1 differential expression indicators (e.g. $\vec{w}_g$ in AW-Fisher) allow further biological investigation on consensus biomarkers as well as study-specific biomarkers. However, when S becomes large, the number of biomarker categories grows exponentially to $(2^S - 1)$ and the biomarker categories become

intractable. In BayesMP, the posterior probability of the differential expression indicator (i.e. $\Pr(Y_{gs}|Z)$) provides probabilistic soft conclusions. After we obtain a list of biomarkers under certain global FDR control, we apply tight clustering algorithm (Tseng and Wong, 2005) to generate data-driven biomarker modules. Tight clustering is a resampling-based algorithm built upon K-means or K-medoids. This method aggregates information from repeated clustering of subsampled data to directly identify tight clusters (i.e. sets of biomarkers with close distances), and does not force every biomarker into a cluster. It can be applied to any dissimilarity matrix if K-medoids is used. Pathway enrichment analysis is then applied to functionally annotate each biomarker module. The resulting biomarker modules of different meta-patterns will greatly facilitate interpretation and hypothesis generation for further biological investigation. In the first two real data applications, for example, heterogeneous meta-patterns of biomarkers are expected from the nature of multi-tissue or multi-brain-region design across studies. Biomarkers up-regulated in one brain region but non-DE (or even down-regulated) in another brain region is of great interest. It should be noted that in the third prostate cancer data application, meta-patterns still help characterize the heterogeneity of different cohorts (e.g. differences of study population and probe design), even though this heterogeneity is hopefully minimal since these studies focus on the same disease with homogeneous tissue type.

To apply tight clustering, we need to define a dissimilarity measure for any pair of genes. Denote by $\vec{U}_{gs}$ the posterior probability vector for Y_{gs}: $\vec{U}_{gs} = (\Pr(Y_{gs} = 1|Z), \Pr(Y_{gs} = -1|Z), \Pr(Y_{gs} = 0|Z))$. For two genes i and j, we first calculate the dissimilarity of $\vec{U}_{is}$ and $\vec{U}_{js}$ in study s and then average over study index s. The dissimilarity measure between $\vec{U}_{is}$ and $\vec{U}_{js}$ we considered includes Cosine dissimilarity, l_2 dissimilarity, $l2_{2D}$ dissimilarity, Symmetric KL dissimilarity and Hellinger dissimilarity. Details of these dissimilarity measurement are described in Supplementary Section I. By using the simulation setting in Section 3.2, we found Cosine dissimilarity outperforms others (see details in Supplementary Figure S2), and hence adopt it in our paper and would recommend it for other applications.

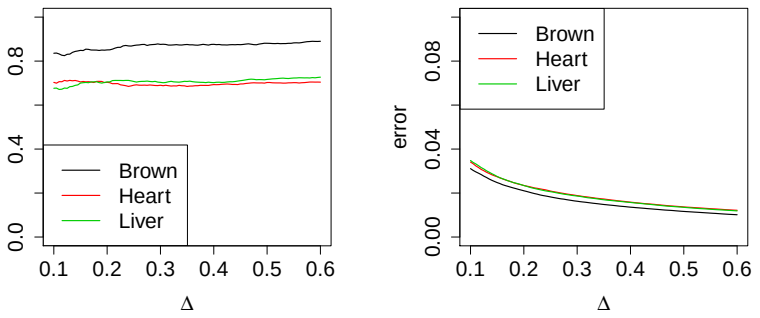
2.5. *Estimating* γ. The proportion of null component is estimated following the procedure by Storey (2002). Given a neighborhood $(-\Delta, \Delta)$ around 0, we estimate the null proportion in study s as

$$\widehat{\tau^{(s)}}(\Delta) = \frac{\sum_{g=1}^{G} \mathbb{I}(|Z_{gs}| < \Delta)}{G[2\Phi(\Delta) - 1]},$$

where $\Phi(\cdot)$ is the CDF of standard Gaussian distribution. Δ is a tuning parameter for this estimation. On one hand, larger Δ generates more conservative (positively biased) estimation; on the other hand, smaller Δ increases variability of the estimation. To facilitate the choice of Δ for bias-variance tradeoff, we can estimate the standard error of the mean (SEM) as

$$\widehat{\varepsilon^{(s)}}(\Delta) = \frac{\mathrm{se}(\widehat{\tau^{(s)}})}{\widehat{\tau^{(s)}}} = \sqrt{\frac{\sum_{g=1}^{G} \mathbb{I}(|Z_{gs}| \geq \Delta)}{G \sum_{g=1}^{G} \mathbb{I}(|Z_{gs}| < \Delta)}}.$$

We suggest users select a minimal Δ to avoid bias as long as $\widehat{\varepsilon^{(s)}}(\Delta)$ is reasonably small. Figure 3 shows an example from mouse metabolism data. Figure 3(a) and Figure 3(b) demonstrate the change of null proportion estimation and SEM respectively according to Δ for three different studies. In this example, we chose $\Delta = 0.4$, such that the SEM is smaller than 0.02 for all three studies. In the Bayesian hierarchical model in Section 2.1, the hyperparameter γ is determined by $\gamma = \frac{\sum_s \widehat{\tau^{(s)}}(\Delta)}{S}$.



(a) null proportion vs Δ (b) error vs Δ

FIG 3. *Parameter estimation and diagnosis result when estimating alternative distributions $f_{\pm 1}^{(s)}$ for multi-tissue microarray studies using metabolism related knockout mice. Figure 3(a) and Figure 3(b) demonstrate the change of null proportion estimation and SEM respectively according to Δ for the three tissues (Brown fat, Heart and Liver).*

3. Simulation results.

3.1. *DE gene detection and FDR control.* To evaluate the performance of the proposed method and compare to other methods, we performed the simulations below.

1. Let S be the number of studies, $G = 10,000$ be the total number of genes, and $N = 20$ be the number of cases and controls ($2N$ samples in total).
2. We firstly focus on simulating gene correlation structure and assume no effect size for all genes in all studies. Sample expression levels with correlated genes following the procedure in Song and Tseng (2014).
   (a) Sample 200 gene clusters with 20 genes in each cluster and the remaining 6,000 genes are uncorrelated. Denote by $C_g \in \{0, 1, \ldots, 200\}$ the cluster membership indicator for gene g, e.g. $C_g = 1$ indicates gene g is in cluster 1, whereas $C_g = 0$ indicates gene g is not in any gene cluster.
   (b) For cluster c and study s, sample $A'_{cs} \sim \mathrm{W}^{-1}(\Psi, 60)$, where $1 \leq c \leq 200$, $\Psi = 0.5I_{20\times 20} + 0.5J_{20\times 20}$, W^{-1} denotes the inverse Wishart distribution, I is the identity matrix and J is the matrix with all elements equal 1. Then A_{cs} is calculated by standardizing A'_{cs} such that the diagonal elements are all 1's. The covariance matrix for gene cluster c in study s is calculated as $\Sigma_{cs} = \sigma^2 A_{cs}$, where σ is a tuning parameter we vary in the evaluation.
   (c) Denote by $g_{c1}, \ldots, g_{c20}$ the indices of the 20 genes in cluster c (i.e. $C_{g_{cj}} = c$, where $1 \leq c \leq 200$ and $1 \leq j \leq 20$). Sample expression levels of genes in cluster c for sample n in study s as $(X'_{g_{c1}sn}, \ldots, X'_{g_{c20}sn}) \sim \mathrm{MVN}(0, \Sigma_{cs})$, where $1 \leq n \leq 2N$ and $1 \leq s \leq S$. For any uncorrelated gene g with $C_g = 0$, sample the expression level for sample n in study s as $X'_{gsn} \sim \mathrm{N}(0, \sigma^2)$, where $1 \leq n \leq 2N$ and $1 \leq s \leq S$.
3. Sample DE genes, effect sizes, and their differential expression directions.
   (a) Assume that the first G_1 genes are DE in at lease one of the combined studies, where $G_1 = 30\% \times G$. For each $1 \leq g \leq G_1$, sample v_g from discrete uniform distribution $v_g \sim \mathrm{UNIF}(1, \ldots, S)$ and then randomly sample subset $\mathbf{v}_g \subseteq \{1, \ldots, S\}$ such that $|\mathbf{v}_g| = v_g$. Here $\mathbf{v}_g$ is the set of studies in which gene g is DE.
   (b) For any DE gene g ($1 \leq g \leq G_1$), sample gene-level effect size $\theta_g \sim \mathrm{N}_{0.5+}(1, 1)$, where N_{a+} denotes the truncated Gaussian distribution within interval (a, ∞). For any $s \in \mathbf{v}_g$, also sample study-specific effect size $\theta_{gs} \sim \mathrm{N}_{0+}(\theta_g, 0.2^2)$.
   (c) Sample $d_g \sim \mathrm{Ber}(0.5)$, where $1 \leq g \leq G_1$ and $s \in \mathbf{v}_g$. Here d_g controls effect size direction for gene g.

4. Add the effect sizes to the gene expression levels sampled in Step 2c. For controls ($1 \leq n \leq N$), set the expression levels as $X_{gsn} = X'_{gsn}$. For cases ($N+1 \leq n \leq 2N$), if $1 \leq g \leq G_1$ and $s \in \mathbf{v}_g$, set the expression levels as $X_{gsn} = X'_{gsn} + (-1)^{d_g}\theta_{gs}$; otherwise, set $X_{gsn} = X'_{gsn}$.

We performed simulation with $S = 3, 5, 10$ and $\sigma = 1, 2, 3$ to account for different numbers of combined studies and various signal/noise ratio. We applied limma to compare the gene expression levels between the control group and case group. We transformed the two-sided p-values from limma to one-sided p-values by taking account of the directions of estimated effect sizes. Then one-sided p-values are transformed to Z statistics. We estimated the null proportion as described in Section 2.5, then BayesMP was applied to Z statistics. BayesMP took 60 minutes on a regular PC with 2.1GHz CPU (i.e. for one simulation with $S = 3$ and $\sigma = 1$) to obtain 10,000 posterior samples using MCMC. Supplementary Figure S1 shows the first 2,000 posterior samples of π_g in two example genes - a DE gene and a non-DE gene. Because the posterior samples converge to a stationary distribution very fast for our method (see examples in Supplementary Figure S1), excluding 500 posterior samples for burn-in is enough for the analyses of this paper. We repeated the simulation for 100 times and averaged the results. We compared the performance of our method and existing methods designed for decision space $\mathcal{D}_{\bar{A}}$ (maxP), $\mathcal{D}_B$ (Fisher's method and AW), and $\mathcal{D}_{\bar{r}}$ (rOP) with $r = \lfloor S/2 \rfloor + 1$ using both FDR and the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. Note that in our comparison, we used $\mathcal{D}_{\bar{A}}$ and $\mathcal{D}_{\bar{r}}$ which are equivalent to the complementary hypothesis testings $\text{HS}_{\bar{A}}$ and $\text{HS}_{\bar{r}}$, and the true number of studies in which a gene is DE can be calculated because the truth is known in simulation. Table 1 compares the actual FDR control and AUC of different methods. Supplementary Table S1 shows the number of genes declared as DE using these methods. The nominal FDR used in these methods is 5%, which is widely accepted in genomic research. For decision space $\mathcal{D}_B$, all the three methods controled FDR around its nominal level, which is anticipated because HS_B is complementary and equivalent to $\mathcal{D}_B$. Fisher and AW were slightly over-conservative in terms of FDR control - Fisher's method and AW controlled FDR at around 3.5%, whereas our BayesMP controlled FDR at around nominal 5%. This phenomenon has also been observed in Song and Tseng (2014) when the genes are correlated. Because BayesMP is less conservative than the other two methods, we were able to detect slightly more genes under $\mathcal{D}_B$. Besides BayesMP achieved similar (or slightly better) AUC with Fisher and AW under $\mathcal{D}_B$. Fisher' s method is known to be almost optimal (known as asymptotic Bahadur optimality (ABO) (Littell

and Folks, 1971)) when effect sizes are consistent and equal for all studies. This indicated BayesMP is also almost optimal for $\mathcal{D}_B$ under the simulation scenario. For decision space $\mathcal{D}_{\bar{A}}$ and $\mathcal{D}_{\bar{r}}$, we observed that maxP and rOP lost control of FDR. As discussed in Section 1, this is caused by the nature that HS_A and HS_r have non-complementary null and alternative spaces. To the contrary, BayesMP still controled FDR close to its nominal level for $\mathcal{D}_{\bar{A}}$ and $\mathcal{D}_{\bar{r}}$. Note that because maxP and rOP were not able to control FDR at its nominal level, the number of genes detected by these methods are not directly comparable to our methods. However, the number of genes detected by BayesMP was only slightly less than the number of genes detect by rOP for $\mathcal{D}_{\bar{r}}$, regardless of the conservative FDR control. When S was large ($S = 10$) in simulation, the FDR control of BayesMP under $\mathcal{D}_{\bar{A}}$ deviated from its nominal level ($\sim 9\%$ instead of 5%). The reason of the anti-conservative control was that the data simulation setting was different from model generative process, thus small errors accumulated when S gets large. However, BayesMP still performed much better than maxP and roP (FDR=0.58 for maxP in $\mathcal{D}_{\bar{A}}$ and FDR=0.23 for rOP in $\mathcal{D}_{\bar{r}}$ setting). In addition, BayesMP achieved much larger AUC than maxP and rOP under $\mathcal{D}_{\bar{A}}$ and $\mathcal{D}_{\bar{r}}$, which indicated better power of BayesMP.

3.2. *Simulation to evaluate meta-pattern gene module detection.* To evaluate the performance of gene module detection, we adopted a simulation procedure similar to Section 3.1. We simulated $S = 4$ studies in total. Among the $G = 10,000$ genes, we set 4% of them as homogeneously concordant DE genes, 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% of all genes 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 "ssp1−"). The rest of the genes are not DE (denoted as "nonDE"). The biological variance σ is set to 1 in this simulation.

We first applied the proposed method to this synthetic dataset. We controlled FDR at 5% under $\mathcal{D}_B$ and obtained 691 genes. These genes were used as input for our gene module detection using tight clustering algorithm. We identified 6 gene modules in these 691 genes. The detected gene modules

TABLE 1

Comparison of different methods by FDR and AUC of ROC curve for decision spaces $\mathcal{D}_{\bar{A}}$, $\mathcal{D}_B$, and $\mathcal{D}_{\bar{r}}$. The nominal FDR is 5% for all compared methods. The mean results and SD (in parentheses) were calculated based on 100 simulations.

			$\mathcal{D}_{\bar{A}}$		$\mathcal{D}_B$			$\mathcal{D}_{\bar{r}}$ $(r = \lfloor S/2 \rfloor + 1)$	
	S	σ	BayesMP	maxP	BayesMP	Fisher	AW	BayesMP	rOP
FDR	3	1	0.054	0.207	0.050	0.035	0.035	0.035	0.087
			(0.008)	(0.013)	(0.006)	(0.005)	(0.004)	(0.005)	(0.008)
		2	0.052	0.199	0.054	0.035	0.035	0.036	0.080
			(0.012)	(0.016)	(0.008)	(0.006)	(0.006)	(0.006)	(0.010)
		3	0.036	0.183	0.050	0.034	0.035	0.031	0.071
			(0.018)	(0.021)	(0.010)	(0.008)	(0.009)	(0.008)	(0.015)
	5	1	0.069	0.358	0.053	0.035	0.034	0.038	0.129
			(0.009)	(0.017)	(0.005)	(0.004)	(0.004)	(0.005)	(0.008)
		2	0.073	0.348	0.055	0.035	0.034	0.041	0.113
			(0.016)	(0.023)	(0.006)	(0.005)	(0.005)	(0.007)	(0.008)
		3	0.054	0.332	0.049	0.035	0.034	0.036	0.098
			(0.032)	(0.035)	(0.008)	(0.008)	(0.008)	(0.008)	(0.013)
	10	1	0.096	0.583	0.061	0.035	0.036	0.049	0.228
			(0.019)	(0.023)	(0.004)	(0.004)	(0.004)	(0.005)	(0.010)
		2	0.108	0.569	0.061	0.035	0.035	0.058	0.197
			(0.029)	(0.027)	(0.006)	(0.005)	(0.005)	(0.009)	(0.012)
		3	0.083	0.553	0.053	0.036	0.036	0.056	0.163
			(0.063)	(0.038)	(0.007)	(0.006)	(0.006)	(0.009)	(0.014)
AUC	3	1	0.977	0.926	0.973	0.973	0.973	0.980	0.972
			(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)
		2	0.907	0.875	0.879	0.877	0.875	0.902	0.873
			(0.006)	(0.007)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)
		3	0.831	0.805	0.787	0.783	0.779	0.819	0.776
			(0.008)	(0.008)	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)
	5	1	0.974	0.920	0.979	0.978	0.979	0.986	0.979
			(0.004)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
		2	0.920	0.890	0.897	0.894	0.892	0.929	0.893
			(0.007)	(0.006)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)
		3	0.864	0.833	0.812	0.806	0.801	0.859	0.802
			(0.009)	(0.009)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)
	10	1	0.964	0.910	0.985	0.983	0.985	0.986	0.986
			(0.007)	(0.003)	(0.001)	(0.002)	(0.001)	(0.002)	(0.002)
		2	0.910	0.905	0.920	0.917	0.917	0.950	0.919
			(0.011)	(0.006)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)
		3	0.875	0.863	0.848	0.839	0.834	0.904	0.838
			(0.013)	(0.010)	(0.004)	(0.004)	(0.005)	(0.005)	(0.006)

TABLE 2
Contingency table of simulation underlying truth and tight clustering result with 6 target modules. 0 represents the scattered gene group. 1 ∼ 6 represent 6 detected modules. Bolded numbers are genes with correct assignment.

Module	homo−	homo+	ssp1−	ssp1+	ssp2−	ssp2+	nonDE
1	**177**	0	0	0	0	0	0
2	0	**164**	0	0	0	0	0
3	0	3	0	0	0	**84**	4
4	0	4	0	**72**	0	0	7
5	1	0	0	0	**69**	0	3
6	0	0	**68**	0	0	0	2
0	6	4	0	0	0	0	**23**

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 detected gene modules clearly correspond to the true modules, and most of the nonDE genes were left to the noises. The heatmaps, confidence scores and DE patterns of these 6 modules are shown in Supplementary Figure S2. An alternative approach is to apply tight clustering directly on the Z-statistics. By comparing the results, we found that the modules detect by this naïve approach are neither pure nor distinguishable under our simulation settings (see details in Supplementary Table S2).

3.3. *Additional simulations.* To assess impact of unbalanced sample size, we simulated the following special scenarios with

1. Different numbers of samples in different studies, and
2. Different numbers of cases and controls in each study.

Below, we followed the simulation setting in Section 3.1 unless otherwise mentioned. In Scenario 1, we allowed different studies to have different numbers of samples. Under this Scenario, we simulated Case (a): numbers of samples (case/control) being 20/20, 30/30, 40/40 for 3 studies respectively; and Case (b): numbers of samples (case/control) being 20/20, 50/50, 100/100 respectively. In Scenario 2, we allowed the numbers of cases and controls being different within each study. Under this Scenario, we simulated Case (c): numbers of samples (case/control) being 60/20, 60/20, 60/20 for 3 studies respectively.

The results of simulation Cases (a)∼(c) are shown in Supplementary Table S3. We observed that under Scenario 1 and Scenario 2, BayesMP controled FDR to its nominal level for $\mathcal{D}_{\bar{A}}$, $\mathcal{D}_B$ and $\mathcal{D}_{\bar{r}}$. These results indicate that our Bayesian model is robust against impact of heterogeneity sample size in a wide spectrum of scenarios.

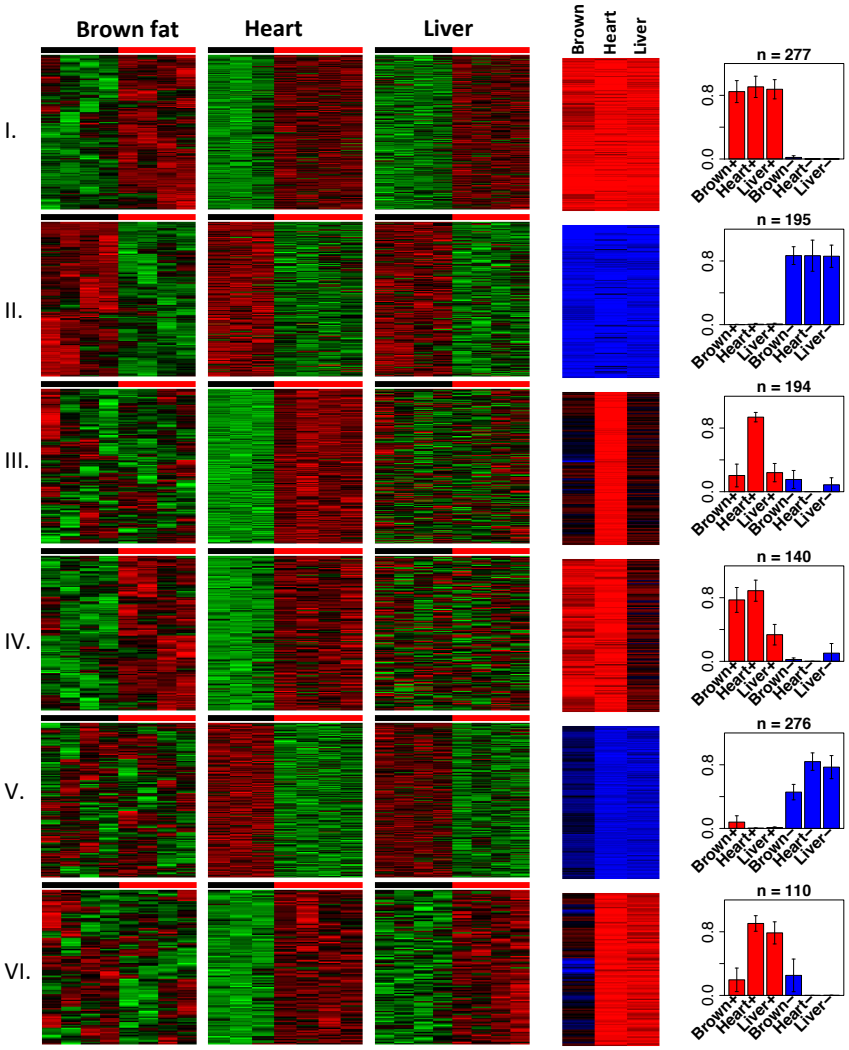
4. Real data applications. To further evaluate our method and demonstrate its usage, we applied BayesMP on three real meta-analysis examples: one on the gene expression of multi-tissue microarray studies using metabolism related knockout mice, one on multi-brain-region RNA-seq studies using HIV transgenic rats and another on transcriptomic prostate cancer studies across multiple platforms. The sample size description is shown in Supplementary Table S5.

4.1. *mouse metabolism data.* Very long-chain acyl-CoA dehydrogenase (VLCAD) deficiency was found to be associated with energy metabolism disorder in children (Li and Tseng, 2011). Two genotypes of the mouse model - wild type (VLCAD +/+) and VLCAD-deficient (VLCAD −/−) - were studied for three types of tissues (brown fat, liver and heart) with 3 to 4 mice in each genotype group. Total number of genes from these three transcriptomic microarray studies is 14,495. Supplementary Table S5(a) shows details of the study design and the data set is available in supplement. Two-sided p-values were calculated using limma by comparing wild-type versus VLCAD-deficient mice in each tissue and one-sided p-values were obtained by considering the effect size direction. We first estimated the null proportion (Shown in Figure 3) as described in Section 2.5. BayesMP took 111 minutes to obtain 10,000 posterior samples using MCMC and the first 500 posterior samples were excluded as burn-in iterations. By controlling gFDR at 5%, we detected 133 probes under $\mathcal{D}_{\bar{A}}$, among them 125 have concordant effect size directions in all the three tissues. The heatmap for the genes detected under $\mathcal{D}_{\bar{A}}$ is shown in Supplementary Figure S4.

Similarly, under $\mathcal{D}_B$ we obtained 1,701 DE genes at gFDR level of 5%. Then we applied the tight clustering algorithm based on the cosine distance as described in Section 2.4 to detect modules in these genes. The results are shown in Figure 4. Using the tight clustering, we were able to detect 6 gene modules with unique patterns. The first two biomarker modules are consensus genes that are up-regulated or down-regulated in all tissues. The next four modules are biomarkers with study-specific differential patterns. For example, DE genes in 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 were created by combining pathways from KEGG, BIOCARTA, REACTOME and GO databases and mapping all the human genes to their orthologs in mouse

using Jackson Laboratory Human and Mouse Orthology Report (http://www.informatics.jax.org/orthology.shtml). At Fisher's exact test FDR 5% cutoff, we summarized the pathway detection result (see supplementary Excel file 1 for detailed pathway information). Among the six gene modules with distinct DE patterns, module I is enriched in enzyme pathways (e.g. KEGG LYSOSOME; $q = 2.8 \times 10^{-4}$), module II is enriched in pathways for hemoglobin oxygen binding (e.g. BIOCARTA AHSP PATHWAY; $q = 0.017$), module III is enriched in immune response related pathways (e.g. DEFENSE RESPONSE; $q = 4.2 \times 10^{-8}$), module IV is enriched in cell cycle related pathways (e.g. BIOCARTA MCM PATHWAY; $q = 3.9 \times 10^{-3}$), module VI is enriched in phagocytosis pathways (e,g, KEGG FC GAMMA R MEDIATED PHAGOCYTOSIS; $q = 0.067$). For module V, we didn't detect any enriched pathways. Remarkably, 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. The biomarker clustering result enhances meta-analysis interpretation and motivates hypothesis for further biological investigation. For example, it is intriguing to understand why VLCAD-mutation impacts DE genes only up-regulated in heart but not in brown fat or liver and why these genes are associated with the DEFENSE RESPONSE pathway.

4.2. *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 http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE47474). The HIV transgenic rat model was designed to study learning, memory, vulnerability to drug addiction and other psychiatryic disorders vulnerable to HIV positive patients. They sequenced RNA transcripts with 12 F334 rats and 12 HIV transgenic rats in prefrontal cortex (PFC), hippocampus (HIP), and striatum (STR) brain regions (See detail in Supplementary Table S5(b)). We applied the same alignment procedure using tophat (Trapnell, Pachter and Salzberg, 2009) adopted by Li et al. (2013) and obtained the RNA-seq count data by bedtools (Quinlan and Hall, 2010) with 16,821 genes. We filtered out gene with less than 100 total counts within any brain region and ended up with 11,824 genes. We removed potential outliers by checking the sample correlation heatmaps (Supplementary Figure S5). We employed R package "edgeR" to perform DE gene detection and obtained two-sided p-values. The one-sided p-values were obtained by considering the effect size directions and further converted to Z statistics. We first estimated the null proportion as described in Section 2.5, then applied BayesMP to the Z statis-



(a) Heatmap (b) CS (c) bar plot

FIG 4. *Six meta-pattern modules of biomarkers from mouse metabolism example. Each row shows a set of detected biomarkers showing similar meta-pattern of differential signals. 4(a) Heatmaps of detected genes (on the rows) and samples (on the columns) for each tissue (brown fat, heart or liver). Black color bar on top represents wild type (control) and red color bar on top represents VLCAD-deficient mice (case). 4(b) Heatmaps of confidence scores (CS) (genes on the rows and three studies on the columns). Confidence score is described in Section 2.3, which ranges from -1 (blue color for down-regulation) to 1 (red color for up-regulation). 4(c) Bar plots of mean posterior probability (error bar represents standard deviation across all genes in the module) of differential expression indicator in each brain region. Up-regulation is shown by red and down-regulation is shown by blue. Number of genes is shown on top of each barplot. In Figure 4(b) and 4(c), we use "Brown" to denote "Brown fat".*

tics. It took 48 minutes to obtain 10,000 posterior samples via MCMC and the first 500 posterior samples were excluded as burn-in iterations. Since it is well known that the post-mortem brain expression profiles generally contain weak signals, we controlled global FDR at 20% in the analysis. Under $\mathcal{D}_{\bar{A}}$, we detected 73 genes, among which all 73 have concordant DE directions. The heatmaps of the expression levels (log of normalized counts) of these genes in the three brain regions are shown in Supplementary Figure S6. Under $\mathcal{D}_B$, we detected 383 genes. We further applied the tight clustering algorithm and obtained 3 gene modules. Their gene expression heatmaps, DE confidence scores and bar plots of posterior probability of differential expression are shown in Figure 1. To examine the biological functions of these modules, we also performed pathway enrichment analysis using the same procedure as in Section 4.1 (see supplementary Excel file 2 for detailed information). According to the results, module I is down-regulated in all the three brain regions, and is enriched in pathways related to immune system (e.g. REACTOME INNATE IMMUNITY SIGNALING; $p = 4.44 \times 10^{-4}$); module II is up-regulated in all the three brain regions, and is enriched in pathways related to activation of adenylate cyclase (e.g. REACTOME G ALPHA S SIGNALLING EVENTS; $p = 1.69 \times 10^{-3}$); module III is down-regulated in HIP, but up-regulated in PFC and STR, and it is enriched in pathways related to synapsis (e.g. SYNAPTIC TRANSMISSION $p = 2.75 \times 10^{-3}$) and neuron connections (e.g. KEGG NEUROACTIVE LIGAND RECEPTOR INTERACTION $p = 6.37 \times 10^{-3}$). Since it is well-known that HIV attacks the immune system (Weiss, 1993), we anticipate that genes for immune response to be down-regulated, as observed in module I. The up-regulation of adenylate cyclase activation we found in module II has been reported to mediate the increasing of cyclic AMP in neurons infected by HIV (Speth et al., 2002). Moreover, because different brain regions have different functions, it is not surprising to discover some neuron-related genes that may respond differently to HIV in different brain region (module III).

4.3. *Prostate Cancer dataset.* In this example, we combined 7 prostate cancer transcriptomic datasets, which study the same biological problem using different gene expression platforms including cDNA array, Affymetrix and RNA-seq. Details of these 7 datasets are described in Supplementary Table S5(c). After matching, 4,235 genes that appeared in all studies were selected for the analysis. For each study, we imputed missing values using K-nearest neighbor in R and obtained two-sided p-values using limma by comparing normal and cancer samples. Then we calculated one-sided p-values by considering direction of the effect sizes. We first estimated the null

proportion as described in Section 2.5, then applied BayesMP. It took 82 minutes to obtain 10,000 posterior samples using MCMC and the first 500 posterior samples were excluded as burn-in iterations.

Since we expect the studies combined in this application to be more homogeneous than previous examples, genes DE in most of the studies are our major interest and worth further investigation. Moreover, because the studies are conducted by different groups using different platforms, we want our analysis to be robust against a couple of studies with poor quality or unspecific probe design. Therefore we chose to use $\mathcal{D}_{\bar{r}}$ ($r = 5$). Under gFDR 0.05, we detected 974 significant genes. Top pathways associated with these genes are: STRUCTURAL MOLECULE ACTIVITY ($p = 5.12 \times 10^{-7}$), REACTOME TRANSLATION ($p = 4.54 \times 10^{-6}$). This result is reasonable because cancer disease may destroy the cell environment and result in molecular and translational behavior. We also applied rOP ($r = 5$) in this meta-analysis, and detected 1,862 genes under FDR 0.05 (following procedure of Benjamini and Hochberg, 1995). This excessive number of genes detected is anticipated because we have shown rOP being anti-conservative by simulation.

Although we are more interested in homogeneous biomarkers in this prostate cancer example, meta-pattern can still help characterize the heterogeneities between cohorts. Meta-pattern result of these 7 studies is shown in Figure 5, with the first 2 modules being homogeneous. In the other modules, we found that Nanni and Tomlins had unique DE pattern, which indicates potential heterogeneity of these studies. This is consistent with what was found in a quality control analysis (Kang et al., 2012), in which Nanni and Tomlins were identified to potentially have problematic quality. They were either old microarray studies or small sample sizes. This shows that BayesMP can help characterize the heterogeneities between cohorts and identify potential problematic studies, even though the studies are expected to be homogeneous.

5. Conclusion. For meta-analysis at genome-wide level, the issues to efficiently integrate information across studies and genes and to quantify homogeneous and heterogeneous DE signals across studies have brought new statistical challenges. Bayesian hierarchical model provides a feasible and effective solution. Compared to traditional hypothesis testing, decision theory framework from Bayesian modeling provides a more flexible inference to determine DE genes from meta-analysis. In this paper, we proposed a Bayesian hierarchical model for general transcriptomic meta-analysis. From posterior distribution of the latent variable (DE indicators Y_{gs}), global FDR is well-controlled and there is no need to select different test statistics for different hypothesis testing settings ($\text{HS}_{\bar{A}}$, HS_B and $\text{HS}_{\bar{r}}$). Post hoc clus-

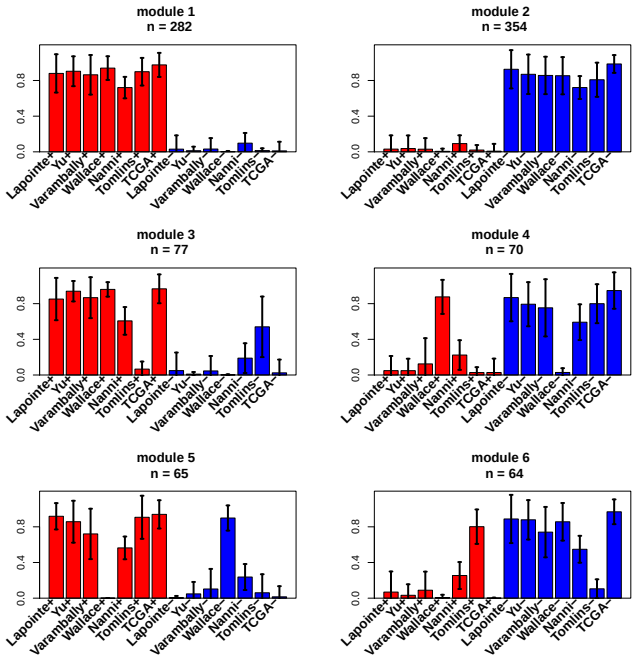


FIG 5. *Bar plots of mean posterior probability (error bar represents standard deviation across all genes in the module) 7 cross platform prostate cancer studies. Up-regulation is shown by red and down-regulation is shown by blue. Number of genes is shown on top of each barplot.*

tering analysis on the detected biomarkers generates biomarker modules of different meta-patterns that facilitate biological interpretation and provides clues for hypothesis generation and biological investigation.

Our proposed BayesMP framework has the following advantages. Firstly, the model is simple, yet practical and powerful. The model is based on one-sided p-values. This allows easy integration of data from different platforms (for example, many different platforms from microarray and RNA-seq). As a contrast, Scharpf et al. (2009) described a full Bayesian hierarchical model for microarray meta-analysis where the input data are microarray raw data (normalized intensities). Although such a full Bayesian model theoretically best integrates all information and can be more powerful, it cannot combine new RNA-seq platforms since RNA-seq generates count data versus continuous intensity measures in microarray. Such a full hierarchical model also runs a greater risk of model mis-specification that increases systemic bias across different microarray platforms. Our framework based on p-values circumvents these difficulties and is powerful as long as the method used to generate p-values in each study is effective. Secondly, we adopted a conjugate Bayesian approach using DPs for alternative distributions $f_{\pm 1}^{(s)}$. DPs is non-parametric thus robust against model assumption and conjugacy guarantee fast computing of our Gibbs sampling procedure. Thirdly, we have shown that decision theory framework from BayesMP provides good global FDR control and power under different hypothesis testings (or decision spaces). Fourthly, in contrast to the "hard" decision of 0-1 weights in AW-Fisher, the posterior distributions of the DE indicators Y_{gs} provides a stochastic quantification and "soft" decision. For example, in the mouse metabolism example, gene Mbnl2 (probeset 1422836_*at*) and gene Bcl2l11 (probeset 1435449_*at*) have very similar p-values in the three studies: $(0.0063, 0.16, 0.097)$ for Mbnl2 and $(0.0070, 0.16, 0.098)$ for Bcl2l11. Using AW, Mbnl2 ended up with a p-value of 0.020 with weights $(1, 0, 1)$ but Bcl2l11 had a p-value of 0.021 with weights $(1, 1, 1)$. Slight change of the p-value in the second study (heart tissue) resulted in different weights. The posterior probabilities for these two genes are, however, very similar, with $(0.827, 0.561, 0.671)$ and $(0.819, 0.561, 0.662)$ respectively and they belonged to the same gene module I in Figure 4. The stochastic quantification avoids sensitive 0-1 weight changes in AW-Fisher. Finally, traditional Fisher's method does not categorize DE genes with homogeneous or heterogeneous DE patterns across studies. In contrast, the improved AW-Fisher method allows categorization of biomarkers but generates up to $2^S - 1$ biomarker clusters, which becomes intractable when S is large. The posterior probability of Y_{gs} from BayesMP allows application of tight clustering to directly identify tight clusters of biomarkers with distinct

DE meta-patterns. Our simulation and three applications have shown good clustering accuracy and improved interpretation of the biomarker modules.

BayesMP potentially has the following potential limitations. Computing is often a consideration for Bayesian approaches. Our experiences have shown that 10,000 simulations are sufficient to generate the posterior probabilities in general and less than 1 hour is enough for combining around $10,000$ genes using a regular desktop. For applications to much larger number of features (e.g. SNPs or methylation sites in methyl-seq), parallel computing and/or faster MCMC techniques will be needed.

Efron (2004) pointed out that it might worth to estimate an empirical null distribution for the Z-statistics, when the null distribution deviates from $N(0,1)$. Therefore we highly recommend the user to estimate the empirical null distribution following Efron (2004), if any evidence suggests that the null distribution of p-values in a single study is not uniform. However, the user should also be careful about the assumptions made when estimating the empirical null distributions. For example, Efron (2004) assumes DE genes are less than a certain proportion (e.g. 10%), which is likely to be violated in our example of prostate cancer.

A prior for σ_c^2 can be given to better characterize the variability in σ_c^2 (e.g. truncated inverse gamma distribution, here truncation such that $\sigma_c^2 \geq 1$ is a sufficient condition such that density function of Z-statistics is monotone with respect to Z). But from computational point of view, such a prior will make the Bayesian procedure lose conjugacy. Thus we keep $\sigma_c^2 = 1$ and believe that the variability of σ_c^2 could be absorbed by the uncertain number of DP components.

BayesMP is implemented in R calling C++. The BayesMP package is publicly available at GitHub `https://github.com/Caleb-Huo/BayesMP` and the authors' websites.

References.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biol* **11** R106.

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. *Journal of the Royal Statistical Society. Series B (Methodological)* 289–300.

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Bhattacharjee, S., Rajaraman, P., Jacobs, K. B., Wheeler, W. A., Melin, B. S., Hartge, P., Yeager, M., Chung, C. C., Chanock, S. J., Chatterjee, N. et al. (2012). A subset-based approach improves power and interpretation for the combined

analysis of genetic association studies of heterogeneous traits. *The American Journal of Human Genetics* **90** 821–835.

Birnbaum, A. (1954). Combining independent tests of significance. *Journal of the American Statistical Association* 559–574.

Chang, L.-C., Lin, H.-M., Sibille, E. and Tseng, G. C. (2013). Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC bioinformatics* **14** 368.

Cooper, H., Hedges, L. V. and Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

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

Efron, B. (2004). Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *Journal of the American Statistical Association* **99** 96–104.

Efron, B. (2009). Empirical Bayes estimates for large-scale prediction problems. *Journal of the American Statistical Association* **104** 1015–1028.

Efron, B. et al. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical science* **23** 1–22.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic epidemiology* **23** 70–86.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association* **90** 577–588.

Fisher, R. A. (1934). Statistical methods for research workers.

Huo, Z., Ding, Y., Liu, S., Oesterreich, S. and Tseng, G. (2016). Meta-Analytic Framework for Sparse K-Means to Identify Disease Subtypes in Multiple Transcriptomic Studies. *Journal of the American Statistical Association* **111** 27–42.

Kang, D. D., Sibille, E., Kaminski, N. and Tseng, G. C. (2012). MetaQC: objective quality control and inclusion/exclusion criteria for genomic meta-analysis. *Nucleic Acids Research* **40** e15–e15.

Li, J. and Tseng, G. C. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics* **5** 994–1019.

Li, M. D., Cao, J., Wang, S., Wang, J., Sarkar, S., Vigorito, M., Ma, J. Z. and Chang, S. L. (2013). Transcriptome sequencing of gene expression in the brain of the HIV-1 transgenic rat. *PloS one* **8** e59582.

Li, Q., Wang, S., Huang, C.-C., Yu, M. and Shao, J. (2014). Meta-analysis based variable selection for gene expression data. *Biometrics* **70** 872–880.

Littell, R. C. and Folks, J. L. (1971). Asymptotic optimality of Fisher's method of combining independent tests. *Journal of the American Statistical Association* 802–806.

Müller, P. and Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical science* 95–110.

Muralidharan, O. (2010). An empirical Bayes mixture method for effect size and false discovery rate estimation. *The Annals of Applied Statistics* 422–438.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics* **9** 249–265.

Newton, M. A., Noueiry, A., Sarkar, D. and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5** 155–176.

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

Ramasamy, A., Mondry, A., Holmes, C. C. and Altman, D. G. (2008). Key issues in

conducting a meta-analysis of gene expression microarray datasets. *PLoS Med* **5** e184.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26** 139–140.

Scharpf, R. B., Tjelmeland, H., Parmigiani, G. and Nobel, A. B. (2009). A Bayesian model for cross-study differential gene expression. *Journal of the American Statistical Association* **104**.

Simon, R. (2005). Development and validation of therapeutically relevant multi-gene biomarker classifiers. *Journal of the National Cancer Institute* **97** 866–867.

Simon, R., Radmacher, M. D., Dobbin, K. and McShane, L. M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute* **95** 14–18.

Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor* 397–420. Springer.

Song, C. and Tseng, G. C. (2014). Hypothesis setting and order statistic for robust genomic meta-analysis. *The annals of applied statistics* **8** 777.

Speth, C., Schabetsberger, T., Mohsenipour, I., Stöckl, G., Würzner, R., Stoiber, H., Lass-Flörl, C. and Dierich, M. P. (2002). Mechanism of human immunodeficiency virus-induced complement expression in astrocytes and neurons. *Journal of virology* **76** 3179–3188.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 479–498.

Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A. and Williams Jr, R. M. (1949). *The American soldier: adjustment during army life.* Princeton Univ. Press.

Trapnell, C., Pachter, L. and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25** 1105–1111.

Tseng, G. C., Ghosh, D. and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research*.

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.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* **98** 5116.

Weiss, R. A. (1993). How does HIV cause AIDS? *Science* **260** 1273–1279.

Zhao, Y., Kang, J. and Yu, T. (2014). A Bayesian nonparametric mixture model for selecting genes and gene subnetworks. *The annals of applied statistics* **8** 999.

Zhiguang Huo
Department of Biostatistics
University of Pittsburgh
Pittsburgh, PA 15261
E-mail: zhh18@pitt.edu

Song Chi
Division of Biostatistics
College of Public Health
The Ohio State University
Columbus, OH 43210
E-mail: song.1188@osu.edu

Geroge Tseng
Department of Biostatistics, Human Genetics
and Computational Biology
University of Pittsburgh
Pittsburgh, PA 15261
E-mail: ctseng@pitt.edu