Siangphoe et al. BMC Bioinformatics (2019) 20:18
https://doi.org/10.1186/s12859-018-2491-9

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Classical and Bayesian random-effects meta-analysis models with sample quality weights in gene expression studies

Uma Siangphoe[1*], Kellie J. Archer[2] and Nitai D. Mukhopadhyay[3]

Abstract

Background: Random-effects (RE) models are commonly applied to account for heterogeneity in effect sizes in gene expression meta-analysis. The degree of heterogeneity may differ due to inconsistencies in sample quality. High heterogeneity can arise in meta-analyses containing poor quality samples. We applied sample-quality weights to adjust the study heterogeneity in the DerSimonian and Laird (DSL) and two-step DSL (DSLR2) RE models and the Bayesian random-effects (BRE) models with unweighted and weighted data, Gibbs and Metropolis-Hasting (MH) sampling algorithms, weighted common effect, and weighted between-study variance. We evaluated the performance of the models through simulations and illustrated application of the methods using Alzheimer's gene expression datasets.

Results: Sample quality adjusting within study variance (w_{P6}) models provided an appropriate reduction of differentially expressed (DE) genes compared to other weighted functions in classical RE models. The BRE model with a uniform(0,1) prior was appropriate for detecting DE genes as compared to the models with other prior distributions. The precision of DE gene detection in the heterogeneous data was increased with the DSLR2w_{P6} weighted model compared to the DSLw_{P6} weighted model. Among the BRE weighted models, the w_{P6}weighted- and unweighted-data models and both Gibbs- and MH-based models performed similarly. The w_{P6} weighted common-effect model performed similarly to the unweighted model in the homogeneous data, but performed worse in the heterogeneous data. The w_{P6}weighted data were appropriate for detecting DE genes with high precision, while the w_{P6}weighted between-study variance models were appropriate for detecting DE genes with high overall accuracy. Without the weight, when the number of genes in microarray increased, the DSLR2 performed stably, while the overall accuracy of the BRE model was reduced. When applying the weighted models in the Alzheimer's gene expression data, the number of DE genes decreased in all metadata sets with the DSLR2w_{P6}weighted and the w_{P6}weighted between study variance models. Four hundred and forty-six DE genes identified by the w_{P6}weighted between study variance model could be potentially down-regulated genes that may contribute to good classification of Alzheimer's samples.

Conclusions: The application of sample quality weights can increase precision and accuracy of the classical RE and BRE models; however, the performance of the models varied depending on data features, levels of sample quality, and adjustment of parameter estimates.

Keywords: Random-effects model, Bayesian random-effects model, Meta-analysis, Study heterogeneity, Gene expression, Sample quality weights, Alzheimer's disease

* Correspondence: siangphoeu@vcu.edu
This publication reflects the views of the author and should not be construed to represent FDA's views or policies.
[1]Office of Biostatistics, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, USA
Full list of author information is available at the end of the article

Background

Although modern sequencing technologies such as ribonucleic acid sequencing and next-generation sequencing have been developed, microarrays have been a widely used high-throughput technology in gathering large amounts of genomic data [1, 2]. Due to small sample sizes in single microarray studies, microarray studies are combined with meta-analytic techniques to increase statistical power and generalizability of the results [1, 3].

Common meta-analysis techniques applied in gene expression studies included combining of *p*-values, rank values, and effect sizes. Examples of the p-value based methods include Fisher's method, Stouffer's method, minimum *p*-value method, maximum p-value method, and adaptively weighted Fisher's method. The rank-based methods include *r*th ordered *p*-value method, naïve sum of ranks, naïve product of ranks, rank product, and rank sum methods. The effect-size based methods include fixed-effects (FE) and random-effects (RE) models.

Appropriateness of the meta-analysis techniques in gene expression data depends on types of hypothesis testing: HSA, HSB, or HSC as described in [4–6]. Maximum p-value and naïve sum of rank methods were appropriate for HSA hypothesis that detected DE genes across all studies. The *r*th ordered *p*-value method and two-step DerSimonian and Laird estimated RE models were appropriate for HSB hypothesis that detected DE genes in one or more studies. DerSimonian and Laird (DSL) and empirical Bayes estimated RE models, including our two-step estimated RE model using DSL and random coefficient of determination (R^2) method were appropriate for HSC hypothesis that detected DE genes in a majority of combined studies [4–6].

Some of these methods may be limited in their application. The p-value based methods are limited in reporting summary effects and addressing study heterogeneity [3, 7–9]. The rank-based methods are robust towards outliers and applied without assuming a known distribution [8, 10]; however, their results are dependent on the influence of other genes included in microarrays [1]. The FE model assumes that total variation is derived from a true effect size and a measurement error [3]; however, the effect may vary across studies in real-world applications. Concurrently, although the RE model can address study-specific effects and accounts for both within and between study variation, the between study variation or the heterogeneity in effect sizes is unknown. Many frequentist-based methods have been developed to estimate the between study variation. More details can be found in [6, 9, 11, 12].

The RE models are commonly applied in gene expression meta-analysis. Classical RE models assume studies are independently and identically sampled from a population of studies. However, an infinite population of studies may not exist and studies may be designed based on results of previous studies, thus potentially violating an independence assumption. Bayesian random-effects (BRE) models have been used to allow for uncertainty of parameters. The uncertainty is expressed through a prior distribution and a summary of evidence provided by the data is expressed by the likelihood of the models. Multiplying the prior distribution and the likelihood function results in a posterior distribution of the parameters [13, 14].

Sample quality has substantial influence on results of gene expression studies [15, 16]. The degree of heterogeneity may differ due to inconsistencies in sample quality. Low heterogeneity can be found in meta-analyses containing good quality samples, while high heterogeneity arises in meta-analyses containing poor quality samples. In our recent study, we evaluated the relationships between DE and heterogeneous genes in meta-analyses of Alzheimer's gene expression data. We detected some overlapped DE and heterogeneous genes in meta-analyses containing borderline quality samples, while no heterogeneous genes were detected in meta-analyses containing good quality samples [6]. Obviously, data obtained from borderline (poor) quality samples can increase study heterogeneity and reduce the efficiency of meta-analyses in detecting DE genes [17, 18].

In this study, we implemented a meta-analytic approach that includes sample-quality weights to take study heterogeneity into account in RE and BRE models. The gene expression data therefore would consist of up-weighted good quality samples and down-weighted borderline quality samples. Therefore in the Methods section we first review quality assessments of microarray samples, sample-quality weights, RE models, BRE models, weighted RE models, and weighted BRE models. We then describe our simulation studies and application data. Our results are then presented followed by discussion and conclusions.

Methods

This section describes quality assessments of microarray samples, sample-quality weights, RE models, BRE models, weighted RE models, and weighted BRE models.

Microarray quality assessments

Affymetrix GeneChips and Illumina BeadArrays have been widely used single channel microarrays. Quality assessments in Affymetrix arrays include the 3′:5′ ratios of two-control genes: beta-actin, and glyceraldehyde-3-phosphate dehydrogenase (GAPDH); the percent of number of genes called present; the array-specific scale factor; and the average background [15, 19]. A 3′:5′ ratio close to 1 indicates a good quality sample while a ratio > 3 suggests a poor quality sample, resulting from problems of RNA extraction, cDNA

synthesis reaction, or conversion to cRNA [15, 20]. Additionally, the percent present calls should be consistent among all arrays hybridized and generally should range from 30 to 60% [21]. The scale factor is used to assess overall expression levels with an acceptable value within 3-fold of one another. The proportion of up- and down-regulated genes should be consistent at the average signal intensity so that the expression among arrays can be comparable. The average background should also be consistent across all arrays [15]. For Illumina BeadArrays, quality assessments include the average and standard deviation of intensities, the detection rate, and the distance of specific probe intensities to the overall mean intensities of all samples [22–24].

Random-effects models

In this section, we provided a brief summary of the random-effects models implemented in this study. The hypothesis settings for detecting DE genes in meta-analysis of gene expression data are described in the supplemental material.

DerSimonian-Laird model (DSL)

An unbiased standardized mean difference in expression between groups (y_{ig}) can be obtained for each gene g as described in Hedges et.al. (1985) and Choi et.al. (2003) as:

$$y_{ig} = y'_{ig} - \frac{3y'_{ig}}{4(n_{ig}-2)-1},\quad y'_{ig} = \frac{\overline{x}_{ig(a)} - \overline{x}_{ig(c)}}{s_{ig}}, \tag{1}$$

$$s^2_{ig} = \frac{(n_{ig(a)}-1)s^2_{ig(a)} + (n_{ig(c)}-1)s^2_{ig(c)}}{n_{ig(a)} + n_{ig(c)}-2}, \tag{2}$$

where $\overline{x}_{ig(a)}$ and $\overline{x}_{ig(c)}$ represent the mean expression of case (a) and control (c) groups in ith study, $i = 1,\ldots,k$; $s_{ig}$and $n_{ig}$are an estimate of the pooled standard deviation across groups and the total sample size in the ith study; and$y_{ig}$is obtained as the correction for sample size bias. The estimated variance of y_{ig} is $\sigma^2_{ig} = (n^{-1}_{ig(a)} + n^{-1}_{ig(c)}) + y^2_{ig}(2(n_{ig(a)} + n_{ig(c)}))^{-1}$. The model of effect-size combination is based on a two-level hierarchical model:

$$\begin{aligned} y_{ig} &= \theta_{ig} + \varepsilon_{ig}, \quad \varepsilon_{ig} \sim N\left(0, \sigma^2_{ig}\right) \\ \theta_{ig} &= \beta_g + \delta_{ig}, \quad \delta_{ig} \sim N\left(0, \tau^2_g\right), \end{aligned} \tag{3}$$

where y_{ig} is the effect for gene g in ith study, $i = 1,\ldots,k$; θ_{ig} is the true difference in mean expression; σ^2_{ig} is the within-study variability representing sampling errors conditional on the ith study; β_g is the common effects or average measure of differential expression across datasets for each gene or the parameter of interest; $\delta_{ig}$is the random effect; and τ^2_g is the between-study variability. The RE model is defined when there is between-study variation [11, 25]. The estimator for τ^2_g is typically obtained using DerSimonian-Laird (DSL) estimator [26, 27] as

$$\hat{\tau}^2_{DSL(g)} = \max\left\{0, \frac{Q_g - (k_g - 1)}{S_{1g} - (S_{2g}/S_{1g})}\right\}, \tag{4}$$

where $Q_g = \sum_{i=1}^{k} w_{ig}(y_{ig} - \hat{\beta}_g)^2$, $w_{ig} = \sigma^{-2}_{ig}$, $\hat{\beta}_g = \frac{\sum_{i=1}^{k} w_{ig} y_{ig}}{\sum_{i=1}^{k} w_{ig}}$, $S_{rg} = \sum_{i=1}^{k} w^r_{ig}$, and $r = \{1, 2\}$. For each gene, we estimated $\hat{\beta}_g(\hat{\tau}^2_{DSL(g)})$ with $w_{ig} = (\sigma^2_{ig} + \hat{\tau}^2_{DSL(g)})^{-1}$ using a generalized least squares method to obtain statistics $z_{DSL(g)}$. More details can be found in [11, 25].

Two-step estimate model (DSLR2)

The $\hat{\tau}^2_{DSLR2(g)}$ was estimated by the DSL method in the first step and iterated with random-effect coefficients of determination ($R^2_{DSL(g)}$) in the second step. In other words, we assumed $\delta_{ig} \sim N(0, R^2_{DSL(g)})$ and replaced $\hat{\tau}^2_{DSL(g)}$ by $R^2_{DSL(g)}$ in the second-step estimation. $\hat{\tau}^2_{\mathrm{DSL}(g)}$ and $R^2_{\mathrm{DSL}(g)}$ are a function of $\tau^2(\mathbf{Y}_g - \boldsymbol{\beta}_g)$, so its bias does not influence the unbiasedness of the treatment and random effects [6, 12]. The $\hat{\tau}^2_{DSLR2(g)}$on the zero-to-one scale provides a lower minimum sum of squared error (MSSE) than the $\hat{\tau}^2_{DSL(g)}$ estimate. The $R^2_{DSL(g)}$ measuring the strength of study heterogeneity can also be used to compare variation of genes in different meta-analyses to decide which studies should be included in the meta-analysis [28]. The estimates of treatment effects, its variance, z-statistics, and random effects are obtained as

$$\hat{\beta}_g\left(R^2_{DSL(g)}\right) = \frac{\sum_{i=1}^{k}\left(\sigma^2_{ig} + R^2_{DSL(g)}\right)^{-1} y_{ig}}{\sum_{i=1}^{k}\left(\sigma^2_{ig} + R^2_{DSL(g)}\right)^{-1}}, \tag{5}$$

$$Var\left[\hat{\beta}_g\left(R^2_{DSL(g)}\right)\right] = \frac{1}{\sum_{i=1}^{k}\left(\sigma^2_{ig} + R^2_{DSL(g)}\right)^{-1}}, \tag{6}$$

$$z_g\left(R^2_{DSL(g)}\right) = \frac{\hat{\beta}_g\left(R^2_{DSL(g)}\right)}{\sqrt{Var\left(\hat{\beta}_g\left(R^2_{DSL(g)}\right)\right)}} \sim N(0,1), \tag{7}$$

$$\hat{\delta}_{ig}\left(R^2_{DSL(g)}\right) = \frac{R^2_{DSL(g)}}{\sigma^2_{ig} + R^2_{DSL(g)}}\left(y_{ig} - \hat{\beta}_g\left(R^2_{DSL(g)}\right)\right) \tag{8}$$

When compared to the DSL method, the DSLR2 method had a relatively better sensitivity and accuracy in detecting DE genes under HSC hypothesis testing and a higher precision when the proportion of truly DE genes

in the metadata was higher [6]. The DSLR2 method performed well with a low computational cost and almost all significantly DE genes identified were genes among the significantly DE genes identified using the DSL method. However, similar to the DSL method, the performance of the DSLR2 method can be reduced when sample sizes in single studies are restricted (e.g., < 60 in both arms) and the normality assumption of the meta-analysis outcome does not hold [6].

The RE models may be inefficient due to improper distributional assumptions. A permutation technique that is not based on a parametric distribution was applied to assess statistical significance of the common effect [11]. A modified BH method was used to control the FDR for multiple testing in the RE models [29]. We obtained the modified FDR by the order statistics of the actual and permuted z-statistics $z_{(g)} = (z_{(1)} \le \cdots \le z_{(G)})$ and $z^r_{(g)} = (z^r_{(1)} \le \cdots \le z^r_{(G)})$ as

$$FDR_g = \frac{(1/R)\sum_{r=1}^{R}\sum_{(g)=1}^{G} I\left(|z^r_{(g)}| \ge z_\alpha\right)}{\sum_{(g)=1}^{G} I\left(|z_{(g)}| \ge z_\alpha\right)}, \quad (9)$$

where α is the significance threshold of the single test, g is an index of genes *1,...,G*, and r is an index of permutation *1,...,R*.

Bayesian random-effects model (BRE)

The BRE models are different from the classical RE model in that the data and model parameters in the BRE models are considered to be random quantities [30]. The BRE models were used to allow for the uncertainty of the between-study variance in this study. The model for gene g is given by

$$\begin{aligned} & y_{ig} \,|\, \theta_{ig} \sim N\left(\theta_{ig}, \sigma^2_{ig}\right), \\ & \theta_{ig} \,\Big|\, \beta_g, \tau_g \sim N\left(\beta_g, \tau^2_g\right), \\ & \beta_g \sim N(0, 1000), \\ & \tau_g \sim uniform\,(a,b)\ and\ gamma\,(\alpha,\beta). \end{aligned} \quad (10)$$

The kernel of the posterior distribution can be written as

$$\begin{aligned} p(\beta_g, \theta_{1g}, \ldots, \theta_{kg}, \tau^2_g) &\propto p(\boldsymbol{\theta}_g|\mathbf{y}_g, \sigma^2_g)\, p(\beta_g, \tau^2_g|\boldsymbol{\theta}_g) \\ &\propto \prod_{i=1}^{k} p(\theta_{ig}|y_{ig}, \sigma^2_{ig})\, p(\theta_{ig}|\beta_g, \tau^2_g)\, \pi(\beta_g)\, \pi(\tau^2_g), \end{aligned} \quad (11)$$

where $\mathbf{y}_g = (y_{1g}, \ldots, y_{kg})$, $\boldsymbol{\sigma}^2_g = (\sigma^2_{1g}, \ldots, \sigma^2_{kg})$, and $\boldsymbol{\theta}_g = (\theta_{1g}, \ldots, \theta_{kg})$ for gene g in the ith study; $i = 1,...,k$. The $\pi(\beta_g)$ and $\pi(\tau^2_g)$ are non-informative priors given as $\beta_g \sim N(0, 1000)$, and τ_g~uniform (a,b) and gamma (α,β).

The choice of prior distributions for scale parameters can affect analysis results, particularly in small samples. With scale parameters, the distributional form and the location of the prior distributions are decided [31]. Uniform distributions are appropriate non-informative priors for τ^2_g [13]. We conducted simulations to select appropriate priors for τ^2_g, allowing the maximum (b) of the uniform distribution to be b∈{0.005, 0.001, 0.05, 0.01, 0.5, 0.1, 1, 5, 10} and b~Gamma(1,2). The potential choices of the appropriate priors were selected based on parameters obtained from an Alzheimer's gene expression data [6] in order to further apply the results.

Sample-quality weights

The quality control (QC) criteria indicative of poor quality samples we used were the 3′:5′ GAPDH ratio > 3 and/or percent of present calls < 30% for Affymetrix arrays; and detection rate < 30% for Illumina BeadArrays, in addition to data visualizations [15, 20]. Poor quality samples were excluded before data preprocessing. Theoretically, an optimal weight for meta-analysis is the inverse of the within-study variance. The variance of weighted mean ($\hat{\beta}_g$) is minimized when the individual weights are taken from the variance of the samples y_{ig}. A high variance therefore gives low weights in meta-analysis [32, 33]. In this study, the weights corresponding to the QC indicators fall into two categories: standardized ratio weights and zero-to-one weights (Table 1).

Standardized ratio weights ($w_{S,ij}$)

$$\begin{aligned} & S_{ij} = \left|\frac{R_{ij}-1}{SD(R_i)}\right| \in (0, \infty), \\ & w_{S,ij} = f\left(S_{ij}, \sigma^2_i, \tau^2\right), \end{aligned} \quad (12)$$

where R_{ij} is a quality indicator, i.e. 3′:5′ GAPDH ratio of the jth sample in the ith study, $SD(R_i)$is the standard deviation of the quality indicator in the ith study, $w_{S1-S3} \in (0, \infty)$, and $w_{S4-S8} \in (0, 1)$. $f(.)$ is a function of sample-quality weights with the within and between study variances as shown in Table 1. A low value of the S_{ij} indicates good quality samples, providing high values of standardized ratio weights ($w_{S,ij}$) to give more weight on the expression data.

Zero-to-one weights ($w_{P,ij}$)

$$\begin{aligned} & P_{ij} = \left\{ \begin{array}{c} \tilde{P}_{ij}(0.01) \\ 2^{-S_{ij}} \end{array} \right\} \in [0.01, \ldots, 1.0], \\ & w_{P,ij} = f\left(P_{ij}, \sigma^2_i, \tau^2\right), \end{aligned} \quad (13)$$

Table 1 List of sample quality weights

Standardized ratio weights ($w_{S,\,ij}$)	Zero-to-one weights ($w_{P,\,ij}$)	
$w_{S1} = (\sigma_g^2 + s_{ij}\hat{\tau}_g^2)^{-1}$	$w_{P1} \in \{2^{-s_{ij}}, 0.01\tilde{p}_{ij}\}$	$w_{P8} = 2^{-(\sigma_{ig}^2+(1-w_{P1})\hat{\tau}_g^2)}$
$w_{S2} = (s_{ij}\sigma_{ig}^2 + \hat{\tau}_g^2)^{-1}$	$w_{P2} = (\sigma_{ig}^2 + (1-w_{P1})\hat{\tau}_g^2)^{-1}$	$w_{P9} = 2^{-((1-w_{P1})\sigma_{ig}^2+\hat{\tau}_g^2)}$
$w_{S3} = (s_{ij}(\sigma_{ig}^2 + \hat{\tau}_g^2))^{-1}$	$w_{P3} = ((1-w_{P1})\sigma_{ig}^2 + \hat{\tau}_g^2)^{-1}$	$w_{P10} = 2^{-((1-w_{P1})(\sigma_{ig}^2+\hat{\tau}_g^2))}$
$w_{S4} = 2^{-(\sigma_{ig}^2+s_{ij}\hat{\tau}_g^2)}$	$w_{P4} = ((1-w_{P1})(\sigma_{ig}^2 + \hat{\tau}_g^2))^{-1}$	$w_{P11} = 2^{-(\sigma_{ig}^2+\hat{\tau}_g^{2(w_{P1})})}$
$w_{S5} = 2^{-(s_{ij}\sigma_{ig}^2+\hat{\tau}_g^2)}$	$w_{P5} = (\sigma_{ig}^2 + \hat{\tau}_g^{2(w_{P1})})^{-1}$	$W_{P12} = 2^{-(\sigma_{ig}^{2(w_{P1})}+\hat{\tau}_g^2)}$
$w_{S6} = 2^{-(s_{ij}(\sigma_{ig}^2+\hat{\tau}_g^2))}$	$w_{P6} = (\sigma_{ig}^{2(w_{P1})} + \hat{\tau}_g^2)^{-1}$	$w_{P13} = 2^{-((\sigma_{ig}^2+\hat{\tau}_g^2)w_{P1})}$
	$w_{P7} = ((\sigma_{ig}^2 + \hat{\tau}_g^2)^{(w_{P1})})^{-1}$	

where $\tilde{P}_{ij}$ and S_{ij} is the percent of present calls and the standardized quality indicators of the jth sample in the ith study, respectively, $w_{P1-P7} \in (0, \infty)$, and $w_{P8-P13} \in (0, 1)$. A high value of the P_{ij} weights indicate good quality samples, providing high values of zero-to-one weights ($w_{P,ij}$) to give more weight on the expression data.

The weights are primarily selected based on availability of quality indicators, such as 3′:5′ GAPDH ratio in Affymetrix arrays or detection rate in Affymetrix arrays and Illumina BeadArrays. Both the 3′:5′ GAPDH ratio and detection rate can be converted to the zero-to-one weights via w_{P1}.

Weighted random-effects models

An appropriate weight was chosen based on the precision and accuracy of the DSL weighted and DSLR2 weighted models in detecting DE genes via simulations and were used to weight the expression data and to adjust the common effect and the between-study variance in the BRE model.

Weighted DSL and DSLR2 models

The $\log_2$ normalized intensity data were weighted with an appropriate weight obtained from the DSL and DSLR2 weighted models. The weighted mean $(\bar{x}_{ig(a)})$ and weighted sample variance $(s_{ig(a)}^2)$ of the normalized intensity data in each group were calculated:

$$\bar{x}_{ig(a)} = \sum_{j=1}^{n_{ig(a)}} w_{ijg(a)} x_{ijg(a)} / \sum_{j=1}^{n_{ig(a)}} w_{ijg(a)}, \tag{14}$$

$$s_{ig(a)}^2 = \frac{\sum_{j=1}^{n_{ig(a)}} w_{ijg(a)} \left(x_{ijg(a)} - \bar{x}_{ig(a)}\right)^2}{S_{1g(a)} - \left(S_{2g(a)}/S_{1g(a)}\right)}; \tag{15}$$

$$S_{rg(a)} = \sum_{j=1}^{n_{ig(a)}} w_{ijg(a)}^r, \quad r = \{1, 2\},$$

$x_{ijg(a)}$ is the $\log_2$ normalized intensity data for gene g of the jth sample in the case (a) group and in the ith study, $n_{ig(a)}$ is the sample size of case (a) group for gene g in the ith study, and $w_{ijg(a)}$ is the sample-quality weight of the jth sample in the case (a) group in the ith study for the gene g. The same calculations were applied for the weighted mean $(\bar{x}_{ig(c)})$ and the weighted sample variance $(s_{ig(c)}^2)$ in the control (c) group. The unbiased standardized mean difference of the expression between groups were re-calculated and re-combined using the DSL and DSLR2 models (Eq.1 and Eq.2).

Weighted common effect model

We adjusted the common effect in the BRE model (Eq.10) by multiplying with an average weight over the total sample in the ith study for gene $g(\bar{w}_{ig} = \sum_{j=1}^{n_{ig(a)}+n_{ig(c)}} w_{ijg}/(n_{ig(a)} + n_{ig(c)}))$. The BRE weighted common effect model for gene g is given by

$$\begin{aligned} &y_{ig} |\ \theta_{ig} \sim N(\theta_{ig}, \sigma_{ig}^2), \\ &\theta_{ig} |\ \beta_g \bar{w}_{ig}, \tau_g \sim N(\beta_g \bar{w}_{ig}, \tau_g^2), \\ &\beta_g \sim N(0, 1000), \\ &\tau_g \sim \text{uniform}(a, b) \text{ and } \text{gamma}(\alpha, \beta) \end{aligned} \tag{16}$$

Weighted between-study variance model

We adjusted the between-study variance in the BRE model (Eq.10) by multiplying with an average weight over the total sample in the ith study for gene g $(\bar{w}_{ig} = \sum_{j=1}^{n_{ig(a)}+n_{ig(c)}} w_{ijg}/(n_{ig(a)} + n_{ig(c)}))$. The BRE weighted between-study variance model for gene g is given by

$$\begin{aligned} &y_{ig} |\ \theta_{ig} \sim N(\theta_{ig}, \sigma_{ig}^2), \\ &\theta_{ig} |\ \beta_g, \tau_g \bar{w}_{ig} \sim N(\beta_g, \tau_g^2 \bar{w}_{ig}), \\ &\beta_g \sim N(0, 1000), \\ &\tau_g \sim \text{uniform}(a, b) \text{ and } \text{gamma}(\alpha, \beta) \end{aligned} \tag{17}$$

Example WinBUGS code appears in the supplemental material.

The weighted common effect and the weighted between study variance in the BRE models with a uniform(0,1) prior were implemented in both unweighted and weighted data using Gibbs and Metropolis-Hasting (MH) sampling algorithms [14, 34]. Two chains each with 20,000 iterations, a 15,000 burn-in period, and a thinning of 3 was performed for all Bayesian models. The convergence of the models was assessed using the Gelman and Rubin diagnostic [34]. Since the posterior distribution was normal and symmetric, the posterior mean was standardized by posterior standard deviation. A Benjamini and Hochberg (BH) procedure was applied to control the false discovery rate (FDR) for multiple gene testing, so that the BRE and classical RE models could be compared throughout the study. Seven BRE models for unweighted and weighted data, Gibbs and MH sampling algorithms, weighted common effect, and weighted between-study variance were implemented as shown in Table 2.

The DE genes were defined as those with FDR less than 5%. Unsupervised hierarchical clustering using Ward's method and one minus Pearson's correlation coefficient for measures of similarities were used to graphically present the DE genes in the individual analysis of Alzheimer's gene expression data using a heatmap.

Simulation setting

Simulated datasets were generated using an algorithm described in previous studies [4–6]. A brief summary of the algorithm is as follows:

1. Five studies each with 2000 genes were generated (800 clustering and 1200 non-clustering genes). The clustering genes with the same correlation pattern within their clusters were equally allocated into 40 clusters.
2. Gene expression levels among clustering and non-clustering genes were assumed to follow a multivariate normal distribution $(X'_{gc1}, \ldots, X'_{gc40})^T \sim MV N(0, \Sigma_{ck})$, $1 \le k \le 5$, $1 \le c \le 40$, $\sum'_{ck} \sim W^{-1}(\psi, 60)$, and $\psi = 0.5I_{20\times 20} + 0.5J_{20\times 20}$, and a standard normal distribution, respectively.
3. Truly DE genes were generated with uniform(0.5,3), accounted for 10% of the total genes, and equally classified into 5 groups ($t_g = 1, \ldots, 5$). On average each group included 200 true genes. As the RE models appropriated under HSC, 120 genes in more than 50% of the combined studies were defined as the truly DE genes.
4. Truly heterogeneous genes constituted 15% of the total genes, implied by the random effects with uniform(0.5,3), and proportionally allocated into truly DE and not truly DE gene groups. The heterogeneous gene was defined by a significant random effect, where the gene expression was not identical across studies.
5. Sample-quality weights were assumed to follow beta distributions($\alpha = 10, \beta = 1$) for the zero-to-one weights and normal distributions $N(0, 0.6)$ for the standardized ratio weights.

Table 2 Bayesian random-effects (BRE) models by data features, sampling algorithms, and weighted inference models

	BRE Models						
	1	2	3	4	5	6	7
Data features							
Unweighted normalized intensity data	✓	✓	✓				
Weighted normalized intensity data				✓	✓	✓	✓
Sampling algorithms							
Gibbs sampling	✓	✓	✓	✓	✓	✓	
Metropolis-Hasting sampling							✓
Weighted inference models							
Unweighted model	✓			✓			✓
Weighted common effect		✓			✓		
Weighted between-study variance			✓			✓	

The N, G, K, and H denote the number of samples, the number of genes, the number of studies, the number of studies containing heterogeneous genes, respectively, all of which varied in different simulations. Because the simulation results under the same algorithms on 2000 and 10,000 genes were similar [6] and implementing Bayesian models requires intensive computations, we conducted the simulations on 2000 genes. Eight simulated metadata sets: two sets for the weighted and unweighted methods in the homogeneous data (H0), and each two of six sets for the weighted and unweighted methods in the heterogeneous data (H1, H2, and H3) were generated. A thousand simulations each with 1000 permutations of group labels were implemented for all DSL and DSLR models, and without permutation for the BRE models with different uniform(0,b) priors; b∈{0.005, 0.001, 0.05, 0.01, 0.5, 0.1, 1, 5, 10, and 100}, and b~Gamma(1,2) prior.

Evaluations of methods in simulations

Because RE models were suitable under HSC hypothesis: detecting DE genes in a majority of combined studies [5, 6], the models were anticipated to detect DE genes in more than 50% of combined studies, $r = 3$ for meta-analysis of five studies. We evaluated the number of detected DE genes, minimum sum squared error (MSSE), precision, accuracy, and area under receiver operating characteristic curve (AUC). Precision was

calculated as the proportion of truly DE genes correctly identified as significant over the total number of genes declared significant. Accuracy was calculated as the proportion of genes correctly identified as being truly DE genes or not truly DE genes over the total of evaluated genes. The accuracy of the tests was also determined using AUC, where $AUC \in (0.5, 0.7]$, $AUC \in (0.7, 0.9]$ and $AUC \in (0.9, 1.0]$ represent low, moderate, and high accuracy, respectively [35, 36]. All statistical methods and simulations were implemented using programs and modified programs from *limma, metafor, GeneMeta, MAMA, Rjags, R2jags, Coda* in the R programming environment.

Four publicly available Alzheimer's disease (AD) gene expression datasets of post-mortem hippocampus brain samples were applied: GSE1297 [37], GSE5281 [38], GSE29378 [39], and GSE48350 [40]. After data preprocessing, quantile normalization, and data aggregating [20, 41–44], our meta-analysis was performed on 12,037 target genes in 131 subjects (68 AD cases and 63 controls). We examined the strength of study heterogeneity by considering five ways of metadata sets as previously described in [6] and defined in the caption of Figs. 5 and 6. The metadata A, B, D, E may contain heterogeneous data due to a relatively high R^2, while the metadata C had a relatively low R^2or contained homogenous data. The 3′:5' GAPDH ratio was used as a quality indicator in this analysis. The 3′:5' GAPDH ratio was converted to the zero-to-one weight, w_{P6}, via w_{P1}.

Results

Table 3 presents the performance of the DSL and DSLR2 models, and the BRE models with different prior distributions. All of the BRE models converged with the potential scale reduction factor close to 1. The BRE model with a uniform(0,1) prior detected more DE genes than the DSL and DSLR2 models. The BRE model with a uniform(0,b) prior where b = {0.001, 0.01, 0.1, 0.005, 0.05, 0.5} detected too many DE genes, particularly in the heterogeneous data, while the BRE model with a uniform(0,5), uniform(0,10), uniform(0,100), and gamma(1,2) prior detected too few DE genes. The DSLR2 model had the lowest MSSE, while the DSL model and the BRE model with a uniform(0,1) prior had similar MSSEs (Additional file 1: Figure S1). In addition, the DSL, DSLR2, BRE with a uniform(0,1) prior detected DE genes with high precision in the homogeneous data, moderate precision in the heterogeneous data, and high accuracy in all datasets. The DSLR2 and BRE with a uniform(0,1) prior had a higher AUC than the DSL model in the heterogeneous data (Fig. 1).

Therefore, the DSLR2 and BRE models with a uniform(0,1) prior were appropriate for detecting DE genes in terms of an appropriate number of DE genes, a lower MSSE, a higher precision, and a higher AUC, particularly in the heterogeneous data. The BRE model with a uniform(0,1) prior particularly performed better than the DSLR2 model in the homogeneous data but performed similarly in the heterogeneous data.

Table 3 Performance of random-effects models applied in simulated data

Model	Prior	No. DE Genes				MSSE				Precision				Accuracy				AUC			
		H0	H1	H2	H3	H0	H1	H2	H3	H0	H1	H2	H3	H0	H1	H2	H3	H0	H1	H2	H3
DSL	–	65	74	92	124	2.9	2.9	2.9	2.9	0.95	0.95	0.91	0.79	0.97	0.97	0.98	0.98	0.76	0.79	0.84	0.90
DSLR2	–	69	104	139	198	1.7	1.7	1.7	1.7	0.95	0.91	0.79	0.59	0.97	0.98	0.98	0.96	0.77	0.89	0.95	0.97
BRE	U(0,0.001)	126	157	254	305	18.1	25.8	33.0	39.9	0.82	0.70	0.45	0.39	0.98	0.97	0.93	0.91	0.93	0.94	0.94	0.94
BRE	U(0,0.01)	218	324	404	436	10.5	16.0	20.0	22.3	0.55	0.37	0.30	0.28	0.95	0.90	0.86	0.84	0.97	0.95	0.92	0.92
BRE	U(0,0.1)	181	269	354	391	9.4	14.3	17.8	19.8	0.66	0.45	0.34	0.31	0.97	0.93	0.88	0.86	0.98	0.96	0.94	0.93
BRE	U(0,1)	80	108	141	203	1.7	2.2	2.4	2.6	1.00	0.94	0.80	0.58	0.98	0.99	0.98	0.96	0.84	0.92	0.96	0.97
BRE	U(0,10)	11	9	9	12	1.0	1.1	1.1	1.1	1.00	1.00	1.00	0.96	0.95	0.94	0.94	0.95	0.54	0.54	0.54	0.55
BRE	U(0,100)	10	8	8	11	1.0	1.0	1.0	1.0	1.00	1.00	1.00	0.96	0.94	0.94	0.94	0.94	0.54	0.53	0.53	0.54
BRE	U(0,0.005)	329	447	520	546	10.6	16.1	20.1	22.4	0.37	0.27	0.23	0.22	0.90	0.84	0.80	0.79	0.94	0.91	0.89	0.89
BRE	U(0,0.05)	184	275	359	395	10.3	15.7	19.6	21.8	0.65	0.44	0.33	0.30	0.97	0.92	0.88	0.86	0.98	0.96	0.94	0.93
BRE	U(0,0.5)	137	167	253	330	3.0	4.4	5.3	5.7	0.86	0.71	0.47	0.36	0.99	0.98	0.93	0.89	0.98	0.98	0.96	0.94
BRE	U(0,5)	13	11	12	17	1.1	1.1	1.1	1.1	1.00	1.00	1.00	0.97	0.95	0.95	0.95	0.95	0.55	0.54	0.55	0.57
BRE	G(1,2)	41	53	69	94	1.7	2.0	2.1	2.1	1.00	1.00	0.97	0.89	0.96	0.97	0.97	0.98	0.67	0.72	0.78	0.84

DE: differentially expressed, MSSE: minimum sum of squared error, AUC: area-under ROC curve, DSL: Dersimonian-Laird model, DSLR2: two-step estimate of Dersimonian-Laird model, BRE: Bayesian random-effects model, U: uniform, and G: gamma. H0, H1, H2, and H3 are the number of {0, 1, 2, and 3} studies containing heterogeneous genes. H0 represents homogenous data. The number of truly DE genes in the simulated data was 120 genes under HSC hypothesis testing

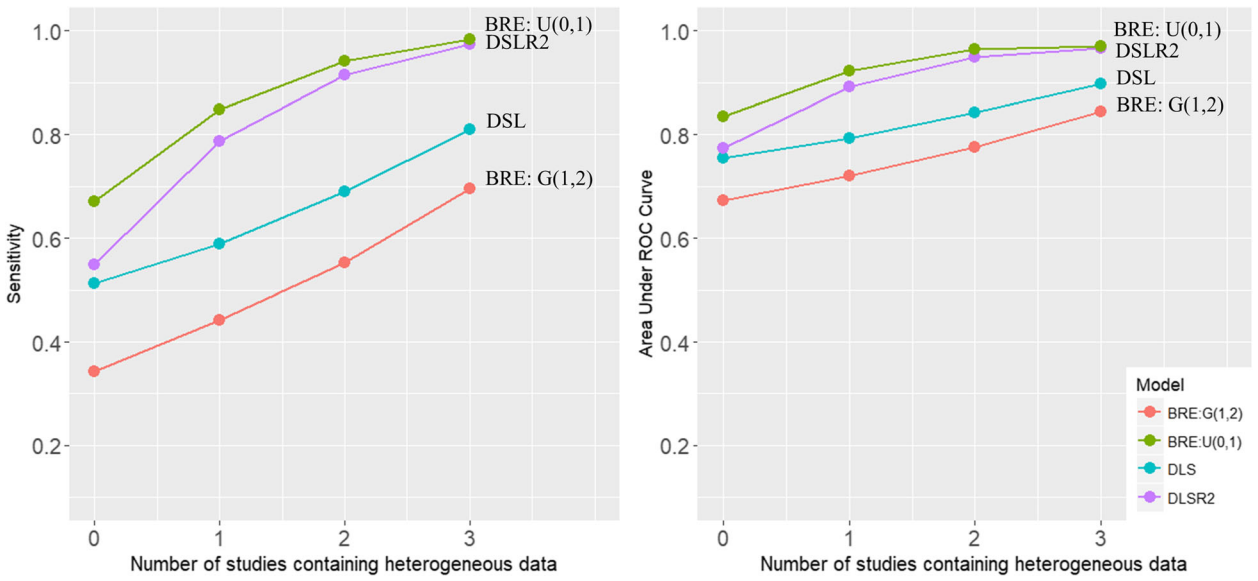

Fig. 1 Sensitivity and area under ROC curve of the random-effects models with Dersimonian-Laird (DSL), two-step (DSLR2), and Bayesian random-effects models (BRE) with uniform(0,1) and gamma(1,2) priors for between-study variance under the HSC hypothesis testing. H0, H1, H2, and H3 are the number of {0, 1, 2, and 3} studies containing heterogeneous genes. H0 represents homogenous data. The number of truly DE genes in the simulated data was 120 genes

Weighted DSL and DSLR2 models

With simulation results, the w_{P6} function was most appropriate for detecting DE genes in the DSL and DSLR2 models. The QC indicators adjusted the within study variance in the weighted function as:

$$w_{P6} = \left(\sigma_{ig}^{2(w_{P1})} + \hat{\tau}_g^2\right)^{-1}, \tag{18}$$

where $w_{P1} \in \{2^{-S_{ij}}, 0.01\tilde{P}_{ij}\}$, $\tilde{P}_{ij}$ denoted percent of present calls, S_{ij} denoted standardized quality indicators of the jth sample in the ith study. Fig. 2 presents the precision of the DSLR2 model with and without the w_{P6} function under different hypotheses in the homogeneous and heterogeneous data. The precision was increased with the DSLR2 weighted model in the heterogeneous data. The w_{P6} model provided an appropriate reduction of detected DE

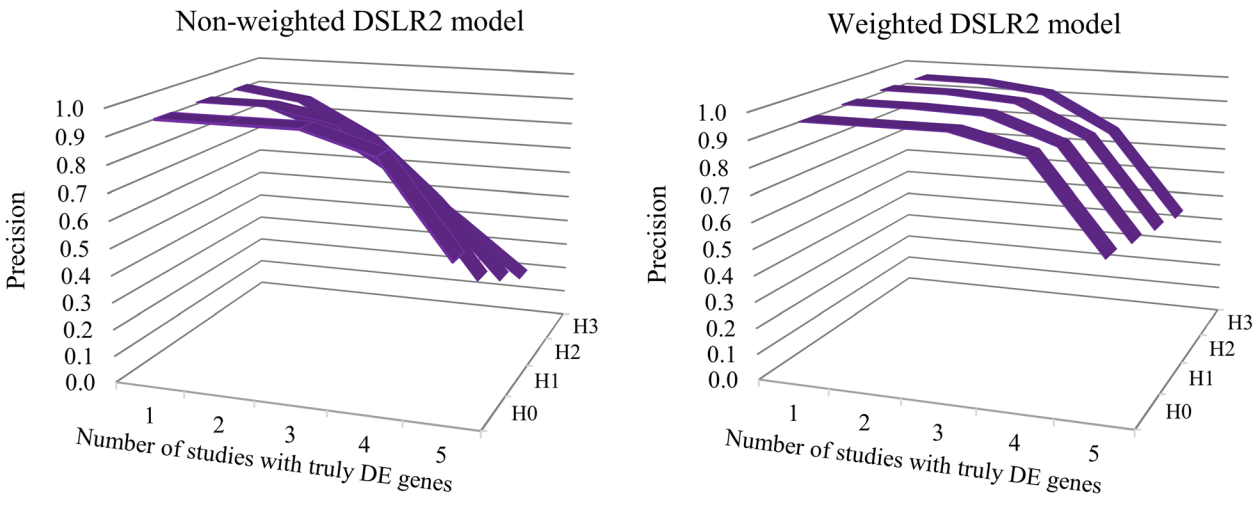

Fig. 2 Precision of two-step random-effects models (DSLR2) with and without the proper weighted function: $w_{P6} = (\sigma_{ig}^{2(w_{P1})} + \hat{\tau}_g^2)^{-1}$), $w_{P1} \in \{2^{-S_{ij}}, 0.01\tilde{P}_{ij}\}$

, $\tilde{P}_{ij}$ denoted percent of present calls, S_{ij} denoted standardized quality indicators of the jth sample in the ith study. H0, H1, H2, and H3 are the number of {0, 1, 2, and 3} studies containing heterogeneous genes. H0 represents homogenous data. The number of truly DE genes in the simulated data was 120 genes under HSC hypothesis testing. DE: differentially expressed

Table 4 Performance of weighted random-effects models applied in simulated data

Model	No. DE Genes				MSSE				Precision				Accuracy				AUC			
	H0	H1	H2	H3	H0	H1	H2	H3	H0	H1	H2	H3	H0	H1	H2	H3	H0	H1	H2	H3
DSLw_{P6}	62	62	64	65	2.9	3.0	3.0	3.0	0.95	0.96	0.96	0.96	0.97	0.97	0.97	0.97	0.75	0.75	0.75	0.76
DSLR2w_{P6}	66	72	78	85	1.6	1.6	1.6	1.6	0.96	0.95	0.94	0.92	0.97	0.97	0.97	0.98	0.76	0.78	0.80	0.82
BRE with a uniform(0,1) prior																				
Model 1: Unweighted data, Gibbs	81	109	140	204	1.7	2.1	2.4	2.6	1.00	0.94	0.81	0.58	0.98	0.99	0.98	0.96	0.84	0.92	0.96	0.97
Model 2: Unweighted data, Gibbs, $\beta\overline{w}_{P6}$	81	66	51	39	6.0	6.2	6.5	6.9	1.00	1.00	0.97	0.93	0.98	0.97	0.96	0.96	0.84	0.77	0.71	0.65
Model 3: Unweighted data, Gibbs, $\tau^2\overline{w}_{P6}$	161	157	151	142	0.8	1.5	2.1	2.7	0.74	0.76	0.77	0.79	0.98	0.98	0.98	0.98	0.99	0.99	0.97	0.96
Model 4: Weighted data, Gibbs	81	87	92	100	1.8	2.2	2.7	3.1	1.00	0.99	0.97	0.93	0.98	0.98	0.98	0.98	0.84	0.86	0.87	0.89
Model 5: Weighted data, Gibbs, $\beta\overline{w}_{P6}$	81	65	51	39	6.3	6.5	6. 9	7.3	1.00	1.00	0.97	0.93	0.98	0.97	0.96	0.96	0.84	0.77	0.70	0.65
Model 6: Weighted data, Gibbs, $\tau^2\overline{w}_{P6}$	162	157	151	142	1.6	2.6	3.6	4.5	0.74	0.76	0.77	0.79	0.98	0.98	0.98	0.98	0.99	0.99	0.97	0.96
Model 7: Weighted data, MH	81	87	93	102	2.2	2.7	3.1	3.5	1.00	0.98	0.97	0.92	0.98	0.98	0.98	0.98	0.84	0.86	0.87	0.89

$\overline{w}_{P6}$ is an average of w_{P6}, $w_{P6} = \left(\sigma_{ig}^{2(w_{P1})} + \hat{\tau}_g^2\right)^{-1}$ over the total samples; $w_{P1} \in \{2^{-S_{ij}}, 0.01\tilde{P}_{ij}\}$, $\tilde{P}_{ij}$ denoted percent of present calls, S_{ij} denoted standardized quality indicators of the *j*th sample in the *i*th study. DE: differentially expressed, MSSE: minimum sum of squared error, AUC: area-under ROC curve, DSL: DerSimonian-Laird model, DSLR2: two-step estimate of DerSimonian-Laird model, BRE: Bayesian random-effects model, U: uniform, G: gamma, MH: Metropolis–Hastings algorithm. H0, H1, H2, and H3 are the number of {0, 1, 2, and 3} studies containing heterogeneous genes. H0 represents homogenous data. The number of truly DE genes in the simulated data was 120 genes under HSC hypothesis testing.

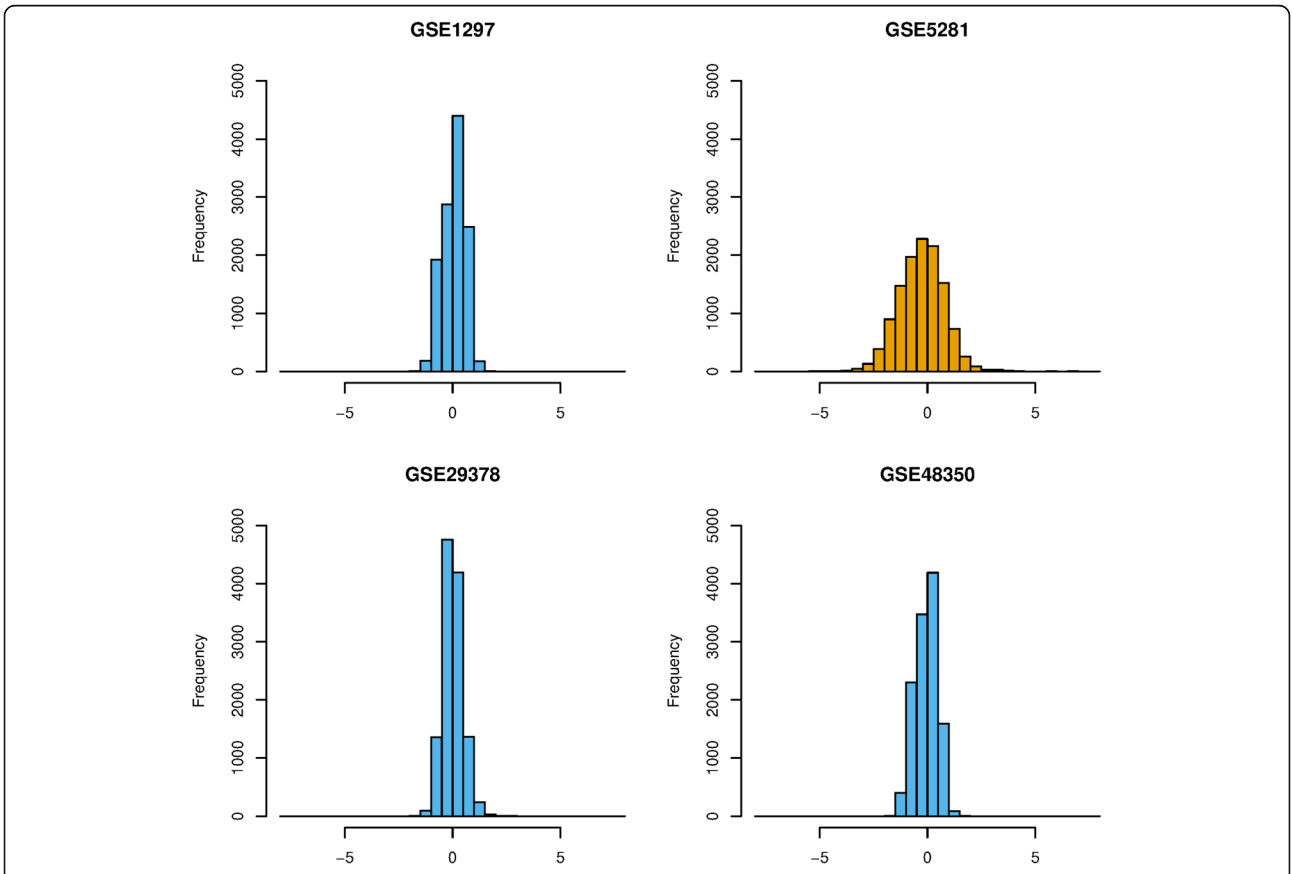

Fig. 3 Distribution of unbiased standardized mean difference of gene expression (x-axis) between Alzheimer's and control groups in GSE1297, GSE5281, GSE29378, and GSE48350 datasets

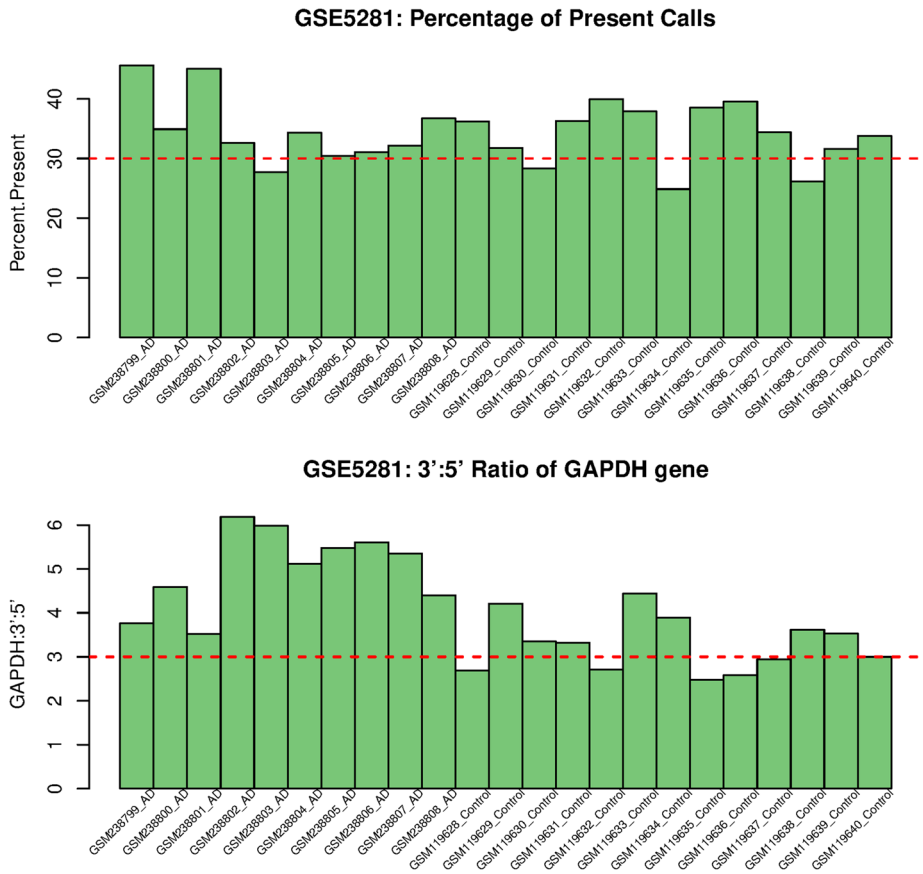

Fig. 4 Percentage of present calls and 3′:5′ GAPDH ratio of GSE5281 samples

genes and MSSEs and higher precision as compared to the other weighted functions (Additional file 1: Tables S1 and S2). Similar results were found under different levels of sample quality (results not shown). The DSLR2 w_{P6} weighted model had a lower MSSE and detected more DE genes than the DSL w_{P6} weighted model in the heterogeneous data.

Weighted Bayesian random-effects models

Table 4 presents the performance of the DSLw_{P6} and DSLR2w_{P6} models, and BRE weighted models. A uniform(0,1) prior for between study variance was applied in all BRE models. The BRE weighted Models 1, 3, 4, 6, and 7 in Table 4 detected more DE genes with a higher AUC than the DSLw_{P6} and DSLR2w_{P6} models. The w_{P6} weighted-data models performed similarly to the unweighted-data models (Models 2 vs. 5 and 3 vs. 6). The w_{P6}weighted common-effect model performed similarly to the unweighted model in the homogeneous data, but performed worse in the heterogeneous data (Models 1 vs. 2). Additionally, the Gibbs- and MH-based models performed similarly on the w_{P6}weighted-data model. The numbers of detected DE genes were reduced close to the number of truly DE genes and the precisions were increased while maintaining a high accuracy as compared to the performance in the unweighted-data Gibbs-based model (Models 4 and 7 vs. 1). For homogeneous and heterogeneous data, the Gibbs- and MH-based models with the w_{P6}weighted-data performed similarly and were most appropriate for detecting DE genes with high precision (Models 4 and 7). The w_{P6}weighted between-study variance models were most appropriate for detecting DE genes with high overall accuracy (Models 3 and 6).

Additional simulation results

Simulations with varying sample size, number of genes, and different levels of sample quality were conducted and some results were presented in the supplemental material. It is noteworthy that the BRE models identified less genes for sample sizes < 60. The DE gene detection and the MSSE were stable for sample sizes > 60. Specifically, the BRE with a U(0,1) had consistently high precisions and was able to maintain overall accuracies for all sample sizes > 60 (Additional file 1: Table S3). As anticipated, these findings were similar to the findings in the classical RE models [6]. When the number of genes in the analyses

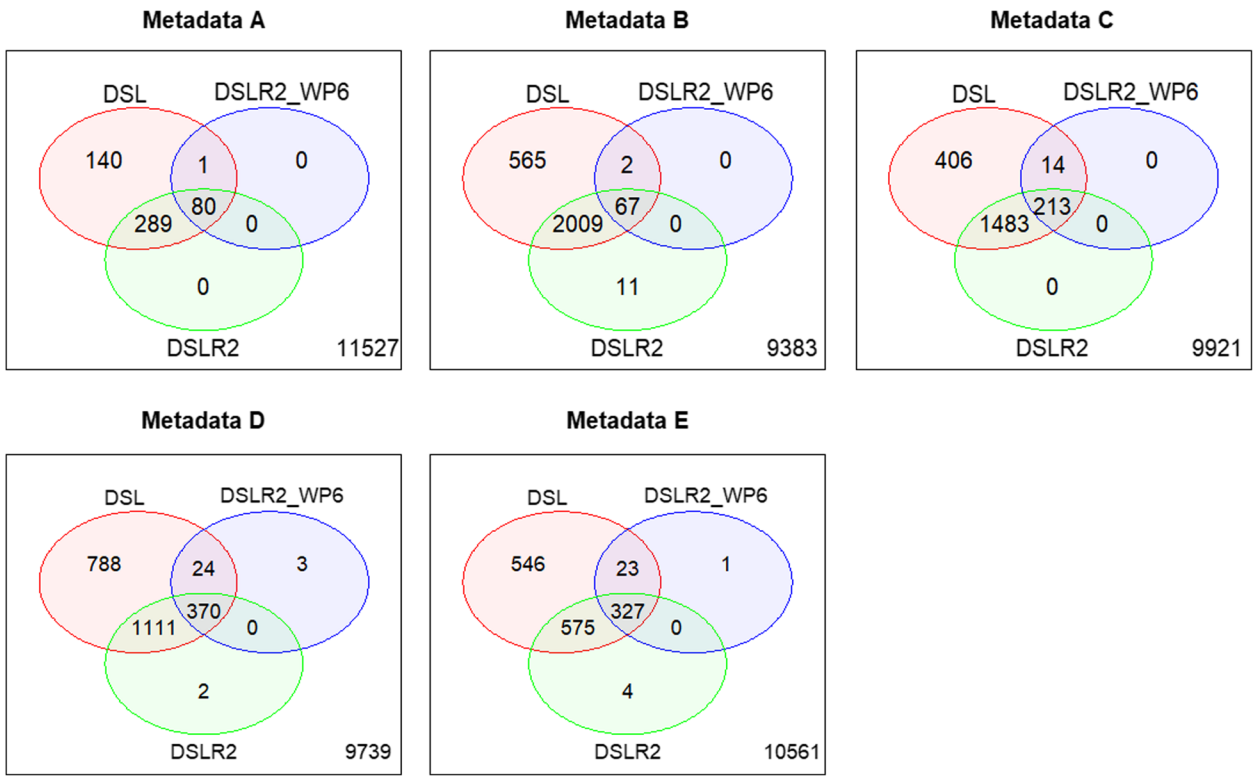

Fig. 5 Venn diagrams present number of differentially expressed genes in Alzheimer's disease as compared to controls in white matter region using classical random-effects models: Dersimonian-Laird (DSL), two-step estimated (DSLR2) random-effects models and with the proper weighted function: $w_{p6} = (\sigma_{ig}^{2(w_{p1})} + \hat{\tau}_g^2)^{-1}$ (DSLR2w_{P6}), where $w_{p1} \in \{s^{-S_{ij}}, 0.01\tilde{P}_{ij}\}$, $\tilde{P}_{ij}$ denoted percent of present calls, S_{ij} denoted standardized quality indicators of the *j*th sample in the *i*th study. Metadata A: GSE1297, GSE5281, and GSE29378; B: GSE1297, GSE5281, and GSE48350; C: GSE1297, GSE29378, and GSE48350; D: GSE1297, GSE5281, GSE29378, and GSE48350; and E: GSE5281, GSE29378, and GSE48350

increased, the classical RE models performed stably, while the overall accuracy in the BRE model with a uniform(0,1) prior was reduced (Additional file 1: Table S4). For different levels of sample quality, the weights with higher sample quality detected more DE genes and had higher overall accuracy than the weights with lower sample quality (Additional file 1: Table S5).

Application in Alzheimer's gene expression data

Our meta-analysis in the Alzheimer's gene expression datasets was performed on 12,037 target genes in 131 subjects (68 AD cases and 63 controls). We primarily examined the strength of study heterogeneity by considering five ways of metadata sets as described in [6]. The metadata A, B, D, E may contain heterogeneous data due to a relatively high R^2, while the metadata C had a relatively low R^2or contained homogenous data. Figure 3 presents distribution of unbiased standardized mean differences of gene expression in the GSE5281 dataset, different from the other datasets. Figure 4 presents the percent of present calls and the 3′:5′ GAPDH ratio of the heterogeneous dataset.

Using the DSLR2w_{P6}weighted model, the number of DE genes decreased in all metadata sets. Almost all the DE genes identified by the weighted model were genes among the significant DE genes identified by the unweighted DSL and DSLR2 models. The DE genes identified using the weighted model in the metadata C concurrently detected approximately 13% of the unweighted DSL and DSLR2 models (266/2116 genes and 213/1696 genes), respectively (Fig. 5). Likewise, the number of DE genes decreases with the w_{P6}weighted between study variance (Models 3 and 6). Those DE genes were genes among the significant DE genes identified by the unweighted model (Model 1). Sixty and 446 DE genes were detected across the three weighted BRE models in the metadata C and D, respectively (Fig. 6). Among the unweighted or weighted classical RE and BRE models, 446 genes could potentially be down-regulated genes that may contribute to good classification of Alzheimer's samples. Additional file 1: Figure S2 presents potential down-regulations of those genes in Alzheimer's samples in each microarray dataset. Of note, no genes were detected using the weighted common-effect models (Models 2 and 5) and the weighted-data model (Models 4 and 7).

The lists of 213 and 446 DE genes can be found in Additional file 1: Tables S6 and S7, respectively, where

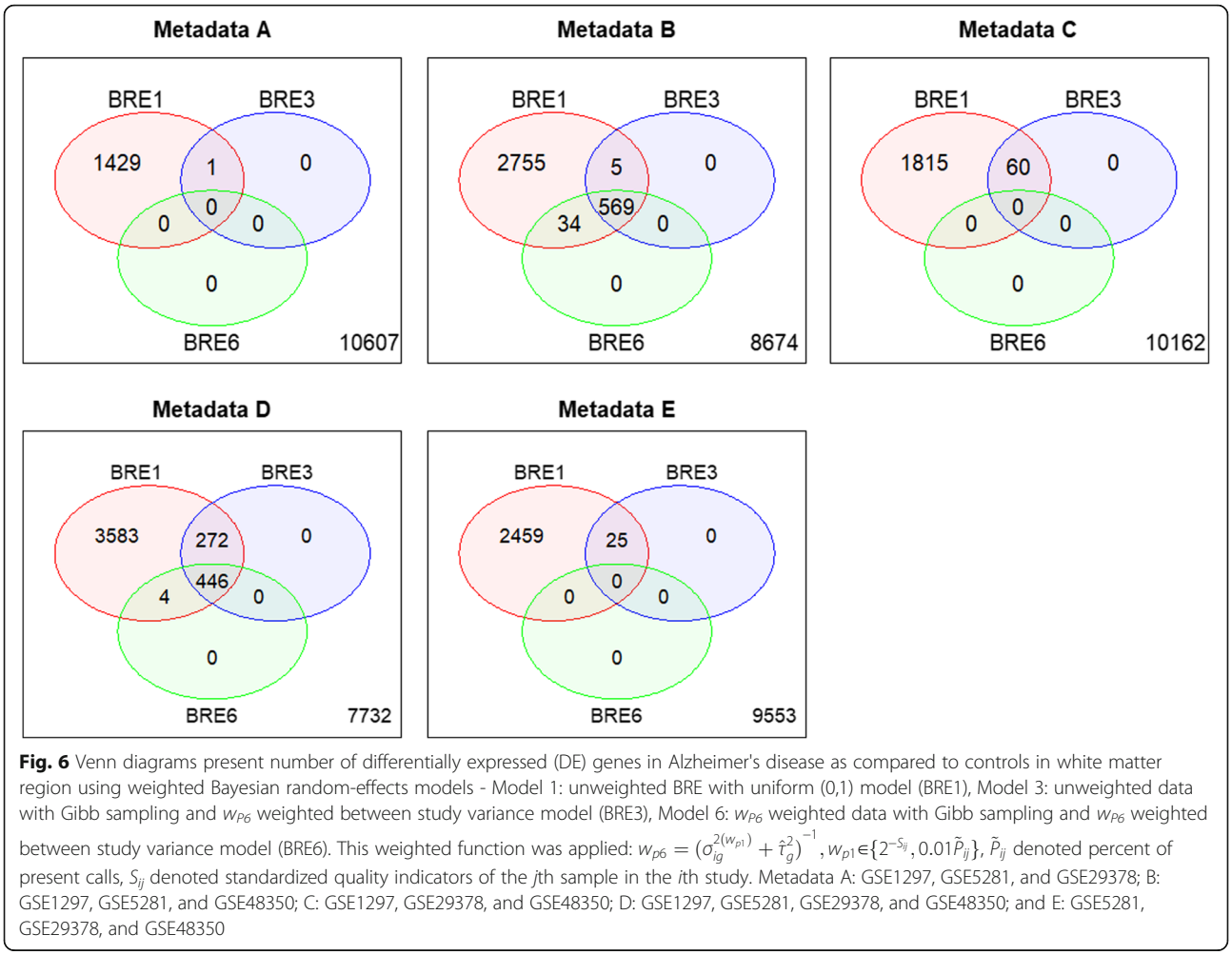

Fig. 6 Venn diagrams present number of differentially expressed (DE) genes in Alzheimer's disease as compared to controls in white matter region using weighted Bayesian random-effects models - Model 1: unweighted BRE with uniform (0,1) model (BRE1), Model 3: unweighted data with Gibb sampling and w_{P6} weighted between study variance model (BRE3), Model 6: w_{P6} weighted data with Gibb sampling and w_{P6} weighted between study variance model (BRE6). This weighted function was applied: $w_{p6} = \left(\sigma_{ig}^{2(w_{p1})} + \hat{\tau}_g^2\right)^{-1}$, $w_{p1} \in \{2^{-S_{ij}}, 0.01\tilde{P}_{ij}\}$, $\tilde{P}_{ij}$ denoted percent of present calls, S_{ij} denoted standardized quality indicators of the jth sample in the ith study. Metadata A: GSE1297, GSE5281, and GSE29378; B: GSE1297, GSE5281, and GSE48350; C: GSE1297, GSE29378, and GSE48350; D: GSE1297, GSE5281, GSE29378, and GSE48350; and E: GSE5281, GSE29378, and GSE48350

98 were the same genes. The identified DE genes participate in significant pathways such as cytoskeleton organization, actin filament bundle organization, synaptic transmission, regulation of biological quality, neutral lipid biosynthetic process, acylglycerol biosynthetic process, intermediate filament-based process, negative regulation of neuron projection development, cell-cell signaling, glutamate decarboxylation to succinate, stress fiber assembly, single-organism behavior, single-organism behavior, response to ethanol, cellular component assembly, neuron projection development, learning and long-term memory.

Discussion

This study presents the performance of the classical RE and BRE models in meta-analysis of gene expression studies. We found the BRE model with a uniform(0,1) prior was appropriate for detecting DE genes as compared to the models with other prior distributions. The BRE model with a uniform(0,1) prior performed better than the DSLR2 model in the homogeneous data, but performed similarly in the heterogeneous data in terms of an appropriate number of detected DE genes, lower MSSE, higher precision, and higher AUC.

This is the first study to reveal an application of sample-quality weights to adjust the study heterogeneity in the classical RE and BRE models in microarray gene expression studies. The DSL and DSLR2 weighted models were implemented for the classical RE models. The unweighted and weighted data, Gibbs and MH sampling algorithms, weighted common effect, and weighted between-study variance were applied for the BRE models. We evaluated the performance of the models through simulation studies and through application to Alzheimer's gene expression datasets.

With simulation results, the sample quality indicators adjusting the within study variance (w_{P6}) in the classical RE models provided an appropriate reduction of detected DE genes and MSSEs, and higher precision as compared to the other weighted functions. The precision in detecting DE genes was increased with the DSLR2 w_{P6} weighted model in the heterogeneous data. The DSLR2 w_{P6} weighted model

had a lower MSSE and detected more DE genes than the DSL w_{P6} weighted model in the heterogeneous data. Among the BRE weighted models, the w_{P6}weighted- and unweighted-data models and both Gibbs- and MH-based models performed similarly. The w_{P6} weighted common-effect model performed similarly to the unweighted model in the homogeneous data, but performed worse in the heterogeneous data. The w_{P6}weighted-data were appropriate for detecting DE genes with high precision, while the w_{P6}weighted between-study variance models were appropriate for detecting DE genes with high overall accuracy.

The sample quality has substantial influence on results of gene expression studies [15]. Because variation of sample quality limited meta-analysis techniques to properly detect DE genes [45, 46] and the classical RE and BRE models allow flexibility in calculating $y_{ig}$and its variance σ_{ig}^2 as well as study-specific adjustments [47], we developed approaches to up-weight good quality samples and down-weight borderline quality samples in the models. This compromised approach utilizes sample-quality information in the meta-analysis of microarray studies in detecting DE genes. The results in this study would benefit microarray gene expression studies because a large amount of microarray data are available in public repositories and unfortunately the data quality are often overlooked. However, the performance of the proposed models depends on not only degree of sample quality but also the number of studies, the number of genes, and sample sizes in the individual studies. The methods for controlling FDR under multiple testing would be another important aspect influencing gene expression results. Further intensive investigation of the topics would be the subject of future research.

The BRE models have the ability to allow for uncertainty of the parameter estimates in the model. Because the classical RE models tended to estimate τ_g^2 as being zero, the variance of $\hat{\beta}_g$ were underestimated. The BRE models, in contrast, used the marginal posterior distribution of τ_g^2 for $\hat{\beta}_g$ estimation, which do not depend on the point estimate of τ_g^2. The BRE models can in turn increase the fitness of the models [48]. To illustrate, the precision was increased in the BREw_{P6}weighted-data models and the accuracy was increased in the BREw_{P6}weighted between-study variance models as compared to the classical RE weighted models. The BRE weighted models could be strengthened further in future research with informative priors using prior knowledge and historical information.

In real-world applications, BRE modeling in gene expression meta-analysis may be computationally intensive. To illustrate, a Gibbs-based model requires approximately 6 h per 10,000 gene set under supercomputers. A MH-based model requires twice longer than a Gibbs-based model. The computational time for a BRE model is highly dependent on not only types of the model, but also computer capacity. Computation time can indeed be another concern for model selection.

Conclusions

This study applies sample-quality weights to adjust the study heterogeneity in the random-effects meta-analysis models. This meta-analytic approach can increase precision and accuracy of the classical and Bayesian random-effects models in gene expression meta-analysis. However, the performance of the weighted models varied depending on data feature, levels of sample quality, and adjustment of parameter estimates.

Additional file

Additional file 1: Table S1. Number of differentially expressed (DE), minimum sum of squared errors (MSSE), precision, and accuracy of non-weighted and weighted random-effects models with Dersimonian-Laird (DSL) estimate applied in simulated data. **Table S2.** Number of differentially expressed (DE), Minimum sum of squared errors (MSSE), precision, and accuracy of non-weighted and weighted random effects meta-analysis model with two-step Dersimonian-Laird (DSLR2) estimate applied in simulated data. **Figure S1.** Number of differentially expressed genes and minimum sum of squared errors of Dersimonian-Laird (DSL), two-step (DSLR2), and Bayesian random-effects (BRE) models with different lengths of uniform priors for between-study variance estimation in simulated data. **Table S3.** Performance of Bayesian random-effects models by different levels of sample sizes (some results from homogenous simulated datasets). **Table S4.** Performance of classical and Bayesian random-effects models by different numbers of genes (some results from H1 heterogeneous simulated datasets). **Table S5.** Performance of weighted random-effects models applied with two levels of sample-quality weights (some simulation results). **Figure S2.** Heatmaps of expression patterns of 446 differentially expressed genes in white matter in Alzheimer's and control samples. The DE genes were detected across the three Bayesian meta-analysis models as shown in metadata D in Fig. 6. **Table S6.** List of 213 significantly differentially expressed genes in Alzheimer's gene expression dataset. The DE genes detected across the DSLR2 w_{P6} weighted and DSLR2 and DSL unweighted models as shown in metadata C in Fig. 5. **Table S7.** List of 446 significantly differentially expressed genes in Alzheimer's gene expression datasets. The DE genes detected across three Bayesian random-effect models (Models 1, 3, and 6) as shown in metadata D in Fig. 6. (PDF 886 kb)

Abbreviations

AUC: Area under receiver operating characteristic curve; BH: Benjamini and Hochberg; BRE: Bayesian random-effects model; DE: Differentially expressed; DSL: DerSimonian and Laird; DSLR2: Two-step DerSimonian and Laird estimate; FDR: False discovery rate; FE: Fixed-effects; GAPDH: Glyceraldehyde-3-phosphate dehydrogenase; MH: Metropolis Hasting; MSSE: Minimum sum of squared error; QC: Quality control; RE: Random-effects; SD: Standard deviation

Acknowledgements

We would like to thank anonymous reviewers for their suggestions and insightful comments.

Funding

Research reported in this work was supported in part by the National Library Of Medicine of the National Institutes of Health under Award Number R01LM011169. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Availability of data and materials
Additional file 1 Supporting online material. PDF document with WinBUGS code, supplementary tables (Additional file 1: Table S1 – S7) and figure (Additional file 1: Figure S1 and S2). The program for implementing the weighted models can be modified from existing R packages and are available upon request.

Authors' contributions
US conceived the study, conducted the simulations, interpreted the results, wrote and edited the manuscript. KA and NM advised the study, interpreted the results, reviewed and edited the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Office of Biostatistics, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, USA. [2]Division of Biostatistics, College of Public Health, The Ohio State University, Columbus, Ohio, USA. [3]Department of Biostatistics, Virginia Commonwealth University, Richmond, Virginia, USA.

Received: 28 August 2018 Accepted: 12 November 2018
Published online: 09 January 2019

References
1. Ramasamy A, Mondry A, Holmes CC, Altman DG. Key issues in conducting a meta-analysis of gene expression microarray datasets. PLoS Med. 2008 Sep 30;5(9):e184.
2. Rung J, Brazma A. Reuse of public genome-wide gene expression data. Nat Rev Genet. 2013;14(2):89–99.
3. Tseng GC, Ghosh D, Feingold E. Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res. 2012 May;40(9):3785–99.
4. Song C, Tseng GC. Hypothesis setting and order statistic for robust genomic meta-analysis. Ann App Stat. 2014;8(2):777.
5. Chang LC, Lin HM, Sibille E, Tseng GC. Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. BMC Bioinformatics. 2013;14: 368,2105–14-368.
6. Siangphoe U, Archer KJ. Estimation of random effects and identifying heterogeneous genes in meta-analysis of gene expression studies. Brief Bioinform. 2017;18(4):602–18.
7. Li, Y, Ghosh, D. Meta-analysis based on weighted ordered P-values for genomic data with heterogeneity. BMC bioinformatics. 2014;(1):226.
8. Wang H, Zheng C, Zhao X. jNMFMA: a joint non-negative matrix factorization meta-analysis of transcriptomics data. Bioinformatics. 2014; 31(4), 572-80.
9. Evangelou E, Ioannidis JP. Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet. 2013;14(6):379–89.
10. Hong F, Breitling R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. Bioinformatics. 2008;24(3):374–82.
11. Choi JK, Yu U, Kim S, Yoo OJ. Combining multiple microarray studies and modeling interstudy variation. Bioinformatics. 2003;19(Suppl 1):i84–90.
12. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. Contemporary Clinical trials. 2007;28(2):105–14.
13. Higgins J, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J Royal Stat Soc: Series A (Statistics in Society). 2009;172(1): 137–59.
14. Ntzoufras I. Bayesian modeling using WinBUGS. New York: Wiley; 2011:698
15. Draghici, Sorin. Statistics and data analysis for microarrays using R and Bioconductor. 2nd Edition ed. New York: Chapman & Hall/CRC Mathematical and Computational Biology; 2010.
16. Siangphoe U, Archer KJ. Gene expression in HIV-associated neurocognitive disorders: a meta-analysis. J Acquir Immune Defic Syndr. 2015;70(5):479–88.
17. Cheng WC, Tsai ML, Chang CW, Huang CL, Chen CR, Shu WY, et al. Microarray meta-analysis database (M(2)DB): a uniformly pre-processed, quality controlled, and manually curated human clinical microarray database. BMC Bioinformatics. 2010;11:421,2105–1-421.
18. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, et al. Multiple-laboratory comparison of microarray platforms. Nat Methods. 2005;2(5):345–50.
19. Asare AL, Gao Z, Carey VJ, Wang R, Seyfert-Margolis V. Power enhancement via multivariate outlier testing with gene expression arrays. Bioinformatics. 2009;25(1):48–53.
20. Dumur CI, Nasim S, Best AM, Archer KJ, Ladd AC, Mas VR, et al. Evaluation of quality-control criteria for microarray gene expression analysis. Clin Chem. 2004;50(11):1994–2002.
21. McClintick JN, Edenberg HJ. Effects of filtering by present call on analysis of microarray experiments. BMC Bioinformatics. 2006;7:49.
22. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics. 2008;24(13):1547–8.
23. Smyth GK. Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor; 2005.
24. Dunning MJ, Smith ML, Ritchie ME, Tavare S. Beadarray: R classes and methods for Illumina bead-based data. Bioinformatics. 2007;23(16):2183–4.
25. HEDGES L, Olkin I. Statistical Methods for Meta-Analysis (Orlando, FL: Academic). HedgesStatistical Methods for Meta-Analysis1985; 1985.
26. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. Introduction to meta-analysis: Chichester: Wiley; 2011.
27. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. Stat Med. 1991;10(11):1665–77.
28. Demidenko E, Sargent J, Onega T. Random effects coefficient of determination for mixed and meta-analysis models. Commun Stat Theory Methods. 2012;41(6):953–69.
29. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc. Series B (Methodological). 1995;51(1):289–300.
30. Alex JS, Keith RA. Bayesian methods in meta-analysis and evidence synthesis. Stat Methods Med Res. 2001;10(4):277–303.
31. Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Stat Med. 2005;24(15):2401–28.
32. Paule RC, Mandel J. Consensus values and weighting factors. J Res Natl Bur Stand. 1982;87(5):377–85.
33. Cochran WG. The combination of estimates from different experiments. Biometrics. 1954;10(1):101–29.
34. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. Texts in statistical science series; 2004.
35. Swets JA. Measuring the accuracy of diagnostic systems. Science. 1988; 240(4857):1285–93.
36. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett. 2006;27(8): 861–74.
37. Blalock EM, Geddes JW, Chen KC, Porter NM, Markesbery WR, Landfield PW. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc Natl Acad Sci U S A. 2004;101(7):2173–8.
38. Liang WS, Dunckley T, Beach TG, Grover A, Mastroeni D, Walker DG, et al. Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. Physiol Genomics. 2007;28(3):311–22.
39. Miller JA, Woltjer RL, Goodenbour JM, Horvath S, Geschwind DH. Genes and pathways underlying regional and cell type changes in Alzheimer's disease. Genome Med. 2013;5(5):48.
40. Blair LJ, Nordhues BA, Hill SE, Scaglione KM, O'Leary JC, Fontaine SN, et al. Accelerated neurodegeneration through chaperone-mediated oligomerization of tau. J Clin Invest. 2013;123(10):4158–69.
41. Davis S, Meltzer PS. GEOquery: a bridge between the gene expression omnibus (GEO) and BioConductor. Bioinformatics. 2007;23(14):1846–7.
42. Parman C, Conrad H, Gentleman R. affyQCReport: QC Report Generation for affyBatch objects. R package version 1.42.0.

43. Gautier L, Cope L, Bolstad BM, Irizarry RA. Affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307–15.
44. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80.
45. Liu R, Holik AZ, Su S, Jansz N, Chen K, Leong HS, et al. Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. Nucleic Acids Res. 2015;43(15):e97.
46. Eijssen LM, Jaillard M, Adriaens ME, Gaj S, de Groot PJ, Muller M, et al. User-friendly solutions for microarray quality control and pre-processing on ArrayAnalysis.org. Nucleic Acids Res. 2013;41(Web Server issue):W71–6.
47. Demidenko E. Mixed models: theory and applications with R: Hoboken: Wiley; 2013.
48. Bodnar O, Link A, Arendacká B, Possolo A, Elster C. Bayesian estimation in random effects meta-analysis using a non-informative prior. Stat Med. 2017; 36(2):378–99.