bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

BAGSE: a Bayesian hierarchical model approach for gene set enrichment analysis

Abhay Hukku[1], Corbin Quick[1], Francesca Luca[2,3], Roger Pique-Regi[2,3], and Xiaoquan Wen*[1]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

[2]Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI 48201, USA

[3]Department of Obstetrics and Gynecology, Wayne State University, Detroit, MI 48201, USA

Abstract

Gene set enrichment analysis has been shown to be effective in identifying relevant biological pathways underlying complex diseases. Existing approaches lack the ability to quantify the enrichment levels accurately, hence preventing the enrichment information to be further utilized in both upstream and downstream analyses. A modernized and rigorous approach for gene set enrichment analysis that emphasizes both hypothesis testing and enrichment estimation is much needed. We propose a novel computational method, Bayesian Analysis of Gene Set Enrichment (BAGSE), for gene set enrichment analysis. BAGSE is built on a natural Bayesian hierarchical model and fully accounts for the uncertainty embedded in the association evidence of individual genes. We adopt an empirical Bayes inference framework to fit the proposed hierarchical model by implementing an efficient EM algorithm. Through simulation studies, we illustrate that BAGSE yields accurate enrichment quantification while achieving similar power as the state-of-the-art methods. Further simulation studies show that BAGSE can effectively utilize the enrichment information to improve the power in gene discovery. Finally, we demonstrate the application of BAGSE in analyzing real data from differential expression experiment and TWAS analysis. BAGSE is implemented using the C++ programming language and is freely available from `https://github.com/xqwen/bagse/`. Simulated and real data used in this paper are also available at the Github repository for the reproducibility purpose.

*xwen@umich.edu

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1 Introduction

Gene set enrichment analysis has become a standard analytic tool in systems biology and bioinformatics. Its primary aim is to identify specific groups of genes in which the association signals are enriched (or depleted) given the association evidence from individual genes. The results from gene set enrichment analysis have implications beyond the association evidence at the single gene level: as the gene set is typically defined by the biological relevance of the member genes, the enrichment of signals in specific gene sets sheds lights on the underlying biological pathways and gene networks, which subsequently helps to uncover relevant molecular mechanisms in a biological system. In practice, gene set enrichment analysis is often conducted downstream of differential expression (DE) analysis and genome-wide genetic association analysis (GWAS). Recently emerged transcriptome-wide association analysis (TWAS) has shown promise in linking causal genes to complex traits utilizing both the data from mapping expression quantitative trait loci (eQTL) and GWAS (Gamazon *et al.*, 2015; Gusev *et al.*, 2018; Zhu *et al.*, 2016). Gene set enrichment analysis based on TWAS results will have the potential to uncover the causal gene networks that lead to complex diseases.

The available gene set enrichment analysis approaches in literature can be roughly classified into two groups. The first group is represented by the popular approach GSEA (Mootha *et al.*, 2003; Subramanian *et al.*, 2005). For each pre-defined gene set, GSEA constructs a ranked list of all member genes based on their association evidence with respect to the phenotype of interest. It then performs a Kolmogorov-Smirnov (KS)-like test to compare the distributions between different gene sets. This procedure has been widely used since its inception, as shown in Keshava Prasad *et al.* (2008); Guttman *et al.* (2009); Shalem *et al.* (2014); Schaub *et al.* (2018). Built upon the algorithm of GSEA, many software packages provide further improvements targeting specific applications (Segrè *et al.*, 2010; Willer *et al.*, 2013; Speliotes *et al.*, 2010). Notably, GSEA-based gene set enrichment analysis has led to breakthroughs in the profiling of cancer cells (Maruschke *et al.*, 2014), and the studies of complex diseases like schizophrenia (Hass *et al.*, 2015) and depression Elovainio *et al.* (2015). A two-stage procedure characterizes the other class of enrichment analysis methods. Taking an example of an enrichment analysis from a DE experiment: in the first stage, genes are classified into either differentially expressed or not based on the association evidence without considering their gene set annotations; in the second stage, a contingency table is constructed according to the DE status and gene set membership of all investigated genes. The resulting contingency table is subsequently used to quantify the enrichment level (by computing the log odds ratio) and testing enrichment (by a chi-squared test or a Fisher's exact test) for a particular gene set. This method has

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

also been widely applied in the recent literature of genomics and complex disease studies (Richiardi *et al.*, 2015; Walter *et al.*, 2015; Chang *et al.*, 2016).

Despite the popularity of both types of methods in gene set enrichment analysis, they both lack the ability of accurate quantification of enrichment levels for gene sets. The GSEA approach is statistically rigorous in performing hypothesis testing; however, it is not designed to provide an estimation of the enrichment level. The two-stage approach is seemingly intuitive; nevertheless, the classification in the first stage ignores the uncertainty of the gene-level association evidence, which leads to biased estimates of enrichment levels (the details will be explained in Section 2.3). We argue that the accurate quantification of enrichment from a gene set enrichment analysis is critical in many bioinformatics applications. Such information is necessary for comparing the relative importance of multiple gene sets in the same disease or comparing the roles of the same gene set in various conditions.

In this paper, we propose an empirical Bayes procedure, Bayesian Analysis of Gene Set Enrichment (BAGSE), for gene set enrichment analysis. Our computational approach is derived from a natural hierarchical model. BAGSE is suitable for not only rigorous hypothesis testing but also accurate quantification of enrichment levels. Additionally, BAGSE can simultaneously handle multiple and/or mutually non-exclusive gene set definitions, a feature currently missing from the existing methods. Finally, we show that within the proposed hierarchical model framework of BAGSE, the gene set enrichment information can be subsequently applied for improving the power in uncovering association evidence at the gene-level. The software package implementing the proposed procedures is made freely available at `https://github.com/xqwen/bagse/`.

2 Methods

2.1 Model and notations

We consider a general setting suitable for analyzing summary-level data generated from both DE and TWAS studies. Specifically, we use β_i to denote the effect size of the association for each gene i. In DE analysis, β typically represents the log fold-change of expression levels under two different experimental conditions; in TWAS, β quantifies the strength of association between the phenotype of interest and the genotype-predicted gene expression levels. Suppose that the analysis of each gene i yields a maximum likelihood estimate of effect size, $\hat{\beta}_i$, along with its standard error, $\hat{s}_i$. With

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

sufficient sample size, it follows that

$$\hat{\beta}_i \sim \mathrm{N}(\beta, \hat{s}_i^2), \tag{1}$$

and we also consider that $(\hat{\beta}_i, \hat{s}_i^2)$ is a sufficient statistic for β_i. Throughout this paper, we assume the observed gene-level association data are summarized by $\boldsymbol{D} := \{(\hat{\beta}_i, \hat{s}_i) : i = 1, \ldots, M\}$ for all M genes, and it is made available for enrichment analysis. In the case that only p-values are made available, we map each p-value to a corresponding z-statistic through a standard normal distribution, hence $\hat{\beta}_i = z_i$ and $\hat{s}_i = 1$ Efron (2012); Stephens (2016).

We define a latent binary indicator $\gamma_i := \mathbf{1}\{\beta_i \neq 0\}$ to represent the true association status of gene i and assume its annotation data, d_i, provides potential prior knowledge on γ_i. For the mathematical convenience of the presentation, unless otherwise specified, we assume a single gene set is pre-defined, and d_i is a binary indicator representing if gene i is annotated. (In Section 2.4 and the Supplementary Material, we relax this restriction and consider multiple overlapping gene sets.) We assume a logistic prior function connecting d_i and γ_i, i.e.,

$$\log \left[\frac{\Pr(\gamma_i = 1)}{\Pr(\gamma_i = 0)} \right] = \alpha_0 + \alpha_1 d_i, \tag{2}$$

where the coefficients $\boldsymbol{\alpha} := (\alpha_0, \alpha_1)$ quantify the enrichment information. For example, if $\alpha_1 > 0$, the genes belonging to the gene set of interest are more likely to be associated.

To complete the hierarchical model, we follow the recently proposed adaptive shrinkage (ASH) method (Stephens, 2016) to model the prior effect size β_i (conditional on $\gamma_i = 1$) using a mixture of K normal distributions, i.e.,

$$\beta_i \mid \gamma_i = 1, d_i \sim \sum_{k=1}^{K} \pi_{k,d_i} N(0, \phi_k^2). \tag{3}$$

Accordingly, conditional on $\gamma_i = 0$, $\beta_i = 0$ by definition.

In practice, we determine the number of the mixing components, K, and corresponding effect size parameters $\{\phi_k^2\}$ using a data-driven approach as described in Stephens (2016). (The technical details are also described in Section 1.4 of the Supplementary material). Importantly, we allow the mixture proportions, i.e., $\boldsymbol{\pi}_{d_i} := \{\pi_{k,d_i}\}$, to vary across different types of annotations, which provides the necessary flexibility to model potentially different effect size distributions for different kinds of gene sets or pathways. We view this feature as an improvement and a generalization of the original ASH model.

The proposed Bayesian hierarchical model can be summarized by the graphical model shown in Fig. 1.

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

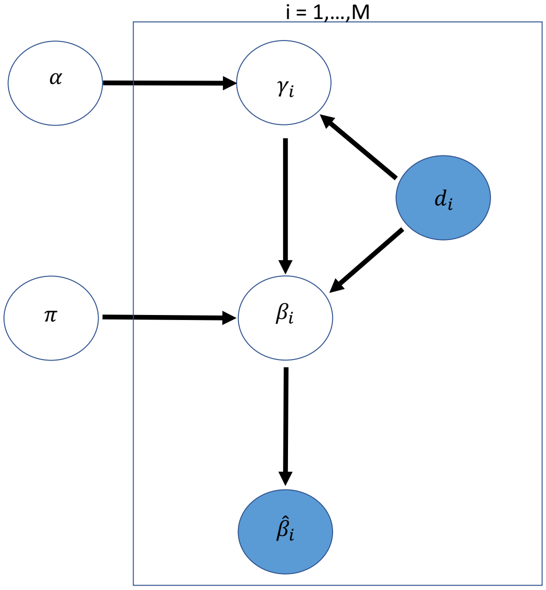

Figure 1: A graphical model representation of the BAGSE model. The directed acyclic graph (DAG) represents a probabilistic generative model. The shaded variables represent data that are observed.

With observed association data $\boldsymbol{D}$ and annotation data $\boldsymbol{d} := (d_1, ..., d_M)$, we frame the problem of enrichment analysis as an inference problem with respect to the enrichment parameter $\boldsymbol{\alpha}$.

2.2 Gene set enrichment estimation

To quantify the enrichment level of annotated genes within a pre-defined gene set, we perform maximum likelihood estimation with respect to $\boldsymbol{\alpha}$ based on the proposed hierarchical model. Particularly, we design an Expectation and Maximization (EM) algorithm to obtain the maximum likelihood estimates (MLEs) for hyperparameters $\boldsymbol{\alpha}, \pi$ by treating the latent binary vector $\boldsymbol{\gamma}$ as missing data.

Briefly, in the E-step of the t-th iteration we evaluate the probability $\Pr(\gamma_i = 1 \mid \boldsymbol{\alpha}^{(t)}, \boldsymbol{\pi}_{d_i}^{(t)}, \boldsymbol{D})$ for all genes (where $\boldsymbol{\alpha}^{(t)}$ and $\boldsymbol{\pi}_{d_i}^{(t)}$ denote the current estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}_{d_i}$, respectively). In this process, the unknown effect size parameters, β_i's, are analytically integrated out. In the M-step, we simply fit a logistic regression model

$$\log\left[\frac{\Pr(\gamma_i = 1|\boldsymbol{\alpha}^{(t)}, \boldsymbol{\pi}_{d_i}^{(t)}, \boldsymbol{D})}{\Pr(\gamma_i = 0|\boldsymbol{\alpha}^{(t)}, \boldsymbol{\pi}_{d_i}^{(t)}, \boldsymbol{D})}\right] = \alpha_0 + \alpha_1 d_i, \tag{4}$$

and use the resulting GLM estimates of α_0 and α_1 to obtain the updated $\boldsymbol{\alpha}^{(t+1)}$. Subsequently,

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$\boldsymbol{\pi}_{d_i}^{(t+1)}$ is computed by maximizing a simple K-dimensional multinomial likelihood function.

We start the algorithm from a set of arbitrary values of $\boldsymbol{\alpha}$ and $\{\boldsymbol{\pi}_l\}$ and iterate between the E and M steps until the pre-defined convergence criteria are met. The standard error of $\hat{\boldsymbol{\alpha}}$ is computed using a profile likelihood approach. The full details of the EM algorithm are provided in the Supplementary Material. Finally, we summarize the result of the gene set enrichment analysis by constructing a 95% confidence interval of α_1 from the EM output. Furthermore, we can obtain a p-value by computing a z statistic from $\hat{\alpha}_1$ and its standard error to test the null hypothesis,

$$H_0 : \alpha_1 = 0. \tag{5}$$

2.3 A latent contingency table interpretation

Here we provide an intuitive and general view of our enrichment analysis model and algorithm. Without loss of generality, we consider a single binary annotation for a particular gene set definition (i.e., a gene is either in or out of the annotated pathway/gene set). Now consider an ideal (but unrealistic) scenario where the true association status of each gene is indeed known. Under this setting, the enrichment analysis can be formulated as a 2×2 contingency table with the 4 cells indicating the 4 possible combinations of association and annotation status. Given the table, it is straightforward to compute the odds ratio to quantify the level of enrichment. It should be known that the enrichment computation from the contingency table is also statistically equivalent to fitting a simple logistic regression model. However in practice, the exact classification of the association status is unknown, and the above simple procedure is not directly applicable.

As mentioned in the introduction, the commonly applied two-stage procedure can be viewed as an ad-hoc procedure to fill in the unobserved 2×2 contingency table based on a simple classification rule. That is, a gene is classified as "associated" if the null hypothesis from a statistical test is rejected, and "unassociated" otherwise. This procedure is intuitive but has some notable caveats. Importantly, it should be clear that the filled" contingency table is not necessarily accurate compared to the underlying true table. This is because adopting hypothesis testing as a classification procedure preferentially restricts type I errors (false positives) but not the overall classification errors, in which type II (false negatives) errors make up a substantial proportion and are not controlled.

To demonstrate this point, we perform simulations and apply the two-stage procedure to estimate the enrichment parameter. In each simulation, we first generate a 2×2 contingency table given the true association and annotation status. We subsequently adjust the cell counts to reflect both the

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

type I and type II errors in the gene-level hypothesis testing. Finally, we compute the enrichment parameter α_1 using the adjusted contingency table. In summary, we find that the two-stage procedure consistently yields the enrichment estimates that are biased toward 0. With the type I error under control, the degree of bias is negatively correlated with the power of gene-level tests. An example from the simulation study is shown in Supplementary Figure 2. Interestingly, the lower power of the gene-level tests is also associated with a higher degree of variation in the enrichment estimates from the adjusted contingency table. Therefore, we conclude that the two-stage procedure can be inaccurate in estimating the enrichment parameter. Nevertheless, for enrichment testing, the direction of the bias from the two-stage procedure seems only to impact power but does not inflate the type I error that asserts enrichment when $\alpha_1 = 0$, as can be noted from results in Supplementary Table 2.

Our proposed EM algorithm, in this case, can be viewed as an iterative approach to fill in the unobserved contingency table, accounting for the uncertainty of the true binary association status. In the E-step of the proposed EM algorithm, we essentially fill in the table with the expected values of each gene. Note that

$$\mathrm{E}\left[\mathbf{1}_{\{\gamma_i=1\}}\right] = \Pr(\gamma_i = 1 \mid d_i = 1, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\pi}^{(t)}). \tag{6}$$

Hence, gene i contributes to the cell count of associated and annotated by an amount of $\Pr(\gamma_i = 1 \mid d_i = 1, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\pi}^{(t)})$, and to the cell count of unassociated and annotated by an amount of $1 - \Pr(\gamma_i = 1 \mid d_i = 1, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\pi}^{(t)})$. An obvious advantage of this approach is that the uncertainty of the association analysis result is naturally accounted for. The M-step is essentially the same statistical procedure given the contingency table is filled with the expected cell counts.

2.4 Use of multi-category gene set annotations

A unique advantage of BAGSE for enrichment analysis is that it can naturally handle multiple gene set annotations. Consider L potentially overlapping gene sets that we wish to evaluate simultaneously. A gene can be independently annotated (i.e., in or out) by an individual annotation. The joint annotation of a gene can be represented by a binary L-vector, which can take 2^L possible values: e.g., $(0, 0, ...0)$ indicates a gene is not annotated in any gene set, and $(1, 1, ...1)$ indicates a gene is annotated in all gene sets. Note that our parametric prior model (2) can naturally accommodate such multi-category annotation by regarding the corresponding d_i as a categorical variable coded by dummy variables (Section 1.2 of the Supplementary Material). For inference, instead of reporting a single enrichment coefficient (α_1), use of multi-category annotation leads to up to $2^L - 1$ enrichment coefficients for different annotation configurations in contrast to the baseline annotation $(0, 0, ..., 0)$.

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The general scheme applies to an arbitrary number of gene sets (L) considered, although a large number of L values may lead to expensive computational costs. It is possible to make simplifying assumptions to reduce computational complexity. For example, let binary indicator $u_{i,l}$ denote if gene i is annotated in the gene set l. We consider an additive prior model

$$\log \left[\frac{\Pr(\gamma_i = 1)}{\Pr(\gamma_i = 0)} \right] = \alpha_0 + \sum_{l=1}^{L} \alpha_l \, u_{i,l}, \tag{7}$$

which is computationally feasible for moderate to large L values. This particular prior model is also implemented in the BAGSE.

2.5 Local fdr control accounting for gene set enrichment

Another unique advantage of BAGSE is that the estimated enrichment information can be subsequently utilized in identifying truly associated candidate genes. Intuitively, accounting for the quantitative enrichment information boosts the power in identifying signals. This is accomplished by expanding a parametric empirical Bayes framework of local false discovery rate control procedure described in Stephens (2016). Specifically, we consider testing the null hypothesis $H_0 : \gamma_i = 0$ for all genes using the enrichment information. For each test, we evaluate the local fdr (lfdr) by plugging in the enrichment estimate $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\pi}}$,

$$\mathit{lfdr}_i := \Pr(\gamma_i = 0 \mid \hat{\beta}_i, \hat{s}_i, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\pi}}_{d_i}), \tag{8}$$

which is a byproduct of our EM algorithm and can be directly applied to control FDR.

Our proposed procedure also provides a principled solution to deal with non-exchangeable multiple hypothesis testing. For example, one may suspect that genes in certain annotated pathways are more likely to be true signals. Such prior expectation can be precisely expressed by equation (2) in our model. In particular, if $\alpha_1 > 0$ and gene i is a member of the gene set of interest, then it has a higher prior probability to be a genuine signal compared to its counterparts absent from the gene set. Furthermore, the enrichment estimation procedure outlined in the EM algorithm allows us to effectively learn the value of α_1 from the observed data. The local fdr computed in (8) combines the enrichment prior and the likelihood information observed from data: if a gene set is estimated to be enriched, the lfdr's of its member genes are down-weighted by the priors.

In addition to controlling FDR for testing presence or absence of signals (i.e., γ_i's), our approach can be extended to control the local false sign rates (lfsr) (Stephens, 2016), which focus on the signals

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

whose effects can be identified robustly. In particular, we compute lfsr for gene i by

$$lfsr_i := \min[\Pr(\beta_i \geq 0 \mid \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\pi}}_{d_i}, \hat{\beta}, \hat{s}), \Pr(\beta_i \leq 0 \mid \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\pi}}_{d_i}, \hat{\beta}, \hat{s})], \tag{9}$$

which is interpreted as the error probability in determining the sign of the effect for gene i.

3 Results

3.1 Simulation studies

We use numerical simulations to benchmark the performance of the proposed Bayesian gene-set enrichment analysis procedure. Our particular focuses are put on examining the accuracy of the enrichment estimates and its performance in enrichment testing.

In each simulated dataset, we consider the analysis of 10,000 genes, each of which is assigned a binary true association status based on the enrichment parameters (α_0, α_1) and pre-defined annotations. For each gene, we assume an association z-score is available for the enrichment analysis: for un-associated genes, we draw the z-scores from the standard normal distribution; for the associated genes, the z-scores are simulated from a t-distribution with the degree of freedom = 10 (which mimics the long-tailed effect size distribution commonly observed in practice, see Supplementary Figure 1 for example). We set the enrichment parameter $\alpha_0 = -1$ throughout, while varying the values of α_1 and the proportion of annotated genes (denoted by q) across all simulations. We vary α_1 values from 0 to 1, due to all investigated methods having power similarly near 100% power as α_1 increases to above 1. We use values for q ranging from 1% to 20%, due to that being a realistic range for q based on the hierarchical nature of popular databases such as KEGG or GO. Each parameter set is used to generate 5000 different datasets.

3.1.1 Evaluation of enrichment parameter estimation

We first examine the point estimate of the enrichment parameter from the BAGSE analysis procedure. For comparison, we also compute the enrichment estimate from the two-stage procedure. The analysis results of the simulated datasets are summarized in Fig. 2 for annotation proportion $q = 10\%$. It is clear that the proposed approach consistently yields unbiased enrichment estimates across all α_1 values. Also consistent with the previous results, the estimates from the two-stage approach are biased towards 0. Across different q values, the estimates from the Bayesian procedure remain

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

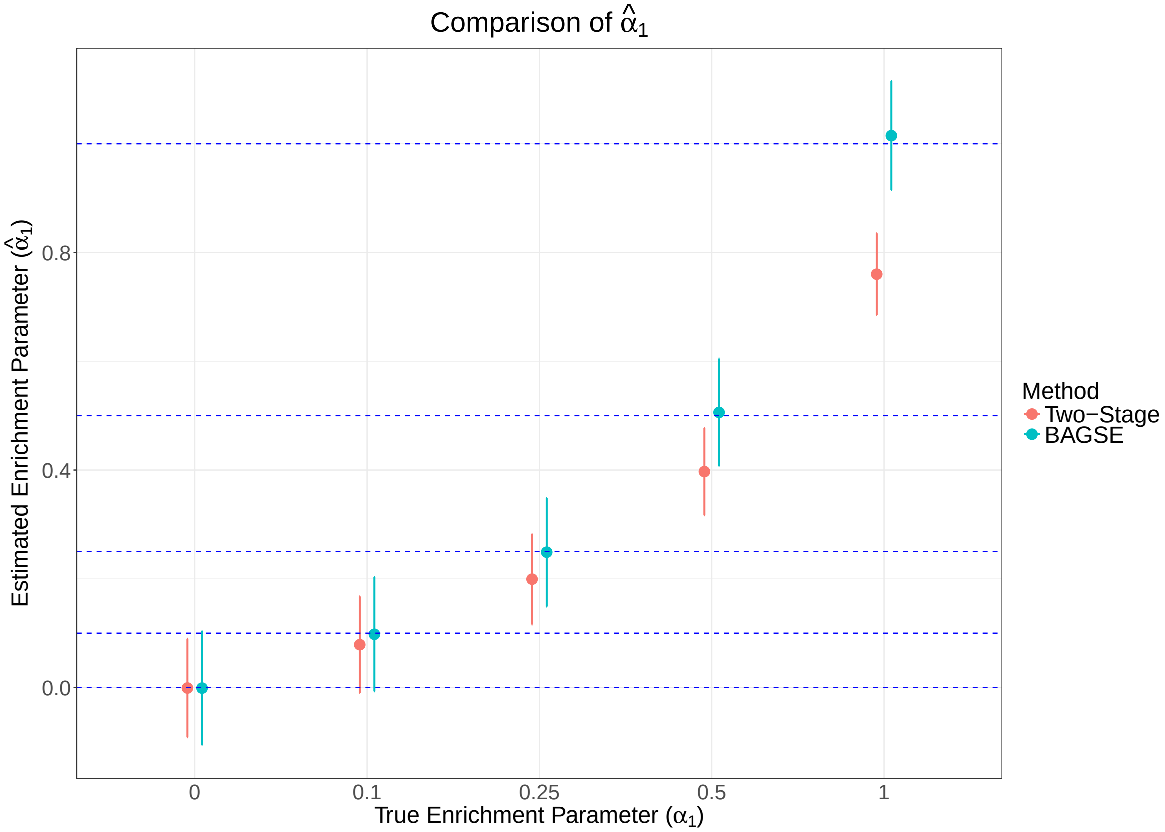

Figure 2: A comparison of the enrichment parameter estimates between the two-stage approach and the proposed method, with standard errors represented as error bars. BAGSE appears to give an unbiased estimate of α_1, while the two-stage approach's estimate grows more severely biased as the enrichment parameter grows higher.

unbiased. But we observe a pattern that shows a lower level of variation in $\hat{\alpha}_1$ when q increases towards 0.5. The latent contingency table interpretation can intuitively explain this phenomenon: as q tends to 0 (or 1), the underlying table becomes more imbalanced, causing increased uncertainty for the enrichment parameter. Results from simulations using all parameter sets can be found in Supplementary Table 1 and Supplementary Figures 3, 4 and 5.

We conduct additional simulations to illustrate the ability of the proposed method in estimating enrichment parameters using annotations from multiple overlapping gene sets. In summary, we find that BAGSE estimates in such scenario remain unbiased and accurate. The details of these additional simulations are described in Section 2 of the Supplementary Material.

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

3.1.2 Power comparison in enrichment testing

Next, we proceed to examine the performance of various methods, including BAGSE, the two-stage approach, and the popular GSEA method, in testing the enrichment hypothesis: $H_0 : \alpha_1 = 0$. We apply both the unweighted and weighted forms of the GSEA procedure. The unweighted GSEA procedure corresponds to a standard Kolmogorov-Smirnov test (by comparing the distribution of the absolute value of z-scores, or p-values, between annotated and unannotated genes). We use the default weight recommended in the original GSEA paper Subramanian *et al.* (2005) to perform the weighted GSEA procedure.

The simulation results for when $q = 10\%$ are summarized in Fig. 3. We conclude that all methods properly control type I errors at 5% level, based on the proportion of tests rejected when $\alpha_1 = 0$ being below 0.05 (denoted by the dotted line) for all methods. BAGSE and the weighted GSEA method are top performers at every combination of simulation parameters, constantly outperforming the unweighted GSEA and the two-stage approaches by a significant margin, especially at the intermediate α_1 values of 0.1 and 0.25. The power difference between BAGSE and the weighted GSEA is generally negligible. In our simulation setting, when $\alpha_1 \to 1$, all methods seemingly achieve the perfect power to reject the null hypothesis. Additionally, we observe that the power to detect enrichment improves as q increases towards 0.5. This phenomenon can be similarly explained by the latent contingency table interpretation of our proposed model: the standard error of the enrichment estimate decreases as the proportion of the annotation increase towards 0.5. Results from simulations using all parameter sets for all methods can be found in Supplementary Table 2 and Supplementary Figures 6, 7 and 8.

3.1.3 Gene discovery incorporating enrichment quantification

Finally, we examine the power in identifying truly associated genes when gene set enrichment information is explicitly considered. Specifically, we compute the local fdr for each gene using (8) where the estimated enrichment parameter is plugged in, and control the overall FDR at 5% level. For a baseline comparison, we use both the q-value procedure Storey *et al.* (2003) and the local fdr procedure (both are implemented in the R package "*qvalue*") to control FDR at the same level ignoring the gene set information.

Our results indicate that all methods properly control the FDR at the desired level. However, the proposed procedure accounting for enrichment quantification consistently outperforms the q-value and the local fdr procedure ignoring the enrichment information in realized power (Figure 4).

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

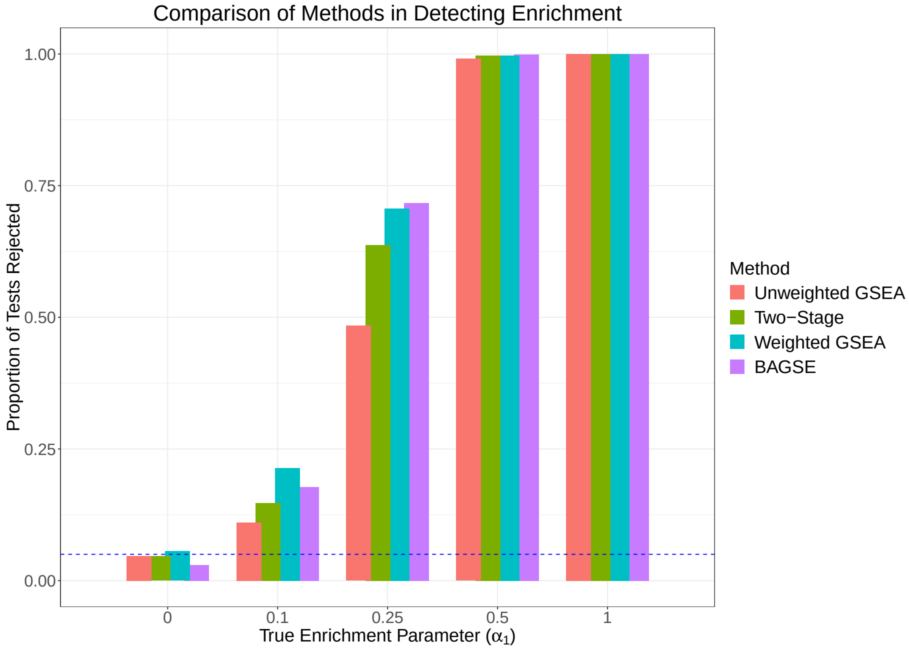

Figure 3: A comparison of type I error and power to detect gene set enrichment using unweighted GSEA, two-stage approach, weighted GSEA, and BAGSE. The y-axis denote the proportions of simulated data sets where the null hypothesis are rejected The leftmost columns, corresponding to the null model $H_0 : \alpha_1 = 0$, represent the type I error rates for the methods. As the enrichment parameter takes non-zero values, the data are simulated under the alternative scenarios, and the corresponding columns represent the power. All methods control the type I errors properly. As expected, power for all methods increases as the true enrichment parameter increases. BAGSE and weighted GSEA outperform the other methods in this regard.

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

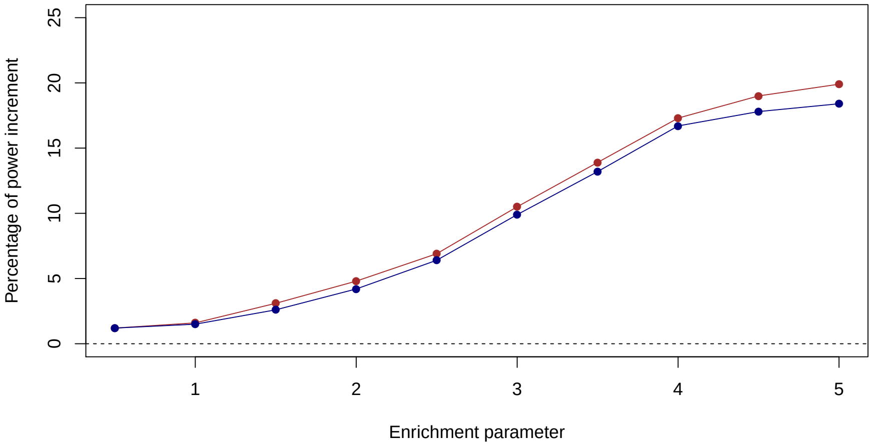

Figure 4: Power increment in gene discovery when accounting for gene set enrichment information. The percentages of power increment by using enrichment information are compared to two standard FDR control procedures: the q-value method (brown line) and the local fdr method (blue line). All methods control FDR at 5% level.

As expected, the improvement of power is positively correlated with the level of enrichment. In our simulation setting, we observe a modest increase in power ($\sim$ 1.5%, or 20 more true positive discoveries per simulated dataset) when the enrichment parameter is ~ 1; as α_1 reaches 5, the power boost becomes much more substantial ($\sim$ 20% improvement in power, or 400 more true discoveries per simulated dataset).

3.2 Real data application I: differential expression experiment

Next, we apply BAGSE to the experimental data from Moyerbrailean *et al.* (2016), which considers the differential expression of genes under the treatment of the glucocorticoid. The study profiles 20,896 genes using RNA-seq. A p-value is obtained for each gene using the software package DESeq2 (Love *et al.*, 2014). We convert the p-values into the z-scores, using the estimated effect sizes from the study to determine the corresponding signs. The experiments are also carried out in multiple tissues. For demonstration purpose, we select the results gathered from the peripheral blood mononuclear cells. Because of the known nature of these cells in their response to glucocorticoid, we expect genes involved in pathways associated with immune response to be enriched.

There are 21 KEGG pathways involved in the immune responses. Of these, 15 contain genes that are in our dataset. We first analyze these 15 pathways separately using BAGSE. The enrichment

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

estimate for each of these pathways is provided in Supplementary Table 3. We observe that the majority of the examined pathways (14 out of 15) are shown to be enriched with DE genes, and their enrichment levels can be straightforwardly compared by utilizing the quantified enrichment estimates. In particular, we find that the Intestinal immune Network for IgA Production pathway seemingly shows an extremely high level of enrichment ($\alpha_1 = 9.34$ with 95% CI [5.38, 13.29]).

Because each pathway only annotates a small proportion of genes, the confidence intervals are typically large (Supplementary Table 3). Following Carbonetto and Stephens (2013), we pool the genes annotated in the 15 KEGG pathways to form a general category of the gene set and examine the enrichment of DE genes in this gene set representing general immune responses. In total, 2.7% of the 20,896 genes are annotated in the aggregated gene set.

We apply BAGSE, the two-stage approach, and GSEA with the two different weighting schemes to conduct gene set enrichment analysis. These results are summarized in Table 1. BAGSE detected strong enrichment for the pooled immune response gene sets, with an enrichment log odds ratio of 1.31 (95% CI [0.96,1.66]), which corresponds to a p-value of 2.2×10^{-13}. The two-stage approach also detects enrichment, with an enrichment log odds ratio of 0.87 (95% CI [0.63, 1.11]). The unweighted GSEA method detects significant enrichment with a p-value of 1×10^{-7}. The weighted GSEA method also detects significant enrichment with an estimated p-value $< 1 \times 10^{-3}$.

Method	P-value	Enrichment estimate (95% CI)
BAGSE	2.2×10^{-13}	1.31 (0.96, 1.66)
Two-Stage	1.2×10^{-12}	0.87 (0.63, 1.11)
Unweighted GSEA	1×10^{-7}	-
Weighted GSEA	¡10^{-3*}	-

Table 1: The comparison of significance in detecting enrichment between methods for data from the Differential Expression of genes under treatment of glucocorticoid. As expected, all methods show significance. Note that the inexact value for the weighted GSEA p-value is due to its calculation using 1,000 permutations (recommended by Subramanian *et al.* (2005)).

Additionally, we examine the power improvement in identifying DE genes by incorporating the quantified enrichment information. Without using the enrichment information, we find that the q-value procedure identifies 1496 genes at 5% FDR level (the local fdr procedure, implemented in the q-value package, identifies 1527 genes). By incorporating the enrichment estimates, BAGSE identifies 1617 genes at the same FDR control level, which amounts to 8% increase in comparison to the q-value procedure (or 6% increase to the local fdr procedure). Overall, these results appear consistent with our simulation studies.

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

3.3 Real data application II: transcriptome-wide association analysis

For the second illustration with real data, we perform gene set enrichment analysis using the association results generated from TWAS. TWAS is a principled approach based on Mendelian randomization (MR), and its units of analysis are genes whose expression profiles are systematically investigated in expression quantitative trait loci (eQTL) studies. The results from TWAS analysis yield insight into the mode of causality from gene expressions to complex traits (Gamazon *et al.*, 2015; Gusev *et al.*, 2018; Zhu *et al.*, 2016). For our illustration, we apply the software package MetaXcan (Barbeira *et al.*, 2016) to examine the association between predicted gene expression levels in the whole blood and the lipid trait of high-density lipoproteins (HDLs). Our TWAS analysis utilizes the whole blood data from the GTEx project (version 6p) (Consortium *et al.*, 2017) and GWAS summary statistics from the Global Lipids Genetics Consortium (Willer *et al.*, 2013). Specifically, we first compute the predicted gene expression levels using the GTEx data using the elastic-net algorithm following Gamazon *et al.* (2015) and obtain the association z-scores between each gene and the HDL trait using MetaXcan.

We aim to examine two KEGG pathways (Neurotrophin signaling pathway and Adipocytokine signaling pathway), which are previously implicated in the GWAS analysis by Willer *et al.* (2013). Note that Willer *et al.* (2013) applied a proximity-based approach linking a GWAS hit to the nearest protein-coding gene and used the MAGENTA procedure (Segrè *et al.*, 2010), which is similar to the two-stage procedure, for the pathway enrichment analysis.

We start our analysis by identifying a group of genes overlapping between the protein-coding genes annotated by the KEGG database and the eGenes (i.e., the genes harboring at least one *cis*-eQTL) implicated by analyzing the GTEx whole blood data. This is because non-trivial expression predictions based on genetic variants are only available for eGenes. In the end, we identify 4,604 genes for the pathway enrichment analysis. We examine these genes for enrichment in the two pathways of interest, studying one pathway at a time. In our data, there are only 15 genes annotated in the Neurotrophin Signaling pathway, and only 12 genes annotated in the Adipocytokine Signaling Pathway. We apply BAGSE (single-category version), the two-stage approach, and the GSEA (weighted and unweighted) method to analyze each pathway. The results for enrichment testing and estimation are summarized in Table 2 and Table 3. In brief, none of the four methods detect significant enrichment (at 5% level) in either pathway. However, the directions of the enrichment levels inferred by BAGSE are consistent with the previous MAGENTA results reported in Willer *et al.* (2013). Due to the unbiased nature of BAGSE's enrichment estimate, as shown in the simulations, we believe this point

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

estimate to have meaning in spite of the lack of statistical significance. We suspect that the lack of significant finding from the TWAS results is likely due to the lack of power in eQTL studies: with a limited number of sample size (343), the eQTL analysis does not have sufficient power to identify eQTLs with small to modest effect sizes. Consequently, the expression levels for a large number of protein-coding genes are poorly (or unable to be) predicted. Nevertheless, we are confident that this issue can be resolved when the well-powered eQTL datasets become available in the near future.

Method	P-value	Enrichment estimate (95% CI)
BAGSE	0.34	1.43 (-1.46, 4.31)
Two-Stage	0.23	1.27 (-0.78, 3.31)
Unweighted GSEA	0.73	-
Weighted GSEA	0.08	-

Table 2: A look at the significance in the testing for enrichment of the neurotrophin signaling pathway, for the four discussed methods. All four methods show no significance at the 5 % significance level, likely due to the relatively small amounts of annotated genes identified from the eQTL analysis.

Method	P-value	Enrichment Estimate (95% CI)
BAGSE	0.35	1.45 (-1.62, 4.52)
Two-Stage	0.15	1.51 (-0.55, 3.56)
Unweighted GSEA	0.48	-
Weighted GSEA	0.14	-

Table 3: A summary of the significance for testing the enrichment of the adipocytokine signaling pathway, for the four methods. Again, all four methods do not show significance at the 5 % significance level. The low number of annotated genes that show evidence of association can severely increase the standard error of the enrichment estimate. This is likely the reason behind the unexpected result of the two-stage point estimate being higher than the BAGSE point estimate here.

4 Conclusion and discussion

In this paper, we have introduced an empirical Bayes procedure, denoted as BAGSE, for gene set enrichment analysis. We have shown, through simulations and real data analysis, that the proposed approach provides a principled inference procedure to estimate the level of enrichment for a gene set, avoiding the caveats of the two-stage procedure. In addition, the proposed Bayesian method maintains strong power in testing enrichment, as compared to other popular approaches to gene set enrichment analysis. Finally, we show that the enrichment estimates from the proposed approach can be subsequently utilized to improve power for gene-level testing.

It should be noted that our approach can be straightforwardly applied to simultaneously estimate

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the enrichment of multiple gene sets with overlapping genes. This unique feature can be critical because, as seen in many commonly used gene pathway definitions (e.g., the KEGG pathway database), there are many genes involved in multiple pathways. This flexibility in gene annotation intrinsic to the proposed Bayesian model gives it a distinct advantage as an enrichment analysis method.

During our simulations, we noted that both power to detect enrichment and enrichment estimation were heavily dependent on the proportion of genes annotated. Our latent contingency table interpretation can partially explain this phenomenon: even if all association status is indeed observed, the standard error of the enrichment estimate is known to be negatively correlated with the smallest cell count, and lower annotation proportion tends to decrease the smallest cell count (which typically is the cell corresponds to the count of both annotated and associated genes). All enrichment testing and estimation methods would be affected by a low proportion of annotated genes, reflected by losing either power or precision. In general, the quality of the gene set definition has a direct impact on results from enrichment analysis. Defining highly specific gene pathways is an ongoing challenge.

In our real data applications, we show the examples of pathway enrichment analysis based on the TWAS result. We expect that this type of analysis will become increasingly popular in the field of systems biology. The TWAS analysis provides a causal inference framework linking individual genes to a complex trait of interest. The gene set enrichment analysis based on TWAS results helps uncover potentially *causal* biological pathways. Although our TWAS enrichment analysis seems inconclusive, we suspect that it is likely due to lack of power in the available eQTL datasets. With the fast growth of available eQTL data, the power of the TWAS analysis is expected to be improved significantly in the near future. Consequently, we expect that the proposed gene set enrichment analysis will help unveil many causal pathways relevant to complex diseases.

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Materials

A EM algorithm for parameter estimation

Recall that $\boldsymbol{D} := \{(\hat{\beta}_i, \hat{s}_i) : i = 1, ..., M\}$ and $\boldsymbol{\gamma} := (\gamma_1, ..., \gamma_M)$. We consider the complete data likelihood $\Pr(\boldsymbol{D}, \boldsymbol{\gamma} \mid \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi})$ by treating $\boldsymbol{\gamma}$ as missing data:

$$\begin{aligned} &\Pr(\boldsymbol{D}, \boldsymbol{\gamma} \mid \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\pi}_d) \\ &= \Pr(\boldsymbol{D} \mid \boldsymbol{\gamma}, \boldsymbol{d}, \boldsymbol{\pi}_d) \Pr(\boldsymbol{\gamma} \mid \boldsymbol{d}, \boldsymbol{\alpha}) \\ &= \prod_{i=1}^{M} P(\hat{\beta}_i, \hat{s}_i \mid \gamma_i, d_i, \boldsymbol{\pi}_{d_i}) \prod_{i=1}^{M} \Pr(\gamma_i \mid \boldsymbol{\alpha}, d_i). \end{aligned} \tag{10}$$

Our model assumes that

$$\hat{\beta}_i \mid \beta_i \sim \mathrm{N}(0, \hat{s}_i^2),$$

and the prior for β_i follows a K-component mixture normal distribution depending on the corresponding gene set annotation, i.e.,

$$\beta_i \mid d_i, \gamma_i \sim \sum_{k=1}^{K} \pi_{d_i,k} \mathrm{N}(\beta_i, \gamma_i\, \phi_k^2).$$

Equivalently, it follows that

$$\hat{\beta}_i \mid d_i, \gamma_i \sim \sum_{k=1}^{K} \pi_{d_i,k} \mathrm{N}(0, \gamma_i\, \phi_k^2 + \hat{s}_i^2).$$

A.1 Model reparameterization

We now consider a general case with L potentially overlapping gene sets. By the coding convention introduced in section 2.4, the gene set annotation d_i is a categorical variable representing R multiple mutually exclusive categories, where $2 \leq R \leq 2^L$. Specifically, we use $d_i = 0$ to denote the baseline level that a gene is not annotated in any of the L gene sets. Under this formulation, the equation (2) in the main text is generalized to

$$\log\left[\frac{\Pr(\gamma_i = 1 \mid d_i = r)}{\Pr(\gamma_i = 0 \mid d_i = r)}\right] = \alpha_r', \quad \text{for } r = 0, 1, ..., R-1, \tag{11}$$

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in this coding system. In case that $L = 1$, this coding system is compatible to equation (2). In particular, $\alpha_0' = \alpha_0$ and $\alpha_1' = \alpha_0 + \alpha_1$. Given K pre-defined mixture components for the effect size distribution in the alternative models, $\boldsymbol{\pi}$ is represented by a $R \times K$ matrix with $(r+1)$-th row vector representing the weights for annotation $d_i = r$ $(r = 0, ..., R-1)$.

For each gene i, we introduce a $[R(K+1)]$-dimension latent indicator, $\boldsymbol{\eta}_i$, to re-parametrize the complete data likelihood function (10). Note that all entries of a valid $\boldsymbol{\eta}_i$ vector take value 0 except that a single entry takes value 1. In particular, we define

$$\eta_{i,\, r(K+1)+1} = 1 \leftrightarrow \{d_i = r;\, \gamma_i = 0\},$$

for $r = 0, ..., R-1$, and

$$\hat{\beta}_i \mid \eta_{i,\, r(K+1)+1} = 1 \sim \mathrm{N}(0, \hat{s}_i^2). \tag{12}$$

That is, this subset of latent indicators and the corresponding parameters define the null (i.e., non-association) models for each annotation category r. Similarly, we set

$$\eta_{i,\, r(K+1)+k+1} = 1 \leftrightarrow \{d_i = r;\, \gamma_i = 1 \text{ and } \beta_i \sim \mathrm{N}(0, \phi_k^2)\,\},$$

for $k = 1, ..., K$, and $r = 0, ..., R-1$. Accordingly,

$$\hat{\beta}_i \mid \eta_{i,\, r(K+1)+k+1} = 1 \sim \mathrm{N}(0, \hat{s}_i^2 + \phi_k^2). \tag{13}$$

Furthermore, the prior weights for the indicator $\boldsymbol{\eta}_i$ are simple functions of the original parameters $(\boldsymbol{\alpha}, \boldsymbol{\pi})$, i.e.,

$$\begin{aligned} w_{i,\, r(K+1)+1} &:= \Pr(\eta_{i,\, r(K+1)+1} = 1) \\ &= \Pr(\gamma_i = 0 \mid d_i = r) \\ &= \frac{1}{1+\exp(\alpha_r')}. \end{aligned} \tag{14}$$

Similarly,

$$\begin{aligned} w_{i,r(K+1)+k+1} &:= \Pr(\eta_{i,\, r(K+1)+k+1} = 1) \\ &= \pi_{\delta(r(K+1)+k+1)} \cdot \frac{\exp(\alpha_r')}{1+\exp(\alpha_r')}, \quad \text{for } k = 1, ..., K. \end{aligned} \tag{15}$$

Note that we use the function $\delta(j) = (d_i = r, k)$ to map the single index $j = r(K+1)+k+1$ to the double indcies $d_i = r$ and k in the original $(\boldsymbol{\alpha}, \boldsymbol{\pi})$ notation for suitable integer $j \in S$, where S denote the index set excluding $j = r(K+1)+1$, for $r = 0, ..., R-1$. Similarly, we use $\delta^{-1}(d_i = r, k) =$

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$r(K+1)+k+1$ to denote the inverse mapping.

Under this alternative parameterization, the complete data likelihood can be expressed by

$$\begin{aligned}
&\prod_{i=1}^{M} P(\hat{\beta}_i, \hat{s}_i \mid \boldsymbol{\eta}_i) \prod_{i=1}^{M} \Pr(\boldsymbol{\eta}_i \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, d_i) \\
&= \prod_{i=1}^{M} \prod_{j=1}^{R(K+1)} \left(P(\hat{\beta}_i, \hat{s}_i \mid \eta_{i,j}=1) \right)^{\mathbf{1}_{\{\eta_{i,j}=1\}}} \qquad (16) \\
&\cdot \prod_{i=1}^{M} \left[\left(\frac{1}{1+\exp(\alpha'_{d_i})} \right)^{\mathbf{1}_{\{\eta_{i,d_i(K+1)+1}=1\}}} \left(\frac{\exp(\alpha'_{d_i})}{1+\exp(\alpha'_{d_i})} \right)^{1-\mathbf{1}_{\{\eta_{i,d_i(K+1)+1}=1\}}} \prod_{j\in S} \pi_{\delta(j)}^{\mathbf{1}_{\{\eta_{i,j}=1\}}} \right],
\end{aligned}$$

Thus, the complete data log-likelihood is given by

$$\begin{aligned}
l(\boldsymbol{\alpha}, \boldsymbol{\pi}) = & \sum_{i=1}^{M} \sum_{j=1}^{R(K+1)} \mathbf{1}_{\{\eta_{i,j}=1\}} \log P(\hat{\beta}_i, \hat{s}_i \mid \eta_{i,j}=1) \\
& + \sum_{i=1}^{M} (1 - \mathbf{1}_{\{\eta_{i,d_i(K+1)+1}=1\}}) \, (\alpha'_{d_i}) \qquad (17) \\
& - \sum_{i=1}^{M} \log \left[1+\exp(\alpha'_{d_i})\right] + \sum_{i=1}^{M} \sum_{j\in S} \mathbf{1}_{\{\eta_{i,j}=1\}} \log \pi_{\delta(j)}
\end{aligned}$$

A.2 E-step

In the E-step for the t-th iteration of the EM algorithm, we compute the expected value of the missing data, $\boldsymbol{\eta}_i$, conditional on the observed data summary statistics, gene set annotation, and the current estimates of the parameters, $\boldsymbol{\alpha}^{(t)}, \boldsymbol{\pi}^{(t)}$, i.e.,

$$\mathrm{E}\left(\mathbf{1}_{\{\eta_{i,j}=1\}} \mid \hat{\beta}_i, \hat{s}_i, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\pi}^{(t)} \right) = \Pr\left(\eta_{i,j}=1 \mid \hat{\beta}_i, \hat{s}_i, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\pi}^{(t)} \right), \forall i, j.$$

The computation can be carried out by the Bayes rule, i.e.,

$$\Pr\left(\eta_{i,j}=1 \mid \hat{\beta}_i, \hat{s}_i, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\pi}^{(t)} \right) = \frac{\Pr(\eta_{i,j}=1 \mid \boldsymbol{\alpha}^{(t)}, \boldsymbol{\pi}^{(t)}) P(\hat{\beta}_i \mid \eta_{i,j})}{\sum_{j'=1}^{R(K+1)} \Pr(\eta_{i,j'}=1 \mid \boldsymbol{\alpha}^{(t)}, \boldsymbol{\pi}^{(t)}) P(\hat{\beta}_i \mid \eta_{i,j'})}, \qquad (18)$$

where the relevant likelihood and prior functions are given by (12), (13), (14), and (15), respectively.

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A.3 M-step

In the M-step for the t-th iteration, we update the estimate of the enrichment parameter by finding

$$\begin{aligned}\boldsymbol{\alpha}^{(t+1)} = \arg\max_{\boldsymbol{\alpha}} \Bigg(& \sum_{i=1}^{M} \Big(1 - \Pr(\eta_{i,d_i(K+1)+1} = 1 \mid \hat{\beta}_i, \hat{s}_i, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\pi}^{(t)})\Big) \cdot (\alpha'_{d_i}) \\ & - \sum_{i=1}^{M} \log\left[1 + \exp(\alpha'_{d_i})\right] \Bigg),\end{aligned} \tag{19}$$

which is equivalent to fitting a logistic regression model with a single categorical covariate, $\boldsymbol{d}$, and the binary response variables, η's, replaced by their corresponding posterior probabilities. Our implementation uses the Newton-Raphson algorithm to perform numerical optimization. Similarly, we update the estimate of $\boldsymbol{\pi}$ by finding

$$\boldsymbol{\pi}^{(t+1)} = \arg\max_{\boldsymbol{\pi}} \left(\sum_{i=1}^{M} \sum_{j \in S} \Pr(\eta_{i,j} = 1 \mid \hat{\beta}_i, \hat{s}_i, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\pi}^{(t)}) \, \log \pi_{\delta(j)} \right), \tag{20}$$

under the constraint

$$\sum_{k=1}^{K} \pi_{r,k} = 1, \forall\, r = 0, 1, ..., R-1.$$

Note that maximization can be achieved analytically. Specifically,

$$\pi_{r,k}^{(t+1)} = \frac{\sum_{i=1}^{M} \Pr(\eta_{i,\delta^{-1}(l,k)} = 1 \mid \hat{\beta}_i, \hat{s}_i, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\pi}^{(t)})}{\sum_{i=1}^{M} \sum_{k'=1}^{K} \Pr(\eta_{i,\delta^{-1}(l,k')} = 1 \mid \hat{\beta}_i, \hat{s}_i, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\pi}^{(t)})}. \tag{21}$$

The software implementation of the EM algorithm can initialize the parameters of interest at arbitrary starting points. By default (i.e., without explicit user specification), we set

$${\alpha'_0}^{(0)} = \cdots = {\alpha'_{R-1}}^{(0)} = 0,$$

and

$$\pi_{r,k}^{(0)} = \frac{1}{K}, \quad \forall\, l, k.$$

We iterate between the E-step and the M-step until pre-defined convergence criteria is satisfied.

A.4 Grid construction

We follow the procedure described in Stephens (2016) to construct a dense set of $\{\phi_k^2\}$ in a data-driven fashion. Specifically, we construct a series of ϕ_k values within the range of $(\phi_{\min}, \phi_{\max})$, where

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$\phi_{\min} = \min(\hat{s}_i)/10$ and $\phi_{\max} = \max(\hat{\beta}_i^2 - \hat{s}_i^2)$. We start with $\phi_1 = \phi_{\min}$ and set

$$\frac{\phi_{k+1}}{\phi_k} = \sqrt{2}, \text{ for } k = 1, 2, ..., \tag{22}$$

until $\phi_{K+1} > \phi_{\max}$.

B Simulation with annotations from multiple gene sets

We perform additional simulations to illustrate the utility of estimating enrichment parameters from multiple overlapping gene sets.

As a proof of concept, we simulate two sets of independent gene set annotations for 10,000 genes in each simulated data set where $q \approx 40\%$ of genes are independently annotated in each gene set. Let binary indicators $u_{i,1}$ and $u_{i,2}$ denote the annotation status of gene i in the two gene sets, respectively. As a result, 36% genes are expected to be not annotated for any of the gene sets (denoted by $[u_{i,1} = 0, u_{i,2} = 0]$ or $d_i = 0$); 24% genes are expected to be annotated only for the first gene set (denoted by $[u_{i,1} = 1, u_{i,2} = 0]$ or $d_i = 1$); 24% genes are expected to be annotated only for the second gene set (denoted by $[u_{i,1} = 0, u_{i,2} = 1]$ or $d_i = 2$); and 16% genes are expected to be annotated for both gene sets (denoted by $[u_{i,1} = 1, u_{i,2} = 1]$ or $d_i = 3$).

For each gene i, we draw its association status γ_i from a Bernoulli(p_i) distribution, where

$$\text{logit}(p_i) = -1 + 0.80\, u_{i,1} + 1.00\, u_{i,2} + 0.55\, u_{i,1} \cdot u_{i,2}. \tag{23}$$

Conditional on $\gamma_i = 1$, the corresponding z score is drawn from a t-distribution with 10 degree of freedom for all d_i values. Otherwise, when $\gamma_i = 0$, the z score is drawn from the standard normal distribution. We generate 1,000 simulated data sets using the above scheme.

When analyzing the simulated data, we do not assume any knowledge of (23) and simply apply the general combinatorial annotation model and the EM algorithm described in section 2.4 of the main text and section 1 of the supplementary material. Specifically, we examine the enrichment estimates in contrast to the baseline annotation (annotation "00"), i.e., the estimates of $\alpha_1 = \alpha_1' - \alpha_0'$ (truth = 0.80) for genes annotated only in set 1 (annotation "10"), $\alpha_2 = \alpha_2' - \alpha_0'$ (truth = 1.00) for genes annotated only in set 2 (annotation "01"), and $\alpha_3 = \alpha_3' - \alpha_0'$ (truth $= 0.80 + 1.00 + 0.55 = 2.35$) for genes annotated in both set 1 and set 2 (annotation "11"), respectively. Across 1,000 simulations, we find that the proposed EM algorithm yields unbiased estimates of enrichment parameters (Figure

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

5). We find that all three estimates are seemingly unbiased: the average point estimates for the three mutually exclusive annotations are 0.806, 1.016 and 2.377, respectively.

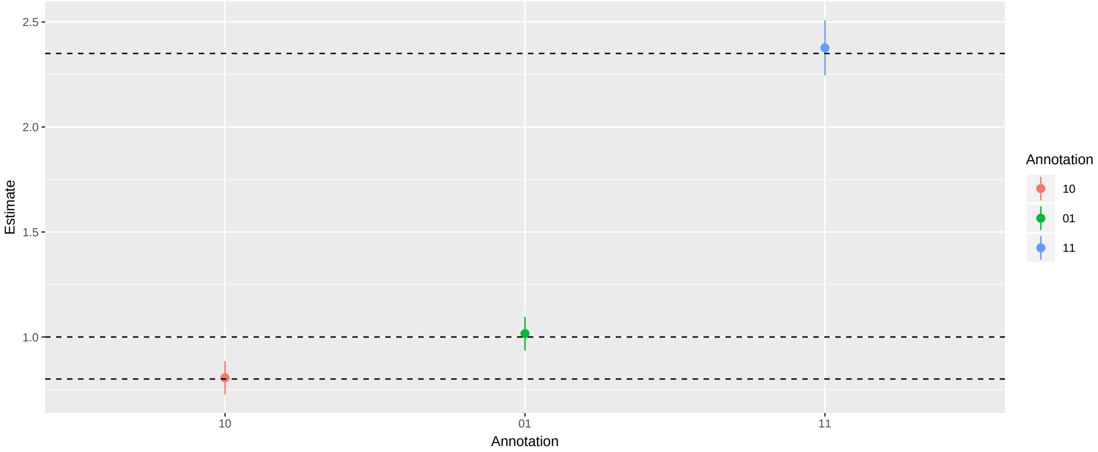

Figure 5: Simultaneous enrichment estimation of overlapping gene sets by BAGSE in simulation studies. Two overlapping gene sets are used in the simulation. Annotation "00" denotes the genes not annotated in any gene set, these genes are the baseline for the enrichment comparison. Annotation "10" denotes the genes annotated only in the first gene set, annotation "01" denotes the genes annotated only in the second gene set, and annotation "11" denotes the genes annotated in both gene sets. The average point estimates and the corresponding standard errors from 1000 independent simulated data sets are plotted. The dashed lines denote the underlying true enrichment levels for each group.

References

Barbeira, A., Shah, K. P., Torres, J. M., Wheeler, H. E., Torstenson, E. S., Edwards, T., Garcia, T., Bell, G. I., Nicolae, D., Cox, N. J., *et al.* (2016). Metaxcan: summary statistics based gene-level association method infers accurate predixcan results. *BioRxiv*, page 045260.

Carbonetto, P. and Stephens, M. (2013). Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for il-2 signaling genes in type 1 diabetes, and cytokine signaling genes in crohn's disease. *PLoS genetics*, **9**(10), e1003770.

Chang, Y., Glass, K., Liu, Y.-Y., Silverman, E. K., Crapo, J. D., Tal-Singer, R., Bowler, R., Dy, J., Cho, M., and Castaldi, P. (2016). Copd subtypes identified by network-based clustering of blood gene expression. *Genomics*, **107**(2), 51–58.

Consortium, G. *et al.* (2017). Genetic effects on gene expression across human tissues. *Nature*, **550**(7675), 204.

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Elovainio, M., Taipale, T., Seppälä, I., Mononen, N., Raitoharju, E., Jokela, M., Pulkki-Råback, L., Illig, T., Waldenberger, M., Hakulinen, C., *et al.* (2015). Activated immune–inflammatory pathways are associated with long-standing depressive symptoms: evidence from gene-set enrichment analyses in the young finns study. *Journal of psychiatric research*, **71**, 120–125.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., *et al.* (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, **47**(9), 1091.

Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H. K., Reshef, Y., Song, L., Safi, A., McCarroll, S., Neale, B. M., *et al.* (2018). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nature genetics*, **50**(4), 538.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., *et al.* (2009). Chromatin signature reveals over a thousand highly conserved large non-coding rnas in mammals. *Nature*, **458**(7235), 223.

Hass, J., Walton, E., Wright, C., Beyer, A., Scholz, M., Turner, J., Liu, J., Smolka, M. N., Roessner, V., Sponheim, S. R., *et al.* (2015). Associations between dna methylation and schizophrenia-related intermediate phenotypesa gene set enrichment analysis. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, **59**, 31–39.

Keshava Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., *et al.* (2008). Human protein reference database2009 update. *Nucleic acids research*, **37**(suppl_1), D767–D772.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, **15**(12), 550.

Maruschke, M., Hakenberg, O. W., Koczan, D., Zimmermann, W., Stief, C. G., and Buchner, A. (2014). Expression profiling of metastatic renal cell carcinoma using gene set enrichment analysis. *International Journal of Urology*, **21**(1), 46–51.

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., *et al.* (2003). Pgc-1α-responsive genes involved in

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, **34**(3), 267.

Moyerbrailean, G. A., Richards, A. L., Kurtz, D., Kalita, C. A., Davis, G. O., Harvey, C. T., Alazizi, A., Watza, D., Sorokin, Y., Hauff, N., *et al.* (2016). High-throughput allele-specific expression across 250 environmental conditions. *Genome research*, pages gr–209759.

Richiardi, J., Altmann, A., Milazzo, A.-C., Chang, C., Chakravarty, M. M., Banaschewski, T., Barker, G. J., Bokde, A. L., Bromberg, U., Büchel, C., *et al.* (2015). Correlated gene expression supports synchronous activity in brain networks. *Science*, **348**(6240), 1241–1244.

Schaub, F. X., Dhankani, V., Berger, A. C., Trivedi, M., Richardson, A. B., Shaw, R., Zhao, W., Zhang, X., Ventura, A., Liu, Y., *et al.* (2018). Pan-cancer alterations of the myc oncogene and its proximal network across the cancer genome atlas. *Cell systems*, **6**(3), 282–300.

Segrè, A. V., Groop, L., Mootha, V. K., Daly, M. J., Altshuler, D., Consortium, D., Investigators, M., *et al.* (2010). Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS genetics*, **6**(8), e1001058.

Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelsen, T. S., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., *et al.* (2014). Genome-scale crispr-cas9 knockout screening in human cells. *Science*, **343**(6166), 84–87.

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Allen, H. L., Lindgren, C. M., Luan, J., Mägi, R., *et al.* (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics*, **42**(11), 937.

Stephens, M. (2016). False discovery rates: a new deal. *Biostatistics*, **18**(2), 275–294.

Storey, J. D. *et al.* (2003). The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics*, **31**(6), 2013–2035.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**(43), 15545–15550.

Walter, N. D., Dolganov, G. M., Garcia, B. J., Worodria, W., Andama, A., Musisi, E., Ayakaka, I., Van, T. T., Voskuil, M. I., De Jong, B. C., *et al.* (2015). Transcriptional adaptation of drug-tolerant

bioRxiv preprint doi: https://doi.org/10.1101/662171; this version posted June 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

mycobacterium tuberculosis during treatment of human tuberculosis. *The Journal of infectious diseases*, **212**(6), 990–998.

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S., *et al.* (2013). Discovery and refinement of loci associated with lipid levels. *Nature genetics*, **45**(11), 1274.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M., *et al.* (2016). Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nature genetics*, **48**(5), 481.