Published in final edited form as:

Stat Med. 2018 December 10; 37(28): 4266–4278. doi:10.1002/sim.7920.

A Bayesian latent variable approach to aggregation of partial and top ranked lists in genomic studies

Xue Li¹, Pankaj Kumar Choudhary², Swati Biswas², and Xinlei Wang¹

¹Department of Statistical Science, Southern Methodist University, Dallas, TX, 75275, USA

²Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX, 75080, USA

Abstract

In genomic research, it is becoming increasingly popular to perform meta-analysis, the practice of combining results from multiple studies that target a common essential biological problem. Rank aggregation (RA), a robust meta-analytic approach, consolidates such studies at the rank level. There exists extensive research on this topic and various methods have been developed in the past. However, these methods have two major limitations when they are applied in the genomic context. First, they are mainly designed to work with full lists, whereas partial and/or top-ranked lists prevail in genomic studies. Second, the component studies are often clustered and the existing methods fail to utilize such information. To address the above concerns, a Bayesian latent variable approach, called BiG, is proposed to formally deal with partial and top-ranked lists and incorporate the effect of clustering. Various reasonable prior specifications for variance parameters in hierarchical models are carefully studied and compared. Simulation results demonstrate the superior performance of BiG compared to other popular RA methods under various practical settings. A non-small-cell lung cancer data example is analyzed for illustration.

Keywords

hierarchical Bayes; clustering effect; coverage rate; meta-analysis; partial/top ranked list; rank aggregation

1 | INTRODUCTION

In genomic studies, a common task is to identify genes that are associated with complex human diseases like cancer. Genomic experiments can be very costly and the obtained data are often noisy. Nevertheless, an enormous amount of genomic data have been generated by different research groups over the past decades. Thus, it is of great interest to integrate results from different studies on the same disease, i.e., performing meta-analysis, in order to improve the analysis power as well as promote the validity and reproducibility of the scientific findings. However, individual studies may not be directly combinable because of between-study heterogeneity due to causes such as different experimental setups, data

*Correspondence: Xinlei Wang, Department of Statistical Science, Southern Methodist University, Dallas, TX 75275, USA. swang@smu.edu.

quality, and type of analyses carried out, etc. A solution to this problem is to integrate the studies at the rank level wherein the results from each study are represented by a ranked gene list. This topic is known as rank aggregation (RA) and many research efforts have been made on this topic.

Among the earliest is Borda's collection¹, which uses summary statistics such as the arithmetic mean (MEAN), geometric mean (GEO) and median (MED) of the individual ranks to obtain the aggregated rank for each item of interest. Robust Rank Aggregation (RRA²) and Stuart's method (Stuart³) utilize the distributions of order statistics and assign a p-value to each item to determine the aggregated rank. The Markov chain (MC) methods^{4,5} construct transition matrices such that items with higher stationary probabilities are ranked higher in the aggregated list with MC1 – MC3⁶ providing three ways of constructing such transition matrices. Lin and Ding⁷ propose Cross Entropy Monte Carlo (CEMC) methods, which involve a stochastic search for the solution to minimizing the distances between the input lists and the aggregated list. Bayesian Aggregation of Rank Data (BARD⁸) assigns aggregated ranks based on the posterior probability that a certain item is relevant. Bayesian Iterative Robust Rank Aggregation (BIRRA⁹) iteratively updates the ranks based on Bayes Factors. For a detailed overview of existing RA methods, see Lin and Ding⁷, Li, Wang and Xiao¹⁰, and the references therein.

One major limitation of existing RA methods is that they mostly focus on aggregating full lists, wherein every individual list provides exact ranks for all items of interest, which is rare in genomic applications¹⁰. More realistically, an individual list often does not include all genes of interest (due to reasons like missing lab measurements for some genes or removals in some preprocessing steps), resulting in a partial list; and/or it may only provide the exact ranks for genes that are ranked at the top, resulting in a top-ranked list. For a top-ranked list, if the set of genes originally studied in the ranking process is available, then one may assume that any genes present in this set but not present in the top-ranked list are ties ranked at the bottom. Several methods claim that they can deal with partial/top-ranked lists via *ad-hoc* adjustments such as replacing the missing ranks with the maximum observed rank plus one, however, they do not differentiate between two different sources of missing ranks – non-inclusion of genes in a study or unreported bottom-ranked genes. Li, Wang, and Xiao¹⁰ conduct a comprehensive comparative study on aggregating partial and top-ranked lists, and show that the performance of RA methods largely depends on the amount of information available from input lists as well as the number of genes that can be considered for follow-up investigation based on resources available. Also, they suggest that information about bottom-ranked genes, when available can be helpful, and hence should be used; moreover, how such information is utilized can make a significant difference. These findings call for a formal and rigorous treatment of partial and top ranked lists.

Another issue that has been overlooked by existing methods in aggregating ranked lists arising from genomic applications is platform bias. Here “platform” can mean any cluster formed by meaningfully grouping studies based on some similarity. Studies could be grouped based on types of “omics” data collected (e.g., gene/mRNA expression, DNA copy number, methylation and microRNA expression profiles), research labs producing raw data, or technologies used like microarray and next-generation sequencing, or even types of

analyses performed, etc. In this paper, a Bayesian latent variable approach, referred to as “Bayesian Aggregation in Genomic applications” (BiG), is proposed to formally deal with partial and top-ranked lists as well as accounting for potential platform bias.

The rest of the paper is organized as follows. In Section 2, we describe the proposed latent variable model and discuss several prior choices for the variance parameters involved. In Section 3, we describe a formal simulation study conducted to evaluate the performance of the proposed method in comparison with other popular methods as well as its robustness to deviations from the model assumptions. In Section 4, the proposed method is illustrated on a non-small cell lung cancer data example. Finally, in Section 5, we provide a summary of the paper and discuss practical issues including computation time and selection of studies to include in the rank aggregation problem. For a list of important notation used in this paper, see Section S1 in Supplementary Material (SM).

2 | BAYESIAN HIERARCHICAL METHODS

2.1 | A latent variable model

A ranked list is often produced based on some underlying variable. For example, the ranking of runners in a race is determined by the time taken to finish the race. In genomic applications, the underlying variable can be some quantity that reflects the strength of association between individual genes and a certain disease, such as (transformed) statistics from commonly used hypothesis tests, however, those values are typically not available when performing RA. This motivates us to consider the following latent variable model.

Suppose there are J studies from a total of P different platforms; and each study generates a ranked list of genes. Let g index genes, j index studies, and p index platforms. For each study j , let $\mathcal{G}_j$ be the index set of genes present in its list. We assume that only the top n_j^T genes in $\mathcal{G}_j$ are ranked explicitly (i.e., their exact ranks are known) and their index set is denoted by $\mathcal{T}_j$. We further assume that the remaining genes in $\mathcal{G}_j$ are known to be ranked lower than any gene in $\mathcal{T}_j$, but their exact ranks are unknown. For the purpose of clarity, higher ranked genes (i.e., genes with lower numeric ranks) are deemed to be more important. Therefore, the genes whose ranks are unknown are referred to as bottom ties and their index set is denoted by $\mathcal{B}_j$, satisfying $\mathcal{G}_j \equiv \mathcal{T}_j \cup \mathcal{B}_j$. Note that $\mathcal{B}_j$ can be the empty set $\emptyset$. Let

$\mathcal{G} = \cup_{j=1}^J \mathcal{G}_j$ be the index set of all genes of interest with $G \equiv |\mathcal{G}|$ genes in total. For genes in $\mathcal{G}_{-j}$ (i.e., those in $\mathcal{G}$ but not in $\mathcal{G}_j$), there exist two cases that can happen in study j : (i) they were not initially investigated in study j and so their ranking information is not available; we refer to these genes as “NA” items of study j ; or (ii) they were initially investigated but not reported probably due to the fact that they are ranked no higher than the lowest ranked gene in $\mathcal{G}_j$. In many practical situations, $\mathcal{G}_j$ is known to contain all genes investigated in study j , and so the genes in $\mathcal{G}_{-j}$ can only be NA’s. Thus, we assume throughout the paper that in every study j , the genes in $\mathcal{G}_{-j}$ are NA’s. However, there exist lists (e.g., top-ranked only lists) where the genes outside could belong to either case, but

which of the two cases is the truth is unknown. For such lists, we should treat the genes in $\mathcal{G}_{-j}$ as either bottom ties or NA's based on our best judgment, update $\mathcal{B}_j$ and $\mathcal{G}_j$ accordingly, and then proceed as before.

Let x_j denote the platform of study j , and let $\mathcal{S}_p = \bigcup_{j: x_j = p} \mathcal{G}_j \subset \{1, \dots, G\}$ be the union of genes present in the lists of studies that belong to platform p , where $p \in \{1, \dots, P\}$. For genes in $\mathcal{G}_j$, let r_{gj} denote the rank of gene g in study j . We assume that r_{gj} 's are determined by an underlying, continuous-valued variable w_{gp} i.e. $r_{gj} > r_{g'j}$ implies $w_{gp} < w_{g'p}$, where w_{gp} reflects the local importance of gene g in study j . Therefore, the gene with the largest local importance has rank 1 in study j .

For genes in $\mathcal{T}_j$, $r_{gj} \in \{1, \dots, n_j^T\}$; and we have $w_{(1,j)} > w_{(2,j)} > \dots > w_{(n_j^T,j)}$, where (g, j) indexes the gene that is the top g th ranked in study j (i.e., $r_{(g,j)} = g$ for $g \leq n_j^T$). For all genes in $\mathcal{B}_j$, $w_{gj} < w_{(n_j^T,j)}$ and r_{gj} can be set to any specific integer larger than n_j^T in our Bayesian approach; for simplicity, we set $r_{gj} \equiv n_j^T + 1$. In study j , let $\mathbf{r}_j$ and $\mathbf{w}_j$ denote the collection of r_{gj} 's and w_{gj} 's for $g \in \mathcal{G}_j$. Thus,

$$p(\mathbf{r}_j | \mathbf{w}_j) = \mathbb{I}\left(w_{(1,j)} > w_{(2,j)} > \dots > w_{(n_j^T,j)}\right) \times \prod_{g \in \mathcal{B}_j} \mathbb{I}\left(w_{gj} < w_{(n_j^T,j)}\right),$$

where $\mathbb{I}(\cdot)$ is the indicator function.

We set up the following linear model for the latent variable w_{gp} given $g \in \mathcal{G}_j$ for $j = 1, \dots, J$:

$$w_{gj} = \mu_g + \kappa_{g,p} + \epsilon_{gj}, \quad (1)$$

where μ_g measures the global importance of gene g , which determines its true rank. The above model assumes that in any study j that belongs to platform p (i.e., $x_j = p$, the importance μ_g , for $g \in \mathcal{G}_j$, is observed with independent platform-specific bias κ_{gp} and study-specific bias ϵ_{gj} , contributing to gene g 's ranking error in study j . We further assume the following independent normal distributions on the three linear components:

$$\mu_g \sim N(0, 1), \kappa_{gp} \sim N(0, \sigma_{\kappa,p}^2), \epsilon_{gj} \sim N(0, \sigma_{\epsilon,j}^2).$$

For the purpose of identifiability, we fix the location and scale parameters of the distribution of μ_g at 0 and 1, respectively. Rather than assuming a common variance for all the platforms or studies, which seems to be restrictive in practice, we allow platform-specific variances $\sigma_{\kappa,p}^2$'s for κ_{gp} 's and study-specific variances $\sigma_{\epsilon,j}^2$'s for ϵ_{gj} 's.

Let $\mathbf{w}$ denote the collection of w_{gj} 's for $g \in \mathcal{G}_j$ and $j = 1, \dots, J$, $\boldsymbol{\mu} = (\mu_g)_{g=1}^G$,

$\boldsymbol{\kappa} = ((\kappa_{gp})_{g=1}^G)_{p=1}^P$, $\sigma_{\epsilon}^2 = (\sigma_{\epsilon \cdot j}^2)_{j=1}^J$, $\sigma_{\kappa}^2 = (\sigma_{\kappa \cdot p}^2)_{p=1}^P$ and $\mathbf{x} = (x_j)_{j=1}^J$. Thus,

$$p(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\kappa} | \mathbf{x}, \sigma_{\epsilon}^2, \sigma_{\kappa}^2) = \prod_{j=1}^J \prod_{g \in \mathcal{G}_j} N(w_{gj} | \mu_g + \kappa_{g, x_j}, \sigma_{\epsilon \cdot j}^2) \times \prod_{g=1}^G N(\mu_g | 0, 1) \times \prod_{p=1}^P \prod_{g \in \mathcal{G}_p} N(\kappa_{gp} | 0, \sigma_{\kappa \cdot p}^2),$$

where $N(x | \mu, \sigma^2)$ denotes the probability density function (pdf) of a normal distribution with mean μ and variance σ^2 , evaluated at x .

2.2 | Prior elicitation for variance parameters

In the literature, how to specify prior distributions for variance parameters in hierarchical models has attracted a lot of attention, and various sensible choices have been suggested, especially when no meaningful information is available for such parameters (e.g., Chapter 5.7 of Spiegelhalter, Abrams and Myles¹¹, and Gelman¹²). Here, we consider and compare several prior choices for $\sigma_{\epsilon \cdot j}^2$'s and $\sigma_{\kappa \cdot p}^2$'s, including two classical approaches using diffuse priors, a fully Bayesian approach suggested by a previous work on a ¹². Note that in the context of RA in genomic applications, no comparison has been made between these priors previously.

Approaches using diffuse priors—First, a common diffuse inverse gamma (IG) prior is considered for the studies and the platforms, respectively, $\sigma_{\epsilon \cdot j}^2 \sim IG(\delta_{\epsilon}, \delta_{\epsilon})$ and

$\sigma_{\kappa \cdot p}^2 \sim IG(\delta_{\kappa}, \delta_{\kappa})$, with typically small values of δ_{ϵ} and δ_{κ} such as 1, 0.1, 0.01, etc. Here, $IG(\alpha, \beta)$ represents an IG distribution with shape parameter α and rate parameter β . The IG priors are conjugate for normal variance parameters, which simplifies the derivation of conditional posterior distributions.

Uniform distributions, another common “noninformative” choice for standard deviation parameters, are also considered, i.e., $\sigma_{\epsilon \cdot j} \sim \text{Uni}(\ell_{\epsilon}, U_{\epsilon})$, $\sigma_{\kappa \cdot p} \sim \text{Uni}(\ell_{\kappa}, U_{\kappa})$. The boundaries of the uniform priors are chosen such that the correlation between w_{gj} 's and μ_g 's is between a reasonably wide range, e.g., [0.05, 0.95].

A fully Bayesian (FB) approach—We also adopt a fully Bayesian approach developed in Johnson et al.¹³ for rank data from primate intelligence experiments. First, a Gamma distribution is used for the reciprocal of each variance parameter, i.e., $1/\sigma_{\epsilon \cdot j}^2 \sim \text{Ga}(\nu_{\epsilon}, \nu_{\epsilon}/\mu_{\epsilon})$ and $1/\sigma_{\kappa \cdot p}^2 \sim \text{Ga}(\nu_{\kappa}, \nu_{\kappa}/\mu_{\kappa})$, where ν_{ϵ} and ν_{κ} are shape parameters and $\nu_{\epsilon}/\mu_{\epsilon}$ and $\nu_{\kappa}/\mu_{\kappa}$ are rate parameters so that μ_{ϵ} and μ_{κ} are means of the Gamma distributions. Next, the following independent hyperpriors are considered for μ_{ϵ} , μ_{κ} , ν_{ϵ} and ν_{κ} : $\mu_{\epsilon}, \mu_{\kappa} \sim \text{Exp}(\delta)$ and $\nu_{\epsilon}, \nu_{\kappa} \sim IG(\alpha, \beta)$ where $\text{Exp}(\delta)$ represents an exponential distribution with rate parameter δ . Following Johnson et al.¹³, we choose δ , α and β so that the Gamma prior for each precision parameter assigns a large percentage of the weight to a reasonably wide interval (e.g. [0.25, 4]).

A data augmentation (DA) approach—Motivated by Gelman¹², we consider a half-t prior for $\sigma_{\kappa.p}$ which can be achieved through the following augmented data model for the platform bias κ_{gp} ,

$$\begin{aligned}\kappa_{gp} &= \alpha_p \cdot \xi_{gp}, \\ \alpha_p &\sim N(0, 1), \xi_{gp} \sim N(0, \sigma_{\xi.p}^2), \\ \sigma_{\xi.p}^2 &\sim IG(\delta_\xi, \delta_\xi),\end{aligned}$$

giving rise to $\sigma_{\kappa.p} = |\alpha_p| \cdot \sigma_{\xi.p}$. Clearly, α_p is a redundant multiplicative parameter, whose standard deviation is not separable from $\sigma_{\xi.p}$, the standard deviation of ξ_{gp} , and so the standard deviation of α_p is fixed at 1 to achieve identifiability. Under the above DA model, the prior distribution of $\sigma_{\kappa.p}$ is equivalent to the distribution of the absolute value of a standard normal random variable divided by the square root of a gamma random variable, which leads to a half-t distribution. For $\sigma_{\epsilon.j}$ we simply use the uniform prior, i.e. $\sigma_{\epsilon.j} \sim \text{Unif}(L_\epsilon, U_\epsilon)$.

The major advantage of introducing a redundant parameter α_p is to obtain conditional conjugacy for all the parameters involved, which simplifies the construction of an MCMC algorithm for implementing the half-t prior setup.

2.3 | Posterior computation and Bayesian inference

We begin with the posterior computation for the approaches using diffuse priors and then briefly discuss modifications for the FB and DA approaches. All technical details for each approach can be found in Sections S2.1–2.3 of SM.

Let $\theta \dots$ denote θ given observed gene ranks from the $\mathbf{J}$ studies and all the other parameters and latent variables. Based on diffuse priors, the full conditional posterior distributions for μ_g and κ_{gp} are both normal, and for $\sigma_{\epsilon.j}^2$ and $\sigma_{\kappa.p}^2$ are (truncated) inverse gamma. For $g = 1, \dots, n_j^T, j = 1, \dots, \mathbf{J}$, we can show that

$$w_{(g,j)} | \dots \sim N(\mu_{(g,j)} + \kappa_{(g,j)}, \sigma_{\epsilon.j}^2) \mathbb{I}(w_{(g+1,j)} < w_{(g,j)} < w_{(g-1,j)}).$$

where (g, j) indexes the gene with the observed rank g in study j for $g \leq n_j^T$, as mentioned before. For $g \in \mathcal{B}_j$,

$$w_{gj} | \dots \sim N(\mu_g + \kappa_{(g,j)}, \sigma_{\epsilon.j}^2) \mathbb{I}\left(w_{gj} < w_{(n_j^T, j)}\right).$$

A Markov chain Monte Carlo (MCMC) algorithm¹⁴ using a Gibbs sampler is implemented for posterior sampling, with five steps in each iteration. In steps 1–4, μ_g 's, κ_{gp} 's, $\sigma_{\epsilon.j}^2$'s and $\sigma_{\kappa.p}^2$'s are directly generated from known distributions as described above. In step 5, we

generate w_{gj} 's, where the challenge is that the generated values need to be consistent with the observed orderings in the J studies. For the top ranked genes in $\mathcal{T}_j$, the ordering of updating the $w_{(g,j)}$'s, $g = 1, \dots, n_j^T$, is based on a random permutation of $\{1, 2, \dots, n_j^T\}$. We propose the following sequential-updating process for w_{gj} 's:

- 5(a)** A permutation of $\{1, 2, \dots, n_j^T\}$, say $\{p_1, p_2, \dots, p_{n_j^T}\}$, is generated.
- 5(b)** $w_{(p_1, j)}$ is updated first, then $w_{(p_2, j)}$ is updated and so on. For each $p_k \in \{1, \dots, n_j^T\}$, we generate $w_{(p_k, j)}$ from truncated $N(\mu_{(p_k, j)} + \kappa_{(g, j)}, \sigma_{\epsilon, j}^2)$ with lower bound $w_{(p_k + 1, j)}$ and upper bound $w_{(p_k + 1, j)}$, where we set $w_{(0, j)}$ and $w_{(n_j^T + 1, j)} \equiv \max_{g \in \mathcal{B}_j} \{w_{gj}\}$.
- 5(c)** After $w_{(g, j)}$'s are generated for genes in $\mathcal{T}_j$, w_{gj} 's for genes in $\mathcal{B}_j$ are generated from truncated $N(\mu_g + \kappa_{(g, j)}, \sigma_{\epsilon, j}^2)$ with lower bound $-\infty$ and upper bound $w_{(n_j^T, j)}$.

For the FB approach, the full conditional distributions for μ_g , κ_{gp} and w_{gj} are the same as above; and the full conditionals for the variance parameters are also inverse gamma (with different parameters). However, the full conditionals for the hyperparameters, μ_ϵ , μ_κ , ν_ϵ , and ν_κ , are not known distributions. So Metropolis-Hastings (M-H) sampling procedures similar to those in Johnson et al.¹³ are employed for these hyperparameters. This yields a hybrid Gibbs sampler with built-in M-H steps.

For the DA approach, although a different parameterization is used, the full conditionals are still known distributions and an MCMC algorithm using a Gibbs sampler is implemented in a manner similar to that for the approaches using diffuse priors.

To start a chain using any of our MCMC algorithms, we need to specify initial values for the parameters. For a detailed discussion, see Section S2.4 of SM. For simplicity, in our subsequent numerical evaluation, we set initial values of any μ_g and κ_{gp} to zero, initial value of any $\sigma_{\epsilon, j}^2$ to 1, and that of any $\sigma_{\xi, p}^2$ to 0.5. Standard diagnostic techniques (Gelman et al.¹⁵) can be used to detect convergence of the MCMC algorithm.

Since μ_g 's measure the global importance of individual genes, the aggregated ranking is obtained by sorting their Bayesian estimates, which are simply the average of posterior draws of μ'_g s (after a burn-in period). Meanwhile, an estimate for the variance of an individual study/platform can be obtained from the median of posterior draws of $\sigma_{\epsilon, j}^2 / \sigma_{\kappa, p}^2$, which indicates the relative quality of the study/platform.

2.4 | Comparison of different prior choices

We simulate data from the latent variable model described in Section 2.1 to evaluate the performance of the proposed method under the different prior choices, where we set the number of genes $G = 500$, the number of studies $J = 6$ and the number of platforms $P = 2$ (each platform has three studies). 100 replicates are generated and used for comparison. For the detailed setup, see Section S3.1 in SM. Posterior samples are drawn for each of the four prior choices, including two diffuse priors (IG and Uniform), FB, and DA (half-t), as discussed in Section 2.2. For the IG prior, we set $\delta_\epsilon = \delta_\kappa = 1$; for the FB approach, we set $\delta = 0.1$, $\alpha = 1.17$ and $\beta = 0.65$ so that the precision values largely concentrate between the interval $[0.25, 4]$ *a priori*; and for the DA approach, we set $\delta_\xi = 1$. We also experiment with other hyperparameter values for each prior choice, and we find that results are either worse than or comparable to those using the values selected.

For the purpose of comparison several aspects are considered. First, the convergence of each algorithm corresponding to one of the four prior choices is assessed through trace plots. For μ_g , the convergence seems satisfactory regardless of the prior used. However, this is not true for the variances $\sigma_{\epsilon,j}^2$ and $\sigma_{\kappa,p}^2$. As shown in Figure S1 of SM, the algorithm using the IG prior demonstrates the best overall behavior in convergence. The FB approach seems to be the most concerning with some parameters clearly failing to converge. Along with the fact that it is much more computationally demanding than the rest (due to the use of multiple M-H steps within each Gibbs iteration), we exclude FB from the subsequent comparisons.

Next, we assess how well the parameters can be estimated based on different prior choices. After a burn-in period of 10,000 iterations, posterior means are computed to estimate μ_g 's and posterior medians to estimate $\sigma_{\epsilon,j}^2$'s based on another 10,000 iterations. Average mean squared errors are calculated across all genes for μ_g 's and across all studies for $\sigma_{\epsilon,j}^2$'s, and are provided in Table 1 (a). The algorithm with the IG prior seems to outperform the others, especially for estimating $\sigma_{\epsilon,j}^2$'s.

Finally, we examine how well the corresponding aggregated ranking performs under each prior choice. In the literature of rank aggregation, metrics commonly used for performance evaluation include Spearman and Kendall distances⁷, Area Under Receiver Operating Characteristic Curve (AUC^{16,2,9}), and coverage rate⁸. The distance measures and AUC are more appropriate for a holistic view of performance across the entire aggregate list; and applying them to only a part of the list can lead to misleading results. In practice, people are often concerned with identifying as many relevant genes as possible with the top ranked genes. Therefore, as in Deng et al.⁸, we report coverage rates (defined by percentage of relevant genes covered by the subset of top-ranked genes in the aggregated ranked list), with different cutoffs, in Table 1 (b). For a detailed comparison and discussion of different performance evaluation measures, see Li et al.¹⁰. Here, the algorithm with the IG prior appears to be slightly better than the other two again. Overall, the IG prior is deemed to be preferable among the four prior choices according to the performance in convergence, parameter estimation and the ability to identify relevant genes as well as its relative

simplicity. Thus, it is used in our performance evaluation when comparing our Bayesian method with competing methods.

3 | SIMULATION

Several simulation studies focusing on partial and top-ranked lists are conducted to evaluate the performance of the proposed method BiG, under various situations. We compare BiG with other popular methods, including MEAN, GEO and MED in Borda's collection, RRA, Stuart, Markov Chain (MC) methods including MC1–MC3, CEMC methods including CEMC.k that uses the weighed Kendall's tau distance as the optimization criterion and CEMC.s that uses the weighted Spearman's footrule distance, BARD, and BIRRA.

As mentioned earlier, most of the existing methods are intended to be used with full lists. Although there are suggestions in literature about how they can be altered to accommodate non-full lists, there are some drawbacks to these accommodations. First, to the best of our knowledge, the possibility of bottom ties has never been acknowledged except in Li, Wang and Xiao¹⁰. Second, with the exception of BARD, which formally deals with items with missing ranks in its MCMC process, most methods that claim to have the ability to handle partial/top-ranked list simply assign them with the maximum observed rank plus 1 (i.e., $n_j^T + 1$ in our earlier notation), without examining the cause of the missingness. On the other hand, recall that the proposed method differentiates NA's from bottom ties. As mentioned in Section 2.1, we choose to use $n_j^T + 1$ as the rank for the bottom ties, too. However, any number between n_j^T and $|\mathcal{G}_j|$ would lead to the same aggregated ranks, whereas changing the value of $n_j^T + 1$ may alter the aggregated ranks for the other methods.

In our simulation, the R package “RobustRankAggreg” is used to implement RRA and Stuart; and the R package “TopKLists” is used to implement MC1–MC3, CEMC.k and CEMC.s. BARD is implemented in a C++ program kindly provided by the authors and BIRRA is implemented in an R package associated with the corresponding paper. For BiG, 20,000 iterations are used in each MCMC run with a 10,000 burnin period. MEAN, GEO and MED are calculated in a straightforward way with “NA”s removed, if present.

3.1 | Performance evaluation

Simulated ranked lists are generated from the latent variable model described in Section 2.1, with a total of $G = 500$ genes, $J = 6$ studies and $P = 2$ platforms (three studies for each platform). Let ρ_j denote the correlation between the w_{gj} 's and μ_g 's, which measures the strength of the linear relationship between a gene's local importance in study j and its global importance. Thus, the quality of each study j is controlled by ρ_j , where $\rho_j^2 = 1/\text{Var}(w_{gj})$ and $\text{Var}(w_{gj}) = 1 + \sigma_{\kappa.p}^2 + \sigma_{\epsilon.j}^2$. In practice, the quality of studies can vary greatly, therefore we randomly selected ρ_j from $\text{Unif}(0.3, 0.9)$ to allow for both poor and excellent quality studies. For different ρ_j values, we fix $\sigma_{\kappa.p}^2$'s by setting $\sigma_{\kappa.1}^2$ and $\sigma_{\kappa.2}^2$ to be 60% and 80% of the upper bound of all possible values for $\sigma_{\kappa.p}^2$'s (i.e., $\sigma_{\kappa.1}^2 = 0.14$ and $\sigma_{\kappa.2}^2 = 0.18$), respectively, where

the upper bound is $\text{Var}(w_{gj}) - 1$ evaluated at $\rho_j = 0.9$. Then for each ρ_j simulated, we solve for $\sigma_{\epsilon, j}^2$.

Additional design parameters are introduced to produce partial and top-ranked lists: n^T ($n^T = 10, 20, 50, 100$) controls the number of top ranked genes; and λ ($\lambda = 0.6, 0.8, 1$), the gene inclusion rate, controls the chance that a gene is not an “NA” item in each input list, where genes that are not ranked in the top n^T and not “NA” items are treated as bottom ties. Here, the same values of n^T and λ are used across studies for every combination.. We also examine a mixed setting, where for each of the six studies n^T is uniformly drawn from $\{10, 20, 50, 100\}$ and λ is drawn from $\text{Unif}(0.6, 0.8)$. We generate 500 replicates for every setting considered.

In this simulation, the underlying truth is continuous, and therefore “relevant” genes are not explicitly defined. We naturally consider the top 10, 20 and 50 genes to be relevant and then use the corresponding cutoffs to calculate coverage rates. Coverage rates of the different methods based on top 20 genes are displayed in Table 2, where the winner is in boldface under each setting. For all the (λ, n^T) combinations, the proposed method BiG uniformly outperforms all the other methods. Under the mixed setting, the performance of most methods is close to the lower end of what we see in the fixed settings; and methods such as MEAN and CEMC.s suffer substantial losses. By contrast, BiG is not as sensitive, and still outperforms the others. Coverage rates using cutoffs of top10 and 50 genes are provided in Tables S2 and S3 of Section S4 of SM, leading to the same conclusions.

Next, we evaluate the effect of each design parameter $(n^T, \lambda, \rho, \sigma_{\kappa,1}^2/\sigma_{\kappa,2}^2)$ on the performance of the different methods. The results with varying n^T while the other parameters are held constant (i.e., $\lambda = 0.8, \rho_j \sim \text{Unif}(0.3, 0.9), \sigma_{\kappa,1}^2 = 0.14$ and $\sigma_{\kappa,2}^2 = 0.18$ are presented in Figure 1. When n^T increases, the exact ranks of more genes become known and as we would expect, the performance of BiG improves, which happens with the other methods also, but with a few exceptions (e.g., MEAN, CEMC.k, CEMC.s). However, the rate at which the coverage rate increases slows down as n^T continues to increase, especially when n^T moves further beyond the cutoff used to determine the coverage rate. This can be seen as the coverage rate levels out much faster for cutoff 10 than for cutoff 50.

The results with varying λ while the other parameters are held constant (i.e., $n^T = 50, \rho_j \sim \text{Unif}(0.3, 0.9), \sigma_{\kappa,1}^2 = 0.14$ and $\sigma_{\kappa,2}^2 = 0.18$) are presented in Figure 2. When λ increases, on average, more and more genes of interest are included in each of the input lists, which should positively influence the performance. As seen from Figure 2, this is indeed the case for nearly all methods, and the increasing trend is approximately linear for most of these methods.

The results with varying n^T for the other λ values and with varying λ for the other n^T values show patterns similar to what has been observed from Figures 1 and 2. Overall, BiG has the best performance for partial and top-ranked lists. After all, it is designed with these issues in mind.

In all of the previous simulation settings, ρ_j 's are randomly generated from $Unif(0.3, 0.9)$. To examine the effect of ρ , we fixed it across all studies (i.e., $\rho_j \equiv \rho$) and vary ρ among the set $\{0.3, 0.5, 0.7, 0.9\}$ while holding the other parameters constant (i.e., $n^T = 50$, $\lambda = 0.8$, $\sigma_{\kappa,1}^2 = 0.14$ and $\sigma_{\kappa,2}^2 = 0.18$). As ρ increases, the ranking quality of each study improves and the performance of rank aggregation methods should improve in turn, which can be seen in Figure S2 for all methods except for BARD. Here, the advantage of BiG over other methods is not as pronounced as when the quality of the studies varies; however, it still has the best or close to the best performance.

Last, to assess the effect of $\sigma_{\kappa,p}^2$'s that control platform bias, we fixed $\rho = 0.7$ and then set $\sigma_{\kappa,1}^2/\sigma_{\kappa,2}^2$ to three levels, 10%/20%, 30%/40%, 10%/40% of $Var(w_{gj}) - 1$, with the other parameters held constant ($n^T = 50$, $\lambda = 0.8$). These levels are chosen to represent the cases of L/L, H/H and L/H variability (L for low and H for high). The coverage rates with cutoff 20 are reported in Table 3, and those with cutoffs 10 and 50 in Tables S4 and S5 of Section S4 in SM. The performance of all methods except for BIRRA deteriorates as the variances move from L/L to L/H to H/H. Although MC2 or MC3 perform well sometimes, BiG is consistently the best or close to the best in all settings.. This is not surprising as BiG is the only method that accounts for potential platform bias explicitly. Note that in the previous settings where ρ_j is from a uniform distribution, we set $\sigma_{\kappa,1}^2 = 0.14$ and $\sigma_{\kappa,2}^2 = 0.18$, which typically account for only a small portion of the total variance $1 + \sigma_{\kappa,p}^2 + \sigma_{\epsilon,j}^2$. Doing so avoids being overly favorable to BiG.

3.2 | Robustness checking

We conduct additional simulations with $\lambda = 0.8$ and $n^T = 50$ to assess the performance of BiG when its underlying assumptions are violated. First, we examine three cases when μ_g 's are not generated from $N(0, 1)$, but instead from (i) Student's t distribution with three degrees of freedom (t_3), which represents a thick-tailed distribution, (ii) $Exp(1)$, which represents a right-skewed distribution, and (iii) folded $N(0, 1)$, or equivalently, the absolute value of a standard normal variable, mimicking realistic situations when the ranking in study j is determined by p-values from a two-sided test. Note that t_3 and folded $N(0, 1)$ are scaled so that they have variance of 1; and $\rho_j \sim Unif(0.3, 0.9)$, $\sigma_{\kappa,1}^2 = 0.14$, $\sigma_{\kappa,2}^2 = 0.18$. Second, we consider the situation where there is no platform bias. Third, a dichotomous model (DM) used in^{9,2}, where the underlying truth is whether a gene is relevant or not, is employed to examine the situation when ranked lists are not generated using the linear latent variable structure. Here, signal genes are generated from $N(1, 1)$ and non-signal genes are generated from $N(0, 1)$ with 5% signal rate.

We report the coverage rates with cutoff 20 in Table 4, those with cutoffs 10 and 50 are in Tables S6 and S7 of Section S4 in SM. The results suggest that in spite of some deviations from the model assumptions, the proposed method maintains its strong performance, compared to the other methods. Note that even when the ranked lists are not generated from a linear latent variable model, BiG is among the top performing group and performs reasonably well.

3.3 | Comparison of two treatments for top-k only lists

In practice, there exist top-k only lists for which we do not have the original gene inclusion information, as mentioned in Section 2.1. That is, for each of these lists, the status of the genes that are not top ranked but are included from the other studies could either be i) originally studied but not ranked in the top or ii) not originally studied. Consequently, there are two ways of treating such genes, either as bottom ties or “NA”s. Simulations are performed to evaluate these two treatments for top-k only lists in different (λ, n^T) settings (the other parameters are the same as in Table 2); and results are reported in Tables S8 – S10 of Section S4 in SM. We find that in general, the bottom tie treatment provides better coverage, under which the performance of BiG is quite competitive. We further note that in all the settings considered, there are more bottom ties than “NA”s, as is typical in whole-genome studies. This may explain why the bottom tie treatment performs better. In case that there are more “NA”s than bottom ties, the NA treatment can be better, under which BARD appears to perform well.

4 | AN EXAMPLE IN NON-SMALL CELL LUNG CANCER

We apply the proposed method BiG as well as its competitors to aggregate ranked lists from seven non-small cell lung cancer (NSCLC) studies based on either microarray or RNA-seq data as described in Table 5. Notice that there are three top-ranked only lists for which we can either take the NA treatment or the bottom tie treatment as discussed above.. Because all three of the original studies from which these lists were produced performed genome-wide experiments, it appears to be reasonable to treat those genes as bottom ties. Nonetheless, the alternative where such genes are treated as “NA”s is also examined for comparison.

The union $\mathcal{G}$ of genes appearing in all input lists contains about 22000 distinct genes, however, many of them are not top ranked in any of the component studies. Intuitively, these genes would not be ranked high in the aggregated list, therefore, are removed to save computing time (about 6000 genes left in $\mathcal{G}$). Also, the CEMC methods are not applied here because it would take up to weeks to run even after the downsizing adjustment and they do not show competitive performance in simulations.

The true ranking of genes is known in our simulation, but not in any real application. To be able to evaluate the relative performance of the different RA methods in this example, we have to specify signal genes (i.e., genes associated with the disease) for computing the coverage rate, of which the complete set is not yet known. Here, we use a collection of 148 genes in Li, Wang and Xiao¹⁰ as the surrogate of the “truth”, which are believed to be highly related to NSCLC according to the lung cancer literature. It contains the cancer gene lists for NSCLC from four sources: the Catalogue Of Somatic Mutations In Cancer (COSMIC), MalaCards, The Cancer Genome Atlas (TCGA), and a similar list in Chen et al.²². For the MCMC algorithm used to implement BiG, 5000 iterations are used and the first half are considered as burnin. Coverage rates based on top 100, 200, 300, 400 and 500 genes are summarized in Table 6. Comparing the two treatments of the genes that do not appear in the top-ranked only lists: “NA”s in Table 6 (a) and bottom ties in Table 6 (b), it is clear that most methods have higher coverage rates using the bottom-tie treatment as we suspect. Now,

focusing on the results from Table 6 (b), it can be seen that BiG appears to have the best overall performance.

We note that the coverage rates appear to be low in Table 6. This is because there are ~22,000 human genes in total, but the number of “signal” genes is very small (~150), due to the fact that our current knowledge about NSCLC is still limited. So if we randomly select 500 genes from the genome, the coverage rate would be about 2.3%. By contrast, all the RA methods in Table 6 (b) achieve much higher coverage with their top 500 genes, in particular, BiG achieves a rate of 14.3%, which is over 6 times higher than 2.3%.

5 | DISCUSSION

In this paper, we propose a Bayesian latent variable approach (BiG) to RA. Unlike existing methods, BiG is model-based and specifically designed to rigorously deal with partial and top-ranked lists as well as potential “platform” bias, which are common in genomic applications. In particular, BiG is the only method that formally distinguishes bottom ties from NA’s. Several sensible prior choices (diffuse inverse gamma/uniform priors, fully Bayes, half-t prior via a DA model) for the variance/standard deviation parameters in hierarchical models, as suggested in the literature, are carefully studied under the context of RA in genomic studies. MCMC algorithms are developed for each of these specifications, and then compared via simulation. We find that the algorithm based on the diffuse IG prior is the most preferred given its relatively better convergence, better estimation of the parameters of interest, and higher coverage in RA as well as conceptual simplicity and ease of implementation. Further, a thorough simulation study is conducted to evaluate the performance of BiG (with the IG prior), and compare it to other popular methods under various realistic settings. BiG yields strong performance in almost all settings tested even when there are deviations from its model assumptions. Finally, the performance of the different RA methods is assessed through an NSCLC data example, which confirms the usefulness of BiG.

With regards to computation time, BiG is Bayesian in nature, and it relies on MCMC algorithms, which tend to be computationally demanding. Under the default setting in our simulation (i.e., $\lambda = 0.8$, $n^T = 50$, $\sigma_{\kappa,1}^2 = 0.14$, $\sigma_{\kappa,2}^2 = 0.18$, $\rho_j \text{Unif}(0.3, 0.9)$, and $\mu_g \sim \mathcal{N}(0, 1)$)

BiG (with the IG prior) has a run time of about 2.5 hours with 20,000 iterations using Scientific Linux 6 (64 bit) operating system with Intel® Xeon® CPU X5560 @ 2.80GHz. Most of the other methods discussed in this paper (MEAN, GEO, MED, RRA, Stuart and BIRRA) finished in seconds. MC methods finished within a minute and BARD finished within a few minutes. The run time of CEMC’s, the most computationally expensive methods included in our comparison, is typically between 3 to 5 hours. A potential modification to reduce the computation time of BiG is to use posterior modes instead of posterior means as the Bayesian estimates, which can be seen as an optimization problem and algorithms like Expectation-Maximization (EM²³), or conditional maximization¹⁵ could be employed.

Under a Bayesian framework, BiG can be modified to accommodate more complex situations and incorporate existing biological knowledge. First, it can handle different levels

of “bottom ties”. For example, a study may report the set of top-ranked genes, the set of differentially expressed (DE) genes, and the set of all genes investigated, where there are two levels of bottom ties: non-top-ranked DE genes and non-DE genes. Clearly, the first level should be ranked higher than the second. BiG can be easily expanded to handle this situation. Second, BiG as presented in this paper assumes that genes are independent of each other. However, many studies have shown that genomic data analysis can be greatly improved by modeling correlation structures and borrowing strength among genes^{24,25,26}. With some added structure in the latent variable model (1), BiG can naturally incorporate chromosomal spatial proximity, gene network topology, pathway or functional annotations into RA. Third, there are often a group of genes that have been revealed to be associated with a certain disease. BiG can make use of such information to arrive at better aggregated lists as well as to gauge the quality of component studies. None of the other methods can be adapted to handle such situations. The proposed method is implemented in an R package “BiG” available on CRAN.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The authors are thankful to two reviewers and editors for their constructive comments and suggestions, which considerably improved the manuscript. This work is supported by the NIH grants R15GM113157 (PI: Xinlei Wang).

References

1. Borda Jean C de. Mémoire sur les élections au scrutin. Histoire de l'Académie Royale des Sciences. 1781;.
2. Kolde Raivo, Laur Sven, Adler Priit, Vilo Jaak. Robust rank aggregation for gene list integration and meta-analysis.. *Bioinformatics*. 2012;28(4):573–580. [PubMed: 22247279]
3. Stuart Joshua M, Segal Eran, Koller Daphne, Kim Stuart K. A gene-coexpression network for global discovery of conserved genetic modules. *Science*. 2003;302(5643):249–255. [PubMed: 12934013]
4. Dwork Cynthia, Kumar Ravi, Naor Moni, Sivakumar Dandapani. Rank aggregation methods for the web. In: :613–622; 2001.
5. DeConde Robert P, Hawley Sarah, Falcon Seth, Clegg Nigel, Knudsen Beatrice, Etzioni Ruth. Combining results of microarray experiments: a rank aggregation approach. *Statistical Applications in Genetics and Molecular Biology*. 2006;5(1).
6. Lin Shili. Rank aggregation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*. 2010;2(5):555–570.
7. Lin Shili, Ding Jie. Integration of ranked lists via Cross Entropy Monte Carlo with applications to mRNA and microRNA studies. *Biometrics*. 2009;65(1):9–18. [PubMed: 18479487]
8. Deng Ke, Han Simeng, Li Kate J, Liu Jun S. Bayesian aggregation of order-based rank data. *Journal of the American Statistical Association*. 2014;109(507):1023–1039.
9. Badgeley Marcus A, Sealfon Stuart C, Chikina Maria D. Hybrid Bayesian-rank integration approach improves the predictive power of genomic dataset aggregation. *Bioinformatics*. 2015;31(2):209–215. [PubMed: 25266226]
10. Li Xue, Wang Xinlei, Xiao Guanghua. A Comparative Study of Rank Aggregation Methods for Partial and Top Ranked Lists in Genomic Applications. *Briefings in Bioinformatics*. 2017;.
11. Spiegelhalter David J, Abrams Keith R, Myles Jonathan P. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. John Wiley & Sons, Hoboken, NJ; 2004.

12. Gelman Andrew. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*. 2006;1(3):515–534.
13. Johnson Valen E, Deane Robert O, Van Schaik Carel P. Bayesian analysis of rank data with application to primate intelligence experiments. *Journal of the American Statistical Association*. 2002;97(457):8–17.
14. Liu Jun S. Monte Carlo strategies in scientific computing. Springer Science & Business Media; 2008.
15. Gelman Andrew, Carlin John B, Stern Hal S, Dunson David B, Vehtari Aki, Rubin Donald B. *Bayesian Data Analysis*. CRC press Boca Raton, FL; third ed 2014.
16. Aerts Stein, Lambrechts Diether, Maity Sunit, et al. Gene prioritization through genomic data fusion. *Nature biotechnology*. 2006;24(5):537–544.
17. Shedden Kerby, Taylor Jeremy M G, Enkemann Steven A, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study.. *Nature Medicine*. 2008;14:822–827.
18. Kerkentzes Konstantinos, Lagani Vincenzo, Tsamardinos Ioannis, Vyberg Mogens, Røe Oluf Dimitri. Hidden treasures in “ancient” microarrays: gene-expression portrays biology and potential resistance pathways of major lung cancer subtypes and normal tissue. *Frontiers in oncology*. 2014;4:251. [PubMed: 25325012]
19. Li Yafang, Xiao Xiangjun, Ji Xuemei, Liu Bin, Amos Christopher I. RNA-seq analysis of lung adenocarcinomas reveals different gene expression profiles between smoking and nonsmoking patients.. *Tumour Biology : the Journal of the International Society for Oncodevelopmental Biology and Medicine*. 2015;36:8993–9003. [PubMed: 26081616]
20. Zhou Y, Frings O, Branca RM, et al. microRNAs with AAGUGC seed motif constitute an integral part of an oncogenic signaling network. *Oncogene*. 2016;36:731–745. [PubMed: 27477696]
21. Kim Sang Cheol, Jung Yeonjoo, Park Jinah, et al. A high-dimensional, deep-sequencing study of lung adenocarcinoma in female never-smokers. *PLOS ONE*. 2013;8(2):e55596. [PubMed: 23405175]
22. Chen Min, Zang Miao, Wang Xinlei, Xiao Guanghua. A powerful Bayesian meta-analysis method to integrate multiple gene set enrichment studies. *Bioinformatics*. 2013;29(7):862–869. [PubMed: 23418184]
23. Dempster Arthur P, Laird Nan M, Rubin Donald B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*. 1977;:1–38.
24. Xiao Guanghua, Wang Xinlei, Khodursky Arkady B. Modeling three-dimensional chromosome structures using gene expression data. *Journal of the American Statistical Association*. 2011;106(493):61–72. [PubMed: 21760653]
25. Wang Xinlei, Zang Miao, Xiao Guanghua. Epigenetic change detection and pattern recognition via Bayesian hierarchical hidden Markov models. *Statistics in Medicine*. 2013;32(13):2292–2307. [PubMed: 23097332]
26. Cheng Yichen, Dai James Y, Kooperberg Charles. Group association test using a hidden Markov model. *Biostatistics*. 2016;17(2):221–234. [PubMed: 26420797]

FIGURE 1.

Coverage rates of different methods with varying number of top ranked genes n^T while other parameters are held constant.

FIGURE 2.

Coverage rates of different methods with varying gene inclusion rate λ while other parameters are held constant.

TABLE 1

Performance comparison of algorithms with different prior choices in (a) estimation efficiency using average mean squared errors for μ'_g 's and $\sigma^2_{e \cdot j}$'s; and (b) rank aggregation using coverage rates with the top 10, 50 and 100 cutoffs.

(a)			(b)			
Prior	$MSE(\mu_g)$	$MSE(\sigma^2_{\epsilon \cdot j})$	Prior	Cov10	Cov50	Cov100
IG	0.593	5.204	IG	0.523	0.557	0.610
Unif	0.599	35.855	Unif	0.504	0.542	0.597
DA	0.595	29.66	DA	0.506	0.554	0.606

Comparison of the different methods using coverage rates (reported in percentage) based on the top 20 genes in the aggregated list when n^T and λ vary.

TABLE 2

λ	n^T	MEAN	GEO	MED	Stuart	RRA	MCI	MC2	MC3	CEMC.k	CEMC.s	BIRRA	BARD	BIG
0.6	10	39.6	39.2	27.6	38.9	14	38.6	39.3	39.2	38.9	30.4	3.4	16.7	42.5
	20	43.3	41.7	41.1	41	26.2	41.1	43.8	43.3	43.3	33.9	3.7	22.5	46.9
	50	45.1	45.3	43.9	45.2	41.2	42.4	47.2	45.5	41.1	33.2	4.4	39.7	49.9
	100	43.2	46.3	44.1	47.4	44.6	41.6	46.7	44.3	35.9	32.1	6.7	43	51.3
0.8	10	41.6	40.5	18.9	39.8	8	41.2	42.3	42.7	41.7	33.4	4.4	24.7	46.2
	20	47.1	44.4	36.1	42.2	16.4	44.2	48.5	48.7	48.6	38.3	5.3	34.7	51.4
	50	51.5	49.8	49.7	48	41.2	45.8	52.5	51.3	48	39.9	10	47.3	55.5
	100	51.1	52	50.1	51.8	48.9	45.2	53.2	50	41.7	40.5	22.9	50.5	56.7
1	10	42.1	41.5	15.4	40.8	6.1	41.6	39.8	43.1	41.1	40.2	NA	42.7	47.3
	20	49.6	45.6	31.1	42.9	11.1	45.1	49.2	51.7	51.5	50.8	51.6	48.7	53.6
	50	54.7	52.1	53.1	48.5	31.5	49.8	55	54.5	52	53	54.6	53.8	58
	100	55.4	55.3	53.2	53.4	50.6	51.1	56.8	53.9	45.1	54.5	54.6	55.4	59.9
Mixed		31	39.5	32.3	41.4	34.6	39.2	46.1	44.3	41.7	30.2	5.8	34.8	48.8

Comparison of the different methods using coverage rates (reported in percentage) calculated based on the top 20 genes in the aggregated list when $\sigma^2_{\kappa,1}$ and $\sigma^2_{\kappa,2}$ vary while fixing $n^T = 50$, $\lambda = 0.8$, and $\rho = 0.7$.

TABLE 3

	MEAN	GEO	MED	Stuart	RRA	MCI	MC2	MC3	CEMC.k	CEMC.s	BIRRA	BARD	BiG
L/L	57.2	55.8	55.6	54.8	49.4	54.8	57.5	57.1	53.1	48	11.9	50.2	57.5
L/H	55.1	54.4	53.8	53.6	48.6	53.5	55.3	55.4	51.6	47.2	12	48.6	55.3
H/H	53.8	52.7	52.7	52.5	48.2	52.2	54	54.2	50.7	46.6	12.2	48.2	54.5

TABLE 4

Comparison of the different methods using coverage rates (in percentage) based on the top 20 genes in the aggregated list when (1) μ_g 's are generated from scaled t_3 , $Exp(1)$, and scaled $|N(0, 1)|$, respectively; (2) there is no platform bias (i.e., $\kappa_{gp} \equiv 0$); and (3) a dichotomous model (DM), 5%. $N(1, 1) + 95\% \cdot N(0, 1)$, is used to generate signal vs. non-signal genes.

	MEAN	GEO	MED	Stuart	RRA	MCI	MC2	MC3	CEMC.k	CEMC.s	BIRRA	BARD	BiG
t_3	60	57.6	58.1	55.5	45.9	54.5	60.5	59.1	56.7	48.3	13	54.8	63.3
$Exp(1)$	71.7	70.8	70.3	69.6	62.5	66.7	73	71	66.6	61.9	16.3	66.2	75.2
$ N(0, 1) $	63.2	61.4	61.3	59.8	53.2	57.8	63.2	62.3	58.3	52.4	12.5	58.3	66
$\kappa_{gp} = 0$	53.1	51.3	50.8	48.4	41.6	46.8	53.5	52.7	48.4	40.6	9.6	48.3	56.2
DM	38.9	35.6	36.8	32.6	25.4	33	39.8	40.1	37	24.1	6.4	31.6	37.4

TABLE 5

Sources of ranked lists in the NSCLC example

Studies (Data Set Name)	Technology	Type of list	n_j	n_j^T
Shedden et al. ¹⁷ (CL)	Microarray	Top-ranked with bottom ties	12992	200
Shedden et al. ¹⁷ (NCI__U133A)	Microarray	Top-ranked with bottom ties	12992	300
Shedden et al. ¹⁷ (Moff)	Microarray	Top-ranked with bottom ties	12992	800
Kerkentzes et al. ¹⁸	Microarray	Top-ranked only	3502	3502
Li et al. ¹⁹	RNA-seq	Top-ranked only	2273	2273
Zhou et al. ²⁰	RNA-seq	Top-ranked only	275	275
Kim et al. ²¹	RNA-seq	Top-ranked with bottom ties	20989	500

TABLE 6

Coverage rates with different cutoffs for the NSCLC Example. In the top-ranked only lists, for genes included from the other lists, (a) summarizes results where they are treated as NA's and (b) summarizes results where they are treated as bottom ties.

Cutoff	(a)										
	MEAN	GEO	MED	RRA	Stuart	MCI	MC2	MC3	BARD	BIRRA	BiG
100	0.0	1.3	1.3	1.3	2.6	2.6	2.6	3.9	0.0	1.3	0.0
200	2.6	2.6	2.6	2.6	2.6	3.9	2.6	3.9	3.9	1.3	3.9
300	3.9	2.6	3.9	2.6	2.6	5.2	5.2	3.9	3.9	1.3	9.1
400	5.2	6.5	5.2	2.6	3.9	10.4	6.5	9.1	7.8	1.3	9.1
500	7.8	7.8	6.5	3.9	3.9	13.0	7.8	10.4	9.1	1.3	10.4

Cutoff	(b)										
	MEAN	GEO	MED	RRA	Stuart	MCI	MC2	MC3	BARD	BIRRA	BiG
100	1.3	2.6	1.3	2.6	2.6	2.6	0.0	2.6	1.3	1.3	3.9
200	3.9	2.6	1.3	2.6	2.6	5.2	5.2	5.2	3.9	3.9	6.5
300	5.2	3.9	2.6	2.6	2.6	6.5	7.8	6.5	5.2	9.1	7.8
400	6.5	7.8	7.8	7.8	2.6	10.4	10.4	6.5	6.5	9.1	13.0
500	9.1	9.1	9.1	10.4	2.6	10.4	13.0	9.1	10.4	10.4	14.3