Bioinformatics, 36(1), 2020, 177–185
doi: 10.1093/bioinformatics/btz471
Advance Access Publication Date: 14 June 2019
Original Paper

Gene expression

An empirical Bayesian ranking method, with applications to high throughput biology

John Ferguson[1,]* and Joseph Chang[2]

[1]Biostatistics Division, HRB Clinical Research Facility, National University of Ireland Galway, Galway, Ireland and [2]Department of Statistics and Data Science, Yale University, New Haven, CT, USA

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on April 4, 2018; revised on April 30, 2019; editorial decision on May 29, 2019; accepted on June 5, 2019

Abstract

Motivation: In bioinformatics, genome-wide experiments look for important biological differences between two groups at a large number of locations in the genome. Often, the final analysis focuses on a *P*-value-based ranking of locations which might then be investigated further in follow-up experiments. However, this strategy may result in small effect sizes, with low *P*-values, being ranked more favorably than larger more scientifically important effects. Bayesian ranking techniques may offer a solution to this problem provided a good prior distribution for the collective distribution of effect sizes is available.

Results: We develop an Empirical Bayes ranking algorithm, using the marginal distribution of the data over all locations to estimate an appropriate prior. In simulations and analysis using real datasets, we demonstrate favorable performance compared to ordering *P*-values and a number of other competing ranking methods. The algorithm is computationally efficient and can be used to rank the entirety of genomic locations or to rank a subset of locations, pre-selected via traditional FWER/FDR methods in a 2-stage analysis.

Availability and implementation: An R-package, *EBrank*, implementing the ranking algorithm is available on CRAN.

Contact: john.ferguson@nuigalway.ie

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

1 Introduction

Traditional statistical hypothesis testing was developed based on the ideas of *P*-values and levels of significance and is still in wide-spread use today, despite well-known caveats regarding its usage (Amrhein *et al.*, 2019; Nickerson, 2000). A recent example is the statistical analysis of large genomic datasets, where perhaps hundreds of thousands of genetic locations are simultaneously tested for systematic differences between two groups. Different tests need to be used depending on the experimental technology that is being used, and a surprising degree of statistical ingenuity has been spent developing appropriate test statistics and *P*-values. Whereas significance testing should do reasonably well at identifying locations where real (although potentially small) biological differences exist, there is no guarantee that small *P*-values correspond to scientifically interesting differences. Nevertheless, *P*-values are frequently used as a ranking criterion to produce lists of the most interesting tests/locations to follow up.

As an example, consider Genome-Wide Association Studies (GWAS). These are large scale observational studies where a set of single nucleotide polymorphisms (SNPs) are genotyped for a group of control individuals, and typically a group of individuals with some disease (although continuous traits such as height and body mass index have also been investigated). The genotype for a given individual at a given SNP can be coded as 0, 1 or 2 depending on the number of copies of the variant allele at that locus for the individual. The associative effect of each SNP on disease risk is measured using the population log-odds ratio (OR) between SNP genotype and disease status. While we would ideally want any list of ‘significant’ SNPs to concentrate on the largest population ORs, the variant allele for some SNPs will be rare, meaning that most individuals have genotype 0, and relatively few have genotypes 1 and 2. As a result, there is low power to detect a disease/control difference in the genotypes at these rare SNPs. For example, suppose that SNP A has a disease/control OR of 1.1 and a minor allele frequency (MAF) in controls of 0.4, whereas SNP B has an OR of 1.5 but the MAF (again in controls) is only 0.01. Assuming the disease is rare; this implies that having the variant allele for SNP A is associated with a 10% increase in the probability of disease, with a more important looking increase of 50% for SNP B. However, in practice the respective *P*-values corresponding to these two SNPs may not reflect this information. Indeed, calculations show that the rare SNP with the higher OR will have the lower *P*-value for a sample of 1000 cases and 1000 controls only 48% of the time, assuming independence of the genotypes at the two SNPs (this independence assumption is referred to as 'linkage equilibrium' in the genetics literature).

Bayesian ranking methods provide an alternative approach to constructing lists of interesting hypotheses, or in the cases considered here DNA locations, which de-emphasizes the influence of varying experimental standard errors in comparison with a *P*-value based ranking. Our work here is motivated by a number of previous authors as far back as 1989, where Laird and Louis (1989) used Empirical Bayes ranking techniques to assess relative educational effectiveness of a number of secondary schools. In this manuscript, we develop a general Bayesian ranking method that is applicable to a wide range of genomic datasets. Our method is related to the recent papers demonstrating the use of Bayesian ranking for differential expression in microarrays (Noma *et al.*, 2010; Noma and Matsui, 2013); however, we use a differing Bayesian model, and demonstrate the use of Bayesian ranking in a wider set of scenarios including determining differential expression from RNA-Sequence data and identifying disease associations in GWAS. Simulations and real-data analysis are used to compare the properties of our method against multiple competing approaches. To facilitate the use of our method, we have developed an R-package, *EBrank*, implementing these methods which is downloadable from the CRAN repository.

The rest of the paper is organized as follows. In Section 2, we discuss ranking problems from a general point of view, including a discussion of appropriate loss functions and Bayesian techniques to produce loss-minimizing rankings. In Section 3, we describe a novel non-parametric Empirical Bayes ranking algorithm that is particularly suited to examples from bioinformatics. In Section 4, we examine the performance of the algorithm using a series of simulated and real-data examples, and compare to other approaches for ranking effect sizes. To conclude, we discuss potential limitations of our methodology, when we would expect it to work well, and suggest some potential extensions.

2 Background

As alluded to in the previous section, using *P*-values to rank experiments in massive parallel testing situations may give unsatisfactory results. *P*-values and Bayesian ranking techniques typically optimize different criteria, with the criterion corresponding to *P*-values not as suited for ranking. To make these criteria and distinctions between them more tangible, it is helpful to begin with a brief discussion of loss functions that can be used for ranking problems. We conclude the section with a description of how Bayesian ranking methods are applied in practice, given the choice of such a loss function.

2.1 Loss functions

Suppose we observe N random variables for which the probability distribution for the ith variable depends on an unknown real-valued parameter $_{\theta i}$, for $i \leq N$. Assume that large values of $|\theta_i|$ correspond to effects of scientific interest, whereas values $\theta_i = 0$ indicate no effect. As a result, we can imagine a 'true' ranking of the distributions, which is any ranking consistent with decreasing values of $|\theta_i|$. In other words, assuming no ties among $\{\theta_1, \ldots, \theta_N\}$, the true rank vector, $R = (R_1, \ldots, R_N)$, is defined as

$$R_i = R_i(\theta_1, \ldots, \theta_N) = \sum_{j \leq N} I\{|\theta_j| \leq |\theta_i|\} \tag{1}$$

for $i \leq N$, with $I(A)$ representing the indicator function for the event A. The goal of a ranking procedure is to produce a vector of estimated ranks, $\hat{R} = (\hat{R}_1, \ldots, \hat{R}_N)$, which is as 'close' as possible to the true ranking. The chosen metric that compares how close $\hat{R}$ and R lie corresponds to a loss function. A variety of possibilities exist for the chosen loss function (Critchlow, 2012), the appropriateness of each depending on the scientific task at hand. For instance, in a GWAS N can be extremely large, and most of the associated $|\theta_i|$ may be either 0 or minute and uninteresting. Correctly estimating the ranks for the largest $|\theta_i|$ is obviously of much greater importance than the ranks corresponding to small $|\theta_i|$. As a more concrete example, perhaps the investigator is only interested in producing a list of the K largest $|\theta_i|$. A loss function that might be minimized to achieve this goal is:

$$L_O((\theta_1, \ldots, \theta_N), \hat{R}) = \frac{100}{K} \sum_{i \leq N} I(R_i > N - K, \hat{R}_i \leq N - K) \tag{2}$$

with the subscript 'O' indicating the goal of maximizing the overlap between the K most highly ranked parameters (by the method) and the K truly largest parameters. Here the optimal ranking will change depending on the value of K. This loss function (and generalizations that also focus on mis-classifications regarding top parameter sets) were introduced in Lin *et al.* (2006). 'R-values' are a related approach to maximize overlap between reported and true top parameter lists, but instead only produce a single ranking and so might be used instead when the rankings of all $(\theta_1, \ldots, \theta_N)$ are equally important, or alternatively when a single percentile, represented by K, is hard to define. In a sense, ranking according to r-values optimizes a modified version of (2) over all K, see Henderson and Newton (2016) for more details. Another approach that focuses on the entire parameter vector is to use an L^p rank-loss such as the following:

$$L^p((\theta_1, \ldots, \theta_N), \hat{R}) = \frac{1}{N} \sum_{i \leq N} |R_i(\theta_1, .., \theta_N) - \hat{R}_i|^p, \tag{3}$$

the most analytically tractable cases being represented by $P = 1$ (this loss is also known as Spearman's footrule) and with $P = 2$, for rank square error loss. In practice, due to the additive form of the loss function, it is likely that optimizing L^p type losses will give good results even when only the top K parameters are of interest. Newton and Henderson mention that in numerical experiments the results from applying r-values is quite similar to ranking by posterior expected ranks, which optimizes (3) when $p = 2$. More recently, Jewett *et al.* (2018) suggested a number of novel loss functions for ranking. The idea here is that the penalty of assigning R_i to position i should depend on the difference between the true parameters: $_{\theta i}$ and $\theta_{(\hat{R}_i)}$, rather than the difference in the true ranks leading to losses of the form:

$$L^J((\theta_1, \ldots, \theta_N), \hat{R}) = \frac{1}{N} \sum_{i \leq N} (\theta_i - \theta_{(\hat{R}_i)})^p. \tag{4}$$

While theoretically appealing, finding rank vectors to optimize the posterior expectation of this loss is somewhat difficult when N is large, as explained in the Supplementary Section 2. L^2 rank-loss is much easier to optimize, and at least in the simulations we have considered generates reasonable rankings. With this in mind, we use L^2 rank-loss as the default ranking criterion in the R-package *EBrank*. This package also reports posterior probabilities for $|\theta_i|$ lying among the K largest $|\theta_j|, j \leq N$, which can be used in minimizing (2) for various values of K. Note that in some situations, a subset of experiments that show high evidence for some treatment effect might be pre-selected; a primary example being GWAS, where the main aim is to find a subset of effects corresponding to non-zero population log-ORs, from an initial array of perhaps hundreds of thousands of SNPs. In such cases, ranking procedures like the one described here can be applied to a reduced subset of experiments, used to inform the prior distribution. Simulations (in the Supplementary Section 2) suggest that pre-selection on P-values using a false discovery rate (or other) threshold before ranking can improve ranking performance (as well as reducing Monte Carlo error and speeding up computation time) when true effect sizes are weak; indicating that this strategy might be used as a default method for GWAS. Pre-selection using false discovery rate, family wise error rate and posterior probability thresholds can be automatically implemented using the R-package *EBrank*.

Ranking with P-values Under certain conditions the ranking vector produced by ordering P-values optimizes the loss function

$$L_{T1}((\theta_1,\ldots,\theta_N),\hat{R}) = \sum_{i\leq N} \hat{R}_i I\{\theta_i = 0\}, \tag{5}$$

indicating that highly ranked parameters should correspond to non-zero parameter values (or alternative hypotheses), or informally that the number of Type 1 (T1) errors among highly ranked parameters should be minimized. Note that (5) does not measure the proximity between the true and estimated rank vectors and as a result estimated rankings derived via minimizing (5) may perform poorly when judged according to loss functions designed for ranking such as (2) or (3). We give further details regarding the connection between optimizing (5) and ordering P-values in the Supplementary Material.

2.2 Finding $\hat{\boldsymbol{R}}$ to minimize loss

Finding rank vectors that optimize loss criteria such as (3) within the Frequentist paradigm is usually difficult, although some progress regarding minimax-optimal solutions has been achieved for special loss functions (Bansal *et al.*, 1997). Bayesian modeling offers a computationally viable and principled alternative, and is the approach that we consider in this article. First consider a prior distribution, $\pi(\theta_1,\ldots,\theta_N|\gamma) = \prod_{i\leq N}\pi(\theta_i|\gamma)$ for the parameter vector $(\theta_1,\ldots,\theta_N)$, γ being a set of hyper-parameters specifying the exact prior chosen from a class of plausible priors. A fully hierarchical approach incorporates an additional hyper-prior for γ. Alternatively, Empirical Bayes first specifies a data-driven estimate $\hat{\gamma}$ for γ and subsequently uses the non-hierarchical prior distribution: $\pi(\theta_1,\ldots,\theta_N|\hat{\gamma}) = \prod_{i\leq N}\pi(\theta_i|\hat{\gamma})$. In either case, we can then derive the posterior distribution, $\pi(\theta_1,\ldots,\theta_N|X)$ for $(\theta_1,\ldots,\theta_N)$, given the data X using simulation or analytic techniques using existing procedures. Sometimes the variance of the posterior distribution may be adjusted upwards to account for uncertainty due to error in estimating γ (Berger, 1985), although the Empirical Bayes algorithm described in Section 3 is 'naive' in the sense that this extra source of variation is not considered.

For certain classical loss functions that we consider here, a posterior-loss-minimizing solution for $\hat{\mathbf{R}}$ can be easily determined, once the posterior distribution is known. For instance, in the case of Equation (2), any rank vector $\hat{\mathbf{R}}$ satisfying $\sum_{i=1}^{N} I(\mathbb{P}(R_j > N-K|\mathbf{X}) \geq \mathbb{P}(R_i > N-K|\mathbf{X})) > N-K$, whenever $\hat{R}_j > N-K$ will minimize posterior expected loss (Shen and Louis, 1998). This solution corresponds to selecting the K experiments with the highest posterior probabilities that the parameter $|\theta_j|$ is among the K largest $|\theta_i|, i \leq N$. As another example in the case of L^p loss, the posterior expectation of (3) can be minimized by setting: $\hat{R}_j = \text{median}(R_j(\theta_1,\ldots,\theta_N)|X)$ for $P = 1$ and $\hat{R}_j = \mathbb{E}(R_j(\theta_1,\ldots,\theta_N)|X)$, when $p=2$ (Laird and Louis, 1989). If Markov Chain Monte Carlo techniques are used to estimate the posterior distributions, ranks can be sampled by transforming the sampled parameters at each iteration, k: $\theta_1^{(k)},\ldots,\theta_N^{(k)}$, for $k = 1,\ldots,M$ (M being the number of MCMC samples drawn) into ranks $(R_1^{(k)},\ldots,R_N^{(k)})$ via:

$$R_i^{(k)} = R_i(\theta_1^{(k)},\ldots,\theta_N^{(k)}). \tag{6}$$

$\hat{R}_j$ can then be estimated easily from the sampled ranks; for instance, we can estimate $\text{median}(R_j(\theta_1,\ldots,\theta_N)|X)$ by the empirical median of $R_j^{(k)}$ over $k \leq M$.

3 An empirical Bayes ranking algorithm

We next develop an Empirical Bayes implementation for the ranking techniques described in the previous section. The described algorithm might be considered an approximation to a fully Bayesian approach, with a clear advantage in computational feasibility. An R-package implementing the algorithm can be downloaded from the CRAN repository. The algorithm is designed to estimate the ranks for the absolute values of N unknown univariate parameters $(\theta_1,\ldots,\theta_N)$ based on noisy estimates $(\hat{\theta}_1,\ldots,\hat{\theta}_N)$. This is a commonly observed set up in bioinformatics, and examples from GWAS and RNA-Sequencing will be described in the following section. While the algorithm can be used in an *ad-hoc* fashion given any set of estimates, a few relative weak assumptions are necessary to justify the likelihood model that is implicitly used in Empirical Bayes estimation. These assumptions are as follows:

- The estimates $(\hat{\theta}_{1,\ldots,}\hat{\theta}_N)$ are conditionally independent, given $(\theta_1,\ldots,\theta_N)$
- $\hat{\theta}_i$ is asymptotically normally distribution in that $\frac{\hat{\theta}_i-\theta_i}{\sigma_i}\underset{\rightarrow}{d}N(0,1)$ as sample size $n \rightarrow \infty$.
- Estimated standard errors: $\hat{\sigma}_i$ are available and satisfy a kind of 'log-consistency', i.e. $\hat{\sigma}_i/\sigma_i P1$
- $\hat{\theta}_i$ and $\hat{\sigma}_i$ are asymptotically independent.

The weakness of these assumptions supports the use of the algorithm in a variety of settings. It should be pointed out that the algorithm learns an appropriate prior using Empirical Bayes technique, using aggregate information in the N random variables $\hat{\theta}_i, i \leq N$. An implicit assumption is that N is sufficiently large to estimate this prior reasonably accurately. When N is small, the learned prior may be a poor reflection of how $(\theta_1,\ldots,\theta_N)$ are distributed in reality, and ranking performance may suffer.

3.1 Derivation of an approximate conditional likelihood

Asymptotic normality implies that $\frac{\hat{\theta}_i}{\sigma_i}\dot{\sim}N(\frac{\theta_i}{\sigma_i},1)$ for large n. So since $\hat{\theta}_i$ and $\hat{\sigma}_i$ are asymptotically independent, conditional on $\hat{\sigma}_i$ we still have $\frac{\hat{\theta}_i}{\sigma_i}\dot{\sim}N(\frac{\theta_i}{\sigma_i},1)$. Now log-consistency of the estimated standard errors implies that $\hat{\sigma}_i = \sigma_i(1+o_P(1))$ so that, multiplying by $\frac{\sigma_i}{\hat{\sigma}_i}$, we get:

$$Z_i = \frac{\hat{\theta}_i}{\hat{\sigma}_i}\dot{\sim}N\left(\frac{\theta_i}{\hat{\sigma}_i},\frac{\sigma_i^2}{\hat{\sigma}_i^2}\right) \approx N(\mu_i,1), \tag{7}$$

defining $\mu_i = \frac{\theta_i}{\hat{\sigma}_i}$. Taking the product of these individual densities over $i \leq N$ gives the conditional approximate likelihood

$L(\mu_1, \ldots, \mu_N) = \prod_{i \le N} \phi(z_i - \mu_i)$, conditioned on the values of the standard errors, $\hat{\sigma}_1, \ldots, \hat{\sigma}_N$, where ϕ denotes the standard normal density function.

3.2 Prior and posterior for μ

We choose a flexible class of prior distributions, where iid univariate mixture normal distributions are specified for the marginals, μ_i for $i \le N$, and independence is assumed for μ_i and μ_j with $i \ne j$. The class of mixture normals is restricted so that the first component is a point mass at $\boldsymbol{\mu} = 0$ and the other $J \ge 1$ components are continuous, so that the overall prior can be written as:

$$\pi(\mu) = \prod_{i \le N} \pi(\mu_i),$$

with

$$\pi(\mu_i) = p_0 \delta\{0\} + \sum_{j=1}^{J} p_j \phi\left(\frac{\mu_i - m_j}{s_j}\right) / s_j, \tag{8}$$

where $(p_0, \ldots, p_J)$ is a vector of probabilities summing to 1, $\delta\{0\}$ is the delta function centered at 0 and m_j and s_j are means and standard deviations, characterizing the composite mixture distributions. The mixture normal family is extremely broad, indeed any continuous density function can be arbitrarily closely approximated by a finite normal mixture provided J is large enough, so in effect (8) can be considered a non-parametric model for the distribution of μ_i.

Proceeding now as if the approximation (7) is exact, the marginal density for z_i can be deduced by noting that $z_i = \mu_i + \epsilon_i$, where $\epsilon_i \sim N(0, 1)$, and has a similar mixture form:

$$f(z_i) = p_0 \phi(z_i) + \sum_{j=1}^{J} p_j \phi\left(\frac{z_i - m_j}{\sqrt{1 + s_j^2}}\right) / \sqrt{1 + s_j^2}.$$

Finally, standard calculations show that the posterior distribution for μ_i given z_i also has a mixture normal form, with a point mass at 0, and another mixture normal with J components, as follows:

$$\begin{aligned} \pi_\mu(\mu_i | z_i) &= \frac{\sum_{j=1}^{J} p_j f_j(z_i) \pi_j(\mu_i | z_i)}{f(z_i)} \mu_i \ne 0, \\ \mathbb{P}(\mu_i = 0 | z_i) &= \frac{p_0 \phi(z_i)}{f(z_i)}, \end{aligned} \tag{9}$$

where $f_j(z_i) = \phi((z_i - m_i)/\sqrt{1 + s_j^2})/\sqrt{1 + s_j^2}$ and $\pi_j(\mu_i | z_i) = \phi((\mu_i - m_j^*(z_i))/s_j^*)$, with $m_j^*(z_i) = (s_j^2 z_i + m_j)/(1 + s_j^2)$ and $s_j^* = \sqrt{s_j^2/(1 + s_j^2)}$. The joint posterior can be derived as a product of these densities over i.

3.3 Empirical Bayes estimation of prior for μ

We utilize a constrained EM algorithm adapted from Muralidharan *et al.* (2010) to estimate the parameters: $\mathbf{m} = (m_1, \ldots, m_J)$, $\mathbf{s} = (s_1, \ldots, s_J)$ and $p = (p_0, .., p_J)$ by maximizing the marginal likelihood $\prod_{i \le N} f(z_i)$. The default algorithm constrains the mean and standard deviation for component 0, i.e. m_0 and s_0, to equal 0 and 1, respectively, whereas the variances are constrained to be at least 1 for components 1,.., J. Finally, for the examples in this manuscript we have forced the restrictions $m_1 = 0$ and $s_1 = 10$. Here, s_1 is constrained so that the algorithm can detect a low-probability mixing component representing isolated extremely large parameters, as sometimes may be observed in genome-wide association applications. If such a component is not included, the estimated effect sizes for these extreme parameters might be aggressively and incorrectly shrunk to an overly-conservative prior. This restriction mostly impacts estimated posterior means rather than estimated ranks and can be removed in *EBrank* if so desired. See Supplementary Section 3 for further discussion. $\hat{J}$ is selected by minimizing BIC over the integers $J \in \{1, \ldots, J_{\max}\}$. The estimated mixing probabilities, $\hat{p}$, and component means, $\hat{\mathbf{m}}$, are found directly from the algorithm—while the estimated variances are found using $\hat{\mathbf{s}}^2 = \hat{\sigma}_{EM}^2 - 1$, where $\hat{\sigma}_{EM}$ is the vector of standard deviations found via the algorithm.

3.4 Estimated posterior for θ_i

The posterior distribution for θ_i is calculated as

$$\begin{aligned} \hat{\pi}(\theta_i | z_i) &= \pi_\mu(\theta_i / SE(\hat{\theta}_i)) / SE(\hat{\theta}_i), \theta_i \ne 0 \\ \mathbb{P}(\theta_i = 0 | z_i) &= \mathbb{P}(\mu_i = 0 | z_i), \end{aligned} \tag{10}$$

where $\pi_\mu(.|z_i)$ is given by (9) with the Empirical Bayes derived parameter estimates $\hat{m}_j$, $\hat{s}_j$ and $\hat{p}_j$ substituted for m_j, s_j and p_j for $j \le \hat{J}$. In effect, the continuous part of the posterior for $_{\theta i}$ is found from (9) by multiplying the means and standard deviations of each mixing component by $\hat{\sigma}_i$. These estimated mixture posterior distributions for differing $i = 1, \ldots, N$ are assumed independent.

3.5 Estimating ranks

While it is possible to calculate mean posterior ranks directly from the posterior via the identity $\mathbb{E}(R_i | X) = \sum_{j=1}^{N} \mathbb{P}\{\theta_i \ge \theta_j | \mathbf{X}\}$, it is easier to simulate from the posterior (9), and then transform the simulated values into ranks using (6). Our *R*-package EBrank uses a default of M =10 000 draws from the posterior. Estimated posterior-loss-minimizing rank vectors are then found by replacing the population solutions by their obvious sample analogs. For instance, $\hat{R}_i = \mathbb{E}(R_i(\theta_1, \ldots, \theta_N) | X)$ is estimated by $\tilde{R}_i = \frac{\sum_{j \le M} R_i^{(j)}}{M}$.

4 Examples

4.1 RNA sequence example simulations

RNA-sequencing has emerged as a competitor to microarrays for quantifying and comparing gene expression under different conditions, e.g. Mortazavi *et al.* (2008). The data are the number of short pieces (reads) of cDNA that have been sequenced, lying in each gene or transcript. Gene expression is indirectly measured from the number of reads that map, or align, to a particular gene, normalized by the gene's length (in nucleotides) and the total count of reads over all genes (known as the library size). Genes are typically identified as being differentially expressed using RNA-seq data when their counts for one condition, normalized only by library size and not by gene lengths, are systematically different than the other. Several competing statistical approaches have been developed to identify genes displaying differentially expressed genes in RNA-Seq data [see for instance Anders and Huber (2010), Robinson et al. (2010) and Hardcastle (2015)]. While these methods assume that the counts over the different samples follow a negative-binomial distribution, the test statistics used to identify differentially expressed genes are generally asymptotically normal, at least in the case of large library sizes, supporting the use of our model. Here, we have used the package the R-package*DESeq*2, v1.18 on default settings, to measure gene-wise fold-changes and standard errors from an RNA-Seq count matrix.

Data simulation

Each scenario simulates negative binomial distributed read-counts for 36 536 genes, and $2n$ samples (n being 10, 20 or 50) divided into two equal size groups. A small proportion $P_{DE} \in \{0.05, 0.1, 0.2\}$ of the genes are differentially expressed, in that the average read count for those genes in group 2 differs from the average read count in group 1 after scaling for differences in library sizes. Let θ_i represent the log (base2) ratio between the mean counts for gene i in group 2 and group 1 (commonly referred to as the log of the 'fold-change' in the expression analysis community). For the genes tagged as differentially expressed θ_i is simulated using a $N(0, \sigma^2)$ distribution, with σ set as either 0.5 or 1, dependent on the simulation. For the non-differentially expressed genes, $\theta_i = 0$. We denote the raw read count for sample j of gene i as G_{ij}. These counts are simulated according to a negative-binomial distribution, given by (11), where $NB(m, d)$ corresponds to a negative-binomial distribution having mean m and dispersion parameter d :

$$G_{ij} \sim \begin{array}{l} NB(S_j m_i, d_i) \text{ for j } \in \text{group } 1 \\ NB(S_j 2^{\theta_i} m_i, d_i) \text{ for j } \in \text{ group } 2. \end{array} \qquad (11)$$

The fixed parameters m_i, S_j and d_i, representing scaled mean counts, relative library sizes and dispersions, were determined using a mouse expression dataset first analyzed in Bottomly *et al.* (2011) and available through the Recount database (Frazee *et al.*, 2011). A more complete description of these simulations is given in the Supplementary Material. After simulating the read-counts, log-fold-change estimates $\hat{\theta}_i$ and their standard errors, $SE(\hat{\theta}_i)$, were calculated with *DESeq*2 using the data for each gene. Estimated ranks for the θ vector were produced using Bayesian Ranking, as described in Section 3, by Bayesian ranking using the Empirical prior estimated using a 'smoothing by roughening' approach as illustrated by Noma and Matsui (2013) and implemented through the *GaussianSBR* function within the R-package *hbsim* (an approach that also uses just the effect size estimates and standard errors), two ranking methods that use the original count matrix: non-negative matrix factorization, as described in Jia *et al.* (2015) and a fold-change based method, FCROS, described by Dembélé and Kastner (2014) and finally by ordering the DESeq2 computed P-values. We used 1000 simulations to estimate mean posterior ranks for each of the Empirical Bayes approaches. To compare the performance of the different methods, we examined the fraction of the genes, having the K largest absolute-value fold-changes that are also included in the top K genes according to a specific ranking. This implies measuring the performance of the estimated ranks, $\hat{R}$ in their estimation of R via the quantity:

$$O_K(R, \hat{R}) = \frac{100}{K} \sum_{i \le N} I(R_i > N - K, \hat{R}_i > N - K) \qquad (12)$$

for various values of K. We call this the 'overlap' between R and $\hat{R}$. Values of K =10 and 100 are examined below.

Each simulation scenario required fixing $p_{DE} \in \{0.05, 0.1, 0.2\}$, $n \in \{10, 20, 50\}$ and average effect size, $\sigma \in \{0.5, 1\}$. Eight independent repeats of each individual scenario were simulated. In Figure 1, the distribution of $O_K(R, \hat{R})$ for Bayesian Ranking and P-values is displayed for $K \in \{10, 100\}$ and $\sigma \in \{0.5, 1, 2\}$, setting $P_{DE} = 0.05$ and averaging across the settings for n. In all but one scenario, estimated overlap is highest according to Bayesian Ranking in Section 3. Results from alternative parameter settings are presented in the Supplementary Material, and show similar conclusions. In the Supplementary Material, we also investigate the effect of a number of other factors that might influence the quality of ranking, including pre-selection of genome-wide significant experiments before the ranking step, loss-function choice (including the Jewett loss and the 'r-values' approach detailed in Section 2) and the effect of Monte Carlo error.

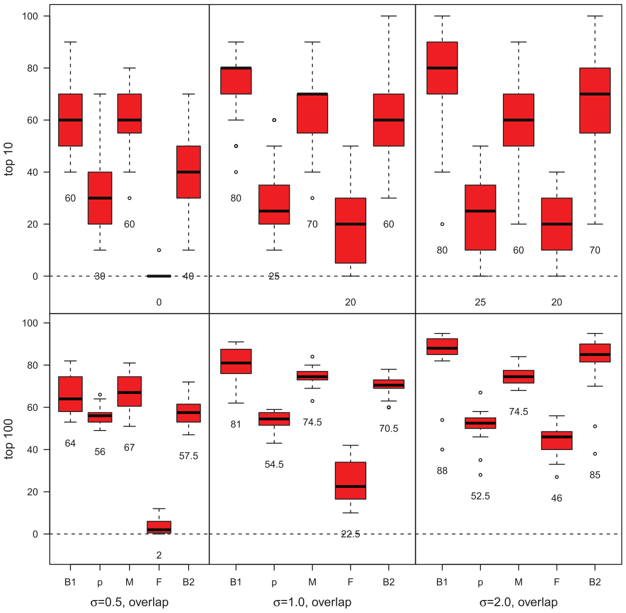

Fig. 1. RNA-seq simulation. In each sub-window, average overlaps over eight simulations are shown for five methods (*B*1: Bayesian ranking as in Section 3, *p*: DESeq2 *P*-values, *M*: non-negative matrix factorization, *F*: fold-change based ranking using FCROS and *B*2: Bayesian ranking using smoothing by roughening). Approximately 5% of genes were simulated to be differentially expressed

Table 1. GWAS studies used in simulations

Disease	# SNPs	# Cases	# Controls	dbGap study accession #
Parkinson's	463 185	1713	3978	phs000501.v1.p1
Crohn's	298 391	968	995	phs000130.v1.p1
Schizophrenia	700 490	1351	1378	phs000021.v1.p1
MS	551 642	978	883	phs000171.v1.p1

Note: Results for Crohn's disease are displayed in Figure 2. See the Supplementary Material for results pertaining to the other diseases.

4.2 Genome-wide association data

To investigate the potential benefits of applying the ranking algorithm in GWAS settings, we downloaded some publicly available SNP data from the DbGAP website (Mailman *et al.*, 2007) for four different diseases. The number of cases and controls and original number of genotyped SNPs for each disease are described in Table 1. For each SNP, disease/control ORs were estimated from the data and shrunk using an Empirical Bayes approach (Ferguson *et al.*, 2013b). Supplementary Figure S2.3 shows both the before and after log-ORs. For each disease, MAF for ~95% of the genotyped SNPs were found using the UCSC genome browser website by matching rsid. The shrunken log-ORs for the N_{all} SNPs where the MAF could be recovered represent true ORs $\mu_1, \ldots, \mu_{N_{all}}$ behind each simulation.

Simulation of data

Two scenarios were investigated: 'weak' signals (n = 2000 cases, n = 2000 controls) and 'strong' signals (n = 10 000 cases,

$n = 10\,000$ controls) via applying the ranking algorithm to 24 independent simulated datasets based on the 'true' $\mu_1, \ldots, \mu_{N_{all}}$ above. Each iteration involved simulating sample ORs and associated P-values for each of the N_{all} SNPs, using an asymptotic approximation for the distribution of the logistic-regression estimated log-OR, given the true MAFs in cases and controls, true μ_i for that SNP and value for n. As a pre-filtering step to ease computational difficulties, we then selected only the subset of $N < N_{all}$ SNPs that had significant P-values after adjusting for a 90% false discovery rate threshold (Benjamini and Hochberg, 1995) for subsequent ranking. This subset of SNPs was ordered based on their P-values and according the Bayesian ranking procedure described in Section 3. However, note that while only a subset of N SNPs was ranked, all N_{all} SNPs were used to calculate the Empirical Bayes informed prior. In the subsequent discussion, we relabel the subset of N parameters that are chosen to be ranked for each simulation as $\theta_1, \ldots, \theta_N$.

Results

Two criteria were used to evaluate the quality of the estimated ranks in recovering the ordering of $(\theta_1, \ldots, \theta_N)$:

1. Overlap of top K SNPs—i.e. the % of the top K ranked SNPs (according to the method either Bayes or P-value) which are in the top K true absolute-value log-ORs (according to the Empirical Bayes shrinkage), given by (12).
2. Percentile Rank of top K SNPs. This is the average percentile of the top K ranked SNPs within the total list of absolute-value logORs as given by: $\frac{1}{K}\sum_{\hat{R}_i > N-K}(100\frac{R_i}{N})$.

Values of K =10 and 100 were considered for both criteria. Boxplots, based on the 24 simulations for each scenario, are displayed in Figure 2 for Crohn's disease. Each pane of the figure corresponds to a particular scenario (either weak or strong signals) and criterion function. The boxplot on the center and left correspond to Bayesian and P-value based rankings. The center boxplot: 'Overall' shows the distribution $P_{Bayes} - P_{pvalue}$ over these eight simulations, where P_x is the value of the criterion function for a particular simulation and ranking method x. These simulations indicate that typically there is a benefit in producing ordered lists of P-values using Bayesian ranking rather than using P-values, sometimes substantially so. For instance, from the 10 SNPs with highest true OR, Bayesian ranking picks out either 9 or 10 of these in the top 10 estimated SNPs compared to only seven when P-values are used. Similar plots for the three other diseases are given in the Supplementary Material.

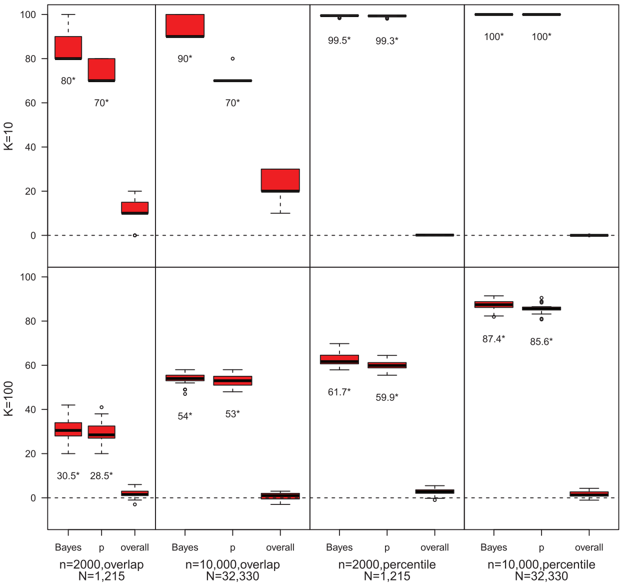

Fig. 2. Boxplots, based on eight simulations for each scenario, are displayed for Crohn's disease. Each pane of the figure corresponds to a particular scenario (either weak or strong signals) and criterion function. The boxplot on the left and center of each pane correspond to Bayesian and P-value based rankings. The boxplots on the right show the distribution of $P_{Bayes} - P_{pvalue}$ over the simulations, where P_x is the value of the criterion function for a particular simulation and ranking method x

4.3 Real-data applications

4.3.1 GWAS

In GWAS, the usual practice was to focus mostly on SNPs that were genome-wide significant, i.e. SNPs that have P-values <0.05 after a Bonferoni adjustment, for follow-up experiments. The reason for this stringent selection was a concern that the selected SNPs might demonstrate significant associations in follow-up studies. As described in Section 4.2, Bayesian ranking can be applied just to a selected group of SNPs, although the Empirical Bayes estimate of the prior must be deduced from the entire set of SNPs. To illustrate this idea, we selected 'genome-wide significant' SNPs for the Crohn's disease data reported in Table 1, which consisted of 968 cases and 995 controls. Our Empirical Bayes ranking procedure, as described in Section 3, was then applied to these 15 SNPs (all SNPs being used to estimate the Empirical Bayes prior, but only the top 15 used in the ranking procedure). Table 2 shows the genome-wide significant SNPs, their ORs, standard errors and ranking according to P-values and estimated Bayesian rankings from the procedure. Note that the SNP ranked number 1 according to Bayesian ranking had the largest absolute-value log-OR from the 15 SNPs; however, the standard error (0.166) is larger than most of the SNPs on the list and as a result, several SNPs have smaller P-values. Note that Bayesian ranking gives almost identical results to ranking by the absolute values of the estimated raw log-ORs in this example, which is an expected feature of the procedure for a group of experiments with low-standard errors. In this case, the Bayesian ranking procedure gives results that are verifiably superior to a P-value based ranking. To show this, we matched SNP ids with the most significant signals from a more recent and comprehensive GWAS of Crohn's disease involving independent discovery and replication datasets of 6333 cases and 15 056 controls and 15 694 cases and 14 026 controls (Franke *et al.*, 2010). While all 15 significant SNPs from the initial study, shown in Table 2, were involved in one of three significant regions in the follow-up study, only rs2076756 and rs11209026 were exact matches with the 71 'top' SNPs reported in Franke *et al.* (2010), with rs11209026 (which was ranked 1 by Bayesian ranking, but only 4 using P-values) having the larger absolute-value log OR.

4.3.2 RNA-Seq

To investigate how well the described Empirical Bayes ranking method deals with ranking the most highly differently expressed genes in real data, we obtained RNA-Seq read-counts of lymphoblastoid cell lines for 89 Yoruban and 91 Central European individuals, sequenced as part of the 1000 Genomes project (1000 Genomes Project Consortium *et al.*, 2012; Lappalainen *et al.*, 2013). The counts were processed from the raw RNA-Seq files using the Recount2 pipeline, and downloaded from the Recount2 repository

Table 2. Ranking of significant SNPs according to Bayesian ranking and *P*-values

SNP	log $\hat{OR}$	SE	*P*-value	Rank (*p*)	Rank (Bayes)	Match
rs2076756	0.55	0.07	1.26×10^{-14}	1	4	* *
rs7517847	−0.50	0.07	2.99×10^{-13}	2	6	—
rs2066843	0.51	0.07	7.87×10^{-13}	3	5	—
rs1343151	−0.48	0.07	1.63×10^{-11}	4	7	—
rs11209026	−1.09	0.17	4.59×10^{-11}	5	1	* *
rs10489629	−0.43	0.07	6.79×10^{-11}	6	10	—
rs10889677	0.44	0.07	9.04×10^{-11}	7	8	—
rs2201841	0.43	0.07	3.57×10^{-10}	8	11	—
rs11465804	−0.96	0.15	3.74×10^{-10}	9	2	—
rs11209032	0.42	0.07	8.64×10^{-10}	10	12	—
rs1004819	0.41	0.07	1.50×10^{-9}	11	13	—
rs8054797	−0.67	0.12	1.11×10^{-8}	12	3	—
rs2241880	−0.36	0.07	4.40×10^{-8}	13	14	—
rs5743289	0.44	0.08	4.65×10^{-8}	14	9	—
rs7194886	−0.35	0.07	1.41×10^{-7}	15	15	—

Note: The SNPs that matched rsids with genome-wide significant SNPs in a much larger follow-up study are marked with asterisks.

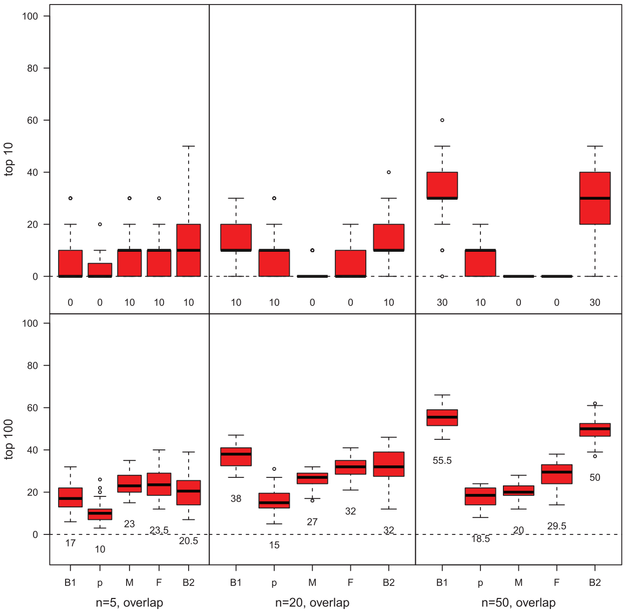

Fig. 3. Each pane of the figure corresponds to a particular Bootstrap sample size ($n = 5$, 20, 50) and top 10 or top 100 overlap when recovering the 'gold standard ranking'. The methods investigated are from right to left *B*1: empirical Bayes algorithm (from Section 3), *p*: *P*-values calculated via DE-Seq2, *M*: non-negative matrix factorization (Jia *et al.*, 2015), *F*: FCROS (Dembélé and Kastner, 2014) and *B*2: empirical Bayes ranking where the prior was estimated using smoothing by roughening

(Collado-Torres *et al.*, 2017). We applied the DESeq2 method (without shrinkage of fold-changes) to estimate fold-changes and standard errors for 58 037 genes in the original sample. Subsequently, we removed 37 597 genes where fewer than 3 samples had read-counts per million sequenced reads exceeding 1, resulting in 20 440 genes remaining. These genes were ranked based on their estimated fold-changes and the resulting ranks were considered as a gold standard in subsequent analyses.

Next, we randomly sampled n ($n = 5$, 20 or 50) individuals from both the Yoruban and the Central European groups. For each random sample of $2 \times n$ individuals, we ran DESeq2 as described before to estimate fold-changes and standard errors for the 20 440 genes under investigation. Subsequently, we employed the same ranking procedures described in Section 4.1 using the associated subsampled RNA-sequence data, fold-changes and standard errors; namely our Empirical Bayes algorithm (from Section 3), *P*-values calculated via DE-Seq2, non-negative matrix factorization (Jia *et al.*, 2015), FCROS (Dembélé and Kastner, 2014) and Empirical Bayes ranking where the prior was estimated using smoothing by roughening (Noma and Matsui, 2013). Both Empirical Bayes methods used 1000 samples from the estimated posterior to estimate mean posterior ranks. This procedure of randomly sampling n Yoruban and Central European individuals from the respective groups, using DESeq2 to calculate fold-changes and standard errors and ranking using the five separate approaches, was repeated independently 40 times for each sceanrio. For each of the 40 simulations, identical overlap statistics to those calculated in Section 4.1 were reported, i.e. the percentage match between the geneids corresponding to the top 10 and 100 estimated ranks and the true ranks that are described in the previous paragraph.

The results, displayed as boxplots in Figure 3, indicate an improvement in the performance of the both Empirical Bayes methods as the sample size (n) increases. This relative improvement with increasing n might be expected as these Empirical Bayes estimates assume the standard error is known, which is only approximately true when n gets large, and rely on accurate estimation of the Empirical Bayes prior, which again is easier for larger n. Overall, the Empirical Bayes methods seem to do best in this example when the sample size is large, but the results are here sensitive to the software that creates the gold standard ranking. In the Supplementary Material, we recreate this figure, but instead using Limma to create the gold standard ranking of fold-change from the original 89 versus 91 comparisons. In this case, non-negative matrix factorization has superior performance, particular at $n = 5$ and $n = 20$. However, we observe a similar pattern of improvement of the Bayesian methods with increasing sample size.

5 Discussion

Bayesian ranking often represents a more appropriate tradeoff between effect size and standard error than do traditional *P*-values when sorting the scientific-importance of differing results in massive multiple hypothesis testing problems. Genome-wide scans for genetic or genomic differences between diseased and healthy individuals (as seen in GWAS and RNA-Seq experiments) are typical examples of such problems. In general the method will work well and be superior to a *P*-value based ranking so long as the standard error of the test statistic varies significantly over differing tests and the proportion of true alternatives (for the algorithm described in this paper, these would be $\{i : |\theta_i| > 0\}$) is large enough. This second condition is necessary so that the learned prior distribution is an effective estimator of the true distribution of parameters over the different tests.

Bayesian Ranking techniques are not a new idea in the statistics literature, an early reference being Laird and Louis (1989), but have only recently be considered as a technique in genomics (Noma *et al.*, 2010). Our work is quite similar to Noma and Matsui (2013) who considered a non-parametric Empirical Bayes ranking method, that implements the smoothing by roughening approach first considered by Shen and Louis (1999), and applied it to microarray data. In this manuscript, we consider a new approach, and extend consideration to other problems in Genomics such as ranking SNPs for disease association and fold-change ranking in RNA-sequence data. In comparison with smoothing by roughening, our approach fared slightly better when ranking the extent of differential expression in RNA-Sequence data. A plausible explanation for this is requires estimating separate parameters for each point on a grid covering the parameter space. In contrast, our approach uses BIC to adaptively choose the number of mixing components necessary to represent the prior. While normal mixture modeling can suffer with identifiability issues, it requires estimating far fewer parameters than smoothing by roughening, and is less likely to give an overfitted estimate of the prior. We implemented smoothing by roughening using the *GaussianSBR* function in the R-package *hhsim*, which uses 200 grid points for the approximation.

In general, we have demonstrated that Bayesian ranking can be helpful in identifying the most important differentially expressed genes (for particular diseases) and SNPs in GWAS to existing procedures. In the case of GWAS, it has been hypothesized that some of the genetic disease heritability, unaccounted for by the accumulation of disease associated SNPs identified by GWAS, is a result of rare variants (Manolio *et al.*, 2009) which tend to go undiscovered due to lower power and the stringent multiple testing corrections. When these variants are included on GWAS chips, they are typically ranked too low to be discovered by conventional methods, although statistics that pool together several rare variants within the same gene to increase power can help with this problem (Ferguson *et al.*, 2013a; Ionita-Laza *et al.*, 2011). Running the algorithm suggested in this manuscript may also help in this regard as it will up-weight the ranking of rare but potentially important SNPs that have large ORs compared to a *P*-value-based ranking. More recently, exon-sequencing and whole genome sequencing studies survey perhaps tens of millions of variant positions in the genome, many of which will be uncommon, heightening the necessity to develop methods that are more flexible in identifying important rare variants. When employing the algorithm to rank SNP locations a couple of points need to be kept in mind. First, we assume that conditional on the true parameter vector $\theta_1, \ldots, \theta_N$, the estimated parameters $\hat{\theta}_1, \ldots, \hat{\theta}_N$ are independent. Technically, SNPs need to be LD-pruned before running the algorithm to approximately satisfy this condition. The algorithm can in theory be modified to run with non-independent effects, provided covariance matrices are known. Second, rankings will likely be more biologically appealing if the input SNP/disease log-ORs are adjusted for other covariates such as age, gender or principle component loadings. Finally, memory and computational issues might prohibit simultaneous genome-wide ranking of all SNPs in GWAS studies. A better strategy is use all SNPs in the Empirical Bayes estimation of the prior, but to pre-select significant SNPs (either based on FDR or FWER rules) before simulation of ranks from the posterior. The actual run time of the algorithm will vary depending on the number of significant SNPs and the number of simulations from the posterior, but using an FDR threshold of 5% and a GWAS of 1 000 000 SNPs, 1000 of which are non-null, one would expect the total run time to be <10 min on a typical desktop PC.

An advantage of Bayesian approaches, in comparison with targeted methods for ranking fold-change parameters such as non-negative matrix factorization and FCROS, is that we can analyze any function $f(\theta_i)$ of the parameter, θ_i, rather than the parameter itself if we so wish. For example, if we sample posterior iterates of θ_i first, we can then transform these iterates using f, and finally rank experiments, $i \leq N$, according to their mean value of $f(\theta_i)$ from the simulations. As a concrete example, in the case of GWAS, population attributable risk (PAR) is a method that can quantify disease impact of a genetic variant on a population level. More specifically, for a given SNP, PAR is the proportion of disease current disease cases that would hypothetically be healthy if everybody had two copies of the major allele at that locus. Provided the disease is rare, PAR is related to the OR, e^{θ_i} and MAF of the SNP (for cases), p_E, in a simple way as $PAR_i \sim 2p_E(1-p_E)e^{\theta_i}/(e^{\theta_i}-1)+ p_E^2 e^{2\theta_i}/(e^{2\theta_i}-1)$, assuming Hardy Weinberg equilibrium in cases (Claus *et al.*, 1996). So, by transforming the simulated true ORs into values into PARs in this way, we can use the procedure to rank PARs.

An R-package, *EBrank*, to run the ranking procedures described in this manuscript can be downloaded from the CRAN repository.

Funding

Dr Ferguson is supported the grant: EIA-2017-017 awarded by the Health Research Board, Ireland.

Conflict of Interest: none declared.

References

1000 Genomes Project Consortium *et al.* (2012) An integrated map of genetic variation from 1, 092 human genomes. *Nature*, **491**, 56.

Amrhein,V. *et al.* (2019) Scientists rise up against statistical significance. *Nature*, **567**, 305–307.

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

Bansal,N.K. *et al.* (1997) On the minimax decision rules in ranking problems. *Stat. Probabil. Lett.*, **34**, 179–186.

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

Berger,J. (1985) *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media, New York.

Bottomly,D. *et al.* (2011) Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PLoS One*, **6**, e17820.

Claus,E.B. *et al.* (1996) The genetic attributable risk of breast and ovarian cancer. *Cancer*, **77**, 2318–2324.

Collado-Torres,L. *et al.* (2017) Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.*, **35**, 319.

Critchlow,D.E. (2012) *Metric Methods for Analyzing Partially Ranked Data*. Vol. **34**. Springer Science & Business Media, Berlin.

Dembélé,D. and Kastner,P. (2014) Fold change rank ordering statistics: a new method for detecting differentially expressed genes. *BMC Bioinformatics*, **15**, 14.

Ferguson,J. *et al.* (2013a) Statistical tests for detecting associations with groups of genetic variants: generalization, evaluation, and implementation. *Eur. J. Hum. Genet.*, **21**, 680–686.

Ferguson,J.P. *et al.* (2013b) Empirical Bayes correction for the Winner's Curse in genetic association studies. *Genet. Epidemiol.*, **37**, 60–68.

Franke,A. *et al.* (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.*, **42**, 1118–1125.

Frazee,A.C. *et al.* (2011) ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, **12**, 449.

Hardcastle,T.J. (2015) Generalized empirical Bayesian methods for discovery of differential data in high-throughput biology. *Bioinformatics*, **32**, 195–202.

Henderson,N.C. and Newton,M.A. (2016) Making the cut: improved ranking and selection for large-scale inference. *J. R. Stat. Soc. Series B Stat. Methodol.*, **78**, 781–804.

Ionita-Laza,I. *et al.* (2011) A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet.*, **7**, e1001289.

Jewett,P.I. *et al.* (2018) Optimal Bayesian point estimates and credible intervals for ranking with application to county health indices. *Stat. Methods Med. Res.* doi: 10.1177/0962280218790104.

Jia,Z. *et al.* (2015) Gene ranking of RNA-seq data via discriminant non-negative matrix factorization. *PLoS One*, **10**, e0137782.

Laird,N.M. and Louis,T.A. (1989) Empirical Bayes ranking methods. *J. Educ. Behav. Stat.*, **14**, 29–46.

Lappalainen,T. *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506.

Lin,R. *et al.* (2006) Loss function based ranking in two-stage, hierarchical models. *Bayesian Anal.*, **1**, 915.

Mailman,M.D. *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.*, **39**, 1181–1186.

Manolio,T.A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods*, **5**, 621–628.

Muralidharan,O. *et al.* (2010) An empirical Bayes mixture method for effect size and false discovery rate estimation. *Ann. Appl. Stat.*, **4**, 422–438.

Nickerson,R.S. (2000) Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol. Methods*, **5**, 241.

Noma,H. and Matsui,S. (2013) Empirical Bayes ranking and selection methods via semiparametric hierarchical mixture models in microarray studies. *Stat. Med.*, **32**, 1904–1916.

Noma,H. *et al.* (2010) Bayesian ranking and selection methods using hierarchical mixture models in microarray studies. *Biostatistics*, **11**, 281–289.

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

Shen,W. and Louis,T.A. (1998) Triple-goal estimates in two-stage hierarchical models. *J. R. Stat. Soc. Series B Stat. Methodol.*, **60**, 455–471.

Shen,W. and Louis,T.A. (1999) Empirical Bayes estimation via the smoothing by roughening approach. *J. Comput. Graph. Stat.*, **8**, 800–823.