7-1-2010

A DECISION-THEORY APPROACH TO INTERPRETABLE SET ANALYSIS FOR HIGH-DIMENSIONAL DATA

Simina Maria Boca
Johns Hopkins University Bloomberg School of Public Health, sboca@jhsph.edu

Hector C. Bravo
University of Maryland, College Park, MD

Brian Caffo
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Jeffrey T. Leek
Johns Hopkins University, Bloomberg School of Public Health

Giovanni Parmigiani
Dana Farber Cancer Institute, Harvard School of Public Health

Abstract

A ubiquitous problem in high-dimensional analysis is the identification of pre-defined sets that are enriched for features showing an association of interest. In this situation, inference is performed on sets, not individual features. We propose an approach which focuses on estimating the fraction of non-null features in a set. We search for unions of disjoint sets (atoms), using as the loss function a weighted average of the number of false and missed discoveries. We prove that the solution is equivalent to thresholding the atomic false discovery rate and that our approach results in a more interpretable set analysis.

1 Introduction

Many scientific studies measure a large number of variables on each sample. These variables are usually measurements of certain physical properties, or "features." In genomics, the features may be genes and the actual variables gene expression measurements; tens or hundreds of thousands of gene expression measurements are taken on only a few subjects. Similarly, in brain imaging studies, tens of thousands of voxel intensities are measured on a small number of study participants. In spatial epidemiology, estimates of disease prevalence are sometimes calculated for a large number of locations, with a small number of individuals in each location. Often, in these cases, it is the sets of variables that are of interest. Gene-set analysis, i.e. the analysis of genomic data using set annotations related to biological processes or pathways, is a particularly important example. Early examples of gene-set analysis include Tavazoie et al. (1999), Mirnics et al. (2000), and Bouton and Pevsner (2002). Set analyses of high-dimensional data have been developed to combine information across features, to infer associations between sets and outcomes or phenotypes. Sets are defined by previous experimentally verified or postulated relationships between features. Usually, a fixed number of sets K is defined in advance. Each of the sets $\mathscr{S} = \{S_1, S_2, \ldots, S_K\}$ is a subset of the collection of features $\mathscr{M} = \{1, 2, \ldots, M\}$. Some sets are based on the spatial or geographic arrangement of features and are naturally disjoint: For instance, a voxel may get mapped to exactly one brain region, and a geographical location will get mapped to exactly one county, but in general features may belong to more than one set. For example, one gene may participate in more than one metabolic pathway.

Set-level inference can be useful in a number of ways. It can increase statistical power, by helping detect correlated changes in features within the same set, which are too subtle to be detected in a marginal analysis (Mootha et al., 2003).

This can be due to coregulation of gene expression within pathways, voxel intensities within brain regions, or disease prevalences in locations within broader geographic regions, such as counties. Another possibility is that the signal is at the level of the sets rather than the features. For example, the same pathway may be altered in different tumors, but the specific genes which are altered may differ between tumors (Parsons et al. (2008)). Inferences at the level of sets are also sought to increase interpretability: For instance, hundreds of genes may be up-regulated in a specific phase of cell division in bacteria, though most of those genes may belong to a small number of known pathways. Knowing that only a few pathways are involved is much more informative than knowing that hundreds of genes show varying levels of expression.

Here we propose a general decision-theoretic approach for set-level analysis which focuses on estimating the fraction of non-null features in a set, rather than using p-values. We focus on the expected number of false and missed discoveries (EFD and EMD), as described in Section 3. Our units of inference are non-overlapping sets, which we call atoms. The EFD and EMD are directly interpretable as measures of enrichment, our approach generally being equivalent to considering the fraction of features within a set that are from the alternative distribution. Our approach combines these key ideas into a decision theory framework where we derive estimators consisting of unions of atoms that minimize a weighted combination of the EFD and EMD. These estimators are functions of the joint posterior distribution of feature-specific parameters and do not require a separate statistical treatment of feature-level and set-level analysis.

The results we present here can be summarized as follows.

1. Section 3 introduces the expected number of false discoveries and the expected number of missed discoveries in a decision-theory framework. In particular, in Theorem 1 we show that the two components of the main loss function we consider, which represent the posterior expected number of false discoveries and the posterior expected number of missed discoveries, may be written in terms of only the marginal feature-level posterior probabilities.
2. Section 4 provides alternative loss functions which may be used, including loss functions with regularization penalties and a loss function which uses the ratio of false discoveries and missed discoveries.
3. Section 5 introduces the concept of atomic FDR and provides algorithms for finding the Bayes estimators for the loss functions described above. For most of the loss functions considered, an analytic closed form is provided. An interpretation of the atomic FDR is also provided.
4. Section 6 outlines an empirical Bayes approach for estimating the feature-level posterior probabilities and applies this approach to simulation and real

gene expression analyses. It also discusses use of the bootstrap in obtaining standard errors for the estimates.

2 Relationship to previous approaches

One common approach to high-dimensional data analysis is to perform inference on a set of features indexed by $\mathscr{M} = \{1, 2, \ldots, M\}$ marginally, one at a time. The data for the mth feature is a $N \times 1$ vector X_m, which is compared to a single $N \times 1$ outcome vector Y. Feature level statistics are functions of the data X_m and the outcome Y. The null hypothesis for a given feature is that it is not strongly associated with the outcome of interest. Thus, each feature can be seen as coming either from the null or an alternative distribution, leading to the commonly used mixture model (Efron et al., 2001; Storey, 2002; Newton et al., 2004):

$$f(x_m|y) = \pi_0 f_0(x_m|y) + (1 - \pi_0) f_1(x_m|y), \tag{1}$$

where f indicates the density for a feature, π_0 indicates the prior probability that a feature is from the null distribution, f_0 indicates the density of the null distribution and f_1 indicates the density of the alternative distribution.

A standard approach for feature-level analysis is to then perform a statistical test for each feature to assess the level of association with the outcome of interest. For example, in genomics a test may be used to decide whether a gene is differentially expressed between two conditions. This can be a t-test or a F-test, or some variation on them (Storey and Tibshirani, 2003a; Baldi and Long, 2001; Cui et al., 2005). Due to the large number of tests performed, a multiple comparison adjustment is needed. The goal is often to control the false discovery rate (FDR) (as in Benjamini and Hochberg (1995), Efron and Tibshirani (2002), Storey (2002), Storey and Tibshirani (2003b), Storey (2003)). Alternatively, the posterior probabilities that features are from the null or alternative distributions, or from mixture components of interest, can be estimated using Bayes (Parmigiani et al., 2002; Do et al., 2005) or Empirical Bayes methods (Efron et al., 2001; Newton et al., 2001; Lönnstedt and Speed, 2002; Efron and Tibshirani, 2002; Newton and Kendziorski, 2003; Gottardo et al., 2003; Newton et al., 2004). These methods generally borrow information across genes, but the inference is still performed one feature at a time (marginally), and no information from outside the experiment is considered.

Most set-level inference methods combine the feature-level statistics into set-level statistics, then perform some hypothesis test for each set. In genomics, one approach is to individually score genes, then separate them into categories based on these scores; the simplest scenario is to choose a single cutoff and declare the genes above the cutoff "differentially expressed" and the genes below the cutoff

"non-differentially expressed." For each set a contingency table is then constructed, which cross-classifies the genes by whether or not they are in the set and by their category and a classical statistical test is performed, such as the hypergeometric test (Tavazoie et al., 1999). Alternatively, a test can be performed to compare the distributions of statistics of the genes within a set to those of all the genes (Mirnics et al., 2000). A method with a somewhat different flavor, *gene-set enrichment analysis* (GSEA), is described in Subramanian et al. (2005) (a preliminary version is employed in Mootha et al. (2003)). For each gene-set, it computes an enrichment score by using a signed variation of the Kolmogorov-Smirnov statistic on a list of ordered gene scores, then calculates a p-value by permuting the phenotype labels. The Wilcoxon test can also be used for performing gene-set analysis, and is one of the standard tools in the `limma` package in `R`, being accessed through the `geneSetTest` function (Michaud et al., 2008)

Our approach focuses on estimating the fraction of features from the alternative distribution using the EFD and EMD (section 3), in contrast to the p-value-based methods described above. Heller et al. (2009) propose estimating a lower bound on the fraction of alternatives in a set by considering the ratio of the number of rejected hypotheses in the set to the total number of hypotheses in the set, using an adjusted p-value threshold which controls for multiple testing. We note that this appears to be a very conservative estimate, as the features which are not rejected do not contribute at all to the estimate, whereas our procedure uses empirical Bayes estimates of the posterior probabilities of features being from the alternative distribution (section 6). Research has also been conducted on using set-level information to improve inference at the feature level, for example, by performing feature-level hypothesis tests on grouped data, as in Cai and Sun (2009). The objective of such a method is different from our present purpose, which is to perform inference at the set level.

Despite the appeal of set level analyses, some difficulties with set analysis and interpretation remain. (1) P-values are not interpretable on an enrichment scale, since they are not estimates of the fractions of false and true discoveries. In particular, p-values have different interpretations for different sample sizes, and generally set annotations cannot be expected to produce sets which all consists of the same number of features. This can lead to sets with very different fractions of alternatives having similar p-values. (2) Another problem results from the common use of "competitive tests" (Goeman and Buhlmann (2007)) in many set-level inference procedures, which pit a set against its complement. This results in the "zero sum problem," where more significant features in one set often lead to higher p-values in other sets. (3) Lastly, most approaches perform separate analyses at the feature and set level, so uncertainty remaining from the feature-level analysis is ignored at the set level.

To expand on the first point, the majority, of methods for set-level inference rely on calculating a p-value for each set. However, p-values do not actually give a direct indication of how much signal is present in a given set. Several common methods for set-level inference in genomics which rely on calculating p-values are reviewed and critiqued in the influential paper of Goeman and Buhlmann (2007). We revisit some of their points here, namely the issues of gene sampling and competitive tests, and show how our approach addresses these problems. Gene sampling methods use permutations of the gene, as opposed to the sample, labels to obtain a p-value. In particular, this means that the sample size is given by the number of genes, as opposed to the number of subjects. One of the practical problems which results from this approach is that p-values have different interpretations for different sample sizes, with larger sample sizes often leading to smaller p-values, even if there is less signal present. Thus, two sets of different sizes with different fractions of alternatives may end up with very similar p-values. In contrast, our method does not suffer from this problem, as the posterior probabilities that features are from the alternative distribution are averaged within atoms (Theorem 4) after being estimated using the data on all the subjects (section 6).

On the other hand, competitive tests consider the null hypothesis that features in a given set are at most as often differentially expressed as features in the complement of the set. This results in a "zero sum problem," where a larger fraction of alternatives in one set may lead to higher p-values in other sets. Methods based on the Fisher exact test, the chi-squared test, the Wilcoxon rank test, and the z-test fall into this class. Consider, for example, the case where the signal in a dataset is represented by 10% of the features. Then take sets consisting of 10, 50, and 100 features, none of which have any signal. The larger sets will then have larger p-values, because as the size of the set increases, the size of its complement will decrease, and therefore the proportion of signal in the complement will increase. Goeman and Buhlmann (2007) suggest a different null hypothesis, namely that no genes in a gene-set represent true signal. However, this does not appear to be the null hypothesis that most scientists would be interested in. We consider that it is more informative, in the case of set-level inference, to move away from a hypothesis framework to an estimation framework and directly estimate the quantity of interest: the fraction of false discoveries or the enrichment of a set.

3 Expected False Discoveries and Expected Missed Discoveries

Let the set of features from the alternative distribution be denoted by $\tau \subset \mathscr{M}$. For example, in the case where the features are gene expression measurements, τ represents the genes which are differentially expressed between the two conditions. If we wanted to estimate τ based on only the feature-level information, we would simply take the estimate to be the set of all features whose statistics in X are within certain limits. For instance, if we performed a two-sample t-test for a differential gene expression problem, we could estimate τ by the set of all genes with p-values below 0.05.

We focus on the case where sets are non-overlapping. For a small number of sets, these can be obtained by breaking down sets into their largest non-overlapping components, called *atoms*, as described in the Appendix. An alternative method, which can be used for a large number of sets (for instance, when considering all the sets in GO or KEGG), involves clustering the features using a dissimilarity measure which depends on the sets in $\mathscr{S}$ which are shared between features (Boca et al., in preparation). We note that in many cases, set annotations are naturally non-overlapping, for instance, when they represent areas of the brain in imagining studies, individual counties in spatial epidemiology, or areas of protein localization in proteomics. Each atom thus consists of a group of features, which may be either from the null or alternative distributions, as seen from Equation (1). A hypothesis testing approach, such as the one taken in Reiner-Benaim et al. (2007), would perform hypothesis tests within each atom, thus considering the atoms to be disjoint families of related hypotheses. Benjamini and Heller (2008) inspect the case where the same features are considered at different "map locations" and hypothesis tests are performed to see whether the fraction of discoveries is above some fixed threshold.

Estimating τ based on just the feature-level data would ignore the scientific background which is represented by the atom-level information. Thus, we seek to estimate τ using the unions of atoms in $\mathscr{A}$. We denote by $\mathscr{U}$ the set of all possible unions of atoms in $\mathscr{A}$. We look for the set in $\mathscr{U}$ which maximizes the overlap with τ by using a relevant loss function. Thus, we are in the situation of providing a scientifically meaningful estimate of τ by solving a constrained estimation problem where we only have elements in $\mathscr{U}$ as possible estimators.

Our method can be placed in the general class of decision-theoretic approaches described in the Appendix. In particular, we consider the scenario where the discrepancy d between features does not take into account how far or close the features are to each other. As shown in Lemma 2 in the Appendix, this is in fact

the only discrepancy measure which allows for a tractable analysis. In this case, the discrepancy is equivalent to the following $0-1$ function which takes as inputs two features m_1 and m_2:

$$d(m_1, m_2) = 1(m_1 \neq m_2). \tag{2}$$

We use the single-linkage (nearest neighbor) function to define the discrepancy between a feature and a set: $d(m, A) = \min\{d(m, m_0) : m_0 \in A\}$.

This is equivalent to saying that the discrepancy between two features is 0 if and only if the features are one and the same, otherwise it is 1. We note that in this case the loss function is reduced to:

$$\begin{aligned} L(\tau, U) &= (1-w) \sum_{m \in U \setminus \tau} d(m, \tau) + w \sum_{m \in \tau \setminus U} d(m, U) \\ &= (1-w) * |U \setminus \tau| + w * |\tau \setminus U| \\ &= (1-w) * \text{Number of false discoveries } + w * \text{ Number of missed discoveries,} \end{aligned}$$

where a false discovery is a feature which is not in τ but is in U and a missed discovery is a feature which is in U but is not in τ. This loss function is similar to the loss function of Cai and Sun (2009), but, while we are performing set-level inference, they are using information on sets of genes to improve gene-level inference. When considering the posterior expected loss function, the two components become the expected number of false discoveries and the expected number of missed discoveries, which we will denote by $\text{EFD}(U)$ and $\text{EMD}(U)$:

$$\begin{aligned} \mathscr{L}(U) &= \sum_{\tau \in 2^{\mathscr{M}}} L(\tau, U) * P(\tau | X, Y) \\ &= (1-w)\text{EFD}(U) + w\text{EMD}(U). \end{aligned}$$

where:

$$\begin{aligned} \text{EFD}(U) &= E_{\tau|X,Y}\{\sum_{m \in U \setminus \tau} d(m, \tau)\} = \\ &= \sum_{\tau \in 2^{\mathscr{M}}} \sum_{m \in U \setminus \tau} p_\tau \\ \text{EMD}(U) &= E_{\tau|X,Y}\{\sum_{m \in \tau \setminus U} d(m, U)\} \\ &= \sum_{\tau \in 2^{\mathscr{M}}} \sum_{m \in \tau \setminus U} p_\tau \end{aligned}$$

Similar loss functions have been used by Genovese and Wasserman (2002), Storey (2003), and Müller et al. (2004), though not in set-level inference.

We use the following notation for the posterior probabilities:

$$p_\tau = P(\tau|X,Y).$$

We also introduce the concept of *marginal feature-level posterior probabilities*:

$$p_U^* = \sum_{\tau \in 2^{\mathscr{M}}, U \subset \tau} p_\tau,$$

which represents the posterior probability that the set U is included in the set of all features from the alternative distribution. For a specific feature m, the *marginal feature-level posterior probability* is denoted:

$$p_m^* = \sum_{m \in \tau} p_\tau,$$

In Theorem 1, we simplify the form of both EFD(U) and EMD(U). In particular, both can be written as affine functions of the marginal feature-level posterior probabilities. This yields major benefits in terms of both modelling and computation.

Theorem 1. *Under the loss function described above and the single linkage property, EFD(U) and EMD(U) can both be written as affine functions of the marginal feature-level posterior probabilities. EFD(U) can be written as the sum of posterior probabilities that the features in U are from the null distribution and EMD(U) can be written as the sum of the posterior probabilities that the features which are not in U are from the alternative distribution:*

$$\begin{aligned} EFD(U) &= \sum_{m \in U} (1 - p_m^*) = |U| - \sum_{m \in U} p_m^* \\ EMD(U) &= \sum_{m \notin U} p_m^* \end{aligned}$$

In Corollary 2, we show that we can also write EFD(U) and EMD(U) in terms of the local false discovery rate (fdr) defined in Efron and Tibshirani (2002).

Corollary 2. *When considered as functions of the data, EFD(U) and EMD(U) can be written in terms of local false discovery rates:*

$$\begin{aligned} EFD(U) &= \sum_{m \in U} fdr(X_m, Y_m) \\ EMD(U) &= \sum_{m \notin U} [1 - fdr(X_m, Y_m)] \end{aligned}$$

4 Alternative loss functions

We consider some variations on the loss function corresponding to the $0-1$ discrepancy measure which was introduced earlier, by looking at what happens when linear constraints are introduced. In order to penalize for a large number of features in the Bayes estimator, we consider the following constrained optimization problem:
Minimize $L(\tau,U)$ with the constraint that $|U| < \rho$, for some $\rho > 0$. This is equivalent to minimizing the loss function:

$$L_f^{\lambda}(\tau,U) \quad = \quad (1-w)*|U \setminus \tau| + w*|\tau \setminus U| + \lambda|U| \text{ for some } \lambda > 0$$

(see for instance Gill et al. (1981)).

Similarly, one may be interested in penalizing for a large number of atoms in the Bayes estimator. For example, in a spatial epidemiology example, one may desire to target an intervention to a fixed number of locations due to issues of cost or work force. In such a case, we would consider the following constrained optimization problem:
Minimize $L(\tau,U)$ with the constraint that $|J| < \eta$, for some $\eta > 0$, where J is the number of atoms in U. This is equivalent to minimizing the loss function:

$$L_a^{\xi}(\tau,U) \quad = \quad (1-w)*|U \setminus \tau| + w*|\tau \setminus U| + \xi|J| \text{ for some } \xi > 0.$$

We remark that we could also penalize unions of atoms with a small number of genes, respectively atoms, by simply changing the signs of λ and ξ. We could also have a loss function which incorporates penalties on both the number of features and the number of atoms.

Alternatively, loss functions which employ the ratio of missed discoveries and false discoveries, instead of the number of missed discoveries and false discoveries, may also be used. Thus, we can consider:

$$\begin{aligned} L_r(\tau,U) &= (1-w)*\frac{|U \setminus \tau|}{|U|} + w*\frac{|\tau \setminus U|}{M-|U|} \\ &= (1-w)*\text{Ratio of false discoveries } + w* \text{ Ratio of missed discoveries} \end{aligned}$$

5 Atomic FDR

5.1 Definition and results

In general, finding the Bayes estimator for the loss functions based on the $0-1$ discrepancy measure can be very computationally challenging, even after the sim-

plification in Theorem 1, since the number of sets in $\mathscr{U}$ which the posterior expected loss needs to be minimized over is 2^L, where L is the total number of atoms. We find an analytic solution for obtaining the Bayes estimator for the loss function which weights the number of false discoveries and missed discoveries, as well as the loss functions with regularization penalties, described in Section 4. This result, established in Theorem 4, shows that the Bayes estimator for the loss L is found by choosing those atoms $A_l \in \mathscr{A}$ with EFD less than or equal to w. This EFD can be thought of as a (Müller et al. (2007)) or Bayesian (Sarkar and Zhou (2008)) false discovery rate for atom A_l, which we denote by Afdr_l, and call the *atomic false discovery rate* (atomic FDR):

$$\text{Afdr}_l = \frac{\text{EFD}(A_l)}{n_l}, \tag{3}$$

where $\text{EFD}(A_l)$ is the expected number of false discoveries in atom A_l and $n_l = |A_l|$, i.e. the number of features in atom l. By Corollary 2, the atomic false discovery rate can be written as the average local false discovery rate of all the features in the atom. Thus, the algorithm for finding the Bayes estimator corresponds to thresholding the atomic FDR at a fixed level determined by the parameter w. For large values of w the procedure allows more false positives, since the EFD is down-weighted in the loss function, while small values of w more strongly weight the EFD and restrict the resulting atomic false discovery rate. In the case where we penalize atoms with many genes, considering L_f^λ, the Bayes estimator is equivalent to thresholding the realized atomic FDR adjusted for a "background rate" of false discoveries λ.

We first prove a result which gives a convenient parametrization of the posterior expected loss using the original loss function:

Lemma 3. *The posterior expected loss $\mathscr{L}(U)$ which results from the $0-1$ dissimilarity measure may be rewritten as:*

$$\delta'\{\mathbf{EFD}_A - w\mathbf{n}\} + w\mathbf{1}'(\mathbf{n} - \mathbf{EFD}_A)$$

where δ, $\mathbf{EFD}_A$, $\mathbf{n}$, and $\mathbf{1}$ are vectors of length $L = |\mathscr{A}|$, with the elements of the first three vectors denoting the indicators of atoms being in the estimator, the expected number of false discoveries per atom, respectively the total number of features in an atom:

$$\begin{aligned}
\delta &= (\delta_1, \delta_2, \ldots, \delta_L)' \\
\mathbf{EFD}_A &= (EFD(A_1), \ldots, EFD(A_L))' \\
\mathbf{n} &= (n_1, n_2, \ldots, n_L)' \\
\mathbf{1} &= (1, 1, \ldots, 1)'
\end{aligned}$$

where δ_l is the indicator of whether atom A_l is part of U.

The posterior expected loss functions $\mathscr{L}_f^\lambda$ and $\mathscr{L}_a^\xi$ which correspond to the loss functions $L_f^\lambda(\tau,U)$ and $L_a^\xi(\tau,U)$ (from Section 4) have similar parametrizations, which are also linear in δ:

$$\mathscr{L}_f^\lambda(\mathbf{t}) = \delta'\{\mathbf{EFD}_A - w\mathbf{n}\} + w\mathbf{1}'(\mathbf{n}-\mathbf{EFD}_A) + \lambda\delta'\mathbf{n} \quad (4)$$

$$\mathscr{L}_a^\xi(\mathbf{t}) = \delta'\{\mathbf{EFD}_A - w\mathbf{n}\} + w\mathbf{1}'(\mathbf{n}-\mathbf{EFD}_A) + \xi\mathbf{1}'\mathbf{n} \quad (5)$$

We provide a straightforward algorithm for obtaining the Bayes estimator for the loss functions L, L_f^λ, and L_a^ξ, for a fixed w between 0 and 1 in Theorem 4.

Theorem 4. *For a fixed value of $w \in [0,1]$, analytic solutions are obtained for the Bayes estimators for the losses L, L_f^λ, and L_a^ξ:*
The indicator δ_l of whether atom A_l is in the Bayes estimator for the loss L is:

$$\delta_l = 1\{Afdr_l \leq w\}$$

The indicator δ_l of whether atom A_l is in the Bayes estimator for the loss L_f^λ is:

$$\delta_l = 1\{Afdr_l + \lambda \leq w\}$$

The indicator δ_l of whether atom A_l is in the Bayes estimator for the loss L_a^ξ is:

$$\delta_l = 1\{Afdr_l + \frac{\xi}{n_l} \leq w\}$$

We now prove a result concerning the loss function L_r from Section 4, which considers the ratio of false discoveries and missed discoveries.

Theorem 5. *The posterior expected loss $\mathscr{L}_r(U)$ which results from the loss function L_r may be rewritten as:*

$$\delta'\left\{\frac{(1-w)}{\delta'\mathbf{n}}\mathbf{EFD}_A - \frac{w}{(M-\delta'\mathbf{n})}(\mathbf{n}-\mathbf{EFD}_A)\right\} + \frac{w}{M-\delta'\mathbf{n}}(\mathbf{n}-\mathbf{EFD}_A)$$

An analytic solution is not available in this case. However, an approximate algorithmic solution is presented below in Algorithm 1, with further details available in the Appendix.

5.2 Interpretation

The interpretation of the atomic FDR (Afdr) is the expected fraction of null features in an atom. Theorem 4 helps us find the Bayes estimator associated with loss function L: it is sufficient to consider atoms which have Afdr below a threshold w. For the loss functions L_f^λ and L_a^ξ, which penalize estimates for including large numbers of features, respectively atoms, the atoms included in the estimator are also characterized solely by the fact that Afdr is below a certain threshold: respectively $w-\lambda$ and $w-\frac{\xi}{n_l}$ for the two loss functions, where n_l is the number of features in the atom. The Afdr also has a direct interpretation in terms of enrichment when it is compared to the average posterior probability that a feature is from the null distribution. Depending on the specific application, not all features may be annotated to a set. For example, not all genes participate in metabolic pathways. In this case, a decision needs to be made whether the comparison should be to the average posterior probability over all features or just over the features annotated to sets. When considering applied examples, it may be more intuitive to look at $1-\text{Afdr}$, which is the expected fraction of alternative features in an atom (or the expected "true discovery rate"). Therefore an estimate $\widehat{\text{Afdr}}$ of Afdr is also an estimate of the fraction of null features in an atom, while an estimate $1-\widehat{\text{Afdr}}$ of $1-\text{Afdr}$ is also an estimate of the fraction of alternative features in an atom.

6 Applications

6.1 Empirical Bayes estimates of posterior probabilities

We initially perform feature-level statistical inference, then translate it to the set level. The feature level statistics represent information from the present study or studies. We combine this information with prior knowledge, represented by annotations placing the features into sets. We only consider non-overlapping sets for set-level inference. If the sets we start out with are overlapping, we use the annotations to obtain non-overlapping sets, which we call atoms (see section 3 and Appendix). We then calculate the expected missed discoveries and false discoveries for every atom, and use this to obtain a Bayes estimator corresponding to a particular loss function. This workflow is summarized in Figure 1.

In our illustrations, we estimated the feature-level posterior probabilities using an Empirical Bayes approach, following the lead of Efron and Tibshirani (2002)

and Newton and Kendziorski (2003) to estimate p_m^* using the form in Remark 2:

$$\widehat{p_m^*} = 1 - \widehat{\text{fdr}(X_m|Y)} = 1 - \hat{\pi}_0 \frac{\widehat{f_0(X_m|Y)}}{f(X_m|Y)}. \tag{6}$$

In particular, we use the nonparametric approach detailed in Storey et al. (2005). The steps are detailed in Algorithm 2. We consider 20 simulations under the null to be sufficient and we take the number of equally spaced knots to be equal to the total number of features.

6.2 Simulations

We carried out simulations which compared our method to a representative competitive testing method which relies on the Wilcoxon rank test and is available in the `limma` package in *R* (accessed through an interface developed for Schaeffer et al. (2008)). We calculated p-values for the `limma` method, as well as q-values, which are adjusted p-values to control the FDR for independent hypothesis tests (Benjamini and Hochberg, 1995). In each simulation, the features which are from the alternative distribution are draws from a normal distribution with mean 1 and variance $1/15$, while remaining features are draws from a normal distribution with mean 0 and variance $1/15$. For each atom l, we also provide the estimate of $1 - \text{Afdr}_l$, $1 - \widehat{\text{Afdr}}_l$, for our method, where Afdr_l is the atomic FDR, as described in Equation (3). $1 - \text{Afdr}_l$ can thus be interpreted as the expected"true discovery rate." An atom is included in the Bayes estimator if $1 - \widehat{\text{Afdr}}_l \geq 1 - w$, for a fixed weight w, as seen from Theorem 4.

We first consider an example where two sets (1 and 2) having 50 and 100 genes have an overlap of 20 genes and 60%, respectively 40%, differentially expressed genes. In scenario (A), there are no differentially expressed genes common to sets 1 and 2, whereas in scenario (B), all of the genes common to sets 1 and 2 are differentially expressed. This example is shown in Figure 1 in the Appendix. We performed 100 simulations with these four sets (1A, 1B, 2A, 2B). We give the results from the `limma` method in Table 1. We note that, since Set 1A and Set 2A and Set 1B and Set 2B are overlapping, the results from this test are difficult to interpret, since it is unclear where the difference in p-values comes from. In fact, Set 1A and Set 2A have no features in common which are from the alternative distribution, while Set 1B and Set 2B have 20 such features in common. In Table 2 we present the results which use our method and are applied on the atoms we obtained from simply taking the intersections and differences of the original sets. The p-values for the individual atoms presented in Table 1 are much more interpretable, and the estimates of $1 - \text{Afdr}$ even more so.

We performed another set of 100 simulations, with 2500 features, 5% of which were from the alternative distribution. These 125 features were distributed among 5 atoms of size 50, which had different fractions of alternatives, of 0.9, 0.7, 0.5, 0.3, and 0.1. We also considered an atom of size 50 with no features from the alternative distribution, for comparison. The remaining features were not placed in any atoms. The mean posterior probability over the 100 simulation runs was estimated to be 0.068, with a standard deviation of 0.021, which is quite close to the true value of 0.05. In Table 3 we present the results which compare our method to the `limma` method. We note that our method provides much more interpretable results. The `limma` method gives an average q-value of 0.011 for a fraction of alternatives of 0.3, with a standard deviation of 0.03, which means that for many of the simulation runs this set would not be considered significant. This is despite the fact that a fraction of alternatives 0.3 is much higher than the background fraction of 0.05. This effect is even more pronounced for a fraction of alternatives of 0.1, where the mean q-value is 0.407.

Boxplots of $1 - \widehat{\text{Afdr}}$ and the q-values versus the true fractions of alternatives for the simulations presented in Table 3 are shown in Figure 2. In the ideal scenario, the mean value of $1 - \widehat{\text{Afdr}}$ would be close to the true fraction of alternatives. Given that the posterior probabilities are estimated to be between 0 and 1, we expect a slight anti-conservative bias for the sets with low fractions of alternatives and a conservative bias for the sets with high fractions of alternatives, which is what we see in the plot in the left panel. The plot in the right panel shows that the p-values have a very wide spread for the low fractions of alternatives, but there is nearly no spread for the higher fractions of alternatives, highlighting the difference in interpretation between estimated fraction of alternatives and significance tests.

We also explored the effect of the atom size on both our method and the `limma` method. We considered 2500 features again, 5% of which are from the alternative distribution. We took 3 atoms, each having a fraction of alternatives of 0.5, but different sizes: 10, 50, and 100 features. The results are displayed in Table 4. For our method, the mean estimate over the 100 runs is similar across the different set sizes. As expected, for the `limma` method, the p-values and q-values for the atom of size 10 is much higher than for the other two atoms. As noted in Section 2 and confirmed in this simulation, the p-values in set inference methods which employ feature sampling have different interpretations for different set sizes. Thus, the results from the `limma` method are not particularly interpretable in this case. Our decision theoretic, estimation-oriented framework provides a much clearer interpretation.

We also considered the impact of the atom size for atoms which have only null features. Once again, we used 2500 features total, 5% of which were from the

alternative distribution. As in the previous example, we considered atoms of sizes 10, 50, and 100. Whereas our method gives an estimated mean $\widehat{\text{Afdr}}$ between 0.961 and 0.985 for each of the three atoms, with standard errors smaller than 0.04, the `limma` metho results in p-values which are increasingly skewed towards 1 as the atom size increases. This is due to its feature-sampling and competitive, as noted in Section 2: for atoms with no features from the alternative, as the atom size increases, the fraction of features from the alternative distribution in the complement of the atom increases. In Figure 3, we show histograms exhibiting this behavior for the different sets, over 100 runs.

We also compared the Bayes estimators resulting from the loss function L which weights the number of false discoveries and missed discoveries to those resulting from the other loss functions we introduced in Section 4 , namely L_f^λ, L_a^ξ, and L_r. We compared the results on 100 simulations with 2250 features, 10% of which were from the alternative distribution. We considered 8 atoms, 4 of size 50 and 4 of size 100. For each of the two sizes considered, atoms had fractions of alternatives of 0, 0.1, 0.5, or 0.9. The results are presented in Table 5. We used $\lambda = 0.20$ and $\xi = 5$. The most frequently selected Bayes estimators from the 100 runs were compared to the ideal scenario, where the features from the alternative and null distributions are given posterior probabilities of 1, respectively 0. We note that the interpretation of w for the loss function L_r is different from that for the other three loss functions. In general, the estimators which were commonly chosen in the simulation runs were subsets of the one in the ideal scenario.

6.3 Gene-set data analysis

We perform two data analyses on genomic data, using standard gene-sets which represent chromosomal regions and KEGG sets. We first present an analysis of a dataset from Subramanian et al. (2005), which compares mRNA expression profiles from lymphoblastoid cell lines of 15 males and 17 females. This data was originally analyzed via the GSEA method from Subramanian et al. (2005), and was later analyzed with a different method, which used t-tests, in Irizarry et al. (2009). The gene-sets used represented chromosomal regions. For illustration, we excluded 40 of the original 212 sets, in order to obtain nonoverlapping atoms. We compared the methods in Subramanian et al. (2005) and Irizarry et al. (2009), as well as the `limma` method, to our method. The results from the top ten sets using our method are presented in Table 6.

The overall estimated expected fraction of false discoveries is 0.976. The set which has the lowest $\widehat{\text{Afdr}}$ with our method is the only set which had a chromosomal

region on the Y chromosome, and it also ranked first in terms of p-values and q-values with the other two methods. Given that the primary difference between the two group is gender, it can serve as a proof of principle. We note that our method is much more interpretable than methods which rely on p-values or q-values: While the estimate of the fraction of alternatives in the set chrYq11 is substantially higher than both the overall estimate and the estimate for the set ranked second, it might not be considered extremely high in absolute terms. Therefore, providing the actual estimate and allowing direct comparisons appears to be much more useful than trying to understand the difference between a q-value of nearly 0 and a q-value of over 0.99.

We further analyzed a dataset from Sotiriou et al. (2006), consisting of expression microarrays from breast tumors. We looked at a subset of untreated tumors, considering the differential expression between ER-positive and ER-negative samples. 10 of the samples were ER-negative, 53 were ER-positive. Using the KEGG annotations for 10 pathways, which resulted in 22 atoms, we note that the strongest signal is found in the atom represented by one of the three-way intersections. Results are shown in Table 7.

6.4 Brain ROI data analysis

We also perform an analysis involving brain imaging data arising from functional magnetic resonance imaging (fMRI), where each set is a region of interest (ROI). A common problem in fMRI is the analysis of so-called group contrast data (Friston et al., 2007). Here, parameters from a first-stage subject-specific regression analysis are compared across subjects in standardized space. Each data point is then a map of regression coefficients, conceptually representing a subject's blood oxygen level dependent (BOLD) response to an experimental stimulus. In this case, the stimulus presented at the subject level was the presentation of famous and non-famous faces (Henson et al., 2002). The contrast of under study compared the appearance of a face regardless of fame status to background. There were 12 subjects and we considered the use of a canonical haemodynamic response function, or HRF, the function that represents the shape of the BOLD response to a stimulus. Our test considered group level activation. That is, considering areas of common response to the task across subjects. Following the SPM paradigm, this implies a one sample t-test applied voxel by voxel to the contrast data. The data are freely available from the Statistical Parametric Mapping (SPM) web site (http://www.fil.ion.ucl.ac.uk/spm/) where further background information and details on preprocessing is given.

In this setting, set-level analysis requires a parcellation of the brain. There are several methods of defining such sets, including: decomposing the imaging

space at multiple resolutions, using functional clustering, and using a labeled anatomical map. We use the latter; specifically, the anatomical decomposition given by Tzourio-Mazoyer et al. (2002). As the anatomical atlas was at a finer resolution than the observed data, we down-sampled it to the lower resolution. This leads to small issues at boundary voxels of the parcellation that were not consequential for the overall analysis by impacting a very small percentage of tests. 22 of the brain regions had $1 - \widehat{\text{Afdr}}$ greater than 0.75 and 11 had $1 - \widehat{\text{Afdr}}$ greater than 0.85. These regions are highlighted in Figure 4. The results show differences in the occipital lobe and parts of the frontal and parietal. These results are not surprising, as the occipital lobe is associated with vision, and the task is visual, while the cortical group activation likely is associated with processing the visual information.

6.5 Obtaining standard errors via the bootstrap

We can easily obtain standard errors for the estimate $1 - \widehat{\text{Afdr}}$ by using a bootstrapping approach. In each bootstrap iteration, we sample with replacement from the cases and, separately, from the controls, re-estimate the feature-level posterior probabilities based on the new feature-level statistics, and obtain bootstrapped values of $1 - \widehat{\text{Afdr}}$. 100 such bootstrap iterations were used with the data from Sotiriou et al. (2006) concerning differential expression between ER-negative and ER-positive breast tumors. The bootstrap standard deviations for the 22 atoms described in the previous section are all between 0.065 and 0.176.

7 Discussion

We introduced a general approach for set-level inference for high-dimensional data, which casts the problem in a decision-theoretic framework and focuses on estimation rather than testing. Set-level inference is an area of increasing interest in many areas of science, because of the necessity of combining quantitative feature-level data with annotations resulting from alternative sources of information. Our method introduces the concept that set-level inference is best performed for disjoint sets (atoms), in order to obtain increased scientific clarity and interpretability. We discuss in detail an implementation that focuses on quantifying the differences between sets based on the expected number of false discoveries (EFD) and the expected number of missed discoveries (EMD). These have a clear interpretation and provide information about the question of greatest interest, which relates to quantifying the fraction of alternatives in each set.

Our approach introduces a new paradigm in set-level inference. Most present methods are based on performing a hypothesis test for each set. The p-values thus obtained are fed into a set-level analysis which in turn requires a multiple testing adjustment. The statistical properties of this overall strategy are difficult to interpret. We provide a rigorous unified framework for feature and set-level analysis. Our estimates have clearly defined optimality properties and are scientifically interpretable.

We show that the loss function defined as the weighted sum of false discoveries and missed discoveries for any union of atoms, can be reduced to a form which depends only on the marginal feature-level posterior probabilities. These probabilities can easily be estimated using existing Empirical Bayes methods. This simplification enables us to obtain an easy algorithm for obtaining the Bayes estimator, which is equivalent to setting a threshold and only letting those atoms whose realized atomic false discovery rate is below it to enter the Bayes estimator. We also provide alternate loss functions: Thus, we may either introduce linear constraints, which are equivalent to a regularization penalty, or consider the fractions of missed discoveries and false discoveries.

References

Anderson, J. and V. Blair (1982): "Penalized maximum likelihood estimation in logistic regression and discrimination," *Biometrika*, 69, 123–136.

Baldi, P. and A. Long (2001): "A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes," *Bioinformatics*, 17, 509–519.

Benjamini, Y. and R. Heller (2008): "Screening for partial conjunction hypotheses," *Biometrics*, 64, 1215–1222.

Benjamini, Y. and Y. Hochberg (1995): "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society B*, 57, 289–300.

Bouton, C. and J. Pevsner (2002): "DRAGON View: information visualization for annotated microarray data," *Bioinformatics*, 18, 323.

Cai, T. and W. Sun (2009): "Simultaneous Testing of Grouped Hypotheses: Finding Needles in Multiple Haystacks," *Journal of the American Statistical Association*, 104, 1467–1481.

Cui, X., J. Hwang, J. Qiu, N. Blades, and G. Churchill (2005): "Improved statistical tests for differential gene expression by shrinking variance components estimates," *Biostatistics*, 6, 59–71.

Do, K., P. Müller, and F. Tang (2005): "A Bayesian mixture model for differential gene expression," *Applied Statistics*, 627–644.

Efron, B. and R. Tibshirani (2002): "Empirical bayes methods and false discovery rates for microarrays," *Genetic Epidemiology*, 23, 70–86.

Efron, B., R. Tibshirani, J. Storey, and V. Tusher (2001): "Empirical Bayes analysis of a microarray experiment," *Journal of the American Statistical Association*, 96, 1151–1160.

Friston, K., J. Ashburner, K. S.J., N. T.E., and P. W.D. (2007): *Parametric mapping: The analysis of functional brain images*, Academic Press.

Garey, M. and D. Johnson (1979): *Computers and intractability: a guide to NP-completeness*, WH Freeman and Company, San Francisco.

Genovese, C. and L. Wasserman (2002): "Operating characteristics and extensions of the false discovery rate procedure," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 499–517.

Gill, P., W. Murray, and M. Wright (1981): *Practical optimization*, London: Academic Press.

Goeman, J. and P. Buhlmann (2007): "Analyzing gene expression data in terms of gene sets: methodological issues," *Bioinformatics*, 23, 980.

Gottardo, R., J. Pannucci, C. Kuske, and T. Brettin (2003): "Statistical analysis of microarray data: a Bayesian approach," *Biostatistics*, 4, 597.

Green, P. and B. Silverman (1994): *Nonparametric regression and generalized linear models: A roughness penalty approach*, New York: Chapman and Hall.

Heller, R., E. Manduchi, G. Grant, and W. Ewens (2009): "A flexible two-stage procedure for identifying gene sets that are differentially expressed," *Bioinformatics*, 25, 1019.

Henson, R., T. Shallice, M. Gorno-Tempini, and R. Dolan (2002): "Face repetition effects in implicit and explicit memory tests as measured by fMRI," *Cerebral Cortex*, 12, 178.

Irizarry, R., C. Wang, Y. Zhou, and T. Speed (2009): "Gene set enrichment analysis made simple," *Johns Hopkins University, Dept. of Biostatistics Working Papers*, 185.

Lönnstedt, I. and T. Speed (2002): "Replicated microarray data," *Statistica Sinica*, 12, 31–46.

Michaud, J., K. Simpson, R. Escher, K. Buchet-Poyau, T. Beissbarth, C. Carmichael, M. Ritchie, F. Sch
"utz, P. Cannon, M. Liu, et al. (2008): "Integrative analysis of RUNX 1 downstream pathways and target genes," *BMC genomics*, 9, 363.

Mirnics, K., F. Middleton, A. Marquez, D. Lewis, and P. Levitt (2000): "Molecular characterization of schizophrenia viewed by microarray analysis of gene expression in prefrontal cortex," *Neuron*, 28, 53–67.

Mootha, V., C. Lindgren, K. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, et al. (2003): "PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes," *Nat. Genet*, 34, 267–273.

Müller, P., G. Parmigiani, and K. Rice (2007): "FDR and Bayesian multiple comparisons rules," *Bayesian statistics*, 8.

Müller, P., G. Parmigiani, C. Robert, and J. Rousseau (2004): "Optimal sample size for multiple testing: the case of gene expression microarrays," *Journal of the American Statistical Association*, 99, 990–1002.

Newton, M. and C. Kendziorski (2003): "Parametric empirical bayes methods for microarrays," in R. I. G. Parmigiani, E.S. Garrett and S. Zeger, eds., *The analysis of gene expression data: methods and software*, New York: Springer Verlag.

Newton, M., C. Kendziorski, C. Richmond, F. Blattner, and K. Tsui (2001): "On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data," *Journal of Computational Biology*, 8, 37–52.

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

Parmigiani, G., E. Garrett, R. Anbazhagan, and E. Gabrielson (2002): "A statistical framework for expression-based molecular classification in cancer," *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 64, 717–736.

Parsons, D., S. Jones, X. Zhang, J. Lin, R. Leary, P. Angenendt, P. Mankoo, H. Carter, I. Siu, et al. (2008): "An Integrated Genomic Analysis of Human Glioblastoma Multiforme," *Science*, 321, 1807.

Reiner-Benaim, A., D. Yekutieli, N. Letwin, G. Elmer, N. Lee, N. Kafkafi, and Y. Benjamini (2007): "Associating quantitative behavioral traits with gene expression in the brain: searching for diamonds in the hay," *Bioinformatics*, 23, 2239.

Sarkar, S. and T. Zhou (2008): "Controlling Bayes directional false discovery rate in random effects model," *Journal of Statistical Planning and Inference*, 138, 682–693.

Schaeffer, E., L. Marchionni, Z. Huang, B. Simons, A. Blackman, W. Yu, G. Parmigiani, and D. Berman (2008): "Androgen-induced programs for prostate epithelial growth and invasion arise in embryogenesis and are reactivated in cancer," *Oncogene*, 27, 7180–7191.

Sotiriou, C., P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, et al. (2006): "Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis," *JNCI Journal of the National Cancer Institute*, 98, 262.

Storey, J. (2002): "A direct approach to false discovery rates," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64, 479–498.

Storey, J. (2003): "The positive false discovery rate: A Bayesian interpretation and the q-value," *The Annals of Statistics*, 31, 2013–2035.

Storey, J., J. Akey, and L. Kruglyak (2005): "Multiple locus linkage analysis of genomewide expression in yeast," *PLoS Biology*, 3.

Storey, J. and R. Tibshirani (2003a): "Sam thresholding and false discovery rates for detecting differential gene expression in dna microarrays," in R. I. G. Parmigiani, E.S. Garrett and S. Zeger, eds., *The analysis of gene expression data: methods and software*, New York: Springer Verlag.

Storey, J. and R. Tibshirani (2003b): "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences of the United States of America*, 100, 9440.

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

Tavazoie, S., J. Hughes, M. Campbell, R. Cho, and G. Church (1999): "Systematic determination of genetic network architecture," *Nature Genetics*, 22, 281–285.

Tzourio-Mazoyer, N., B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot (2002): "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *Neuroimage*, 15, 273–289.

Wright, S. and J. Nocedal (2006): *Numerical optimization*, Springer.

Appendix A: Proofs of results from main manuscript

Proof of Theorem 1. Using Lemma 6 from the Appendix C:

$$\begin{aligned}
\text{EFD}(U) &= \sum_{\tau\in 2^{\mathscr{M}}} \sum_{m\in U\setminus\tau} p_\tau = \sum_{m\in U} \sum_{\tau\in 2^{\mathscr{M}}, m\notin\tau} p_\tau = \sum_{m\in U}(1-p_m^*) = |U| - \sum_{m\in U} p_m^* \\
\text{EMD}(U) &= \sum_{\tau\in 2^{\mathscr{M}}} \sum_{m\in\tau\setminus U} p_\tau = \sum_{m\notin U} p_m^*
\end{aligned}$$

□

Proof of Corollary 2. By the mixture model in Equation (1), when considered as functions of the data, the marginal feature-level posterior probabilities can be rewritten as:

$$\begin{aligned}
p_m^* &= P(\text{feature } m \text{ is from the alternative distribution}|X,Y) \\
&= P(\text{feature } m \text{ is from the alternative distribution}|X_m,Y) \\
&= 1 - P(\text{feature } m \text{ is from the null distribution}|X_m,Y) \\
&= 1 - \pi_0 \frac{f_0(X_m|Y)}{f(X_m|Y)} \\
&= 1 - \text{fdr}(X_m|Y)
\end{aligned}$$

Thus, the results of Theorem 1 may be rewritten as:

$$\begin{aligned}
\text{EFD}(U) &= \sum_{m\in U} \text{fdr}(X_m,Y_m) \\
\text{EMD}(U) &= \sum_{m\notin U} [1-\text{fdr}(X_m,Y_m)]
\end{aligned}$$

□

Proof of Lemma 3.

$$\begin{aligned}
\mathscr{L}(\delta) &= (1-w)\sum_{m\in U}(1-p_m^*) + w\sum_{m\notin U} p_m^* \\
&= (1-w)\sum_{A_l\in U} EFD(A_l) + w\sum_{A_l\notin U}(n_l - EFD(A_l)) \\
&= (1-w)\delta'\mathbf{EFD}_A + w(\mathbf{1}-\delta)'(\mathbf{n}-\mathbf{EFD}_A) \\
&= \delta'\{\mathbf{EFD}_A - w\mathbf{n}\} + w\mathbf{1}'(\mathbf{n}-\mathbf{EFD}_A)
\end{aligned}$$

□

Proof of Theorem 4. The posterior expected losses of L, L_f^λ, and L_a^ξ can be parametrized as affine functions of δ, for a fixed w between 0 and 1. Any affine function of δ, $h(\delta) = \delta'\mathbf{a} + b, \delta \in \{0,1\}^J, \mathbf{a} \in \mathbb{R}^J, b \in \mathbb{R}$ is minimized when $\delta_j = 1\{a_j \leq 0\}$, since $h(t)$ is minimized when $\delta'\mathbf{a}$ is minimized. This is a linear function in each component δ_j of δ, and if we minimize it in each component we also minimize it overall. As a result, it is minimized by choosing to sum only over those components of $\mathbf{a}$ which are negative or zero. □

Proof of Theorem 5.

$$\begin{aligned}
\mathscr{L}_r(\delta) &= (1-w)\frac{\sum_{m\in U}(1-p_m^*)}{|U|} + w\frac{\sum_{n\notin U} p_m^*}{M-|U|} \\
&= (1-w)\frac{\sum_{A_l\in U}\mathrm{EFD}(A_l)}{\sum_{A_j\in U} n_l} + w\frac{\sum_{A_l\notin U}(n_l - \mathrm{EFD}(A_l))}{\sum_{A_l\notin U} n_l} \\
&= (1-w)\frac{\delta'\mathbf{EFD}_A}{\delta'\mathbf{n}} + w\frac{(\mathbf{1}-\delta)'(\mathbf{n}-\mathbf{EFD}_A)}{M-\delta'\mathbf{n}} \\
&= \delta'\left\{\frac{(1-w)}{\delta'\mathbf{n}}\mathbf{EFD}_A - \frac{w}{(M-\delta'\mathbf{n})}(\mathbf{n}-\mathbf{EFD}_A)\right\} + \frac{w}{M-\delta'\mathbf{n}}(\mathbf{n}-\mathbf{EFD}_A)
\end{aligned}$$

□

Appendix B: Atoms

Overlapping sets of features may result in erroneous statistical inferences by inducing correlations between sets. For example, in the case of methods which are based on calculating a p-value for each set, multiple testing adjustments can be complicated by correlations between the hypotheses. Overlap can also cause issues with interpretability, e.g. if a large set in $\mathscr{S}$ gets a poor score, but a smaller set which shares a large fraction of its genes with the large gene-set gets a good score, it is unclear where this difference comes from. See Figure 5 for an example in terms of gene-sets and their fraction of differentially expressed genes: In both (A) and (B), the smaller set (set 1) has 50 genes, 30 of which are differentially expressed, the larger set (set 2) has 100 genes, 40 of which are differentially expressed, and their intersection has 20 genes. However, in panel (A), none of the genes in the intersection is differentially expressed, whereas in (B), they are all differentially expressed. The percent of differentially expressed genes in the difference between sets 2 and 1 varied greatly between the two cases (50% and 25%, respectively).

Here we consider only non-overlapping sets, which we call *atoms*. By considering only non-overlapping sets we avoid the difficulties in analysis and interpretation from overlap. We let $\mathscr{A}$ be the set of atoms which we obtain from the

sets, $\mathscr{S}$ where $\mathscr{A} = \{A_1, \ldots, A_L\}$ has L different non-overlapping sets. Each element of $\mathscr{A}$ is a subset of an element in $\mathscr{S}$ and the intersection of any two elements in $\mathscr{A}$ is empty. One way to obtain atoms is to divide the sets into their largest non-overlapping subunits, so that the collection of atoms $\mathscr{A}$ is defined as the set of minimal cardinality which has the following properties:

1. Given any set $S \in \mathscr{S}$, there exists $\mathscr{I} \in \{1, \ldots, L\}$ such that $S = \cup_{i \in \mathscr{I}} A_i$.
2. Given any atoms A_i and A_j in $\mathscr{A}$ with $i \neq j$, $A_i \cap A_j = \emptyset$.

In the examples shown in Figure 5 (A) and (B), this is a very natural step to take, as we can simply consider three atoms, consisting of the intersection between sets 1 and 2, the difference between set 1 and set 2, and the difference between set 2 and set 1. In particular, in Figure 5 (A), there appear to be two disjoint sets with large fractions of alternatives, whereas in Figure 5 (B), there seems to be a common signal between sets 1 and 2. Hence, it is more informative to split the overlapping sets.

Appendix C: Decision theory framework

We consider a decision theory framework to integrate the feature-level and set-level inference. While the concept is general, we detail it here in terms of a loss function that is linear in two components, one relating to false discoveries and the other to missed discoveries. Within this context we study conditions under which the posterior expected loss can be written in terms of feature-level posterior probabilities.

The discrepancy between τ and a candidate estimator U in $\mathscr{U}$ can be represented by a loss function. We consider the following general class of loss functions, which depend on a discrepancy function d and a fixed constant $w \in [0,1]$:

$$L(\tau, U) \quad = \quad (1-w) \sum_{m \in U \setminus \tau} d(m, \tau) + w \sum_{m \in \tau \setminus U} d(m, U) \tag{7}$$

for all $U \in \mathscr{U}$. We note that $U \setminus \tau$ represents the set of *false discoveries,* while $\tau \setminus U$ represents the set of *missed discoveries* if we were to estimate τ by U. Thus, the loss function is linear in two components, the first one measuring how close features which are false discoveries are to the set of features from the alternative distribution (τ), the second one measuring how close features which are missed discoveries are to the candidate estimator (U).

We consider general discrepancy measures d between features in $\mathscr{M}$ which satisfy:

$$d : \mathscr{M} \times \mathscr{M} \quad -> \quad [0, \infty)$$

$$
\begin{aligned}
d(m,m) &= 0 \text{ for any } m \in \mathscr{M} \\
d(m_1,m_2) &= d(m_2,m_1) \text{ for any } m_1,m_2 \in \mathscr{M}
\end{aligned}
$$

We could use for example the single-linkage (nearest neighbor) function to define the discrepancy between a feature and a set $d(m,A) := \min\{d(m,m_0) : m_0 \in A\}$.

Based on the loss function L, we get the following posterior expected loss:

$$
\mathscr{L}(U) = \sum_{\tau \in 2^{\mathscr{M}}} L(\tau,U) * P(\tau|X,Y)
$$

where $2^{\mathscr{M}}$ is the power set of $\mathscr{M}$ and $P(\tau|X,Y)$ is the posterior probability of set τ exactly representing the set of features which are from the alternative distribution. In particular, $P(\tau|X,Y)$ can be seen as a discrete probability distribution with support $2^{\mathscr{M}}$. It fully reflects uncertainty from the feature-level modeling and dependencies between features.

We use the following notation for the posterior probabilities, for simplicity:

$$
p_\tau = P(\tau|X,Y)
$$

Thus, the posterior expected loss may be written as:

$$
\begin{aligned}
\mathscr{L}(U) &= (1-w) \sum_{\tau \in 2^{\mathscr{M}}} \sum_{n \in U \setminus \tau} d(m,\tau) p_\tau + w \sum_{\tau \in 2^{\mathscr{M}}} \sum_{n \in \tau \setminus U} d(m,U) p_\tau \\
&= (1-w) E_{\tau|X,Y}\{ \sum_{m \in U \setminus \tau} d(m,\tau)\} + w E_{\tau|X,Y}\{ \sum_{m \in \tau \setminus U} d(m,U)\}
\end{aligned}
$$

The two components may be interpreted as the posterior expected value of the sum of distances from each false discovery ($m \in U \setminus \tau$) to the set of features from the alternative distribution (τ) and the posterior expected value of the sum of distances from each missed discovery ($m \in \tau \setminus U$) to the set of all discoveries (U).

The posterior expected loss is written in terms of posterior set-level probabilities. In order to obtain a Bayes estimator, we would need to minimize it, which could be extremely complicated from a modeling point of view, because we would need to model the joint distribution of all features. It would also be extremely computationally intensive, as the number of unions of atoms is 2^L, where we recall that L is the total number of atoms.

It would be much easier to estimate the posterior expected loss for a particular set and to find the union of atoms which minimizes it if we could write it in terms of the posterior probabilities of individual features being from the alternative distribution. We call these probabilities *marginal feature-level posterior probabilities*. They can be estimated in a relatively straight-forward manner, by building

a probability model and using a fully Bayesian framework or an Empirical Bayes (EB) framework.

We introduce the following notation for the *marginal posterior probability for a set U*, which is the sum of all posterior probabilities of sets which include U:

$$p_U^* = \sum_{\tau \in 2^{\mathscr{M}}, U \subset \tau} p_\tau$$

It represents the posterior probability that the set U is included in the set of all features from the alternative distribution. For specific cases we can simply write out the features in the set, e.g. $p_{12} = p_{21} = p_{\{1,2\}}$. The marginal feature-level posterior probabilities are a specific case of this, where the set represents a single feature, i.e.:

$$p_m^* = \sum_{m \in \tau} p_\tau$$

We prove a lemma which simplifies the form of the two components of the posterior expected loss function. Note, in particular, that $E_{\tau|X,Y}\{\sum_{m \in \tau \setminus U} d(m,U)\}$ can be written as a linear function of marginal feature-level posterior probabilities.

Lemma 6. *Under the loss function described in equation (7), in the case of a general discrepancy measure d and single linkage, the following simplified forms of $E_{\tau|X,Y}\{\sum_{m \in U \setminus \tau} d(m,\tau)\}$ and $E_{\tau|X,Y}\{\sum_{m \in \tau \setminus U} d(m,U)\}$ are obtained:*

$$\begin{aligned} E_{\tau|X,Y}\{\sum_{m \in U \setminus \tau} d(m,\tau)\} &= \sum_{\tau \in 2^{\mathscr{M}}} \sum_{m \in U \setminus \tau} d(m,\tau) p_\tau = \sum_{m \in U} \sum_{\tau \in 2^{\mathscr{M}}, m \notin \tau} d(m,\tau) p_\tau \\ E_{\tau|X,Y}\{\sum_{m \in \tau \setminus U} d(m,U)\} &= \sum_{\tau \in 2^{\mathscr{M}}} \sum_{m \in \tau \setminus U} d(m,U) p_\tau = \sum_{m \notin U} d(m,U) p_m^* \end{aligned}$$

Proof.

$$\begin{aligned} \sum_{m \in U} \sum_{\tau \in 2^{\mathscr{M}}, m \notin \tau} d(m,\tau) p_\tau &= \sum_{\tau \in 2^{\mathscr{M}}, m \in U \setminus \tau} d(m,\tau) p_\tau = \sum_{\tau \in 2^{\mathscr{M}}} \sum_{m \in U \setminus \tau} d(m,\tau) p_\tau \\ \sum_{m \notin U} d(m,U) p_m^* &= \sum_{m \notin U} d(m,U) \sum_{\tau \in 2^{\mathscr{M}}, m \in \tau} p_\tau = \sum_{m \in \tau \setminus U, \tau \in 2^{\mathscr{M}}} d(m,U) p_\tau \\ &= \sum_{\tau \in 2^{\mathscr{M}}} \sum_{m \in \tau \setminus U} d(m,U) p_\tau \end{aligned}$$

□

We now show that $E_{\tau|X,Y}\{\sum_{m \in U \setminus \tau} d(m,\tau)\}$ cannot be written as an affine function of marginal feature-level posterior probabilities for a general discrepancy

measure d which takes into account how far or close features are to each other, in the case where the single linkage property holds. Thus, in general, to calculate the posterior expected loss, we need to model the joint distribution of all the features. This also leads to much more complex computations.

Lemma 7. *Under the loss function described in equation (7), in the case of a general discrepancy measure d and single linkage, $E_{\tau|X,Y}\{\sum_{m\in U\setminus\tau} d(m,\tau)\}$ cannot be written as an affine function of marginal feature-level posterior probabilities. Therefore $\mathscr{L}(U)$ also cannot be written as an affine function of marginal feature-level posterior probabilities.*

Proof. We show that we cannot write $E_{\tau|X,Y}\{\sum_{m\in U\setminus\tau} d(m,\tau)\}$ as an affine function of the marginal feature-level posterior probabilities.

Step 1. We first show that, for any proper subset $\nu \subsetneq 2^{\mathscr{M}}$, setting any affine of the posterior probabilities of the sets in ν equal to 0 forces all the coefficients to be 0, i.e.:

Denote the elements in ν by $\tau_{l_1},\ldots,\tau_{l_{|\nu|}}$. We will show that setting any affine function of these elements to 0 implies that all the coefficients are 0. Thus, we have:

$$\sum_{\tau\in\nu} a_\tau p_\tau + b = 0 \tag{8}$$

We note that if $\nu = \{\tau_{l_1},\ldots,\tau_{l_{|\nu|}}\} \subsetneq 2^{\mathscr{M}}$, then $p_{\tau_{l_1}} + \ldots + p_{\tau_{l_{|\nu|}}} \leq 1$ and $p_{\tau_{l_1}} \geq 0,\ldots,p_{\tau_{l_{|\nu|}}} \geq 0$. Plugging in $p_{\tau_{l_1}} = 1, p_{\tau_{l_2}} = \ldots = p_{\tau_{l_{|\nu|}}} = 0$, followed by $p_{\tau_{l_1}} = \frac{1}{2}, p_{\tau_{l_2}} = \ldots = p_{\tau_{l_{|\nu|}}} = 0$, and solving the resulting system of equations in $a_{\tau_{l_1}}$ and b results in $a_{\tau_{l_1}} = b = 0$. From here on, plugging in only one non-zero probability for each $\tau \in \nu$ in turn will result in $a_{\tau_{l_2}} = \ldots = a_{\tau_{l_{|\nu|}}} = 0$.

Step 2. We now apply the result in *Step 1* to show that $E_{\tau|X,Y}\{\sum_{m\in U\setminus\tau} d(m,\tau)\}$ can in general not be written as an affine function of the marginal feature-level posterior probabilities. We note that we have:

$$E_{\tau|X,Y}\{\sum_{m\in U\setminus\tau} d(m,\tau)\} = \sum_{\tau\in 2^{\mathscr{M}}}\sum_{m\in U\setminus\tau} d(m,\tau)p_\tau = \sum_{\tau\in 2^{\mathscr{M}},\tau\neq U}\{\sum_{m\in U\setminus\tau} d(m,\tau)\}p_\tau \tag{9}$$

since $d(m,\tau) = 0$ if $m \in \tau$. Using a simple transformation, we note that showing that $E_{\tau|X,Y}\{\sum_{m\in U\setminus\tau} d(m,\tau)\}$ we need to show that we can find a_m and b such that:

$$E_{\tau|X,Y}\{\sum_{m\in U\setminus\tau} d(m,\tau)\} \quad = \quad \sum_{m\in\mathscr{M}} a_m(1-p^*_m) + b \tag{10}$$

$$
\begin{aligned}
&= \sum_{m \in \mathscr{M}} a_m \sum_{\tau \in 2^{\mathscr{M}}, m \notin \tau} p_\tau + b \\
&= \sum_{\tau \in 2^{\mathscr{M}}} \sum_{m \notin \mathscr{M} \setminus \tau} a_m p_\tau + b \\
&= \sum_{\tau \in 2^{\mathscr{M}}, \tau \neq \mathscr{M}} \{ \sum_{m \in \mathscr{M} \setminus \tau} a_m \} p_\tau + b
\end{aligned}
$$

The coefficients a_m and b are more accurately written as $a_m(U)$ and $b(U)$, but we use the simpler notation here. Setting the expressions in 9 and 10 equal to each other, we get:

$$
\sum_{\tau \in 2^{\mathscr{M}}, \tau \neq U} \{ \sum_{m \in U \setminus \tau} d(m, \tau) \} p_\tau = \sum_{\tau \in 2^{\mathscr{M}}, \tau \neq \mathscr{M}} \{ \sum_{m \in \mathscr{M} \setminus \tau} a_m \} p_\tau + b
$$

We may now take $\nu = 2^{\mathscr{M}} \setminus \mathscr{M}$, which is a proper subset of $2^{\mathscr{M}}$. Using the result in *Step 1*, we get:

$$
\sum_{m \in U \setminus \tau} a_m \quad = \quad \sum_{m \in U \setminus \tau} d(m, \tau),
$$

all the other coefficients being 0. Now consider cycling through all the sets τ such that $U \setminus \tau$ consists of a single element. We thus obtain:

$$
d(m, \tau) = a_m \text{ for all } m \in U \setminus \tau
$$

regardless of how many elements there are in $\tau \setminus U$ and how far away they are from the elements in $U \setminus \tau$.

To illustrate this last portion of the proof, consider $\mathscr{M} = \{1, 2, 3\}$ and $U = \{1, 2\}$. Then:

$$
\begin{aligned}
\tau = \{2\} \quad &=> \quad d(1, \{2\}) = a_1 \\
\tau = \{2, 3\} \quad &=> \quad d(1, \{2, 3\}) = a_1
\end{aligned}
$$

Given our use of the single-linkage property, $d(1, \{2, 3\}) = min\{d(1, \{2\}), d(1, \{3\})\}$. So if $d(1, \{3\}) \geq d(1, \{2\})$, then $d(1, \{2\}) = d(1, \{3\})$, which means that the discrepancy measure d does not take into account how far or close features are to each other. $\square$

Appendix D: Algorithmic solution for the minimization of $\mathscr{L}_r(\delta)$

We first consider the solution to the problem where we constrain the size of the Bayes estimator $|U| = \delta'\mathbf{n}$ to some size ρ. In this case, we get the following constrained linear binary problem:

$$\begin{aligned} \min_{\delta} \quad & \delta' \left\{ \frac{(1-w)}{\rho}\mathbf{EFD}_A - \frac{w}{(M-\rho)}(n - \mathbf{EFD}_A) \right\} \\ \text{s.t.} \quad & \delta'\mathbf{n} = \rho \end{aligned} \tag{11}$$

This is an instance of the well-known 0-1 knapsack problem Garey and Johnson (1979), which can be solved approximately by Dantzig's greedy algorithm. This uses a sorting strategy where atoms are sorted increasingly by the quantity

$$\frac{(1-w)}{\rho}Afdr_l - \frac{w}{(M-\rho)}(1 - Afdr_l)$$

and t_l is set to 1, in order, until $\delta'\mathbf{n} = \rho$. Note that when $\rho = M/2$, atoms are sorted according to $Afdr_l$ as in Theorem 4.

In principle, to solve the fractional problem, one can solve the 0-1 knapsack problem for each possible value of $\rho = |U|$ and select the best solution. Since ρ can range over a large number of possible values, we use a strategy based on the projected gradient of the fractional function at a given point to find a small number of estimator sizes $|U|$ to test.

In summary, the algorithm is as follows: (1) initialize ρ; (2) find solution δ_ρ by solving the 0-1 knapsack problem; (3) find point $\mathbf{s}$ as the minimizer of $\mathscr{L}_r(\delta)$ along the linear-piecewise path $\delta_\rho - \alpha(\nabla_{\boldsymbol{\delta}}\mathscr{L}_r(\delta_\rho))_+$ where $(\nabla_{\boldsymbol{\delta}}\mathscr{L}_r(\delta_\rho))_+$ is the projected gradient at δ_ρ (see Wright and Nocedal (2006)); (4) stop if $\mathbf{s}'\mathbf{n} = \rho$, otherwise set $\rho = \mathbf{s}'\mathbf{n}$ and repeat step (2). Step (3) is a univariate optimization problem which can be solved using any univariate numerical minimization technique (golden-section method, for instance).

Algorithms

Algorithm 1 Algorithm to obtain the Bayes estimator for the loss function L_r.

$\rho \leftarrow \min\{\mathbf{n}\}$.
$I_{iter} \leftarrow 0$.
while $I_{iter} \leq I_{iter}^{max}$ **do**
  Find $\delta_\rho = \min_\delta \delta' \left\{ \frac{(1-w)}{\rho} \mathbf{EFD}_A - \frac{w}{(M-\rho)}(n - \mathbf{EFD}_A) \right\}$, along $\delta'\mathbf{n} = \rho$.
  Find $\mathbf{s} = \min_\delta \mathscr{L}_r(\delta)$, along $\delta = \delta_\rho - \alpha(\nabla_\delta L_r(\delta_\rho))_+$.
  if $\mathbf{s}'\mathbf{n} = \rho$ **then**
    stop.
  else
    $\rho \leftarrow \mathbf{s}'\mathbf{n}$.
  end if
  $I_{iter} \leftarrow I_{iter} + 1$
end while

Algorithm 2 Algorithm to obtain empirical Bayes estimates of the feature-level posterior probabilies.

1: Obtain B sets of null statistics by using distributional assumptions.
2: Denote the null statistics by X_{m0}^{b}, where m indexes the features, $1 \leq m \leq M$ and b the simulations under the null, $1 \leq b \leq B$.
3: Estimate the probability π_0 of a randomly selected feature being from the null distribution for a series of thresholds c by $\hat{\pi}_0(c) = \frac{\#\{F_m \leq c\}}{\#\{F_m^{0b} \leq c\}/B}$.
4: Choose the estimate $\hat{\pi}_0$ by smoothing over $\hat{\pi}_0(c)$, employing the approach detailed in Storey and Tibshirani (2003b). If the number of features is less than 500, make the conservative assumption that π_0 is 1.
5: Estimate the ratio $f_0(X_m)/f(X_m)$ for every feature m by logistic regression, considering the observed statistics (X_m) as "successes" and the nulll statistics (X_{m0}^{b}) as "failures," employing the approach in Anderson and Blair (1982), with a natural cubic spline using a fixed number of equally spaced knots, as in Green and Silverman (1994).
6: Estimate the posterior probability of a specific feature m being from the alternative distribution by using the estimates for π_0 and f_0/f from steps 4 and 5 above and the plug-in formula in Equation (6).

Figures

Figure 1: Workflow diagram. Note that we are combining two sources of information: The feature-level information from the present study, and the annotation information, which represents a distillation of prior scientific knowledge. They are combined to obtain the atom-level EFD (expected false discoveries) and EMD (expected missed discoveries), and finally, the Bayes estimator.

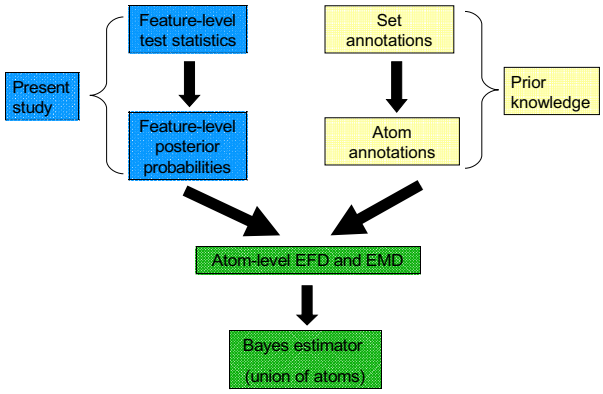


Figure 2: True fractions of alternatives for the simulated datasets described in Table 3. Each boxplot shows the values over the 100 simulation runs for $1 - \widehat{\text{Afdr}}$ (left panel) and for the q-values (right panel) for different fractions of alternatives. The blue line in the left panel represents the ideal scenario, where $1 - \widehat{\text{Afdr}}$ perfectly estimates the fraction of alternatives.

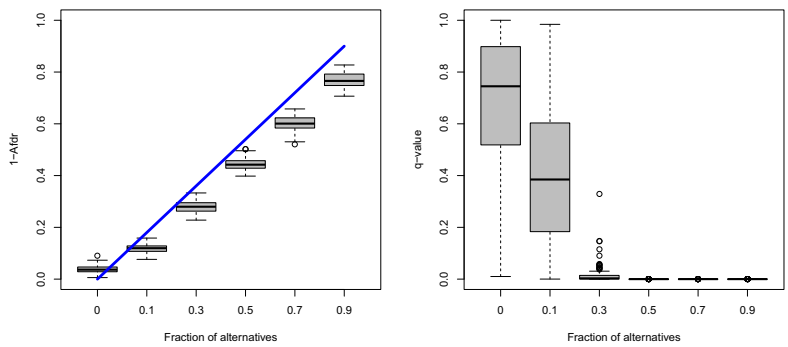


Figure 3: Histograms of the p-values obtained for the `limma` method for atoms of sizes 10, 50, and 100 which have no features from the alternative distribution. Note as the set size increases, the histograms are increasingly skewed towards 1. The mean p-value for the atom of size 10 is 0.606, while for the atom of size 50 it is 0.646, and for the atom of size 100 it is 0.751.

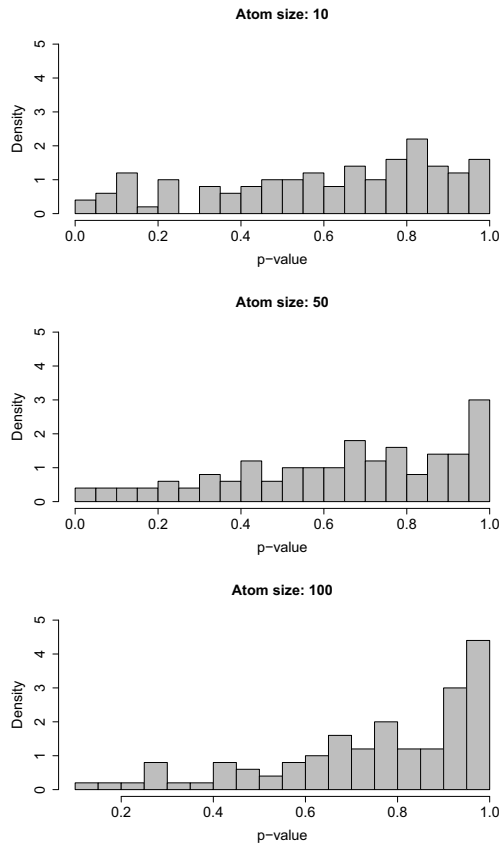


Figure 4: Rendered brain images in three orientations with the highlighted regions being those with $1 - \widehat{\text{Afdr}}$ greater than 0.75 (in yellow), respectively 0.85 (in red).

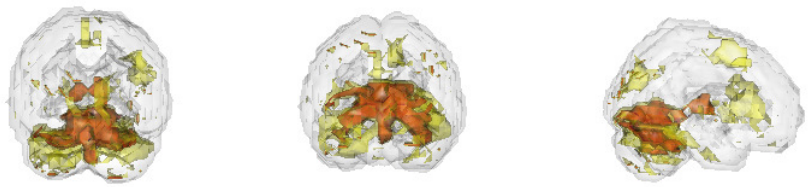
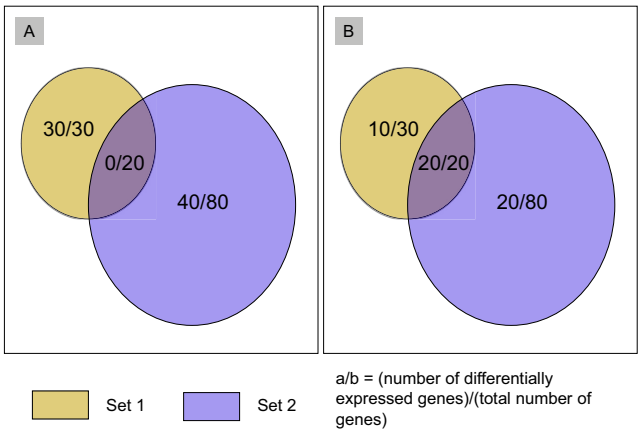


Figure 5: Set 1 has 60% of genes differentially expressed (30/50), while set 2 has 40% of genes differentially expressed (40/100). However, in (A) there are no differentially expressed genes common to sets 1 and 2, and although set 2 has a lower percentage of differentially expressed genes compared to set 1, the percentage of differentially expressed genes which are in set 2 but not in set 1 is higher than the percentage of differentially expressed genes which are in set 2 (50% compared to 40%). In (B), the percentage of differentially expressed genes which are in set 2 but not in set 1 (25%) is lower than the percentage of differentially expressed genes in set 2 (40%).



Tables

Table 1: Summary from 100 simulations using sets 1A, 1B, 2A, and 2B. using the `limma` method. The mean and standard deviation of the p-values and q-values are calculated over the 100 runs. Set 1A and Set 2A have a fraction of alternatives of 0.6, while Set 1B and Set 2B have a fraction of alternatives of 0.4. Since the sets are overlapping, the test gives no information about where the large difference in p-values come from.

	Set	Fract. of alt.	mean p-value	sd p-value	mean q-value	sd q-value
1	1A	0.6	0.045	0.057	0.109	0.115
2	2A	0.4	0.887	0.099	0.96	0.049
3	1B	0.6	0.05	0.054	0.119	0.115
4	2B	0.4	0.883	0.089	0.968	0.04

Table 2: Summary from 100 simulations using sets 1A, 1B, 2A, and 2B, employing our method with the atoms obtained from the intersections and differences of the original sets. The mean and standard deviation of the estimated $1 - \text{Afdr}$ and the mean and standard deviation of the p-values and q-values are calculated over the 100 runs. We note that the results are much more interpretable than those in Table 1.

	Atom	Fract. of alt.	mean $1 - \widehat{\text{Afdr}}$	sd $1 - \widehat{\text{Afdr}}$	mean p-value	sd p-value	mean q-value	sd q-value
1	$1A \setminus 2A$	1	0.916	0.038	<0.001	<0.001	<0.001	<0.001
2	$1A \cap 2A$	0	0.024	0.03	0.999	0.001	1	<0.001
3	$2A \setminus 1A$	0.5	0.467	0.019	0.294	0.166	0.572	0.301
4	$1B \setminus 2B$	0.33	0.318	0.028	0.864	0.112	0.993	0.033
5	$1B \cap 2B$	1	0.914	0.044	<0.001	<0.001	<0.001	<0.001
6	$2B \setminus 1B$	0.25	0.245	0.018	1	<0.001	1	<0.001

Table 3: Summary from 100 simulations with 6 atoms of size 50. There are 2500 features total, the 2200 which are not distributed among the atoms all being from the null distribution. Thus, the overall percentage of features which are from the alternative distribution is 5%. The mean and standard deviation of the estimated $1 - \text{Afdr}$, as well as the mean and standard deviation of the p-values and q-values from the `limma` method, are calculated over the 100 runs.

	Fract. of alt.	mean $1-\widehat{\text{Afdr}}$	sd $1-\widehat{\text{Afdr}}$	mean p-value	sd p-value	mean q-value	sd q-value
1	0.9	0.746	0.039	<0.001	<0.001	<0.001	<0.001
2	0.7	0.59	0.034	<0.001	<0.001	<0.001	<0.001
3	0.5	0.43	0.032	<0.001	<0.001	<0.001	<0.001
4	0.3	0.271	0.026	0.007	0.021	0.011	0.03
5	0.1	0.107	0.027	0.36	0.263	0.407	0.285
6	0	0.029	0.027	0.672	0.247	0.686	0.238

Table 4: Summary from 100 simulations with 3 atoms having fractions of alternatives of 0.5, but different set sizes. The mean and standard deviation of the estimated $1 - \text{Afdr}$, as well as the mean and standard deviation of the p-values and q-values from the `limma` method, are calculated over the 100 runs.

	Size	mean $1-\widehat{\text{Afdr}}$	sd $1-\widehat{\text{Afdr}}$	mean p-value	sd p-value	mean q-value	sd q-value
1	10	0.44	0.059	0.019	0.031	0.019	0.031
2	50	0.436	0.034	<0.001	<0.001	<0.001	<0.001
3	100	0.438	0.026	<0.001	<0.001	<0.001	<0.001

Table 5: Bayes estimators obtained from different loss functions over 100 runs. Atom names use the fraction of alternatives and the set size; for example atom 0.5 - 50 has a fraction of alternatives 0.5 and 50 features. The ideal scenario (in bold), gives posterior probabilities of 1 and 0 to the features from the alternative, respectively from the null distribution. It is compared to the simulation results. In parentheses are listed the number of simulations in which a particular union of atoms is the Bayes estimator (only if it appeared in at least 20 out of 100 runs.)

w	L	L_f^{λ}	L_a^{ξ}	L_r
0.25	**0.9 - 50, 0.9 - 100**	**Empty set**	**0.9 - 50, 0.9 - 100**	**0.9 - 50, 0.9 - 100**
	0.9 - 50, 0.9 - 100 (79)	Empty set (100)	Empty set (68)	0.9 - 50, 0.9 - 100 (61)
			0.9 - 100 (27)	0.9 - 50 (37)
0.5	**0.9 - 50, 0.5 - 50, 0.9 - 100, 0.5 - 100**	**0.9 - 50, 0.9 - 100**	**0.9 - 50, 0.9 - 100**	**0.9 - 50, 0.9 - 100**
	0.9 - 50, 0.9 - 100 (87)	0.9 - 50, 0.9 - 100 (97)	0.9 - 50, 0.9 - 100 (100)	0.9 - 50, 0.9 - 100 (84)
0.67	**0.9 - 50, 0.5 - 50, 0.9 - 100, 0.5 - 100**	**0.9 - 50, 0.9 - 100**	**0.9 - 50, 0.5 - 50, 0.9 - 100, 0.5 - 100**	**0.9 - 50, 0.5 - 50, 0.1 - 50, 0.9 - 100, 0.5 - 100, 0.1 - 100**
	0.9 - 50, 0.5 - 50, 0.9 - 100, 0.5 - 100 (100)	0.9 - 50, 0.9 - 100 (99)	0.9 - 50, 0.5 - 50, 0.9 - 100, 0.5 - 100 (79)	0.9 - 50, 0.5 - 50, 0.1 - 50, 0.9 - 100, 0.5 - 100, 0.1 - 100 (52)
				0.9 - 50 0.9 - 100 (36)

Table 6: Comparison of our method to the GSEA, t-test, and Wilcoxon rank test methods on a dataset from Subramanian et al. (2005). We present the results for the top 10 sets using our method.

	Set	$1-\widehat{\text{Afdr}}$	`limma` p-values	`limma` q-value	GSEA p-value	GSEA q-value	t-test p-value	t-test q-value
1	chrYq11	0.621	<0.001	0.014	<0.001	<0.001	<0.001	<0.001
2	chr4q22	0.956	0.154	0.995	0.195	0.996	0.176	0.92
3	chr2q14	0.958	0.867	0.995	0.411	0.996	0.302	0.92
4	chrXp11	0.959	0.204	0.995	0.149	0.996	0.058	0.641
5	chrXq26	0.963	0.999	0.999	0.677	0.996	0.79	0.984
6	chr14q21	0.965	0.939	0.995	0.902	0.996	0.742	0.984
7	chr10q21	0.966	0.939	0.995	0.54	0.996	0.358	0.92
8	chr7p22	0.966	0.92	0.995	0.891	0.996	0.986	0.992
9	chr5q33	0.966	0.872	0.995	0.263	0.996	0.328	0.92
10	chr12p12	0.967	0.103	0.995	0.636	0.996	0.893	0.984

Table 7: The estimated fraction of alternatives ($1 - \widehat{\text{Afdr}}$) for the 22 atoms resulting from 10 KEGG pathways, using data from Sotiriou et al. (2006). Atoms are labeled based on the KEGG sets they are in, each set having a numerical identifier. Note that the highest estimated fraction is in atom 04310,04110,04120 which consists of 6 genes and represents an intersection of three pathways (the Wnt signaling pathway, the cell cycle, and ubiquitin mediated proteolysis). As a comparison, the range of the estimated fraction of true discoveries for the original pathways was approximately 0.15 to 0.37.

Atom	Size	$1 - \widehat{\text{Afdr}}$
04310	223	0.22
04310,04110	23	0.06
04310,04120	7	0.12
04310,04110,04120	**6**	**0.54**
04110	174	0.34
04110,04120	30	0.22
04120	189	0.17
00010,00071	20	0.15
00010	40	0.24
00010,00030	18	0.22
00010,00020	11	0.25
00010,00230	4	0.00
00020	41	0.12
00030	16	0.25
00030,00230	4	0.11
00061	11	0.34
00071	51	0.20
03020,00230,00240	32	0.33
03022	43	0.23
00230	130	0.19
00230,00240	40	0.38
00240	34	0.38