Published in final edited form as:

Annu Rev Stat Appl. 2015 April ; 2: 95–111. doi:10.1146/annurev-statistics-010814-020335.

Multiset Statistics for Gene Set Analysis

Michael A. Newton^{1,2} and Zhishi Wang¹

¹Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706

²Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin 53706

Abstract

An important data analysis task in statistical genomics involves the integration of genome-wide gene-level measurements with preexisting data on the same genes. A wide variety of statistical methodologies and computational tools have been developed for this general task. We emphasize one particular distinction among methodologies, namely whether they process gene sets one at a time (uniset) or simultaneously via some multiset technique. Owing to the complexity of collections of gene sets, the multiset approach offers some advantages, as it naturally accommodates set-size variations and among-set overlaps. However, this approach presents both computational and inferential challenges. After reviewing some statistical issues that arise in uniset analysis, we examine two model-based multiset methods for gene list data.

Keywords

gene set enrichment; role model; statistical genomics

1. WHAT IS GENE SET ANALYSIS?

Biological systems are so complex that even our data-rich research environment fails in making many predictions, such as those regarding the fate of a tissue (e.g., will it become cancerous?) or the consequence of an intervention (e.g., what effect will a drug have on viral infection?). Biomedical researchers who address such challenges have at their disposal assays that yield genome-wide gene-level measurements on, for example, the expression of the RNA or protein products of each gene, structural properties of the chromatin harboring each gene, other factors affecting the regulation of each gene, and associations between variants of each gene and various disease states. After great expense and effort the researcher, has obtained her genome-wide gene-level data, say, D and is challenged to make sense of it via various forms of data analysis. One central analytical challenge is how to relate the local, endogenous D to all of the other exogenous knowledge, say, K that has so far been compiled on the same genes. Great efforts are underway to encode this exogenous

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

knowledge in ways that facilitate data analysis, and the use of the resulting knowledge resources is becoming an essential component of biological research. One encoding of K is via collections of gene sets, each of which is an unordered set of genes identified by some other evidence as having some specific biological property. Gene set analysis refers to a host of strategies and procedures for integrating the observed D with available gene set information in the pursuit of further knowledge. This type of analysis addresses several important challenges in genomic data analysis, including the following:

1. Gene sets enable data reduction, allowing the organization, simplification, and explanation of a high-dimensional signal. This reduction is especially useful when gene-level signals are relatively strong (e.g., when many genes are differentially expressed between two cellular states) and when a concise description of the functional content of such signals can be derived from the gene sets.
2. Gene sets improve sensitivity compared with gene-level analysis in cases in which genes in the same set have consistent but weak signals.
3. Gene sets structure gene-level data in a way that may improve the prediction of other phenotypes (e.g., regression, biomarker development).

The content of a gene set analysis depends on (a) the structures of data D and knowledge K , (b) the bioinformatic and statistical tools being applied, and (c) the context of the driving problem. In Section 2, we review some of the major efforts at encoding K in publicly available data resources. Further, a large number of bioinformatic and statistical tools are available for integrating D with K . As a reference point, we recognize useful reviews of gene set analysis methodology. Khatri et al. (2012) present a comprehensive review from the perspective of computational biology. Their review describes three major classes of analysis techniques: the earliest enrichment methods (overrepresentation analysis), more quantitative functional class scoring methods, and methods that recognize internal relationships within each set. Goeman & Bühlmann (2007), Barry et al. (2008), Ackermann & Strimmer (2009), and Maciejewski (2014) survey statistical considerations, especially for Khatri's first two classes. Their papers shed light on hypothesis testing issues and on multivariate analysis issues; the articles also shed light on the various ways that permutations and bootstraps may be used to assess statistical significance. Virtually all of the reviewed methods focus on gene sets considered one at a time, and in this sense they are uniset methods. Of course, individual gene sets contain multiple genes, so uniset methods are multivariate from the perspective of gene-level data. Analysts recognize that gene sets are organized in larger assemblages, but this fact does not inform uniset inference computations, except possibly through post hoc adjustment for false discovery rate (FDR) control. We review uniset statistics and calibration issues in Section 3.

More recently, methods have emerged that aim to improve overall inference by simultaneously processing all of the sets in a collection. Simultaneous analysis is known to have benefits in other domains of high-dimensional inference (e.g., empirical Bayes estimation), and this multiset approach has the potential to deal with the persistent inference challenges of gene set analysis. One challenge is that set size affects testing power, so even if procedures are calibrated on a null hypothesis, a power imbalance on the alternative

hypothesis remains (Newton et al. 2007). This imbalance affects how we prioritize or rank order the most interesting gene sets in a given analysis. For instance we might rank by uniset p -values, looking at the sets having the smallest p -values as the ones worth reporting. An artifact emerges because, all else being equal, large sets more easily yield small p -values than do small sets.

A more substantial challenge is the set overlap problem. In addition to statistical dependencies between sets sharing genes, there may be spurious associations and problems with redundancy. To see why, note that many methods focus on the multivariate joint distribution of gene-level measurements within a fixed set and use hypothesis testing to detect associations of this distribution with sample characteristics (e.g., expression measurements in different cell types). These statistical methods consider a gene set S to be nonnull (and thus worthy of reporting) if any gene contained within it exhibits a difference in mean expression between two cellular states. The cause of the differential expression of gene $g \in S$ might be the altered activity of a molecular pathway that is encoded by the set S . Now any other set S' that happens to contain g is nonnull by association, even if the function represented by S' is unaltered. Thus, the multifunctionality of individual genes, called pleiotropy, may lead to spurious gene set associations (Bauer et al. 2010, Newton et al. 2012).

Although pleiotropy is one expression of the set overlap problem, extensive overlap also happens because biological properties can be expressed at different levels of resolution and from different perspectives (e.g., Gillis & Pavlidis 2013). The upshot for a uniset statistical analysis is that reported findings may exhibit substantial redundancy, and one fails to achieve a concise functional description of the gene-level data D . Some bioinformatic tools address this problem by working with trimmed collections of gene sets (e.g., GO SLIMS), by invoking post hoc heuristics to simplify reporting (e.g., Supek et al. 2011), or by adjusting a uniset p -value on the basis of closely related sets (e.g., Alexa et al. 2006, Grossmann et al. 2007). None of these solutions is entirely satisfactory. We lose the full information content and fine granularity of knowledge K when using a trimmed collection, we risk masking functional signals when using the rule-based clustering schemes available for postprocessing uniset output, and available p -value adjustment schemes are cumbersome and use only a fraction of the actual set overlap information.

In Section 4, we discuss some multiset statistical approaches aimed at detecting functional signals via genome-wide modeling of the gene-level data. These methods adhere to the following basic premise: In addition to using gene-level data on genes within a given gene set, inference regarding this set should incorporate knowledge about other sets containing these same genes. This incorporation occurs explicitly through a parsimonious probability model describing how gene-level data are produced as a consequence of activity patterns of biological functions represented by the gene sets. As with any model-based approach, a cost in terms of model simplifications exists that may be at odds with the reality of data generation. The model assumptions are explicit, however, and so may be evaluated; the inference summaries may be relatively insensitive to these assumptions; and the joint modeling approach may be the only way to enable probabilistic reasoning for multiset

analysis. We illustrate two model-based multiset calculations in a gene expression study of nasopharyngeal carcinoma in Section 4.2.

2. GENE SET REPOSITORIES

When evidence links genes by virtue of some shared biological property, these genes compose an archivable gene set. Large-scale collaborative research efforts are underway to encode such biological knowledge. Other efforts, notably BIOCONDUCTOR (Gentleman et al. 2004) provide open software systems for accessing these data.

2.1. Gene Ontology

The Gene Ontology (GO) project from the Gene Ontology Consortium (2000) aims to provide a structured and controlled vocabulary for describing genes and their products across different species. This project has three main goals: to compile and maintain the structured and controlled ontologies by virtue of shared model organism databases, to describe the roles of genes and their products using these ontologies, and to develop tools for querying and manipulating the ontologies (e.g., the GO Consortium has developed AmiGO for searching and browsing GO). Each term (i.e., set) in GO is assigned to one of three root ontologies: molecular function (the function of a gene product at the biochemical level), cellular component (the place in the cell where a gene product is active), or biological process (a biological objective to which the gene product contributes). These three domains are organized into three directed acyclic graphs, wherein each node is a functional category (GO term) and directed edges convey proper-subset information. The terms in each ontology are linked to one another by two types of relationships: “is-a,” which represents a simple class–subclass relationship, and “part-of,” which refers to a component relationship. GO is not static. The initial GO project involved only three model organism databases: the *Saccharomyces* genome database (SGD), the *Drosophila* genome database (FlyBase), and the mouse genome database (MGD). Since the initial project, more and more model organism databases have been incorporated, and GO has now become a standard knowledge base for integration and interpretation of large-scale molecular datasets. At writing, GO contains over 34,400 terms, reflecting diverse biological function in a large number of organisms (BIOCONDUCTOR, GO.DB, version 2.10.1). **Figure 1** renders a small piece of GO: it presents the molecular functions containing 5–10 genes, together with their associated genes, in a bipartite graphical representation.

2.2. Kyoto Encyclopedia of Genes and Genomes

The Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto 2000) is an effort to link genomic and molecular information with high-level functional information (e.g., cell, organism, and ecosystem) and, in doing so, to produce a computer representation of the biological system. KEGG was initiated by the Japanese Human Genome Program in 1995, and it has since been developed and extended to now maintain 16 main databases. These databases are related to different aspects of the biological system, including systems information, genomic information, chemical information, and health information. KEGG has been widely used to infer the functional significance within large-scale genetic studies.

2.3. Reactome

REACTOME (Joshi-Tope et al. 2005) is a curated, peer-reviewed knowledge base of human biological processes. REACTOME aims to provide an integrated view of biological processes that captures what is already known about the interactions between genes, proteins, and molecules by using a data model that is accessible to computation. The REACTOME project also provides tools for visualizing and interpreting large-scale experimental data sets using REACTOME pathways. The core unit in the REACTOME data model is a reaction in which physical input entities are converted to output entities; reactions linked together by shared physical entities then form a biological pathway. These reactions are gathered by experts in the field, peer-reviewed, and edited by REACTOME team members prior to being published in the database. According to Croft et al. (2013), the latest version of REACTOME includes 7,088 human proteins; based on data extracted from 15,107 research publications with PubMed links, these proteins participated in 6,744 reactions.

2.4. Molecular Signature Database

The Molecular Signature Database (MSigDB) (Liberzon et al. 2011) focuses mainly on human gene sets and is a curated database created particularly for the GSEA method (Subramanian et al. 2005). Compared with other gene set sources, MSigDB includes more diverse and more types of gene sets. These include not only biological pathways collected from original publications but also gene sets derived directly from other known data sources such as GO, KEGG, and REACTOME. MSigDB version 4.0 includes 10,925 gene sets and a richer set of annotations.

3. UNISET STATISTICS

The structure of genome-wide gene-level data $D = \{D_g\}$ has considerable bearing on what statistics we compute on a set of genes. At one extreme, the data D_g associated with gene g record raw data from a multisample experiment, such as expression levels from a number of microarrays on cells under various experimental conditions. At another extreme, D_g may only record whether or not gene g was output on a list of genes deemed relevant in the study; for example, D_g may report the decision from a gene-level hypothesis test. Data integration efforts support the first extreme, whereby as much available raw data as possible are incorporated into the set-level analyses. Practical considerations may limit access to such raw data, however; such considerations may also present substantial complexities if we attempt to model them, and these considerations then support analyses using simple gene lists. Both extremes, as well as various intermediate cases, actually occur in practice.

Regardless of the structure of data D_g , two quite different modes of comparison are possible when evaluating a gene set S . As articulated by Goeman & Bühlmann (2007), we can compare (a) what is in S with what is outside of S or (b) what is in S with what we might have expected to be in S if had some null hypothesis had held on the distribution of D_g for $g \in S$. The former comparison is called competitive; the latter is self-contained. Tian et al. 2005 make a similar distinction, which is relevant not only to what test statistic we use to score a set S but also to how we calibrate the statistical significance of that score.

Many methods for the construction of set-level test statistics follow a two-stage format (e.g., Barry et al. 2008, Ackermann & Strimmer 2009). Data on gene g are reduced to a local real-valued statistic, say, T_g , and local statistics are then combined into a real-valued set-level statistic, say, U_S , either by averaging or by some rank-based combination. The SAFE score (Barry et al. 2005) and the GSEA score (Subramanian et al. 2005) are prominent examples; the max-mean statistic is another (Efron & Tibshirani 2007). Recent improvements include the self-contained ROAST score (Wu et al. 2010) and the competitive CAMERA score (Wu & Smyth 2012). In a curious reversal, the score proposed by Sartor et al. (2009) treats the set as a binary response and the gene-level scores T_g as predictor values in a logistic regression model. Other methods are more explicitly multivariate: They focus on the vectors holding gene-level scores in each biological sample, and they address possible effects on the joint distribution of this vector caused by various treatment factors. These methods have not proven more useful than the two-stage scores described above, owing to challenges with the estimation of covariance (Ackermann & Strimmer 2009). Both the two-stage and multivariate methods are examples of functional class scoring statistics according to the classification by Khatri et al. (2012). If gene-level scores T_g are simple indicators that gene g is reported on a short list of interesting genes, we are in a somewhat simpler domain: The only useful second-stage score is a count, such as $U_S = \sum_{g \in S} T_g$. When U_S is unusually large, we say S is enriched for interesting genes, or, equivalently, the list of interesting genes is enriched for type S genes. Methods based on assessing overrepresentation remain popular, partly owing to their simplicity (Khatri & Drăghici 2005, Khatri et al. 2012).

Having constructed a set-level score $U_S(D)$, the next question is how to calibrate it. Answering this question tells us something about how interesting the data in S are, either in isolation or in comparison with those in other sets. Considerable attention has been paid to the calibration problem. Possible solutions include permuting samples, permuting genes, bootstrapping, or invoking some other randomization that aims to respect whatever null hypothesis is in view. These solutions are discussed in the following subsections.

3.1. Sample Permutation

Sample permutation is applicable when the data structure involves multiple samples labeled by some kind of phenotypic response, such as cases in which multiple expression profiles are labeled by the cell type or by the treatment applied to the cells under investigation. In such cases, the local T_g is some test statistic measuring association between the gene-level data and the label. Sample permutation amounts to shuffling the sample labels and recomputing the T_g values to develop a reference distribution for $U_S(D)$. Here, the set S is fixed, as are all of the gene-level data D_g except the sample labels. Sample permutation has the benefit that any dependencies among genes within S are respected. A weakness of this approach lies in the stringency of the underlying null hypothesis, which asserts that all genes $g \in S$ present data that are independent of the phenotypic label. Problems with pleiotropic effects and power imbalance complicate the interpretation of such self-contained analyses when applied to multiple sets.

3.2. Gene Permutation

Gene permutation is applicable over a wider range of data structures, as it requires only gene-level scores $\{T_g\}$. This approach is not intended to work on the same stringent null hypothesis of sample permutation; indeed, it acknowledges that some fraction of genes g in set S may have an association between genomic data and sample label. At issue, rather, is whether the extent of this association is unusually large in comparison with the amount of association in the whole genome. A weakness of gene permutation is that it ignores any particular internal structure of data D_g (e.g., dependence among genes within each set), and thus type I error rates can be inflated. Yet, this competitive approach remains a dominant one, possibly because of its flexibility with data structures or possibly owing to the exploratory manner in which it is deployed. If gene-level data are fixed, gene permutation simply asks how the observed $U_S(D)$ compares with the same statistic computed on random sets S^* of the same size as S .

3.3. Other Calibration Methods

The ROAST method (Wu et al. 2010) is a self-contained calibration approach that utilizes a form of conditional inference suitable to ANOVA-style decompositions of the gene-level data. Unlike sample permutation, this approach relies on an interesting scheme by which simulated data have the same covariance pattern as observed data within the set. The nonparametric bootstrap is a competitive approach that can accommodate dependence-respecting tests, such as tests of the null hypothesis that the average gene-level effect within a set S is the same as the average effect over the whole genome (Barry et al. 2008; Dudoit & Van der Laan 2007). The small sample sizes associated with many genomic studies correspond to high approximation error for bootstrap methods. Alternatively, Wu & Smyth. (2012) derive competitive statistics $U_S(D)$ by first estimating the variance inflation effect caused by dependence and by then referring to a suitable t -approximating null distribution.

4. MULTISSET STATISTICS

4.1. Role Modeling

GENGO (Lu et al. 2008), model-based gene set analysis (MGSA) (Bauer et al. 2010, 2011), and multifunctional analysis (MFA) (Wang et al. 2013) are three distinct inference tools for multiset analysis. These tools are not readily described as either competitive or self-contained; rather, they aim to identify interesting gene sets via modeling of gene-level data in terms of latent states of biological functions associated with the gene sets, which they infer simultaneously by some version of Bayesian inference. Their utility rests partly with the simplicity of both the model structure and the gene-level data structure, although more fully elaborated models could further the general data integration task. Owing to their multivariate stance, GENGO, MGSA, and MFA offer solutions to difficulties caused by power imbalance and gene set overlap.

MGSA and MFA rest on the same generative probability model for gene-list data, and the GENGO model is very similar to those underlying MGSA and MFA. The biggest differences among the three methods have to do with their operating characteristics and how inference is deployed. We first describe the basic generative model. We assume that the genome-wide

gene-level data D have been reduced to a gene list; in other words gene-level indicators $D = \{D_g\}$, where

$$D_g = \begin{cases} 1 & \text{if gene } g \text{ exhibits an interesting feature in endogenous data} \\ 0 & \text{if not.} \end{cases}$$

As with any gene-list data, the D_g may be the result of a gene-level hypothesis test computed from more extensive data. The context guides the language, but we might say that D_g indicates whether or not gene g seems to be activated in the experimental situation, or, equivalently, if it seems to be nonnull. It is also helpful to recognize that hypothesis tests may incur either false positive or false negative errors, and so the actual null/nonnull or inactive/active status is recorded in unobserved indicators:

$$A_g = \begin{cases} 1 & \text{if gene } g \text{ is active (nonnull)} \\ 0 & \text{if not.} \end{cases}$$

The first part of the role model simply recognizes gene-level hypothesis testing errors:

$$D_g | A_g \sim \text{Bernoulli} \begin{cases} \alpha & \text{if } A_g = 0 \\ \gamma & \text{if } A_g = 1 \end{cases}.$$

These parameters record the false positive rate α and the true positive rate γ ; together represented as $\theta = (\alpha, \gamma)$. Some presentations use $\beta = 1 - \gamma$ as the false negative rate. Typically, we have $\alpha < \gamma$, so truly active genes are reported to the list at a higher rate than are truly inactive genes. MGSA and MFA treat the estimation of the error rate parameters differently: MFA takes user input, recognizing the utility of external information, for example about the targeted false discovery rate of the gene-level tests. MGSA allows user input, but in the absence of such input, it produces estimates from fitting the gene-list data. Both methods use the same Bernoulli sampling model for D_g given A_g . They further assume mutual independence within $D = \{D_g\}$ conditionally upon the full list of activations $A = \{A_g\}$; dependencies are assumed to arise only from the joint distribution of A . Future elaborations of these tools could accommodate error dependencies. In both tools, the log likelihood from the observation component of the sampling model is as follows:

$$\begin{aligned} l_{\text{obs}}(A, \theta) &= \log \Pr(D|A, \theta) \\ &= \log \prod_g \Pr(D_g | A_g, \theta) \\ &= \sum_g \{A_g [D_g \log(\gamma) + (1 - D_g) \log(1 - \gamma)] + (1 - A_g) [D_g \log(\alpha) + (1 - D_g) \log(1 - \alpha)]\} \quad (1) \\ &= N_{1,1} \log \gamma + N_{1,0} \log(1 - \gamma) + N_{0,1} \log \alpha + N_{0,0} \log(1 - \alpha). \end{aligned}$$

Here, $N_{1,1} = \sum_g A_g D_g$ counts the truly active genes that appear active (i.e., are reported), $N_{1,0} = \sum_g A_g (1 - D_g)$ counts the truly active genes that are not reported,

$N_{0,1} = \sum_g (1 - A_g) D_g$ counts the truly inactive genes that are falsely reported, and
 $N_{0,0} = \sum_g (1 - A_g) (1 - D_g)$ counts the truly inactive genes that are correctly not reported.

The central structure of the role model relates gene activities $A = \{A_g\}$ to more primitive binary set activities $Z = \{Z_S\}$:

$$Z_S = \begin{cases} 1 & \text{if set } S \text{ is active (nonnull)} \\ 0 & \text{if not} \end{cases}.$$

The notion is that $Z_S = 1$ corresponds to an active role played by genes in S within the cells on test, and the gene's activity is inherited from any sets to which it belongs (roles it may have):

$$A_g = \begin{cases} 1 & \text{if set } Z_S = 1 \text{ for any } S \text{ with } g \in S \\ 0 & \text{if not} \end{cases}. \quad (2)$$

More concisely, $A_g = \max_{S: g \in S} Z_S$. Thus, the gene activities $A = A(Z)$ arise as a mapping from the set activities. **Figure 2** expresses this inheritance on a small component of the bipartite graph expressing GO molecular functions.

In MGSA, the unknown Z_S are modeled as independent and identically distributed (i.i.d.) Bernoulli variables, with success probability π , thus contributing the following log probability mass:

$$l_{\text{hidden}}(Z, \pi) = \sum_S \{Z_S \log(\pi) + (1 - Z_S) \log(1 - \pi)\}. \quad (3)$$

In a Bayesian analysis, the joint posterior of all unknowns has the following logarithm:

$$l_{\text{posterior}}(Z, \theta, \pi) = l_{\text{obs}}(A(Z), \theta) + l_{\text{hidden}}(Z, \pi) + \text{constant}. \quad (4)$$

The MGSA tool samples the joint posterior via Markov chain Monte Carlo (MCMC) and scores sets according to marginal posterior activation probability:

$$U_S(D) = P(Z_S = 1 | D). \quad (5)$$

Empirical evidence shows that scoring sets by marginal posterior probability substantially reduces the redundancy problem and results in reported sets with minimal overlap (Bauer et al. 2010, Wang et al. 2013).

It is helpful to reflect on why the computation in Equation 5 is difficult in complex systems of sets. Overlap and size variation limit the utility of uniset analysis, but these challenges simply complicate the computations of multiset analysis. This is because the apparent activation $D_g = 1$ of some gene, supposing it is not a false positive, must be attributable to the nonnull activity of some containing set. In evaluating a given set S , the calculation uses all of the D_g values for $g \in S$, but it must also recognize other sets S' containing such genes

g , as the activation of such sets might better explain the data. Of course, data on other genes in such sets S' have bearing on the issue of the activation of S' . We quickly find ourselves regressing away from the original set S in trying to fully evaluate the evidence supporting $Z_S = 1$; only a posterior probability computation enables this assessment. An article by Newton et al. (2012) investigated the complexity of the posterior computation; when the intersection graph of the sets is sufficiently simple, message-passing algorithms (e.g., the junction tree algorithm) enable exact computation of the marginal posterior. In systems as complex as GO, these exact algorithms are no longer practical.

Newton et al. (2012) also investigated the mapping $Z \rightarrow A$ in Equation 2, observing that although in general, this mapping is not invertible, it becomes invertible when we refine the definition of activity. The activation hypothesis (AH) asserts that a set of genes is active (nonnull) if and only if each gene in the set is active. Under the AH, the inverse mapping $A \rightarrow Z$ is as follows:

$$Z_S = \min_{g \in S} A_g, \quad (6)$$

which finds that the set must be active if it contains only active genes. Indeed this assumption is quite distinct from that given by Equation 2; the mathematical asymmetry is intended to reflect a biological primacy of functions (i.e., sets) compared with individual genes, and it has the useful consequence that multiple sets can be dealt with simultaneously. More practically, the AH provides a framework for managing a large number of interrelated statistical hypotheses.

Considering the directed graphical structure of GO, it may be reasonable to assert the AH. Take two sets: a more-specific set A and a less-specific set B , with $A \subset B$. The role model assumption (Equation 2) transfers activity to all genes $g \in A$ in case the larger set is nonnull: $Z_B = 1$. Genes $g \in A$ inherit their inclusion in set/property/term B by virtue of B describing a less specific annotation. In the absence of constraints, our collection of hypotheses could assert that (a) all genes g with property A are nonnull and (b) the set of genes with property A is null. To make this assertion more concrete, suppose that B describes genes encoding a class of soluble proteins and that A is a subset of B corresponding to the heaviest 20% of these proteins. Knowing that B is active could mean, for example, that all genes in B are differentially expressed between two cell types, implying that all genes for heavy soluble proteins are differentially expressed. How could this hypothesis be consistent with the assertion that the set of genes for heavy soluble proteins is equivalently expressed between cell types? The AH is intended to constrain the collection of hypotheses to avoid this sort of conundrum.

The noninvertibility of the $Z \rightarrow A$ mapping affects the operating characteristics of the posterior summaries. To measure the effect, Wang et al. (2013) formulated a slightly different prior, P_2 , as an alternative to the i.i.d. Bernoulli prior P_1 for $Z = \{Z_S\}$ encoded in Equation 3. Namely,

$$P_2(Z=z) = P_1(Z=z | Z \in AH),$$

where $z = \{z_S\}$ is a vector of possible binary set-level activation states. By conditioning the i.i.d. prior on the constraint AH, one greatly alters the distribution of probability mass over the high-dimensional space of latent activity states. The P_1 probability of AH is small for complex collections [and getting smaller, according to the analysis in an article by Wang et al. (2013)]. It is as if, like water leaking from a system of pipes, the prior mass distributed by P_1 has leaked into regions of the state space that ought to be avoided. The cost of fixing this leak is that posterior computations under the P_2 prior must respect the constraints imposed by the AH. Wang et al. (2013) showed that the AH is equivalent to a set of linear inequality constraints on the joint collection of activities $\{\{A_g\}, \{Z_S\}\}$ and further that the log posterior given by Equation 4 is a linear function of this augmented set of variables. The MFA tool utilizes a novel MCMC scheme to approximate $U_S(D) = P_2(Z_S = 1|D) = P_1(Z_S = 1|D, \text{AH})$; this tool also deploys integer-linear programming (ILP) to find the joint maximum a posteriori (MAP) estimate $\hat{Z} = \{\hat{Z}_S\}$ by maximizing a linear function in binary variables subject to linear inequality constraints. Initial numerical studies reported in the aforementioned article by Wang et al. (2013) suggest that MFA has improved sensitivity compared with MGSA.

4.2. Example: Gene Expression Changes Associated with Viral Infection

Sengupta et al. (2006) reported a genome-wide expression study of nasopharyngeal carcinoma (NPC), paying particular attention to unusually extensive negative associations between host genes in the NPC cells and the expression of a key gene in the infecting Epstein–Barr virus (EBV). Gene set analysis was used in that study to guide the interpretation of genome-wide findings and to inform follow up experiments. The gene-level Spearman correlations between host genes and the EBV gene are available in the `allegR` package, which performs basic uniset computations (Newton et al. 2007). Here we reconsider these data in order to illustrate multiset computations.

Genes showing extreme negative Spearman correlations were entered into the gene list; in total, the 5% FDR-controlled list comprised 438 genes (Entrez ID). We integrated this list with GO[5:50], a collection of 5,994 GO terms that hold between 5 and 50 human genes. GO[5:50] itself annotates only about half of the human genes (10,293), and it annotates only 232 of the genes showing NPC–EBV association. We applied MFA and MGSA to assess the functional content of this reduced NPC–EBV gene list. For both methods, we fixed the false positive rate α and the true positive rate γ using a simple mixture argument applied to available gene-level data. The mixture argument goes as follows: The marginal probability $P(D_g = 1)$ is estimated by the relative size of the gene list; the marginal inactivation rate $P(A_g = 0)$ is estimated by the average, over all genes, of the local FDR $P(A_g = 0 | \text{correlation}_g)$, as computed by the `locfdr` R package (Efron 2004). Restriction of this averaging to genes reported on the top list gives an estimate of $P(A_g = 0 | D_g = 1)$, and restricting to those not so reported gives an estimate of $P(A_g = 0 | D_g = 0)$. The true positive and false positive rates follow by applying Bayes's rule, as shown by the following **R** code:

```
## R version 3.0.1; alleg version 2.0.1
data(npc) ## NPC Spearman correlations between host expression and virus
```

```

scores
<- (1/2)*sqrt(28)*log((1-npc)/(1+npc)) ## variance stabilized
pv <- pnorm(scores, lower.tail = FALSE) ## nominally N(0,1) on H0
qv <- p.adjust(pv, method = ``BH``) ## Benjamini Hochberg
ok <- (qv <= 0.05) ### probe sets at 5% FDR by Benjamini Hochberg
## now consider locfdr; for mixture estimation
library(locfdr)
fit <- locfdr(scores, nulltype = 0, plot = 0)
u <- fit$fdr
## A = gene level activity indicator
## D = gene level call
pD1 <- mean(ok) ## about 1% of genes reported
pA0 <- mean(u) ## about 69% of genes are truly null
pA0gD1 <- mean(u[ok]) ## P(A = 0|D = 1), FDR about 4% (good, since BH says
5%!)
pA0gD0 <- mean(u[!ok]) ## P(A = 0|D = 0), about 70%
## now Bayes rule to get P(D|A)
pD1gA0 <- pA0gD1*pD1/pA0 ## alpha = false positive rate 0.00061
pD1gA1 <- (1-pA0gD1)*pD1/(1-pA0) ## gamma = true positive rate 0.03297

```

The list of annotated EBV-associated genes was integrated with GO, fixing the error rates estimated above in both MGSA and MFA. **Table 1** summarizes the gene sets inferred to be active by MFA–ILP, that is, the MAP estimate of the active states. 44 sets constitute the MAP estimate. The uniset analysis previously reported in an article by Sengupta et al. (2006) is not directly comparable with **Table 1**, as it worked with a simpler collection that included relatively larger sets. We do not attempt a detailed analysis of the identified functions, except to say that (a) some of them recapitulate in further detail the broad classes previously reported (e.g., GO:0005031 and GO:2001238 are particular aspects of the cell death function) and (b) understanding the role of these functions in NPC is part of ongoing research. Our primary objective, rather, is to use the NPC example as a platform to compare MFA and MGSA, which rely on similar statistical assumptions but yield different findings.

To control the comparison, we used the MGSA estimate of π to drive the MFA calculations, including both optimization and marginal posterior computation. **Table 1** reports MGSA-computed marginal posterior probabilities assigned to sets in MAP estimate by MFA. Considering the characteristics of MCMC, very little of the discrepancy can be explained by Monte Carlo error, and this discrepancy therefore represents differences induced by the different prior assumptions. As was also reported by Wang et al. (2013), both MFA and MGSA produce relatively nonoverlapping gene sets (those with very little redundancy). An interesting feature of MFA is the coverage of the 232 NPC-associated genes:

$$\begin{aligned}\text{coverage} &= \sum_g \hat{A}_g D_g \\ \text{miscoverage} &= \sum_g \hat{A}_g (1 - D_g),\end{aligned}$$

where $\hat{A}_g = 1$ if gene g is part of any of the inferred active sets. MFA covers 99 genes; the MGSA estimate covers 72 genes using 25 sets (14 in common with MFA; see **Table 2**). MFA miscovers more genes than MGSA does (789 versus 673), but the coverage/miscoverage ratio favors MFA, as it has done in all of the examples considered to date. Both methods cover the reported gene list much better than the standard uniset method (Fisher's test) does. For example, if we rank sets by Fisher's p -value, then we would miscover 869 genes in accumulating enough sets to cover 99 NPC-associated genes.

5. SUMMARY

The integration of experimental genome-wide gene-level data with preexisting data on the functional properties of the same genes remains a central challenge in genomic data analysis. International efforts, including GO and BIOCONDUCTOR, continue to develop computational knowledge systems for the organization and analysis of genomic data. These systems rely on statistical methodology to enable the design of effective data analysis tools. From the extensive literature on such methodology, we discussed both uniset and multiset statistics in this brief report: Uniset statistics process gene-level data within each set to derive set-level scores; these scores are calibrated according to some reference distribution and possibly corrected for multiple testing as a last step. Multiset statistics are multivariate in the sets themselves; model-based multiset statistics entail jointly modeling genome-wide gene-level data. On the one hand, these statistics present an opportunity to overcome persistent difficulties caused by gene set overlap and set-size variation; on the other hand, the current formulations remain extremely simplistic in their modeling of data integration. Further research is needed to better understand the effects of (a) prior structure, (b) constraints, (c) error dependencies, (d) gene-level data reduction, and (e) annotation uncertainty, among other issues. On (a), for example, the i.i.d. Bernoulli prior of MGSA for set activities induces a non-i.i.d. prior on gene activities. We know that constraining the i.i.d. prior has a substantial effect, so the uniformity scale (set versus gene) may also have an effect that requires examination. On (b), the AH constraint induces identifiability of the set-level hypothesis, although one could entertain alternative formulations. For example, activation of a set might induce activation of a subset of its genes, rather than all of them. There may not be too much to gain by this approach, as it may do little other than altering the accounting of what are true versus false positives and what are true versus false negatives. In any case, other invertible gene set activity mappings may exist that would enable simultaneous consideration of multiple sets.

As with so much model-based analysis, the scope of potential computations and diagnostics is limited only by the imagination; the multiset modeling approach will gain traction via its successful application to genomic examples. Examples are still too few at present; many opportunities present themselves, and many more will do so given the continual elaboration of the genomic knowledge base. We hope that the development of open software tools (R packages cited in Bauer et al. 2011, Wang et al. 2013), will allow more extensive empirical investigation that in turn will yield further insights into this challenging data integration problem.

ACKNOWLEDGMENTS

M.A.N. was supported in part by grants from the National Institutes of Health: R21HG006568 and U54AI117924. Z.W. was supported by a fellowship from the Morgridge Institute of Research.

LITERATURE CITED

- Ackermann M, Strimmer K. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*. 2009; 10(1):47. [PubMed: 19192285]
- Alexa A, Rahnenfuhrer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*. 2006; 22:1600–7. [PubMed: 16606683]
- Ashburner M, Bell CA, Blake JA, Botstein D, Butler H, Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nat. Genet*. 2000; 25:25–29. <http://www.geneontology.org>. [PubMed: 10802651]
- Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*. 2005; 21:1943–49. [PubMed: 15647293]
- Barry WT, Nobel AB, Wright FA. A statistical framework for testing functional categories in microarray data. *Ann. Appl. Stat*. 2008; 2:286–315.
- Bauer S, Gagneur J, Robinson PN. GOing Bayesian: model-based gene set analysis of genome-scale data. *Nucleic Acids Res*. 2010; 38(11):3523–32. [PubMed: 20172960]
- Bauer S, Robinson PN, Gagneur J. Model-based gene set analysis for Bioconductor. *Bioinformatics*. 2011; 27(13):1882–83. [PubMed: 21561920]
- Croft D, Mundo AF, Haw R, Milacic M, Weiser J, et al. The Reactome pathway knowledgebase. *Nucleic Acids Res*. 2014; 42(1):D472–77. [PubMed: 24243840]
- Dudoit, S.; van der Laan, MJ. *Multiple Testing Procedures with Applications to Genomics*. Springer; New York: 2007.
- Efron B. Large-scale simultaneous hypothesis testing: the choice of the null hypothesis. *J. Am. Stat. Assoc*. 2004; 99:96–104.
- Efron B, Tibshirani R. On testing the significance of sets of genes. *Ann. Appl. Stat*. 2007; 1:107–29.
- Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004; 5(10):R80. [PubMed: 15461798]
- Gillis J, Pavlidis P. Assessing identity, redundancy and confounds in Gene Ontology annotations over time. *Bioinformatics*. 2013; 29(4):476–82. [PubMed: 23297035]
- Goeman JJ, Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*. 2007; 23(8):980–87. [PubMed: 17303618]
- Grossmann S, Bauer S, Robinson PN, Vingron M. Improved detection of overrepresentation of Gene-Ontology annotations with parent–child analysis. *Bioinformatics*. 2007; 23:3024–31. [PubMed: 17848398]
- Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, et al. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*. 2005; 33:D428–32. <http://www.reactome.org>. [PubMed: 15608231]
- Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 2000; 28:27–30. <http://www.genome.jp/kegg>. [PubMed: 10592173]
- Khatri P, Ghisic S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*. 2005; 21:3587–95. [PubMed: 15994189]
- Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLOS Comput. Biol*. 2012; 8:e1002375. [PubMed: 22383865]
- Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular Signatures Database (MSigDB) 3.0. *Bioinformatics*. 2011; 27:1739–40. <http://www.broadinstitute.org/gsea/msigdb>. [PubMed: 21546393]

- Lu Y, Rosenfeld R, Simon I, Nau GJ, Bar-Joseph Z. A probabilistic generative model for GO enrichment analysis. *Nucleic Acids Res.* 2008; 36:e109. [PubMed: 18676451]
- Maciejewski H. Gene set analysis methods: statistical models and methodological differences. *Brief. Bioinform.* 2014; 15:504–18.
- Newton MA, He Q, Kendzierski C. A model-based analysis to infer the functional content of a gene list. *Stat. Appl. Genet. Mol. Biol.* 2012; 11(2) doi:10.2202/1544-6115.1716.
- Newton MA, Quintana FA, den Boon JA, Sengupta S, Ahlquist P. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann. Appl. Stat.* 2007; 1(1):85–106.
- R Core Team. R: A Language and Environment for Statistical Computing. R Found. Stat. Comput.; Vienna: 2014. <http://www.R-project.org>
- Sartor MA, Leikauf GD, Medvedovic M. LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics.* 2009; 25:211–17.
- Sengupta S, den Boon JA, Chen I-H, Newton MA, Dahl DB, et al. Genome-wide expression profiling reveals EBV-associated inhibition of MHC class I expression in nasopharyngeal carcinoma. *Cancer Res.* 2006; 66(16):7999–8006. [PubMed: 16912175]
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS.* 2005; 102:15545–50. [PubMed: 16199517]
- Supek F, Bošnjak M, Skunca N, Smuc T. REVIGO summarizes and visualizes long lists of Gene Ontology terms. *PLOS ONE.* 2011; 6(7):e21800. [PubMed: 21789182]
- Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, et al. Discovering statistically significant pathways in expression profiling studies. *PNAS.* 2005; 102:13544–49. [PubMed: 16174746]
- Wang Z, He Q, Larget B, Newton MA. A multi-functional analyzer uses parameter constraints to improve the efficiency of model-based gene-set analysis. *Ann. Appl. Stat.* 2013 In press. arXiv: 1310.6322 [stat.ME].
- Wu D, Lim E, Vaillant F, Asselin-Labat ML, Visvader JE, Smyth GK. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics.* 2010; 26(17):2176–82. [PubMed: 20610611]
- Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.* 2012; 40(17):e133–33. [PubMed: 22638577]

SUMMARY POINTS

1. Characterizing the biological functions that underlie genomic data is essential for understanding gene-level findings and for guiding subsequent experimentation.
2. International efforts such as the Gene Ontology project provide structured knowledge bases that express biological functions as collections of gene sets, making statistical data integration possible.
3. Widely used uniset statistical methods tackle the integration problem in various ways but often do not incorporate the functional profiles of genes in a given set when scoring the statistical significance of that set. Variable set size and among-set overlaps complicate the interpretation of uniset results.
4. Model-based statistical approaches circumvent the size/overlap problems via explicit representation of the gene-level data in terms of latent binary activity states of the gene sets. These multiset methods involve relatively sophisticated computations, Monte Carlo integrations and constrained optimizations, and they have demonstrated improved performance over their uniset counterparts.
5. Model-based multiset analysis presents a compelling case in high-dimensional inference, as dependencies among units (genes) are explicit and induced by the interrelationships of the collected gene sets.

FUTURE ISSUES

1. Biology knowledge bases continue to mature. Although the number of genes in our genome has stabilized, the number of functions ascribable to genes continues to increase.
2. Current model formulations (the role model) enable multiset analysis but remain woefully simplistic in their ability to deal with the vagaries of genomic data. Elaborations of the role model, such as to better model gene dependencies, should further improve performance.
3. The continued elaboration of software systems both to visualize genomic data in the context of multiset information and to support nimble posterior computation will further enhance statistical solutions to the gene set analysis problem.

Figure 1. 543 Gene Ontology (GO) terms (*red*) representing all molecular functions (MFs) holding between 5 and 10 human genes and the associated 2,613 genes (*green*), presented as a bipartite graph. The edges (*blue*) connect genes to the sets that contain them. This graph has 134 components, including one large component with over 1,700 nodes. (Based on data in org.Hs.eg.db version 2.8.0.).

Figure 2.

One component from **Figure 1** involving 3 Gene Ontology (GO) terms and 17 genes (*left*). The right panel shows one realization of latent activity states, with an activated term (*black*) inducing activity on all the contained genes (*green*). Both model-based gene-set analysis (MGSA) and multifunctional analysis (MFA) use the set-to-gene activity mapping expressed in the right panel. MFA constrains latent activities to guarantee a unique gene-to-set mapping; obtaining such a mapping becomes an issue in more complex domains of the graph. Gene-level data are Bernoulli trials having a low rate (*white*) or a higher rate (*gray*).

Table 1

Nasopharyngeal carcinoma (NPC) example: 44 Gene Ontology (GO) terms inferred to be active by the multifunctional analysis (MFA) maximum a posteriori (MAP) estimate and ordered by marginal posterior activation probability

GO ID	Term (up to 40 chars)	Statistics	P.MFA	P.MGSA
GO:0005070	SH3/SH2 adaptor activity	4/49	1.00	0.98
GO:0030672	synaptic vesicle membrane	4/45	0.96	0.77
GO:0015377	cation:chloride symporter activity	2/8	0.93	0.16
GO:0033179	proton-transporting V-type ATPase, V0 do	2/10	0.93	0.20
GO:0005031	tumor necrosis factor-activated receptor	2/12	0.93	0.04
GO:0005154	epidermal growth factor receptor binding	3/20	0.92	0.84
GO:0015459	potassium channel regulator activity	2/27	0.91	0.62
GO:0004957	prostaglandin E receptor activity	2/5	0.89	0.01
GO:0005545	1-phosphatidylinositol binding	2/20	0.88	0.64
GO:0004708	MAP kinase kinase activity	2/15	0.85	0.18
GO:0030127	COPII vesicle coat	1/9	0.84	0.00
GO:0051642	centrosome localization	2/9	0.83	0.64
GO:0004091	carboxylesterase activity	2/29	0.81	0.61
GO:0005544	calcium-dependent phospholipid binding	3/22	0.79	0.54
GO:0031235	intrinsic to internal side of plasma mem	2/14	0.78	0.18
GO:0005245	voltage-gated calcium channel activity	2/33	0.78	0.23
GO:0012507	ER to Golgi transport vesicle membrane	6/25	0.74	0.51
GO:0001937	negative regulation of endothelial cell	2/23	0.72	0.22
GO:0005044	scavenger receptor activity	2/44	0.70	0.53
GO:0019388	galactose catabolic process	2/5	0.68	0.44
GO:0043027	cysteine-type endopeptidase inhibitor ac	2/21	0.67	0.29
GO:0001725	stress fiber	2/42	0.66	0.11
GO:0051016	barbed-end actin filament capping	2/7	0.64	0.06
GO:0030574	collagen catabolic process	2/22	0.64	0.34
GO:0035085	cilium axoneme	2/47	0.63	0.25
GO:0002020	protease binding	2/50	0.63	0.26
GO:0005801	cis-Golgi network	2/26	0.61	0.26
GO:0003950	NAD+ ADP-ribosyltransferase activity	2/22	0.58	0.45
GO:0060397	JAK-STAT cascade involved in growth horm	3/24	0.57	0.16
GO:0045648	positive regulation of erythrocyte diffe	3/15	0.55	0.15
GO:0043401	steroid hormone mediated signaling pathw	2/14	0.55	0.12
GO:0032747	positive regulation of interleukin-23 pr	2/5	0.52	0.23
GO:0050957	equilibrioception	1/7	0.46	0.01
GO:0006685	sphingomyelin catabolic process	2/5	0.44	0.01
GO:2001238	positive regulation of extrinsic apoptot	2/15	0.42	0.17

GO ID	Term (up to 40 chars)	Statistics	P.MFA	P.MGSA
GO:0001961	positive regulation of cytokine-mediated	2/16	0.40	0.09
GO:0030175	filopodium	4/50	0.37	0.93
GO:0090184	positive regulation of kidney developmen	2/13	0.36	0.06
GO:0050885	neuromuscular process controlling balanc	4/41	0.35	0.90
GO:0032266	phosphatidylinositol-3-phosphate binding	3/17	0.34	0.93
GO:0071300	cellular response to retinoic acid	3/46	0.33	0.09
GO:2000725	regulation of cardiac muscle cell differ	2/7	0.26	0.00
GO:0001916	positive regulation of T cell mediated c	2/11	0.03	0.02

Basic statistics on these terms are provided (number of NPC-associated genes/set size). The rightmost columns give the Markov chain Monte Carlo-computed marginal posterior activation probabilities for these terms using model-based gene set analysis (MGSA) and MFA.

Table 2

Nasopharyngeal carcinoma example: 25 Gene Ontology (GO) terms inferred to be active by model-based gene set analysis (MGSA)

GO ID	Term (up to 40 chars)	Statistics	P.MGSA	P.MFA
GO:0005070	SH3/SH2 adaptor activity	4/49	0.98	1.00
GO:0030175	filopodium	4/50	0.93	0.37
GO:0032266	phosphatidylinositol-3-phosphate binding	3/17	0.93	0.34
GO:0031228	intrinsic to Golgi membrane	3/48	0.91	0.00
GO:0050885	neuromuscular process controlling balance	4/41	0.90	0.35
GO:0033209	tumor necrosis factor-mediated signaling	3/36	0.89	0.00
GO:0005154	epidermal growth factor receptor binding	3/20	0.84	0.92
GO:0030672	synaptic vesicle membrane	4/45	0.77	0.96
GO:0015296	anion:cation symporter activity	3/31	0.75	0.00
GO:0015932	nucleobase-containing compound transmembrane	3/19	0.71	0.00
GO:0008484	sulfuric ester hydrolase activity	2/16	0.67	0.02
GO:0034109	homotypic cell-cell adhesion	3/37	0.66	0.00
GO:0005545	1-phosphatidylinositol binding	2/20	0.64	0.88
GO:0051642	centrosome localization	2/9	0.64	0.83
GO:0021954	central nervous system neuron development	3/43	0.62	0.00
GO:0015459	potassium channel regulator activity	2/27	0.62	0.91
GO:0004091	carboxylesterase activity	2/29	0.61	0.81
GO:0006595	polyamine metabolic process	2/20	0.61	0.00
GO:0005100	Rho GTPase activator activity	4/35	0.58	0.06
GO:0017022	myosin binding	3/28	0.58	0.03
GO:0043392	negative regulation of DNA binding	2/32	0.56	0.07
GO:0005544	calcium-dependent phospholipid binding	3/22	0.54	0.79
GO:0005044	scavenger receptor activity	2/44	0.53	0.70
GO:0004629	phospholipase C activity	2/36	0.51	0.02
GO:0012507	ER to Golgi transport vesicle membrane	6/25	0.51	0.74

Abbreviation: MFA, multifunctional analysis.