Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–17

DOI: 10.1177/0962280217712271
journals.sagepub.com/home/smm
SAGE

Testing for differentially expressed genetic pathways with single-subject N-of-1 data in the presence of inter-gene correlation

A Grant Schissler,[1,2,3,4] Walter W Piegorsch[1,2,3,5] and Yves A Lussier[1,2,3,4]

Abstract

Modern precision medicine increasingly relies on molecular data analytics, wherein development of interpretable single-subject ("N-of-1") signals is a challenging goal. A previously developed global framework, N-of-1-*pathways*, employs single-subject gene expression data to identify differentially expressed gene set pathways in an individual patient. Unfortunately, the limited amount of data within the single-subject, N-of-1 setting makes construction of suitable statistical inferences for identifying differentially expressed gene set pathways difficult, especially when non-trivial inter-gene correlation is present. We propose a method that exploits external information on gene expression correlations to cluster positively co-expressed genes within pathways, then assesses differential expression across the clusters within a pathway. A simulation study illustrates that the cluster-based approach exhibits satisfactory false-positive error control and reasonable power to detect differentially expressed gene set pathways. An example with a single N-of-1 patient's triple negative breast cancer data illustrates use of the methodology.

Keywords

Gene expression data, RNA-seq, gene set, inter-gene correlation, N-of-1, single-subject inference, precision medicine, triple negative breast cancer, affinity propagation clustering, exemplar learning

1 Introduction

1.1 Background

Modern precision medicine increasingly relies on molecular data analytics, where standard approaches accumulate large amounts of molecular data from multiple patients.[1,2] Within this context, an experiential paradigm that has garnered recent attention is development of interpretable single-subject signals to truly focus on the individual patient.[3,4] A novel approach to generating single-subject information is known as the *N-of-1* trial,[5] where the individual patient is the sole source/unit of observation and any statistical descriptions and inferences are intended to relate only to that patient.

While the broad N-of-1 strategy possesses obvious impact for advancing personalized medicine, in certain settings, the approach can exhibit greatly expanded capabilities. Consider, e.g., a molecular N-of-1 analysis where data on a gene's messenger RNA (mRNA) expression are sampled from a patient's diseased organ or tissue. Rather than study expression levels across the patient's entire transcriptome *en masse*, we previously proposed a framework, called N-of-1-*pathways*, that applies the N-of-1 strategy to pertinent gene pathways, i.e. to preassigned collections of gene sets within which the genes are assumed to have associated mechanisms or functions.[6–8] Gene set (pathway) membership is typically defined using curated knowledgebases; here, we employ the gene ontology (GO)[9] knowledgebase and its corresponding GO-biological processes (GO-BP) gene sets.

[1]Interdisciplinary Program in Statistics, The University of Arizona, Tucson, AZ, USA
[2]Center for Biomedical Informatics and Biostatistics (CB2), The University of Arizona, Tucson, AZ, USA
[3]BIO5 Institute, The University of Arizona, Tucson, AZ, USA
[4]Department of Medicine, The University of Arizona, Tucson, AZ, USA
[5]Department of Mathematics, The University of Arizona, Tucson, AZ, USA

Corresponding author:
A Grant Schissler, BIO5 Institute, 1657 E Helen St, The University of Arizona, Tucson, AZ 85721, USA.
Email: grantschissler@email.arizona.edu

Table 1. Paired mRNA ($\log_2$) expression data in pathway GO:0003071 (renal system process involved in regulation of systemic arterial blood pressure), from a breast cancer patient.

Gene	Case/cancer expression	Normal tissue expression	Difference
CYP4A11	0.00	3.71	−3.71
AGTR1	6.13	7.86	−1.73
OR51E2	2.90	1.54	1.36
CYP11B2	0.00	0.00	0.00
PTPRO	3.72	6.22	−2.50
CYP4F2	0.00	0.40	−0.40
AGT	8.40	7.89	0.52
PCSK5	6.68	6.92	−0.25
PDGFB	8.82	9.56	−0.74
F2RL1	8.25	7.91	0.35
EMP2	12.54	12.50	0.04
GAS6	11.89	12.35	−0.47
CYP4F12	1.43	4.72	−3.29
F2R	9.15	9.77	−0.61
SERPINF2	6.38	9.57	−3.19

Note: See text for details.

The aim in our previous works[6–8] was to improve and enhance mechanism-anchored, single-subject, gene expression analysis. In practice, the tactic assumes that the mRNA expressions are collected across all pathways of interest from a single patient under two different paired conditions—e.g. baseline and case, unaffected tissue vs. tumor tissue, before and after treatment, etc. The goal is then to quantify and test for differential pathway expression for that patient using the data derived under the two paired transcriptomes from those different conditions. The framework represents a substantial opportunity for a clinically actionable and economically efficient approach to personalized medicine.[10]

1.2 Breast cancer example

To fix the concepts and also motivate our methodological development, consider the following example. Table 1 gives mRNA expression data from a female subject exhibiting triple negative breast cancer, or TNBC. (TNBC is often more difficult to treat than other breast cancers and thus demands a poorer prognosis; it also disproportionately affects minorities.[11]) The table presents matched normal-tumor pairs of mRNA expression outcomes from a sample of the patient's healthy breast tissue (left column) and of her cancerous breast tissue (middle column) in a specific gene pathway, renal system process involved in regulation of systemic arterial blood pressure (GO:0003071), made up of 15 individual genes. The data were taken from a large collection of RNA-seq data sets obtained from the Cancer Genome Atlas (TCGA; see Acknowledgments). The collection yielded 20,051 mRNA counts, as mapped to HUGO gene symbols[12] for both normal and tumor samples derived from this patient. Following standard practice, a stabilizing transformation was applied by adding 1 to each normalized count and then taking a base-2 logarithm.

Of interest here is whether or not the paired observations in this pathway exhibit significant differential expression between normal and tumor tissues *in this single patient*. It is of further interest to also repeat the calculations across the larger collection of all pathways under study. (A total of 3411 pathways are available with this patient's data; see Section 4.1. As noted above, we define the pathways using the GO[9] knowledgebase via GO-BP gene sets.) By isolating significant, differentially expressed pathways (DEPs) for this patient, the N-of-1-*pathways* approach offers a glimpse into the individual, dysregulated, cellular mechanisms between her tumor and normal breast tissue control.

1.3 Identifying DEPs

Generically, for this N-of-1-*pathways* setting, we assume the matched-pair data are collected as expression levels from the g^{th} gene ($g = 1, \ldots, G$) in some pre-specified pathway with G genes. Denote each gene's ($\log_2$-

transformed) expression under the baseline condition as B_g and its matched case expression as C_g. The paired differences become $D_g = C_g - B_g$, with sample mean $\bar{D}$ and standard error $s_{\bar{D}} = s_D/\sqrt{G}$, where s_D is the sample standard deviation of the D_gs.

The sample mean $\bar{D}$ of the difference in log-expression between case and baseline conditions is a natural statistic to use for quantifying any observed dysregulation in a given pathway. We have found that for sufficiently large pathways, the sampling distribution of $\bar{D}$ presents as roughly symmetric and bell-shaped, consistent with standard central limit results on sums of random variables.[13] Thus, one might consider testing for no difference between baseline and case conditions using the differences D_g and applying the usual paired, $G - 1$ d.f. t-test[14] via the test statistic $t = \bar{D}/(s_D/\sqrt{G})$. If questions exist regarding the t-test's reliance on normal (Gaussian) distribution sampling, appeal can be made instead to its well-established nonparametric analog, the Wilcoxon signed-rank test for two-sample matched-pairs data.[14] Applied over many different, preselected pathways, corrections for multiple false discoveries to these simple per-pathway inferences may also be included.[15]

Additional complexities occur with matched-pair RNA-seq N-of-1 data, however. It is important to recognize that expressions across genes within a given pathway are likely to be both heterogeneous and, more critically, correlated. When testing for differences between the two conditions, inter-gene heterogeneity may not detrimentally affect the null distribution of the test statistic; however, experience with cohort-based gene set testing has shown that inter-gene correlation is non-trivial and should be accounted for in order to maintain the nominal operating characteristics of any inferential procedure.[16,17] Essentially, both the naïve t-test and the signed-rank test assume independence among the pathway's gene expressions and as we will see below, in the presence of non-zero inter-gene correlation, they generally fail to contain the test's false-positive error rate at nominal levels.

Indeed, an unanswered problem when testing for DEPs with single-patient N-of-1 data is how to incorporate non-zero inter-gene correlation(s) from a single N-of-1 sample of G gene expressions. Herein, we propose use of external information to identify clusters of correlated genes within a pathway, and then manipulate that information to build a correlation-adjusted, cluster-based test statistic for assessing differential N-of-1 expression in that pathway. (Mention of external information naturally leads one to consider some form of hierarchical Bayesian model. While we do not dismiss this possibility, it is not our goal in the methods we present below to employ a Bayesian approach. Our attention focuses instead on developing frequentist strategies for the DEP testing scenario.) We begin in Section 2 with a description of our model and clustering strategy, along with a suggested algorithm for its implementation. Section 3 follows with a short simulation study of the method's operating characteristics, with focus on its performance in the presence of inter-gene correlation. Section 4 illustrates the approach by returning to the TNBC data from Section 1.2, while Section 5 ends with an overview and discussion. Note that all calculations we present below are performed in the **R** statistical programming environment.[18]

2 Methodological strategy and clustering algorithm

Our motivating goal is to identify DEPs within a single N-of-1 subject while recognizing that the genes' expressions within a pathway are likely correlated. We restrict attention in this section to a single target pathway, studied under two paired conditions—generically listed as "Case" and "Baseline"—as is typical in N-of-1-*pathways* applications. (Consideration of more than two conditions is certainly possible, but we have not seen any instances of such in practice.)

Our experience indicates that without some form of external information on the within-pathway correlation(s), it is difficult to develop a test of differential pathway expression for a single, N-of-1 sample. Incorporating external gene expression data in testing procedures for pathway dysregulation in a single sample is rare, but not unexplored.[19] To address the issue, we appeal to the growing amount of transcriptomic information being uploaded to modern data storehouses/knowledgebases, and in particular aim to extract from these sources relevant biological-context input to aid in the test's construction. Our approach has a bioinformatic flavor, and it attempts to turn the problem on its head: rather than allow the inter-gene correlations to stymie DEP assessment, we use the external correlation information to aggregate non-negatively correlated genes (assumed strictly co-expressed, not anti-expressed) into correlated clusters within a pathway. We propose the following strategy: (a) identify an existing knowledgebase from which all $\binom{G}{2}$ inter-gene correlations can be determined/calculated for the G genes within our target gene set (pathway), (b) apply an existing clustering algorithm to group

the G genes into clusters based on the external inter-gene correlation data, then (c) incorporate the derived cluster information into a test for differential expression. The latter effort is facilitated by a useful cluster-adjusted t-test given by Williams.[20] The following sub-sections provide details for each step.

2.1 Correlation-based clustering of genes within a pathway

To define gene clusters within a pathway, we devised a two-stage algorithm: (i) estimate inter-gene correlation via independent source(s), then (ii) conduct unsupervised clustering of those genes using the external correlations to define gene-gene similarity.

In the first stage, we identify biological context-relevant mRNA expression data from some external knowledgebase(s). Many different possibilities exist and users can choose any external database that suits their needs. Some caution may be necessary; however, many sources provide a wide selection of gene-pair correlations, derived from multiple inputs. When these are collected to define inter-gene correlations for the G genes within a specific pathway, the consequent correlation matrix need not be positive definite, as the original observations likely did not come from a single, coherent sample. This can lead to computational (and interpretational) problems. To avoid this concern, we recommend use where possible of knowledgebases that contain multiple gene-expression samples from independent individuals. We give an example of such in Section 4.2.

Thus, suppose that mRNA expression measurements from P independent patients are taken on the G genes under study. Form the corresponding matrix of ($\log_2$-transformed) gene expression data $\mathbf{X}_{G\times P} = \{x_{ih}\}$ and use it to compute the usual Pearson product-moment correlation coefficients

$$r_{ih} = \frac{1}{P-1}\sum_{\ell=1}^{P}\left(\frac{x_{i\ell}-\bar{x}_i}{s_i}\right)\left(\frac{x_{h\ell}-\bar{x}_h}{s_h}\right) \tag{1}$$

for each $(i,h)^{\text{th}}$ pair of the G genes ($i \neq h$), where $\bar{x}_i$, s_i, $\bar{x}_h$, and s_h are the sample means and standard deviations across patients for the i^{th} and h^{th} genes, respectively. For convention, we define any gene with zero standard deviation (likely to be caused by a gene with null expression across all samples) as uncorrelated with every other gene.

2.2 Gene clustering via affinity propagation

We employ the information from the externally derived inter-gene correlations r_{ih} to define a set of gene clusters for our target pathway. To perform the clustering, we favor use of affinity propagation (AP)[21] although this warrants further discussion, see Section 5.2. AP clustering exhibits similarities with the well-known k-means clustering approach: it forms clusters around multiple center points, called *exemplars*, that are assessed for cluster similarity in an iterative fashion; Frey and Dueck[21] gave full details. Our strategy shares features with a widely used framework known as correlation network analysis, whose applications in biology include clustering of genes into densely connected sets.[22]

To cluster via AP, we require a *similarity metric* $s(g_i, g_h)$ to define the suitability of a data point (gene) g_h to serve as the cluster exemplar for another data point (gene) g_i ($h \neq i$). Since our goal is to cluster on correlations, we use $s(g_i, g_h) = r_{ih}$, i.e. the recorded correlation between genes g_i and g_h in the external knowledgebase from Section 2.1. Rather than pre-specify the number of clusters, say, m, AP allows for as many (or as few) clusters as the similarity values encourage: higher input similarities increase the likelihood that a gene serves as a cluster exemplar. To add a level of performance stability, however, we recommend that each final cluster should contain at least four genes, unless AP requires smaller cluster sizes in order to converge. We impose that requirement here.

We enlist the **R** package *apcluster*[23] to implement the AP algorithm. We generally accept most of the package defaults, although we manipulate one selected "input preference" parameter, q: assuming all genes are equally valid candidates for the cluster exemplars, the program uses the q^{th} sample similarity quantile as the shared input preference that a gene is chosen as such an exemplar. Informally speaking, this input preference is the propensity of a gene to become an exemplar for itself, i.e. $q = s(g_i, g_i)$. (See the *apcluster* documentation at https://cran.r-project.org/package=apcluster for further details.) In effect, q acts as a tuning parameter that affects the number of clusters in the final AP solution: $q = 0$ tends to produce fewer numbers of clusters, while $q = 1$ tends towards increasing numbers of clusters. The *apcluster* default is the median at $q = \frac{1}{2}$.

To select the tuning parameter q, we appeal to the external correlation data. We begin with a definition for the separation distance between any two points (genes), where greater "distance" implies poorer rationale for assigning two genes to the same cluster. Here, we take the distance metric as simple Pearson distance[24] scaled to the unit interval: $\Delta(g_i, g_h) = \frac{1}{2}(1 - r_{ih})$. For a given input q, suppose AP has converged to a solution of m_q clusters, with $n_j \geq 4$ genes assigned to each j^{th} cluster among the $G = \sum_{j=1}^{m_q} n_j$ genes in the target pathway. Denote the corresponding collection of clusters as $\{\mathcal{C}_1, \ldots, \mathcal{C}_{m_q}\}$. To keep the calculations manageable, we vary q over some range of values such as $q = 0, 0.05, 0.10, \ldots, 1.0$, and borrow from methods of partitioned cluster analysis[25]: first, calculate the cumulative within-cluster variation

$$\omega_q(\mathcal{C}_1, \ldots, \mathcal{C}_{m_q}) = \sum_{j=1}^{m_q} \sum_{\substack{i \in \mathcal{C}_j \\ h \in \mathcal{C}_j}} \Delta(g_i, g_h) \tag{2}$$

and then the corresponding between-cluster variation

$$\beta_q(\mathcal{C}_1, \ldots, \mathcal{C}_{m_q}) = \sum_{j=1}^{m_q} \sum_{\substack{i \in \mathcal{C}_j \\ h \notin \mathcal{C}_j}} \Delta(g_i, g_h) \tag{3}$$

A popular measure for exploring the quality of different clustering solutions is the "pseudo-F" statistic[26]

$$F_q(\mathcal{C}_1, \ldots, \mathcal{C}_{m_q}) = \frac{(\tau_q - \omega_q)/(m_q - 1)}{\omega_q/(G - m_q)} \tag{4}$$

where $\tau_q = \omega_q + \beta_q$. Since it mimics the F-ratio from an analysis of variance, large values of equation (4) suggest higher pertinence for the given cluster assignment, or translated here, for the input value of the corresponding q. Thus, we maximize equation (4) to help select an operable value for q.

With all these components in place, our correlation-based clustering algorithm takes the following steps:

Step 1. Set $q = 0$, $q_{\text{step}} = 0.05$, and define `min.clust.size` $= 4$.
Step 2. Cluster the G genes within the target pathway using AP, supplying the similarity matrix of correlations, $\mathbf{R} = \{r_{\text{ih}}\}$, and current input preference q. To begin, start with $q = 0$.
Step 3. Compute the pseudo-F statistic from equation (4) for the given cluster assignment.
Step 4. Check that the `min.clust.size` restriction was met. If so, increment q by $q_{\text{step}} = 0.05$ and repeat Steps 2–4. Find the largest value of F_q and select that cluster solution; let m be the corresponding number of clusters.
Step 5. If the `min.clust.size` restriction fails, or if $q = 1$, then select the cluster solution from the largest value of F_q where `min.clust.size` is met or $q = 0$. Let m be the corresponding number of clusters.
Step 6. Return the gene list $\mathcal{G}$ with cluster assignments $\{\mathcal{C}_1, \ldots, \mathcal{C}_m\}$.

2.3 Assessing pathway differential expression

Given the external, correlation-based cluster assignments from Section 2.2, we next return to the current N-of-1 data set of gene expressions from our single subject. Let $D_{jk} = C_{jk} - B_{jk}$ be the k^{th} gene-wise case-to-baseline difference ($k = 1, 2, \ldots, n_j$) in the j^{th} cluster ($j = 1, 2, \ldots, m$) from a total of $G = \sum_j n_j$ genes in the target pathway. Also, let $D_{++} = \sum_j \sum_k D_{jk}$ be the grand sum, $D_{j+} = \sum_k D_{jk}$ be the j^{th} cluster sum, $\bar{D} = \sum_j D_{j+}/m$ be the mean difference across clusters, and $\bar{\bar{D}} = D_{++}/G$ be the grand mean. Lastly, take S_D^2 as the sample variance $\frac{1}{m-1}\sum_j (D_{j+} - \bar{D})^2$ across clusters. (If desired, these expressions translate in a straightforward fashion when pre-determined, cluster-specific weights, w_j, are available to differentially weight the clusters, e.g. use $\bar{D}_w = \sum_j w_j D_{j+} / \sum_j w_j$, etc.)

We take the parameter $\mu_{DEP} = E(\bar{\bar{D}})$ to represent the true magnitude of pathway differential expression. The statistical hypotheses of interest are

$$\begin{aligned} H_0 &: \quad \mu_{DEP} = 0 \\ H_a &: \quad \mu_{DEP} \neq 0 \end{aligned} \tag{5}$$

Our aim is to employ the externally derived cluster assignments to construct a test statistic from which we can assess if our N-of-1 data can identify the target pathway as a significant DEP. The grand mean $\overline{\overline{D}}$ serves as an obvious starting point, since it measures the extent of differential expression in the pathway. We prefer, however, to base our test statistic's construction on the mean cluster difference $\bar{D}$: although less interpretable than $\overline{\overline{D}}$, operating with $\bar{D}$ provides for some useful computational simplification, e.g. testing H_0 in equation (5) reduces to calculations only involving $\bar{D}$.

To estimate $\text{Var}[\bar{D}]$, we turn to a robust, unbiased variance estimator offered by Williams[20] for cluster-correlated data. Given externally defined clusters $\{\mathcal{C}_1, \ldots, \mathcal{C}_m\}$, Williams showed that under our notation, the variance of the grand sum can be estimated as

$$\widehat{\text{Var}}[D_{++}] = \frac{m}{m-1}\sum_{j=1}^{m}(D_j - \bar{D})^2 \tag{6}$$

where $\bar{D} = \sum_j D_{j+}/m$ was defined above. Notice that equation (6) is simply $\widehat{\text{Var}}[D_{++}] = mS_D^2$. With it, the standard error of $\bar{D}$ becomes $S_D/\sqrt{m}$.

Now, suppose under H_0 that the within-pathway cluster sums are distributed as $D_{j+} \sim N(0, \sigma^2)$, with variances σ^2 taken as constant across clusters. The use of a stabilizing log-transform on the original read counts supports the homogeneous-variance normal/Gaussian assumption here, at least approximately, when clusters are sufficiently large and roughly equal in size.[27] The hyperbolic arc-sine can provide an even-more-stable transformation with read counts such as these, although the improved stability it provides may not be sufficient to support its additional complexity.[28] We do not generally see it applied to RNA-seq read counts in practice.

Conditional on the external definitions for the clusters, it is straightforward then to show that the t-statistic

$$T = \frac{\bar{D}}{S_D/\sqrt{m}} \tag{7}$$

is distributed as $T \sim t(m-1)$ under H_0. A (two-sided) p-value for testing H_0 against H_a is simply $p = 2\text{P}[t(m-1) \geq |T|]$. Reject H_0 and report the target pathway as differentially expressed when p is smaller than some pre-specified α-level.

3 Simulation study

3.1 Simulation design and generation

The test statistic for DEP assessment constructed in Section 2.3 relies on approximating arguments to be valid with N-of-1-*pathways* gene expression data. To investigate the performance of our test methodology in practice, we conducted a series of Monte Carlo evaluations.

We examined how four different inputs affect the test's operating characteristics: (1) pathway size; (2) proportion, say, π, of DEGs within the pathway; (3) the fold-change increase, say, ψ, in mean differential expression from the Case to the Baseline sample; and (4) the inter-gene dependence structure as quantified by a within-pathway correlation matrix $\mathbf{R}$. We explain each of these components in detail below.

Our focal setting for building the simulation configurations is the pathway size G. We studied five different sizes: $G = 15, 30, 50, 100, 200, 400$. In order to represent true practical settings, we associated each different pathway size with an existing pathway annotation from the GO-BP knowledgebase[9] (downloaded 1 September 2015). We focused on matched-pair breast cancer outcomes, following on our example from Section 1.2. We were able to identify 110 patients with RNA-sequencing from TCGA's breast adenocarcinoma (BRCA) database and modified existing GO-BP gene set definitions to include gene symbols that were measured in their BRCA data sets (20,501 genes for the 110 subjects' matched pairs; downloaded 27 Jul 2016). This resulted in 634 distinct mRNAs being filtered out from GO-BP (11,530 down to 10,896 genes). Next, we randomly selected a pathway meeting these requirements at each of the six sample sizes. Table 2 displays the selected pathways and five-number summary statistics for the $\binom{G}{2}$ inter-gene correlations within each. We note from the summary data that these pathways exhibit a broad range of inter-gene correlations, centered at or near zero and often with a slight positive skew. Indeed, every correlation matrix $\mathbf{R}$ for the six pathways departs significantly from a diagonal matrix: using a nonparametric sphericity test from Chen et al.,[29] none of the six corresponding (pointwise) p-values from the test is larger than 10^{-6}. Thus, we feel confident that each chosen pathway exhibits non-trivial inter-gene correlation.

Table 2. Description of the six pathways selected for our simulation study.

Gene set identifier	Description	G	m	r_{min}	r_{Q1}	r_{Q2}	$\bar{r}$	r_{Q3}	r_{max}
GO:0060350	Endochondral bone morphogenesis	15	2	−0.66	−0.14	−0.02	0.0030	0.120	0.60
GO:0016925	Protein sumoylation	30	6	−0.75	−0.18	−0.00	0.0269	0.259	0.89
GO:0048565	Digestive tract development	50	4	−0.73	−0.15	−0.01	0.0175	0.174	0.87
GO:0045185	Maintenance of protein location	100	9	−0.81	−0.20	0.00	0.0179	0.228	0.95
GO:0002683	Negative regulation of immune system process	200	12	−0.79	−0.12	0.00	0.0310	0.169	0.99
GO:0002520	Immune system development	400	28	−0.86	−0.14	0.00	0.0220	0.172	0.98

Note: G is the total number of genes for which the TCGA BRCA data have measurements for each pathway; m is the number of clusters in each particular pathway determined by the correlation clustering algorithm (see text). The remaining columns contain the five-number summary and mean for the inter-gene correlations, r, observed in TCGA BRCA normal tissue RNA-seq data on an available sample of 110 patients.

Table 2 also contains a column reporting the number of separate clusters, m, for each pathway determined by our correlation clustering algorithm from Section 2.2. Not surprisingly, the number of clusters is low for small pathway sizes and increases as G grows very large. We give the complete list of gene cluster assignments in a supplementary document: see Supplemental Table S1. Supplemental Figure S1 also gives histograms for each set of $\binom{G}{2}$ correlations.

To model gene expression in the g^{th} gene in any of the six selected pathways ($g = 1, \ldots, G$), we followed established practice[30,31] and assumed a negative binomial (NB) distribution for the frequency of reads, Y_g, with mean expression μ_g and dispersion parameter δ_g such that $\text{Var}[Y_g] = \mu_g + \delta_g \mu_g^2$. To determine rough estimates of both parameters for each g^{th} gene, we returned to the TCGA BRCA database and retrieved RNA-seq read counts on normal breast tissue samples from the same 110 independent patients. For each gene in the specified pathways from Table 2, we found corresponding estimates $\hat{\mu}_g$ and $\hat{\delta}_g$ via the method of moments. (If $\hat{\delta}_g$ indicated under-dispersion compared to a Poisson random variable, we conservatively deemed the gene's read count to be Poisson distributed with mean $\hat{\mu}_g$.) These values defined the NB distribution for Baseline responses of gene g in our simulations.

To generate the corresponding, affected, Case responses in gene g, we imposed a ψ-fold increase in its mean response and assigned that gene a differentially expressed mean of $\psi\hat{\mu}_g$. We varied this fold-change over the range $\psi = 1.5, 2, 4$. We combined the fold-change setting with a "DEG proportion" setting in order to study how changes in a pathway's differential expression are identified by our proposed methodology. That is, we allowed the proportion of DEGs in a pathway to vary over the range $\pi = 0, 0.3, 0.6, 0.9$. When $\pi = 0$, no DEGs were present, and both the Baseline and Case mean responses were taken as $\hat{\mu}_g$. This studied the false-positive error rate of our cluster-based test procedure. When $\pi > 0$, however, that proportion of the G genes in the pathway was assigned a differential mean response equal to $\psi\hat{\mu}_g$, as described above. (We rounded πG to the nearest integer.) This latter case was designed to study the power, our test procedure exhibits to identify DEPs.

Lastly, we studied the effects of inter-gene dependence by manipulating the inter-gene correlations. This was quantified by a within-pathway correlation matrix $\mathbf{R}$ consisting of all pairwise correlations r_{ih} between genes g_i and g_h in the given pathway. We used the externally determined correlations calculated from the TCGA database (summarized in Table 2) and constructed a $G \times G$ $\mathbf{R}$ matrix for the pathway's G genes using the corresponding r_{ih} values. For each pathway, we considered three forms for the correlation structure:

(i) Cluster-correlated data that imposes the AP cluster assignments from Section 2.2, such that pairs of genes within each cluster receive their corresponding correlation of r_{ih} from the full set of pairwise correlations, and genes between clusters (within a pathway) receive a correlation of zero. This is intended to reproduce the clustered correlation structure assumed by our testing methodology. We refer to this as "block" correlation, since $\mathbf{R}$ takes on a block-diagonal form.
(ii) Cluster-correlated data where gene pairings receive their corresponding correlations of r_{ih} irrespective of their imposed, within-pathway, cluster assignments. This is intended to study our method's robustness when the cluster assignments are not correctly specified. We refer to this as the "all" correlation case.
(iii) An "independence" assumption among all genes, such that $\mathbf{R}$ equals the $G \times G$ identity matrix $\mathbf{I}$. This studies the operating characteristics of our method under the supposition of no within-pathway correlation.

Correlated RNA-seq gene expressions were then simulated under our negative binomial model using the correlations in **R**. The negative binomial read counts were simulated via **R**'s `rnbinom` pseudo-random variate generator. We employed copulas[32,33] in order to 'tie' together the marginal negative binomial distributions and form a multivariate construction. We found that a few simulation replicates could result in poor performance in the implementation[33] if generated serially, so we conducted all the simulations at once. We set the number of replicate, simulated, N-of-1 data sets to 2000, and first generated an entire non-DEG simulated data collection of size $G \times 2000$. Next, an entire DEG simulated data collection of size $G \times 2000$ was generated with every marginal gene mean μ multiplied by the current fold change ψ. The two data collections were merged based on the selected DEG genes (see above) to form the complete simulated data set for the particular configuration with the appropriate proportion of DEGs, π. This ensured that the desired correlation structures were properly represented.

Combining the six different pathway sizes with three types of fold change, four different proportions of DEGs/pathways, and three different correlation structures produced 216 separate simulation configurations for study. (When $\pi = 0$ the value of ψ is irrelevant, but for completeness' sake, we ran separate simulations at each value of $\psi = 1.5, 2, 4$ even for $\pi = 0$. We expect the results to show essentially similar performance in all three cases.) At each configuration, the simulated N-of-1 samples represented paired pseudo-random NB read counts $Y_{Bg} \sim \text{NB}(\hat{\mu}_g, \hat{\delta}_g)$ and $Y_{Cg} \sim \text{NB}(\psi\hat{\mu}_g, \hat{\delta}_g)$ for the specified pathway of size G. To impose **R**, we used a standard Gaussian copula given the clustering solution for the specified pathway (Supplemental Table S1) and the simulation correlation setting (''block'', ''all'', or ''independent''). Then we transformed the individual NB counts, Y, to the stabilized values $X = \log_2(Y + 1)$. Lastly, we tested for differential expression using the AP-based clustered-T statistic in equation (7). For comparison purposes, we also applied the Wilcoxon signed-rank test and the naïve t-test discussed in Section 1.3 to assess their operating performances.

For each test statistic, we set our pointwise false-positive rate to $\alpha = 0.05$ and recorded the proportion of cases where the test rejected the null hypothesis of no differential gene expression. Notice that with 2000 simulations per configuration, the approximate standard error of our empirical Monte Carlo rejection rates at the nominal 5% level is $\sqrt{(0.05)(0.95)/2000} = 0.005$ and it never exceeds $\sqrt{(0.5)(0.5)/2000} = 0.011$.

3.2 Simulation results

3.2.1 False-positive rates

Empirical false-positive error rates corresponding to $\pi = 0$ from our simulations are graphed in Figure 1. The figure presents results for $\psi = 1.5$ at $\alpha = 0.05$. We note that results with $\psi = 2, 4$ were essentially similar, as was expected. The display cross-classifies the results by correlation structure **R** and pathway size G. We see that for the independence case ($\mathbf{R} = \mathbf{I}$; far left column of graphs), all three methods control error rates at or slightly below the nominal 5% level within Monte Carlo sampling variability. Since we argue that genes within pathways should not be assumed independent, however, results from this independence case are primarily valid only as a proof of concept.

For the more practical case with non-zero inter-gene correlations, Figure 1 (right two columns) shows that our cluster-based approach using equation (7) exhibits reasonable false-positive error performance. Under the cluster-correlated (''block'') structure for which it is designed, its empirical error rates are essentially at or slightly below the nominal rate for all pathway sizes. Interestingly, under the ''all'' correlation structure, our cluster-based test also shows good performance, either maintaining the nominal false-positive rate or becoming slightly conservative. We view this as a form of robustness to misspecified clustering, at least of the form we impose in our ''all'' correlation setting. The feature might be explained by the propensity of positively correlated genes to cluster together in the clustering algorithm (Section 2.2), which would tend to drive inter-cluster correlations to be negative (depending on the assumed correlation distribution) rather than close to zero. This could then lead to over-estimation of the standard errors and produce slight drops in sensitivity in the test statistic.

By contrast, both the signed-rank and naïve t-test exhibit substandard false-positive error performance when the data incorporate either form of inter-gene correlation. Error rates run upwards of 60% in some cases, although for the $G = 15$ pathway under the all-correlation structure, all methods appear to control false-positive error. This may be due to that pathway's relatively symmetric and tighter pattern of inter-gene correlations (see Table 2). The result is not reproduced with the block correlation structure, however, so we do not place much value in this indication.

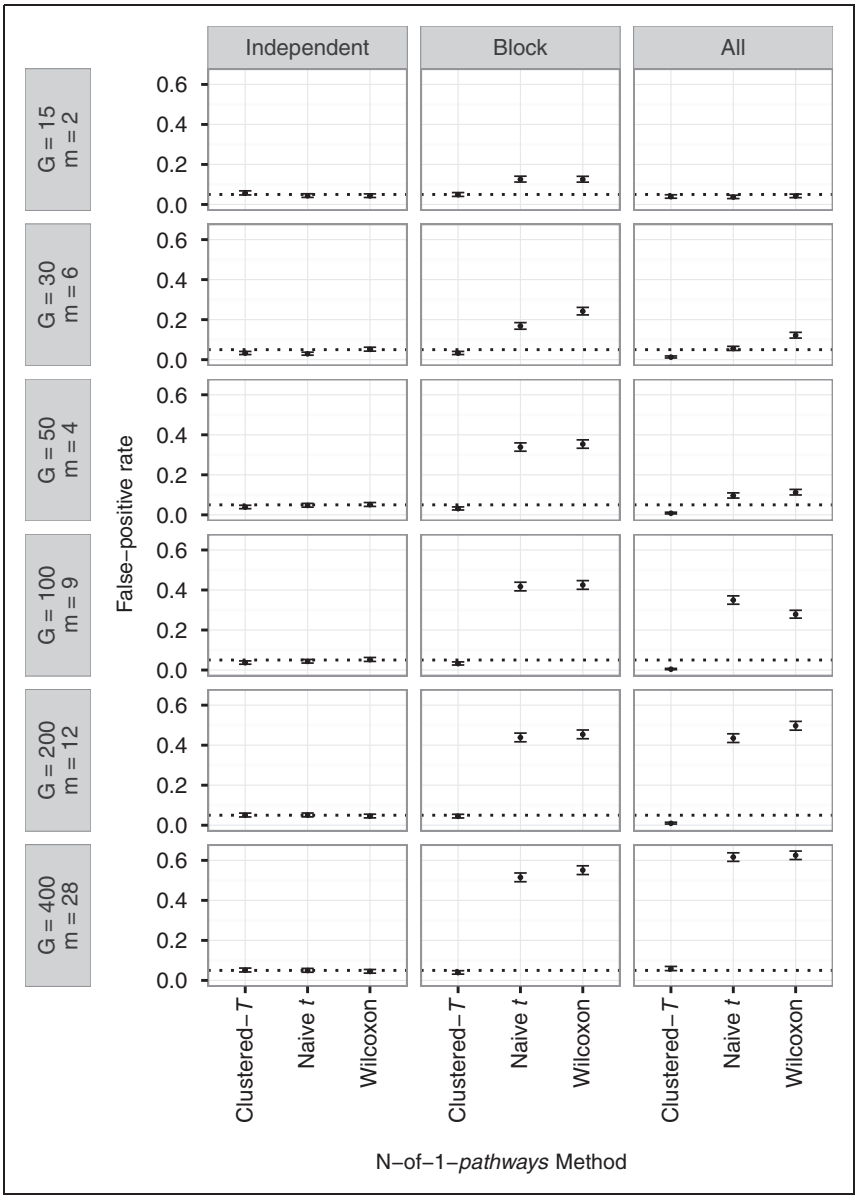

Figure 1. Empirical false-positive rates (dots) based on 2000 simulated N-of-1-*pathways* data sets for three competing testing procedures (lower horizontal axis: Clustered-T = proposed test, naïve t = standard t test, Wilcoxon = signed-rank test), cross-classified by correlation structure (top: Independent = uncorrelated mRNA expression, Block = cluster-correlated expression, All = unconstrained inter-gene correlation) and pathway size G (left). The corresponding cluster numbers, m, for each pathway are also listed; see Table 2. Nominal significance level is set to $\alpha = 0.05$ (dotted horizontal lines). Results reported for $\psi = 1.5$ (see text). Horizontal bars are pointwise 95% Agresti–Coull[34] confidence intervals for the underlying false-positive rate based on each set of 2000 simulated samples.

3.2.2 *Simulated power*

Empirical powers, as rejection rates, from our simulations are graphed in Figure 2. The figure presents the results as a function of the DEG proportion π and stratifies power curves by the fold change parameter ψ. (To provide a baseline, the figure also includes the false-positive results at $\pi = 0$.) As in Figure 1, the display cross-classifies the results by correlation structure $\mathbf{R}$ and pathway size G, and the nominal significance level was again set to $\alpha = 0.05$. Only results for our AP-based clustered-T statistic in equation (7) are presented, since the other, independence-based methods were seen in Figure 1 to be substandard.

The patterns in Figure 2 show a trend towards increasing power with increasing DEG proportion π (i.e. greater departure from H_0). The effect is more pronounced as the fold-change increases from $\psi = 1.5$ (dot-dashes) to $\psi = 4$

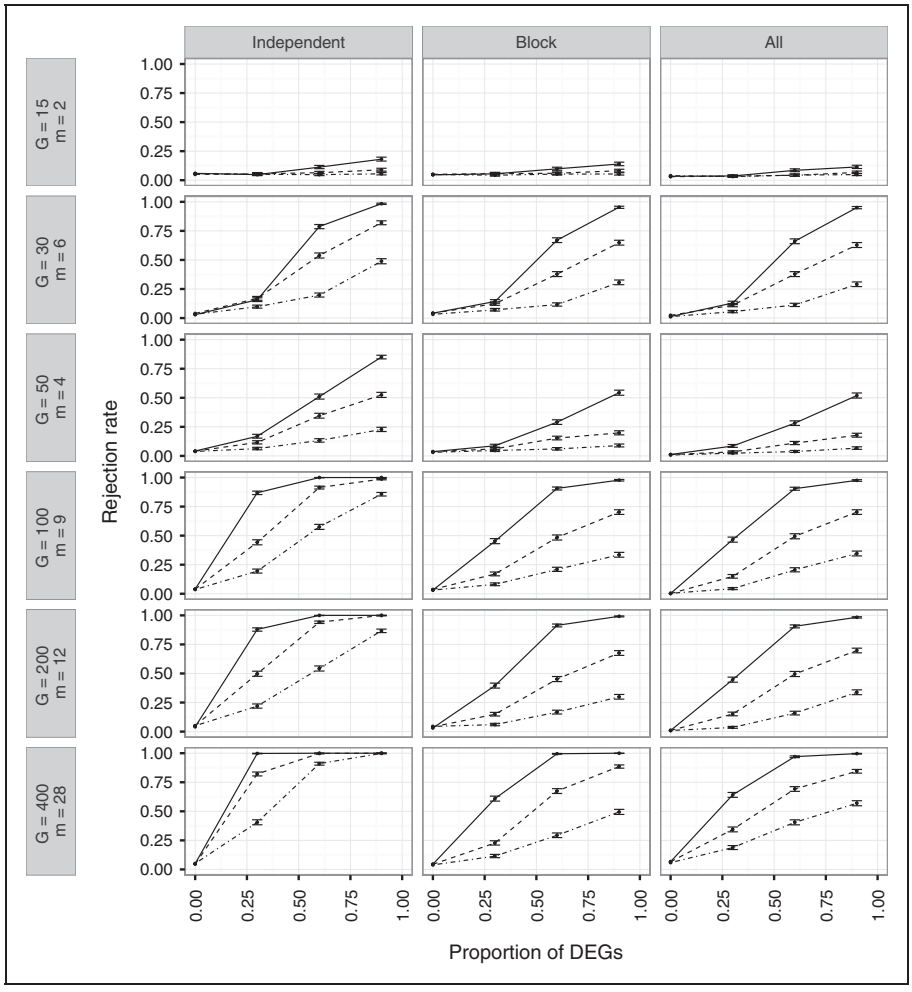

Figure 2. Empirical rejection probabilities ("power") for the AP-based cluster approach using equation (7), based on 2000 simulated N-of-1-*pathways* data sets. Results are presented as a function of DEG proportion π (lower horizontal axis). Displays are cross-classified by correlation structure (top: Independent = uncorrelated mRNA expression, Block = cluster-correlated expression, All = unconstrained inter-gene correlation) and pathway size G (left). The corresponding cluster numbers, m, for each pathway are also listed; see Table 2. Simulated fold change is indicated by line styling: $\psi = 4$ (solid lines), $\psi = 2$ (dashes), and $\psi = 1.5$ (dot-dashes). Nominal significance level is set to $\alpha = 0.05$. Horizontal bars are pointwise 95% Agresti–Coull[34] confidence intervals for the underlying rejection rate based on each set of 2000 simulated samples.

(solid lines). Also, faster rises in power are indicated for the independence case ($\mathbf{R} = \mathbf{I}$; far left column of graphs). In all instances, these phenomena are not unexpected. Interestingly, for the smallest pathway at $G = 15$, the curves were all visibly dampened and exhibited limited power. We attribute this more to the small number of clusters ($m = 2$) than to the small pathway size, since the cluster-based t-test then operates with only $m - 1 = 1$ d.f. The two features are related, of course, since it is difficult to generate a large number of clusters with only 15 separate genes. Indeed, for the larger pathways with $G \geq 100$, the power curves rise far more dramatically; less so for the smaller pathways with fewer clusters. Supplementary Tables S2 to S3 contain full numeric summaries of the rejection rates associated with this simulation study.

In general, these results suggest that the AP-based clustered-T statistic in equation (7) exhibits good false-positive error control and reasonable power, at least under the settings chosen for these simulations.

4 Breast cancer example revisited

To illustrate use of our cluster-based methodology, we return to the breast cancer example from Section 1.2. Recall therein that a female patient exhibiting triple negative breast cancer (TNBC) provided matched samples from both

her healthy breast tissue and her cancerous breast tissue. (The complete data for this patient—identifiable only as "TCGA-A7-A0CE"—were quarried from the Cancer Genome Atlas/TGCA.)

Our goal is not to present a comprehensive investigation for further medical treatment(s) with this particular patient, rather, simply to exhibit how our cluster-based statistics are calculated in practice and to suggest possible avenues for their further use. To start, gene expressions on all available genes for the patient were normalized to account for library size (via transcripts per million) and then transformed via $\log_2(y_g + 1)$ for each g^{th} gene's expression count y_g in both the Baseline (normal tissue) and Case (cancerous tissue) groups. As above, we defined our pathways from a standard gene set knowledgebase, chosen here to be the GO-BP ontology.[9] Recall that example data from one such pathway, GO:0003071 (renal system process involved in regulation of systemic arterial blood pressure), were presented in Table 1. To make the process manageable and results more interpretable, we retained the pathways that had no fewer than 15 and no more than 500 genes annotated and provided a measured RNA-seq data set. This produced 3411 different GO-BP pathways.

4.1 Clustering of gene ontology-biological processes pathways

To construct our inter-gene, correlation-based clusters to correspond with this patient's TNBC data, we returned to the same 110 independent, external subjects (these excluded TCGA-A7-A0CE's expression data) in the TGCA BRCA database employed in our simulation study (Section 3.1). Gene expressions in normal breast tissue from these women were available across the same genes listed in the 3411 pathways chosen above. Within each of these pathways, inter-gene correlations r_{ih} for the expressions were calculated from the 110 samples. We verified that each pathway's associated correlation matrix represented non-trivial inter-gene correlation, using the sphericity test from Chen et al.[29] Only four of the 3411 pathways yielded p-values above 0.05. Since this was such a small number, we chose not to distinguish these four pathways from the rest. We next assigned within-pathway gene clusters to all pathways via our AP clustering algorithm as described in Section 2.2, using the calculated correlations r_{ih}.

The external clustering proceeded in a manageable fashion. Employing *R* without parallelization, the algorithm processed all pathways in slightly under 30 min on a MacBook Pro carrying a 2.5 GHz Intel Core i7 processor and 16 GB of 1600 MHz DDR3 RAM. Note that the external clustering need to be completed only once for a given ontology and biological context/target tissue.

Most pathways yielded numbers of clusters m from the high single-digits to upwards of 30. For five of the pathways, however, we did experience an anomaly: the algorithm converged to a solution with m equal to 1. (That is, it assigned all genes in the pathway into a single cluster.) Since our AP-based clustered-T statistic in equation (7) requires at least $m = 2$ clusters to operate, single-cluster pathways cannot be assessed with our method. Thus, we chose to remove these five pathways from consideration and not report a p-value. (One might instead manually inspect each single-cluster pathway and reason out some other, plausible, $m \geq 2$ clustering assignment. This would require additional, subjective, domain-specific judgment on a case-by-case basis, however.) The five-number summary for m across the remaining 3406 scored pathways was {2, 3, 5, 9, 41} with the average cluster number equaling 6.97. The distribution was strongly skewed to the right.

4.2 Differentially expressed pathways for patient TCGA-A7-A0CE

Conditional on the external cluster assignments, we calculated our AP-based clustered-T statistic from equation (7) and found its pointwise, two-sided p-value, corresponding to a $t(m-1)$ reference distribution, for each of the scored pathways (including the four pathways exhibiting no inter-gene correlation). Pathways could be classified as either differentially expressed (DEP) or non-dysregulated, depending on whether the null hypothesis was rejected or not, respectively. One goal of the N-of-1-*pathways* strategy is to use such information to isolate possible pathways that affect or associate with the patient's disease outcome, i.e. here with TNBC. For example, a non-dysregulated pathway that is a target of a therapeutic drug could be indicative of a patient's poor response to therapy.

For patient TCGA-A7-A0CE, 601 of her scored pathways gave a pointwise, unadjusted p-value at or below 5%. To correct for multiplicity, we further applied a Benjamini–Hochberg[15] false-discovery adjustment. This resulted in 80 DEPs with a false discovery rate (FDR) less than 15%. (Among these was included the GO:0003071 pathway in Table 1.) Table 3 displays the 10 top-identified dysregulated pathways, ordered by their original, cluster-based p-values. Supplemental Table S4 lists all these 80 DEPs.

Table 3. The 10 top-hit (smallest p-value) differentially expressed pathways and corresponding summary statistics for TNBC patient TCGA-A7-A0CE, after application of the AP-based clustered-*T* procedure.

Gene set identifier	Description	$\bar{\bar{D}}$	*T*-statistic	*p*	*G*	*m*
GO:0045785	Positive regulation of cell adhesion	−0.75	−4.92	0.00011	226	19
GO:0032101	Regulation of response to external stimulus	−0.47	−4.42	0.00015	458	28
GO:0070124	Mitochondrial translational initiation	0.28	7.55	0.00028	84	7
GO:0010769	regulation of cell morphogenesis involved in differentiation	−0.51	−4.80	0.00028	168	15
GO:0030155	Regulation of cell adhesion	−0.51	−4.24	0.00037	384	22
GO:1902532	Negative regulation of intracellular signal transduction	−0.42	−4.37	0.00042	283	18
GO:0009306	Protein secretion	−0.65	−4.68	0.00043	272	14
GO:0034142	Toll-like receptor 4 signaling pathway	−0.33	−5.35	0.00046	105	10
GO:0051223	Regulation of protein transport	−0.56	−3.94	0.00047	452	30
GO:0051092	Positive regulation of NF-kappaB transcription factor activity	−0.54	−5.06	0.00049	111	11

Note: See text for details.

Further study of patient TCGA-A7-A0CE's, 80 DEPs may offer insights into disease pathogenesis, progression, and possible therapeutic targets. For example, the dysregulated molecular pathways identified in in Table 3 are known to be associated with breast cancer progression, including cell adhesion,[35] signal transduction,[36] NF-κB,[37] and Toll-like receptors.[38,39] The latter (Toll-like receptor) has particularly promising immunotherapy potential.[40]

5 Discussion

We have introduced a method for scoring and testing differentially expressed pathways (DEPs) from a single-subject, matched-pair collection of gene expression data under the N-of-1-*pathways* paradigm.[6] By focusing on gene sets/pathways, this N-of-1 strategy provides a powerful tool for development of subject-specific precision medicine. Unfortunately, the limited amount of data within the N-of-1, single-subject setting makes construction of suitable statistical inferences difficult. In particular, we illustrated that the presence of inter-gene correlation within a pathway undermines the ability of standard paired-testing methods to control false-positive error. We argue instead for incorporation of external information into the significance test procedure. Toward this end, we propose a novel correlation-based clustering algorithm that employs external gene expression data to aggregate positively co-expressed genes into clusters within pathways. To our knowledge, this is the first single-subject gene set testing procedure that accounts for inter-gene correlation.

In a short simulation study, the method exhibited satisfactory false-positive error control and robustness against certain modeling assumptions, including the form of inter-cluster correlation. The method also powerfully detected DEPs when the number of clusters and fold-change of differentially expressed genes (DEGs) was large. In addition, it is worth mentioning that our cluster-based approach does not require estimation of inter-gene correlations for the single N-of-1 subject being studied, as we show in an example with a TNBC patient.

While the method presents ample potential, there are, of course, some caveats. One potential drawback is the availability of external, context-relevant, gene expression data. In the cancer literature, there are a growing number of openly available patient data sets and cell lines that could be recruited for use. Other biomedical contexts may prove less accommodating, however. Furthermore, we implicitly assume that co-expression of genes within the same tissue is stable across the external database, but further study is warranted to test the validity of this supposition. Depending on the estimated, external, inter-gene correlations and the chosen gene set ontology, some pathways may exhibit resistance to accurate and effective clustering. This could result in small cluster numbers or large, positive, inter-cluster correlation, and potentially lower power under our cluster-based strategy.

5.1 Extensions

With the above limitations in mind, one might consider a number of possible extensions and variations to our cluster-based approach. We propose what is, in effect, a self-contained gene set test,[41] meaning that only genes in the pathway impact the score. A test that employs inter-gene correlations across the entire transcriptome could present a competitive alternative approach when many pathways fail to cluster under a chosen ontology. Indeed,

correlation-based clustering is an active area of research across many disciplines.[42] And, as noted above, future N-of-1 data may present more than two conditions beyond our Case-vs.-Baseline pairing. The statistical details for identifying DEPs under multiple conditions would make for an interesting extension of our cluster-based approach.

It is also natural when considering incorporation of external information to consider some form of Bayesian methodology.[43] The nature and form of any such model would be highly complex, and it is unclear how much external information or subjective input would be required to implement any eventual construct. Still, the richness of the Bayesian paradigm offers many avenues for study and it may provide a useful arena for development with N-of-1, precision-medicine inferences.

5.2 Choice of clustering algorithm

As a referee has noted, selection of the core clustering algorithm in Section 2.2 is a fundamental linchpin for creating our external gene clusters and implementing the test statistic in equation (7). While we recommend the use of AP clustering, many other algorithms could be applied in its place[44] and whether any of these would provide enhanced (or inferior) performance over AP is an open question. Indeed, one would expect results from any reasonable alternative clustering procedure to be at least roughly comparable to those from AP if the clustering pattern is sufficiently strong.

To examine this aspect, we considered an alternative clustering algorithm from one of the more important classes of affinity-based clustering strategies, *spectral clustering*, see von Luxburg[45] for an instructive introduction. We applied von Luxburg 's recommended settings to the spectral clustering (SC) implementation from the *kknn* package[46] in **R**. Then, we simply replaced our calls to the *apcluster* package, and associated support coding, with calls to the *kknn* `specClust` function in our **R** code. We also chose to employ eigengap heuristics specific for the SC approach to select the number of clusters, m, in place of the pseudo-F statistic applied with AP's q input preference parameter. These heuristics find m as that number of clusters with a maximal difference (or 'gap') between adjoining-ordered eigenvalues of the data's graph Laplacian, see von Luxburg[45] for complete details. Lastly, we restricted the algorithm to require, as above, $n_j \geq 4\ \forall j$.

To study the performance of this alternative SC algorithm, we returned to our Monte Carlo evaluations from Section 3. We implemented spectral clustering to redefine the clusters used for constructing the statistic in equation (7) and then applied the procedure at the various parameter configurations and other input settings described in Section 3.1. We recorded the instances of significant pathways identified by the corresponding test. The resulting, empirical, false-positive rates at $\alpha = 0.05$ were generally satisfactory: the average false-positive rate for spectral clustering was 0.0347 across all simulation conditions. By comparison, from Figure 1, the average was 0.0351 for AP clustering. (The complete false-positive results appear in Supplemental Table S2, reference to which corroborates that the SC error rates are comparable to those from AP clustering.) In the corresponding power simulations, AP clustering yielded greater power for some settings, while for others, SC was more powerful with similar performance in the aggregate. A graphic illustrating this similarity is given as Supplemental Figure S2, displayed are the power curves under spectral clustering and those under AP clustering from Figure 2. It is seen that both algorithms perform well as G grows, and that both AP and SC appear to operate in a similar fashion, at least for the limited simulation settings we consider.

To study the effect of clustering-algorithm choice in practice, we returned to the TNBC data from Section 4 and applied spectral clustering as described above in place of AP to define the gene clusters. After application of the clustered-T statistic from equation (7), this produced an alterative ordered list of p-values and corresponding pathways, the results from which indicated reasonably strong similarities between the two clustering algorithms. For example, among the 3351 pathways that produced p-values under both methods—see below—the top-identified dysregulated pathway using AP clustering, positive regulation of cell adhesion (GO:0045785), was also top ranked under SC. At a broader level, 2870 of the 3351 pathways indicated qualitative agreement at the traditional 5% level: 361 give both p_{AP} and p_{SC} below 0.05, while 2509 show p_{AP} and p_{SC} above 0.05.

Digging deeper, however, it is possible to identify important, information-based, genetic overlap between some of the remaining 481 pathways whose p-values remained discordant. To do so, we turned to the concept of information-theoretic similarity (ITS). Briefly, ITS quantifies the degree of semantic similarity between two candidate pathways by comparing the information content encoded in the ontology structure based on the mechanisms annotated to the genes comprising each pathway, see Tao et al.[47] for details. The result is a similarity score between 0 and 1: values closer to 1 indicate greater genetic similarity between the pathways. Ongoing experience has shown that pairs of pathways with ITS scores above 0.7 exhibit high structural

Table 4. Categorization outcome agreement between spectral clustering and AP clustering in the clustered-*T* test of equation (7) when applied to TNBC data from Section 4, as a function of chosen significance level.

Significance	Positive	Negative	ITS	No	No match
Level (%)	Overlaps	Overlaps	Matches	Matches	Percentage
10	684	2116	485	66	$\frac{66}{3351} = 1.97\%$
5	361	2509	396	85	$\frac{85}{3351} = 2.54\%$
1	88	3046	150	67	$\frac{67}{3351} = 2.00\%$

Note: Categories are: (i) positive overlaps (*p*-values both below significance level), (ii) negative overlaps (*p*-values both above significance level), (iii) *p*-value discords with information-theoretic similarity (ITS) match (high informatic similarity), and (iv) *p*-value discords with no ITS match. See text for details.

overlap,[7,48] enough to label the pathways as 'matches' from an informatic perspective. (To provide a sense of the rarity of obtaining an ITS match by chance alone, note that the proportion of all candidate GO-BP pairs whose ITS exceed 0.7 is $35,287/6,056,940 = 0.00583$.) We adopted this criterion and queried whether any discordant pairs of pathways with *p*-values resting on either side of the 5% cutoff might still exhibit ITS matches.

Following the protocol developed by Gardeux et al.,[7] we began by isolating all those pathways identified by AP as significant at the 5% level—there are 601 such pathways for patient TCGA-A7-A0CE. Next, we isolated the analogous list of pathways identified by SC as significant at the 5% level—there are 602 of these. Then we asked, for a given pathway in the AP list do any pathways in the SC list possess an ITS score of 0.7 or higher to produce an ITS match? If any SC pathway did so, we scored the pathway from the AP list as an ITS AP-SC match. If none of the 602 SC pathways achieved the 0.7 threshold, then that AP pathway was recorded as a pure "no match" discord.

Collecting all this together produced four categorizations: (i) those where both p_{AP} and p_{SC} dropped below 0.05 (positive outcome "overlap," with 361 pathways), (ii) those where both p_{AP} and p_{SC} exceeded 0.05 (negative overlap, with 2509 pathways), (iii) those where only one *p*-value dropped below 0.05 but the pathway exhibited information-theoretic similarity ("ITS match," with 396 pathways), and (iv) those where only one *p*-value dropped below 0.05 and there was no ITS match ("no match," with 85 pathways). Clearly, the categorical similarity here is strong, with only 85/3351 (2.5%) pathways showing no form of overlap or match. Moving to a different significance level changes the counts, but not the mismatch pattern. Table 4 summarizes the results at the popular 10%, 5%, and 1% levels.

Figure 3 visualizes the relationship at the 5% level by plotting the 3351 paired *p*-values, as $-\log\{p\}$, from both the algorithms. The figure distinguishes the four overlap groups via the plotting character: (i) dark gray dots for positive overlaps, (ii) white dots for negative overlaps, (iii) light gray dots for ITS matches, and (iv) black dots for no match. The pattern shows a clear progression along the 45° line of pure agreement, with a spread that widens somewhat but then stabilizes as $-\log\{p\}$ grows. As expected, the two overlap groups lie in diagonal quadrants along the 45° line to the lower left and upper right of the plot. Notably, the black-dot "no match" group intermixes with the light-gray ITS-match group in the other quadrants of the display, with no otherwise-remarkable pattern. On balance we see that the two clustering algorithms exhibit reasonable outcome similarities when applied to this patient's data, but that their *p*-values can nonetheless vary somewhat.

Going further with these data, the list of DEPs found via SC at FDR < 15% extended to 266 pathways; a somewhat greater subset than the 80 found using AP clustering, but not a large percentage difference when recalling that the full collection of pathways under consideration numbers 3351. Indeed, no matter the underlying pattern, it is difficult to imagine that any two clustering-based, ordered lists of over 3300 pathways would be identical. One could ask, however, how comparable do the orderings appear? A similar bioinformatic question arises when comparing ordered gene lists, a useful quantification for which is described by Yang et al.[49] They manipulate ranks of the ordered *p*-values to build a similarity score, weighted to emphasize greater overlap at the top/most-dysregulated portion of both lists. (One could instead emphasize agreement at bottom, or at both top and bottom, but a top-focused weighting seems most appropriate for our DEP application.) The orderings are then permuted to produce an empirical *p*-value measuring how similar the two lists appear; small values indicate strong overlap. For long lists such as ours, Monte Carlo permutation is recommended. A Bioconductor package, *OrderedList*,[50] facilitates the calculations.

We computed the Yang et al. overlap metric to compare the rankings of our two common-DEP lists, using 1,000,000 Monte Carlo permutations, and otherwise accepting default settings in *OrderedList* (aside from

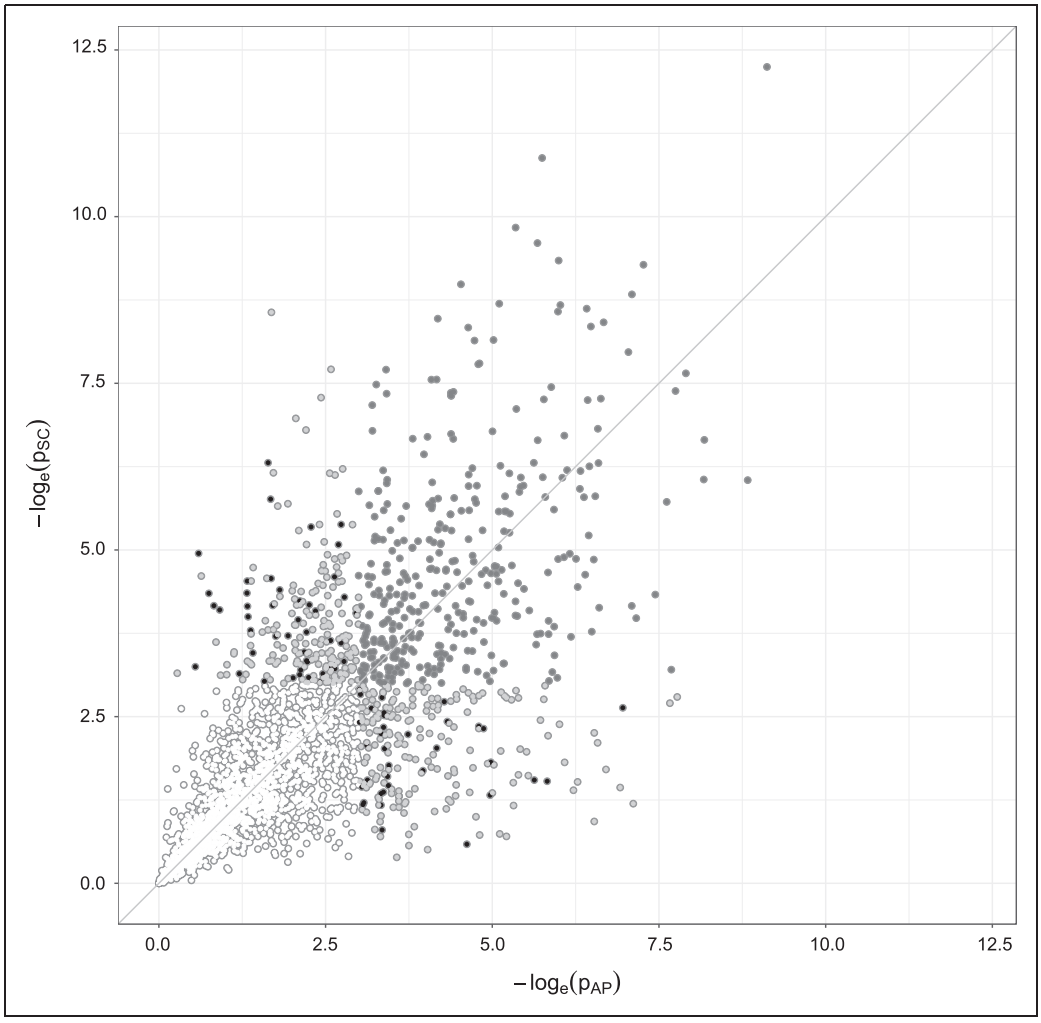

Figure 3. Comparison of $-\log\{p\}$ values from spectral clustering (SC) vs. AP clustering in the clustered-*T* test of equation (7) when applied to TNBC data from Section 4. Dot color indicates *p*-value overlap status: (i) dark gray dots for significant *p*-value overlaps (both below 5% cutoff), (ii) white dots for insignificant *p*-value overlaps (both above 5% cutoff), (iii) light gray dots for *p*-value discords with ITS match (high informatic similarity), and (iv) black dots for *p*-value discords with no ITS match. See text for details.

changing the top-and-bottom weighting to top-weighted, as discussed above). This resulted in a top-weighted overlap score of 25,706. Referenced to all permutations of the ranked lists, the empirical *p*-value was essentially zero, suggesting a strong, top-weighted, dysregulation-identifying similarity between the two clustering algorithms for these data.

We also recorded the patterns of cluster sizes, m, the two algorithms produced. Excluding cases with $m = 1$, the average number of clusters per pathway was 6.97 for AP clustering and 6.32 for spectral clustering. The corresponding five-number summaries were {2, 3, 5, 9, 41} for AP clustering (seen earlier), and {2, 3, 4, 8, 41} for spectral clustering. Despite these further similarities, we noticed one complicating feature in our analysis: when we permitted it, spectral clustering exhibited a greater tendency to produce clusters of size G—producing only $m = 1$ cluster of genes within the affected pathway—more often than AP clustering. While AP clustering produced five cases of $m = 1$ among the original 3411 pathways, spectral clustering gave 56. (There was one pathway where both methods selected $m = 1$, hence the number of common scored pathways was 3351.) Once again, these are trifling percentages when viewed in the context of 3351 scored pathways, but the 10-fold difference did attract our attention.

More generally, while both algorithms yielded roughly similar patterns in the cluster sizes here, the potential exists for them to produce different cluster solutions with a given historical data source. A particular pathway therein may possess an affinity pattern which engenders several essentially-optimal clustering solutions and

assignments for m. Different clustering algorithms might settle on different choices for these cluster solutions, which in turn could lead to very similar or very different test outcomes under our clustered-T strategy. When the pathway's cluster pattern is strong and a single optimal solution stands out, any differences should be minimal. In cases where the pattern is more ambivalent, however, choosing a different cluster algorithm may affect the nature of the final outcome.

Cluster algorithm selection is clearly a non-trivial component of our larger strategy and a need exists for more-extensive study of it and of the larger clustered-T methodology. We are exploring all the various issues discussed here and in Section 5.1 above, and we hope to report on them in future manuscripts.

Acknowledgements

The results published here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. We gratefully acknowledge the kind and helpful input of Mr. Qike Li, Dr. Joanne Berghout, Dr. Ikbel Achour, Dr. Colleen Kenost, Dr. Haiquan Li, Dr. Nima Pouladi, Dr. Ryan Gutenkunst, and Dr. Joseph Watkins. In addition, thanks are due the Editor and two anonymous referees for insightful comments that greatly improved the quality of the manuscript. This work represents a portion of the first author's Ph.D. dissertation from the University of Arizona Graduate Interdisciplinary Program in Statistics.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This material is based upon work supported by the U.S. National Science Foundation under Grant No. 1228509 and by the U.S. National Institutes of Health under Grant No. R03ES027394.

Supplemental material

Various supplemental tables and graphics, referenced above, are available with this paper at the journal's website.

References

1. van't Veer LJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; **415**: 530–536.
2. Lang JE, et al. Expression profiling of circulating tumor cells in metastatic breast cancer. *Breast Cancer Res Treat* 2014; **149**: 121–131.
3. Yang X, et al. Single sample expression-anchored mechanisms predict survival in head and neck cancer. *PLoS Comput Biol* 2012; **8**: e1002350.
4. Perez-Rathke A, Li H and Lussier YA. Interpreting personal transcriptomes: personalized mechanism-scale profiling of RNA-seq data. *Pac Symp Biocomput* 2013; **18**: 159–170.
5. Lillie EO, et al. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Per Med* 2011; **8**: 161–173.
6. Gardeux V, et al. 'N-of-1-pathways' unveils personal deregulated mechanisms from a single pair of RNA-Seq samples: towards precision medicine. *J Am Med Inform Assoc* 2014; **21**: 1015–1025.
7. Gardeux V, et al. Concordance of deregulated mechanisms unveiled in underpowered experiments: PTBP1 knockdown case study. *BMC Med Genomics* 2014; **7**(Suppl 1): Article S1.
8. Schissler AG, et al. Dynamic changes of RNA-sequencing expression for precision medicine: N-of-1-pathways Mahalanobis distance within pathways of single subjects predicts breast cancer survival. *Bioinformatics* 2015; **31**: i293–i302.
9. Ashburner, et al. Gene ontology: tool for the unification of biology. *Nat Genet* 2000; **25**: 25–29.
10. Hamburg MA and Collins FS. The path to personalized medicine. *N Engl J Med* 2010; **363**: 301–304.
11. Foulkes WD. Triple-negative breast cancer. *N Engl J Med* 2010; **363**: 1938–1948.
12. Povey S, et al. The HUGO gene nomenclature committee (HGNC). *Hum Genet* 2001; **109**: 678–80.
13. Resnick SI. *A probability path*. [Reprint of the 2005 edition]. New York: Springer Science & Business Media, 2014, Sec. 8.6.
14. Moore DS. *Introduction to the practice of statistics*. 8th ed. New York: WH Freeman & Co, 2014, Sec. 7.1 and Sec. 15.2.
15. Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc. Ser B* 1995; **57**: 289–300.

16. Wu D and Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res* 2012; **40**: e133.
17. Tamayo, et al. The limitations of simple gene set enrichment analysis assuming gene independence. *Stat Methods Med Res* 2016; **25**: 472–487.
18. R Core Team. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, 2016.
19. Wang H, et al. Individualized identification of disease-associated pathways with disrupted coordination of gene expression. *Brief Bioinform* 2015; **17**: 78–87.
20. Williams RL. A note on robust variance estimation for cluster-correlated data. *Biometrics* 2000; **56**: 645–646.
21. Frey BJ and Dueck D. Clustering by passing messages between data points. *Science* 2007; **315**: 972–976.
22. Zhang B and Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 2005; **4**: Article17.
23. Bodenhofer U, et al. *APCluster*: an R package for affinity propagation clustering. *Bioinformatics* 2011; **27**: 2463–2464.
24. Prasad TV and Ahso SI. Data mining for bioinformatics – microarray data. In: Fulekar MH (ed.) *Bioinformatics: applications in life and environmental sciences*, 1st ed. Dordrecht, The Netherlands: Springer, 2009, pp.77–144.
25. Piegorsch WW. *Statistical data analytics: foundations for data mining, informatics, and knowledge discovery*, 1st ed. Chichester: John Wiley & Sons, 2015, Sec. 11.1.2.
26. Caliński T and Harabasz J. A dendrite method for cluster analysis. *Commun Stat* 1974; **3**: 1–27.
27. Green RH. *Sampling design and statistical methods for environmental biologists*, 1st ed. New York: John Wiley & Sons, 1979, p.51.
28. Anscombe FJ. The statistical analysis of insect counts based on the negative binomial distribution. *Biometrics* 1949; **5**: 165–173.
29. Chen SX, Zhang LX and Zhong PS. Tests for high-dimensional covariance matrices. *J Am Stat Assoc* 2010; **105**: 810–819.
30. Pookhao N, et al. A two-stage statistical procedure for feature selection and comparison in functional analysis of metagenomes. *Bioinformatics* 2015; **31**: 158–165.
31. Di Y, et al. Higher order asymptotics for negative binomial regression inferences from RNA-sequencing data. *Statist Appl Genet Molec Biol* 2013; **12**: 49–70.
32. Genest C and Neslehova J. A primer on copulas for count data. *ASTIN Bull* 2007; **37**: 475–515.
33. Yan J. Enjoy the joy of copulas: with a package copula. *J Stat Softw* 2007; **21**: 1–21.
34. Agresti A and Coull BA. Approximate is better than ''exact'' for interval estimation of binomial proportions. *Am Stat* 1998; **52**: 119–126.
35. Li DM and Feng YM. Signaling mechanism of cell adhesion molecules in breast cancer metastasis: potential therapeutic targets. *Breast Cancer Res Treat* 2011; **128**: 7–21.
36. Rosen LS, Ashurst HL and Chap L. Targeting signal transduction pathways in metastatic breast cancer: a comprehensive review. *Oncologist* 2010; **15**: 216–235.
37. Brown KD. Transglutaminase 2 and NF-κB: an odd couple that shapes breast cancer phenotype. *Breast Cancer Res Treat* 2013; **137**: 329–336.
38. Yang H, et al. Toll-like receptor 4 prompts human breast cancer cells invasiveness via lipopolysaccharide stimulation and is overexpressed in patients with lymph node metastasis. *PLoS One* 2014; **9**: e109980.
39. Kidd LCR, et al. Contribution of toll-like receptor signaling pathways to breast tumorigenesis and treatment. *Breast Cancer* 2013; **5**: 43–51.
40. Zhang C, et al. Immunotherapeutic impact of Toll-like receptor agonists in breast cancer. *Anticancer Agents Med Chem* 2015; **15**: 1134–1140.
41. Goeman JJ and Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 2007; **23**: 980–987.
42. Kriegel HP, Peer K and Arthur Z. Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans Knowl Discov Data* 2009; **3**: Article1.
43. Christensen R, et al. *Bayesian ideas and data analysis: an introduction for scientists and statisticians*, 1st ed. Boca Raton, FL: CRC Press, 2011.
44. Everitt BS, et al. *Cluster analysis*, 5th ed. Chichester: John Wiley & Sons, 2011.
45. von Luxburg U. A tutorial on spectral clustering. *Stat Comput* 2007; **17**: 39–5416.
46. Hechenbichler K and Schliep KP. *Weighted k-nearest-neighbor techniques and ordinal classification*. [Discussion Paper 399, SFB 386]. Munich: Ludwig-Maximilians University of Munich, 2004.
47. Tao Y, et al. Information theory applied to the sparse gene ontology annotation network to predict novel gene function. *Bioinformatics* 2007; **23**: i529–i538.
48. Gardeux V, et al. Towards a PBMC ''virogram assay'' for precision medicine: concordance between ex vivo and in vivo viral infection transcriptomes. *J Biomed Inform* 2015; **55**: 94–103.
49. Yang X, et al. Similarities of ordered gene lists. *J Bioinform Comput Biol* 2006; **4**: 693–708.
50. Scheid S, et al. *Similarities of ordered gene lists. User's guide to the bioconductor package ordered list 1.11.3*. Technical Report Nr. 2006/01, Max Planck Institute for Molecular Genetics, Berlin, Germany, 2006.