Published in final edited form as:

Comput Stat Data Anal. 2009 January 15; 53(3): 788–800. doi:10.1016/j.csda.2008.07.004.

Internal validation inferences of significant genomic features in genome-wide screening

Cheng Cheng¹

Department of Biostatistics, St. Jude Children's Research Hospital, 332 N. Lauderdale Street, Memphis, TN 38105-2794

Abstract

Although validation of classification and prediction models has been a long-standing topic in Statistics and computer learning, the concept of statistical validation in genome-wide screening studies has been vague. Internal validation generally refers to validation procedures solely based on the study dataset. A popular approach to internal validation of identified genomic features has been the split-dataset validation. Contrast to this approach, internal validation in genome-wide association screening studies is precisely defined through the concepts of association profile and profile significance. A general procedure and two specific profile significance measures are developed and are compared with the split-dataset validation approach by a simulation study. The simulation results clearly demonstrate the strength and limitations of the profile significance approach to internal validation, especially its enormous gain in sensitivity (power) and stability over the split-dataset validation. The proposed methodology is illustrated by an example of genome-wide SNP association analysis in genetic epidemiology.

Keywords

validation; gene expression profiling; genome-wide association analysis; association profile; profile significance; permutation test; microarray; genomics

1. Introduction

Validation in microarray studies means, generally speaking, to confirm to some degree of satisfaction, the findings from a study in external dataset(s) that is similar or identical to the study dataset in aspects most important to the study. We need to distinguish biological validation of a specific finding (a gene, for example) by additional wet-lab experiments on independent samples, from statistical validation of the findings as a whole. This paper discusses the latter, and more specifically, internal validation to be precisely defined later.

1.1 Basic types of microarray data analyses

To discuss statistical validation for genome-wide studies, it is instructive to begin by recalling the three basic types of microarray analyses: class prediction, class discovery, and association analysis. Class prediction aims at building a classifier or prediction model for a phenotype of

¹Email cheng.cheng@stjude.org; Tel. 1-901-595-2935; FAX 1-901-595-8843.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

interest, using the gene expressions in an expression profile as features. Here the final study findings should be one or a few classifiers built from the study data. Success of a classifier relies on many factors. For a therapy outcome predictor, these factors include but not limited to, the existence of features (gene expression patterns) representative of the phenotype to classify (predict), a reasonably large sample size of the study dataset so that there is sufficient statistical power to identify such features, and most importantly, the therapeutic relevance of the study data and the features that make up the classifier (Simon, 2005a). Identification of representative features can usually be accomplished by an association analysis. Then various classification and statistical modeling methods can be applied to build the classifiers. Simon (2005b) outlines a road map for building a classifier. Multiple rounds of cross validation (CV) on the study dataset (training set) is usually used to tune the classifier's parameters and to obtain an estimate of the classification accuracy when the classifier is applied to future samples. The final classifier however, should be built using the entire study dataset. Concepts of validity and methods of classifier validation have been studied in several papers. Insightful discussions can be found, for example in Altman and Royston (2000) and Simon (2005a, b). A classifier can only be genuinely validated externally, i.e., on prospective samples external to the study dataset. Before its application in clinical practice the classifier must be tested and its accuracy assessed and compared to existing classification rules (if any), on a large number of prospective samples in the same therapeutic context.

Class discovery applies various clustering techniques such as the hierarchical clustering method to the study dataset to find sample and gene clusters novel to the existing biological knowledge. Here the findings are such clusters. Obviously these clusters cannot be meaningfully validated on the study dataset. The only genuine validation of such novel clusters is to see if they can be reasonably identified on a set of samples external to the study data. To date there has been no rigorous statistical inference procedure for this purpose.

Association analysis aims at identifying genomic features (gene expressions, copy number variations, polymorphisms, haplotypes, etc.), Gene Ontology (GO) groups, or pathways that are associated with the phenotype of interest. For a genome-wide gene expression study the basic findings are an expression profile consisting of a set of probes or probesets of genes. Statistical validity and validation of a given gene expression profile have not been clearly defined in the literature, although numerous intuitions and heuristics do exist. Certainly an ideal validation would be to see if the association between the expression profile and the phenotype still exists on external dataset(s) similar to the study dataset, but exactly how this can be done is unclear at this point because there is a large number of plausible options. Yet certain options may not be appropriate in general. For example, it is not wise to insist that each gene expression reaches the same or higher statistical significance level on the validation set, as the association between a particular gene expression and the phenotype may not be stable across datasets. Simon (2005b) briefly mentions this point in the context of classifier validation, and the same logic applies to association analysis as well. It is the association between the phenotype and the expression profile as a whole that should be validated. This thus raises a multivariate analysis problem that requires dimension reduction strategies, for which many plausible options are available.

1.2 Split-dataset validation (SDV)

Perhaps initially driven by the desire of building a therapy outcome classifier and realizing later the limitation of the study dataset and the high demand of classifier validation, some association studies in the literature employ a popular type of association analysis that has a flavor of classifier building and validation. Roughly speaking this approach consists of three steps. First, the study dataset is split randomly into two parts at a subjectively chosen ratio (e.g., 2:1 or 1:1). Next, one (the larger) part is used as a "training set" to identify a gene

expression profile. This step may consist of a few complex sub-steps involving association analysis, classifier-building, and multiple-round CV, all operating on the training set. The end product is an expression profile consisting of a set of gene expressions, and possibly a classifier. Finally, a cluster analysis using the identified genes as features or the classifier built on the test set is applied to the other part of the study dataset (the “validation set”), and statistical significance (P value) of the association between the phenotype and the cluster labels or the class labels predicted by the classifier is computed, often via testing a two- or multi-group comparison hypothesis or stochastic dependence in a contingency table. Figure 3 of Bullinger *et al.* (2004) gives a good illustration of this approach, with specific choices of methods in each step. By doing so the gene expression profile is considered somehow validated. There are two obvious drawbacks in this approach. First, splitting the study dataset into two can greatly reduce the statistical power to detect genes related to the phenotype. Consequently, there may not be adequate power on either the “training set” or the “validation set” to perform the required hypotheses tests. Second, the statistical significance (P value) of the association between the cluster (class) labels and the phenotype so observed on the “validation set” is difficult to interpret. One simply wonders what level of significance (P value) will be observed if the whole study dataset is randomly split again and the entire analysis is exactly repeated; and for the third time, for the fourth time, etc. This is a question about the variability of the observed significance level of association on the “validation set”, and it can only be answered by randomly splitting the study dataset a large number of times and repeating the entire analysis on each split. However, if this multiple-round splitting is performed, then the procedure does not, and can not, validate any *specific* gene expression profile because different sets of genes (expression profile) are obtained in different rounds. The result from such a multiple-round splitting analysis may be loosely interpreted as a multiple-round CV test of the complete null hypothesis that no association exists against the general alternative that some association exist, but clearly this is not a validation of any specific gene expression profile. The large variability in the results observed from randomly split datasets was recently demonstrated by Bøvelstad *et al.* (2007; Table 1).

1.3 Internal validation

In some applications of association analysis it may be quite difficult to either obtain an external validation dataset, or to split the study dataset and still maintain adequate statistical power. This is particularly the case when an infrequent phenotype is being studied (e.g., Bogni *et al.* 2006). It is thus desirable to have a sound and interpretable procedure that can meaningfully validate the identified expression profile on the study dataset itself (an internal validation). Simon (2005b) regards internal validation of a classifier essentially as cross validations on the study dataset to produce estimates of the error probability that the classifier incurs when it is applied to future samples. Such error probability estimates provide an assessment of the generalizability of the classifier beyond the study dataset. Internal validation of a specific (fixed) gene expression profile should have the same objective, i.e. assessing the generalizability of the gene expression profile; without considering prediction however, a different conceptual framework is necessary. The basic idea of hypothesis testing is useful here. The internal validation of a gene expression profile can be regarded as an inference to demonstrate that the identification of this gene expression profile on the study dataset is unlikely due to chance completely; thus the gene expression profile carries a reasonable degree of generalizability beyond the study dataset. This can be accomplished by a rigorous conditional inference to assess the likelihood of observing the same or stronger findings when the complete null hypothesis stating “no association exists” is true (i.e., all null hypotheses are true), given the data. One approach is to construct statistics that measure the strength of the association between the expression profile as a whole and the phenotype, and then assessing the statistical significance of this strength of association by a permutation test — to estimate how often one would observe the same or stronger overall level of association when in fact no

association exists (i.e., by chance completely). This approach requires dimension reduction strategies and measures of stochastic dependence, for which a large number of plausible options exist. An implementation of this type of internal validation that combines cluster analysis, two-sample comparison statistics, and permutation test can be found in Edick *et al.* (2005), Bogni *et al.* (2006), and the respective supplemental documents.

Goeman *et al.* (2004) introduce a test of global association of a (large) group of genes with a phenotype. By using all genes on the array as the gene group their procedure provides a test of the complete null hypothesis against the *general* alternative that some gene-phenotype association exist, but not a specific (conditional) alternative that gene-phenotype association exists in a given group of genes derived from an association inference (selection) procedure. As the authors pointed out, the gene group in the global test must be defined *a priori* without using the phenotype; this is exactly the opposite to the gene group identified from a selection inference that specifically search for gene-phenotype association (the association profile, see below). Thus the global test by Goeman *et al.* and the proposed internal validation inference address two fundamentally different issues.

At issue then is the need for (1) a precise definition of internal validation in gene expression profiling (or other genome-wide association analyses alike), and (2) concrete methods to perform the internal validation. This paper introduces the concepts *association profile* and *profile significance* by extending the traditional P value to massive multiple tests. An association profile is a set of genomic features declared as significantly associated with the phenotype by an inference procedure. For example, in gene differential expression profiling an association profile is a set of genes (probes) found to be significantly associated with the phenotype (differentially expressed) by a large number of association tests, one test per gene, often with some adjustment for multiple testing (Cheng and Pounds 2007 and references therein). Some genes in this set could be highly co-expressed if they are in a pathway activated by the biological process. In genome-wide association analysis with SNP arrays an association profile is a set of SNPs identified by association tests, one test per SNP, and some SNPs in the association profile could be in high linkage disequilibrium. Profile significance of an association profile is the probability of obtaining the same or higher overall level of association contained in the association profile by chance when the complete null hypothesis is true. The internal validation of an association profile is precisely an inference procedure to obtain the profile significance. Such an inference procedure can be implemented by permutation tests or other re-sampling methods when appropriate. A simulation study that uses a gene differential expression model mimicking effects by a gene signaling pathway is conducted. The results clearly show that for small to moderate sized gene expression profiles the proposed internal validation inference is much more powerful (sensitive) than the split-dataset approach, while holding the nominal levels fairly well. For large gene expression profiles the power of the proposed procedures is diluted, however in this case the split-dataset procedure actually “validates” gene expression profiles of much poorer statistical quality as measured by false discovery rate (FDR) and the false negative rate due to the reduction in power for identifying truly differentially expressed genes based on the “training set” only.

The subsequent sections are organized as follows: In Section 2 association profile and profile significance are formally defined to provide a rigorous concept of internal validation in an association study. A permutation inference procedure for internal validation based on the profile significance is constructed in its full generality, followed by introducing two specific profile significance measurements. In Section 3 a simulation study is conducted to assess the operating characteristics of the proposed method and compare its performance with the popular split-dataset approach. A real-data example from a genome-wide association study is discussed in Section 4. Finally discussion and concluding remarks are made in Section 5.

2. Profile significance

2.1 Concepts and general procedure

The concepts of profile significance and internal validation are developed in this section. Following the introduction of association profile in multiple tests, profile significance is defined via a generalization of the traditional P value; then internal validation is defined as an inference procedure based on the profile significance. Having precisely defined the concepts, a procedure of internal validation inference is introduced in its full generality. Because a gene expression profile (or a list of SNPs) is usually generated from probe-by-probe (SNP-by-SNP) tests of association (or differential expression), it is conceptually helpful to begin by considering general massive multiple hypothesis tests for association in the context of microarray analysis. Let $X_1, \dots, X_m$ be m random variables and Y be a p -dimensional random vector (often $p = 1$). Suppose, in a single study m hypotheses are tested: H_{0i} : X_i is not associated with Y vs. H_{1i} : not H_{0i} , $i = 1, \dots, m$. The specific form of “associated” is application and model dependent. For example, in gene differential expression analysis, X_i is the expression of the i th probe and Y is a categorical variable with say k levels. If the one-way ANOVA model is applied to each X_i , i.e., $X_i = \mu_{it} + \varepsilon_i$, $t = 1, \dots, k$, H_{0i} reduces to $\mu_{i1} = \dots = \mu_{ik}$. If Y is binary and the logistic regression model is applied, i.e., $\text{logit}(Y = 1) = \mu_i + \beta_i X_i$, H_{0i} reduces to $\beta_i = 0$. Other frequently used models include multifactor ANOVA possibly with interaction effects, multiple linear regression models, and relative hazard (Cox) regression when Y is a failure time subject to right-censoring or competing risks.

Suppose $[X_1, \dots, X_m, Y]$ is observed on n subjects. Let (X_{ij}, Y_j) be the j th observation of (X_i, Y) , let $\mathbf{X}$ be the m -by- n matrix formed by X_{ij} ($i = 1, \dots, m, j = 1, \dots, n$), and let $\mathbf{Y} = [Y_1, \dots, Y_n]$. In conjunction with a statistical model (parametric or nonparametric), a test is applied m times, once for each pair (H_{0i}, H_{1i}) to test H_{0i} vs. H_{1i} based on the data $\{(X_{ij}, Y_j), j = 1, \dots, n\}$, and then a significance rule is applied to decide if H_{0i} should be rejected in favor of H_{1i} , often with some adjustment for multiplicity. This inference is henceforth denoted by $\mathcal{Q}$ and the result of this inference is henceforth denoted by $\mathcal{Q}(\mathbf{X}, \mathbf{Y})$, which is a set of indexes telling which H_{0i} is rejected in favor of H_{1i} :

$$\mathcal{I}(\mathbf{X}, \mathbf{Y}) = \{i: H_{0i} \text{ is rejected}, 1 \leq i \leq m\}.$$

The corresponding dataset $\mathcal{R} = \mathcal{R}(\mathcal{Q}(\mathbf{X}, \mathbf{Y})) = \{(X_{ij}, Y_j), i \in \mathcal{Q}(\mathbf{X}, \mathbf{Y}), j = 1, \dots, n\}$ will be called an *association profile*. Always set $\mathcal{R} = \emptyset$ (empty set) if $\mathcal{Q}(\mathbf{X}, \mathbf{Y}) = \emptyset$ (i.e., no null hypothesis is rejected). Suppose that the hypotheses H_{0i} vs. H_{1i} ($i = 1, \dots, m$) are tested on two datasets $(\mathbf{X}_1, \mathbf{Y}_1)$ and $(\mathbf{X}_2, \mathbf{Y}_2)$. Given two association profiles $\mathcal{R}_1 = \mathcal{R}(\mathcal{Q}(\mathbf{X}_1, \mathbf{Y}_1))$ and $\mathcal{R}_2 = \mathcal{R}(\mathcal{Q}(\mathbf{X}_2, \mathbf{Y}_2))$ (note in general $\mathcal{Q}(\mathbf{X}_1, \mathbf{Y}_1) \neq \mathcal{Q}(\mathbf{X}_2, \mathbf{Y}_2)$), say $\mathcal{R}_1$ is stronger than $\mathcal{R}_2$ by comparison $\mathcal{C}$ and write $\mathcal{R}_1 \geq \mathcal{C} \mathcal{R}_2$ if and only if $\mathcal{R}_1$ shows stronger associations between Y and the X 's indexed in $\mathcal{Q}(\mathbf{X}_1, \mathbf{Y}_1)$ than does $\mathcal{R}_2$ for the variables indexed in $\mathcal{Q}(\mathbf{X}_2, \mathbf{Y}_2)$, according to a given comparison rule $\mathcal{C}$ which will be detailed in Section 2.2. Always regard $\emptyset \geq \mathcal{C} \mathcal{R}$ by convention. For interpretability $\mathcal{C}$ is required to satisfy the following conditions: For (almost) all realizations of data, (i) $\mathcal{R} \geq \mathcal{C} \mathcal{R}$ (reflexive), and (ii) if $\mathcal{R}_1 \geq \mathcal{C} \mathcal{R}_2$ and $\mathcal{R}_2 \geq \mathcal{C} \mathcal{R}_3$ then $\mathcal{R}_1 \geq \mathcal{C} \mathcal{R}_3$ (transitive).

Define the *complete null hypothesis* as H_0^* : no association exists; or equivalently, all null

hypotheses are true — $H_0^* = \bigwedge_{i=1}^m H_{0i}$, where $\wedge$ denotes the logical “and.” This complete null hypothesis has been used in producing estimators of parameters related to the number of erroneous rejections; see for example, Owen (2005). Let $(\mathbf{X}_{obs}, \mathbf{Y}_{obs})$ be the observed data and let $\mathcal{R}_{obs} = \mathcal{R}(\mathcal{Q}(\mathbf{X}_{obs}, \mathbf{Y}_{obs}))$ be an observed association profile generated from testing the hypotheses on $(\mathbf{X}_{obs}, \mathbf{Y}_{obs})$. The *profile significance* of $\mathcal{R}_{obs}$ with respect to the comparison

rule $\mathcal{C}$, denoted henceforth by $PS_{\mathcal{C}}$, is defined to be the probability of $\mathcal{R} \geq_{\mathcal{C}} \mathcal{R}_{obs}$ under the complete null hypothesis H_0^* for a generic association profile $\mathcal{R} = \mathcal{R}(\mathcal{Q}(\mathbf{X}, \mathbf{Y}))$:

$$PS_{\mathcal{C}} = \Pr(\mathcal{R} \geq_{\mathcal{C}} \mathcal{R}_{obs}; H_0^*).$$

The profile significance serves as a measure of statistical significance of the observed association profile given a comparison rule $\mathcal{C}$. Estimation of the profile significance $PS_{\mathcal{C}}$ defines the *internal validation* of the association profile $\mathcal{R}_{obs}$. Similar to the quantitative performance measurements of a classifier on a validation dataset (e.g., the receiver operation characteristic curve), the profile significance, a probability by definition, gives a quantitative level of validation, instead of simply saying the association profile is valid or not. For most applications having a quantitative, meaningful level of validation is more informative and may be more desirable than a simple yes-no conclusion.

In most applications the experimental design allows for estimating $PS_{\mathcal{C}}$ by a permutation or some other re-sampling procedure that simulates data accurately under the complete null hypothesis conditional on the observed data. For example, when each Y_j ($j = 1, \dots, n$) is a scalar variable, it often suffices to randomly permute the entries in $\mathbf{Y}_{obs}$, or to generate all possible permutations if n is small. A general procedure is outlined below. Suppose that $\mathcal{R}_{obs} \neq \emptyset$; i.e., at least one null hypothesis is rejected by the observed data.

Algorithm 1—Permutation Estimation of The Profile Significance

Let L be the number of permutation (or re-sampling) runs. Repeat for $\ell = 1, 2, \dots, L$:

1. Simulate a dataset $(\mathbf{X}_{\ell}, \mathbf{Y}_{\ell})$ under H_0^* by permutation or re-sampling;
Apply the same inference to the data $(\mathbf{X}_{\ell}, \mathbf{Y}_{\ell})$ to generate $\mathcal{R}_{\ell} = \mathcal{R}(\mathcal{Q}(\mathbf{X}_{\ell}, \mathbf{Y}_{\ell}))$.

A permutation estimator of $PS_{\mathcal{C}}$ is given by

$$\widehat{PS}_{\mathcal{C}} = \frac{1}{L} \sum_{\ell=1}^L I(\mathcal{R}_{\ell} \geq_{\mathcal{C}} \mathcal{R}_{obs}),$$

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

It should be noted that the simulation model under the complete null hypothesis H_0^* used in Step 1 can be tailored to each particular application. Re-sampling methods other than permutation can be appropriate as well for a particular application. This will be illustrated by a real data example in Section 4. Clearly the conclusion hinges on the comparison rule $\mathcal{C}$. The construction and specific form of this comparison rule may vary from application to application. Generally speaking in an application $\mathcal{C}$ must avoid algorithmic artifact and tuning parameters as much as possible, and be justifiably fair. Moreover, because the comparison rule is evaluated in each permutation (re-sampling) round and thus for a large number of times, it is desirable that its evaluation is computationally efficient and numerically robust. Two concrete comparison rules meeting these criteria and general enough to cover most of the applications are described below.

2.2 Two specific comparison rules

Total P-value Score—Suppose testing H_{0i} vs. H_{1i} generates a P value P_i ($i = 1, \dots, m$), and the conclusion consists of $\mathcal{Q}_{\gamma}(\mathbf{X}, \mathbf{Y}) = \{i : P_i \leq \gamma, i = 1, \dots, m\}$ for some significance level γ . In

practice the significance level γ can be determined in various ways, often with some adjustment for massive multiplicity (Benjamini and Hochberg 1995; Benjamini *et al.* 2006; Storey *et al.* 2003; Storey 2002; Allison *et al.* 2002; Pounds and Cheng 2006; Cheng *et al.* 2004; Cheng 2006). Let $\mathcal{R}(\gamma) = \mathcal{R}(\mathcal{Q}_\gamma(\mathbf{X}, \mathbf{Y}))$ be the corresponding association profile. Because P_i carries the weight of evidence supporting H_{1i} (X_i is associated with Y), a proper aggregation of the P values would well reflect the overall strength of the X – Y associations in $\mathcal{R}(\gamma)$. Thus the *total P value score* aggregates the “marginal” (an individual X with Y) association strength in the association profile:

$$S(\gamma) = - \sum_{i=1}^m I(P_i \leq \gamma) \log(P_i).$$

Suppose H_{0i} vs. H_{1i} ($i = 1, \dots, m$) are tested on two datasets $(\mathbf{X}_1, \mathbf{Y}_1)$ and $(\mathbf{X}_2, \mathbf{Y}_2)$, giving two association profiles $\mathcal{R}_k(\gamma) = \mathcal{R}(\mathcal{Q}_\gamma(\mathbf{X}_k, \mathbf{Y}_k))$, and two total P value scores $S_k(\gamma)$ computed in the same way as defined above, $k = 1, 2$. Then the comparison rule $\mathcal{C}$ says $\mathcal{R}_1(\gamma) \geq \mathcal{C} \mathcal{R}_2(\gamma)$ if and only if $S_1(\gamma) \geq S_2(\gamma)$. It is straightforward to verify that this comparison rule is reflexive and transitive.

It should be emphasized that in the estimation procedure (Algorithm 1) of the profile significance $PS_{\mathcal{C}}$, exactly the same inference as that applied to the multiple hypothesis tests on the original data must be used in all the permutation (re-sampling) rounds. Additionally, note that this comparison rule does not depend on any information beyond the P values, avoiding tuning parameters and artifacts, and is straightforward to compute. Internal validation using this type of score statistics will be called *profile significance by total P value score* (PSTPS).

Supervised clustering for dimension reduction—The PSTPS procedure measures the global level of association between the genomic features in an association profile and the phenotype by aggregating the evidence of association from each individual feature. Such a statistic will have good power when each individual feature exerts substantial association with the phenotype, but jointly these features do not necessarily provide accurate prediction (classification) of the phenotype. On the other hand, the opposite has also been empirically observed (Bogni *et al.* 2006) — each individual feature exhibits relatively weak association with the phenotype but all the features in the association profile jointly exert rather strong evidence of global association with the phenotype after a dimension reduction by clustering. Thus for determination of the profile significance it is desirable to have another dimension reduction method complementary to PSTPS. An illustrative example in low dimension (2 features in the association profile) will be discussed after describing the general procedure.

When the phenotype Y is a categorical variable or a failure time variable subject to right-censoring, the idea of clustering can be borrowed to derive a dimension reduction technique. Consider the categorical case first, and without loss of generality code the categories (equivalently, distinct values of Y) as $1, 2, \dots, K$. Given an association profile $\mathcal{R} = \mathcal{R}(\mathcal{Q}(\mathbf{X}, \mathbf{Y})) = \{(X_{ij}, Y_j) : i \in \mathcal{Q}(\mathbf{X}, \mathbf{Y}), j = 1, \dots, n\}$, for convenience renumber the X variables indexed in $\mathcal{Q}(\mathbf{X}, \mathbf{Y})$ as $X_1, \dots, X_p$ and rewrite the X_{ij} ’s into p -vectors $\mathbf{x}_j = [X_{1j}, \dots, X_{pj}]$, $j = 1, \dots, n$.

Algorithm 2: Supervised clustering: categorical phenotype

1. Partition the X data into K sets: $\mathcal{X}_k = \{\mathbf{x}_j : Y_j = k, j = 1, \dots, n\}$, $k = 1, \dots, K$.

2. Compute a centroid for $\mathcal{X}_k$ by the mean vector $\mathbf{x}_k^* = N_k^{-1} \sum_{j=1}^n I(Y_j=k) \mathbf{x}_j$ with $N_k = \sum_{j=1}^n I(Y_j=k)$, or the component-wise median $\mathbf{x}_k^* = [\hat{\xi}_1, \dots, \hat{\xi}_p]$ with $\hat{\xi}_i = \{X_{ij} : Y_j = k, j = 1, \dots, n\}$.
3. Choose a proper distance (or dissimilarity) function $d(\cdot, \cdot) : \mathbb{R}^p \times \mathbb{R}^p \mapsto [0, \infty)$. Assign sample j the cluster code k if $d(\mathbf{x}_j, \mathbf{x}_k^*) = \min(d(\mathbf{x}_j, \mathbf{x}_r^*), r=1, \dots, K)$.

It should be emphasized that the goal here is not to build a classifier although this algorithm is essentially a nearest-centroid classification procedure. This algorithm merely serves as a dimension reduction technique. Consequently, measures of classification accuracy (such as false positive rate) are not appropriate statistics for the purpose of estimating the profile significance. Rather general association statistics should be considered here. Thus let C_j be the cluster code assigned to sample j . Then the measurement of association becomes a measure of concordance in the pairs $(C_j, Y_j), j = 1, \dots, n$. There are several plausible options here. The most straightforward measurement would be the truncated sample correlation coefficient $S = \max(R, 0)$ with R the sample correlation coefficient of the C - Y pairs. Other options include statistics based on contingency tables; for example, when $K = 2$, S can be set to $-\log(P)$ of the two-sided P value of Fisher's exact test; for larger K (and n) Pearson's Chi-square statistic can be used. Finally the comparison rule $\mathcal{C}$ says $\mathcal{R}_1 \geq \mathcal{C} \mathcal{R}_2$ if and only if $S_1 \geq S_2$. Clearly this comparison is reflexive and transitive.

Algorithm 2 can be extended to handle a censored continuous phenotype. A right-censored failure time phenotype is defined by a pair $Y = (T, \delta)$, where $T = \min(T_f, C)$, T_f is the actual time to failure, C is the censoring time, $\delta = 0$ (censored; failure is not observed) if $T_f > C$, and $\delta = 1$ (failure is observed) if $T_f \leq C$. In the presence of censoring only the pair (T, δ) is observable. The phenotype data now consist of n i.i.d. pairs $\mathbf{Y} = [(T_1, \delta_1), \dots, (T_n, \delta_n)]$. The supervised clustering procedure operates on an association profile $\mathcal{R} = \mathcal{R}(\mathbf{X}, \mathbf{Y}) = \{(X_{ij}, T_j, \delta_j) : i \in \mathcal{Q}(\mathbf{X}, \mathbf{Y}), j = 1, \dots, n\}$ in exactly the same way as the categorical case with $K = 2$; the only modification is to replace Y_j by $\delta_j + 1$ throughout the algorithm. The supervised clustering produces two groups $G_k = \{j : \text{sample } j \text{ is assigned cluster code } k\}, k = 1, 2$. With the two groups so generated, the failure time phenotype data are partitioned into two groups as well: $\mathcal{Y}_k = \{(T_j, \delta_j) : j \in G_k\}, k = 1, 2$. Then compute a two-sample test statistic of comparing the failure time distributions between the two groups, e.g., the log-rank statistic. Let S_1, S_2 be two test statistics so generated for two association profiles $\mathcal{R}_1$ and $\mathcal{R}_2$. Then the comparison rule $\mathcal{C}$ says $\mathcal{R}_1 \geq \mathcal{C} \mathcal{R}_2$ if and only if $|S_1| \geq |S_2|$. It is straightforward to show that this comparison is reflexive and transitive.

Note that apart from the choice of the distance (dissimilarity) function, the comparison rule based on supervised clustering does not contain any tuning parameter and algorithmic artifact. A natural choice of the distance function usually exists in most of the applications. It is imperative that exactly the same distance function is applied to both the original data and in every permutation (re-sampling) round. Internal validation using the above procedures will be called *profile significance by supervised clustering* (PSSC).

To illustrate the scenario described earlier, consider a low dimensional example with 2 continuous genomic features X_1 and X_2 in the association profile and a binary phenotype, shown in Figure 1. Neither feature is very strongly associated with the binary phenotype individually — the two-sample t-test P value is 0.0505 and 0.0786 respectively. However jointly these two features show extremely strong association with the phenotype. Based on 2,000 permutations, PSTPS is estimated as 0.0465; whereas using Mahalanobis distance in the clustering procedure, PSSC is estimated as < 0.0005 ($1/2000$) by either the correlation statistic or the score statistic from Fisher's exact test as described above. Therefore PSSC can be much more powerful than

PSTPS in such scenario, and in applications these two types of profile significance can be complementary to each other.

3. A simulation study

3.1 Simulation model and set up

Simulation model—Using the notation introduced in Section 2, a dataset of m genes and n samples consists of (X_{ij}, Y_j) , $i = 1, \dots, m, j = 1, \dots, n$. In the simulation model the phenotype data are binary and fixed: the first n_1 are 1 and the remaining are 0: $Y_1 = Y_2 = \dots = Y_{n_1} = 1$ and $Y_{n_1+1} = \dots = Y_n = 0$; thus defining two groups, with size n_1 and $n_2 = n - n_1$ respectively.

In the differential expression model the first m_1 genes are set differentially expressed. Differential expression and correlations among expression data are generated by the “latent variable” model as follows: Generate for $j = 1, \dots, n_1$, Z_j i.i.d. $N(\mu, \sigma^2)$, Z'_j i.i.d. $N(a\mu, \sigma^2)$, and for $j = n_1 + 1, \dots, n$, Z_j i.i.d. $N(0, \sigma^2)$, Z'_j i.i.d. $N(0, \sigma^2)$. For $i = 1, \dots, m_1$ set

$$X_{ij} = c(i - Z_j) + \varepsilon_{ij}, \quad j = 1, \dots, n_1,$$

and $X_{ij} = Z_j/b + \varepsilon_{ij}/b$, $j = n_1 + 1, \dots, n$. Henceforth the ε_{ij} 's are i.i.d. $N(0, \sigma_e^2)$. Thus $E(X_{ij}) = c(i - \mu)$ in the $Y = 1$ group and $E(X_{ij}) = 0$ in the $Y = 0$ group. The expressions are correlated: $\text{Cov}(X_{ij}, X_{i'j}) = c^2\sigma^2$ in the $Y = 1$ group ($j = 1, \dots, n_1$) and $\text{Cov}(X_{ij}, X_{i'j}) = b^{-2}\sigma^2$ in the $Y = 0$ group ($j = n_1 + 1, \dots, n$).

The next $m_2 - m_1 + 1$ genes are not differentially expressed but are correlated among themselves and with the first m_1 genes: For $i = m_1 + 1, \dots, m_2$, set $X_{ij} = Z_j - Z'_j/a + \varepsilon_{ij}/b$, $j = 1, \dots, n_1, \dots, n$. Thus $E(X_{ij}) = 0$ in both groups; $\text{Cov}(X_{ij}, X_{i'j}) = (1 + 1/a)\sigma^2$, $m_1 + 1 \leq i \neq i' \leq m_2$, and for $1 \leq i \leq m_1$, $m_1 + 1 \leq i' \leq m_2$, $\text{Cov}(X_{ij}, X_{i'j}) = -c\sigma^2$ in the $Y = 1$ group, $\text{Cov}(X_{ij}, X_{i'j}) = -b^{-2}\sigma^2$ in the $Y = 0$ group. The remaining gene expressions are noise: For $i = m_2 + 1, \dots, m$ and $j = 1, \dots, n_1, \dots, n$, set X_{ij} i.i.d. $N(0, \sigma^2 + \sigma_e^2)$.

This type of model mimics the activation of a biological pathway under a condition (represented by $Y = 1$) that leads to gene differential expression. The latent variables Z and Z' represent an underlying response to this condition, resulting in the correlated differential expressions and non-differential expressions. The block-wise correlated gene expressions popular in simulation studies of FDR methods can be generated using this type of model by introducing a few more latent variables, one for each correlation block.

Gene expressions under the complete null hypothesis of no differential expression are generated by replacing the index i with μ (mean of Z) in $c(i - Z_j)$.

Simulation configuration—The sample size $n = 120$; size of the $Y = 1$ group $n_1 = 24$; the number of genes $m = 10,000$; the number of differentially expressed and correlated genes $m_1 = 100$; $m_2 = 5,000$, giving 4,900 non-differentially expressed and correlated genes. A fairly high-noise, high-correlation scenario is considered: $\sigma = \sigma_e = 2$, $\mu = 60$, $a = 10$, $b = 2$, and $c = 4\sqrt{5/96}/30 \approx 0.03$. Four pre-specified P value cutoffs (γ values), 0.0005, 0.001, 0.005, and 0.01 are used in the screening of differentially expressed genes. In computing the profile significance $L = 400$ permutations are used. Differential expression and the complete null cases are both simulated with generating $N = 1,000$ datasets in each case.

Procedures—Screening of differentially expressed genes is performed by computing the two-sided P value of the two-sample t test for each gene, and rejecting the null hypothesis if $P \leq \gamma$ for a pre-specified significance γ . Three internal validation procedures are evaluated, PSTPS, PSSC, and the split-dataset validation (SDV). On each simulated dataset PSTPS and PSSC are computed using permutation as described in Section 2.2. In computing PSSC, the same P value cutoff γ used on the unpermuted dataset is used in each permutation round to generate a feature (gene) set for clustering. The test statistics used in PSSC is the truncated sample correlation between the cluster labels and Y , as described in Section 2.2.

In SDV, each simulated dataset is randomly spitted into two parts at the 1:1 ratio. The first half is used as the training set and the second half as the validation set. Screening of differentially expressed genes is performed only on the training set, by computing the two-sided P value of the two-sample t test for each gene, and rejecting the null hypothesis if $P < \gamma$ for a pre-specified significance γ . Then using the “differentially expressed” genes so found as features, a k -mean cluster with $k = 2$ is performed on the validation set, and the two-sided P value of Fisher’s exact test of independence between the cluster label and Y is computed to define the validation; this P value will be referred as the *validation P value*.

Performance measurements—Operating characteristics of PSTPS, PSSC, and SDV are assessed by empirical ROC curves (Figure 2). For each γ significance level used in the screening step, let $P_1(\gamma), \dots, P_N(\gamma)$ be the values of a profile significance or the validation P values computed on the N simulated datasets. Let $P_N(t; \gamma)$ be the empirical distribution function of

these observations, i.e., $\tilde{P}_N(t; \gamma) = N^{-1} \sum_{i=1}^N I(P_i(\gamma) \leq t)$, $-\infty < t < \infty$. Then for a given $\alpha \in [0, 1]$, $\tilde{P}_N(\alpha; \gamma)$ is the estimated “power” of the procedure at the α nominal significance level under the differential expression model (alternative), and under the complete null hypothesis it is the estimated actual significance level corresponding to α . The empirical ROC curve is generated by plotting the estimated power against the estimated actual significance level on the grid of nominal levels from 0.0 to 0.3 with 1000 grid points.

Error behavior of the screening multiple tests under the differential expression model is assessed by estimating the actual FDR and the actual false negative rate (FNR). For a given γ significance level, let $R_i(\gamma)$ be the total number of genes declared as differentially expressed, and $V_i(\gamma)$ be the number of genes erroneously declared as differentially expressed, on the i th simulated dataset ($i = 1, \dots, N$). Then the estimated actual FDR is

$$\widehat{FDR}(\gamma) = N^{-1} \sum_{i=1}^N (V_i(\gamma) / R_i(\gamma)) \text{ and the estimated actual FNR is } \widehat{FNR}(\gamma) = N^{-1} \sum_{i=1}^N [1 - m_1^{-1}(R_i(\gamma) - V_i(\gamma))].$$

To relate the internal validation procedures with these error measurements, descriptive statistics of $P_1(\gamma), \dots, P_N(\gamma)$, including sample mean, standard deviation, 10th, 50th, and 90th percentile are plotted against $\widehat{FDR}(\gamma)$ and $\widehat{FNR}(\gamma)$ for each γ value (Figure 3). Additionally, details of error proportions are examined for the simulated datasets on which the SDV procedure produced a more significant result than the PSSC (i.e., the validation P value is $\leq$ the PSSC), for $\gamma = 0.005, 0.01$. For each such dataset define the false positive proportion as $FPP_i(\gamma) = V_i(\gamma) / R_i(\gamma)$, and the false negative proportion as $FNP_i(\gamma) = [m_1 - (R_i(\gamma) - V_i(\gamma))] / m_1$. The FPP (FNP) on the training set used in SDV is plotted against the FPP (FNP) on the whole dataset, showing their empirical conditional joint distribution given that SDV is more significant than PSSC. The proportion below the diagonal line gives a point estimate of the conditional probability that the error level FPP (FNP) of the association profile obtained from the whole dataset is higher than that obtained from the training set, given that SDV is more significant than PSSC (Figure 4).

3.2 Results

The empirical ROC curves are given in Figure 2. PSTPS (solid line) has excellent power when the significance level is greater than 5% for three out of four γ values. This procedure is slightly liberal; for example, for $\gamma = 0.001$, at the 1%, 5%, and 10% nominal level the estimated power is 0.412, 1.00, 1.00 respectively, and the estimated actual level is 0.014, 0.074, 0.149 respectively. The power at the 5% level is diluted for less stringent screening ($\gamma = 0.01$), but at the 10% nominal level the PSTPS procedure still maintains 0.967 estimated power and 0.109 estimated actual level. For stringent screening ($\gamma = 0.0005, 0.001$), although slightly liberal PSSC (long-dashed line) has excellent power; for example, for $\gamma = 0.001$, at the 1%, 5%, and 10% nominal level the estimated power is 0.778, 0.924, 0.959 respectively, and the estimated actual level is 0.015, 0.059, 0.113 respectively. At the less stringent screening levels ($\gamma = 0.005, 0.01$) the power of PSSC is diluted and the procedure does not hold the nominal levels well. The SDV procedure (dash-dot line) is conservative and has low power in all cases; for $\gamma = 0.001$, at the 1%, 5%, and 10% nominal level the estimated power is 0.429, 0.559, 0.634 respectively, and the estimated actual level is 0.006, 0.026, 0.060 respectively.

For less stringent screening ($\gamma = 0.01$) SDV is more powerful than both PSTPS and PSSC at any significance level $\leq 5\%$. However a question is whether SDV is “validating” an association profile of good statistical quality as measured by FDR and FNR. Figure 3 shows plots of the across-simulation summary statistics of the simulated profile significance values and the SDV validation P values against the estimated actual FDR and FNR. At any given screening significance threshold γ the FDR and FNR of the association profile obtained on the entire dataset are always lower than those obtained on the training set which is of half the size of the whole dataset — note the points representing SDV are all shifted to the right. For example, at the less stringent $\gamma = 0.01$, the estimated FDR of the association profile obtained on the whole data set is 0.580, compared to 0.696 obtained on the training set; the estimated FNR on the whole dataset is 0.582, compared to 0.748 on the training set. The increase in both types of errors and reduction in power resulted from splitting the dataset into one half is rather evident. The standard deviation and percentiles show that PSTPS is quite stable despite of being slightly liberal or lack of power at 1% level in a few cases seen in Figure 2. PSSC is stable for stringent screening significance levels ($\gamma = 0.0005, 0.001$), and becomes more variable when the screening becomes less stringent ($\gamma = 0.005, 0.01$). Similar to a recent finding (Bøvelstad *et al.* 2007; Table 1), splitting the dataset (SDV) produces validation P values that are highly variable regardless the stringency in screening, as shown by the standard deviation and percentiles.

As a more detailed investigation, Figure 4 shows for $\gamma = 0.005, 0.01$ the empirical conditional joint distribution of FPP (FNP) on the whole dataset and on the training set, given that SDV is more significant than PSSC. The proportion of points below the diagonal line is an estimate of the conditional probability that FPP on the whole dataset is higher, given SDV is more significant than PSSC. Thus even if given SDV being more significant, only about 10% to 12% of the times SDV is validating a better association profile as measured by FPP, and is extremely unlikely (estimated conditional probability = 0) validating a better association profile in terms of FNP.

3.3 Discussion

Figure 2 suggests a complementarity between PSSC and PSTPS in their performance. For association profiles generated by stringent screening PSSC outperforms PSTPS, and for less stringent screening, at 5% to 10% significance levels PSTPS outperforms PSSC. Both procedures experience power dilution as the screening significance level becomes less stringent, and the power dilution of PSTPS goes slower. SDV is consistently conservative regardless the stringency of the screening, although it becomes more powerful than PSTPS and

PSSC at the $\leq 5\%$ significance levels when the screening is not stringent ($\gamma = 0.01$, actually rather loose for $m = 10000$ tests); in this case however PSTPS performs quite well at the 10% nominal significance level. Given the large gains in power, the small amount of liberalness of the PSTPS procedure may be a worthy trade off, especially for association profiles generated by less stringent screening. Figures 3 and 4 clearly demonstrate the increase in the levels of false positive and false negative errors (thus reduction in power) incurred by the SDV approach. Consequently, even if when SDV is more significant than profile significance, very frequently SDV is validating an association profile that is of much poorer statistical quality than that produced using the whole study dataset.

4. An example in genetic epidemiology

Wellcome Trust Case Control Consortium (WTCCC) and The Australo-Anglo-American Spondylitis Consortium (AAASC) recently published a genome-wide association scan of 14,500 nonsynonymous single nucleotide polymorphisms (nsSNPs) in four diseases (WTCCC & AAASC 2007). The proposed methodology is now illustrated by a screening for nsSNPs associated with ankylosing spondylitis (AS), using for most of the analyzable nsSNPs, 922 cases and 1,500 controls.

After filtering for missing genotype data and monomorphism, 10,985 nsSNPs were left analyzable by Cochran-Armitage trend test (χ^2_{slope} , Fleiss *et al.* 2003, P.194). First, P value of the χ^2_{slope} test was calculated for each of the 10,985 nsSNP. These P values were then analyzed by the FDR estimation and the profile information criterion I_p developed in Cheng *et al.* (2004). The estimated proportion of the analyzed nsSNPs not associated with AS (i.e., the proportion of true null hypotheses) was $\hat{\pi}_0 = 0.8825$. At the prespecified per-test significance level $\gamma = 0.0001, 0.00025, 0.0005, 0.00075, 0.001, 0.0025, 0.005, 0.0075$, and 0.01 , respectively 392, 426, 452, 467, 478, 517, 569, 626, and 669 nsSNPs were significant, with estimated FDR 0.0035, 0.0080, 0.0140, 0.0190, 0.0236, 0.0480, 0.0868, 0.1205, and 0.1502. The I_p criterion suggested the per-test significance threshold $\gamma^* = 0.005843$ with estimated FDR at this P value threshold as 0.0988. The PSTPS statistics (Section 2.2) were computed next. At the above γ values the PSTPS of the association profiles were 3703.94, 4002.58, 4209.78, 4320.26, 4397.30, 4649.62, 4941.07, 5229.34, and 5433.68 respectively. At the I_p -determined $\gamma^* = 0.005843$ the total P value score of the association profile was 5008.93.

A simulation of the complete null hypothesis was carried out next to estimate the pro-file significance for (i.e., internally validate) these association profiles. For each nsSNP information about association with AS was contained in the 3×2 contingency table, with rows corresponding to the three genotypes and column 1 (column 2) corresponding to AS (Control). Without loss generality, code the genotypes (rows) into ordinal values 0, 1, 2. In this particular design, the number of cases and the number of controls (i.e., the column totals) can be considered as fixed, but the row totals (genotype counts) should not necessarily be fixed. Random permutation of the group labels keeps both the column and the row totals unchanged for each nsSNP. The simulation of the no-trend null hypothesis herein relaxed the constraint of fixed row totals by conditioning on the observed marginal row proportions and the column totals. Specifically the sampling procedure proceeded as follows. For an nsSNP let p_i ($i = 0, 1, 2$) be the observed row proportions, let C_j ($j = 1, 2$) be the column totals, and let N_{ij} be the count in the (i, j) th cell that is to be generated. First generate N_{01} and N_{02} according to the Binomial(C_1, p_0) and Binomial(C_2, p_0) distribution respectively; next conditional on N_{01} (resp. N_{02}), generate N_{11} (N_{12}) according to Binomial($C_1 - N_{01}, p_1/(p_1 + p_2)$) (resp. Binomial($C_2 - N_{02}, p_1/(p_1 + p_2)$)); finally set $N_{21} = C_1 - N_{01} - N_{11}$ and $N_{22} = C_2 - N_{02} - N_{12}$. Using this model for the no-trend null hypothesis 4,000 rounds of simulation were run. In each simulation round a 3×2 table was generated as described for each nsSNP, exactly the same analyses by

P value cutoff (the γ values) and by I_p were repeated using the simulated (null) tables, and the total P value scores were computed.

For every association profile generated by a prespecified per-test significance level (i.e., γ value), a total P value score greater than or equal to the corresponding one observed on the original data was observed in none of the 4,000 simulations; hence the estimated profile significance $\widehat{PS}(\gamma) < 0.00025$ for every association profile generated at the prespecified per-test significance level γ . The same was true for the association profile identified by the I_p criterion; the 5-number summary (sample minimum, three quartiles, maximum) of the total P value scores across the 4,000 simulations was (0.00, 581.35, 845.13, 1047.64, 1289.44, 3557.67). Additionally, Bonferroni-adjusting for ten parameter estimates (9 γ values and I_p), the $100(1 - 0.01/10)\% = 99.9\%$ Clopper-Pearson exact confidence interval for the PSTPS was [0.0000, 0.0019]. Therefore there was strong statistical evidence that identification of these association profiles was not completely attributable to chance and thus these association profiles contained nsSNPs truly associated with AS.

5. Concluding remarks

In this paper internal validation for genome-wide screening studies is precisely defined through the concepts of association profile and profile significance. A general procedure and two specific profile significance measures are developed and compared with the popular split-dataset validation approach by a simulation study. The simulation results clearly demonstrate the strength and limitations of the profile significance approach to internal validation.

Although developed in the context of gene expression profiling, the profile significance is a general concept for any massive multiple tests problem, and its extension to other genome-wide association studies is rather straightforward. The gene expression matrix $\mathbf{X}$ can be generalized to a matrix of any genomic features (e.g., SNPs, haplotypes, copy number variations, etc.). In a genome-wide association analysis using SNP arrays, the interest may be in identifying single SNPs associated with the phenotype, or sets of SNPs linked to a gene associated with the phenotype. In either case, as illustrated by the example in Section 4, it is worthwhile to assess the total weight of evidence for genotype-phenotype associations in the list of SNPs determined from SNP-by-SNP test for association. PSTPS is applicable whenever P values can be computed for testing the X - Y associations. The PSSC procedure does not require X to be continuous; this procedure is applicable whenever a proper real-valued distance function $d(\cdot, \cdot)$ can be defined on the genomic feature space. One other case often seen in applications but not considered in Section 2 is continuous X and continuous Y , for which supervised principal components (Bair *et al.* 2006) is an obvious choice.

In extending Algorithm 2 to handle censored failure time phenotype, the censored samples are regarded as if they never fail; this is certainly incorrect for building a prediction model. It should be noted however, that the purpose of this procedure is not to build a prognostic model for the phenotype based on the gene expression profile; rather this algorithm serves as a dimension reduction mechanism for constructing a global association statistic to measure the X - Y stochastic dependence in the association profile. The global association statistic is constructed in a way accounting for time to failure and censoring, using for example the log-rank test statistic. The global test of Goeman *et al.* (2004) would appear to be applicable to this kind of phenotype via for example Cox's regression model. As discussed in Introduction however, the global test is only applicable to either the set of all genes (features) on the array or a set of genomic features prespecified without involving the phenotype, which is fundamentally different from an association profile.

Operationally, the SDV approach is essentially a one-round cross validation, whereas the profile significance approach uses permutation (or re-sampling). Advantages and drawbacks of these two approaches are investigated on a few aspects in this paper, and considering these aspects overall, the results would be in favor of the latter. This is a first step toward developing meaningful and rigorous inferences for statistical validation in genome-wide association studies. A number of open problems remain for future research.

The power dilution of PSSC and PSTPS seen in the simulation study reflects an aspect of “the curse of high dimensionality.” As the screening gets less stringent the size of the association profile (number of genes declared as associated with the phenotype) increase: with $m = 10000$, under the complete null hypothesis simulated by the permutation, the expected sizes of the association profiles are 5, 10, 50, 100 respectively for $\gamma = 0.0005, 0.001, 0.005, 0.01$; for example on the 1000th simulated dataset, the average size of the association profile obtained from the 400 permuted datasets is 57.8 for $\gamma = 0.005$. So the number of features (dimension) used in the clustering procedure or the P value score statistic increase; yet the total sample size $n = 120$ and the number of samples with phenotype $Y = 1$, $n_1 = 24$ hold constant. When a massive number of possible features are available ($m = 10000$), even if there is no association exists (simulated by permutation), simply by chance one can frequently find enough number of features that separate the two phenotypic groups of the limited number of samples equally well or even better than what is observed on the original dataset. Modifications of these procedures to overcome the curse of dimensionality will be studied in the future.

In practice whenever allowed by the sample size, a large number of permutations or re-sampling (at least 2000) should be performed for the profile significance procedures. PSTPS is almost trivial to compute and does not require much computing time even for a large number of permutation (simulation) runs. The current implementation of PSSC takes more computing time due to the overhead in computing the distances and cluster centroids; empirically for 10,000 or so genomic features, 300 or so samples and 2,000 simulation rounds it may take a couple of hours. A more efficient implementation is underway.

In the simulation study only 400 permutations are used for sake of the computing time to finish the simulation study. The small number of permutations could be one of the reasons for the liberalness of PSTPS and PSSC seen in the simulation results. Furthermore, Efron (2004) cautions the inability of simple permutations to adequately cover the variability under the null hypothesis and proposes an alternative for deriving the null distribution; this aspect in the context of profile significance estimation needs to be studied in the future.

Figure 3 suggests that in practice one can plot the profile significance against the estimated FDR (or q value, or local FDR) for a grid of significance threshold values γ on the $[0, 1] \times [0, 1]$ square; this would provide a more comprehensive assessment of the association profiles generated from various significance thresholds, in addition to the q -value plot (Storey and Tibshirani, 2003). Intuitively one would like to select γ values that give points close to the lower-left corner; developing criteria for such a selection is worth future efforts.

A theoretical question relevant to practice is to what asymptotic distribution the estimated profile significance $\widehat{PS}_c$ converges as $m \rightarrow \infty$, and $n \rightarrow \infty$ at a slower rate (so $m/n \rightarrow \infty$) and the number of permutations $L/n \rightarrow \infty$. This will provide a clear understanding of the operating characteristics observed in the simulation results. The problem may not be analytically tractable in general but it can be feasible to obtain an analytical answer for more structured comparison rules such as PSTPS.

Acknowledgments

This research is funded in part by NIH/NIGMS Pharmacogenetics Research Network and Database (U01 GM61393, U01GM61374, <http://pharmgkb.org/>) from the National Institutes of Health, Cancer Center Support Grant P30 CA-21765, and the American Lebanese and Syrian Associated Charities (ALSAC). The author is grateful to Dr. Stan Pounds and two referees for their comments and suggestions that greatly helped strengthen this paper.

References

- Allison DB, Gadbury GL, Heo M, Fernandez JR, Lee C-K, Prolla TA, Weindrich R. A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis* 2002;39:1–20.
- Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine* 2000;19:453–473. [PubMed: 10694730]
- Bair E, Hastie T, Paul D, Tibshirani R. Prediction by supervised principal components. *Journal of the American Statistical Association* 2006;101:119–137.
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Ser B* 1995;57:289–300.
- Benjamini Y, Krieger AM, Yekutieli D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* 2006;93:491507.
- Bogni A, Cheng C, Liu W, Yang W, Pfeffer J, Mukatira S, French D, Downing JR, Pui C-H, Relling MV. Genome-Wide Approach to Identify Risk Factors for Therapy-Related Myeloid Leukemia. *Leukemia* 2006;20:239–246. [PubMed: 16341039]
- Bøvelstad HM, Nygard S, Støvdal HL, Aldrin M, Borgan Ø, Frigessi A, Lingjæde OC. Predicting survival from microarray data – a comparative study. *Bioinformatics* 2007;23:2080–2087. [PubMed: 17553857]
- Bullinger L, Döhner K, Bair E, Fröhling S, Schlenk RF, Tibshirani R, Döhner H, Pollack JR. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *New England Journal of Medicine* 2004;350:1605–1616. [PubMed: 15084693]
- Cheng, C. Optimality: The Second Erich L. Lehmann Symposium. Vol. 49. Institute of Mathematical Statistics; Beachwood, OH, USA: 2006. An adaptive significance threshold criterion for massive multiple hypothesis testing; p. 51-76.
- Cheng C, Pounds S. False discovery rate paradigms for statistical analyses of microarray gene expression data. *Bioinformatics* 2007;1:436–446. [PubMed: 17597936]
- Cheng, C.; Pounds, S.; Boyett, JM.; Pei, D.; Kuo, M-L.; Roussel, MF. Statistical significance threshold criteria for analysis of microarray gene expression data. *Statistical Applications in Genetics and Molecular Biology*. 2004. 3, Article 36. URL/www.bepress.com/sagmb/vol3/iss1/art36
- Edick MJ, Cheng C, Yang W, Cheok M, Wilkinson MR, Pei D, Evans WE, Kun LE, Pui C-H, Relling MV. Lymphoid Gene Expression as a predictor of risk of secondary brain tumors. *Genes, chromosomes & Cancer* 2005;42:107–116. [PubMed: 15543619]
- Efron B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of The American Statistical Association* 2004;99:96–104.
- Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 2001;96:1151–1160.
- Fleiss, JL.; Levin, B.; Paik, MC. *Statistical Methods for Rates and Proportions*. Vol. 3. John Wiley & Sons Inc; Hoboken, New Jersey: 2003.
- Goeman JJ, van der Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004;20:93–99. [PubMed: 14693814]
- Owen A. Variance of the number of false discoveries. *Journal of Royal Statistical Society, Ser B* 2005;67:411–426.
- Pounds S, Cheng C. Robust estimation of the false discovery rate. *Bioinformatics* 2006;22:1979–1987. [PubMed: 16777905]
- Simon R. Development and Validation of Therapeutically Relevant Multi-Gene Biomarker Classifiers. *Journal of the National Cancer Institute* 2005a;97:866–867. [PubMed: 15956642]

- Simon R. Roadmap for Developing and Validating Therapeutically Relevant Genomic Classifiers. *Journal of Clinical Oncology* 2005b;23:7332–7341. [PubMed: 16145063]
- Storey JD, Taylor J, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of Royal Statistical Society, Ser B* 2003;66:187–205.
- Storey JD. A direct approach to false discovery rates. *Journal of Royal Statistical Society, Ser B* 2002;64:479–498.
- Storey JD, Tibshirani R. Statistical significance for genome-wide studies. *Proceedings of National Academy of Science* 2003;100:9440–9445.
- WTCCC & AAASC. Association of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nature Genetics* 2007;39:1329–1337. [PubMed: 17952073]

Figure 1.
An example of 2 continuous genomic features in the association profile and a binary phenotype depicted by bullets and circles.

Figure 2.

Empirical ROC curves of the PSTPS, PSSC, and SDV procedures: PSTPS, PSSC, and SDV each is represented by solid, long-dash, and dash-dot line respectively. The 1%, 5% and 10% nominal levels are indicated by the three vertical dotted lines, and on each curve points corresponding to these nominal levels each is indicated by circle, triangle, and diamond respectively.

Figure 3.

Across-simulation sample summary statistics of profile significance values and validation P values plotted against estimated actual FDR and FNR at each γ value. The summary statistics are mean, one standard deviation above the mean, and 10%, 50%, and 90% percentiles. Each point corresponds to a γ value and the four points representing mean or median are connected by line segments. Solid line and bullet represent PSTPS, long-dash line and square represent PSSC, and dash-dot line and triangle represent validation P value (SDV). The change in the mean validation P value from one γ value to another is too subtle to show up in graph.

Figure 4.

Empirical conditional joint distribution of FPP (FNP) on the whole dataset and training set given SDV is more significant than PSSC, for less stringent screening ($\gamma = 0.005, 0.01$). Points falling below the diagonal line indicate the cases that the association profile obtained on the whole dataset has higher error level.