Published in final edited form as:

Stat Appl Genet Mol Biol. 2014 April 1; 13(2): 217–226. doi:10.1515/sagmb-2013-0062.

Efficient identification of context dependent subgroups of risk from genome wide association studies

Greg Dyson¹ and Charles F. Sing^{2,*}

¹Department of Oncology, Wayne State University, Detroit, Michigan, USA

²Department of Human Genetics, University of Michigan, Ann Arbor, Michigan, USA

Abstract

We have developed a modified Patient Rule-Induction Method (PRIM) as an alternative strategy for analyzing representative samples of non-experimental human data to estimate and test the role of genomic variations as predictors of disease risk in etiologically heterogeneous sub-samples. A computational limit of the proposed strategy is encountered when the number of genomic variations (predictor variables) under study is large (> 500) because permutations are used to generate a null distribution to test the significance of a term (defined by values of particular variables) that characterizes a sub-sample of individuals through the peeling and pasting processes. As an alternative, in this paper we introduce a theoretical strategy that facilitates the quick calculation of Type I and Type II errors in the evaluation of terms in the peeling and pasting processes carried out in the execution of a PRIM analysis that are underestimated and non-existent, respectively, when a permutation-based hypothesis test is employed. The resultant savings in computational time makes possible the consideration of larger numbers of genomic variations (an example genome wide association study is given) in the selection of statistically significant terms in the formulation of PRIM prediction models.

Keywords

PRIM; classification; GWAS

Introduction

We have been modifying the Patient Rule-Induction Method (PRIM), introduced by (Friedman and Fisher, 1999), to simultaneously address the problems of predictor selection, etiological heterogeneity and non-additivity of predictor effects (Dyson, et al., 2007; Dyson, et al., 2009). The augmented PRIM is a non-traditional, non-parametric, alternative statistical strategy for building multiple prediction models that takes into account the reality that the etiology of a common human disease is heterogeneous in the population at large and provides an evaluation of the role of the contribution of non-additive effects of predictor variables. The PRIM produces the subset of predictor variables and their values that define each of the mutually exclusive partitions of the total sample of individuals that vary

Correspondence to: Charles F. Sing., Department of Human Genetics, University of Michigan, 1241 E. Catherine Street, 5928 Buhl Building, Ann Arbor, MI 48109-0618, USA, Ph: 734-764-5492; Fax: 734-763-5277; csing@umich.edu.

significantly in risk of disease. The multiplicity of the partitions of individuals identified by this analytical strategy is consistent with the assumption that there are multiple etiological pathways leading to disease and that one predictive model is not sufficient for describing all cases of disease. The multiple models of varying risk of disease produced by PRIM allow more individualized prediction of risk for particular subsets of individuals in the population under study.

We have previously applied the augmented PRIM to predict ischemic heart disease incidence, illustrating the context-dependent added predictive value of genetic variations, using data collected by the Copenhagen City Heart Study (Dyson, et al., 2007; Dyson, et al., 2009). Other authors have used a PRIM approach to predict differential gene expression (Cole, et al., 2003), define prognostic groups for myeloma (LeBlanc, et al., 2002), and to identify haplotypes that are most likely to share a common ancestor (Yu, et al., 2004). Additionally, another research group published two papers using a PRIM approach to predict mortality in elderly individuals and hyperglycemia in intensive care unit patients (Nannings, et al., 2008; Nannings, et al., 2008). In a previous paper we introduced a permutation strategy to test the statistical significance of each term (defined by selected values of a predictor variable) considered in defining a partition (Dyson, et al., 2009). The objective of this manuscript is to present an alternative theoretical framework for hypothesis testing of each term considered in the peeling and pasting processes that are central to the PRIM. Two simulated datasets will be used to illustrate the difference between a distribution-based and permutation-based approach to compute p-values. An analysis of the prostate cancer case control genome-wide database is also reported.

Methods

PRIM background

Mutually exclusive partitions of individuals, characterized by terms defined by selected values of predictor variables, with a significantly higher proportion (θ) of 1's for a 0–1 binary response variable than expected under the null distribution are produced by the PRIM. The partitioning (or classification) is executed through repeated implementations of the peeling and pasting algorithms constrained by the minimum proportion of unassigned individuals that must be included in the definition of a valid partition, denoted as the support parameter (β). Terms defined by combinations of values of multiple predictor variables constructed using the Combinatorial Partitioning Method (Nelson, et al., 2001) facilitate the inclusion of non-additive effects in defining partitions. Dyson, et al. (2007) present a detailed description of the peeling and pasting algorithms and a strategy for selection of the support parameter via a grid search. Briefly, peeling is an iterative process that creates a partition by excluding individuals with particular values of predictor variables (creating a peeling term), while pasting iteratively amends individuals to the partition, also based upon values of predictor variables (creating a pasting term), after the peeling stage has been completed (Dyson, et al., 2007). At each step when a term may be used to (further) define a partition, a test is employed to determine if the respective increase in the mean of the response variable when utilizing the current term to define the current partition compared to the partition without the current term defining it is statistically significant. A permutation

algorithm for generating a null distribution of θ for testing the statistical significance of each term generated, a multiple testing adjustment to correct for all of the hypotheses (terms) tested and a confidence interval for the estimate of θ associated with each partition, that enables one to make comparisons of estimates across partitions, have been incorporated into the PRIM by Dyson, et al. (2009).

Approximately 90% of the computing time required for carrying out a PRIM analysis is devoted to calculating the null distribution of the maximum θ , via permutations, for each of the sequential terms that is considered in the peeling and pasting steps for each β examined. As the number of predictor variables increases, the number of terms tested by the PRIM, in both the original analysis and each of the permutation analyses for each β examined, increases combinatorially. Therefore the evaluation of a large number of predictors (such as in a genome-wide association study) dictates that an alternative to permutation testing be developed. Below we present a realistic alternative based on a derivation of the theoretical distribution of the maximum θ .

Peeling hypothesis testing modification

a) Type I error—The objective of a call to the PRIM peeling algorithm when considering a binary (0 or 1) response in an input population of size n is to create a sample of size m with a statistically significantly higher proportion of 1's than observed in the input population. Let m_1 denote the number of 1's in a sample of size m taken without replacement from the population of size n which has n_1 1's. The PRIM peeling algorithm will select the sample which results in the largest $m_1/m = \hat{\theta}$ among all potential samples that fulfill the requirements for testing: $m \geq n^*\beta$ and $m_1/m > n_1/n$. The minimum size of a potential sample is determined by β , i.e. $m \geq n^*\beta$, while the requirement that $m_1/m > n_1/n$ is a consequence of the PRIM peeling objective. The metric that is used to evaluate whether a potential sample results in a statistically significantly higher proportion of 1's is the largest $\hat{\theta}$. Thus the null distribution for θ is required to derive the null distribution for the maximum θ . We can decompose the probability density function of the θ under the null distribution as follows using V to indicate that both $m_1/m > n_1/n$ and $m \geq n^*\beta$ are met:

$$\begin{aligned}
f(\theta|V) &= \sum_{j=\lceil \beta \times n \rceil}^{n=1} \{f(\theta, m \\
&= j|V)\} \\
&= \sum_{j=\lceil \beta \times n \rceil}^{n=1} \{f(\theta|m \\
&= j, V) \\
&\times p(m \\
&= j|V)\} = \sum_{j=\lceil \beta \times n \rceil}^{n=1} \{f(\frac{m_1}{m}|m \\
&= j, V) \\
&\times p(m \\
&= j|V)\} \\
&= \sum_{j=\lceil \beta \times n \rceil}^{n=1} \{f(m_1|m \\
&= j, V) \times p(m=j|V)\}
\end{aligned} \tag{1}$$

We assume that the unconditional marginal distribution of m is discrete uniform for this derivation. The second component of the last line of equation (1) can be computed analytically by calculating the empirical probability mass function of the m values from all potential samples that satisfy V . The first component in the last line of equation (1) can only be evaluated from its density function. The probability mass function for m_1 follows a non-central hypergeometric distribution (Fog, 2008; Wallenius, 1963):

$$f(m_1; n_1, n, m, \omega) = \binom{n_1}{m_1} \binom{n - n_1}{m - m_1} \int_0^1 (1 - x^{\omega/D})^{m_1} (1 - x^{1/D})^{m - m_1} dx, \tag{2}$$

where $D = \omega(n_1 - m_1) + (n - n_1) - (m - m_1)$ and ω is the relative odds of selecting a 1 to selecting a 0. Under the null distribution of unbiased selection of 1's or 0's (equivalently $\omega = 1$) equation (2) reduces to a hypergeometric distribution with parameters m, n_1, n . Note that only combinations of values of m and m_1 that satisfy V are included in the calculation of the first component of equation (1).

The number of possible θ values is large and countable. To ease the burden of computation we used 10,000 values evenly spaced between 0.0001 and 1.0 in estimating the probability density function of θ . Probabilities computed from equation (1) were rounded up to the nearest 1/10,000 decimal category. The result is a conservative approximation of the null distribution of a single draw from all possible samples that fulfill V . Since the PRIM peeling algorithm selects the sample with the largest $\hat{\theta}$ from the N samples that satisfy V , it follows that the null probability distribution function of the largest $\hat{\theta}$ is given by the probability

density function of the maximum order statistic for equation (1), namely $N \times f(\theta|V) \times F(\theta|V)^{N-1}$ where $F(\theta|V)$ is the cumulative distribution function for $f(\theta|V)$.

b) Type II error—The derivation of the probability density function of maximum θ presented above for computing the Type I error can also be used to compute an alternative theoretical distribution of the maximum θ , indexed by ω . A value of ω greater than 1.0 in equation (2) results in a probability density function for θ where the odds of drawing a 1 is greater than the odds of drawing a 0 - exactly the non-central situation we wish to model under the alternative to the null distribution. The selection of ω will depend on the amount of bias in the selection mechanism that is considered under the alternative hypothesis.

Pasting hypothesis testing modification

a) Type I error—We begin with a sample of size k (consisting of k_1 1's) from a population of size n (consisting of n_1 1's) that has been defined by one or more applications of the peeling algorithm and zero or multiple applications of the pasting algorithm. The objective of a call to the PRIM pasting algorithm is to determine whether adding, i.e. pasting, a sample from the remainder population of size $n - k$ to the existing sample of size k will result in a statistically significantly higher proportion of 1's than observed in the k responses. Next, let t_1 denote the number of 1's out of a sample of size t taken without replacement from the remainder population of size $n - k$ which has $n_1 - k_1$ 1's. A valid pasting sample will have the requirements that $(k_1 + t_1) / (k + t) > k_1 / k$ and $t_1 / (k - k_1 + t) < k_1 / k$. The first requirement ensures that the proportion of 1's is larger than the proportion of 1's in the current sample; while the second requirement ensures that the sample of size t could not have resulted in a larger $\hat{\theta}$ than from the current sample of size k (i.e., it would have already been chosen to define the current sample). The PRIM algorithm will select the amended sample which results in the largest $(k_1 + t_1) / (k + t) = \hat{\theta}$ among all potential amended samples. The metric that is used to evaluate whether the amended sample results in a statistically significantly higher proportion of 1's is this largest $\hat{\theta}$.

As is the case in the application of the peeling algorithm, the null distribution for the θ is required to test the statistical significance of the largest $\hat{\theta}$. We can decompose the probability density function of θ under the null distribution as follows using W to indicate that $(k_1 + t_1) / (k + t) > k_1 / k$ which reduces to $t_1 / t > k_1 / k$ and $t_1 / (k - k_1 + t) < k_1 / k$ are satisfied:

$$\begin{aligned}
g(\theta|W) &= \sum_{j=1}^{n-k-1} \{g(\theta, t \\
&= j|W)\} \\
&= \sum_{j=1}^{n-k-1} \{g(\theta|t \\
&= j, W) \\
&\times p(t \\
&= j|W)\} = \sum_{j=1}^{n-k-1} \{g(\frac{k_1+t_1}{k+t}|t \\
&= j, W) \\
&\times p(t \\
&= j|W)\} \\
&= \sum_{j=1}^{n-k-1} \{g(t_1|t \\
&= j, W) \times p(t=j|W)\}
\end{aligned} \tag{3}$$

We assume that the unconditional marginal distribution of t is discrete uniform for this derivation. The second component of the last line of equation (3) can be derived analytically by calculating the empirical probability mass function of the t values from all potential samples that satisfy W . Again, as is the case for the peeling algorithm, the first component of the last line of equation (3) can only be evaluated from its density function. The probability mass function for t_1 follows a non-central hypergeometric distribution (Fog, 2008; Wallenius, 1963):

$$f(t_1; n_1 - k_1, n - k, t, \omega) = \binom{n_1 - k_1}{t_1} \binom{(n - k) - (n_1 - k_1)}{t - t_1} \int_0^1 (1 - x^{\omega/D})^{t_1} (1 - x^{1/D})^{t-t_1} dx, \tag{4}$$

where $D = \omega((n_1 - k_1) - t_1) + ((n - k) - (n_1 - k_1)) - (t - t_1)$ and ω is the relative odds of selecting a 1 to selecting a 0. Under the null distribution of unbiased selection of 1's or 0's (equivalently $\omega = 1$), equation (4) reduces to a hypergeometric distribution with parameters t , $n_1 - k_1$ and $n - k$. Note that only combinations of values of t and t_1 that satisfy W are included in the calculation of the first component of equation (3).

The probability density function of θ produced by a single draw from all possible samples that satisfy W is estimated in the same manner as described above for the application of the peeling algorithm. Again, as in the application of the peeling algorithm, the PRIM pasting algorithm chooses the sample from among all possible samples (N) which results in the largest $\hat{\theta}$, we use the density of the maximum order statistic for equation (3), namely $N \times$

$g(\theta|W) \times G(\theta|W)^{N-1}$ where $G(\theta|W)$ is the cumulative distribution function for $g(\theta|V)$, as the probability density function for testing the statistical significance of the largest $\hat{\theta}$.

b) Type II error—The derivation of the probability density function of maximum θ presented above for computing the Type I error can also be used to compute an alternative distribution, indexed by ω . A value of ω greater than 1 in equation (4) results in a probability density function for the maximum θ where the odds of selecting a 1 is greater than the odds selecting a 0, exactly the non-central situation we wish to model under the alternative distribution. The selection of ω will depend on the prior assumptions about the amount of bias in the selection of 1's that is considered under the alternative hypothesis.

Databases

Simulated databases

We compared the permutation and theoretical strategies for generating a null distribution and calculating a p-value for a term using two simulated data sets each with 200 categorical predictor variables (with a random number of categories), $n = 1000$ and $n_I = 150$. The program for creating the simulated data using the R programming environment (RCoreTeam, 2013). Both datasets were constructed so the response variable was, in theory, not significantly associated with any of the categorical predictor variables. The first dataset was created to compare the null distributions of the maximum θ produced by the two strategies in the situation where the predictor variables are, in theory, all mutually independent. The second dataset was constructed to illustrate the impact on the comparison of the two null distributions of the maximum θ when the predictor variables are not mutually independent. For this purpose, the simulation was used to create a dataset in which each predictor variable is highly associated with one other predictor variable (at least 98 % equivalent values). The Epanechnikov kernel, with a bandwidth of 0.002 was used to estimate the smoothed null probability distributions generated by the permutation and theoretical strategies.

We use the same two datasets generated above for demonstrating the computation for a peeling term to generate the null distribution of maximum θ and compare the distribution of the p-values produced by the theoretical and permutation strategies for a term in the pasting process. We assume that the peeling process is completed and we have obtained a sample defined by one or more terms which has 90 1's and 210 0's ($\hat{\theta} = 0.30$) from a population of 240 1's and 1060 0's ($\theta = 0.18$), which is significant with $p < 0.0001$ using the theoretical strategy. We proceed to illustrate the computation of p-values for a single pasting term assuming that the remaining 150 1's and 850 0's not assigned to the sample produced by the peeling algorithm are composed of either dataset 1 or 2, as previously described. The Epanechnikov kernel, with a bandwidth of 0.0003 was used to estimate the smoothed null probability distributions generated by the permutation and theoretical strategies.

Prostate cancer database

We next applied this new PRIM algorithm using the distribution-based p-value calculation to the controlled access dbGAP database (Mailman, et al., 2007): Cancer Genetic Markers of

Susceptibility Prostate Cancer Genome-Wide Association Study (GWAS) - Primary Scan (Stage 1) - PLCO Screening Trial (phs000207.v1.p1) (Yeager, et al., 2007). The database consists of 317,402 non-monomorphic SNPs measured in 1101 controls and 492 non-aggressive (Gleason Score < 7 and Stage < III) and 659 aggressive (Gleason Score 7 or Stage III) prostate cancer cases. All participants were non-Hispanic whites. For the purposes of this analysis, we combine the cases into a single response group. As this is a case-control database, care must be taken when interpreting the resultant context-dependent partitions identified by the PRIM. To that end, we will produce odds ratios for each of the terminal partitions relative to the remainder partition.

Analysis of simulated dataset 1—In this dataset there are 282 possible peeling terms, using $\beta = 0.10$ and one predictor variable, that may be used to define the first term in the first partition. One of the predictor variables could produce 5 possible terms, 1 could produce 4 possible terms, 14 could produce 3 possible terms, 56 could produce 2 possible terms, 119 could produce 1 term only and 9 could produce 0 terms. Among these 282 possible terms, the proportion of 1's was maximized, 31 1's and 123 0's, using a term defined by outcome 1 of variable 187. Using a nominal significance level of 0.0232, which corresponds to an experiment-wise Type I error rate of 0.05, the critical test value defined by the null distribution of the maximum θ obtained using the theoretical strategy was 0.2584 (Figure 1A). The largest observed $\hat{\theta} = 31/154 = 0.2013$ is less than this critical test value and thus not statistically significant, $p = 0.9277$. The choice of 0.2543 as the critical test value under the null hypothesis results in at least 0.80 power when the alternative theoretical distribution of the maximum θ is defined by an ω of 1.4880 or greater (Figure 1A).

To estimate the null distribution of the maximum θ for the first peeling term of the first partition for dataset 1 using the permutation method, 1,000,000 permutations of the response variable were assembled and a one term PRIM analysis was performed on each permuted sample using the original predictor set. The critical test value based on the permutation strategy using a nominal significance level of 0.0232 was 0.2548 and the p-value obtained for the largest observed $\hat{\theta}$ of 0.2013 was 0.9438. Although the statistical decision in this example would be identical using either strategy, there are subtle differences in the shapes of the distribution functions (Figure 1A). Some of this difference is due to the fact that the number of potential terms varies among the 1,000,000 permutations. In this dataset the number of terms generated by the permutation strategy was approximately normally distributed with a mean of 279.2 and standard deviation of 5.0.

For dataset 1 there are 9 possible pasting terms (1 predictor variable could produce 2 possible terms, 7 could produce 1 possible term and 192 could produce 0 possible terms) that satisfied the validity criteria. Among these 9 terms, the term defined by outcome 0 for variable 62 (consisting of 2 1's and 2 0's) resulted in an amended sample with the largest updated $\hat{\theta}$ of $(90+2) / (300+4) = 0.3026$. Using a nominal significance level of 0.0232, which corresponds to an experiment-wise Type I error rate of 0.05, the critical test value utilizing the theoretical strategy is 0.3107 and there is power of 0.80 to detect an effect with ω 2.3572 (Figure 1C). The largest observed $\hat{\theta}$ of 0.3026 is less than the critical test value and thus not statistically significant, $p = 0.9470$.

The critical test value for the permutation-based strategy (1,000,000 permutations used) using a nominal significance level of 0.0232 is 0.3092, while the p-value for significance of the observed $\hat{\theta}$ is 0.8089. The average number of terms in each permutation is 6.6 (standard deviation = 2.5). The shape of the distribution function for the permutation-based strategy (Figure 1C) has multiple peaks due to the discreteness of the resultant maximum θ found in the permutations. Only 50 of 22,164 possible terms that satisfy the validity criteria were observed in the 1,000,000 permutations.

Analysis of simulated dataset 2—In this dataset, there are 271 possible terms, again using $\beta = 0.10$ and one predictor variable, that may be used to define the first term of the first partition. The largest $\hat{\theta} = 0.1929$ was obtained in the sub-sample defined by outcome 0 of variable 63, which consisted of 27 1's and 113 0's. The critical test values for the theoretical and permutation strategies are 0.2538 and 0.2524, respectively using a nominal significance level of 0.0232. The choice of 0.2538 as the critical test value under the null hypothesis results in at least 0.80 power when the alternative theoretical distribution of the maximum θ is defined by an ω of 1.4883 or greater (Figure 1B). The resultant p-values for the theoretical and permutation strategies associated with the largest observed $\hat{\theta}$ are 0.9937, and 0.9624, respectively. The average number of terms in each permutation was 280.2 (SD = 11.2). Figure 1B shows that the permutation distribution has moved farther away from the theoretical distribution for the analysis of this dataset. This is due to the fact that the effective number of terms considered by the permutation method is approximately 100, not 200, because of the associations between predictor variables programmed into dataset 2.

Among the 6 possible pasting terms in dataset 2, the term defined by outcome 0 of variable 162 (consisting of 2 1's and 2 0's) resulted in an amended sample with the largest updated $\hat{\theta}$ of $(90+2) / (300+4) = 0.3026$. Using a nominal significance level of 0.0232, the critical test value is 0.3101 and there is power of 0.80 to detect an effect with $\omega = 2.5339$ employing the theoretical strategy. The resultant p-value for significance of this pasting term is 0.8590. The critical test value for the permutation-based strategy (derived from 1,000,000 permutations) using a nominal significance level of 0.0232 is 0.3092 while the p-value for significance of the observed term is 0.6448. The average number of terms in each permutation is 4.2 (standard deviation = 2.6). The shape of the permutation-based null distribution (Figure 1D) is again discrete due to having only 50 unique values of the maximum θ to create the null distribution.

Analysis prostate cancer GWAS—For the analysis of the prostate cancer database, we examined each SNP one-at-a-time. The optimal support parameter was chosen to be 0.215, using an experiment-wise error rate of 0.05. There was only one significant partition which consisted of only one peeling term (p-value = 0.0058). SNP rs4721099 = TT or TC was defined as the high risk subgroup of 503 individuals. The remainder group thus consisted of the 1749 individuals with either CC or missing for rs4721099. The resultant odds ratio (95% confidence interval) of this high risk partition compared to the remainder partition computed using logistic regression was 1.68 (1.37, 2.06); $p < 5 \times 10^{-7}$. This SNP is in the inter-genic region on chromosome 7 between LOC100419641 and LOC389465, with no prior basis for

a functional deleterious effect. This SNP association has not been replicated by any subsequent prostate cancer studies.

Discussion

The application of the PRIM to datasets with only a few thousand predictor variables and a few thousand observations is computationally infeasible when a permutation test for each peeling and pasting term is employed. For instance, the GWAS dataset analyzed in this paper with $n = 2252$ and $m = 317,402$, 2,000 permutations for one term using one variable to define a term for one support parameter takes an average of 7.5 days. If we assume that there is an average of 4 terms tested for each of the 91 potential support parameters (0.05 to 0.50 in increments of 0.005), it would take 30 days to finish one PRIM analysis on 91 CPUs. Therefore, we turned to developing an alternative theoretical distribution-based framework that utilizes properties of the non-central hypergeometric distribution for hypothesis testing that eases the computational burden (the above illustration takes 6.5 hours to complete using the distribution-based approach). By using the theoretical alternative, the computational time is reduced by approximately 90%, allowing larger datasets to be analyzed. Now the most computationally intensive component of the algorithm is searching through all of the input variables for the maximum θ . In addition this new approach enables one to compute power (not possible using a permutation-based approach) and produces more accurate p-values than the permutation-based hypothesis testing strategy.

A permutation-based strategy for computing the probability density function for maximum θ is highly influenced by the relative frequencies of the levels of the input predictor variables, their correlations with each other and their relationships with the response variable; implying that a permutation-based derivation may not estimate an appropriate null distribution (Benjamini, et al., 2008; Efron, 2008). This follows because only a limited number of possible terms can be created by the permutation strategy, whereas the theoretical strategy introduced in this paper derives a null distribution formed from all possible terms. This subtle, but important, difference is illustrated in the analyses of simulated dataset 2 relative to dataset 1. Since all predictor variables in dataset 2 are highly related to one other predictor variable, the true number of 'effective' terms is much less than the 271 generated by dataset 1, leading to the permutation-based distribution deviating farther from the theoretical distribution in dataset 2. Reality will be somewhere between these two extreme dataset examples (near-complete independent and near-complete dependent predictor variables).

The theoretical calculation considers the impact of unobserved or unmeasured predictor variables by enumerating and deriving the distribution of the resultant θ from all possible valid sub-samples. The permutation-based computation only explores a fraction of the possible subsamples that are observable when utilizing the theoretical calculation. For example, there are 22,164 unique combinations of t and t_1 that satisfy W for the inclusion of one pasting term for both simulated datasets, of which, only 50 (not the same 50 in the two analyses) were observed in the 1,000,000 permutations for each dataset. Consequently terms, and the associated θ s, that could not be observed using a permutation-based method are not included in the calculation of the null distribution, leading to the discrete appearance

of the null distributions in Figures 1C and 1D. The number of potential terms is a random variable when using a permutation-based p-value calculation which impacts the null distribution that is produced. As indicated by (Benjamini, et al., 2008), the number of potential terms will vary across permutations (just like the number of tested hypotheses varies in his example). In particular, in analyses using the PRIM there are typically few potential pasting terms across all partitions. Thus there is a potential that, in a pasting analysis permutation, there are no valid pasting terms. In fact, of the 1,000,000 permutations for the pasting component using dataset 2, 19,519 (2 %) had no possible pasting terms.

The GWAS example given in this paper had only a little over 300,000 predictor variables examined for inclusion in the model one variable at a time. The software that we have developed allows for an almost unlimited number of variables examined singly or 2 or more at one time; i.e., a non-additive interaction effect may be the optimal predictor of risk, while also controlling for experiment-wise error (Dyson, et al., 2009). The original publication attached to this database (Yeager, et al., 2007) focused on SNPs in 8q24. The PRIM algorithm did not identify any SNPs in 8q24, partially because the PRIM methodology addresses a different question than standard analysis approaches. The objective of a PRIM analysis is to identify subsets of the input sample defined by variables (e.g., SNPs) with a higher prevalence of a given phenotype; while the objective of standard GWAS analysis is to identify variations are associated with a particular phenotype using the entire input sample. With the use of PRIM, we were able to rapidly interrogate the entire breadth of the measured SNPs (utilizing an experiment-wise error rate correction mechanism embedded in the algorithm) to identify a high-risk partition and implicate a SNP associated with prostate cancer risk.

Acknowledgments

The authors wish to thank Richard Merkle and Alex Sadovsky for their assistance in programming.

Funding This work was supported by National Institutes of Health grant number HL072905, National Institute of General Medical Science grant number GM065509 and National Cancer Institute grant number 5P30CA022453.

Bibliography

- Benjamini Y, et al. Comment: Microarrays, empirical Bayes and the two-groups model. *Stat. Sci.* 2008; 23:23–47.
- Cole SW, Galic Z, Zack JA. Controlling false-negative errors in microarray differential expression analysis: a PRIM approach. *Bioinformatics.* 2003; 19:1808–1816. [PubMed: 14512352]
- Dyson G, et al. An application of the patient rule-induction method for evaluating the contribution of the Apolipoprotein E and Lipoprotein Lipase genes to predicting ischemic heart disease. *Genetic Epidemiology.* 2007; 31:515–527. [PubMed: 17436307]
- Dyson G, et al. Modifications to the Patient Rule-Induction Method That Utilize Non- Additive Combinations of Genetic and Environmental Effects to Define Partitions That Predict Ischemic Heart Disease. *Genetic Epidemiology.* 2009; 33:317–324. [PubMed: 19025787]
- Efron B. Microarrays, empirical Bayes and the two-groups model. *Stat. Sci.* 2008; 23:1–22.
- Fog A. Calculation methods for Wallenius' noncentral hypergeometric distribution. *Communications in Statistics-Simulation and Computation.* 2008; 37:258–273.
- Friedman JH, Fisher NI. Bump hunting in high-dimensional data. *Stat. Comput.* 1999; 9:123–143.
- LeBlanc M, Jacobson J, Crowley J. Partitioning and peeling for constructing prognostic groups. *Stat. Methods Med. Res.* 2002; 11:247–274. [PubMed: 12094758]

- Mailman MD, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nature genetics*. 2007; 39:1181–1186. [PubMed: 17898773]
- Nannings B, Abu-Hanna A, de Jonge E. Applying PRIM (Patient Rule Induction Method) and logistic regression for selecting high-risk subgroups in very elderly ICU patients. *Int. J. Med. Inform.* 2008; 77:272–279. [PubMed: 17646128]
- Nannings B, Bosman RJ, Abu-Hanna A. A Subgroup Discovery Approach for Scrutinizing Blood Glucose Management Guidelines by the Identification of Hyperglycemia Determinants in ICU Patients. *Methods of Information in Medicine*. 2008; 47:480–488. [PubMed: 19057804]
- Nelson MR, et al. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res*. 2001; 11:458–470. [PubMed: 11230170]
- RCoreTeam. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria: 2013.
- Wallenius, K. Department of Statistics. Stanford University; 1963. Biased Sampling: The Non-central Hypergeometric Probability Distribution.
- Yeager M, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature genetics*. 2007; 39:645–649. [PubMed: 17401363]
- Yu K, Martin RB, Whittemore AS. Classifying disease chromosomes arising from multiple founders, with application to fine-scale haplotype mapping. *Genetic Epidemiology*. 2004; 27:173–181. [PubMed: 15389930]

Figure 1.

Density plots of the maximum θ for the theoretical distribution derived from the 1,000,000 permutation analyses and theoretical alternative distribution (with θ chosen to correspond to 0.80 power) for (A) one peeling term from simulated dataset 1; (B) one peeling term from simulated dataset 2; (C) one pasting term from simulated dataset 1; and (D) one pasting term from simulated dataset 2. The critical values corresponding to 0.05 experiment-wise error

rate for both the theoretical and permutation-based distributions are indicated with vertical lines in each plot.