Published in final edited form as:

Genet Epidemiol. 2016 February ; 40(2): 113–122. doi:10.1002/gepi.21950.

Adaptive Set-Based Methods for Association Testing

Yu-Chen Su, W. James Gauderman, Berhane Kiros, and Juan Pablo Lewinger

Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA

Abstract

With a typical sample size of a few thousand subjects, a single genomewide association study (GWAS) using traditional one-SNP-at-a-time methods can only detect genetic variants conferring a sizable effect on disease risk. Set-based methods, which analyze sets of SNPs jointly, can detect variants with smaller effects acting within a gene, a pathway, or other biologically relevant sets. While self-contained set-based methods (those that test sets of variants without regard to variants not in the set) are generally more powerful than competitive set-based approaches (those that rely on comparison of variants in the set of interest with variants not in the set), there is no consensus as to which self-contained methods are best. In particular, several self-contained set tests have been proposed to directly or indirectly ‘adapt’ to the *a priori* unknown proportion and distribution of effects of the truly associated SNPs in the set, which is a major determinant of their power. A popular adaptive set-based test is the adaptive rank truncated product (ARTP), which seeks the set of SNPs that yields the best-combined evidence of association. We compared the standard ARTP, several ARTP variations we introduced, and other adaptive methods in a comprehensive simulation study to evaluate their performance. We used permutations to assess significance for all the methods and thus provide a level playing field for comparison. We found the standard ARTP test to have the highest power across our simulations followed closely by the global model of random effects (GMRE) and a LASSO based test.

Keywords

Set-tests; Global tests; Gene Set Enrichment; Pathways; Adaptive Methods

INTRODUCTION

The standard approach to the analysis of genomewide association studies (GWAS) is to individually test the marginal effect of each SNP in a large genotyping panel containing from hundreds of thousands to a few million SNPs. With the typical sample sizes of individual studies in the few thousands of subjects, a SNP variant must confer a sizable effect on risk in order to be detected at the stringent genome-wide multiple testing threshold of $p\text{-value} < 5 \times 10^{-8}$. Because SNPs with smaller effects may not be detectable individually, methods that group SNPs into biologically relevant sets such as genes or pathways in the hope of detecting the combined effect of weaker associations, have received increasing

research attention. Set-based methods can be classified as competitive or self-contained, according to how the null hypothesis of no association is formulated. Competitive methods rely on previously computed individual SNP statistics or p-values and test whether their distribution in the pathway of interest differs from their distribution in other pathways or the full complement of SNPs not in the pathway [Fridley and Biernacka 2011; Fridley, et al. 2010]. Self-contained methods test directly whether the SNPs in the pathway are jointly associated with the disease trait with no reference to an external set. Self-contained set-based methods are generally more powerful than competitive set-based approaches [Fridley, et al. 2010], although some authors have argued that competitive tests are more appropriate in some situations [Evangelou, et al. 2012]. A perceived advantage of competitive methods is that they do not require additional computation at the individual-level data as they are based solely on summary statistics or p-values from the original GWAS. However, it has been shown that properly accounting for the correlations among statistics to control the type-I error of competitive tests does require use of the individual-level data [Gatti, et al. 2011]. In this paper we focus exclusively on self-contained methods.

Self-contained set-based methods can in turn be classified into marginal and joint approaches depending on whether they model the effect of individual SNPs or multiple SNPs jointly. Marginal approaches combine individual SNP-based p-values into a summary statistic for the entire set using some p-value combination method. Among the many existing p-value combination methods, several have been specifically used for combining SNP p-values, such as choosing the smallest p-value in the set (MinP), Fisher's p-value combination method, and the gamma method [Biernacka, et al. 2013]. By contrast, joint approaches rely on multivariate regression to jointly model and test the effect of all the SNPs in a set. Joint approaches have been proposed based on classical multivariate methods (e.g. Hotelling's T^2) [Hotelling 1931; Tsai and Chen 2009], random and mixed effects models (e.g. global model of random effects (GMRE)) [Goeman, et al. 2004], and modern regularized regression methods (e.g. LASSO) [Chen, et al. 2010]. Since the asymptotic distributions of the test statistics based on some of these regression methods are either complex or unknown, many joint approaches require permutation methods to assess significance. Permutation methods are also required for marginal approaches, since the null distributions of the p-value combination statistics are typically unknown when the p-values are correlated.

A challenge to set-based testing is that the proportion of associated SNPs within a set is unknown and can vary substantially from set to set and with the disease or trait under study. In 'high sparsity' scenarios only a small proportion of the features may be associated and in 'low to medium sparsity' scenarios a larger proportion of the features may be implicated. Methods that are powerful in high sparsity scenarios may generally perform sub-optimally in low sparsity scenarios and vice-versa. It is therefore desirable to have tests that can perform well at any level of sparsity. A simple marginal approach specifically proposed to address this issue is the truncated product (TP), which uses Fisher's p-value combination method to summarize the variants with p-values below a pre-specified threshold [Zaykin, et al. 2002]. By changing the threshold one can adjust to the level of sparsity and control the performance of the resulting test. Dudbridge and Koeleman proposed the rank truncated product (RTP), an extension of the TP that selects the top p-values ranking below a given rank threshold

rather than those below a p-value threshold[Dudbridge and Koeleman 2003]. Other rules of thumb for choosing the proportion of top variants to combine have been proposed [Fridley and Biernacka 2011]. However, if one chooses the ‘wrong’ number of SNPs to combine, i.e. a small number of SNPs when a large proportion of the SNPs in the set are associated or to combine a large number of SNPs when a small proportion are associated, the resulting test can have low power relative to choosing the optimal number.

Since the number of associated SNPs is never known in advance, it is desirable to have set-based tests that are power robust, i.e. with good power regardless of the true unknown scenario. This can be achieved by choosing the number of top variants to combine in an adaptive, data-driven fashion. One such ‘adaptive’ marginal test is the RTP with a variable truncation point discussed by Dudbridge and Koeleman[Dudbridge and Koeleman 2004] and then extended and named adaptive rank truncated product (ARTP) by Yu et al [Yu, et al. 2009]. ARTP modifies the RTP to combine the variants that give the best evidence of association rather than pre-specifying the proportion of top variants to combine. As part of our simulations below we show that ARTP can indeed achieve the desired robustness to the unknown number of associated SNPs.

In addition to the ARTP, several joint approaches have been proposed to explicitly or implicitly adapt to the unknown sparsity of the association signal in the SNP set. One of these proposals is to use the least absolute shrinkage and selection operator (LASSO) regression [Tibshirani 1996] to choose which SNPs are to be included in the model, using the property that the LASSO shrinks many of the SNP effect estimates to zero, effectively excluding them. The degree of shrinkage is controlled by a penalty tuning parameter that is typically chosen based on the data using cross validation. Chen et al.[Chen, et al. 2010] have shown that the standard LASSO adjusted for the linkage disequilibrium (LD) correlation structure among SNPs is more powerful than non-adaptive marginal methods such as using the smallest p-value to summarize the set. Tsai et al.[Tsai and Chen 2009] also compared joint approaches including traditional Hotelling’s T^2 with ANCOVA, principal component (PC) analysis, and other competitive methods and showed that Hotelling’s T^2 has the most power in certain scenarios while PC analysis has the lowest power overall. Another joint method that can be considered implicitly adaptive is the global model of random effects (GMRE) by Goeman et al. [Goeman, et al. 2004]. It models the SNP effects as random rather than fixed effects, sampled from a common distribution with mean zero and variance τ^2 . The null hypothesis of no association becomes $\tau^2=0$, which is tested using a score test. The GMRE method has been shown to have more power than non-adaptive marginal and competitive methods[Goeman and Buhlmann 2007; Goeman, et al. 2004]. A few additional publications have provided power comparisons of various subsets of the above methods[Fridley, et al. 2010; Tsai and Chen 2009; Wang, et al. 2010]. While adaptive methods have been shown to be generally more powerful than non-adaptive methods[Fridley and Biernacka 2011], no study has compared the performance of adaptive methods. In this paper, we perform an extensive comparison of adaptive set-based tests. We compare the standard ARTP, several ARTP variations we introduced, including adaptive versions of partial least squares (PLS) and principal components, the LASSO and GMRE. We also considered a method based on stepwise model selection, which can be also regarded as

adaptive. In our simulation, we have not endeavored to include every existing set-based method but rather to choose a representative method from each class of similar approaches.

METHODS

We begin by briefly describing each of the existing methods that we compare in our simulation study and then introduce several natural variants or extensions of ARTP that we also compare. We assume a set of $P > 1$ SNPs belonging to a gene, a genetic pathway, or a particular set of interest based on prior biological information. The SNP set is to be tested jointly for association with a trait of interest. We will describe the methods for a binary case-control trait because a case-control study is the most common GWAS design, but for most methods described below there is no conceptual difference if the outcome were not binary. Multivariate traits do present additional complexities that deserve separate exploration. We let $\mathbf{Y} = (Y_1, \dots, Y_N)$ be a $N \times 1$ vector of 1's and 0s representing case-control status, where n is the total number of subjects. The matrix of genotypes is denoted by $\mathbf{S} = [\mathbf{S}_1, \dots, \mathbf{S}_P]$ where each column $\mathbf{S}_j$ is the $N \times 1$ vector of coded genotypes at SNP j for all N subjects. For simplicity we assume a coding of genotypes requiring a single variable per SNP (e.g. additive, recessive, dominant) though all the methods described below can deal with a co-dominant coding requiring two variables per SNP or markers with more than two alleles. The matrix $\mathbf{X} = [\mathbf{1} \ \mathbf{S}_1 \dots \mathbf{S}_P]$ is the matrix of genotypes, augmented with a column of ones representing an intercept term to be used for regression modeling.

A commonly used model for analyzing SNP data with a binary outcome is a logistic regression, which, for testing all SNPs jointly in an omnibus test, its has the form:

$$\text{logit}[P(Y=1)] = \alpha + S\beta \quad (1)$$

Here, $\beta = [\beta_1 \dots \beta_P]'$ is a P -dimensional vector of parameters where β_j corresponds to the log-odds ratio of SNP j , $j = 1, \dots, P$ (additional adjustment covariates including variables that capture population structure are typically included in the logistic model used for analyzing GWAS but here we have omitted them for simplicity). If the outcome were quantitative, repeated measures, or survival data, then one could use standard linear regression, mixed linear models, and Cox regression, respectively. The global null hypothesis of no association based on model (1) is obtained by setting $\beta = \mathbf{0}$ and can be tested using a Wald, likelihood ratio, or score statistic. For example, if $L(\alpha, \beta)$ is the likelihood for the logistic model in (1) the likelihood ratio test (LRT) is obtained by comparing the full model with unconstrained β against the null model that sets $\beta = \mathbf{0}$:

$$LRT = -2 \ln \left(\frac{\text{likelihood of null model}}{\text{likelihood of alternative model}} \right) = -2 \ln \left\{ \frac{L(\hat{\alpha}, \beta=0)}{L(\hat{\alpha}, \hat{\beta})} \right\} \quad (2)$$

In our simulations we include several methods that rely on logistic regression of the outcome on subsets of the SNP genotypes or transformed variables constructed from the SNP genotypes (e.g. principal components). In the standard *non-adaptive* setting where each SNP

or transformed variable that enters the regression model is *not* selected based on the trait $\mathbf{Y}$, the LRT statistic (and the equivalent Wald and score statistics) has a chi-square distribution with P degrees of freedom and therefore significance can be easily assessed. In the adaptive setting, where the SNP or transformed variables are selected using information on the trait $\mathbf{Y}$, the LRT will no longer have an asymptotic chi-square distribution. In our simulations below we do not rely on asymptotic distributions to assess significance, but rather we use a permutation approach for all the methods (described below). This ensures a ‘level playing field’ for comparing methods, because the permutation approach guarantees control of the type I error at the desired target level, and therefore power differences between methods cannot be due to differences in attained type I error.

Joint Modeling Approaches

Global model of random effects (GMRE)—Logistic regression fails when the binary outcome can be perfectly predicted by the covariates [Albert and Anderson 1984]. More generally, classical regression methods can become unstable in the presence of multicollinearity and completely break down when the number of covariates exceeds the number of observations [Farrar and Glauber 1967]. This is common when analyzing pathways, which typically comprise hundreds or thousands of correlated SNPs. When $P > N$, the null hypothesis becomes unidentifiable since $\mathbf{X}\beta=0$ has multiple solutions. An alternative in this situation is to treat β as random. Specifically, one can assume that β is sampled from some distribution with expectation $\mathbf{0}$ and covariance matrix $\tau^2\mathbf{\Sigma}$. The null hypothesis of no association between the SNPs in the set and the outcome becomes $H_0: \tau^2=0$. GMRE is one such method that assumes β are random effects sampled from a common distribution. GMRE uses a score statistic for τ^2 to test the global null hypothesis. In our simulations we compared the GMRE as implemented in R-package ‘globaltest’ [Friedman, et al. 2010; Goeman, et al. 2004]. However, we did not rely on the asymptotic distribution of the test statistic to compute p-values for GMRE but rather used the permutation approach described below.

LASSO—LASSO regression [Tibshirani 1996] overcomes the multicollinearity problem by introducing a regularization L_1 penalty on the β parameters in model (1). The effect of the L_1 penalty is to shrink some of the β_j coefficients to exactly zero, effectively performing parameter estimation and variable selection simultaneously. The tuning parameter controlling the degree of shrinkage is usually selected based on the data using cross-validation. The LASSO can also be interpreted as a random effect model with β sampled from a double exponential (Laplace) distribution. In our simulations we used the implementation of the LASSO in the ‘*glmnet*’ R-package [Friedman, et al. 2010]. The L_1 tuning parameter is chosen by a k-fold cross-validation based on the deviance of the logistic regression model. The test statistic we used for the LASSO is the likelihood ratio defined above in (2), except that the model parameters are estimated by the LASSO rather than maximum likelihood. Again, we assessed the significance of the LR statistic based on the LASSO using the permutation approach described below.

Hotelling’s T^2 —For a case-control study the classical Hotelling’s T test [Hotelling 1931] can be used to test for association between case-control status and the genotypes.

Specifically, the Hotelling's T^2 test is a multivariate generalization of the standard t-test and can be used to test for differences in means between the genotypes of cases and of controls. It should be noted that Hotelling's T^2 test does require $N < P$. We included Hotelling's T^2 test in our simulations to illustrate the performance of non-adaptive methods.

Marginal P-values Approaches

In the set-based approaches described above, the trait is jointly modeled as a function of all SNPs using a multiple regression model. By contrast, marginal set methods separately test each SNP and later combine the SNP-specific p-values into a single summary statistic for the set. P-value combination for a pathway, which is the typical SNP set of interest, can be performed in a non-hierarchical fashion by a simple combination of all the p-values corresponding to SNPs mapped to the pathway, or in a hierarchical fashion by first combining SNP-based p-values within each gene into gene-based statistics, and then the gene-based p-values into pathway level statistics. Yu et al. showed that when only SNPs in a small proportion of genes within a pathway are involved, the hierarchical approach is preferred [Yu, et al. 2009]. We describe the marginal p-values methods included in our simulations using a non-hierarchical p-value combination but the methods apply equally with a hierarchical combination. We assume that for each SNP in the set of interest with a p-value p_j , $j=1, \dots, P$ has been computed using any standard test such as a Wald or likelihood ratio test based on a simple logistic regression for the SNP. There are many ways of combining p-values [De la Cruz, et al. 2010; Fridley and Biernacka 2011; Peng, et al. 2010] into a summary statistic, but in the pathway-based analysis context Fisher's method [Fisher 1925] has been proposed as part of the RTP and ARTP methods described below. Fisher's combination statistic is given by $-2 \sum_{j=1}^P \log(p_j)$.

When the p-values being summarized are independent, Fisher's combination statistic has a chi-square distribution with $2P$ degrees of freedom under the global null hypothesis that no SNP in the set is associated with the outcome. However, since SNPs within a gene or a pathway are typically in linkage disequilibrium (LD) and therefore the corresponding p-values are not independent, Fisher's method will not have an asymptotic chi-square distribution under the null. It is then common to rely on permutations to assess the significance of marginal set-based tests. In our simulation we use permutations for all the joint and marginal methods we compare.

Truncation Product and Rank Truncation Product (RTP) methods—The rank truncation product method [Dudbridge and Koeleman 2003] uses Fisher's method to combine the top most significant p-values rather than all p-values. The idea is to maximize the evidence of association by discarding 'noise' unassociated SNPs. Intuitively, one should aim to combine only the SNPs that are associated with the outcome, and these are found predominantly among the top SNPs. To apply RTP the p-values are ordered from most to least significant and the top k most significant p-values are combined into the summary statistic for the set, where k is a predefined 'truncation point' between 1 and P . One way of choosing the truncation point is by selecting those SNPs with p-values below a preset threshold (eg. p-value $< 5 \times 10^{-6}$). De la Cruz et al. [De la Cruz, et al. 2010] proposed

selecting truncation point at 0.1 quantile of the distribution of the smallest p-values obtained from permutations. Other methods of combining SNPs and/or choosing a truncation point have also been proposed[Peng, et al. 2010]. Because the asymptotic distribution of the RTP statistic does not have a simple form when the p-values that are combined are not independent, significance for RTP is typically assessed using permutations.

Minimum P-value (MinP)—Use of the smallest p-value, MinP, has been proposed as a simple way to summarize a set of p-values[Dudbridge and Koeleman 2003]. This is a particular case of RTP, obtained with a truncation point of 1. Intuitively, MinP should perform well when a single SNP or a small number of SNPs in the set are associated with the trait.

Adaptive Rank Truncation Product (ARTP)—Choosing a single truncation point in RTP may lead to suboptimal power when there is a mismatch between the number of SNPs combined and the number that are truly associated. A more robust approach, akin to model selection, is to choose the truncation point that yields the strongest evidence of association (smallest p-value) among all possible truncation points. This adaptive extension of RTP introduced by Yu et al.[Yu, et al. 2009] is called the adaptive rank truncated product. Assessing the significance of the ARTP test requires in principle a second layer of permutations over RTP. However, Yu et al.[Yu, et al. 2009] showed that permutations can be ‘recycled’ and thus nested permutations can be avoided. A further computational speedup can be achieved by examining a subset of all possible truncation points. Yu et al.[Yu, et al. 2009] showed that one can use as few as 5 truncation points without sacrificing performance. In our simulation, we used 20 truncation points to compute the adaptive rank truncation product.

Variations and Novel Methods

In addition to the existing methods described above, in our simulations we compared several natural extensions of ARTP that combine the idea of adaptation in ARTP with dimensionality reduction. We also considered an approach that uses stepwise model selection as an alternative form of adaptation and which has not been studied before in the context of set-based testing. PCA analysis has been used extensively in genetic applications[Price, et al. 2006]. In the association context, the idea is to approximate the original genotypic data using a lower dimensional representation based on the first top principal components (PCs) (e.g. the top components explaining at least 80% of the total variance in the genotype data). Association with the phenotype is then assessed by replacing the original genotypes in the regression model in (1) by the PC’s. The hope is that by capturing most of the genetic variability with fewer variables the resulting reduction in degrees of freedom would yield a more powerful test. Principal-components regression has been proposed by several authors for testing pathway-based associations[Cai, et al. 2013; Gauderman, et al. 2007]but it has been shown to have low power in certain scenarios[Fridley, et al. 2010]. We compare three methods based on PCA that we briefly describe below.

MinP on Principal Components (MinP PC)—Analogously to MinP, we perform a marginal effect test on each PC and choose the minimum p-value as the summary statistic for the entire set of SNPs.

Adaptive Rank Truncation Product on Principal Components (ARTP PC)—Instead of selecting a predefined number of top PCs, one can select the PCs adaptively as in ARTP by including the top PCs that maximize the evidence of association. Specifically, from a grid of values representing 75, 80, 85, 90 and 95% of the proportion of the total variance explained by the PCs; we choose the cutoff that yields the smallest p-value in a logistic regression model that includes the corresponding PCs.

Stepwise Regression on Principal Components (Step PC)—Rather selecting the PCs that enter the logistic regression model using ARTP we use stepwise model selection on the PCs. Specifically, we perform forward stepwise logistic regression which starts with no PC in the model and systematically tests for improvement to the fit based on the Akaike's Information Criterion, AIC when adding one PC at a time. To limit the computational burden we only consider the top set of PC's explaining at least 95% of the total variance in the SNPs for inclusion in the model. Using stepwise regression to select which PCs to include in the model is a data adaptive alternative to ARTP. As test statistic for the SNP set we used the LRT based on the logistic regression model with the selected PC's. Because the PCs entering the model are selected using the phenotypic data, the distribution of the final likelihood ratio test statistic does not follow a standard asymptotic chi-square distribution. Instead, we assessed significance by permutations to account for the data driven model selection.

Adaptive Rank Truncation Product on Partial Least Squares (ARTP PLS)—Partial least squares (PLS) is a dimension reduction technique similar to PCA. Like PCA, PLS partitions the data into orthogonal linear combinations of the original variables, but unlike PCA, which seeks linear combinations that maximize the variance, PLS seeks linear combinations that maximize the covariance with the outcome [Sun, et al. 2009]. We used an extension of PLS for logistic regression implemented in the R-package 'pls' [Bjørn-Helge, et al. 2011]. As test statistic for testing a set of SNPs we used the likelihood ratio based on logistic regression of the case-control status on the selected PLS components. In our simulations, we investigated an adaptive version of PLS regression analogous to the ARTP PC above but using the PLS components of the genotypes instead of the PCs.

Cluster Regression (CR)—The idea of this method is to first cluster SNPs according to LD, compute a p-value for each cluster and then adaptively select the top clusters to summarize the set. Since nearby SNPs are generally in LD, if one SNP is associated with the disease of interest, then those SNPs in LD with it will also be associated. We use single linkage hierarchical clustering to group SNPs that are in high LD with each other. Specifically, we cluster based on a distance between SNPs, defined as $1 - r_{ij}^2$ where r_{ij} is the correlation between alleles (coded as 0-1) in SNPs i and j in a haplotype. We choose an r^2 cutoff of 0.7 so that all the SNPs within each cluster will have a square correlation of at least 0.7. We summarize each cluster using multiple logistic regression of all the SNPs in the

cluster. We then apply the ARTP algorithm to the sorted cluster p-values to obtain a pathway-based statistic. We also consider a variant of this procedure, which uses a single degree of freedom test for each cluster by constraining all the SNPs within each cluster to have the same regression coefficient. This is equivalent to regressing the case-control outcome on the ‘average’ genotype of all the SNPs within each cluster. To apply this method all pairwise correlations of SNPs in the cluster must be positive. Because the clusters are chosen so that all SNPs within a cluster are in high LD with each other, this can be always achieved by ‘flipping’ the allelic coding of some SNPs if necessary. We refer to the clustering approach with a single degree of freedom test as *CR-1df*.

Assessing significance

To assess the statistical significance of the method presented above, the null distribution of the corresponding tests statistics is required. However, except for GMRE, the distribution of the test statistic for each of the methods above is unknown and/or difficult to derive, even asymptotically. This is due both to the presence of LD among SNPs that induces correlations between SNP p-values and also to the data driven adaption of many of the methods proposed. For example, the sum of the most significant log p-values in the ARTP related methods does not have the standard chi-square distribution that would result from combining independent and unsorted p-values. For this reason, the majority of the original tests above use some form of permutation testing to assess significance. In our simulation we used permutation to assess significance for all the methods, which guarantees the control of the type I error at the nominal level and provides a level playing field for comparison of methods, ensuring that any difference in performance is due to true differences in power and not differences in test size.

The permutation scheme to assess significance is as follows: the phenotypes Y_i , $i=1, \dots, N$ are shuffled while keeping the matrix of genotypes $\mathbf{X}$ fixed, and the test statistic is re-computed in turn for each shuffled data. This permutation scheme ‘breaks’ any association present between genotypes and outcome while preserving the LD structure. Statistical significance of the original data test statistic is obtained by comparing with the distribution of the permutation based statistics.

Simulation Study

We conducted a simulation study to evaluate the performance of the set-based tests described above in a series of scenarios based on sets with independent SNPs or sets with a realistic linkage disequilibrium structure among the SNPs. For the latter, we simulated genotypes based on real genotypic data corresponding to the ~254kb long prostaglandin E receptor 3 (PTGER3) gene from 1904 non-Hispanic white subjects from the Children Health Study (CHS)[Gauderman, et al. 2004] GWAS (Respiratory outcomes are the main focus of the CHS and the prostaglandin E receptor 3 (PTGER3) gene is part of the prostanoid receptor family known to be associated with aspirin-intolerant asthma[Kim, et al. 2007]). Within PTGER3, the CHS data had 162 typed SNPs, and we also considered 1093 additional untyped SNPs with a minor allele frequency (maf) > 5% for a total of 1255 SNPs. The LD structure of this region is fairly typical with 15 LD blocks. We imputed the untyped 1093 SNPs using the CEU HapMap[International HapMap 2005] as a reference and haplotyped

all 1904 individuals using the software BEAGLE[Browning and Browning 2007]. This gave us a ‘pool’ of 3808 haplotypes with a realistic LD structure among the SNPs. We assigned the original 162 typed SNPs to represent typed SNPs in the simulated data and the originally untyped 1093 SNPs to represent untyped SNPs in the simulated data. We designated sets of causal SNPs chosen from either the typed SNPs or among the untyped SNPs in the region. For sets of independent SNPs and for sets of SNPs in LD, we considered scenarios with a small number of causal SNPs ranging from 0 to 10 causal SNPs, and scenarios with a relatively high number of causal SNPs ranging from 20 to 140 causal SNPs (in increments of 10). Across simulation replicates the set of causal SNPs remained fixed. To simulate an individual’s genotype we sampled a random pair of haplotypes from the pool of 3,808 haplotypes.

We generated each individual’s case-control phenotype based on a logistic regression model for the disease probability of the form:

$$Pr(Y=1|S)=\frac{\exp(S\beta)}{1+\exp(S\beta)} \quad (3)$$

where S is the matrix of the genotypes (coded 0,1,2 according to the number of minor alleles) of the designated causal SNPs, augmented with a vector of ones for the intercept. We generated 1,000 cases and 1,000 controls for each scenario.

The number of causal SNPs, K , was the primary parameter of interest in the simulations. We set the effect size (log odds-ratio) of each of the causal SNPs so that the power to detect at least one causal SNP was not too high or too low. Having power in the midrange allows for performance differences between methods to become apparent, as power differences are “squeezed” together when overall power is either too close to one or too close to zero. To ensure simulated scenarios yielding power in the desired midrange, we set the log-odds ratios for the causal SNPs, $\beta_1 \dots \beta_K$ so that, based on their corresponding allele frequencies, each causal SNP would have the *same* power to be detected if individually tested with a simple univariate association test. For the scenarios with low numbers of causal SNPs we assigned 20% marginal power to detect each causal SNP at the 5% level. For the scenarios with large numbers of causal SNP we assigned 10% marginal power to each causal SNP. The intercept of the logistic regression model for the case-control status in (4) was set to yield a population disease prevalence of approximately 1%. In addition to the simulation scenarios described above, in which each SNP has the same power to be detected if tested marginally, we also investigated the performance of set-based tests in a set of simulations where the marginal power to detect each individual SNP is not constant across the causal SNPs, i.e. there are easier and harder to detect causal SNPs. In the independent SNP scenarios, similarly to the scenario with LD, we varied the number of causal SNPs from 0 to 10 in increments of 1 and from 20 to 140. The complete range of simulation sets scenarios is summarized in Table 1.

For each scenario we computed an upper bound to the achievable power to get a sense of how close the power of each method was to this optimal level. This upper bound was

obtained by testing the SNP set using a likelihood ratio test based on a logistic regression model containing only the known causal SNPs.

For each scenario examined we computed power based on 1,000 replicates and within each replicate p-values were computed using 1,000 permutations. Each scenario took approximately 15 hours of computation using parallelization with a high-performance computing cluster.

RESULTS

As expected based on the theoretical guarantees offered by the permutation-based assessment of significance used in our study, all methods preserved the type I error at the target level for the scenarios with no simulated causal SNPs (Fig 1 and Fig 3). As mentioned earlier, the key implication of the accurate type I error control is that any observed power differences between methods in the simulation study can be directly attributed to intrinsic differences in performance.

Figure 2 highlights the importance of adaptation to the unknown proportion of associated SNPs. In the two scenarios depicted (8 and 50 simulated causal SNPs out of a total of 162 SNPs respectively), the power attained by the non-adaptive RTP varies widely with the number of top most significant p-values combined to construct the statistic for the set (i.e. the truncation point). (RTP attains maximum power when combining a number of p-values far in excess to the actual number of simulated causal SNPs because of the many additional associations induced by LD). Because the optimal truncation point is not known in advance, using RTP with a fixed truncation point results in a sizable loss of power relative to the maximum attainable. This inefficiency is particularly pronounced in scenarios with a higher proportion of associated SNPs, whereby combining only a few top p-values results in a drastic power loss. By contrast, by *adaptively* choosing the number of top SNPs to combine, ARTP achieves power that is very close to the maximum power achieved by RTP but without any prior knowledge of the optimal number. These results underscore the need for adapting set-based tests so that they become robust to the unknown number of associated SNPs.

In agreement with the observation above, the adaptive methods ARTP, ARTP-PC, LASSO and GMRE evaluated in the simulation showed the expected robustness of power, performing well (relative to the maximum achievable power) for the entire range of simulated causal SNPs, LD structure (LD vs. independent), distribution of effect sizes (uniform vs. non-uniform power; not show), and causal SNPs (typed vs. un-typed; not shown). The two non-adaptive methods, minP and Hotelling's T^2 test also performed in accordance with the observation that the non-adaptive RTP lacks power robustness: MinP had good power (again relative to the maximum achievable power) in scenarios with few causal SNPs and poor power in scenarios with many causal SNPs, while the reverse was true for Hotelling's T^2 , which had good power in scenarios with many causal SNPs and poor power in scenarios with few causal SNPs.

The main finding of our simulation study is that in all scenarios ARTP outperformed the other set-based methods, including the adaptive ones (Fig 2 and Fig 3). This was a somewhat surprising result as we expected that the set tests based on the LASSO and GMRE, which are not only adaptive but model all SNPs in the set jointly, would generally have better power than ARTP, which while adaptive, models each SNP in the set marginally. However, although GMRE and LASSO performed well (generally tracking each other closely, GMRE with slightly better power), they consistently had lower power than ARTP.

As a group, the methods that incorporate some form of dimensionality reduction (ARTP-PC, minP-PC, CR-1df, ARTP-PLS) -- with or without additional adaptation -- did not offer a power advantage over the methods that did not incorporate dimensionality reduction. For example, although ARTP-PC performed well, it had consistently lower power than the original ARTP, and MinP-PC only performed better than its non-dimension reduction counterpart MinP in scenarios with a large proportion of associated SNPs. CR-1df, which achieves dimensionality reduction by initial clustering of SNPs based on LD, performed well but has lower power than ARTP or the other adaptive methods. ARTP-PLS was a particularly poor performer among the methods incorporating dimensionality reduction.

DISCUSSION

Self-contained approaches for testing sets of SNPs in genes and pathways are generally more powerful than competitive methods such as gene-set enrichment and its related variants [Fridley et al., 2010; Wang et al., 2010]. However, non-adaptive self-contained methods can be non-robust to the usually unknown proportion of associated SNPs in the set being tested. For example, the common practice of using the smallest p-value (minP) to summarize an entire set of SNPs would yield suboptimal power when there are more than a few SNPs that are associated. Even if a single SNP is causal, multiple SNPs will be associated in the presence of LD. Conversely, summarizing a set of SNPs by combining all the p-values in the set (e.g. sum log p-values) will yield suboptimal power when the proportion of associated SNPs is small. While adaptive set-based tests can be robust to the proportion of associated SNPs, it has been unclear which adaptive method is best. We compared several adaptive and non-adaptive set-based methods representative of the main types proposed in the literature. We simulated scenarios where the SNPs in the set of interest are in LD, representing a contiguous genomic region, and also scenarios with independent SNPs. All set-based tests were evaluated using a permutation approach that preserves the type I error at the target level, allowing us to ascribe differences in power to real differences in performance. We found the ARTP test to consistently outperform all other methods in scenarios ranging from a single associated SNP to scenarios where a large proportion of the tested SNPs were associated with the trait. The power advantage of ARTP held both in scenarios with and without LD. In practice, most scenarios of interest correspond to a mixture of the LD and independent SNP scenarios. For example, SNPs mapped to a common pathway with multiple genes will contain clusters of SNPs in LD with each other (e.g. within a gene) with little or no LD between clusters. Our simulation results directly apply to these more complex scenarios, since power for testing a SNP set consisting of independent clusters of SNPs in LD is a direct function of the power for testing each individual cluster.

Except for Hotelling's test, all the approaches we considered in our evaluation are capable of incorporating covariate adjustments by potential confounders. In particular, ARTP is automatically covariate-adjusted when the p-values it combines are themselves adjusted (e.g. when derived from adjusted regressions). In the absence of covariates, the null hypothesis of no association can be formally expressed as statistical independence between phenotype and genotypes. For testing this null hypothesis, a permutation strategy where the rows of Y are permuted while keeping G fixed or, equivalently, the rows of G are permuted while keeping Y fixed, yields a test that is guaranteed to control the type I error at the target level regardless of the distribution of the data. Such a permutation test is nonparametric in that no distributional assumptions are required except for exchangeability of subjects under the null hypothesis. However, when testing for genetic associations, researchers almost invariably want to control for potential confounders or effect modifiers Z (e.g. covariates capturing population strata). The null hypothesis of no association between Y and G 'adjusted for' Z , becomes the conditional independence between Y and G given Z , rather than unconditional independence. However, only in very limited circumstances exact permutation strategies can be adapted for testing conditional independence. For example, if an adjusting covariate Z is binary with values 0 and 1, testing for conditional independence can be accomplished by performing permutations within subjects with $Z=0$ and subjects with $Z=1$. Often however, the adjustment covariate is continuous or there are multiple discrete adjustment covariates. In these situations, testing for conditional independence using restricted within-group permutations is not possible because each unique combination of the adjustment covariates Z will be typically represented by a single subject[4] and therefore no permutation other than the trivial one would be available. However, when exact testing of conditional independence via restricted within- group permutations is not possible, an alternative strategy is to perform a standard unconditional permutation scheme (i.e. permute Y while leaving G and Z fixed) but using a test statistic that is 'already adjusted for Z '[5]. This is typically accomplished by using a regression model with Y as the outcome and G and Z as covariates, and using a test statistic based on the regression parameters for G , adjusted for Z . Strictly speaking, the resulting procedure tests the unconditional null hypothesis of independence between Y and the pair (G, Z) , not the conditional independence of Y and G given Z . However, being 'already adjusted for Z ', the test statistic will be nearly independent of Z , and the resulting test will effectively testing conditional independence. In particular, such a permutation test will have little or no power to detect a confounded association driven exclusively by the relationship between Y and Z . Though rarely explicitly stated, this latter approach of using unconditional permutation testing with an adjusted test statistic—the approach we used in our simulations—is the most common way permutations are used when testing genetic associations. Although it does not enjoy the theoretical error control guarantees of the standard permutation method, it has been extensively shown empirically to yield good control of the type I error and it is also much more widely applicable.

Based on the earlier results of Yu et al.[Yu, et al. 2009], we expected ARTP to perform well, but we did not anticipate it to be consistently better than joint methods such as GMRE or those based on the LASSO. We expected that modeling the effect of all SNPs jointly would offer superior power and should be preferred over marginal methods that simply combine individual SNP p-values. Instead, although LASSO and GMRE performed well, they were

consistently less powerful than ARTP, even in scenarios where the logistic regression model used by the LASSO and GMRE matched the simulated model. ARTP performs better since it forms a model selection that is geared towards hypothesis testing (by maximizing the statistical evidence) while LASSO performs model selection through tuning the penalty to maximize deviance. Alternative ways of choosing the LASSO tuning parameter to improve the performance of the LASSO as a set-test its worth investigating.

Our findings are aligned with those of Evangelou et al. who found ARTP to perform best in a comparison of several competitive approaches to set-based testing that included ARTP implemented as a competitive rather than self-contained method. In their comparison they also included the tail strength measure [Taylor and Tibshirani 2006] —which we did not include in ours —but found it to perform poorly.

In addition to the standard ARTP, we proposed two new variants that incorporate principal components or partial least squares. The goal was to gain additional power by combining dimensionality reduction with data adaptation. In general we found that these approaches did not perform better than the standard ARTP even when the components were chosen adaptively. However, this may have more to do with the particular way PC and PLS reduce dimensionality, (e.g. PLS chooses components that maximize the covariance between genotypes and outcome), which may not translate into components of maximal association. Canonical correlations, a related classical dimensionality reduction method that seeks components that maximize the correlation rather than the covariance, may offer better performance, particularly in the multivariate trait context. We will explore the use of canonical correlations for set-based testing with multiple traits in future work. As an alternative form of ‘adaptation’ we also evaluated the use of step model selection on principal components but it performed quite poorly. This shows that the adaptation step in ARTP is not equivalent to other forms of model selection.

We have shown that ARTP has a clear power advantage over other adaptive methods. Although ARTP is computationally intensive this is a disadvantage shared with other adaptive methods that require permutations (GMRE being an exception as there is an asymptotic approximation that can be used to assess its significance). Recently a sequential version of ARTP has been introduced, but it does not have a power advantage over the standard ARTP [Chen, et al. 2013]. Also, Zhang et al. derived the analytic distribution of the ARTP test in the case of independent p-values [Zhang, et al. 2013]. Unfortunately, their sensitivity study shows that under departure of independence the type-I error can be anticonservative; permutations remain for now the only viable option for assessing the significance of ARTP with guaranteed control of the type I error. Also, as originally proposed, ARTP can only be used for testing genetic main effects, as permutations are not suitable for testing interactions [Buzkova, et al. 2011]. In future work we plan to develop faster algorithms to perform ARTP without the need of permutations and extend ARTP for testing GxE interactions.

Finally, although we focused on SNPs, ARTP can be applied with many other types of genomic features such as gene expression, methylation or rare variants.

Acknowledgments

This research was supported in part by NIH grants 5R01ES019876-05, 5R01HD061968-05, 5R21HL115606-03, P30ES007048, 5P30CA014089 and 1U19CA148107. We also thank Duncan Thomas for valuable comments on the manuscript.

References

- Albert A, Anderson JA. On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*. 1984; 71(1):1–10.
- Biernacka JM, Jenkins GD, Wang L, Moyer AM, Fridley BL. Use of the gamma method for self-contained gene-set analysis of SNP data. *Eur J Hum Genet*. 2013; 20(5):565–71.
- Bjørn-Helge, M., Wehrens, R., Kristian Hovde, L. pls: Partial Least Squares and Principal Component regression. R package version 2.3-0. 2011. <http://CRAN.R-project.org/package=pls>
- Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*. 2007; 81(5):1084–97. [PubMed: 17924348]
- Buzkova P, Lumley T, Rice K. Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions. *Ann Hum Genet*. 2011; 75(1):36–45. [PubMed: 20384625]
- Cai M, Dai H, Qiu Y, Zhao Y, Zhang R, Chu M, Dai J, Hu Z, Shen H, Chen F. SNP set association analysis for genome-wide association studies. *PLoS One*. 2013; 8(5):e62495. [PubMed: 23658731]
- Chen HS, Pfeiffer RM, Zhang SP. A Powerful Method for Combining P-Values in Genomic Studies. *Genetic Epidemiology*. 2013; 37(8):814–819. [PubMed: 23959976]
- Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, Peters U, Hsu L. Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *Am J Hum Genet*. 2010; 86(6):860–71. [PubMed: 20560206]
- De la Cruz O, Wen X, Ke B, Song M, Nicolae DL. Gene, region and pathway level analyses in whole-genome studies. *Genet Epidemiol*. 2010; 34(3):222–31. [PubMed: 20013942]
- Dudbridge F, Koeleman BP. Rank truncated product of P-values, with application to genomewide association scans. *Genet Epidemiol*. 2003; 25(4):360–6. [PubMed: 14639705]
- Dudbridge F, Koeleman BP. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *Am J Hum Genet*. 2004; 75(3):424–35. [PubMed: 15266393]
- Evangeliou M, Rendon A, Ouwehand WH, Wernisch L, Dudbridge F. Comparison of methods for competitive tests of pathway analysis. *PLoS One*. 2012; 7(7):e41018. [PubMed: 22859961]
- Farrar DE, Glauber RR. Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics*. 1967; 49(1):92–107.
- Fisher, RA. Statistical methods for research workers. Edinburgh, London: Oliver and Boyd; 1925.
- Fridley BL, Biernacka JM. Gene set analysis of SNP data: benefits, challenges, and future directions. *Eur J Hum Genet*. 2011; 19(8):837–43. [PubMed: 21487444]
- Fridley BL, Jenkins GD, Biernacka JM. Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. *PLoS One*. 2010; 5(9)
- Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*. 2010; 33(1):1–12. [PubMed: 20808728]
- Gatti DM, Barry WT, Nobel AB, Rusyn I, Wright FA. Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics*. 2011; 11:574.
- Gauderman WJ, Avol E, Gilliland F, Vora H, Thomas D, Berhane K, McConnell R, Kuenzli N, Lurmann F, Rappaport E, et al. The effect of air pollution on lung development from 10 to 18 years of age. *N Engl J Med*. 2004; 351(11):1057–67. [PubMed: 15356303]
- Gauderman WJ, Murcray C, Gilliland F, Conti DV. Testing association between disease and multiple SNPs in a candidate gene. *Genet Epidemiol*. 2007; 31(5):383–95. [PubMed: 17410554]
- Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*. 2007; 23(8):980–7. [PubMed: 17303618]

- Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*. 2004; 20(1):93–9. [PubMed: 14693814]
- Hotelling H. The Generalization of Student's Ratio. *The Annals of Mathematical Statistics*. 1931; 2(3): 360–378.
- International HapMap C. A haplotype map of the human genome. *Nature*. 2005; 437(7063):1299–320. [PubMed: 16255080]
- Kim SH, Kim YK, Park HW, Jee YK, Kim SH, Bahn JW, Chang YS, Kim SH, Ye YM, Shin ES, et al. Association between polymorphisms in prostanoid receptor genes and aspirin-intolerant asthma. *Pharmacogenet Genomics*. 2007; 17(4):295–304. [PubMed: 17496729]
- Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, Zhao J, Zhou X, Reville JD, Jin L, et al. Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur J Hum Genet*. 2010; 18(1):111–7. [PubMed: 19584899]
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006; 38(8): 904–9. [PubMed: 16862161]
- Sun, LA., Ji, SW., Yu, SP., Ye, JP. On the Equivalence Between Canonical Correlation Analysis and Orthonormalized Partial Least Squares. 21st International Joint Conference on Artificial Intelligence (Ijcai-09), Proceedings; 2009. p. 1230-1235.
- Taylor J, Tibshirani R. A tail strength measure for assessing the overall univariate significance in a dataset. *Biostatistics*. 2006; 7(2):167–81. [PubMed: 16332926]
- Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B*. 1996; 58(1)
- Tsai CA, Chen JJ. Multivariate analysis of variance test for gene set analysis. *Bioinformatics*. 2009; 25(7):897–903. [PubMed: 19254923]
- Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet*. 2010; 11(12):843–54. [PubMed: 21085203]
- Yu K, Li Q, Bergen AW, Pfeiffer RM, Rosenberg PS, Caporaso N, Kraft P, Chatterjee N. Pathway analysis by adaptive combination of P-values. *Genet Epidemiol*. 2009; 33(8):700–9. [PubMed: 19333968]
- Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. Truncated product method for combining P-values. *Genet Epidemiol*. 2002; 22(2):170–85. [PubMed: 11788962]
- Zhang SP, Chen HS, Pfeiffer RM. A combined p-value test for multiple hypothesis testing. *Journal of Statistical Planning and Inference*. 2013; 143(4):764–770.

Figure 1.

Power of set-based methods as a function of the number of causal SNPs in scenarios with independent SNPs. Panel a) low proportion of causal SNPs. Panel b) High proportion of causal SNPs. CR-1df, PCA and PLS based methods were not included in this simulation scenarios since without LD among SNPs dimension reduction is not possible.

Figure 2.

Combining too few or too many p-values using the RTP method results in a sizable loss of power. By contrast, ARTP is robust to the unknown number of associated SNPs.

Figure 3. Power of set-based methods as a function of the number of causal SNPs in scenarios with LD structure from the PTGER3 gene.

Table 1

Summary of simulation scenarios.

Number of Causal SNPs (out of 162 in the set)		Marginal Power
No LD	0-10 (Low Proportion)	50%
	20-140 (High Proportion)	10%
LD	0-10 (Low Proportion)	20%
	4	Varying power across causal SNPs