Published in final edited form as:

Genet Epidemiol. 2019 February ; 43(1): 102–111. doi:10.1002/gepi.22168.

ComPaSS-GWAS: A method to reduce type I error in genome-wide association studies when replication data are not available

Jeremy A. Sabourin^{#1}, Cheryl D. Cropp^{#2,3,4}, Heejong Sung¹, Lawrence C. Brody⁵, Joan E. Bailey-Wilson², and Alexander F. Wilson¹

¹Genometrics Section, Computational and Statistical Genomics Branch, NHGRI/NIH, Baltimore, MD

²Statistical Genetics Section, Computational and Statistical Genomics Branch, NHGRI/NIH, Baltimore, MD

³Translational Genomics Research Institute (TGen), Integrated Cancer Genomics Division, Phoenix, AZ

⁴Samford University, McWhorter School of Pharmacy, Birmingham, AL

⁵Medical Genomics and Metabolic Genetics Branch, NHGRI/NIH, Bethesda, MD

These authors contributed equally to this work.

Abstract

Results from association studies are traditionally corroborated by replicating the findings in an independent dataset. Although replication studies may be comparable for the main trait or phenotype of interest, it is unlikely that secondary phenotypes will be comparable across studies, making replication problematic. Alternatively, there may simply not be a replication sample available because of the nature or frequency of the phenotype. In these situations, an approach based on complementary-pairs stability selection, ComPaSS-GWAS, is proposed as an ad-hoc alternative to replication. In this method, the sample is randomly split into two conditionally independent halves multiple times (resamples) and a GWAS is performed on each half in each resample. Similar in spirit to testing for association with independent discovery and replication samples, a marker is corroborated if its p -value is significant in both halves of the resample. Simulation experiments were performed for both non-genetic and genetic models. The type I error rate and power of ComPaSS-GWAS were determined and compared to the statistical properties of a traditional GWAS. Simulation results show that the type I error rate decreased as the number of resamples increased with only a small reduction in power and that these results were comparable to those from a traditional GWAS. Blood levels of vitamin pyridoxal 5'-phosphate (PLP) from the Trinity Student Study (TSS) were used to validate this approach. The results from the validation study were compared to, and were consistent with, those obtained from previously published independent replication data and functional studies.

Keywords

GWAS; replication; corroboration; type I error; power

INTRODUCTION

Standard statistical analyses are designed to maximize the power of a test while minimizing the type I error rate. However, because of the highly non-independent nature of the variants in a genome, the assumptions underlying the methods used are generally not met and the type I error rate is inflated [Sung et al. 2011, Wilson and Ziegler 2011]. Meta-analyses of independent data sets are often used to ameliorate the inflated type I error rates and low power to detect variants with low locus-specific heritabilities in GWAS [Cantor et al. 2010, Begum et al. 2012]. Although meta-analyses can reduce type I error rates and increase power beyond that of a single study, the approach is limited to those studies where independent replication samples are available. However, for many studies, replication data may only be available for the primary trait or phenotype of interest and secondary phenotypes may not be available or be quite different in the replication sample. In other studies, there may simply not be a replication sample available because of the nature or frequency of the phenotype of interest.

Methods proposed to address the absence of independent replication data include sample splitting and resampling. Skol et al. [2006] proposed splitting samples into a discovery and replication set for a two-stage GWAS where the discovery set was completely genotyped and the replication set was genotyped only on SNPs that were suggestive in the discovery set, but this was less powerful than analyzing the full data set. Over time, the cost of genotyping decreased and the preferred approach was to analyze the entire sample rather than split the sample due to the loss of power of the two-stage discovery/replication approach [McCarthy et al. 2008].

Resampling methods use the available data and a repeated sampling technique to obtain estimates that may be less sensitive to sampling variability than those obtained from a single analysis of the entire study. Approaches that use resampling theory include bootstrap aggregation (bagging) [Breiman 1996] and subsample aggregation (sub-bagging) [Buhlmann and Yu 2002]. These methods have been adapted to variable selection procedures in stability selection (SS) [Meinshausen and Buhlmann 2010] and resample model averaging (RMA) [Valdar et al. 2009]. Stability selection has been explored for GWAS [Alexander and Lange 2011] and RMA has been proposed for fine mapping [Valdar et al. 2012]. The frameworks of SS and RMA were extended in complementary pairs stability selection [Shah and Samworth 2013] where each resample splits the data into complementary halves and a measure of the corroboration of findings across the split is determined.

In this study, we propose an approach that combines the corroboration aspect of complementary pairs stability selection with simple linear regression based GWAS analysis, denoted ComPaSS-GWAS. ComPaSS-GWAS selects SNPs that are corroborated across each random split of the data over a large number of resamples. This method is intended to be used as a follow-up analysis to GWAS, in situations when an independent replication sample

is not available. Simulation experiments are used in this study to determine the statistical properties of the method and these results compared to those from a traditional GWAS. In addition, the method is applied to blood levels of vitamin pyridoxal 5'-phosphate (PLP) from the Trinity Student Study (TSS) [Carter et al. 2015]. A number of the SNPs in the PLP GWAS analysis have been independently replicated in the literature, and these replicated SNPs are used to validate the ComPaSS-GWAS method.

METHODS

A traditional GWAS approach is described followed by a description of ComPaSS-GWAS to illustrate differences between the approaches. The notation assumes a quantitative trait in a linear model with the residual error following a Gaussian distribution, but the method can be modified to allow for more complex modeling methods, e.g., a generalized linear model or a linear mixed model.

For a traditional GWAS, let n represent the number of samples (individuals) and m represent the number of markers or variants, e.g., single-nucleotide polymorphisms (SNPs) or more generally sequence variants. Let $\mathbf{y} = (y_1, \dots, y_n)$ denote a vector of quantitative trait values that have been appropriately transformed for normality, $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_m]$ denote a $n \times m$ matrix of SNP genotypes, and $\mathbf{D} = \{\mathbf{y}, \mathbf{X}\}$. For each of the m SNPs, a regression based GWAS considers the simple linear model,

$$y_i = \mu + x_{ij}\beta_j + \epsilon_i \quad (1)$$

where μ is the model intercept, x_{ij} is the genotype of the i 'th individual for SNP j , β_j is the effect of the SNP j , and $\epsilon_j \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian error term. Each SNP is tested (assuming independence) to determine if β_j is significantly different from 0.

ComPaSS-GWAS

In a discovery/replication study design, an analysis is performed on the discovery sample data set $\mathbf{D}_{disc}$ and then on an independent replication sample data set $\mathbf{D}_{rep}$ and results across the data sets are compared to identify significant SNPs [Begum et al. 2012]. ComPaSS-GWAS is based on the framework of complementary pairs stability selection [Shah and Samworth 2013] which combines both sample splitting and resampling techniques. In ComPaSS-GWAS, the data $\mathbf{D}$ are randomly split into two complementary halves $\mathbf{D}_{kA}$ and $\mathbf{D}_{kB}$ in each of $k = 1 \dots K$ resamples. Each split may be done completely at random, or it may be done to preserve aspects of the data that may impact the analysis such as gender ratio or stratification. Each half of a split of the data is a conditionally independent sub-sample given the available data $\mathbf{D}$ and the two halves are analogous to randomly selected discovery and replication samples. Under this assumption, each half is analyzed with GWAS methods and the results from each half are compared to see whether the variants are corroborated across the split. For each split, SNPs corroboration evidence is identified in matrix $\mathbf{\Gamma}$ as

$$\Gamma_{jk} = \begin{cases} 1 & \text{if } \max(p_{jkA}, p_{jkB}) \leq \alpha \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where j is the j 'th SNP, k is the k 'th random split, p_{jkA} and p_{jkB} are the GWAS p-values of SNP j based on halves A and B of split k respectively and α is the critical value. ComPaSS-GWAS calculates a final score aggregating over all K resamples for SNP j , γ_j , where

$$\gamma_j = \frac{1}{K} \sum_{k=1}^K \Gamma_{jk}. \quad (3)$$

The range of γ_j is from 0 to 1, with higher values giving more support for corroboration of the SNP. The larger the number of random splits, K , used, the more precise the value of γ_j will be; $K = 100$ will be used throughout unless otherwise noted. The set of selected SNPs are defined as those SNPs where $\gamma_j \geq \eta$, where η is a predetermined ‘‘corroboration’’ parameter specifying the proportion of random splits where a SNP was corroborated. Thus, the identification of corroborated SNPs in ComPaSS-GWAS depends on two parameters; the critical value used for corroboration on each split, α , and the proportion of splits that are corroborated, η . As the value of η increases, the more confident one can be of the corroborated SNPs association with the trait for the given α .

Simulation Experiments

The genotype data used in the simulation experiments were based on the TSS genotypes. The Trinity Student Study was designed to identify genetic determinants involved in the variation of traits related to folate and vitamin B₁₂ metabolism in healthy young students at Trinity College Dublin in Ireland [Desch et al. 2013, Molloy et al 2016]. The GWAS data set included 2,232 unrelated participants with 757,533 autosomal SNPs with MAF ≥ 0.01 . Genotyping data was identical across all resamples. Traits were simulated based on three different models. In the first, a completely non-genetic trait (as assumed under the null hypothesis of a GWAS) was generated that followed a standard normal distribution, $y_i \sim N(0,1)$ in order to determine type I error under the null hypothesis. Under this model the heritability of the trait (h^2) was 0.0. In the second, the alternate (causal) model was based on three independent causal SNPs selected on chromosome 4 with minor allele frequencies (MAFs) of 0.05. The phenotype, y , was defined as

$$y_i = \sum_{j=1}^3 g_{ij} \beta_j + \epsilon_i$$

where y_i is the quantitative phenotypic value of subject i , g_{ij} is the genotype value of the j 'th causal SNP for subject i coded as the number of copies of the minor allele, β_j is the effect of the j 'th causal SNP and $\epsilon_i \sim N(0, \sigma^2)$ is a Gaussian error term. The phenotypes were generated such that the locus specific heritability (h_L^2) of the causal SNP was 0.01, 0.02 and 0.03, respectively and σ^2 chosen such that the total heritability of the simulated trait (h^2) was

0.06. This model was used to determine type I error under the alternative model by excluding all the genotypes on chromosome 4 (which includes causal SNPs and SNPs in LD with the causal SNPs that may skew the estimate if included) and power, by considering each of the 3 causal variants on chromosome 4 separately. Each simulation experiment was based on 1000 independent replicates.

Because many quantitative traits, including the original TSS traits, are skewed and have to be transformed to reduce non-normality, the third trait model was simulated to explore the sensitivity of ComPaSS-GWAS to minor violation of the normality assumption of the error terms. Under this model the null trait was generated from a gamma distribution with a shape and scale parameter of 3 and 20 respectively, $y_i \sim \text{Gamma}(3, 20)$, and then $\log_{10}$ transformed to be approximately normal. As in the first null model, the heritability of the trait (h^2) was 0.0.

First, simulation experiments were performed to determine empirically derived parameter values for ComPaSS-GWAS in order to determine a type I error that is comparable to that of a traditional GWAS with critical value of 5×10^{-8} . Various values of the within split critical value parameter, α , and the corroboration parameter, η , were considered under the null hypothesis. Second, type I error and power were determined for selected parameter settings under the alternate hypothesis. And third, type I error rate and power were determined comparing a traditional GWAS, a two-stage model (i.e, ComPaSS-GWAS with a single split), and ComPaSS-GWAS with the number of splits, K , ranging from 1 to 100.

The type I error rates were calculated as the proportion of non-causal SNPs selected of the total number of SNPs tested averaged over replicates. For the alternate model with causal SNPs, the type I error rates were calculated with SNPs from chromosomes other than chromosome 4, which contained the causal SNPs. Power was calculated as the proportion of replicates where each specific causal SNP was identified. Summary statistics were determined for ComPaSS-GWAS under a variety of parameter values and for a traditional GWAS of the full sample using PLINK [Purcell et al. 2007] with a traditional genome-wide critical value of 5×10^{-8} .

Validation

In order to validate the method, ComPaSS-GWAS should be able to identify, in a single study, associations that have been previously replicated in other studies. To demonstrate this, Pyridoxal 5'-phosphate (PLP), a form of vitamin B₆ measured in the TSS [Carter et al. 2015], was reanalyzed to validate ComPaSS-GWAS because several previously published trait-marker associations have been replicated. Following the approach of Carter et al. [2015], raw PLP values were $\log_{10}$ transformed and pre-adjusted for age, sex, and vitamin B₆ intake; the residuals from this pre-adjusted model were used for analysis. As in the original analysis, seventy-four subjects with extreme values of vitamin B₆ intake (greater than 11000 $\mu\text{g}/\text{day}$ from fortified foods/supplements) were excluded from the analyses reducing the sample size to 2,158 subjects. These same data were analyzed with ComPaSS-GWAS with splits that control for the gender ratio and results obtained from ComPaSS-GWAS were compared to results from the literature.

RESULTS

Simulation Results

The type I error rates of ComPaSS-GWAS under a null model following normal distribution are presented in Table I for several combinations of the critical value parameter, α , and the corroboration parameter, η . Several of these combinations had no type I errors observed in any of the replicates. The critical value and corroboration parameters of $\alpha = 1 \times 10^{-3}$ and $\eta = 0.60$, had a type I error rate of 2.90×10^{-8} which was nearly identical to the type I error rates obtained from the traditional GWAS with a critical value of 5×10^{-8} on these same data (3.83×10^{-8}). A paired t-test comparing the type I error rates from ComPaSS-GWAS (for $K = 100$ splits) at critical value and corroboration parameters of $\alpha = 1 \times 10^{-3}$ and $\eta = 0.60$ with the type I error rates from a traditional GWAS of the full sample for 1000 replicates was not significant ($p = 0.145$).

The type I error rates for the model with $\log_{10}$ transformed gamma distributed traits are presented in Supplementary Table I. These results are nearly identical to those from the null hypothesis under the normal distribution (Table I). This suggests that the non-normal distribution and $\log_{10}$ transformation of the traits used in the original TSS analysis of PLP [Carter et al. 2015] were appropriate for the validation of ComPaSS-GWAS.

The power to detect a SNP with locus specific heritabilities of 0.01, 0.02 and 0.3 was determined for the causal model. Table II presents the observed type I error and power for a traditional GWAS and selected critical value and corroboration parameter combinations of ComPaSS-GWAS. Again, the type I error rate and power for ComPaSS-GWAS with a critical value and corroboration parameter of $\alpha = 1 \times 10^{-3}$ and $\eta = 0.60$ was comparable to that for the traditional GWAS. Observed type I error rates for ComPaSS-GWAS parameters combinations are presented in Supplemental Table IIa for additional parameter combinations. Observed power is presented in Supplemental Tables II b–d for traits with locus-specific heritabilities of 0.01, 0.02 and 0.03, respectively.

The impact of resampling

In ComPaSS-GWAS, as the number of repeated random splits increases, the type I error rate decreases. This reduction of type I error allows for a less stringent critical value for a resample than for a two-stage GWAS based on a single split, thus increasing power. Figure I presents the type I error rate of a traditional GWAS of the full sample with a critical value of 5×10^{-8} and ComPaSS-GWAS with $\alpha = 10^{-3}$ and $\eta = 0.6$ as a function of the number of splits used. Because power is dependent on the type I error rate, the critical value parameter of ComPaSS-GWAS, α , was adjusted so that the type I error rate was approximately the nominal rate of 5×10^{-8} . Figure II presents the power of a traditional GWAS of the full sample with a critical value of 5×10^{-8} and ComPaSS-GWAS with the critical value parameter adjusted for comparable type I error rates. As the number of splits increased, the power also increased. The type I error and power with 100 splits, i.e. resamples, appear to be nearly identical to those for a traditional GWAS.

Table III presents the comparisons of the type I error and power between a single sample split and ComPaSS-GWAS with 100 random splits. For a single split (analogous to a two-

stage GWAS with a critical value of 10^{-3} , the observed type I error rate was approximately $\alpha^2 = 10^{-6}$, consistent with regression theory. Thus, an α of $\sqrt{5 \times 10^{-8}}$ would be required to achieve the desired rate of 5×10^{-8} for a single split. The power of ComPaSS-GWAS increased compared to that of a single split (two-stage) approach with similar type I error rates for locus specific heritabilities of 0.01 and 0.02. The power is nearly 1.0 when the locus-specific heritability is 0.03 or above, regardless of the critical value or method used.

Validation of ComPaSS-GWAS with the TSS data

As reported in Carter et al. [2015] nine SNPs in and around the *ALPL* gene had genome-wide significant associations with PLP ($p < 5 \times 10^{-8}$). These same data were analyzed with ComPaSS-GWAS with $K = 100$ splits that controlled for the gender ratio. Table IV presents information on the nine SNPs identified by Carter et al. [2015] including the SNP name, location, minor allele frequency, GWAS p -value, functional information on *ALPL* (the gene suggested to be biologically involved in variation of PLP values by Carter et al. [2015]) obtained from WANNVAR [Yang and Wang 2015] and the University of California at Santa Cruz genome browser [Kent et al. 2002], and the reference for the four replicated SNPs; as well as the scores from ComPaSS-GWAS for α values of 1×10^{-3} and 1×10^{-4} for each of these SNPs. All of these nine genome-wide significant SNPs were identified by ComPaSS-GWAS with parameters of $\alpha = 1 \times 10^{-3}$ and $\eta = 0.6$. ComPaSS-GWAS did not detect any other additional SNPs. Four of the nine SNPs identified in the TSS data were previously reported in the literature (rs1780324 [Yuan et al. 2008], rs1697421 [Keene et al. 2014], rs1780316 [Keene et al. 2014], rs1256335 [Harza et al. 2009]) and can be taken to be independent replications (the “replicated SNP set”). For a parameter set of $\alpha = 1 \times 10^{-3}$ and $\eta = 0.60$, the parameter combination with a type I error and power similar to that of a genome-wide significance level in a traditional GWAS, ComPaSS-GWAS identified all nine SNPs reported in Carter et al. [2015] including all four SNPs in the replicated SNP set. When the parameter set was more conservative ($\alpha = 1 \times 10^{-4}$ and $\eta = 0.5$), five of the nine SNPs were identified with ComPaSS-GWAS. Three of these five SNPs (rs1780324, rs1697421, and rs1256335) were in the replicated SNP set; two (rs1256335 and rs1772719) were not in the replication set, but still located in or near the *ALPL* gene.

DISCUSSION

The use of an independent replication study is taken to be the preferred means of corroborating results. In the absence of an independent replication sample, a more stringent critical value could be used, but this still fails to address the replication issue. Alternatively, a two-stage GWAS that splits the (single sample) data into a discovery and replication sample could be used for replication, but this has been shown to be less powerful than the single sample approach. Here, we present ComPaSS-GWAS, a method based on the framework of complementary pairs stability selection [Shah and Samworth 2013], that has better type I error control than data split into a single discovery and replication sample, with power roughly equal to that of a single sample approach. Based on resampling theory, ComPaSS-GWAS takes an approach that reduces the variability of statistics due to sampling variability [Breiman 1996, Buhlmann and Yu 2002] and improves error control. In this study, the critical value and corroboration parameters of $\alpha = 1 \times 10^{-3}$ and $\eta = 0.60$ were

shown to have a type I error rate and power similar to that of that obtained in a traditional GWAS, and we recommend them as parameters for analyzing GWAS data.

As shown in Figure I, the mean and confidence intervals for the type I error rate for ComPaSS-GWAS decreased and narrowed, as the number of splits increased from 1 to 100 and stabilized at about 30 resamples. Furthermore, when the number of resamples were greater than 30, the mean and confidence intervals were nearly identical to those for a traditional GWAS, at least for samples of the size considered here. When the type I error rate was adjusted to maintain the nominal rate for different numbers of splits, the power increased with the number of random splits (Figure II). Several factors can influence the number of resamples required to provide stability for the type I error rate and power, particularly the overall size of the sample and criteria used to determine the splits. We believe that a minimum of 100 resamples should be considered to provide stable statistical properties. Although performing a GWAS on 100 splits of the data took substantially longer than a traditional one-sample GWAS, the ComPaSS-GWAS analysis for the PLP phenotype in the TSS data (757533 SNPs on 2158 subjects) took an average of 121.8 minutes (sd=0.24 min) to run on a UNIX server using a single 2.6GHz Intel(R) Xeon(R) CPU.

The repeated sampling aspect of ComPaSS-GWAS ameliorates the limitations of the single split sample (or two-stage) approach and the corroboration aspect of complementary pairs stability selection can be taken to be an alternative method of corroboration similar in spirit to that using a discovery and replication study design. In addition, the use of resampling techniques was shown to be somewhat more powerful than approaches that are based on a single split of the data such as a two stage GWAS, best illustrated by causative SNPs with locus specific heritabilities of 0.01 and 0.02. The SNPs selected by ComPaSS-GWAS may be more likely to be representative of the actual effect that could be replicated given appropriate replication data.

Carter et al. [2015] performed a traditional GWAS with a critical value of 5×10^{-8} and identified nine SNPs in the PLP analysis in the TSS. All of these nine genome-wide significant SNPs were identified by ComPaSS-GWAS with our recommended parameters of $\alpha = 1 \times 10^{-3}$ and $\eta = 0.6$, with no other SNPs identified. Four of the SNPs identified in the TSS data were previously reported in the literature (rs1780324 [Yuan et al. 2008], rs1697421 [Keene et al. 2014], rs1780316 [Keene et al. 2014], rs1256335 [Harza et al. 2009]) and can be taken to be independent replications of each SNP (the replicated SNP set). ComPaSS-GWAS thus identified all of the SNPs in the replication SNP set. With a more conservative set of parameters ($\alpha = 1 \times 10^{-4}$ and $\eta = 0.5$) ComPaSS-GWAS identified a set of five SNPs with three of these five SNPs (rs1780324, rs1697421, and rs1256335) being in the replicated SNP set. Assuming that the four replicated SNPs are in fact true replications, ComPaSS-GWAS identified four out of four (100%) of the SNPs in the replicated SNP set with our recommended critical values and three out of four (75%) SNPs in the replicated SNP set with a more conservative set of critical values. Thus, it appears that ComPaSS-GWAS can be used as an ad-hoc method of corroboration when independent replication data are not available.

Although ComPaSS-GWAS was designed with the focus on analysis of data sets without a replication data set itself, it may also be useful on data sets where replication data are available but not ideal. For example, some replication data sets are small enough that combined analysis with a meta-analysis may be primarily driven by the larger cohort, and in other cases heterogeneity between the discovery and replication cohort may make meta-analysis results problematic for at least some regions of the genome. In these cases, the additional insight from the resampling based approximation to replication that ComPaSS-GWAS provides may be invaluable. An example of the use of ComPaSS-GWAS on this situation is presented in Szekely et al. [2018].

ComPaSS-GWAS is presented here under the framework of a simple linear regression based GWAS for an unrelated population based GWAS, but the ComPaSS framework is general enough to adapt to different analysis methods for each split. For example, if minor population structure was present in the data, the simple linear regression based GWAS may be used by including principle components in either a pre-regression model or as covariates in the GWAS. If population structure is a larger concern, the traditional simple linear regression GWAS analysis may be replaced with an analysis such as EMMAX [Kang et al. 2010] to account for the population structure in the data. An example of where EMMAX was used in place of the traditional GWAS is Szekely et al. [2018]. When such a substitution is made, the parameter values that were found to be appropriate in the simulations presented here may not be optimal, and additional simulations may be needed to select appropriate parameter values for the modified analysis.

The use of ComPaSS-GWAS may be problematic when rare SNPs are present in the data. A rare SNP can by chance have all subjects with the minor allele placed into the same half of the split thus ensuring it cannot be corroborated. The probability of this happening on a single split can be modeled by a binomial distribution with the number of trials corresponding to the number of subjects, n , observing the minor allele (approximately $2 \times \text{MAF} \times n$) and success rate of 0.5 (equal probability of being in each half of a split). Specifically the probability that this would happen is $2 \times P(0 \text{ subjects with minor allele in only one half})$ which may be computed as $2 \times 0.5^{\tilde{n}}$ where $\tilde{n}$ is the number of subjects with the minor allele which is approximately the floor of $2 \times \text{MAF} \times n$ for rare SNPs. As the probability of a random split assigning all subjects with the minor allele to one subsample increases (smaller n or rarer MAF), the $\hat{\gamma}$ values of these SNPs can be lowered due to the inability to replicate the SNP on these splits. If this probability is too high, the $\hat{\gamma}$ value may be reduced enough to be problematic. For example, in a small cohort with 500 subjects where a standard GWAS MAF filtering of 1% is applied, the rarest SNP in the analysis would have a probability of 0.19% of having all subjects with the minor allele in one half of the data, which would not be problematic. However, if the same cohort was genotyped on a chip focusing on rare variants and a lower MAF threshold of 0.5% was applied, the probability increases to 6.25%, which can have a significant impact on the score of the rarer SNPs. Although the loss of the rare SNP in a sample split is a plausible reason that it may not be replicated, this is not a desirable aspect for an analysis' procedure. One possible solution to the issue of losing a rare SNP on a split due to resampling may be to use an approach similar to the fractional resampling scheme used in Sabourin et al. [2015]. With

this type of approach, rather than splitting the samples, each individual would have a regression weight ($w_i \sim U(0,1)$) determining the contribution for analysis on D_A and have a complementary weight ($1 - w_i$) for the analysis on D_B . Such a modification would ensure each SNP is included on each analysis, but the conditional independence between D_A and D_B on each split is lost. Sabourin et al. [2015] demonstrated that in a resampling procedure based on just D_A , fractional resampling performed nearly identical to subsampling half the data, but further testing would be necessary to see how it would affect the complementary pairs based analysis.

In summary, ComPaSS-GWAS is a method that can be applied to traits (including case control binary traits) where a traditional regression based GWAS would be an appropriate analysis, e.g., traits that could be analyzed in PLINK [Purcell et al. 2007], but where an appropriate independent replication sample may not be available. Alternatively, it would be useful in prioritizing a large number of significant or nearly significant results for follow-up. Because the framework of ComPaSS-GWAS is based on complementary pairs stability selection, all that is required is a method to select SNPs in each half of the data. The ComPaSS framework is not limited to a traditional regression based GWAS analysis (simple linear regression and logistic regression for quantitative and case/control traits respectively); alternative methods could be implemented if required. Software for ComPaSS-GWAS will be available as an R package, *r/ComPaSS*, which will utilize PLINK for efficient GWAS analyses.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This project was supported in part by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health. None of the authors has any source of conflict of interest.

REFERENCES

- Alexander DH, & Lange K (2011). Stability selection for genome-wide association. *Genet Epidemiol*, 35(7), 722–728. [PubMed: 22009793]
- Begum F, Ghosh D, Tseng GC, & Feingold E (2012). Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Res*, 40(9), 3777–3784. [PubMed: 22241776]
- Breiman L (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140.
- Buhlmann P, & Yu B (2002). Analyzing bagging. *Annals of Statistics*, 30(4), 927–961.
- Cantor RM, Lange K, & Sinsheimer JS (2010). Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *Am J Hum Genet*, 86(1), 6–22. [PubMed: 20074509]
- Carter TC, Pangilinan F, Molloy AM, Fan R, Wang Y, Shane B, ... Mills JL (2015). Common Variants at Putative Regulatory Sites of the Tissue Nonspecific Alkaline Phosphatase Gene Influence Circulating Pyridoxal 5'-Phosphate Concentration in Healthy Adults. *J Nutr*, 145(7), 1386–1393. [PubMed: 25972531]
- Desch KC, Ozel AB, Siemieniak D, Kalish Y, Shavit JA, Thornburg CD, ... Ginsburg D (2013). Linkage analysis identifies a locus for plasma von Willebrand factor undetected by genome-wide association. *Proc Natl Acad Sci U S A*, 110(2), 588–593. [PubMed: 23267103]

- Hazra A, Kraft P, Lazarus R, Chen C, Chanock SJ, Jacques P, ... Hunter DJ (2009). Genome-wide significant predictors of metabolites in the one-carbon metabolism pathway. *Hum Mol Genet*, 18(23), 4677–4687. [PubMed: 19744961]
- Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, ... Eskin E (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*, 42(4), 348–354. [PubMed: 20208533]
- Keene KL, Chen WM, Chen F, Williams SR, Elkhatib SD, Hsu FC, ... Sale MM (2014). Genetic Associations with Plasma B12, B6, and Folate Levels in an Ischemic Stroke Population from the Vitamin Intervention for Stroke Prevention (VISP) Trial. *Front Public Health*, 2(112), 112. [PubMed: 25147783]
- Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, & Haussler D (2002). The human genome browser at UCSC. *Genome Res*, 12(6), 996–1006. [PubMed: 12045153]
- McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, & Hirschhorn JN (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, 9(5), 356–369. [PubMed: 18398418]
- Meinshausen N, & Bühlmann P (2010). Stability selection. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 72(4), 417–473.
- Molloy AM, Pangilinan F, Mills JL, Shane B, O'Neill MB, McGaughey DM, ... Brody LC (2016). A Common Polymorphism in HIBCH Influences Methylmalonic Acid Concentrations in Blood Independently of Cobalamin. *Am J Hum Genet*, 98(5), 869–882. [PubMed: 27132595]
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, ... Sham PC (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3), 559–575. [PubMed: 17701901]
- Sabourin J, Nobel AB, & Valdar W (2015). Fine-mapping additive and dominant SNP effects using group-LASSO and fractional resample model averaging. *Genet Epidemiol*, 39(2), 77–88. [PubMed: 25417853]
- Schwantes-An TH, Sung H, Sabourin JA, Justice CM, Sorant AJM, & Wilson AF (2016). Type I error rates of rare single nucleotide variants are inflated in tests of association with non-normally distributed traits using simple linear regression methods. *BMC Proc*, 10(Suppl 7), 385–388. [PubMed: 27980666]
- Shah RD, & Samworth RJ (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 75(1), 55–80.
- Skol AD, Scott LJ, Abecasis GR, & Boehnke M (2007). Optimal designs for two-stage genome-wide association studies. *Genet Epidemiol*, 31(7), 776–788. [PubMed: 17549752]
- Sung H, Kim Y, Cai J, Cropp CD, Simpson CL, Li Q, ... Wilson AF (2011). Comparison of results from tests of association in unrelated individuals with uncollapsed and collapsed sequence variants using tiled regression. *BMC Proc*, 5 Suppl 9(9), S15.
- Szekely E, Schwantes-An TL, Justice CM, Sabourin JA, Jansen PR, Muetzel RL, ... Shaw P (2018). Genetic associations with childhood brain growth, defined in two longitudinal cohorts. *Genet Epidemiol*, 42(4), 405–414. [PubMed: 29682794]
- Valdar W, Holmes CC, Mott R, & Flint J (2009). Mapping in structured populations by resample model averaging. *Genetics*, 182(4), 1263–1277. [PubMed: 19474203]
- Valdar W, Sabourin J, Nobel A, & Holmes CC (2012). Reprioritizing genetic associations in hit regions using LASSO-based resample model averaging. *Genet Epidemiol*, 36(5), 451–462. [PubMed: 22549815]
- Wilson AF, & Ziegler A (2011). Lessons learned from Genetic Analysis Workshop 17: transitioning from genome-wide association studies to whole-genome statistical genetic analysis. *Genet Epidemiol*, 35 Suppl 1(S1), S107–114. [PubMed: 22128050]
- Yang H, & Wang K (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat Protoc*, 10(10), 1556–1566. [PubMed: 26379229]
- Yuan X, Waterworth D, Perry JR, Lim N, Song K, Chambers JC, ... Mooser V (2008). Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes. *Am J Hum Genet*, 83(4), 520–528. [PubMed: 18940312]

Type I error rate of a GWAS and ComPaSS for different numbers of random splits

Figure I:

95% confidence intervals for the observed Type I error rate for a traditional GWAS of the full sample with a critical value of 5×10^{-8} and ComPaSS-GWAS with different numbers of random splits for parameters $\alpha = 10^{-3}$ and $\eta = 0.6$.

Figure II:

95% confidence intervals for the observed power for a traditional GWAS of the full sample with critical value of 5×10^{-8} and ComPaSS-GWAS with different numbers of random splits for parameters $\eta = 0.6$ with α adjusted to maintain the nominal type I error rate of 5×10^{-8} .

Table I.

Estimated type I error rates for a ComPaSS-GWAS analysis of a normally distributed null trait. The empiric type I error rate for a GWAS with a critical value of 5×10^{-8} was 3.83×10^{-8} .

η	Split critical value (α)				
	1×10^{-2}	1×10^{-3}	1×10^{-4}	1×10^{-5}	1×10^{-6}
0.2	9.44×10^{-5}	9.16×10^{-7}	7.92×10^{-9}	0	0
0.4	3.06×10^{-5}	1.91×10^{-7}	1.32×10^{-9}	0	0
0.5	1.61×10^{-5}	7.66×10^{-8}	0	0	0
0.6	7.62×10^{-6}	2.90×10^{-8}	0	0	0
0.8	9.37×10^{-7}	1.32×10^{-9}	0	0	0

Table II.

Type I error rate and power under the causal model for a traditional GWAS and ComPaSS-GWAS with indicated parameter settings.

Method	Type I error rate	Power for SNP with h_L^2 of		
		0.01	0.02	0.03
GWAS (critical value 5×10^{-8})	5.05×10^{-8}	0.231	0.938	0.999
ComPaSS ($\alpha = 1 \times 10^{-3}$; $\eta = 0.5$)	1.28×10^{-7}	0.263	0.956	1
ComPaSS ($\alpha = 1 \times 10^{-3}$; $\eta = 0.6$)	6.03×10^{-8}	0.212	0.903	0.999
ComPaSS ($\alpha = 1 \times 10^{-4}$; $\eta = 0.2$)	7.01×10^{-9}	0.151	0.881	0.999

Table III.

Type I error rate and power under the causal model for a single split of the data and ComPaSS-GWAS using 100 random splits.

Method	Type I error rate	Power for SNP with h_L^2 of		
		0.01	0.02	0.03
Single split ($\alpha = 1 \times 10^{-3}$)	9.74×10^{-7}	0.261	0.891	0.990
ComPaSS ($\alpha = 1 \times 10^{-3}$; $\eta = 0.6$)	6.03×10^{-8}	0.212	0.903	0.999
Single split ($\alpha = \sqrt{5} \times 10^{-8}$)	5.75×10^{-8}	0.127	0.785	0.971

Table IV.

Reported genome-wide significance SNPs from traditional GWAS with critical value of 5×10^{-8} of the PLP trait in the TSS from Carter et al. [2015] and ComPaSS-GWAS selected SNPs.

SNP	Chr	Mb	MAF	GWAS p-value	ComPaSS-GWAS Score		ALPL	Association reported prior to Carter et al. [2015]
					$\alpha = 1 \times 10^{-3}$	$\alpha = 1 \times 10^{-4}$		
rs1780324	1	21.694	0.49	1.3×10^{-11}	0.95 [†]	0.84 [*]	Intergenic	Yuan et al. [2008]
rs1697421	1	21.696	0.49	3.5×10^{-11}	0.94 [†]	0.77 [*]	Intergenic	Keene et al. [2014]
rs1780321	1	21.697	0.33	2.8×10^{-12}	0.98 [†]	0.87 [*]	Intergenic	
rs4021228	1	21.736	0.47	4.8×10^{-9}	0.83 [†]	0.33	Intronic	
rs1780316	1	21.762	0.07	8.7×10^{-9}	0.72 [†]	0.16	Exonic	Keene et al. [2014]
rs1256335	1	21.763	0.22	5.0×10^{-14}	1.00 [†]	0.98 [*]	Intronic	Harza et al. [2009]
rs2275370	1	21.773	0.21	1.8×10^{-9}	0.78 [†]	0.33	Intronic	
rs1772719	1	21.777	0.23	2.5×10^{-16}	1.00 [†]	1.00 [*]	UTR3	
rs1772720	1	21.778	0.08	8.8×10^{-10}	0.86 [†]	0.46	Downstream	

[†] selected for ComPaSS-GWAS parameter set $\alpha = 1 \times 10^{-3}$ and $\eta = 0.6$

^{*} selected for ComPaSS-GWAS parameter set $\alpha = 1 \times 10^{-4}$ and $\eta = 0.5$