HHS Public Access

Author manuscript

J Hum Genet. Author manuscript; available in PMC 2016 May 07.

Published in final edited form as:

J Hum Genet. 2015 December ; 60(12): 729–738. doi:10.1038/jhg.2015.110.

Evaluation of a two-step iterative resampling procedure for internal validation of genome-wide association studies

Guolian Kang[1], **Wei Liu**[1], **Cheng Cheng**[1], **Carmen L. Wilson**[2], **Geoffrey Neale**[3], **Jun J. Yang**[4], **Kirsten K. Ness**[2], **Leslie L. Robison**[2], **Melissa M Hudson**[2], and **Deo Kumar Srivastava**[1,#]

[1]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, USA

[2]Department of Epidemiology and Cancer Control, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, USA

[3]Hartwell Center for Bioinformatics and Biotechnology, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, USA

[4]Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, USA

Abstract

Genome-wide association studies (GWAS) have successfully identified many common genetic variants associated with complex diseases over the past decade. The "gold standard" method for validating the top single nucleotide polymorphisms (SNPs) identified in GWAS is to independently replicate the findings in similar or diverse large-scale external cohorts. However, for rare diseases, it can be difficult to find an external validation cohort within a reasonable timeframe. In such situations, resampling methods, such as the two-step iterative resampling (TSIR) approach have been used to identify SNPs associated with the outcome of interest. However, the TSIR approach involves choosing several parameters in each step, which can influence the performance of the approach. In this paper, we undertook extensive simulation studies to assess the effect of choice of different parameters on the type I error and power for both binary and continuous phenotypes and also compared the TSIR approach with the traditional one-stage (OS) and two-stage (TS) GWAS analysis. We illustrate the usefulness of the TSIR approach by applying it to a GWAS of childhood cancer survivors. Our results indicate that the TSIR approach with an at least 70:30 split and a cut-off of discovering and replicating SNPs at least 20 times in 100 replications provides conservative type I error control and has near "optimal" power for internally validated SNPs. Its performance is comparable to the TS GWAS for which an external validation cohort is available with only slight reduction in power in some situations. It has almost the same power as OS GWAS with conservative type I error which leads to fewer false positive findings. TSIR is a

[#]Address for correspondence: Deo Kumar Srivastava, Ph.D., Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, USA, Phone: +1-901-595-3372, Fax: +1-901-595-8843, kumar.srivastava@stjude.org.

Conflict of Interest

The authors have no conflict of interests to declare.

powerful and efficient method for identifying and internally validating SNPs for GWAS when independent cohorts for external validation may not be available.

Keywords

two-step iterative resampling approach; Genome-wide association studies; internal validation

Introduction

Under the common disease common variants (CDCV) hypothesis, genome-wide association studies (GWAS) have successfully identified associations between common genetic variants, such as single nucleotide polymorphisms (SNPs) with complex diseases[1–3]. A two-stage[4] or multiple-stage design[5–6] has been commonly applied to design GWAS to detect SNPs associated with complex diseases. For the two-stage design, the whole cohort is divided into discovery and replication/validation cohorts. In Stage I, the top signals/SNPs are identified in the discovery cohort using well-defined a priori criterion that are then replicated/validated in Stage II using an "independent" replication cohort, i.e. independent of the discovery cohort. For multiple-stage designs with more than two stages, after the first stage, stages II and beyond are usually employed to validate the top most significant markers for downstream analyses.

In GWAS involving rare diseases or outcomes in pediatric cancers we often aim to identify biologic markers that can predict treatment outcomes, help explain treatment-related toxicities, or help us understand the effects of treatment modalities on different subtypes of disease. Because these diseases are rare, some with prevalence rates of 1 per million, e.g. retinoblastoma and Ewing's Sarcoma[7,8], it may not be possible to find an external cohort to validate the top SNPs within a reasonable timeframe. Even when disease outcomes are not rare, it can also be hard to find a suitable external validation cohort. An example is the evaluation of genetic predictors of clinically ascertained outcomes in the SJLIFE cohort[9], a study among childhood cancer survivors treated at St. Jude Children's Research Hospital (SJCRH), who have survived 10 or more years from diagnosis and are at least 18 years of age. Because this study includes the largest cohort of childhood cancer survivors with prospective medical/clinical evaluation of health outcomes, it is extremely hard to find another cohort that has similarly ascertained health phenotypes[10]. In such situations, it is imperative that an innovative and robust internal validation approach is undertaken to validate the top SNPs identified through GWAS[10,11].

The current research was motivated by a study within the SJLIFE cohort designed to identify the SNPs associated with the obesity phenotype (evaluated as a binary outcome measure) in survivors of childhood cancer treated with cranial radiation for which an external cohort to validate our findings was not available[10]. Thus, we considered an internal validation approach, namely the two-step iterative re-sampling (TSIR) approach, used by Yang et al. [11] for identifying SNPs associated with the risk of relapse in children treated for acute lymphoblastic leukemia. An alternative approach would be to use a permutation approach[12,13], which is particularly suited to situations where the prevalence of the binary

outcome is low and the number of cases is small. Another permutation-based internal validation approach is the "profile significance"[14], particularly suited for situations where the global level of association between genomic features may be of interest. For large sample sizes these approaches can be computationally intensive and time consuming. The focus of this manuscript is to describe the operating characteristics of the TSIR approach.

The TSIR approach used in Yang et al.[11] can roughly be described as follows. The original cohort is split, using a $\pi : (1-\pi)$ ratio, with $\pi=0.5$, into discovery and replication cohorts. Using the discovery cohort, SNPs are individually tested for association with the outcome using Fine and Gray's hazard regression model. All SNPs that are significant at α_1 ($4.4\text{x}10^{-3}$) are carried forward to the replication step. A SNP identified in the discovery cohort is considered to be replicated if the same SNP is associated with the outcome in the replication step at α_2 (=0.05) significance level. This discovery-replication process is repeated 100 times and a particular SNP is designated as "associated" or "internally validated" with the outcome if it is discovered/replicated at least 10 times.

In the approach described by Yang et al.[11], there was no rationale or statistical justification provided for the following: (1) rationale for the 50:50 split of the original cohort into discovery and replication cohorts (2) the choice of $\alpha_1= 4.4\text{x}10^{-3}$ with α_2 is fixed at level 0.05 (α_2 =0.05)(3) a cut-off of 10 in the discovery-replication process. We were interested in assessing how the various choices in (1)–(3) above affect the statistical properties of the TSIR approach, how the TSIR approach performs for continuous and binary outcomes and finally how the performance of TSIR approach compares to the OS and TS GWAS analysis?

The research presented here, supported by extensive simulation studies, is designed to guide researchers to employ the appropriate choice of parameters when using the TSIR approach for their genomics research involving GWAS when external validation cohorts are not available. The usefulness of the TSIR approach is further demonstrated by applying it to data reported in Wilson et al.[10].

Method

Two-step iterative resampling (TSIR) procedure

The TSIR described by Yang et al.[11] was used in the context of survival data. However, in the current analysis we were interested in binary as well as continuous phenotypes. Accordingly, we discuss evaluation of binary and continuous end points in parallel.

We assume that, for GWAS in a one-stage design, there are N_0 controls and N_1 cases in a case-control genetic association study (total sample size $N = N_0 + N_1$) or N individuals in a genetic association study of a continuous phenotype and that the SNP of interest is biallelic. The 2 alleles at a SNP are denoted as A and a, where A is the minor allele and the three genotypes are AA, Aa, and aa. Suppose that observations (s_i, X_i, G_i), $i = 1, 2, \ldots N$, are available for N individuals, s_i is the indicator of case-control status or the quantitative value of the continuous phenotype of the subject i; $X_i = [x_{i1}, x_{i2}, \ldots, x_{im}]^T$ is the vector of m covariates to be included in the model (e.g., demographic or clinical variables); and $G_i = 0$, 1, or 2 is the numerical coding of the 3 genotype aa, Aa or AA of the SNP for an individual.

For the TSIR approach the original cohort is randomly split into discovery and replication set by the ratio π:(1-π). A SNP is considered discovered and replicated if its association testing p-values are statistically significant at levels α_1 and α_2 in the discovery and replication steps, respectively (Figure 1). This process is repeated r=100 times and a SNP is considered to be "associated" with the phenotype if the SNP is discovered and replicated at least r times in 100 repetitions.

It may be noted that if we conduct the association analysis, logistic regression or linear regression, with the entire cohort (sample size N), i.e. without splitting the sample into discovery and replication cohorts then we are conducting traditional OS GWAS. However, without having an independent validation cohort there is always a concern of false discoveries and the discovered SNPs are subject to suspicion and criticism. In such situations the proposed TSIR approach overcomes this limitation and the simulations studies suggest that the results based on TSIR approach are more believable and defensible.

Traditional Two-stage GWAS design

The traditional two-stage design was introduced as an efficient alternative to conducting a single GWAS analysis (one-stage design) that includes all genotyped participants[4,15]. The two-stage (TS) design was proposed as a way to economize on the cost of genotyping, which were quite high when initial GWAS studies were undertaken. In a TS design, N_1^{TS} and N_2^{TS} are the number of individuals available for genetic analysis in each of the two stages with $N^{TS}=N_1^{TS}+N_2^{TS}$ being the total sample size. In stage I, a small set of the individuals $N_1 = N^* \pi$($\pi < 0.3$, π is the proportion of participants included in stage I) would be used as the discovery cohort for whole-genome genotyping and the promising markers at liberal levels of type I error control ($\alpha_1 \geq 0.01$) would be identified. Then, in Stage II, a larger cohort of individuals, independent of those in Stage I, of size $N_2 = N^*(1-\pi)$, with $(1 - \pi) > 0.7$, would be used as a validation cohort for genotyping the markers selected in stage I. The final list of markers would be determined based on the results from the Stage II data or in combined Stages I and II data at more stringent levels of type I error control ($\alpha_2=\alpha/M$, where M is the number of markers associated with phenotype in Stage I and α is the genomewide significance level)[4]. However, as genotyping costs have decreased over time, the design of TS GWAS has also changed accordingly. Importantly, many more individuals are genotyped for markers spread across the genome in stage I and tested for association with the phenotype of interest at increasingly stringent level of type I error (α_1), while a more liberal level of type I error control (α_2) is used in stage II in smaller cohort of individuals, independent of those in Stage I, as the validation cohort[15,16]. In order to compare the traditional TS GWAS with the proposed TSIR approach in the current analysis, the parameters for the TS design were chosen to reflect the set-up of the TSIR approach.

Simulation studies

We performed extensive simulation studies to evaluate the empirical power and type I error rate of the TSIR procedure for testing associations of SNPs with binary and continuous phenotypes for different parameter combinations as shown below. To evaluate the merits of TSIR, we varied the proportion (π) of individuals included in the discovery cohort from 0.3

to 0.9 in increments of 0.1, and chose α_1 = 0.01, 0.001, and 0.0001 (with α_2 fixed at 0.05) for both binary and continuous phenotypes. The prevalence of disease was set at 0.01, 0.1 and 0.3 for the binary phenotype.

Data generation

Genotype generations—Given the minor allele frequency (MAF) p_A of minor allele A (major allele a), the genotype frequencies $p(G=g)$ were calculated according to Hardy–Weinberg equilibrium (HWE) law, that is, $p(G=0)=(1-p_A)^2$, $p(G=1)=2p_A(1-p_A)$, $p(G=2)=(p_A)^2$. Two covariates were considered in our models: x_1 a binary variable that takes value of 1 with a probability of 0.5 and 0 otherwise, and x_2 a continuous variable that follows a standard normal distribution. Based on these assumed distributions, the complete data on the genotypes and 2 covariates for a population of 2,000,000 individuals was generated.

Phenotype generation

Binary phenotype: The case-control status was determined from the generated genotype and covariate data according to the model similar to Kang et al.[17]:

$$\Pr(s_i=1|G_i,\ x_{i1},\ x_{i2})=\frac{\exp(\alpha_0+\theta G_i+0.5x_{i1}+0.5x_{i2})}{1+\exp(\alpha_0+\theta G_i+0.5x_{i1}+0.5x_{i2})}. \quad (1)$$

We controlled the baseline disease prevalence by setting α_0 to 0.3, 0.1 and 0.01 to represent high, moderate and low disease prevalence in the case where all three regression coefficients corresponding to SNP, x_{i1} and x_{i2} are 0.

Continuous phenotype: The continuous phenotype was generated from the generated genotype and covariate data according to the model outlined in Wu et al. (2011):

$$s_i=\theta G_i+0.5x_{i1}+0.5x_{i2}+e_i, \quad (2)$$

where e_i is the random error following a standard normal distribution.

Using the models proposed in (1) and (2), N_1 cases and N_0 controls or N samples were randomly generated from the simulated population of 2,000,000 individuals for binary or continuous outcomes, respectively.

Assessment of Type I error probability

Two values for the MAFs considered were 0.05 and 0.2 in our evaluation of type I error. The case-control status or the continuous phenotype was determined from the generated genotype and covariate data by using their respective models in (1) and (2), with $\theta = 0$. To estimate the type I error rate of the TSIR approach, 10,000,000 replicated datasets were simulated for the case-control model, with 250, 350, 450, 550 and 1000 cases and 1, 2.5 and 4 times the numbers of independent controls under the null hypothesis of H_0: $\theta = 0$, respectively. The same numbers of replicated datasets were simulated for the continuous phenotype study, with 500, 700, 900, 1100, and 2000 samples under the null hypothesis of

H_0: $\theta = 0$. We used the number of successful replication r = 10, 20 and 25 to estimate the type I error rate of the TSIR procedure. TSIR was applied to each replicate dataset and the empirical type I error rate was estimated as the proportion of replicates in which the tested SNP was identified as "associated" with the phenotype using TSIR procedure.

Assessment of power

Three genetic disease models were considered: additive, dominant, and recessive with their corresponding genotype codings of 0, 1, 2; 0, 1, 1; and 0, 0, 1 for three genotypes aa, Aa and AA. The case-control status or the continuous phenotype was determined from the generated genotype and covariate data according to the simulation methods given above, with $\theta = 0.2$, 0.4 and 0.7 to mimic a small, moderate and larger effect sizes, respectively. Datasets were generated 10,000 times for each configuration. TSIR used for the type I error simulation was applied to each replicate data-set, and power was estimated as the proportion of replicates in which the tested SNP was identified as "validated." Based on type I error simulation results, we used n=20 in the power estimation of TSIR procedure, as it seemed to control the type I error rate at the desirable levels such as $5x10^{-5}$ or $5x10^{-6}$.

Comparison with the two-stage (TS) design

To investigate the performance of TSIR, we compared the power of TSIR with that of the TS design under two scenarios based on the different sample sizes. Under the fir the number of individuals in stage I is the same as those for the TSIR approach. Under the second scenario it is assumed that we have another independent replication cohort but the sample sizes in the two stages are similar to the sample sizes in the discovery and replication stages of the TSIR approach. The TS designs under two scenarios are denoted by TS_1 and TS_2, respectively. To make comparisons reasonable, we selected a significance level combination of $\alpha_1=10^{-4}$ and $\alpha_2=0.05$ for the two-stage design to ensure an overall type I error rate per SNP of 5×10^{-6} (Table 1)[4]. Based on power simulation results above, power was optimized for the TSIR when the ratio of individuals in the discovery and replication cohorts was 70:30 and assuming that both the "discovery" and "validation" cohorts were sampled from the same homogenous population. For the TSIR approach, we considered the number of cases for the binary phenotype to be N_1= 280, 560, and 1120 and the number of controls to be 2.5 times the number of cases (N_2= 700, 1400, and 2800), with total sample size of $N=(N_1+N_2)$ = 980, 1960 and 3920. For TS1, we considered $N_1^{TS}=N$ and $N_2^{TS}=\frac{3}{7}N_1^{TS}=\frac{3}{7}N$ ($N^{TS}=N_1^{TS}+N_2^{TS}=N+\frac{3N}{7}=\frac{10N}{7}$) to be the number of individuals in Stage I (discovery) and Stage II (validation) of the two-stage design. We then randomly sampled N_1 and N_0 individuals ($N=N_1+N_0$) from the general population of 2,000,000 individuals for the TSIR approach and also used the same sample as the Stage I discovery cohort ($N_1^{TS}=N=980$, 1960 and 3920) for the TS analysis. To create a validation cohort for Stage II for the TS approach, we randomly sampled $N_2^{TS}=\frac{3}{7}N_1^{TS}=420$, 840, and 1680 individuals from the general population of 2,000,000, this kept the ratio of participants in the discovery to validation datasets (0.7:0.3) consistent with the TSIR approach. For TS_2, we considered the same sample as that for TSIR but mimicked the features of TS design by

splitting N individuals into $N_1^{TS}=\frac{7}{10}N$ for stage I discovery cohort and $N_2^{TS}=\frac{3}{10}N$ for stage II validation cohort and then applied association analysis methods to these two cohorts.

When considering the continuous phenotype, the number of individuals included in the analyses for the TSIR approach were N = 700, 1400, and 2800. A similar approach was adopted for analysis of the TS design with the continuous phenotype. Datasets were generated 10,000 times for each configuration. The power of TS_1 and TS_2, was estimated as the proportion of replicates in which a SNP was discovered in stage I at $p < \alpha_1$ (where $\alpha_1=10^{-4}$) and validated in stage II at $p < \alpha_2$ (where $\alpha_2=0.05$).

The power properties of TSIR approach were also compared to OS GWAS for simulated N individuals for both binary and continuous outcomes. The power of OS procedure was estimated as the proportion of replicates in which a SNP was statistically significant at a level of $\alpha_1\alpha_2$.

Simulation results

Empirical type I error rate of TSIR—Table 1 and Supplementary Table S1 display the empirical type I error rates when r=20.

When evaluating the binary phenotype using the TSIR approach, as π increased so did the empirical type I error, however, the type I error was still maintained at a level of $\alpha_1 \times \alpha_2$ per SNP. If $\alpha_1 = 10^{-3}$ and 10^{-4}, the medians of empirical type one error rate were 0.000033 (range: 0 0.00007), and 0.000001 (range: 0.000000 0.0000056), respectively. The TSIR procedure controlled type I error per SNP at 5×10^{-5} and 5×10^{-6} if $\alpha_1 = 10^{-3}$ and 10^{-4}, respectively, which is the same as the type I error control ($\alpha_1 \times \alpha_2$) seen in the TS approach. For smaller sample sizes, such as for N_1=250, it was seen that, irrespective of the prevalence, the type I error rate was much closer to $\alpha_1\alpha_2$ with discovery cohort proportions of π = 0.6 and 0.7. However, with the discovery cohort proportion of π=0.7, the type I error rates were much better compared to discovery cohort proportions less than 0.7, particularly for more stringent values of α_1, e.g. for α_1=0.0001. The type I error rates corresponding to π=0.60 and 0.70 are 1.8 and 2.7, 2.8 and 4.2, and 4.1 and 4.9 corresponding to sample sizes of 250, 550 and 1000, respectively.

It is seen that as the sample size increases the type I error control improves the proportion of individuals allocated to the discovery stage relative to the validation stage is minimal when the type I error control used in the discovery phase is somewhat larger $\alpha_1 \geq 0.001$. But, for more stringent values of α_1, such as α_1=0.0001, the TSIR approach with 70% in the discovery cohort still provides qualitatively superior type I error control. From Table 1, similar conclusions can be drawn when the phenotype is continuous.

The type I error rate per SNP was not maintained at $\alpha_1 \times \alpha_2$ level when r=10 was chosen as the validation cutoff (Supplementary Table S2). Similarly, the type I error rate per SNP was too conservative when r=25 was chosen as the validation cutoff (Supplementary Table S3). In addition, with the discovery cohort proportion of π=0.8 and 0.9, the type I error rates were close to those when π=0.7 for $\alpha_1 > 0.0001$, but were a little higher than those when π

=0.7 for α_1 = 0.0001 (Supplementary Table S3). However, the power when π =0.7 for all simulated α_1 plateaued (see below for empirical power). Thus, for power estimation or comparison, we will plot the results for π up to 0.7.

Empirical power of TSIR—All power evaluations, discussed below, were conducted using r=20 as the validation cutoff. Based on the extensive power simulation studies the following conclusions can be drawn:

Binary phenotype: From Figures 1–2 and Supplementary Figures S1–S2 it is seen that when α_1 =0.01, the power of the TSIR approach for detecting a SNP with a MAF of 0.2 is not affected by the proportion of individuals (π) included in Stage I (discovery cohort). However, as expected, for more conservative values of α_1, that is, $\alpha_1 \leq 10^{-3}$, the power of TSIR approach first increased sharply then plateaued with the increasing values of π. Also, not surprising, as α_1 became more conservative the power of TSIR approach decreased.

When both the effect size of the SNP and the sample size were small or very large, the proportion of individuals included in discovery cohort had little effect on the power of the TSIR approach. In contrast, if the effect size was moderate or small but the sample size was large, or the effect size was large but the sample size was small, then the power estimates were optimized when π ranged between 0.5 and 0.7. However, for stringent values of α_1, π=0.7 for the discovery cohort provided consistently better power. Neither the prevalence of disease nor the MAF affected the power of the TSIR approach (supplementary Figures 1–2).

Continuous phenotype: As seen in Figure 4, for a SNP with large effect size, e.g. θ=0.7, and MAF=0.2, the power of TSIR was close to 1 regardless of π, α_1 and sample size (≥500). Similarly, as seen in Figure 3, for a SNP with a small effect size of θ=0.2 and MAF=0.05, the power of TSIR approach was close to 0 regardless of π, α_1 and sample size (≤2000). As seen with the binary phenotype, if $\alpha_1 = 10^{-3}$, 10^{-4}, and 10^{-5}, with the increasing proportion π of individuals included in step I, "discovery cohort," the power of TSIR first increased sharply then became plateaued around π=0.6 and 0.7, which is particularly true for smaller values of α_1.

Power comparisons among OS, TSIR, TS_1 and TS_2—It is clear from Figure 5 that, not surprisingly, the power of TS_2 was larger due to the fact that TS_2 procedure used more individuals, and the power of TS_1 was lower than OS even though TS_1 uses the same number of individuals as OS but, under OS procedure, the analysis is conducted only once. The power of TS_2 was larger than that of the TSIR approach especially when the sample size and effect size were moderate. However, this has to be balanced by the fact that the TS procedure used 30% more individuals (for the validation cohort) than those used for TSIR approach. For the binary phenotype the largest difference in power estimates between both approaches was seen to be 0.14 when there were 560 cases and 1400 controls corresponding to a SNP with MAF of 0.2 and effect size of 0.4. For other situations, corresponding to large effect sizes or small/large sample sizes the power estimates for the two approaches were comparable and reasonably close. The power of TSIR was almost identical to that of OS, which is expected, since for TSIR and OS the sample was the same but TSIR used a re-sampling statistical technique to better control possible false positives (the simulated type I

error rate of TSIR was smaller than $\alpha_1 \times \alpha_2$ which is the theoretical type I error rate for OS) at the same time without sacrificing the power as TS_1 did.

All simulation results for TS_1, TS_2 and TSIR were conducted using the two-sided test in stage 2 for TS_1 and TS_2 or in step 2 for TSIR, which will have slightly reduced power since it ignores the direction of association. We re-ran all simulations using exactly the same parameters as those for Figure 5 and re-calculated the power for TS_1, TS_2 and TSIR but used one-sided test in stage 2 or step 2. For the binary phenotype, the maximum gain in power for TSIR with one-sided test was 0.01. But for the TS_1 and TS_2, the maximum gain in power was 0.052 and 0.053, respectively. The very similar conclusions held for continuous phenotype. One-sided test did improve the power of TSIR but the power increase was relatively small which means TSIR approach is relatively robust to one-sided or two-sided test due to 100-round iterative resampling. For TS_1 and TS_2, though one-sided test improved their power at about 5%, our simulations suggest that, in general, TS_1 had smaller power than TSIR and TS_2 but TS_2 is not feasible due to lack of availability of an external validation cohort. Thus the results further confirm the good performance and the practical usefulness of TSIR compared to OS or TS with or without the availability of additional validation cohort in ongoing and future GWAS or NGS.

Simulation studies for the obesity SNPs—Simulation studies were also conducted to estimate the empirical power for detecting association between SNPs identified for the obesity phenotype in Table 2 using the TSIR approach[10]. The simulation parameters were taken to reflect the MAF, prevalence and effect size (in terms of odds ratios) observed in a cohort of cancer survivors exposed to cranial radiation therapy (CRT) cohort (Table 2). Specifically, for each SNP, we first generated genotype data under Hardy–Weinberg equilibrium with MAF similar to that observed in the survivor cohort for a population with 2,000,000 individuals as above; then generated phenotype (case-control) data from the generated genotype dataset using the model above with OR and prevalence of the disease same as those observed for the survivor cohort. Finally, a sample of 365 cases and 411 controls was randomly drawn from the population and analyzed using the TSIR approach. This process was repeated 10,000,000 and 10,000 times for the estimation of empirical type I error and power, respectively. The empirical type I error rate was estimated as the proportion of times the SNP associated with obesity was validated wrongly and the empirical power was estimated as the proportion of times the SNP was validated correctly. For example, for SNP rs2769921 with MAF of 0.43, there was 69% power using TSIR approach to detect if the SNP was truly associated with obesity in cancer survivors with an odds ratio of 0.577; however, there was only 3.4×10^{-6} chance to wrongly identify that this SNP was associated with obesity in cancer survivors (Table 2). Similarly, for SNP rs4971486 with MAF of 0.22, the power to detect it was 0.69 if it was truly associated with obesity with an OR of 1.9 and the type I error was 4.51×10^{-6} if it were not associated with obesity.

Discussion

It is well recognized that the top signals emerging from GWAS or next generation sequencing must be validated in independent cohorts[18,19]. However, independent external

validation cohorts among those with rare diseases can be difficult to find within a limited timeframe. In such situations, two stage resampling approaches have been used to identify and validate the SNPs associated with binary phenotypes of interest if the number of cases is not small. One such approach, namely the TSIR approach, has been proposed and we evaluated its operating characteristics through extensive simulation studies. These studies suggest that the TSIR approach, with the choice of 7:3 partitioning of the original cohort into "discovery" and "replication" cohorts, a cut-off of r=20 for identifying SNPs associated with the phenotype in 100 replications, and strictly controlling the type I error rate below $\alpha_1 \times \alpha_2$ provide good type I error control and near optimal power. In our analyses, using the parameters above the power of the TSIR approach was found to be slightly lower than that observed for the TS_2 approach, but this is due to the fact that fewer individuals were included in the analyses of the TSIR approach than in the TS approach. Interestingly, with same sample sizes, the power of the TSIR was almost identical to that of OS, but TSIR approach had a conservative type I error control than OS. It is often not possible to obtain an external cohort for validation for very rare diseases and unique cohorts. Thus, based on our analyses, we recommend the use of the TSIR approach for identifying the top candidate SNPs associated with a particular phenotype of interest. Identification of SNPs using the TSIR approach may help prioritize those candidate SNPs that should be evaluated in laboratory studies. However, it should be noted that the TSIR approach is only applicable when the size of the population of interest is sufficiently large for sample splitting.

In GWAS, the first step prior to statistical genetic association testing is quality control analysis which includes Hardy-Weinburg Disequilibrium (HWD) test to remove markers departing from HWE[20]. Thus, in our TSIR simulations the genotype data is generated by assuming HWE. However, if we are concerned about HWD in a GWAS, then some statistical association testing method[21] that can adjust for HWD is available and can be used to replace the logistic regression in TSIR but we would expect the conclusions drawn above would still hold. Furthermore, in our simulations we used logistic regression. In the literature, there are many statistical methods available for genetic association testing, which can also be employed in TSIR approach[18]. We would expect that the conclusions drawn above would still hold. The common SNPs with MAFs of 0.2 and 0.05 in GWAS were investigated in this study. Currently rare variant association identification in the next generation sequencing studies is highly in demand due to missing inheritability of complex trait post-GWAS[22]. If the sample size of the study is large enough so that the splitting of the cohort is reasonable, then the TSIR approach allows for sufficient statistical power to detect the rare variants in both steps[23]. With smaller sample sizes where splitting is unreasonable, a permutation test may be applied. However, for rare variants association, we often conduct gene (set)-based analysis[24,25], not single SNP-based analysis. This way we can employ TSIR procedure as an internal validation method if there is no external validation cohort available.

If the individuals in the study cohort are from different populations, we can just simply adjust for population stratification by including genetic ancestry score as covariates in the logistic regression model[26]. Here our interest was on detecting genetic effect on the binary outcome. In Post-GWAS, besides rare variant associations above, gene-environment interaction also plays an important role in finding missing inheritability for complex trait[27].

They are worthy of investigation by simulations but we would expect that similar conclusions would hold.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank two reviewers for their helpful comments which have significantly improved the paper. This research was supported by St. Jude Children's Research Hospital Cancer Center Support (CORE) grant CA21765 from the National Cancer Institute and by the American Lebanese and Syrian Associated Charities (ALSAC).

References

1. Klein RJ1, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J. Complement factor H polymorphism in age-related macular degeneration. Science. 2005; 308(5720):385–389. [PubMed: 15761122]
2. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2007; 445:881–885. [PubMed: 17293876]
3. The Wellcome Trust Case Control Consortium (WTCCC). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678. [PubMed: 17554300]
4. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet. 2006; 38:209–213. [PubMed: 16415888]
5. Pahl, Roman; Schäfer, Helmut; Müller, Hans-Helge. Optimal multistage designs—a general framework for efficient genome-wide association studies. Biostatistics. 2008; 10:297–309. [PubMed: 19075295]
6. Rothman N, Garcia-Closas M, Chatterjee N, Malats N, Wu X, Figueroa JD, Real FX, Van Den Berg D, Matullo G, Baris D, et al. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. Nat Genet. 2010; 42(11):978–984. [PubMed: 20972438]
7. Gurney JG, Severson RK, Davis S, Robison LL. Incidence of cancer in children in the United States. Sex-, race-, and 1-year age-specific rates by histologic type. Cancer. 1995; 75:2186–95. [PubMed: 7697611]
8. Wheeler HE, Maitland ML, Dolan ME, Cox NJ, Ratain MJ. Cancer pharmacogenomics: strategies and challenges. Nat Rev Genet. 2013; 14(1):23–34. [PubMed: 23183705]
9. Hudson MM, Ness KK, Nolan VG, Armstrong GT, Green DM, Morris EB, Spunt SL, Metzger ML, Krull KR, Klosky JL, Srivastava DK, Robison LL. Prospective medical assessment of adults surviving childhood cancer: study design, cohort characteristics, and feasibility of the St. Jude Lifetime Cohort Study. Pediatr Blood Cancer. 2011; 56(5):825–836. [PubMed: 21370418]
10. Wilson CL, Liu L, Yang JJ, Kang G, Ojha RP, Neale G, Srivastava DK, Gurney JG, Hudson MM, Robison LL, Ness KK. Genetic and clinical factors associated with obesity among adult survivors of childhood cancer: a report from the St. Jude Lifetime cohort. Cancer. 2015 In press.
11. Yang JJ, Cheng C, Devidas M, Cao X, Campana D, Yang W, Fan Y, Neale G, Cox N, Scheet P, et al. Genome-wide association study identifies germline polymorphisms associated with relapse of childhood acute lymphoblastic leukemia. Blood. 2012; 120(20):4197–4204. [PubMed: 23007406]
12. Elliott KS, Chapman K, Day-Williams A, Panoutsopoulou K, Southam L, Lindgren CM, Arden N, Aslam N, Birrell F, Carluke I, Carr A, Deloukas P, Doherty M, Loughlin J, McCaskie A, Ollier WE, Rai A, Ralston S, Reed MR, Spector TD, Valdes AM, Wallis GA, Wilkinson M, Zeggini E. GIANT consortium; arcOGEN consortium. Evaluation of the genetic overlap between

osteoarthritis with body mass index and height using genome-wide association scan data. Ann Rheum Dis. 2013; 72:935–941. [PubMed: 22956599]

13. Hayes MG, Pluzhnikov A, Miyake K, Sun Y, Ng MC, Roe CA, Below JE, Nicolae RI, Konkashbaev A, Bell GI, Cox NJ, Hanis CL. Identification of type 2 diabetes genes in Mexican Americans through genome-wide association studies. Diabetes. 2007; 56(12):3033–3044. [PubMed: 17846124]
14. Cheng C. Internal validation inferences of significant genomic features in genome-wide screening. Computational Statistics and Data Analysis. 2009; 53:788–800. [PubMed: 20084293]
15. Simón-Sánchez J, Schulte C, Bras JM, Sharma M, Gibbs JR, Berg D, Paisan-Ruiz C, Lichtner P, Scholz SW, Hernandez DG, Krüger R, Federoff M, Klein C, Goate A, Perlmutter J, Bonin M, Nalls MA, Illig T, Gieger C, Houlden H, Steffens M, Okun MS, Racette BA, Cookson MR, Foote KD, Fernandez HH, Traynor BJ, Schreiber S, Arepalli S, Zonozi R, Gwinn K, van der Brug M, Lopez G, Chanock SJ, Schatzkin A, Park Y, Hollenbeck A, Gao J, Huang X, Wood NW, Lorenz D, Deuschl G, Chen H, Riess O, Hardy JA, Singleton AB, Gasser T. Genome-wide association study reveals genetic risk underlying Parkinson's disease. Nat Genet. 2009; 41(12):1308–1312. [PubMed: 19915575]
16. Yue WH, Wang HF, Sun LD, Tang FL, Liu ZH, Zhang HX, Li WQ, Zhang YL, Zhang Y, Ma CC, Du B, Wang LF, Ren YQ, Yang YF, Hu XF, Wang Y, Deng W, Tan LW, Tan YL, Chen Q, Xu GM, Yang GG, Zuo XB, Yan H, Ruan YY, Lu TL, Han X, Ma XH, Wang Y, Cai LW, Jin C, Zhang HY, Yan J, Mi WF, Yin XY, Ma WB, Liu Q, Kang L, Sun W, Pan CY, Shuang M, Yang FD, Wang CY, Yang JL, Li KQ, Ma X, Li LJ, Yu X, Li QZ, Huang X, Lv LX, Li T, Zhao GP, Huang W, Zhang XJ, Zhang D. Genome-wide association study identifies a susceptibility locus for schizophrenia in Han Chinese at 11p11.2. Nat Genet. 2011; 43(12):1228–1231. [PubMed: 22037552]
17. Kang G, Bi W, Zhao Y, Zhang JF, Yang JJ, Xu H, Loh ML, Hunger SP, Relling MV, Pounds S, Cheng C. A New System Identification Approach to Identify Genetic Variants in Sequencing Studies for a Binary Phenotype. Hum Hered. 2014; 78:104–116. [PubMed: 25096228]
18. Igl BW, Konig IR, Ziegler A. What do we mean by 'replication' and 'validation' in genome-wide association studies? Human Heredity. 2009; 67:66–68. [PubMed: 18931511]
19. Ioannidis JPA, Gilles T, Daly MJ. Validating, augmenting and refining genome-wide association signals. Nature Review Genetics. 2009; 10:318–329.
20. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics. 2007; 81(3):559–575. [PubMed: 17701901]
21. Song K, Elston RC. A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies. Stat Med. 2006; 25(1):105–126. [PubMed: 16220513]
22. Freidlin B, Zheng G, Li Z, Gastwirth JL. Trend tests for case-control studies of genetic markers: power, sample size and robustness. Hum Hered. 2002; 53:146–152. Erratum in Hum Hered 2009; 68: 220. [PubMed: 12145550]
23. Kang G, Lin D, Hakonarson H, Chen J. Two-stage extreme phenotype sequencing design for discovering and testing common and rare genetic variants: efficiency and power. Human Heredity. 2012; 73:139–147. [PubMed: 22678112]
24. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, Lin X. NHLBI GO Exome Sequencing Project—ESP Lung Project Team. Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. The American Journal of Human Genetics. 2012; 91(2):224–237. [PubMed: 22863193]
25. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011; 89(1):82–93. [PubMed: 21737059]
26. Kang G, Gao G, Shete S, Redden DT, Chang B-L, Rebbeck TR, Barnholtz-Sloan JS, Patterson N, Pajewski NM, Allison DB. Capitalizing on admixture in genome-wide association studies: A two-stage testing procedure and application to height in African-Americans. Front Genet. 2011; 2:11.10.3389/fgene.2011.00011

27. Chen J, Kang G, VanderWeele T, Zhang C, Mukherjee B. Efficient designs of gene-environment interaction studies: implications of Hardy-Weinberg equilibrium and gene-environment independence. Statistics in Medicine. 2012; 31(22):2516–2530. [PubMed: 22362617]

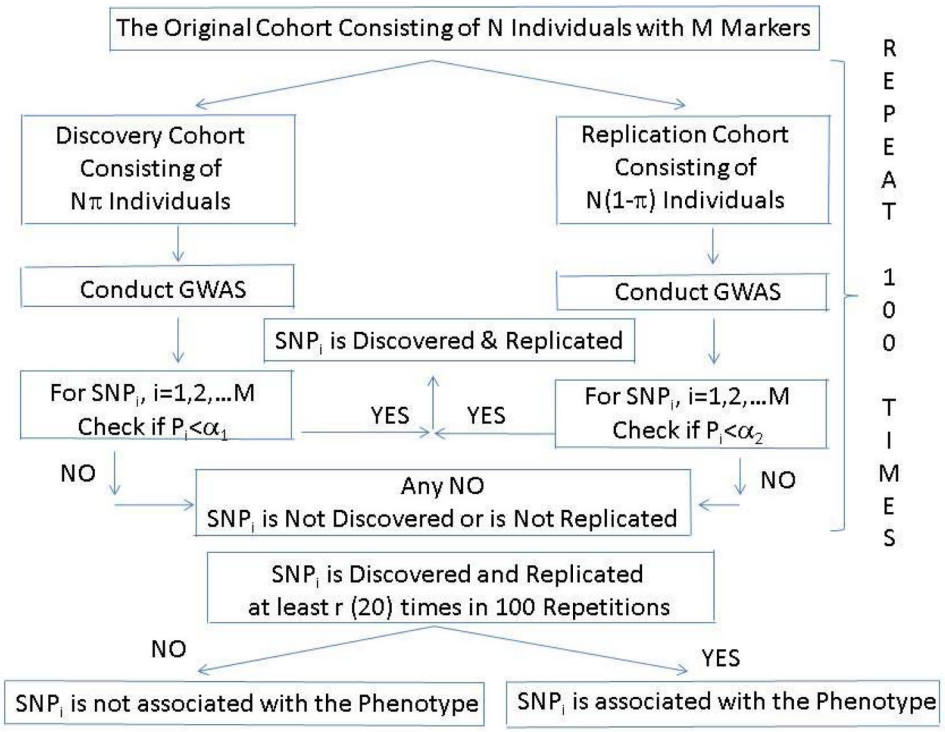

Figure 1.
The two-step iterative re-sampling procedure for GWAS

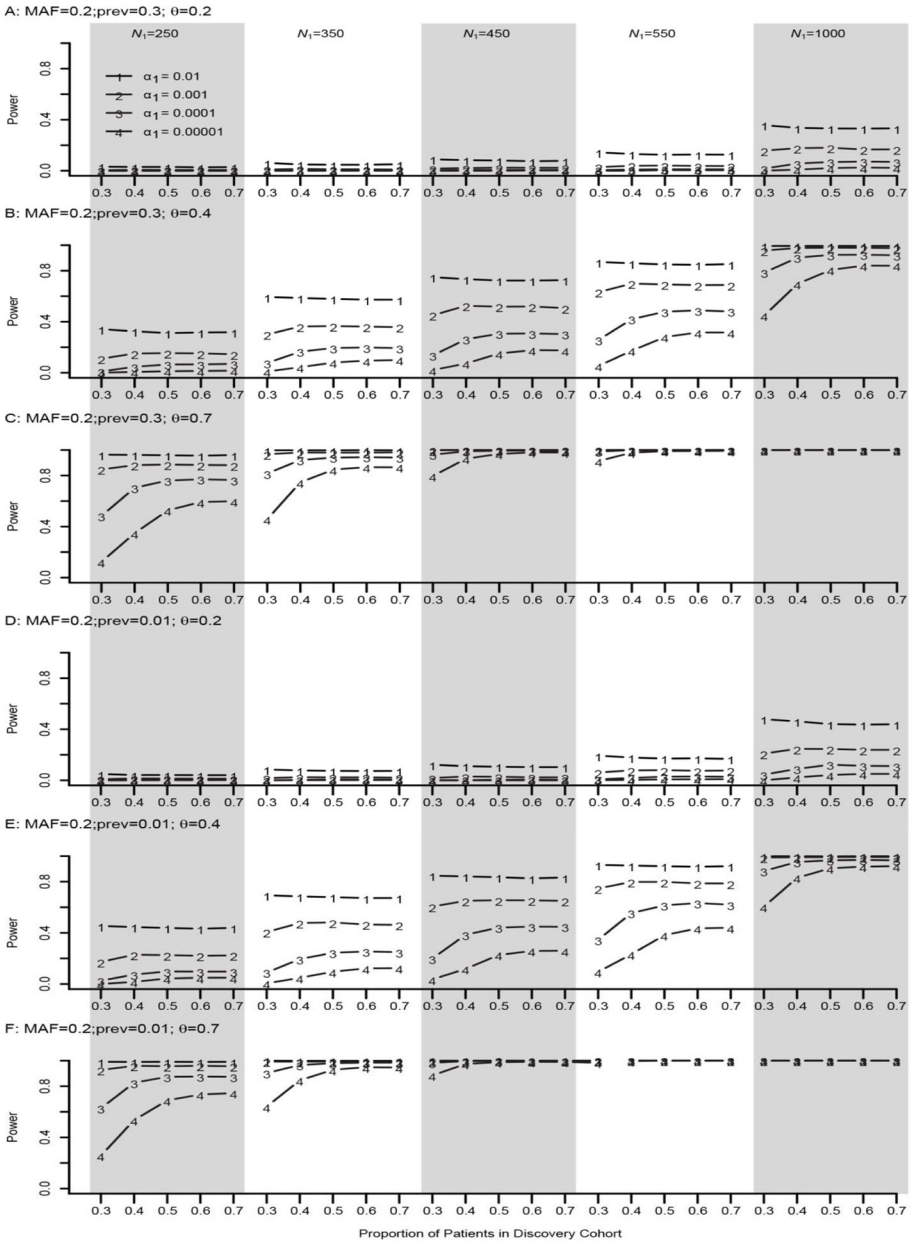

Figure 2. Empirical power of TSIR for detecting a SNP with a MAF of 0.2 for a binary phenotype

A, B and C are for small effect size $\theta = 0.2$, moderate effect size 0.4, and large effect size 0.7 for a large prevalence of 0.3, respectively. D, E and F are for small effect size $\theta = 0.2$, moderate effect size 0.4, and large effect size 0.7 for a small prevalence of 0.01, respectively. The solid lines with the numbers of 1–4 correspond to α_1 = 0.01, 0.001, 0.0001, and 0.00001, respectively.

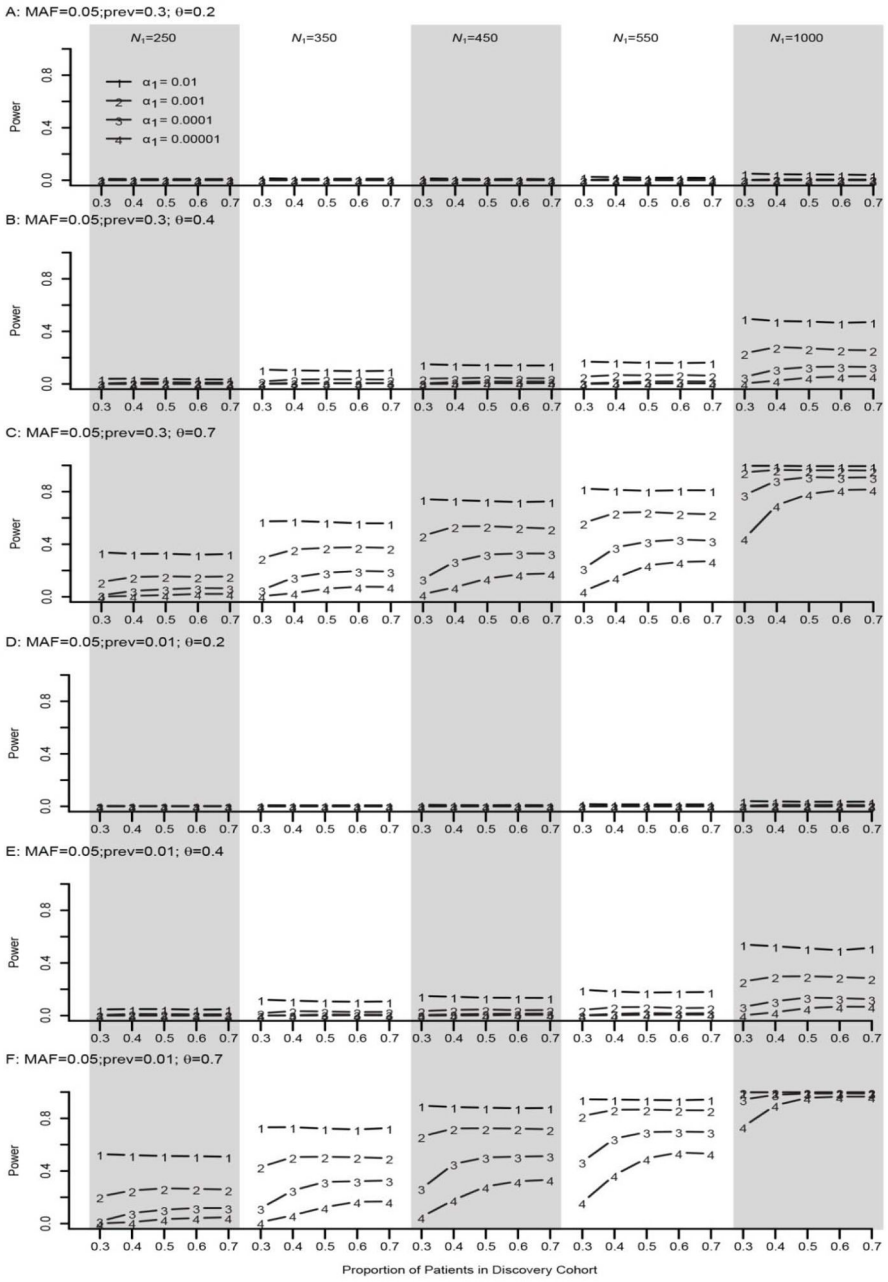

Figure 3. Empirical power of TSIR for detecting a SNP with a MAF of 0.05 for a binary phenotype

A, B and C are for small effect size $\theta = 0.2$, moderate effect size 0.4, and large effect size 0.7 for a large prevalence of 0.3, respectively. D, E and F are for small effect size $\theta = 0.2$, moderate effect size 0.4, and large effect size 0.7 for a small prevalence of 0.01, respectively. The solid lines with the numbers of 1–4 correspond to α_1 = 0.01, 0.001, 0.0001, and 0.00001, respectively.

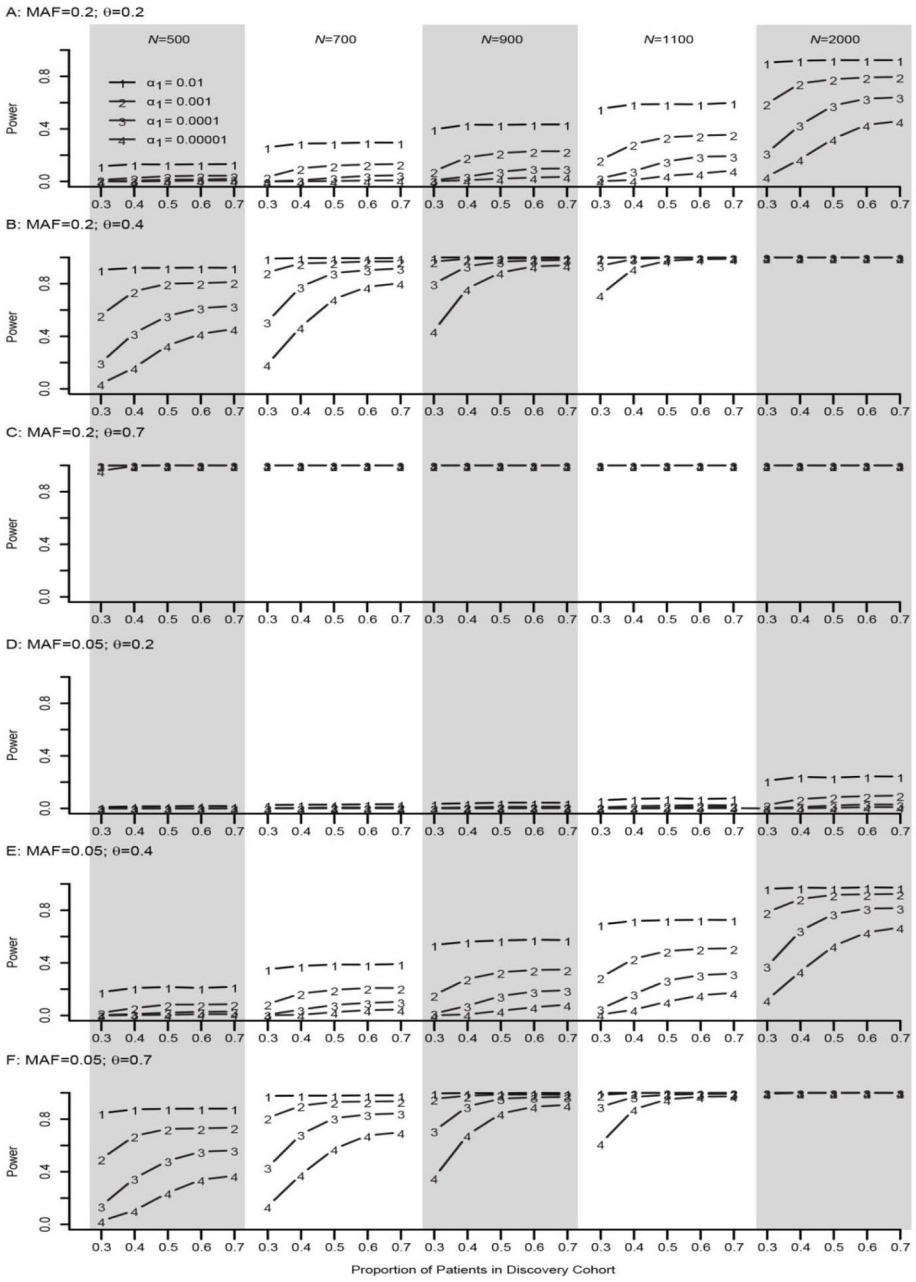

Figure 4. Empirical power of TSIR for detecting a SNP with a MAF of 0.05 (A–C) and 0.2 (D–E) for a continuous phenotype

A, B and C are for small effect size $\theta = 0.2$, moderate effect size 0.4, and large effect size 0.7, respectively. The solid lines with the numbers of 1–4 correspond to $\alpha_1 = 0.01$, 0.001, 0.0001, and 0.00001, respectively.

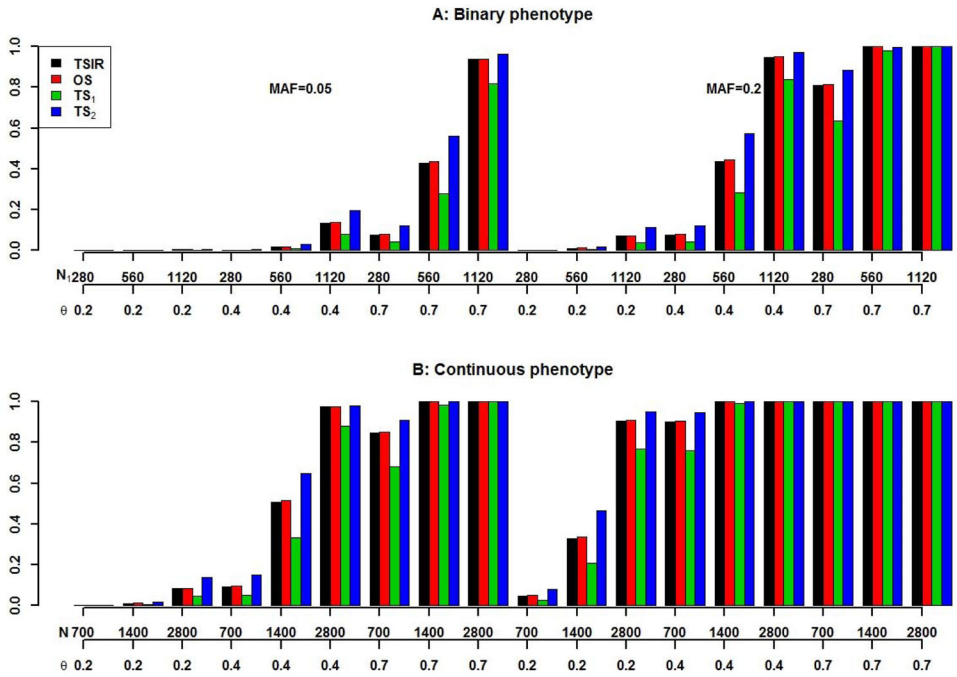

Figure 5. Power comparisons between TSIR, OS and TSs for detecting a SNP with MAFs of 0.05 and 0.2 associated with binary (A) and continuous (B) phenotypes

The first x-axis for A and B is for the number of cases (the number of controls is 2.5 times of the number of cases) and the number of individuals, respectively. The second x-axis is for θ. The four bars are for TSIR, OS, TS_1 and TS_2, respectively.

Table 1

The empirical type I error rates×10^5, and 10^6 of TSIR for identifying a CV with MAF of 0.2 associated with the binary and continuous phenotype at α_1=0.001 and 0.0001, respectively and using cut-off of t=20 in discovery-replication process

Binary Phenotype													
α_1				0.001					0.0001				
				Proportion in Discovery Cohort					Proportion in Discovery Cohort				
N_1	N_0/N_1	p_A	Prev.	0.3	0.4	0.5	0.6	0.7	0.3	0.4	0.5	0.6	0.7
250	1.0	0.2	0.30	0.07	1.3	3.1	3.6	3.9	0	0.00	0.6	2.1	2.3
250	2.5	0.2	0.30	0.18	1.9	3.5	4.0	4.3	0	0.40	1.6	2.5	3.6
250	4.0	0.2	0.30	0.33	2.0	3.4	4.2	4.3	0	0.10	1.8	3.2	4.3
250	1.0	0.2	0.10	0.05	1.3	3.0	3.6	3.9	0	0.10	0.6	1.7	3.0
250	2.5	0.2	0.10	0.20	1.6	3.2	3.9	4.2	0	0.00	1.2	2.8	3.4
250	4.0	0.2	0.10	0.26	1.7	3.4	3.9	4.3	0	0.40	1.3	2.2	2.7
250	1.0	0.2	0.01	0.07	1.2	2.8	3.6	3.9	0	0.00	0.8	1.8	2.7
250	2.5	0.2	0.01	0.26	2.0	3.7	4.4	4.7	0	0.30	1.8	4.0	4.4
250	4.0	0.2	0.01	0.39	2.3	4.1	4.6	4.8	0	0.00	1.1	3.2	3.9
550	1.0	0.2	0.30	0.36	1.3	2.6	3.1	3.2	0	0.00	0.2	0.6	1.3
550	2.5	0.2	0.30	0.25	2.1	3.8	4.4	4.6	0	0.30	1.9	3.4	3.9
550	4.0	0.2	0.30	0.25	2.1	3.8	4.4	4.7	0	0.20	2.2	2.9	4.0
550	1.0	0.2	0.10	0.20	1.8	3.5	4.0	4.1	0	0.10	1.3	2.3	3.3
550	2.5	0.2	0.10	0.32	2.2	4.0	4.5	4.9	0	0.30	1.8	2.9	3.4
550	4.0	0.2	0.10	0.41	2.2	3.9	4.3	4.6	0.1	0.40	2.3	4	5.0
550	1.0	0.2	0.01	0.22	1.9	3.3	4.0	4.2	0	0.30	1.7	2.8	4.2
550	2.5	0.2	0.01	0.42	2.5	4.2	4.7	5.0	0	0.40	2.0	3.4	4.1
550	4.0	0.2	0.01	0.54	2.5	4.3	4.8	5.1	0	0.80	2.6	3.9	4.3
1000	1.0	0.2	0.30	0.24	2.0	3.9	4.5	4.6	0	0.20	1.9	3.3	4.0
1000	2.5	0.2	0.30	0.37	2.2	3.9	4.5	4.5	0	0.40	1.9	3.6	3.6
1000	4.0	0.2	0.30	0.40	2.2	3.9	4.7	4.7	0	0.24	1.5	3.8	4.3
1000	1.0	0.2	0.10	0.32	2.0	3.7	4.3	4.5	0.1	0.20	2.0	3.3	4.6

Binary Phenotype															
			α_1			0.001					0.0001				
						Proportion in Discovery Cohort					Proportion in Discovery Cohort				
N_1	N_0/N_1	p_A	Prev.	0.3	0.4	0.5	0.6	0.7	0.3	0.4	0.5	0.6	0.7
1000	2.5	0.2	0.10	0.32	2.1	3.6	4.2	4.4	0	0.10	1.8	3.0	4.3
1000	4.0	0.2	0.10	0.35	2.4	3.9	4.5	4.8	0	0.60	2.0	4.5	4.4
1000	1.0	0.2	0.01	0.32	2.2	3.8	4.4	4.8	0	0.50	2.2	4.1	4.9
1000	2.5	0.2	0.01	0.46	2.4	3.9	4.6	4.7	0	0.10	2.3	3.9	4.9
1000	4.0	0.2	0.01	0.42	2.6	4.3	4.7	4.9	0	0.40	2.2	3.5	4.1

Continuous Phenotype													
	α_1	0.001						0.0001					
		Proportion in Discovery Cohort						Proportion in Discovery Cohort					
N	p_A	0.3	0.4	0.5	0.6	0.7		0.3	0.4	0.5	0.6	0.7	
500	0.2	0.37	2.3	4.3	4.6	4.9		0.1	0.80	2.6	3.4	4.7	
700	0.2	0.33	2.5	4.0	4.6	4.8		0	0.50	2.4	3.9	4.9	
900	0.2	0.35	2.4	4.2	4.9	5.1		0	0.20	1.5	3.6	4.3	
1100	0.2	0.35	2.2	4.0	4.5	4.7		0	0.20	2.1	3.3	4.6	
2000	0.2	0.39	2.2	4.0	4.4	4.7		0	0.60	2.6	4.3	4.3	
500	0.05	0.45	2.4	4.0	4.8	4.9		0.12	0.47	3.1	5.2	6.0	
700	0.05	0.45	2.3	3.9	4.6	4.9		0	0.80	2.5	3.7	4.0	
900	0.05	0.52	2.7	4.6	5.3	5.4		0	0.70	2.7	5.1	5.6	
1100	0.05	0.37	2.3	3.9	4.5	4.7		0	0.50	1.7	3.0	3.6	
2000	0.05	0.46	2.4	4.4	5.1	5.2		0	0.50	1.9	3.4	3.9	

Table 2

Empirical power and type I error rate ($\times 10^6$) simulation results for 21 SNPs identified associated with BMI.

SNP[1]	Chr.[2]	Location[3]	MA[4]	OR[5]	MAF[6]	Empirical power	Empirical type I error rate
rs4971486	2	4895318	G	1.936	0.2223	0.686	4.51
rs6745523	2	4908703	A	1.939	0.1768	0.518	3.30
rs1371477	2	4909920	T	1.867	0.1761	0.438	3.61
rs12648678	4	175598280	G	0.5091	0.1699	0.516	3.70
rs2171139	4	175624314	C	0.5196	0.1735	0.503	3.80
rs2443547	5	18173672	C	0.5911	0.4329	0.621	3.90
rs2923765	5	18177252	T	1.682	0.4588	0.588	3.70
rs2972927	5	18178225	T	0.5846	0.4381	0.626	4.60
rs2923756	5	18192581	G	1.836	0.1972	0.443	3.40
rs12514191	5	18198169	G	1.804	0.194	0.404	2.90
rs315825	5	18198934	A	0.595	0.4334	0.612	2.91
rs2938412	5	18205107	A	1.603	0.4472	0.434	4.20
rs453891	5	18212493	T	0.5946	0.4361	0.606	3.60
rs1316610	5	18230131	C	1.835	0.1893	0.453	2.60
rs2972892	5	18234352	T	0.5945	0.4548	0.633	3.30
rs2938451	5	18236104	A	0.5798	0.4186	0.672	3.10
rs2962166	5	18258575	T	0.5867	0.4183	0.627	3.90
rs2972911	5	18258703	A	0.5932	0.4432	0.605	2.70
rs2019973	5	134573992	T	0.6216	0.4691	0.441	3.41
rs2769921	13	107794883	C	0.577	0.43	0.690	3.40
rs12709954	19	56716579	T	0.5784	0.2874	0.515	3.22

[1] SNP identifier according to the dbSNP database

[2] Chromosome

[3] Physical location of SNP based on human gene assembly 19

[4] Minor allele

Author Manuscript

Author Manuscript

Author Manuscript

5 Odds ratios

6 Minor allele frequency