UW Biostatistics Working Paper Series

6-22-2016

A Powerful Statistical Framework for Generalization Testing in GWAS, with Application to the HCHS/SOL

Tamar Sofer
University of Washington, tsofer@uw.edu

Ruth Heller
Tel-Aviv University, ruheller@post.tau.ac.il

Marina Bogomolov
Technion-Israel Institute of Technology, marinabo@tx.technion.ac.il

Christy L. Avery
University of North Carolina at Chapel Hill, christy_avery@unc.edu

Mariaelisa Graff
University of N Carolina, Chapel Hill, migraff@email.unc.edu

See next page for additional authors

Suggested Citation

Sofer, Tamar; Heller, Ruth; Bogomolov, Marina; Avery, Christy L.; Graff, Mariaelisa; North, Kari E.; Reiner, Alex; Thornton, Timothy A.; Rice, Kenneth; Benjamini, Yoav; Laurie, Cathy C.; and Kerr, Kathleen F., "A Powerful Statistical Framework for Generalization Testing in GWAS, with Application to the HCHS/SOL" (June 2016). *UW Biostatistics Working Paper Series.* Working Paper 414.
http://biostats.bepress.com/uwbiostat/paper414

Authors
Tamar Sofer, Ruth Heller, Marina Bogomolov, Christy L. Avery, Mariaelisa Graff, Kari E. North, Alex Reiner, Timothy A. Thornton, Kenneth Rice, Yoav Benjamini, Cathy C. Laurie, and Kathleen F. Kerr

A Powerful Statistical Framework for Generalization Testing in GWAS, with Application to the HCHS/SOL

Tamar Sofer,[1*] Ruth Heller,[2] Marina Bogomolov,[3] Christy L. Avery,[4] Mariaelisa Graff,[4] Kari E. North,[4] Alex P. Reiner,[5] Timothy A. Thornton,[1] Kenneth Rice,[1] Yoav Benjamini,[1] Cathy C. Laurie,[1] and Kathleen F. Kerr[1]

[1]Department of Biostatistics, University of Washington, Seattle, WA 98105, USA
[2]Department of Statistics and Operations Research, Tel-Aviv University, Tel-Aviv, 6997801, Israel
[3]Faculty of Industrial Engineering and Management, Technion-Israel Institute of Technology, Haifa 3200003, Israel
[4]Department of Epidemiology, University of North Carolina, Chapel Hill, NC 27514, USA
[5]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98195, USA

Abstract

In GWAS, "generalization" is the replication of genotype-phenotype association in a population with different ancestry than the population in which it was first identified. The standard for reporting findings from a GWAS requires a two-stage design, in which discovered associations are replicated in an independent follow-up study. Current practices for declaring generalizations rely on testing associations while controlling the Family Wise Error Rate (FWER) in the discovery study, then separately controlling error measures in the follow-up study. While this approach limits false generalizations, we show that it does not guarantee control over the FWER or False Discovery Rate (FDR) of the generalization null hypotheses. In addition, it fails to leverage the two-stage design to increase power for detecting generalized associations. We develop a formal statistical framework for quantifying the evidence of generalization that accounts for the (in)consistency between the directions of associations in the discovery and follow-up studies. We develop the directional generalization FWER (FWER_g) and FDR (FDR_g) controlling r-values, which are used to declare associations as generalized. This framework extends to generalization testing when applied to a published list of SNP-trait associations. We show that our framework accommodates various SNP selection rules for generalization testing based on p-values in the discovery study, and still control FWER_g or FDR_g. A key finding is that it is often beneficial to use a more lenient p-value threshold then the genome-wide significance threshold. For instance, in a GWAS of Total Cholesterol (TC) in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL), when testing all SNPs with p-values$< 5 \times 10^{-8}$ (15 genomic

*Correspondence: tsofer@uw.edu

regions) for generalization in a large GWAS of whites, we generalized SNPs from 15 regions. But when testing all SNPs with p-values$< 6.6\times10^{-5}$ (89 regions), we generalized SNPs from 27 regions.

Introduction

When presenting results from genome-wide association studies (GWAS), current standards require a "two-stage design" in which possible discoveries in the first stage are replicated in an independent study (Cohen, 1999). 'Generalization' is the replication of a genotype-phenotype association in a population with different ancestry (or other characteristics) than the population in which it was first identified. Increasingly, generalization testing is performed as part of this two-stage design, primarily because GWAS is expanding into populations of diverse ancestry. First, with non-white discovery populations, there tend to be fewer similar studies available, so only generalization and not replication is feasible. Second, if the discovery study population is admixed (e.g. Hispanics/Latinos), it is customary to seek generalization in some of its parental populations.

Interestingly, even though the current standard for GWAS mandates replication, error controlling multiple testing adjustment procedures are often applied separately in the discovery and follow-up stages, without employing a replication- or generalization- based statistical framework. Bogomolov and Heller (2013) have shown that such approaches do not guarantee control over false generalization claims. Let the generalization null hypothesis state that a SNP is not associated with the trait in the discovery population, the follow-up population or both; and this null is rejected if evidence of association exists for both populations. Define generalization testing as any multiple testing adjustment pro-

cedure that controls measures of generalization error such as the Family-Wise Error Rate (FWER_g) or the False Discovery Rate (FDR_g). In this paper, we propose methods to test the generalization null hypotheses in GWAS, by expanding and adapting recent statistical methods developed for replication.

Bogomolov and Heller (2013) considered replication testing using discovery and follow-up studies, and developed multiple testing procedures with protection against erroneous replicability claims by controlling the FWER_g or the FDR_g. They showed that one must account for multiple testing in both the discovery and the follow-up studies to avoid a high number of erroneous replicability claims. Heller et al. (2014) suggested improvements to these procedures when used for GWAS, and developed r-values to quantify the evidence for replication while controlling FWER_g or FDR_g in GWAS. However, the r-values in Heller et al. (2014) do not account for the direction of the observed association. In this work we extend the r-values approach to incorporate the direction of observed associations. This acknowledges that we do not want to claim that an association generalizes if the direction of effect is different in the two populations. Our procedure performs directional control by using one-sided p-values to compute directional r-values at the generalization testing stage, despite using two-sided tests in the discovery stage. This makes our procedure more powerful than the procedure of Heller et al. (2014) for discovering associations with the same direction in both studies. We perform extensive simulations to study fixed and data-adaptive rules for selecting SNPs based on their p-values in the discovery study, and compare multiple-testing adjustment procedures in combination with these selection rules.

Materials and Methods

The generalization multiple testing framework

Expanding on the formal framework for replication of Heller et al. (2014), here we describe the generalization null hypothesis, propose a multiple testing adjustment procedure for generalization analysis, and contrast it with procedures currently used for single-stage studies.

Generalization versus discovery null hypotheses

There are two crucial differences between testing SNPs in a single-stage versus two-stage study design. In a single-stage design, all eligible SNPs in a single study are tested (after quality control filters). In contrast, in a two-stage design (1) the set of SNPs considered for generalization testing is based on results from the discovery study, and (2) tests of the null hypotheses are based on association analysis results from both the discovery and the generalization studies. Thus, suppose that m SNPs are tested in the discovery study. In a single-stage design, the discovery null is rejected for all significant associations in this study. However, in a two-stage design, the generalization null hypothesis is rejected when a SNP is associated with the trait in the generalization study as well, and the directions of association are the same.

Measures of false generalization

In multiple testing, there are two common measures of error: the FWER, and the FDR. In a single-stage GWAS, FWER is the probability of rejecting at least one null hypothesis

corresponding to a SNP not associated with the trait. FDR is the expected proportion of false null rejections out of all rejections, i.e. the expected proportion of falsely detected SNPs out of all those reported as associated with the trait. We describe the FWER and FDR for generalization testing.

Define the left-sided (right-sided) alternative as the scenario in which a given SNP allele is negatively (positively) associated with the trait in a given study (either discovery or follow-up). Let

$$H_{ij} = \begin{cases} 1 & \text{if the right-sided alternative is true for SNP } j \text{ in population } i \\ 0 & \text{if the null hypothesis of no association is true for SNP } j \text{ in population } i \\ -1 & \text{if the left-sided alternative is true for SNP } j \text{ in population } i \end{cases}$$

Let $\mathcal{H}_j = \{\mathbf{h} = (h_{1j}, h_{2j}) : h_{ij} \in \{-1, 0, 1\}\}$ be the set of 9 possible configurations of the vector $\mathbf{H}_j = (H_{1j}, H_{2j})$ for two-sided alternatives for SNP j. The set of possible configurations is depicted in Figure 1. The generalization null hypothesis for SNP j is true if $\mathbf{H}_j$ belongs to the set $\mathcal{H}^0 = \{(-1,1), (-1,0), (1,-1), (1,0),\ (0,0), (0,-1), (0,1)\}$. A SNP for which the generalization null is false has $\mathbf{H}_j \in \mathcal{H}^A = \{(1,1), (-1,-1)\}$.

For a SNP j, denote by R_j^R and R_j^L the indicators of whether a generalization null rejection ("generalization claim") is made in the right or left direction, respectively. Suppose that R generalization claims are made by an analysis. The number of true generalization claims is

$$S = \sum_{\{j:\mathbf{H}_j=(1,1)\}} R_j^R + \sum_{\{j:\mathbf{H}_j=(-1,-1)\}} R_j^L,$$

and $R - S$ is the number of false generalization claims.

The directional generalization (and replication) FWER and FDR are given by:

$$\begin{aligned} \text{FWER}_g &= Pr\left(R - S > 0\right), \\ \text{FDR}_g &= E\left(\frac{R-S}{\max(R,1)}\right). \end{aligned}$$

Controlling for false generalizations

Heller et al. (2014) proposed r-values for testing associations in both the discovery and generalization (or replication) studies. Notably, Heller et al. (2014)'s procedure is not concerned with directional consistency. We now extend the procedure proposed by Heller et al. (2014) to the directional r-values framework and procedure for directional control in generalization testing. Following the definitions given below, the procedures are provided, and the proofs that these procedures control $\text{FDR}_g/\text{FWER}_g$ are relegated to the supplemental material.

Definition: *The directional $FDR_g/FWER_g$ r-value for a SNP is the lowest FDR/FWER level at which we can say that the SNP association is generalized with the same direction of association in both the discovery and generalizing studies.*

The directional p-values: Denote the left- and right-sided p-values for SNP association j in study $i \in \{1, 2\}$ by p_{ij}^L, p_{ij}^R respectively. For continuous test statistics, $p_{ij}^R = 1 - p_{ij}^L$. The p-values (p'_{1j}, p'_{2j}) corresponding to variant j used in generalization analysis are defined as:

$$p'_{1j} = \begin{cases} p_{1j}^L & \text{if } p_{1j}^L < p_{1j}^R \\ p_{1j}^R & \text{if } p_{1j}^L > p_{1j}^R \end{cases} \qquad p'_{2j} = \begin{cases} p_{2j}^L & \text{if } p_{1j}^L < p_{1j}^R \\ p_{2j}^R & \text{if } p_{1j}^L > p_{1j}^R. \end{cases}$$

Note that the one-sided p-values from both studies are guided by the estimated direction of association in the discovery study, so that if the association of SNP j was in the negative (positive) direction, then $p'_{1j} < 0.5$ is the left (right) sided hypothesis p-value in the discovery study, and p'_{2j} is the left (right) sided hypothesis p-value in the follow-up study. Therefore, if the evidence towards association is in the same direction in both studies, both $p'_{1j}, p'_{2j} < 0.5$, but if the estimated associations are in opposite directions in the two studies, then $p'_{2j} > 0.5$.

Data and parameters required for $\mathbf{FDR}_g$/$\mathbf{FWER}_g$ r-values computation:

1. m, the number of SNPs examined in the discovery study.

2. $\mathcal{R}_1$, the set of SNPs selected for follow-up based on discovery study results. Let $R_1 = |\mathcal{R}_1|$ be their number.

3. The directional p-values for the followed-up SNPs $\{(p'_{1j}, p'_{2j}) : j \in \mathcal{R}_1\}$.

4. $l_{00} \in [0, 1)$, the user-specified lower bound on the fraction of SNP associations, out of the m SNPs examined in the discovery study, that are null in both studies. Default value for a GWAS is $l_{00} = 0.8$, following Heller et al. (2014).

5. $c_2 \in (0, 1)$, the emphasis given to the follow-up study (see Section Variations in Heller et al. (2014)), default value is $c_2 = 0.5$.

Computation of the $\mathbf{FDR}_g$/$\mathbf{FWER}_g$ r-values

1. Defining functions $f_i^{FDR}(x)/f_i^{FWER}(x), i \in \mathcal{R}_1, x \in (0, 1)$:

(a) Compute $c_1(x) = \frac{1-c_2}{1-l_{00}(1-c_2 x)}$, the inverse weight function for the p-values from the discovery study.

(b) For every SNP $j \in \mathcal{R}_1$ compute the following e-values:

$$e_j(x) = \max\left(\frac{m}{c_1(x)} p'_{1j},\ \frac{R_1}{c_2} p'_{2j}\right), \quad j \in \mathcal{R}_1.$$

(C) [FDR] Let $f_i^{FDR}(x) = \min_{\{j: e_j(x) \geq e_i(x), j \in \mathcal{R}_1\}} \frac{e_j(x)}{rank[e_j(x)]}$, where $rank[e_j(x)]$ is the rank of the e-value for a SNP $j \in \mathcal{R}_1$ (with maximum rank for ties).

(C) [FWER] Let $f_j^{FWER}(x) = e_j(x)$.

2. The FDR_g (FWER_g) r-value for SNP $i \in \mathcal{R}_1$ is the solution to $f_i^{FDR}(r_i) = r_i$ ($f_i^{FWER}(r_i) = r_i$) if a solution exists in $(0, 1)$, and 1 otherwise. The solution is unique, see Lemma S1.1 in Heller et al. (2014).

The directional procedure for establishing generalization with FDR_g/FWER_g control at level q.

1. Compute the FDR_g/FWER_g r-values.

2. Declare as generalized all SNPs with FDR_g/FWER_g r-value at most q. Denote this set of SNPs by $\mathcal{R}_2$.

3. If a SNP $j \in \mathcal{R}_2$ has $p'_{1j} = p^L_{1j}$, then declare this SNP as having a generalized left-sided alternative; If SNP $j \in \mathcal{R}_2$ has $p'_{1j} = p^R_{1j}$, then declare the SNP as having a generalized right-sided alternative.

Selection rules

In generalization analysis, SNP associations are first tested in the discovery study, and then a smaller subset of these SNPs is selected for testing in the follow-up study, according to a selection rule. The most well known selection rule is the one that selects SNPs with p-value$< 5 \times 10^{-8}$ in the discovery study. This selection rule originates in single-stage designs, in which one has to report significant findings in the discovery study while controlling the FWER, assuming that $m = 10^6$. We consider other selection rules for a generalization/replication-based study design where the null hypothesis is the generalization null hypothesis.

1. Selection rule 1, recommended by Heller et al. (2014) for FDR_g control. Apply the FDR controlling BH procedure (Benjamini and Hochberg, 1995) on all p-values from the discovery study to obtain BH-adjusted p-values. Choose all SNPs with BH-adjusted p-value$\leq t$, where

$$\begin{aligned} t &= c_1(q) * q, \text{ with} \\ c_1(x) &= \frac{0.5}{1 - l_{00}(1 - 0.5x)}. \end{aligned} \tag{1}$$

   Use $q = 0.05$ to control FDR_g at the 0.05 level. The rationale behind selection rule 1 is that every SNP with BH-adjusted discovery p-value larger than t has no chance of generalizing. Heller et al. (2014) applied this selection rule in settings where either both discovery and replication used two-sided p-values, or both used one-sided p-values with pre-determined direction. We can also apply it on one-sided p-values used for generalization testing when the discovery study hypotheses were two-sided.

2. Selection rule 2, recommended by Heller et al. (2014) for FWER_g control. This rule selects all SNPs with discovery p-value$\leq t'$, where

$$t' = c_1(q) \times q/m. \tag{2}$$

As in selection rule 1, SNPs with p-value$> t'$ have no chance of generalizing using the FWER_g controlling procedure and selecting them may only reduce power. Again, we can apply this selection rule on the one-sided p-values used for generalization testing.

Note that selection rule 1 is data adaptive, and depends on the distribution of signals in the discovery study. Selection rule 2 is a fixed threshold rule. In selection rule 2, if $l_{00} = 0.8, q = 0.05$, and $m = 10^6$, we get $t' = 1.14 \times 10^{-7}$. When one-sided p-values are used for generalization testing, the original two-sided p-values passing this threshold are $\leq 2.28 \times 10^{-7}$.

Linkage Disequilibrium (LD)

GWAS datasets may potentially contain tens of millions of genotyped and imputed SNPs. Many of these SNPs are in linkage disequilibrium; that is, allelic variation within one SNP is correlated with allelic variation in another SNP. Often, when a discovery study association is detected, the locus may contain tens of correlated SNPs with low p-value. All of the significant SNPs may be tagging the same underlying causal genetic variant, or there may be multiple causal variants (referred to as "allelic heterogeneity"). Usually, the SNP with lowest p-value among the set of correlated SNPs is identified as the "lead", or "sentinel", SNP and is the one reported. There are two important issues in generalization testing of SNPs in the presence of LD:

1. When testing all associated SNPs (compared to a single SNP from any set of SNPs in LD), the multiple testing burden is larger than when testing only independent SNPs. In applying FWER-type control, this issue is sometimes handled by calculating the effective number of independent SNPs, and using this number in applying a Bonferroni correction. In contrast, this is not a problem for FDR control, which is concerned with fractions of false positives.

2. The pattern of LD usually differ between two different populations, and consequently different SNPs may best tag the underlying causal genetic variation.

Therefore, if appropriate information is available (e.g. the LD matrix of the SNPs in the generalization study), we recommend following-up on all SNPs passing the appropriate p-value threshold and using the effective number of tests instead of the actual number of selected SNPs in computing r-values.

Simulation studies: discovery and generalization GWAS

In this section, we examine the performance of our methods when the discovery study hypotheses tests are two-sided. In particular, we examine the gain in power from using one-sided p-values guided by the evidence in the discovery study, compared to applying the two-sided p-values in the procedure suggested by Heller et al. (2014). We also assess the impact of using different selection rules on generalization power.

In the simulation study described below, we directly simulated test statistics for two studies. This approach allowed us to conduct a large number of simulations, and study the effect of test statistics' inflation under the null (e.g. due to ancestry confounding or

low minor allele counts), where inflation means that the p-values distribution is left-skewed rather than uniform under the null. First, test statistics were calculated in the discovery study, and then SNPs were selected for generalization testing based on several selection rules. We used selection rules 1 (for FDR_g control) and 2 (for FWER_g control), applied on both one- and two-sided p-values. We also used the selection rules that take all SNPs with p-values$< 1 \times 10^{-6}, 1 \times 10^{-7}$, and $< 5 \times 10^{-8}$ in the discovery study.

In an additional simulation study, we investigated GWAS of cohorts designed to mimic realistic data sets with differences in LD structure and MAFs between the discovery and the generalization cohort (supplemental material) in a smaller number of simulations. There, we also compared generalization testing of all SNPs satisfying the selection rule with the testing of only the lead SNPs from each of the detected loci.

Simulating test statistics with null inflation

In each of 1,000 repetitions of the simulations, we sampled a million independent test statistics for both the discovery and the follow-up studies. Of these SNPs, 100 were causal in the discovery study and 100 were causal in the follow-up study. 50 of the causal SNPs overlapped between the studies. We considered two common generalization scenarios. In the first setting the discovery study had relatively low power, and the follow-up study has high power. For instance, this happens when discovery is performed in the HCHS/SOL, and follow-up is pursued in a large meta-analysis GWAS in individuals of Europeans ancestry. In the second setting the discovery study had high power, and the follow-up study had low power. This happens when investigators in the HCHS/SOL study the generalization

to Hispanics/Latinos of associations that were formerly reported in large meta-analyses of whites.

In both settings, the test statistics corresponding to causal SNPs in the discovery study 1 were sampled, in each simulation, as $z_{1,j} \sim \mathcal{N}(u_{1j}, 1)$ where u_{1j} is a realization of a random variable sampled from unif(u_l, u_h) distribution. When the discovery study had lower power we set $u_l = 4, u_h = 5$. Corresponding two-sided p-values had a median p-value of 7×10^{-6}, with an inter-quartile range of $[2 \times 10^{-7}, 1 \times 10^{-4}]$. When the discovery study had high power we set $u_l = 5, u_h = 6$. Corresponding two-sided p-values had a median p-value of 4×10^{-8}, with an inter-quartile range of $[5 \times 10^{-10}, 2 \times 10^{-6}]$. The 100 test statistics corresponding to the causal SNPs in study 2 were similarly sampled as $z_{2,j} \sim \mathcal{N}(u_{2j}, 1)$ where $u_{2j} \sim$ unif$(5, 6)$ when the follow-up study had high power, and $u_{2j} \sim$ unif$(3, 4)$ when the follow-up study had low power; the latter had corresponding two-sided p-values with a median p-value of 5×10^{-4} and an inter-quartile range of $[3 \times 10^{-5}, 5 \times 10^{-3}]$. Finally, we generated inflation via a simple procedure in which the test statistics corresponding to non-causal (null) SNPs, in both the discovery and the generalizing cohorts, were independently sampled from a Normal distribution with mean of zero and variance of 1.21, corresponding to $\lambda_{gc} = 1.21$ (Devlin and Roeder, 1999).

We studied additional simulation settings: a 90% overlap of the causal SNPs between the two populations, a larger number of causal SNPs (1,000 and 10,000), and a discovery study with $u_{1j} \sim$ unif$(4, 5)$, corresponding to two-sided p-values having a median p-value of 7×10^{-6}, with an inter-quartile range of $[2 \times 10^{-7}, 1 \times 10^{-4}]$, and follow-up study with $u_{2j} \sim$ unif$(3, 4)$. Overall, our simulations covered many plausible scenarios of the power

of the discovery and follow-up studies (high, medium, and low discovery power, and high and low follow-up study power), and reasonable assumptions on the overlap between the genetic component of the two populations, and on the number of SNPs associated with the trait. Finally, we also studied the effect of setting l_{00} to $0.9, 0.95$.

The HCHS/SOL

The HCHS/SOL (LaVange et al., 2010; Sorlie et al., 2010), is a community based cohort study, following self-identified Hispanic individuals from four field centers (Chicago, IL; Miami, FL; Bronx, NY; and San Diego, CA). Individuals were sampled via a two-stage sampling scheme, in which households were randomly sampled from sampled block groups. Almost 13,000 study participants consented for genotyping. This study was approved by the institutional review boards at each field center, where all subjects gave written informed consent.

Genotyping, imputation and quality control

Blood samples from HCHS/SOL individuals were genotyped on a custom array consisting of Illumina Omni 2.5M content plus ∼150,000 custom markers selected to include ancestry-informative markers, variants characteristic of Amerindian populations, known GWAS hits and other candidate gene polymorphisms. Quality control was similar to the procedure described in Laurie et al. (2010), and included checks for sample identity, batch effects, missing call rate, chromosomal anomalies (Laurie et al., 2012), deviation from Hardy-Weinberg equilibrium, Mendelian errors, and duplicate sample discordance. 12,803 samples passed

quality control, and 2,232,944 SNPs passed quality filters. Pairwise kinship coefficients and principal components reflecting ancestry were estimated in an iterative procedure which accounts for admixture (Conomos, 2014; Conomos et al., 2016). Genome-wide imputation was done using the 1000 Genomes Project phase 1 reference panel.1000 Genomes Project Consortium (2012) Genotypes were first pre-phased with SHAPEIT2 (Delaneau et al., 2013) (v2.r644) and then imputed with IMPUTE2 (Howie et al., 2009) (v2.3.0).

Identifying SNP-TC associations in the HCHS/SOL

To study our proposed methods for identifying SNP-trait associations, we performed a GWAS of TC in the HCHS/SOL followed by generalization testing using publicly available GWAS results. Our goal was to demonstrate that by using selection rules that are geared towards a two-stage study design, we can identify more generalized associations of independent loci, compared to selection rules that are based on a single-stage design.

The analysis was adjusted for sex, age, 5 principal components to control for confounding bias due to ancestry, and study design variables (e.g. study center, sampling weights). Analysis was performed using linear mixed effect models, with random effects corresponding to block groups, households, and kinship. As advocated by Kraft et al. (2009), the HCHS/SOL analyses were matched to the published analyses, so that the same trait transformation was used in both analyses. Thus, we first regressed TC values on covariates, and then applied a rank-based inverse normal transformation on the residuals. We then used the transformed residuals as the trait in the association testing.

We utilize the availability of complete results from the Global Lipids Genetics Con-

sortium (GLGC) TC GWAS (Willer et al., 2013), conducted in a large meta-analysis of multiple cohorts of European ancestry comprising of over 180,000 individuals, to compare multiple generalization analyses of TC. First, we considered generalization geared towards establishing new associations, in which we perform a GWAS in the HCHS/SOL as a discovery stage, with generalizations to the independent GLGC data set. Second, we consider generalization with the goal of testing whether previously established associations extend to the HCHS/SOL. For this, we selected SNPs published by Teslovich et al. (2010) and Willer et al. (2013) to generalize to the HCHS/SOL.

As mentioned previously, we expect multiple SNPs from each detected locus to be associated with TC, due to LD. Therefore, we define a locus as the region of size 1Mb around a SNP. We tested all SNPs satisfying the selection rule criterion, even if they were in LD with each other. However, we report generalization results both in terms of SNP associations, and by loci: after generalization testing, we identified the first locus by taking the SNP with smallest discovery p-value (lead discovery SNP) to represent it. We then "removed" all SNPs in its vicinity, and continued to select other SNPs in a similar manner. A locus with any SNP that generalized is declared a generalized locus.

Results

Simulations

Tables 1 and 2 provide the simulation results when the discovery power was low and the follow-up study had high power, when the goal was to control FDR_g and FWER_g,

respectively. Tables 3 and 4 provide similar results for the setting where the discovery study power was high and the follow-up study had low power. In all tables we omitted the results for the selection rules that selected SNPs for follow-up based on discovery two-sided p-value$\leq 10^{-7}$, as this resulted in "intermediate" results in terms of power between selection rules of higher and lower p-value thresholds, and is less beneficial than other selection rules.

For each selection rule, the characteristics of the selected SNP sets and generalization tests are provided, averaged across the iterations of simulations. The latter are provided in terms of estimated power, calculated as the average proportion of generalized SNPs, out of all generalizable SNPs in the simulation, false positives (FP) as the average number of generalizations of SNPs that are not in fact generalizable. In addition, when the selection rules and multiple testing adjustment methods were aimed at FDR_g control (Tables 1 and 3), we also provide false discovery proportion (FDP_g), which is the average proportion of false positives out of all generalized SNPs, and estimates FDR_g, and the standard deviation of the false discovery proportion across all the simulations, $\text{SD}(\text{FDP}_g)$. When the selection rules and multiple testing adjustment methods were aimed at FWER_g control (Tables 2 and 4), we provide the estimated FWER_g, as the proportion of simulations having at least one false positive generalization, i.e. the mean of $I_{[V>0]}$, the indicator function of having at least one false generalization, i.e. $V = R - S > 0$, and also $\text{SD}(I_{[V>0]})$. The standard errors of all measures are also provided.

As expected throughout, the higher the p-value threshold implied by the selection rule, the larger the number of selected SNPs, and the larger the number of true generalizable

SNPs selected. As expected by chance, 50% of the non-generalizable candidate SNPs have different direction of estimated effects in the two studies, so the one-sided p-values from the generalization study for these SNPs are higher than 0.5. Therefore, it is not surprising to see fewer false positive generalizations under directional control (using one-sided p-values). In both simulation settings and under both FDR_g and FWER_g control, directional control also had higher generalization power compared to using two-sided p-values, with less difference when the selection rule had very low p-values, or in other words, when fewer SNPs were under the null. In the settings in which the discovery study had high discovery power there was consequently higher generalization power, but also slightly higher error rates. Importantly, both FDR_g and FWER_g r-values always protected their target error measures.

FDR_g control: Focusing on directional FDR_g r-values, selection rule 1 applied with either one- or two-sided p-values was most powerful in the low discovery power setting, and selection rule 1 applied on two-sided p-values was most powerful in the high discovery power setting. Generalization testing using BH on the follow-up study alone did not control FDR_g when selection rule 1 was applied on one-sided p-values, and it also did not control FDR_g when applied on two-sided p-values. In lower p-value thresholds, when a high proportion of the tested SNPs were under the alternative, FDR_g was controlled when BH was used on the follow-up study alone. Since the r-values approach is slightly more stringent than the BH on the follow-up approach, FDR_g r-values are expected to be somewhat less powerful. The difference in power is small when the follow-up SNPs were highly significant. More specifically, the power is identical when the discovery power is low

and the selection rule is discovery p-value$\leq 10^{-6}$ or $\leq 5 \times 10^{-8}$, and the power differed by only 0.03 for the same selection rules, when the discovery power is high.

FWER$_g$ control: Selection rule 2 applied on one-sided p-values was the most powerful selection rule in both settings. Generalization testing using Bonferroni correction on the follow-up study alone never controlled FWER$_g$, though error rates were slightly improved by using one-sided p-values.

Finally, additional simulations (results unreported) revealed the same pattern of results, overall suggesting that selection rule 1 applied on two-sided p-values is the most powerful for FDR$_g$ control, and selection rule 2 applied on one-sided p-values is the most powerful for FWER$_g$ control. Setting l_{00} to higher values $\{0.9, 0.95\}$ had almost no effect on the results when selection rules with two-side discovery p-values$\leq 10^{-6}$ (or lower) were used, and had mixed effects on power when selection rule 1 was used (beneficial in the low discovery power setting, but less powerful in the high discovery power setting).

The HCHS/SOL TC GWAS

HCHS/SOL as the primary discovery study in a two-stage design

We performed a GWAS of TC in the HCHS/SOL, to establish generalized SNP-TC associations. We test using both FDR$_g$ and FWER$_g$ controlling r-values, and compare them in combination with different selection rules. In Table 5, for each combination of selection rule and multiple testing adjustment method, we report the number of SNPs followed-up that are available in both the HCHS/SOL and the GLGC TC GWAS, the number of loci they correspond to, the number of generalized SNPs and generalized loci, and the number

of loci with none of the SNPs having p-value$< 5 \times 10^{-8}$ in Willer et al. (2013)'s GWAS.

As expected, the number of SNPs selected for follow-up increased as the p-value threshold became higher; usually, the number of generalized loci increased as well. When the selection rule was all SNPs with p-value$< 5 \times 10^{-8}$, the followed-up SNPs corresponded to 15 loci, all of which generalized under FWER_g (and FDR_g) control. For FWER_g control, the number of generalized loci was the same, and maximal (17 loci), with selection rule 2 (on one-sided p-values) and p-value$< 10^{-6}$. This is consistent with the rationale behind selection rule 2, because SNPs with HCHS/SOL two-sided p-value$> 2.28 \times 10^{-7}$ cannot be generalized under FWER_g control.

In the FDR_g-controlling analysis applied on SNPs satisfying selection rule 1 on two-sided p-values, 21 loci generalized. These included a single generalized locus that would not be reported in either the HCHS/SOL or the GLGC GWAS alone. The lead SNP, rs870992 on chromosome 5, had r-value$= 0.008$, HCHS/SOL p-value$= 2 \times 10^{-5}$, and GLGC p-value$= 5.2 \times 10^{-5}$. This SNP was formerly associated with concentration of liver enzymes in plasma in a GWAS(Chambers et al., 2011). In the FDR_g-controlling analysis applied on SNPs satisfying selection rule 1 with one-sided p-values, there were 22 loci with strong evidence of association in the GLGC GWAS (SNPs with p-values$< 5 \times 10^{-8}$), and 5 loci generalized that would not been detected in the HCHS/SOL or GLGC GWAS alone. One of them was the locus that includes rs870992. Another SNP, rs2072781 in chromosome 6, had r-value$= 0.009$ (HCHS/SOL p-value$= 2.1 \times 10^{-5}$, GLGC p-value$= 1 \times 10^{-4}$). This SNP is in the MYLIP gene, formerly associated with high TC in Mexicans (Weissglas-Volkov et al., 2011). Three additional loci had relatively higher p-values in the GLGC GWAS

(0.007-0.05) and r-values in the range 0.01-0.05. The five loci are reported in Table S9 in the supplemental material.

In all analyses, there was a generalized locus in which the HCHS/SOL lead SNP did not generalize, as it had p-value= 0.92 in Willer et al. (2013), but a different SNP in the same locus had p-value= 1.4×10^{-46} in Willer et al. (2013) and generalized. This supports a strategy that analyzes all SNPs, rather than an LD-pruned set.

Generalizing previously reported TC-SNP associations

There are 74 SNPs previously reported as associated with TC with p-values$\leq 5 \times 10^{-8}$ and are available for generalization testing in the HCHS/SOL data set. Teslovich et al. (2010), in a meta-analysis of more than 100,000 individuals, reported 51 SNPs, that were later replicated in Willer et al.,Willer et al. (2013) which further meta-analyzed their association testing results with additional results from the GLGC study. Willer et al. (2013) reported an additional set of 23 SNPs (that were not meta-analyzed with Teslovich et al.'sTeslovich et al. (2010) results). Therefore, we performed two generalization analyses: one for the 51 SNPs that were replicated and had meta-analysis results combining the two studies, and one for the set of 23 SNPs reported only in Willer et al.Willer et al. (2013) In this analysis, 33 of the formerly replicated SNPs generalized to the HCHS/SOL, while none of the SNPs that were only reported in Willer et al. (2013) generalized. This is likely due to their low effect sizes.

In the supplemental material, we provide an additional analysis in which we follow-up for generalization testing all SNPs with p-value$< 10^{-6}$ in the GLGC GWAS, without any

SNP pruning. This analysis generalized 9 more loci than the analysis that tested only the published lead SNPs.

Discussion

In this work, we propose to leverage two-stage design to increase generalization power in GWAS. We show that by using a multiple testing adjustment framework tailored to the two-stage study design we can combine testing results from the discovery and follow-up studies to increase power with essentially no increase in the rate of false positive findings. We introduce procedures for calculating directional FDR_g and FWER_g r-values, computed based on one-sided p-values. We prove that r-values control their directional error measures when there is no genomic inflation, and show via simulation that errors are controlled in the presence of inflation. These procedures are, by construction, more powerful than those based on two-sided p-values when the direction of association is consistent between discovery and follow-up populations. We studied SNP selection rules that are geared towards generalization-based designs, and found in simulation studies that by choosing SNPs for generalization testing based on p-values less conservative than the genome-wide significance threshold, e.g. selection rules 1 and 2 for FDR_g control and FWER_g control, respectively, we are able to generalize more SNPs while controlling the desired error rate. Finally, we demonstrated our procedure on a GWAS of TC in the HCHS/SOL. First, we consider the scenario in which HCHS/SOL is a discovery study and generalization is required for reporting a significant finding. Second, we considered the scenario in which there are established SNP-trait associations that we want to generalize to the HCHS/SOL.

An approach that was promoted in the past to increase power in a two-stage design was to perform a joint analysis of the two studies via meta-analysis (Skol et al., 2006). However, this approach does not test the generalization null hypothesis, and an association may appear significant even if it exists only in one population. In contrast, our approach is focused in generalization testing; generalizations makes stronger statements on the underlying similarity in genetic associations between populations.

We provide practical recommendations based on our results. First, in terms of selection rules, we recommend selection rule 2 for FWER_g control at the α level, which selects SNPs with two-sided discovery p-value$< 2.28 \times 10^{-7}$ for $\alpha = 0.05$. For FDR_g control at the $\alpha = 0.05$ level, we recommend selecting SNPs with discovery p-value$\leq 10^{-6}$, or based on selection rule 1 if it is more conservative. If selection rule 1 applied to one-sided p-values is conservative, it is preferable to other selection rules since it limits the set of SNPs to those that can potentially be generalized. Second, we recommend follow-up on all SNPs satisfying the selection rule. Limiting follow-up to lead SNPs from the discovery study may reduce generalization power due to different LD patterns between the discovery and follow-up populations, in which the best tag SNP in one may not be the best tag SNP in the other. Finally, while FDR_g control allows for more false positive generalizations compared to FWER_g control, it also allows for more generalizations. This is well known, and the GWAS culture favors caution and prioritizes FWER control. In generalization, however, FDR_g may be more appropriate than FWER_g, since the investigator may be willing to tolerate a small fraction of false positives among the generalizations, as the overall number of reported false associations may already be dramatically reduced by generalization testing, compared

to reported associations from a discovery GWAS alone.

In this work we focus on generalization testing of associations from European ancestry populations to Hispanics/Latinos, and the other way around. Hispanics/Latinos are admixed and have large proportion of European ancestry; therefore we expect a large overlap in genetic architecture between the two populations. However, we do expect our conclusions to hold also when studying generalizations from Africans to Europeans, and other population as well. We performed additional simulations studies in similar scenarios reported in this manuscript but with varying degrees of overlap between causal SNPs and distributions of test statistics, corresponding to many plausible generalization scenarios. The conclusions remained the same.

While our methodology focuses on generalization of variants, in the data analysis we reported results by loci. The loci generalization framework still needs to be developed. Consider the null hypothesis of no generalization of a locus that states that none of the SNPs in the locus generalized. We here reported a locus as generalized if at least one of its associated SNPs generalized, but we did not offer a measure of locus-generalization evidence. Assigning a r-value for this null hypothesis is a topic of future work.

Supplemental Data

Supplemental Data include the description of an additional simulation study and its results in eight tables, a table of data analysis results, and mathematical derivations.

Software

An R package to perform generalization analysis can be installed using the R commands

```
library(devtools)
install_github("tamartsi/generalize", subdir = "generalize")
```

Also, a web applet that computes r-values based on one-sided p-values from the discovery and follow-up study, and does not require any software installation, is available in

`http://www.math.tau.ac.il/~ruheller/App.html`

Acknowledgements

The authors thank the staff and participants of HCHS/SOL for their important contributions. This work was supported in part by NHLBI HHSN268201300005C. The Hispanic Community Health Study/Study of Latinos was carried out as a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (N01-HC65233), University of Miami (N01-HC65234), Albert Einstein College of Medicine (N01-HC65235), Northwestern University (N01-HC65236), and San Diego State University (N01-HC65237). The following Institutes/Centers/Offices contribute to the HCHS/SOL through a transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke, NIH Institution-Office of Dietary Supplements. The Genetic Analysis Center at the University of Washington was supported by NHLBI and NIDCR

contracts (HHSN268201300005C AM03 and MOD03). The research of YB has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° [294519] (PSARPS).

References

1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491** 56–65.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 289–300.

Bogomolov, M. and Heller, R. (2013). Discovering findings that replicate from a primary study of high dimension to a follow-up study. *Journal of the American Statistical Association*, **108** 1480–1492.

Chambers, J. C., Zhang, W., Sehmi, J., Li, X., Wass, M. N., Van der Harst, P., Holm, H., Sanna, S., Kavousi, M., Baumeister, S. E., Coin, L. J., Deng, G., Gieger, C., Heard-Costa, N. L., Hottenga, J.-J., Kuhnel, B., Kumar, V., Lagou, V., Liang, L., Luan, J., Vidal, P. M., Mateo Leach, I., O'Reilly, P. F., Peden, J. F., Rahmioglu, N., Soininen, P., Speliotes, E. K., Yuan, X., Thorleifsson, G., Alizadeh, B. Z., Atwood, L. D., Borecki, I. B., Brown, M. J., Charoen, P., Cucca, F., Das, D., de Geus, E. J. C., Dixon, A. L., Doring, A., Ehret, G., Eyjolfsson, G. I., Farrall, M., Forouhi, N. G., Friedrich,

N., Goessling, W., Gudbjartsson, D. F., Harris, T. B., Hartikainen, A.-L., Heath, S., Hirschfield, G. M., Hofman, A., Homuth, G., Hypponen, E., Janssen, H. L. A., Johnson, T., Kangas, A. J., Kema, I. P., Kuhn, J. P., Lai, S., Lathrop, M., Lerch, M. M., Li, Y., Liang, T. J., Lin, J.-P., Loos, R. J. F., Martin, N. G., Moffatt, M. F., Montgomery, G. W., Munroe, P. B., Musunuru, K., Nakamura, Y., O'Donnell, C. J., Olafsson, I., Penninx, B. W., Pouta, A., Prins, B. P., Prokopenko, I., Puls, R., Ruokonen, A., Savolainen, M. J., Schlessinger, D., Schouten, J. N. L., Seedorf, U., Sen-Chowdhry, S., Siminovitch, K. A., Smit, J. H., Spector, T. D., Tan, W., Teslovich, T. M., Tukiainen, T., Uitterlinden, A. G., Van der Klauw, M. M., Vasan, R. S., Wallace, C., Wallaschofski, H., Wichmann, H.-E., Willemsen, G., Wurtz, P., Xu, C., Yerges-Armstrong, L. M., Abecasis, G. R., Ahmadi, K. R., Boomsma, D. I., Caulfield, M., Cookson, W. O., van Duijn, C. M., Froguel, P., Matsuda, K., McCarthy, M. I., Meisinger, C., Mooser, V., Pietilainen, K. H., Schumann, G., Snieder, H., Sternberg, M. J. E., Stolk, R. P., Thomas, H. C., Thorsteinsdottir, U., Uda, M., Waeber, G., Wareham, N. J., Waterworth, D. M., Watkins, H., Whitfield, J. B., Witteman, J. C. M., Wolffenbuttel, B. H. R., Fox, C. S., Ala-Korpela, M., Stefansson, K., Vollenweider, P., Volzke, H., Schadt, E. E., Scott, J., Jarvelin, M.-R., Elliott, P. and Kooner, J. S. (2011). Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nat Genet*, **43** 1131–1138.

Cohen, B. (1999). Freely associating. *Nat Genet*, **22** 1–2.

Conomos, M., Laurie, C., Stilp, A., Gogarten, S., McHugh, C., Nelson, S., Sofer, T., Fernandez-Rhodes, L., Justice, A., Graff, M., Young, K., Seyerle, A., Avery, C., Taylor, K., Rotter, J., Talavera, G., Daviglus, M., Wassertheil-Smoller, S., Schneiderman, N., Heiss, G., Kaplan, R., Franceschini, N., Reiner, A., Shaffer, J., Barr, R., Kerr, K., Browning, S., Browning, B., Weir, B., Avilés-Santa, M., Papanicolaou, G., Lumley, T., Szpiro, A., North, K., Rice, K., Thornton, T. and Laurie, C. (2016). Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos. *The American Journal of Human Genetics*, **98** 165 – 184.

Conomos, M. P. (2014). *Inferring, estimating and accounting for population and pedigree structure in genetic analyses*. Ph.D. thesis, University of Washington, Seattle.

Delaneau, O., Zagury, J.-F. and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods*, **10** 5–6.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, **55** 997–1004.

Heller, R., Bogomolov, M. and Benjamini, Y. (2014). Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proceedings of the National Academy of Sciences*, **111** 16262–16267.

Howie, B. N., Donnelly, P. and Marchini, J. (2009). A flexible and accurate genotype

imputation method for the next generation of genome-wide association studies. *PLoS Genet*, **5** e1000529.

Kraft, P., Zeggini, E. and Ioannidis, J. P. (2009). Replication in genome-wide association studies. *Statistical Science: A review journal of the Institute of Mathematical Statistics*, **24** 561.

Laurie, C. et al. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, **34** 591–602.

Laurie, C. C., Laurie, C. A. et al. (2012). Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nature Genetics*, **44** 642–650.

LaVange, L. M., Kalsbeek, W. D., Sorlie, P. D., Avilés-Santa, L. M., Kaplan, R. C., Barnhart, J., Liu, K., Giachello, A., Lee, D. J., Ryan, J. et al. (2010). Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. *Annals of epidemiology*, **20** 642–649.

Skol, A. D., Scott, L. J., Abecasis, G. R. and Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature genetics*, **38** 209–213.

Sorlie, P. D., Avilés-Santa, L. M., Wassertheil-Smoller, S., Kaplan, R. C., Daviglus, M. L., Giachello, A. L., Schneiderman, N., Raij, L., Talavera, G., Allison, M. et al. (2010). Design and implementation of the hispanic community health study/study of latinos. *Annals of epidemiology*, **20** 629–641.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J. et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466** 707–713.

Weissglas-Volkov, D., Calkin, A. C., Tusie-Luna, T., Sinsheimer, J. S., Zelcer, N., Riba, L., Tino, A. M. V., Ordoñez-Sánchez, M. L., Cruz-Bautista, I., Aguilar-Salinas, C. A. et al. (2011). The N342S MYLIP polymorphism is associated with high total cholesterol and increased LDL receptor degradation in humans. *The Journal of clinical investigation*, **121** 3062–3071.

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S. et al. (2013). Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, **45** 1274 – 1283.

Follow-up \ Discovery	Left	Null	Right
Left	$(-1,-1)$	$(0,-1)$	$(1,-1)$
Null	$(-1,0)$	$(0,0)$	$(1,0)$
Right	$(-1,1)$	$(0,1)$	$(1,1)$

Figure 1: The set of possible configuration of the vector $\mathbf{H}_j = (H_{1j}, H_{2j})$. The association of SNP j with the trait is defined as generalized association (marked as gray) when both alternatives are either left (negative direction of allele-trait association, $H_j = (-1,-1)$), or right (positive direction of allele-trait association, $H_j = (1,1)$).

Num	True Disc	True Gen	adjustment	power (SE)	FP (SE)	FDP_g (SE)	SD (FDP_g)
Selection rule 1 (one-sided), on average 1.4×10^{-4}							
600.72	74.39	37.15	BH (one-sided)	0.74 (0.00)	4.15 (0.07)	0.08 (0.00)	0.04
			BH (two-sided)	0.74 (0.00)	4.70 (0.08)	0.08 (0.00)	0.04
			FDR_g r-values (one-sided)	0.48 (0.00)	0.16 (0.01)	0.00 (0.00)	0.01
			FDR_g r-values (two-sided)	0.41 (0.00)	0.10 (0.01)	0.00 (0.00)	0.01
Selection rule 1 (two-sided), on average 1.7×10^{-5}							
151.03	57.80	28.84	BH (one-sided)	0.58 (0.00)	2.18 (0.05)	0.05 (0.00)	0.03
			BH (two-sided)	0.58 (0.00)	2.50 (0.05)	0.06 (0.00)	0.04
			FDR_g r-values (one-sided)	0.48 (0.00)	0.55 (0.02)	0.01 (0.00)	0.02
			FDR_g r-values (two-sided)	0.42 (0.00)	0.34 (0.02)	0.01 (0.00)	0.02
10^{-6}							
44.12	35.48	17.63	BH (one-sided)	0.35 (0.00)	0.91 (0.03)	0.03 (0.00)	0.03
			BH (two-sided)	0.35 (0.00)	0.98 (0.03)	0.03 (0.00)	0.03
			FDR_g r-values (one-sided)	0.35 (0.00)	0.50 (0.02)	0.02 (0.00)	0.02
			FDR_g r-values (two-sided)	0.35 (0.00)	0.54 (0.02)	0.02 (0.00)	0.03
5×10^{-8} (two-sided)							
18.87	18.13	9.09	BH (one-sided)	0.18 (0.00)	0.40 (0.02)	0.02 (0.00)	0.03
			BH (two-sided)	0.18 (0.00)	0.43 (0.02)	0.02 (0.00)	0.03
			FDR_g r-values (one-sided)	0.18 (0.00)	0.23 (0.01)	0.01 (0.00)	0.02
			FDR_g r-values (two-sided)	0.18 (0.00)	0.23 (0.02)	0.01 (0.00)	0.02

Table 1: Simulations characteristics and generalization testing results from 1,000 simulations in which the discovery study had low power and the follow-up study had high power, when the goal is to control FDR_g. Num is the average number of followed-up SNPs. "True Disc" and "True Gen" are the average number of true effect SNPs in the discovery, and in both the discovery and follow-up study, respectively, of those selected. Selection rules were applied on either two- or one-sided p-values. For testing, the compared methods are BH on the follow-up study alone, and FDR_g r-values. For both methods we compared standard analysis without directional control by using two-sided p-values, and directional control via one-sided p-values. Power is the average proportion of generalized SNPs out of the truly generalized SNPs, FP is the average number of falsely generalized SNPs, FDP_g is the average false discovery proportion, and SD (FDP_g) is the standard deviation of the FDP_g across simulations. Standard errors are in parentheses.

Num	True Disc	True Gen	adjustment	power (SE)	FP (SE)	FWER_g (SE)	SD ($I_{[V>0]}$)
			10^{-6}				
44.12	35.48	17.63	Bonferroni (one-sided)	0.35 (0.00)	0.08 (0.01)	0.07 (0.01)	0.07 (0.26)
			Bonferroni (two-sided)	0.35 (0.00)	0.09 (0.01)	0.08 (0.01)	0.08 (0.28)
			FWER_g r-values (one-sided)	0.25 (0.00)	0.03 (0.01)	0.03 (0.01)	0.03 (0.17)
			FWER_g r-values (two-sided)	0.21 (0.00)	0.02 (0.00)	0.02 (0.00)	0.02 (0.14)
			Selection rule 2 (ond-sided)				
28.38	25.86	12.88	Bonferroni (one-sided)	0.26 (0.00)	0.06 (0.01)	0.06 (0.01)	0.06 (0.24)
			Bonferroni (two-sided)	0.25 (0.00)	0.06 (0.01)	0.06 (0.01)	0.06 (0.24)
			FWER_g r-values (one-sided)	0.25 (0.00)	0.04 (0.01)	0.04 (0.01)	0.04 (0.18)
			FWER_g r-values (two-sided)	0.22 (0.00)	0.03 (0.01)	0.03 (0.01)	0.03 (0.18)
			Selection rule 2 (two-sided)				
23.51	22.06	11.01	Bonferroni (one-sided)	0.22 (0.00)	0.06 (0.01)	0.06 (0.01)	0.06 (0.24)
			Bonferroni (two-sided)	0.22 (0.00)	0.06 (0.01)	0.06 (0.01)	0.06 (0.24)
			FWER_g r-values (one-sided)	0.22 (0.00)	0.04 (0.01)	0.04 (0.01)	0.04 (0.19)
			FWER_g r-values (two-sided)	0.22 (0.00)	0.04 (0.01)	0.04 (0.01)	0.04 (0.19)
			5×10^{-8} (two-sided)				
18.87	18.13	9.09	Bonferroni (one-sided)	0.18 (0.00)	0.06 (0.01)	0.06 (0.01)	0.06 (0.24)
			Bonferroni (two-sided)	0.18 (0.00)	0.07 (0.01)	0.06 (0.01)	0.06 (0.24)
			FWER_g r-values (one-sided)	0.18 (0.00)	0.03 (0.01)	0.03 (0.01)	0.03 (0.18)
			FWER_g r-values (two-sided)	0.18 (0.00)	0.03 (0.01)	0.03 (0.01)	0.03 (0.18)

Table 2: Simulations characteristics and generalization testing results from 1,000 simulations in which the discovery study had low power and the follow-up study had high power, when the goal is to control FWER_g. Num is the average number of followed-up SNPs. "True Disc" and "True Gen" are the average number of true effect SNPs in the discovery, and in both the discovery and follow-up study, respectively, of those selected. Selection rules were applied on either two- or one-sided p-values. For testing, the compared methods are Bonferroni correction on the follow-up study alone, and FWER_g r-values. For both methods we compared standard analysis without directional control by using two-sided p-values, and directional control via one-sided p-values. Power is the average proportion of generalized SNPs out of the truly generalized SNPs, FP is the average number of falsely generalized SNPs, and FWER_g is the average number of simulations with any false positive generalization (the mean of $I_{[V>0]}$), where $I_{[V>0]}$ is the indicator of at least one false generalization across simulations. Standard errors are in parentheses.

Num	True Disc	True Gen	adjustment	power (SE)	FP (SE)	FDP_g (SE)	SD (FDP_g)
Selection rule 1 (one-sided), on average 1.6×10^{-4}							
673.96	94.92	47.41	BH (one-sided)	0.72 (0.00)	4.14 (0.07)	0.08 (0.00)	0.04
			BH (two-sided)	0.65 (0.00)	4.32 (0.07)	0.08 (0.00)	0.04
			FDR_g r-values (one-sided)	0.54 (0.00)	0.21 (0.01)	0.01 (0.00)	0.01
			FDR_g r-values (two-sided)	0.43 (0.00)	0.15 (0.01)	0.00 (0.00)	0.01
Selection rule 1 (two-sided), on average 2.4×10^{-5}							
212.73	88.92	44.43	BH (one-sided)	0.77 (0.00)	2.86 (0.05)	0.05 (0.00)	0.03
			BH (two-sided)	0.71 (0.00)	3.12 (0.06)	0.06 (0.00)	0.03
			FDR_g r-values (one-sided)	0.66 (0.00)	0.77 (0.03)	0.02 (0.00)	0.02
			FDR_g r-values (two-sided)	0.56 (0.00)	0.54 (0.02)	0.01 (0.00)	0.02
10^{-6}							
80.64	72.00	35.99	BH (one-sided)	0.66 (0.00)	1.49 (0.04)	0.03 (0.00)	0.02
			BH (two-sided)	0.63 (0.00)	1.59 (0.04)	0.04 (0.00)	0.03
			FDR_g r-values (one-sided)	0.63 (0.00)	0.80 (0.03)	0.02 (0.00)	0.02
			FDR_g r-values (two-sided)	0.58 (0.00)	0.84 (0.03)	0.02 (0.00)	0.02
5×10^{-8} (two-sided)							
52.76	52.02	25.92	BH (one-sided)	0.48 (0.00)	0.98 (0.03)	0.03 (0.00)	0.03
			BH (two-sided)	0.46 (0.00)	1.05 (0.03)	0.03 (0.00)	0.03
			FDR_g r-values (one-sided)	0.45 (0.00)	0.53 (0.02)	0.02 (0.00)	0.02
			FDR_g r-values (two-sided)	0.42 (0.00)	0.57 (0.02)	0.02 (0.00)	0.02

Table 3: Simulations characteristics and generalization testing results from 1,000 simulations in which the discovery study had high power and the follow-up study had low power, when the goal is to control FDR_g. Num is the average number of followed-up SNPs. "True Disc" and "True Gen" are the average number of true effect SNPs in the discovery, and in both the discovery and follow-up study, respectively, of those selected. Selection rules were applied on either two- or one-sided p-values. For testing, the compared methods are BH on the follow-up study alone, and FDR_g r-values. For both methods we compared standard analysis without directional control by using two-sided p-values, and directional control via one-sided p-values. Power is the average proportion of generalized SNPs out of the truly generalized SNPs, FP is the average number of falsely generalized SNPs, FDP_g is the average false discovery proportion, and SD (FDP_g) is the standard deviation of the FDP_g across simulations. Standard errors are in parentheses.

Num	True Disc	True Gen	adjustment	power (SE)	FP (SE)	FWER_g (SE)	SD ($I_{[V>0]}$)
			10^{-6}				
80.64	72.00	35.99	Bonferroni (one-sided)	0.43 (0.00)	0.09 (0.01)	0.08 (0.01)	0.08 (0.28)
			Bonferroni (two-sided)	0.38 (0.00)	0.09 (0.01)	0.09 (0.01)	0.09 (0.28)
			FWER_g r-values (one-sided)	0.33 (0.00)	0.04 (0.01)	0.04 (0.01)	0.04 (0.19)
			FWER_g r-values (two-sided)	0.26 (0.00)	0.04 (0.01)	0.04 (0.01)	0.04 (0.19)
			Selection rule 2 (ond-sided)				
64.92	62.39	31.13	Bonferroni (one-sided)	0.39 (0.00)	0.07 (0.01)	0.07 (0.01)	0.07 (0.26)
			Bonferroni (two-sided)	0.34 (0.00)	0.09 (0.01)	0.08 (0.01)	0.08 (0.28)
			FWER_g r-values (one-sided)	0.34 (0.00)	0.05 (0.01)	0.05 (0.01)	0.05 (0.21)
			FWER_g r-values (two-sided)	0.27 (0.00)	0.04 (0.01)	0.04 (0.01)	0.04 (0.20)
			Selection rule 2 (two-sided)				
59.10	57.66	28.76	Bonferroni (one-sided)	0.37 (0.00)	0.07 (0.01)	0.07 (0.01)	0.07 (0.26)
			Bonferroni (two-sided)	0.32 (0.00)	0.08 (0.01)	0.08 (0.01)	0.08 (0.28)
			FWER_g r-values (one-sided)	0.32 (0.00)	0.05 (0.01)	0.05 (0.01)	0.05 (0.22)
			FWER_g r-values (two-sided)	0.28 (0.00)	0.04 (0.01)	0.04 (0.01)	0.04 (0.20)
			5×10^{-8} (two-sided)				
52.76	52.02	25.92	Bonferroni (one-sided)	0.33 (0.00)	0.08 (0.01)	0.07 (0.01)	0.07 (0.26)
			Bonferroni (two-sided)	0.3 (0.00)	0.08 (0.01)	0.08 (0.01)	0.08 (0.27)
			FWER_g r-values (one-sided)	0.3 (0.00)	0.05 (0.01)	0.05 (0.01)	0.05 (0.22)
			FWER_g r-values (two-sided)	0.26 (0.00)	0.04 (0.01)	0.04 (0.01)	0.04 (0.20)

Table 4: Simulations characteristics and generalization testing results from 1,000 simulations in which the discovery study had high power and the follow-up study had low power, when the goal is to control FWER_g. Num is the average number of followed-up SNPs. "True Disc" and "True Gen" are the average number of true effect SNPs in the discovery, and in both the discovery and follow-up study, respectively, of those selected. Selection rules were applied on either two- or one-sided p-values. For testing, the compared methods are Bonferroni correction on the follow-up study alone, and FWER_g r-values. For both methods we compared standard analysis without directional control by using two-sided p-values, and directional control via one-sided p-values. Power is the average proportion of generalized SNPs out of the truly generalized SNPs, FP is the average number of falsely generalized SNPs, FWER_g is the average number of simulations with any false positive generalization (the mean of $I_{[V>0]}$), where $I_{[V>0]}$ is the indicator of at least one false generalization across simulations. Standard errors are in parentheses.

Selection rule	selected SNPs	loci	adjustment	gen SNPs	gen loci	# loci not sig Willer
Selection rule 1 - one-sided	1,662	89	FDR$_g$ r-values	1,352	27	5
Selection rule 1 - two-sided	1,208	51	FDR$_g$ r-values	1,076	21	1
10^{-6}	742	18	FDR$_g$ r-values	706	17	0
10^{-6}	742	18	FWER$_g$ r-values	583	17	0
Selection rule 2 - one-sided	627	17	FWER$_g$ r-values	583	17	0
Selection rule 2 - two-sided	574	16	FWER$_g$ r-values	538	16	0
5×10^{-8}	546	15	FWER$_g$ r-values	514	15	0

Table 5: Generalization testing results from a set of analyses based on a HCHS/SOL GWAS as the discovery study, and GLGC GWAS as the follow-up study. For each selection rule we report the number of SNPs selected for follow-up testing, and the number of loci containing these SNPs. For combinations of selection rules and multiple testing adjustment method we report the number of generalized loci, and the number of generalized loci that did not contain any SNP with p-value$< 5 \times 10^{-8}$ in the GLGC GWAS.

Supplementary Material: A Powerful Statistical Framework for Generalization Testing in GWAS, with Application to the HCHS/SOL

Tamar Sofer, Ruth Heller, Marina Bogomolov, Christy L. Avery, Mariaelisa Graff, Kari E. North, Alex P. Reiner, Timothy A. Thornton, Kenneth Rice, Yoav Benjamini, Cathy C. Laurie, and Kathleen F. Kerr

Contents

1 Additional simulation study: simulating diverse cohorts

1.1 Simulation set-up

Using Hapgen2 (Su et al., 2011), we simulated two populations, one of 20,000 Europeans, derived from the CEU Hapmap (Gibbs et al., 2003) sample, that represented the discovery cohort, and one of 10,000 Mexicans derived from the MEX Hapmap sample that represented the generalizing cohort. The smaller MEX population size reflects the fact that often, cohorts of diverse ethnicities are smaller than those of Europeans. For each population, we simulated 90 causal SNPs affecting a quantitative outcome, of which 45 overlapped, in 5 different simulation scenarios. The 5 simulation scenarios differed only by the list of causal SNPs, to allow for potential differences in generalization power due to difference in LD structure. The MAFs of the causal SNPs in the CEU ranged between 0.04 to 0.49, and were different in the MEX for the same SNPs, since they were the Hapmap MAFs for these populations. The outcome model was $y_{pi} = \boldsymbol{g}_{pi}^T \boldsymbol{\beta}_p + \epsilon_{pi}$, with $\boldsymbol{g}_{pi}$ being the vector of 90 allelic counts of individual i in population p, corresponding to the causal SNPs in this population. $\boldsymbol{\beta}_p$ was the vector of SNP effects of population p, and $\epsilon_{pi} \sim \mathcal{N}(0, 1)$ was the residual error. The median simulated β_j in CEU was 0.07, and the largest effect sizes were 0.20 and 0.25. Of the 45 simulated causal SNPs that overlapped between populations, 12 had the same effect size in CEU and MEX so that $\beta_{CEU,k} = \beta_{MEX,k}$ for $k = 1, \ldots, 12$, and 33 had effect sizes in MEX sampled from a uniform distribution around the CEU effect, so that $\beta_{MEX,k} \sim \text{unif}(0.2 \times \beta_{CEU,k}, 1.8 \times \beta_{CEU,k})$.

From each of the 5 simulation settings we generated 20 simulations, to a total of 100

simulations of GWAS in two cohorts. In each simulation, we tested about 800,000 SNPs were tested for association with the simulated outcome. According to the GWAS results in the discovery population (either CEU or MEX), we performed a look-up of results in the follow-up population (either MEX or CEU). For the two combinations of discovery and follow-up populations, we report two sets of results. In the first analysis, SNPs that were followed up were pruned, so that no two SNPs closer than 1M base pairs to each other were followed-up (i.e. we follow-up for generalization testing only lead SNPs). We determined if the SNPs was a "true signal" if the correlation (due to LD) between the detected SNP and any simulated causal SNP was higher than 0.5. In the second set of results, we follow-up all SNPs satisfying the selection rules and tested all. We then determined how many loci generalized by defining loci as regions of 1M SNPs (here we did not use LD information, to reduce computations).

1.2 Results - generalization testing of CEU results in MEX

To study the instance in which the first stage of the study performs a GWAS in a large study of European individuals, and the follow-up study is a smaller study of Hispanic/Latino individuals, we provide generalization testing results for the case were the GWAS in the CEU is treated as the discovery study, and the GWAS in the MEX population as the follow-up. Results are given in Tables S1-S4. To summarize the conclusions from these simulations, FDR_g and FWER_g r-values provide better control agains false positive generalization claims compared to procedures that limit the FWER/FDR on the follow-up study alone. With FDR control, it is more powerful to follow all SNPs satisfying the selection rule

compared to pruning SNPs, especially when applying the more lenient selection rules. The difference in power diminishes as the selection rule becomes more stringent. However, the number of false positives also increases somewhat when SNPs are not pruned. For FWER control, it is more powerful to follow only lead SNPs. With any method of error control, and with and without pruning of SNPs, it was beneficial to follow-up on a larger set of SNPs than that dictated by the genome-wide significance level. In particular, selection rules 1 and 2 are powerful.

Similar simulations were performed with a smaller population in the follow-up study of 6,000 MEX individuals. The conclusions remained the same, only the generalization power decreased.

1.3 Results - generalization testing of MEX results in CEU

To study the instance in which the first stage of the study performs a GWAS in a relatively small study of Hispanic/Latino individuals (or other diverse, non-European population), and the follow-up study is a larger study, we provide generalization testing results for the case were the GWAS in the MEX is treated as the discovery study, and the GWAS in the CEU population as the follow-up. Results are given in Tables S5-S8. To summarize the conclusions from these simulations, FDR_g and FWER_g r-values provide better control agains false positive generalization claims compared to procedures that limit the FWER/FDR on the follow-up study alone. Not pruning SNPs is slightly more powerful (in terms of power) than pruning SNPs when applying FDR_g control, but this difference is essentially non-existent in when FWER_g is controlled. With any method of error control,

and with and without pruning of SNPs, it was beneficial to follow-up on a larger set of SNPs than that dictated by the genome-wide significance level. In particular, selection rules 1 and 2 are powerful.

Compare to generalizing results from CEU to MEX, here we have lower power, as expected, since less discoveries are made in the first study. In addition, it is striking that when implementing FDR_g control and following-up on all SNPs satisfying the selection rule, with no further pruning, the number of false positives is much larger when generalizing from CEU to MEX, than the other way around. This may also be due to the higher power of the CEU GWAS.

Adjustment	Loci	True gen loci	Gen loci	FP	Power
1×10^{-6}					
Bonferroni (one sided)	61.78	22.75	22.49	0.80	0.48
Bonferroni (two sided)	61.78	22.75	21.34	0.74	0.46
FWER$_g$ r-values (one sided)	61.78	22.75	20.40	0.70	0.44
FWER$_g$ r-values (two sided)	61.78	22.75	19.09	0.62	0.41
Selection rule 2 - one sided					
Bonferroni (one sided)	56.55	21.91	21.62	0.77	0.46
Bonferroni (two sided)	56.55	21.91	20.55	0.72	0.44
FWER$_g$ r-values (one sided)	56.55	21.91	20.54	0.71	0.44
FWER$_g$ r-values (two sided)	56.55	21.91	19.16	0.62	0.41
Selection rule 2 - two sided					
Bonferroni (one sided)	53.50	21.35	21.02	0.71	0.45
Bonferroni (two sided)	53.50	21.35	20.09	0.67	0.43
FWER$_g$ r-values (one sided)	53.50	21.35	20.08	0.66	0.43
FWER$_g$ r-values (two sided)	53.50	21.35	19.21	0.62	0.41
5×10^{-8}					
Bonferroni (one sided)	49.66	20.19	19.95	0.68	0.43
Bonferroni (two sided)	49.66	20.19	19.16	0.65	0.41
FWER$_g$ r-values (one sided)	49.66	20.19	19.15	0.64	0.41
FWER$_g$ r-values (two sided)	49.66	20.19	18.37	0.61	0.39

Table S1: Averaged generalization testing results of CEU associations in MEX, given by loci, when SNPs passing the selection rule are pruned by distance to the lead SNP. The controlled error measure was FWER$_g$. We compare the Bonferroni adjustment on the follow-up study alone with FWER$_g$ r-values, both with and without directional control implemented with one-sided p-values.

Adjustment	SNPs	Loci	True gen loci	Gen loci	FP	Power
			1×10^{-6}			
Bonferroni (one sided)	801.63	61.78	31.02	20.23	0.66	0.43
Bonferroni (two sided)	801.63	61.78	31.02	19.33	0.63	0.42
FWER$_g$ r-values (one sided)	801.63	61.78	31.02	18.42	0.57	0.40
FWER$_g$ r-values(two sided)	801.63	61.78	31.02	17.25	0.50	0.37
			Selection rule 2 - one sided			
Bonferroni (one sided)	681.07	56.55	29.00	19.47	0.65	0.42
Bonferroni (two sided)	681.07	56.55	29.00	18.65	0.60	0.40
FWER$_g$ r-values (one sided)	681.07	56.55	29.00	18.64	0.59	0.40
FWER$_g$ r-values (two sided)	681.07	56.55	29.00	17.44	0.51	0.38
			Selection rule 2 - two sided			
Bonferroni (one sided)	624.80	53.50	27.72	19.09	0.61	0.41
Bonferroni(two sided)	624.80	53.50	27.72	18.28	0.55	0.39
FWER$_g$ r-values (one sided)	624.80	53.50	27.72	18.27	0.54	0.39
FWER$_g$ r-values (two sided)	624.80	53.50	27.72	17.56	0.53	0.38
			5×10^{-8}			
Bonferroni (one sided)	554.81	49.66	25.90	18.27	0.60	0.39
Bonferroni (two sided)	554.81	49.66	25.90	17.51	0.54	0.38
FWER$_g$ r-values (one sided)	554.81	49.66	25.90	17.50	0.53	0.38
FWER$_g$ r-values (two sided)	554.81	49.66	25.90	16.87	0.51	0.36

Table S2: Averaged generalization testing results of CEU associations in MEX in simulations, given by loci, when all SNPs passing the selection rule are followed-up and the controlled error measured was FWER$_g$. We compare the Bonferroni adjustment on the follow-up study alone with FWER$_g$ r-values, both with and without directional control implemented with one-sided p-values.

Adjustment	Loci	True gen loci	Gen loci	FP	Power
Selection rule 1 - one sided					
BH (one sided)	271.41	23.08	32.07	2.58	0.66
BH (two sided)	271.41	23.08	30.59	2.58	0.62
FDR_g r-values (one sided)	271.41	23.08	26.17	1.00	0.56
FDR_g r-values (two sided)	271.41	23.08	24.26	0.90	0.52
Selection rule 1 - two sided					
BH (one sided)	168.09	23.72	32.62	2.46	0.67
BH (two sided)	168.09	23.72	30.94	2.33	0.64
FDR_g r-values (one sided)	168.09	23.72	27.09	1.11	0.58
FDR_g r-values (two sided)	168.09	23.72	25.15	0.96	0.54
1×10^{-6}					
BH (one sided)	61.78	22.75	28.11	1.68	0.59
BH (two sided)	61.78	22.75	27.05	1.64	0.56
FDR_g r-values (one sided)	61.78	22.75	20.40	0.70	0.44
FDR_g r-values (two sided)	61.78	22.75	25.59	1.26	0.54
5×10^{-8}					
BH (one sided)	49.66	20.19	23.95	1.37	0.50
BH (two sided)	49.66	20.19	23.27	1.38	0.49
FDR_g r-values (one sided)	49.66	20.19	22.99	1.10	0.49
FDR_g r-values (two sided)	49.66	20.19	22.20	1.10	0.47

Table S3: Averaged generalization testing results of CEU associations in MEX in simulations, given by loci, when SNPs passing the 8 selection rule are pruned by distance to the lead SNP. The controlled error measure was FDR_g. We compare the BH adjustment on the follow-up study alone with FDR_g r-values, both with and without directional control implemented with one-sided p-values.

Adjustment	SNPs	Loci	True gen loci	Gen loci	FP	Power
Selection rule 1 - one sided						
BH (one sided)	3014.62	271.41	40.99	42.87	8.93	0.75
BH (two sided)	3014.62	271.41	40.99	41.37	8.92	0.72
FDR_g r-values (one sided)	3014.62	271.41	40.99	32.67	2.30	0.67
FDR_g r-values (two sided)	3014.62	271.41	40.99	30.13	1.96	0.63
Selection rule 1 - two sided						
BH (one sided)	2123.12	168.09	39.93	40.13	6.57	0.75
BH (two sided)	2123.12	168.09	39.93	38.73	6.63	0.71
FDR_g r-values (one sided)	2123.12	168.09	39.93	33.98	2.93	0.69
FDR_g r-values (two sided)	2123.12	168.09	39.93	31.37	2.42	0.64
1×10^{-6}						
BH (one sided)	801.63	61.78	31.02	30.32	3.04	0.61
BH (two sided)	801.63	61.78	31.02	29.79	3.28	0.59
FDR_g r-values (one sided)	801.63	61.78	31.02	28.58	2.15	0.59
FDR_g r-values(two sided)	801.63	61.78	31.02	27.58	2.15	0.57
5×10^{-8}						
BH (one sided)	554.81	49.66	25.90	25.45	2.31	0.51
BH (two sided)	554.81	49.66	25.90	25.17	2.58	0.50
FDR_g r-values (one sided)	554.81	49.66	25.90	24.26	1.73	0.50
FDR_g r-values (two sided)	554.81	49.66	25.90	23.52	1.73	0.48

Table S4: Averaged generalization testing results of MEX associations in CEU in simulations, given by loci, when all SNPs passing the selection rule are followed-up and the controlled error measured was FDR_g. We compare the BH adjustment on the follow-up study alone with FDR_g r-values, both with and without directional control implemented with one-sided p-values.

Adjustment	Loci	True gen loci	Gen loci	FP	Power
1×10^{-6}					
Bonferroni (one sided)	20.99	19.44	16.88	0.69	0.36
Bonferroni (two sided)	20.99	19.44	16.80	0.69	0.36
FWER$_g$ r-values (one sided)	20.99	19.44	15.38	0.61	0.33
FWER$_g$ r-values (two sided)	20.99	19.44	14.64	0.54	0.31
Selection rule 2 - one sided					
Bonferroni (one sided)	18.66	17.64	15.44	0.61	0.33
Bonferroni (two sided)	18.66	17.64	15.38	0.61	0.33
FWER$_g$ r-values (one sided)	18.66	17.64	15.38	0.61	0.33
FWER$_g$ r-values (two sided)	18.66	17.64	14.65	0.54	0.31
Selection rule 2 - two sided					
Bonferroni (one sided)	17.68	16.91	14.71	0.54	0.31
Bonferroni (two sided)	17.68	16.91	14.68	0.54	0.31
FWER$_g$ r-values (one sided)	17.68	16.91	14.68	0.54	0.31
FWER$_g$ r-values (two sided)	17.68	16.91	14.66	0.54	0.31
5×10^{-8}					
Bonferroni (one sided)	16.51	15.88	13.81	0.46	0.30
Bonferroni (two sided)	16.51	15.88	13.78	0.46	0.30
FWER$_g$ r-values (one sided)	16.51	15.88	13.78	0.46	0.30
FWER$_g$ r-values (two sided)	16.51	15.88	13.76	0.46	0.30

Table S5: Averaged generalization testing results of MEX associations in CEU in simulations, given by loci, when SNPs passing the selection rule are pruned by distance into the lead SNPs only. We compare the Bonferroni adjustment on the follow-up study alone with FWER$_g$ r-values, both with and without directional control implemented with one-sided p-values.

Adjustment	SNP	Loci	True gen loci	Gen loci	FP	Power
			1×10^{-6}			
Bonferroni (one sided)	287.36	20.99	16.75	16.74	0.51	0.36
Bonferroni (two sided)	287.36	20.99	16.75	16.66	0.51	0.36
FWER$_g$ r-values (one sided)	287.36	20.99	16.75	15.28	0.46	0.33
FWER$_g$ r-values (two sided)	287.36	20.99	16.75	14.47	0.44	0.31
			Selection rule 2 - one sided			
Bonferroni (one sided)	243.00	18.66	15.22	15.33	0.46	0.33
Bonferroni (two sided)	243.00	18.66	15.22	15.30	0.46	0.33
FWER$_g$ r-values (one sided)	243.00	18.66	15.22	15.30	0.46	0.33
FWER$_g$ r-values (two sided)	243.00	18.66	15.22	14.52	0.45	0.31
			Selection rule 2 - two sided			
Bonferroni (one sided)	224.44	17.68	14.47	14.62	0.45	0.31
Bonferroni (two sided)	224.44	17.68	14.47	14.59	0.45	0.31
FWER$_g$ r-values (one sided)	224.44	17.68	14.47	14.59	0.45	0.31
FWER$_g$ r-values (two sided)	224.44	17.68	14.47	14.53	0.45	0.31
			5×10^{-8}			
Bonferroni (one sided)	204.14	16.51	13.59	13.74	0.40	0.30
Bonferroni (two sided)	204.14	16.51	13.59	13.72	0.40	0.30
FWER$_g$ r-values (one sided)	204.14	16.51	13.59	13.72	0.40	0.30
FWER$_g$ r-values (two sided)	204.14	16.51	13.59	13.66	0.40	0.29

Table S6: Averaged generalization testing results of MEX associations in CEU in simulations, given by loci, when all SNPs passing the selection rule are followed-up and the controlled error measure is FWER$_g$. We compare the Bonferroni adjustment on the follow-up study alone with FWER$_g$ r-values, both with and without directional control implemented with one-sided p-values.

Adjustment	Loci	True gen loci	Gen loci	FP	Power
Selection rule 1 - one sided					
BH (one sided)	111.95	28.30	26.28	2.94	0.52
BH (two sided)	111.95	28.30	26.01	2.72	0.52
FDR_g r-values (one sided)	111.95	28.30	18.78	0.82	0.40
FDR_g r-values (two sided)	111.95	28.30	17.88	0.72	0.38
Selection rule 1 - two sided					
BH (one sided)	62.56	26.56	24.38	2.37	0.49
BH (two sided)	62.56	26.56	24.16	2.20	0.49
FDR_g r-values (one sided)	62.56	26.56	18.85	0.86	0.40
FDR_g r-values (two sided)	62.56	26.56	17.94	0.72	0.38
1×10^{-6}					
BH (one sided)	20.99	19.44	17.18	0.79	0.36
BH (two sided)	20.99	19.44	17.10	0.76	0.36
FDR_g r-values (one sided)	20.99	19.44	17.06	0.75	0.36
FDR_g r-values (two sided)	20.99	19.44	16.98	0.71	0.36
5×10^{-8}					
BH (one sided)	16.51	15.88	13.94	0.48	0.30
BH (two sided)	16.51	15.88	13.90	0.47	0.30
FDR_g r-values (one sided)	16.51	15.88	13.88	0.47	0.30
FDR_g r-values (two sided)	16.51	15.88	13.86	0.46	0.30

Table S7: Averaged generalization testing results of MEX associations in CEU in simulations, given by loci, when SNPs passing the selection rule are pruned by distance to the lead SNP. The controlled error measure was FDR_g We compare the BH adjustment on the follow-up study alone with FDR_g r-values, both with and without directional control implemented with one-sided p-values.

Adjustment	SNP	Loci	True gen loci	Gen loci	FP	Power
Selection rule 1 - one sided						
BH (one sided)	869.93	111.95	28.84	30.14	4.89	0.56
BH (two sided)	869.93	111.95	28.84	29.93	4.80	0.56
FDR_g r-values (one sided)	869.93	111.95	28.84	24.51	1.58	0.51
FDR_g r-values (two sided)	869.93	111.95	28.84	22.47	1.03	0.48
Selection rule 1 - two sided						
BH (one sided)	625.74	62.56	25.12	25.99	2.80	0.52
BH (two sided)	625.74	62.56	25.12	25.71	2.60	0.51
FDR_g r-values (one sided)	625.74	62.56	25.12	24.69	1.73	0.51
FDR_g r-values (two sided)	625.74	62.56	25.12	22.63	1.13	0.48
1×10^{-6}						
BH (one sided)	287.36	20.99	16.75	16.74	0.51	0.36
BH (two sided)	287.36	20.99	16.75	17.39	0.88	0.37
FDR_g r-values (one sided)	287.36	20.99	16.75	17.29	0.78	0.37
FDR_g r-values (two sided)	287.36	20.99	16.75	17.13	0.66	0.37
5×10^{-8}						
BH (one sided)	204.14	16.51	13.59	14.19	0.74	0.30
BH (two sided)	204.14	16.51	13.59	14.11	0.67	0.30
FDR_g r-values (one sided)	204.14	16.51	13.59	14.01	0.57	0.30
FDR_g r-values (two sided)	204.14	16.51	13.59	13.93	0.49	0.30

Table S8: Averaged generalization testing results of MEX associations in CEU in simulations, given by loci, when all SNPs passing the selection rule are followed-up and the controlled error measure is FDR_g. We compare the BH adjustment on the follow-up study alone with FDR_g r-values, both with and without directional control implemented with one-sided p-values.

2 Additional data analysis results

2.1 SNPs that generalized in the FDR$_g$ directional r-values TC analysis but were not discovered in HCHS/SOL or GLGC GWAS alone

rsID	Chr	Position	effect allele	other allele	SOL MAF	SOL Beta	SOL SE	SOL Pval	Willer Beta	Willer SE	Willer Pval	FDR_g rval
rs2286779	10	118394551	G	C	0.47	-0.05	0.01	3×10^{-5}	-0.02	0.01	8×10^{-3}	0.014
rs703225	13	33042802	T	C	0.30	0.06	0.01	5×10^{-5}	0.02	0.01	2×10^{-2}	0.022
rs870992	5	52193237	A	G	0.10	-0.09	0.02	2×10^{-5}	-0.03	0.01	5×10^{-5}	0.008
rs12514413	5	107323866	T	C	0.17	-0.07	0.02	4×10^{-5}	-0.01	0.01	3×10^{-2}	0.042
rs2072781	6	16147349	T	C	0.09	0.10	0.02	2×10^{-5}	0.04	0.01	1×10^{-4}	0.009

Table S9: Generalized SNP associations, that were not previously known, in the TC generalization analysis based on selection rule 1 applied on 1-sided p-values of the HCHS/SOL discovery study. Generalization testing was based on directional FDRg r-values. The generalizing study was the GLGC GWAS.

2.2 Generalization of total cholesterol SNPs discovered in Europeans - without SNP pruning

In this analysis we tested all SNPs with p-value$< 10^{-6}$ in the GLGC GWAS. There were 2.4 million genotyped SNPs with association testing results in Willer et al. (2013), and 5,399 SNPs had p-value$< 10^{-6}$ and were available in the HCHS/SOL. Of these SNPs 2,418 of the SNPs generalized, which includes the 33 SNPs that were generalized in Analysis A. In addition, another one of the SNPs reported in Willer et al. (2013) generalized. Other generalized SNPs were not specifically reported in the papers. However, we defined loci as 1MB regions around the known loci, and found that all SNPs that generalized in Analysis B where located at loci around reported SNPs. In particular, there were 9 loci in which the reported SNP did not generalize, but other SNPs did. These generalizations did not occur in the analysis reported in the main manuscript, in whic these SNPs were not tested.

3 Mathematical derivations

Definition. *A stable selection rule satisfies the following condition: for any $j \in \mathcal{R}_1$, changing p^L_{1j} so that j is still selected while all other discovery study p-values are held fixed, will not change the set $\mathcal{R}_1$.*

Stable selection rules include selecting the hypotheses with two-sided discovery p-values below a certain cut-off, or by a non-adaptive multiple testing procedure on the discovery study two-sided p-values such as the BH procedure for FDR control or the Bonferroni procedure for FWER control, or selecting the k hypotheses with the smallest two-sided

p-values, where k is fixed in advance.

Theorem 1 *Let f_{00} be the true fraction of the m SNPs investigated in the discovery study that are null in both studies. The level q directional procedure based on FDR_g r-values in Section 2.1.5 in the manuscript controls the directional FDR_g at level at most q if the following conditions are satisfied: the rule by which the set $\mathcal{R}_1$ is selected is a stable selection rule; $l_{00} \leq f_{00}$; the p-values within the follow-up study are jointly independent or are positive regression dependent on the subset of p-values corresponding to true null hypotheses (property PRDS); for SNPs with $\mathbf{H}_j \notin \{(1,1),(-1,-1)\}$ the follow-up study p-values are independent of the discovery study p-values; and in addition one of items 1-3 below is satisfied.*

1. *The p-values within the discovery study are independent.*

2. *Arbitrary dependence among the p-values within the discovery study, when in the computation of the FDR_g r-values (section 2.1.4 in the main manuscript) m is replaced by $m^* = m\sum_{i=1}^{m} 1/i$.*

3. *Arbitrary dependence among the p-values within the discovery study, and the selection rule is such that the discovery study p-values of the SNPs that are selected for follow-up are at most a fixed threshold $t \in (0,1)$, when c_1 computed in Step 3(a) is replaced by*

$$\tilde{c}_1(x) = \max\{a : a(1 + \sum_{i=1}^{\lceil tm/(ax)-1 \rceil} 1/i) = c_1(x)\}.$$

Steps 3(b) and 3(c) remain unchanged. In step 4, the FDR r-value for feature $i \in \mathcal{R}_1$ is $r_i = \min\{x : f_i(x) \leq x\}$ if a solution exists in $(0,1)$, and one otherwise.

The implication of item 3 is that for generalization controlling FDR_g at level q, if $t \leq c_1(q)q/m$, no modification is required, so the procedure that declares as generalized all SNPs with r-values at most q controls the FDR_g at level q any type of dependency in the discovery study. Note that the modification in item 3 will lead to more generalization than the modification in item 2 only if $t < \frac{c_1(q)q}{1+\sum_{i=1}^{m-1} 1/i}$.

From simulation study 2, even if the discovery study p-values are not independent, the conservative modifications of the r-value computation in items 2-3 are unnecessary for FDR_g control in GWAS.

Theorem 2 *The level q directional procedure based on $FWER_g$ r-values controls the directional $FWER_g$ at level q if $l_{00} \leq f_{00}$, and if for SNPs with $\mathbf{H}_j \notin \{(1,1),(-1,-1)\}$ the follow-up study p-values are independent of the discovery study p-values.*

3.1 Proof of Theorem 1

We first show that the following procedure is identical to that of declaring the set of SNPs with FDR r-values at most q as generalized. First, compute the number of generalization claims at level q as follows:

$$R_2 \triangleq \max\left\{ r : \sum_{j \in \mathcal{R}_1} \mathbf{I}\left[(p'_{1j}, p'_{2j}) \leq \left(\frac{r}{m} c_1(q) q, \frac{r}{R_1} c_2 q \right) \right] = r \right\}.$$

Next, declare as generalized SNPs the set

$$\mathcal{R}_2 = \left\{ j : (p'_{1j}, p'_{2j}) \leq \left(\frac{R_2}{m} c_1(q) q, \frac{R_2}{R_1} c_2 q \right), j \in \mathcal{R}_1 \right\}.$$

It was shown in Lemma S1.1 in Heller et al. (2014), without directional control, that this procedure is identical to declaring the set of SNPs with FDR r-values at most q as

generalized. It is straightforward to see that the proof of Lemma S1.1 in Heller et al. (2014) remains unchanged when the p-values are replaced by (p'_{1j}, p'_{2j}), therefore the above procedure is identical to that of declaring the set of SNPs with FDR_g r-values at most q as generalized.

We will now prove that under the conditions of items 1-3 of Theorem 1 the directional procedure based on FDR_g r-values controls FDR_g at a level which is smaller or equal to

$$\begin{aligned} &c_1(q)c_2q^2(|j : \mathbf{H}_j \in \{(-1,0),(1,0),(0,0)\}|)/m + \\ &c_1(q)q|j : \mathbf{H}_j \in \{(0,1),(0,-1),(-1,-1),(1,1),(-1,1),(1,-1)\}|/m + \\ &c_2qE[|\mathcal{R}_1 \cap \{j : \mathbf{H}_j \in \{(-1,0),(1,0),(-1,1),(1,-1),(0,1),(0,-1),(0,0)\}|/|\mathcal{R}_1|], \end{aligned} \tag{1}$$

where the cardinalities are over the sets containing all m SNPs, i.e. $j = 1, \ldots, m$. Note that this expression is at most q if $l_{00} \leq f_{00}$. To see this, note that

$$|j : \mathbf{H}_j \in \{(-1,0),(1,0),(0,0)\}|/m = f_{\cdot 0},$$

and

$$|j : \mathbf{H}_j \in \{(0,1),(0,-1),(-1,-1),(1,1),(-1,1),(1,-1)\}|/m = 1 - f_{\cdot 0}.$$

Moreover,

$$E[|\mathcal{R}_1 \cap \{j : \mathbf{H}_j \in \{(-1,0),(1,0),(-1,1),(1,-1),(0,1),(0,-1),(0,0)\}|/|\mathcal{R}_1|] \leq 1.$$

Therefore, expression (1) is at most

$$
\begin{aligned}
& c_1(q)c_2q^2f_{\cdot 0} + c_1(q)q(1-f_{\cdot 0}) + c_2q \\
& = c_1(q)q - f_{\cdot 0}c_1(q)q(1-c_2q) + c_2q \\
& \leq c_1(q)q - l_{00}c_1(q)q(1-c_2q) + c_2q \\
& = c_1(q)q[1 - l_{00}(1-c_2q)] + c_2q \\
& = (1-c_2)q + c_2q = q.
\end{aligned}
$$

We will now prove that the expression in (1) is an upper bound for FDR_g, which is

$$
\begin{aligned}
& E\left(\frac{R-S}{\max(R,1)}\right) = \\
& \sum_{\{j:\mathbf{H}_j\in\{(0,-1),(0,1),(0,0),(1,0),(-1,0),(1,-1),(-1,1)\}\}} E\left(\frac{R_j^L + R_j^R}{\max(R,1)}\right) + \\
& \sum_{\{j:\mathbf{H}_j=(1,1)\}} E\left(\frac{R_j^L}{\max(R,1)}\right) + \sum_{\{j:\mathbf{H}_j=(-1,-1)\}} E\left(\frac{R_j^R}{\max(R,1)}\right). \qquad (2)
\end{aligned}
$$

For each $j \in \{1,\ldots,m\}$, we define $C_r^{(j)}$ as the event in which if j is declared generalized, r hypotheses are declared generalized including j, which amounts to the definition given in the proof of Theorem 1 in Supplementary Material of Heller et al. (2014), where the one-sided p-values (p_{1j}, p_{2j}) are replaced by (p'_{1j}, p'_{2j}). Note that for any given realization of $|\mathcal{R}_1|$ and value of r such that $r > |\mathcal{R}_1|$, $C_r^{(j)} = \emptyset$.

From the equivalent procedure above we get the following equality,

$$
\begin{aligned}
E\left(\frac{R_j^L}{\max(R,1)}\right) &= \sum_{r=1}^{m}\frac{1}{r}\Pr\left(j\in\mathcal{R}_1, P_{1j}^L \leq \min\left(\frac{rc_1(q)q}{m}, 0.5\right), P_{2j}^L \leq \frac{rc_2q}{\max(|\mathcal{R}_1|,1)}, C_r^{(j)}\right) \\
&\leq \sum_{r=1}^{m}\frac{1}{r}\Pr\left(P_{1j}^L \leq \frac{rc_1(q)q}{m}, P_{2j}^L \leq c_2q, C_r^{(j)}\right), \qquad (3)
\end{aligned}
$$

where the equality follows from the fact that a generalization claim is made in the left direction only if $P_{1j}^L \leq P_{1j}^R$, i.e. only if $P_{1j}^L < 0.5$. Similarly,

$$E\left(\frac{R_j^R}{\max(R,1)}\right) \leq \sum_{r=1}^{m} \frac{1}{r} \Pr\left(P_{1j}^R \leq \frac{rc_1(q)q}{m}, P_{2j}^R \leq c_2 q, C_r^{(j)}\right). \tag{4}$$

Using inequalities (3) and (4), and the facts that P_{1j}^L and P_{1j}^R are uniform for $j \in \{j : H_{1j} = 0\}$ and are stochastically larger than uniform for $j \in \{j : H_{1j} = 1\}$ and $j \in \{j : H_{1j} = -1\}$ respectively, we obtain the following inequalities:

$$E\left(\frac{R_j^L}{\max(R,1)}\right) \leq \begin{cases} c_1(q)q/m & \text{if } \mathbf{H}_j \in \{(0,-1),(1,-1),(1,1)\}, \\ c_2 q E[I(j \in \mathcal{R}_1)/|\mathcal{R}_1|] & \text{if } \mathbf{H}_j \in \{(-1,0),(0,1),(-1,1)\}, \\ c_1(q)c_2 q^2/m & \text{if } \mathbf{H}_j \in \{(0,0),(1,0)\}, \end{cases}$$

$$E\left(\frac{R_j^R}{\max(R,1)}\right) \leq \begin{cases} c_1(q)q/m & \text{if } \mathbf{H}_j \in \{(0,1),(-1,1),(-1,-1)\}, \\ c_2 q E[I(j \in \mathcal{R}_1)/|\mathcal{R}_1|] & \text{if } \mathbf{H}_j \in \{(1,0),(0,-1),(1,-1),(0,0)\}, \\ c_1(q)c_2 q^2/m & \text{if } \mathbf{H}_j \in \{(-1,0)\}. \end{cases}$$

These upper bounds for items 1-3 of Theorem 1 follow from similar derivations to these given in the proof of items (i)-(iii) of Theorem 1 in Heller et al. (2014), respectively. Specifically, for each of the items, the upper bounds $c_1(q)q/m$, $c_2 q E[I(j \in \mathcal{R}_1)/|\mathcal{R}_1|]$ and $c_1(q)c_2 q^2/m$ are derived similarly to inequalities [S3], [S4], and [S5] in the proof of Theorem 1 in Heller et al. (2014), respectively. Thus we obtain

$$E\left(\frac{R_j^R + R_j^L}{\max(R,1)}\right) \leq \begin{cases} c_2 q E[I(j \in \mathcal{R}_1)/|\mathcal{R}_1|] + c_1(q)c_2 q^2/m & \text{if } \mathbf{H}_j = (0,0), \\ c_2 q E[I(j \in \mathcal{R}_1)/|\mathcal{R}_1|] + c_1(q)c_2 q^2/m & \text{if } \mathbf{H}_j \in \{(1,0),(-1,0)\}, \\ c_1(q)q/m + c_2 q E[I(j \in \mathcal{R}_1)/|\mathcal{R}_1|] & \text{if } \mathbf{H}_j \in \{(0,1),(0,-1)\}, \\ c_2 q E[I(j \in \mathcal{R}_1)/|\mathcal{R}_1|] + c_1(q)q/m & \text{if } \mathbf{H}_j \in \{(1,-1),(-1,1)\}, \end{cases}$$

and for the directional error terms:

$$E\left(\frac{R_j^L}{\max(R,1)}\right) \leq \frac{c_1(q)q}{m}, \quad \text{for } j \text{ with } \quad \mathbf{H}_j = (1,1)$$
$$E\left(\frac{R_j^R}{\max(R,1)}\right) \leq \frac{c_1(q)q}{m}, \quad \text{for } j \text{ with } \quad \mathbf{H}_j = (-1,-1).$$

The result follows from using expression (2) for FDR_g, and summing up over the above upper bounds.

3.2 Proof of Theorem 2

It is easy to show that the procedure in Section 2.1.5 of the main manuscript is unchanged if we replace Step 2 by the following: all SNPs with $f_j^{FWER}(q) \leq q$ are declared generalized. The equivalence follows from the facts that $f_j^{FWER}(x)$ is a continuous function of x and $f_j^{FWER}(x)/x$ is strictly monotone decreasing (this result follows from the proof of Lemma S1.1 in the SI of Heller et al. (2014) and it is straightforward to show that it continues to hold in the directional generalization analysis).

We will now prove that the expression in (1) with q replaced by α is an upper bound for the directional FWER_g, which is $\Pr(R-S>0)$. It was shown in the proof of Theorem 1 that this expression is at most α if $l_{00} \leq f_{00}$. Note that

$$\Pr(R-S>0) \leq E(R-S) \leq \sum_{\{j:\mathbf{H}_j=(1,1)\}} E(R_j^L) + \sum_{\{j:\mathbf{H}_j=(-1,-1)\}} E(R_j^R)$$
$$+ \sum_{\{j:\mathbf{H}_j\in\{(0,-1),(0,1),(0,0),(1,0),(-1,0),(1,-1),(-1,1)\}\}} E(R_j^R + R_j^L).$$

We consider the procedure that replaces Step 2 by declaring SNPs with $f_j^{FWER}(\alpha) \leq \alpha$ as generalized (as discussed above). The directional error terms (declaring that a SNP

association is generalized in one direction, when in fact the association is in the other direction) in the first two sums above are bounded by:

$$E\left(R_j^L\right) \leq \frac{c_1(\alpha)\alpha}{m}, \quad \text{for } j \text{ with } \quad \mathbf{H}_j = (1,1)$$
$$E\left(R_j^R\right) \leq \frac{c_1(\alpha)\alpha}{m}, \quad \text{for } j \text{ with } \quad \mathbf{H}_j = (-1,-1)$$

These bounds hold since (without loss of generality), for j with $\mathbf{H}_j = (1,1)$

$$E\left(R_j^L\right) \leq \Pr(P_{1j}^L \leq \min(c_1(\alpha)\alpha/m, 0.5), P_{2j}^L \leq c_2\alpha/R_1)$$
$$\leq \Pr(P_{1j}^L \leq c_1(\alpha)\alpha/m) \leq c_1\alpha/m,$$

where the first inequality follows from the fact that a generalization claim is made in the left direction only if $P_{1j}^L \leq P_{1j}^R$, i.e., only if $P_{1j}^L < 0.5$, and the last inequality follows that the fact that for $H_{1j} = 1$, P_{1j}^L is stochastically larger than uniform.

All remaining errors are false generalization claims that are not directional errors. Clearly,

$$E(R_j^R + R_j^L) = \Pr(\min(P_{1j}^L, P_{1j}^R) \leq c_1(\alpha)\alpha/m, P'_{2j} \leq c_2\alpha/|\mathcal{R}_1|, j \in \mathcal{R}_1).$$

It is simple to show (using similar derivations to these in the proof of Theorem S6.1 in the SI of Heller et al. (2014)) that the right hand side is at most the following upper bounds:

$$E(R_j^R + R_j^L) \leq \begin{cases} c_2\alpha E[I(j \in \mathcal{R}_1)/|\mathcal{R}_1|] + c_1(\alpha)\alpha/m \times c_2\alpha & \text{if } \mathbf{H}_j = (0,0), \\ c_2\alpha E[I(j \in \mathcal{R}_1)/|\mathcal{R}_1|] + c_1(\alpha)\alpha/m \times c_2\alpha & \text{if } \mathbf{H}_j \in \{(1,0),(-1,0)\}, \\ c_1(\alpha)\alpha/m + c_2\alpha E[I(j \in \mathcal{R}_1)/|\mathcal{R}_1|] & \text{if } \mathbf{H}_j \in \{(0,1),(0,-1)\}, \\ c_2\alpha E[I(j \in \mathcal{R}_1)/|\mathcal{R}_1|] + c_1(\alpha)\alpha/m & \text{if } \mathbf{H}_j \in \{(1,-1),(-1,1)\}. \end{cases}$$

The result follows from summing over these upper bounds.

References

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y. et al. (2003). The international HapMap project. *Nature*, **426** 789–796.

Heller, R., Bogomolov, M. and Benjamini, Y. (2014). Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proceedings of the National Academy of Sciences*, **111** 16262–16267.

Su, Z., Marchini, J. and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, **27** 2304–2305.

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S. et al. (2013). Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, **45** 1274 – 1283.