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 information about the simulations described in the manuscript

1.1 Generating test statistics

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 in both the discovery and the generalizing cohorts by sampling the

non-causal test statistics from a Normal distribution with mean of zero and variance of 1.21, corresponding to $\lambda_{gc} = 1.21$ (Devlin and Roeder, 1999). Inflation may exist due to ancestry confounding or low minor allele counts.

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 $u_{1j} \sim \text{unif}(4, 5)$, 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 $u_{2j} \sim \text{unif}(3, 4)$. 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 $0.9, 0.95$.

1.2 Additional results

Tables S1 and S2 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 S3 and S4 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 S1 and S3), 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 S2 and S4), 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.

1.3 Conclusions from unreported simulations

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 $u_{1j} \sim \text{unif}(4, 5)$, 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 $u_{2j} \sim \text{unif}(3, 4)$. 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 $0.9, 0.95$.

These simulations 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).

2 Additional simulation study: simulating diverse cohorts

2.1 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).

2.2 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 S5-S8. To summarize the conclusions from these simulations, FDR_g and FWER_g r-values provide better control agains false positive gener-

alization 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.

2.3 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 S9-S12. 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.

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 S1: 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 S2: 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 S3: 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 S4: 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.

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 S5: 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 S6: 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 S7: Averaged generalization testing results of CEU associations in MEX 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	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 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 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 S9: 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 S10: 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 S11: 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 S12: 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.

3 Additional data analysis results

3.1 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 S13: 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.

3.2 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.

4 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.*

4.1 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,

$$E\left(\frac{R_j^R}{\max(R,1)}\right) \leq \sum_{r=1}^{m} \frac{1}{r} \Pr\left(P_{1j}^R \leq \frac{r c_1(q) q}{m}, P_{2j}^R \leq c_2 q, C_r^{(j)}\right). \tag{4}$$

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.

4.2 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

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

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.