simGWAS: a fast method for simulation of large scale case-control GWAS summary statistics

Mary D. Fortune[1,2] (mdf34@cam.ac.uk) Chris Wallace[1,2] (cew54@cam.ac.uk)

May 2, 2018

1. MRC Biostatistics Unit, University of Cambridge, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge, CB2 0SR, United Kingdom

2. Department of Medicine, University of Cambridge, Addenbrooke's Hospital, Hills Rd, Cambridge CB2 0SP

Funding MF and CW are funded by the Wellcome Trust (WT099772, WT107881) and CW by the MRC (MC_UU_00002/4). MF is currently funded by Dementia Platforms UK.

Abstract

Methods for analysis of GWAS summary statistics have encouraged data sharing and democratised the analysis of different diseases. Ideal validation for such methods is application to simulated data, where some "truth" is known. As GWAS increase in size, so does the computational complexity of such evaluations; standard practice repeatedly simulates and analyses genotype data for all individuals in an example study. We have developed a novel method based on an alternative approach, directly simulating GWAS summary data, without individual data as an intermediate step. We mathematically derive the expected statistics for any set of causal variants and their effect sizes, conditional upon control haplotype frequencies (available from public reference datasets). Simulation of GWAS summary output can be conducted independently of sample size by simulating random variates about these expected values. Across a range of scenarios, our method, available as an open source R package, produces very similar output to that from simulating individual genotypes with a substantial gain in speed even for modest sample sizes. Fast simulation of GWAS summary statistics will enable more complete and rapid evaluation of summary statistic methods as well as opening new potential avenues of research in fine mapping and gene set enrichment analysis.

Keywords Genome-wide association studies (GWAS); simulation; case-control studies

Background

The genome wide association study design is now more than a decade old (Visscher et al., 2017), and the size of GWAS cohorts has continued to grow, from 1000s to, now, 100,000s of individuals. Given the competing demands of open science and privacy concerns (P3G Consortium et al., 2009), it has become standard to share data in the form of summary statistics (allelic effect sizes and standard errors, or simply p values) more readily than the full genotype data. A wealth of methods have been developed to operate directly on the summary statistics, from for fine mapping of genetic causal variants (e.g. PAINTOR (Kichaev et al., 2014), CAVIARBF (Chen et al., 2015) and JAM (Newcombe et al., 2016)) to (co-)heritability estimation (Bulik-Sullivan et al., 2015) and integration of GWAS results from different traits (Giambartolomei et al., 2014; Zhu et al., 2016). Such methods are often derived through approximating a multivariate linear regression likelihood by incorporating information about correlation structures (linkage disequilibrium, LD) from reference populations. While GWAS of binary traits (eg case-control data) typically adopt a logistic regression approach to correctly model the increase in risk on the log odds scale, these summary statistic methods, often originally derived for linear regression models, have been applied to both quantitative and binary data.

As Biobank-sized datasets come to fruition, such summary statistic methods are likely to become even more important, since, for such large numbers of samples, operating on the complete genotype data matrices for efforts such as Bayesian fine mapping of causal variants is computationally prohibitive. Indeed, GWAS summary statistics for multiple traits from UK-Biobank have already been made freely available (Canela-Xandri et al., 2017). Although for speed these are generated by linear regression on 0/1 coded outcome variables, methods exist to also convert these to log odds ratios in the case of binary data (Lloyd-Jones et al., 2018). While Biobanks tend to adopt a cohort design, meta-GWAS studies continue to over-sample cases compared to controls, in order to increase the available power, and are now exceeding 100,000 cases and controls in single studies (Michailidou et al., 2017).

The gold standard for evaluating performance of summary statistic methods is through analysis of simulated data, allowing inference to be compared to a known "truth". GWAS simulators typically proceed by adding phenotypes to a sample of genotype data that is either simulated or from a reference population ("forward simulation") or by simulating or sampling genotype data conditional on a supposed distribution of phenotypes, typically used for case-control data structures ("conditional simulation"). Forward approaches in particular can be used very flexibly for generating multiple (quantitative) phenotypes, a design also common to Biobank datasets (Meyer and Birney, 2018). However, conditional approaches are generally required for case-control data, because typical GWAS designs recruit cases disproportionately to their frequency in the population in order to increase power. Simulation options in this case are more limited because the problem is mathematically harder. For single causal variant scenarios, resampling from a reference population conditional on allele frequencies at a target variant may be used. For more complicated causal models, involving multiple variants potentially in LD, GWAsimulator (Li and Li, 2008), TriadSim (Shi et al., 2018) or HAPGEN (Su et al., 2011) can very efficiently simulate haplotypes for cases and controls in small genomic regions. In particular, by incorporating mutations and recombinations, HAPGEN can simulate large populations with only a few hundred reference haplotypes. However, the generation of GWAS summary statistics, eg using SNPTEST (Marchini et al., 2007), requires analysis of the individual level data which can be slow, particularly for logistic models which require iterative optimization at each SNP.

The general approach of simulating both genotype and phenotype on an individual level cannot scale well for Biobank-scale or large meta-GWAS situations, because of the number of individuals required. It is also potentially wasteful - the individual level data are not required when the goal is to evaluate methods that work on summary statistics.

Here, we present an alternative approach, which simulates summary statistics directly, without needing to ever generate genotype data. It scales as a function of the number of SNPs, but is constant with regards to the number of samples, thus making it ideally designed for simulation of summary statistics for large case-control studies.

Results

Overview of our approach

We first introduce the mathematical calculations which underpin our method. Given a causal model specifying which SNPs are causal and their effects on disease in the form of odds ratios, and reference data on allele and haplotype frequencies in controls, we calculate the expected Z score under an additive model at each SNP from a Cochran Armitage score test. (Cochran Armitage score tests have been used for GWAS because of their computational simplicity, requiring no iterative maximisation procedure, and because they allow for additive, dominant or recessive coding, although additive coding is the most commonly used (Sasieni, 1997)).

Simulated Z scores can then be derived by multivariate normal simulation using standard software. This suffices in the case where the summary statistic methods to be used work upon Z scores alone. However, when log odds ratios and their standard errors are required, we appeal to the asymptotic similarity of score tests and Wald tests, and simulate standard errors under the causal model. Together with simulated Z scores, we can then back-calculate the log odds ratios as the product of simulated Z scores and standard errors. An outline description of our calculations follows; full details are given in the Appendix.

Let $Y_i \in \{0,1\}$ denote the indicator of disease status for the ith of N individuals sampled according to case-control status (N_1 cases, $Y_i = 1$; N_0 controls, $Y_i = 0$). Let n be the total number of SNPs. For any SNP X, write G_i^X for its genotype coding $\in \{0,1,2\}$ at sample i. Then, for the commonly used Cochran-Armitage score test, the Z-Score at SNP X is computed as:

$$Z_X = \frac{\sum_{i=1}^{N}((G_i^X - \overline{G^X})(Y_i - \bar{Y}))}{\sqrt{\frac{N_0 N_1}{N(N-1)} \sum_{i=1}^{N}(G_i^X - \overline{G^X})^2}} = \frac{U_X}{\sqrt{(N-1)V_X V_Y}}$$

where V_X, V_Y denote var(X), var(Y), respectively.

Write $\mathbf{W} = (W_1, ..., W_m)^T$ for the vector of causal SNPs and $\gamma = (\gamma_1, ..., \gamma_m)^T$ for their log odds ratios of effect. We assume that Y_i given $G_i^{\mathbf{W}}$ can be modelled as a binomial logistic regression:

$$\mathbb{P}_{sam}(Y_i = 1 | G_i^{\mathbf{W}} = \mathbf{w}) = \frac{e^{\gamma_0 + \gamma_1 w_1 + ... + \gamma_m w_m}}{1 + e^{\gamma_0 + \gamma_1 w_1 + ... + \gamma_m w_m}}$$

where $\mathbb{P}_{sam}()$ denotes that this is the probability within the GWAS sample and γ_0 is chosen such that $\mathbb{P}_{sam}(Y_i = 1) = \frac{N_1}{N}$. The conditioning is required because allele frequencies vary between cases and controls at causal variants and those in LD with them, meaning the overall allele frequencies in our sample differ from those in the population as a whole. By specifically distinguishing between $\mathbb{P}_{sam}()$ and the more general $\mathbb{P}()$, we can condition on having chosen N_0 cases and N_1 controls and thus perform the conditional simulation needed for case-control studies.

By conditioning upon the values of $G^{\mathbf{W}}$ and Y, we obtain the expected value of U_X, the covariance between G^X and Y:

$$\begin{aligned}\mathbb{E}(U_X) \quad = \quad & \frac{(N-1)N_0 N_1}{N^2} \sum_{\mathbf{w} \in \mathbb{Z}_3^m} \left[\left(\frac{N_0}{N_1} e^{\gamma_0 + \gamma_1 w_1 + ... + \gamma_m w_m} - 1 \right) \right] \times \\ & \left[2\mathbb{P}(G_i^X = 2 \cap G_i^{\mathbf{W}} = \mathbf{w}) + \mathbb{P}(G_i^X = 1 \cap G_i^{\mathbf{W}} = \mathbf{w}) \right] \qquad (1)\end{aligned}$$

The variance of U_X is $V_X V_Y$ where $V_Y = \frac{N_0 N_1}{N(N-1)}$ and V_X is the variance of G^X. As V_X is a variance, a natural model is an inverse gamma distribution, $V_X \sim \Gamma^{-1}(\alpha, \beta)$. By similar conditioning upon $G^{\mathbf{W}}$ and Y, we show that the parameters of this distribution are

$$\alpha = \frac{2\mathbb{E}(V_X^2) - (\mathbb{E}(V_X))^2}{\mathbb{E}(V_X^2) - (\mathbb{E}(V_X))^2} \tag{2}$$

$$\beta = \frac{\mathbb{E}(V_X)\mathbb{E}(V_X^2)}{\mathbb{E}(V_X^2) - (\mathbb{E}(V_X))^2} \tag{3}$$

(the derivation of this and expressions for the first two moments of V_X are given in the Appendix). This means we can either simulate V_X from its distribution or calculate

$$\mathbb{E}\left(\frac{1}{\sqrt{V_X}}\right) = \frac{1}{\sqrt{\beta}} \frac{\Gamma(\frac{2\alpha+1}{2})}{\Gamma(\alpha)}$$

so that, to a first order approximation,

$$\mathbb{E}(Z_X) \approx \mathbb{E}(U_X) \times \mathbb{E}\left(\frac{1}{\sqrt{V_X}}\right) \times \sqrt{\frac{N_0 N_1}{N(N-1)}}$$

Putting this together, we can now calculate the expected Z score, $\mathbf{Z}^{\mathrm{E}}$, across a set of SNPs, given a causal model and some phased reference data with which to calculate the probabilities in (1). Note that the computational complexity of this calculation is independent of both disease frequency and the number of samples required.

For some applications, the expected Z Score may suffice. However, note that the expected GWAS p value is not the p value associated with the expected Z score. Instead, we must simulated "observed" GWAS results which vary randomly about $\mathbf{Z}^{\mathrm{E}}$, with variance 1, such that the correlation between the Z score at two SNPs is equal to the correlation between their genotypes (Burren et al., 2014). It is hence computationally simple to simulate multiple realisations of GWAS Z scores as $\mathbf{Z}^* \sim \mathrm{MVN}(\mathbf{Z}^{\mathrm{E}}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a matrix describing correlation between SNPs for the region, again estimated from the reference panel.

To generate log odds ratios, γ, and their standard errors, σ, we appeal to the asymptotic similarity of Wald tests from a logistic regression model to the Cochran Armitage score test, and the result that the variance of the score statistic U_X is the inverse of the variance of the estimated γ, under the null (McCullagh and Nelder, 1983). Thus, we simulate $V_X^* \sim$

Inverse Gamma(α, β) with (α, β) given by (2,3) and hence $V(U_X)^* = V_X^* V_Y$. Finally, we set $\sigma^* = \sqrt{1/V(U_X)^*}$ and calculate $\gamma^* = \sigma^* \mathbf{Z}^*$.

Validation of simulated summary statistics

We visually confirmed that the calculated $\mathbf{Z}^{\mathrm{E}}$ appeared sensible for one to four causal SNP models in a single region (Figure S1). We next simulated data using our method or an individual-based method (HAPGEN+SNPTEST) for five scenarios (Table 1), and observed that both the distributions of Z scores and log odds ratios were similar for both strategies, either for sample sizes of 1000 cases and 1000 controls or 5000 cases and 5000 controls, both at the causal variants (Figure 1) and at variants in weak LD with them (Figure S2). Note in particular the difference between scenarios 4 and 5. In scenario 4, two variants in weak LD each have a log odds ratio of $\log(1.2) = 0.18$ or $\log(1/1.2) = -0.18$. In this case, marginal estimate of odds ratios are close to these values, and Z scores are highly significant. In scenario 5, the pair of odds ratios are the same, but at strongly linked variants ($r^{2=0.8}$). This would be expected to cause the effect of one to be "cancelled" by the other in the marginal associations, so that estimates of log OR are attenuated towards 1 and significance is dramatically lower, as seen for both HAPGEN+SNPTEST and simGWAS simulations.

Finally, we compared simulation speed of each strategy as the number of causal variants, the number of samples and the number of replicates varied. For a region with 1000 SNPs using AFR data from 1000 Genomes ($\sim$ 600 samples), both methods were very fast ($<$ 30 seconds) for the simplest scenario of 1000 cases and 1000 controls. We found that both methods required slightly, but negligibly, more time as the number of causal variants increased from one to six (Figure 2a). As expected, HAPGEN+SNPTEST scaled linearly with either the number of replications or sample size, whereas simGWAS timings were independent of either factor (Figure 2b–c). This emphasizes the potential for fast simulation of summary statistics for very large case-control datasets.

Discussion

Simulating GWAS summary statistics in the context of case-control studies, for any required causal model and set of odds ratios, has several potential applications. Primarily, simulated GWAS results have become the accepted gold standard for validating newly developed statistical models for the analysis of GWAS data. Our intent is to enable the faster simulation of summary statistics compared to individual level data simulation, while at the same time using considerably less disk space.

In addition to supporting method development, simulation of GWAS statistics is also used in tests aggregating information across sets of SNPs, e.g. for pathway analysis. Pathway analysis can test either the *global null*, of no association between any SNP and phenotype, or the *competitive null*, which assumes there are some truly associated SNPs, but that these are randomly distributed amongst the sets of SNPs considered (i.e. those near genes in or out of the pathway under test, or those corresponding to presence or absence of a feature of interest). The second seems more appropriate, because it acknowledges that enrichment tests are performed in the context of genome-wide significant associations having been already found. However, the second is also much harder to simulate.

A common technique for simulating under a competitive null is permutation testing; the underlying dataset is maintained, and labels are permuted to generate new datasets where traits are still associated, but there is no possible correlation to the feature of interest. However, doing this so as not to destroy the genomic structure within the region, can require inventive generation of null distributions, for example, by circularisation and permutation of genomic features to allow empirical null distributions to be calculated under a competitive null (Trynka et al., 2015). While these are efficient, they can only be used for features that span shorter distances than LD - e.g. for chromatin mark enrichment but not genes collected in pathways.

To allow more simple simulation techniques to be used, pathway-based tests of the competitive null have been adapted to have the same expected null distribution as tests of the global null. This requires replacing p values for individual genes by their ranks (Evangelou et al., 2012) which loses distributional information.

There is therefore potential to further develop pathway or enrichment test methodology if the distribution of test statistics under a competitive null hypothesis could be derived. Our method would naturally allow simulation of GWAS summary data under a specific hypothesis about the location and magnitude of genetic effects, in order to generate empirical null distributions for tests of the competitive null, preserving genomic structure even when analysis is performed across multiple regions.

Finally, our method could be used to evaluate output of fine-mapping applied to real data. Particularly in regions where the patterns of LD between putative associated SNPs are complex, it can be hard to dissect what the true causal variants are. Different fine mapping methods make different assumptions about the number and independence of causal variants, which can produce conflicting results (Wallace et al., 2015; Newcombe et al., 2016). By generating expected summary statistics under alternative fine-mapped solutions, it may be possible to see whether one or another is more compatible with observed data.

Our method enables faster simulation of GWAS case-control summary statistics compared to individual level data simulation, at the same time using considerably less disk space. This should facilitate computationally simpler evaluation of existing and new summary GWAS methods and has the potential to underpin new method development in other areas.

Methods

Simulations to validate summary statistics

We evaluated our proposed method by simulating summary statistics in parallel using simGWAS (our method) and the same settings with HAPGEN2 + SNPTEST2, using reference data from 1000 Genomes Phase 3 (1000 Genomes Project Consortium et al., 2015) (AFR cohort, $\sim$ 600 subjects). Reference data was downloaded from `https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference`. We visually compared distributions of summary statistics, as well as time to create the statistics under different scenarios. Full code to run these simulations is available from `http://github.com/chr1swallace/simgwas-paper`.

Availability of data and material A software package making our method available is at `http://github.com/chr1swallace/simGWAS` and code used to produce the figures and results in this paper is available at `http://github.com/chr1swallace/simgwas-paper`.

Conflict of interest None to declare

Acknowledgements We thank Marcus Klarqvist for advice on efficient C coding.

References

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., and Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, 47(3):291–295.

Burren, O. S., Guo, H., and Wallace, C. (2014). VSEAMS: a pipeline for variant set enrichment analysis using summary GWAS data identifies IKZF3, BATF and ESRRA as key transcription factors in type 1 diabetes. *Bioinformatics (Oxford, England)*, 30(23):3342–8.

Canela-Xandri, O., Rawlik, K., and Tenesa, A. (2017). An atlas of genetic associations in UK Biobank.

Chen, W., Larrabee, B. R., Ovsyannikova, I. G., Kennedy, R. B., Haralambieva, I. H., Poland, G. A., and Schaid, D. J. (2015). Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics*, 200(3):719–736.

Evangelou, M., Rendon, A., Ouwehand, W. H., Wernisch, L., and Dudbridge, F. (2012). Comparison of Methods for Competitive Tests of Pathway Analysis. *PLoS ONE*, 7(7):e41018.

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.*, 10(5):e1004383.

Kichaev, G., Yang, W. Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P., and Pasaniuc, B. (2014). Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLoS Genetics*, 10(10).

Li, C. and Li, M. (2008). GWAsimulator: a rapid whole-genome simulation program. *Bioinformatics*, 24(1):140–142.

Lloyd-Jones, L. R., Robinson, M. R., Yang, J., and Visscher, P. M. (2018). Transformation of Summary Statistics from Linear Mixed Model Association on All-or-None Traits to Odds Ratio. *Genetics*, (February):genetics.300360.2017.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, 39(7):906–913.

McCullagh, P. and Nelder, J. (1983). *Generalized Linear Models*. Generalized Linear Models, Chapman & Hall.

Meyer, H. V. and Birney, E. (2018). PhenotypeSimulator: a comprehensive framework for multi-trait, multi-locus phenotype to genotype simulation. *Bioinformatics*.

Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A., Bolla, M. K., Wang, Q., Tyrer, J., Dicks, E., Lee, A., Wang, Z., Allen, J., Keeman, R., Eilber, U., French, J. D., Qing Chen, X., Fachal, L., McCue, K., McCart Reed, A. E., Ghoussaini, M., Carroll, J. S., Jiang, X., Finucane, H., Adams, M., Adank, M. A., Ahsan, H., Aittomäki, K., Anton-Culver, H., Antonenkova, N. N., Arndt, V., Aronson, K. J., Arun, B., Auer, P. L., Bacot, F., Barrdahl, M., Baynes, C., Beckmann, M. W., Behrens, S., Benitez, J., Bermisheva, M., Bernstein, L., Blomqvist, C., Bogdanova, N. V., Bojesen, S. E., Bonanni, B., Børresen-Dale, A.-L., Brand, J. S., Brauch, H., Brennan, P., Brenner, H., Brinton, L., Broberg, P., Brock, I. W., Broeks, A., Brooks-Wilson, A., Brucker, S. Y., Brüning, T., Burwinkel, B., Butterbach, K., Cai, Q., Cai, H., Caldés, T., Canzian, F., Carracedo, A., Carter, B. D., Castelao, J. E., Chan, T. L., David Cheng, T.-Y., Seng Chia, K., Choi, J.-Y., Christiansen, H., Clarke, C. L., NBCS Collaborators, Collée, M., Conroy, D. M., Cordina-Duverger, E., Cornelissen, S., Cox, D. G., Cox, A., Cross, S. S., Cunningham, J. M., Czene, K., Daly, M. B., Devilee, P., Doheny, K. F., Dörk, T., Dos-Santos-Silva, I., Dumont, M., Durcan, L., Dwek, M., Eccles, D. M., Ekici, A. B., Eliassen, A. H., Ellberg, C., Elvira, M., Engel, C., Eriksson, M., Fasching, P. A., Figueroa, J., Flesch-Janys, D., Fletcher, O., Flyger, H., Fritschi, L., Gaborieau, V., Gabrielson, M., Gago-Dominguez, M., Gao, Y.-T., Gapstur, S. M., García-Sáenz, J. A., Gaudet, M. M., Georgoulias, V., Giles, G. G., Glendon, G., Goldberg, M. S., Goldgar, D. E., González-Neira, A., Grenaker Alnæs, G. I., Grip, M., Gronwald, J., Grundy, A., Guénel, P., Haeberle, L., Hahnen, E., Haiman, C. A., Håkansson, N., Hamann, U., Hamel, N., Hankinson, S., Harrington, P., Hart, S. N., Hartikainen, J. M., Hartman, M., Hein, A., Heyworth, J., Hicks, B., Hillemanns, P., Ho, D. N., Hollestelle, A., Hooning, M. J., Hoover, R. N., Hopper, J. L., Hou, M.-F., Hsiung, C.-N., Huang, G., Humphreys, K., Ishiguro, J., Ito, H., Iwasaki, M., Iwata, H., Jakubowska, A., Janni, W., John, E. M., Johnson, N., Jones, K., Jones, M., Jukkola-Vuorinen, A., Kaaks, R., Kabisch, M., Kaczmarek, K., Kang, D., Kasuga, Y., Kerin, M. J., Khan, S., Khusnutdinova, E., Kiiski, J. I., Kim, S.-W., Knight, J. A., Kosma, V.-M., Kristensen, V. N., Krüger, U., Kwong, A., Lambrechts, D., Le Marchand, L., Lee, E., Lee, M. H., Lee, J. W., Neng Lee, C., Lejbkowicz, F., Li, J., Lilyquist, J., Lindblom, A., Lissowska, J., Lo, W.-Y., Loibl, S., Long, J., Lophatananon, A., Lubinski, J., Luccarini, C., Lux, M. P., Ma, E. S. K., MacInnis, R. J., Maishman, T., Makalic, E., Malone, K. E., Kostovska, I. M., Mannermaa, A., Manoukian, S., Manson, J. E., Margolin, S., Mariapun, S., Martinez, M. E., Matsuo, K., Mavroudis, D., McKay, J., McLean, C., Meijers-Heijboer, H., Meindl, A., Menéndez, P., Menon, U., Meyer, J., Miao, H., Miller, N., Taib, N. A. M., Muir, K., Mulligan, A. M., Mulot, C., Neuhausen, S. L., Nevanlinna, H., Neven, P., Nielsen, S. F., Noh, D.-Y., Nordestgaard, B. G., Norman, A., Olopade, O. I., Olson, J. E., Olsson, H., Olswold, C., Orr, N., Pankratz, V. S., Park, S. K., Park-Simon, T.-W., Lloyd, R., Perez, J. I. A., Peterlongo, P., Peto, J., Phillips, K.-A., Pinchev, M., Plaseska-Karanfilska, D., Prentice, R., Presneau, N., Prokofyeva, D., Pugh, E., Pylkäs, K., Rack, B., Radice, P., Rahman, N., Rennert, G., Rennert, H. S., Rhenius, V., Romero, A., Romm, J., Ruddy, K. J., Rüdiger, T., Rudolph, A., Ruebner, M., Rutgers, E. J. T., Saloustros, E., Sandler, D. P., Sangrajrang, S., Sawyer, E. J., Schmidt, D. F., Schmutzler, R. K., Schneeweiss, A., Schoemaker, M. J., Schumacher, F., Schürmann, P., Scott, R. J., Scott, C., Seal, S., Seynaeve, C., Shah, M., Sharma, P., Shen, C.-Y., Sheng, G., Sherman, M. E., Shrubsole, M. J., Shu, X.-O., Smeets, A., Sohn, C., Southey, M. C., Spinelli, J. J., Stegmaier, C., Stewart-Brown, S., Stone, J., Stram, D. O., Surowy, H., Swerdlow, A., Tamimi, R., Taylor, J. A., Tengström, M., Teo, S. H., Beth Terry, M., Tessier, D. C., Thanasitthichai, S., Thöne, K., Tollenaar, R. A. E. M., Tomlinson, I., Tong, L., Torres, D., Truong, T., Tseng, C.-C., Tsugane, S., Ulmer, H.-U., Ursin, G., Untch, M., Vachon, C., van Asperen, C. J., Van Den Berg, D., van den Ouweland, A. M. W., van der Kolk, L., van der Luijt, R. B., Vincent, D., Vollenweider, J., Waisfisz, Q., Wang-Gohrke, S., Weinberg, C. R., Wendt, C., Whittemore, A. S., Wildiers, H., Willett, W., Winqvist, R., Wolk, A., Wu, A. H., Xia, L., Yamaji, T., Yang, X. R., Har Yip, C., Yoo, K.-Y., Yu, J.-C., Zheng, W., Zheng, Y., Zhu, B., Ziogas, A., Ziv, E., ABCTB Investigators, ConFab/AOCS Investigators, Lakhani, S. R., Antoniou, A. C., Droit, A., Andrulis, I. L., Amos, C. I., Couch, F. J., Pharoah, P. D. P., Chang-Claude, J., Hall, P., Hunter, D. J., Milne, R. L., García-Closas, M., Schmidt, M. K., Chanock, S. J., Dunning, A. M., Edwards, S. L., Bader, G. D., Chenevix-Trench, G., Simard, J., Kraft, P., and Easton, D. F. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551(7678):92–94.

Newcombe, P. J., Conti, D. V., and Richardson, S. (2016). JAM: A Scalable Bayesian Framework for Joint Analysis of Marginal SNP Effects. *Genetic Epidemiology*, 40(3):188–201.

P3G Consortium, Church, G., Heeney, C., Hawkins, N., de Vries, J., Boddington, P., Kaye, J., Bobrow, M., and Weir, B. (2009). Public access to genome-wide data: five views on balancing research with privacy and protection. *PLoS Genet.*, 5(10):e1000665.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411.

Sasieni, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics*, 53(4):1253–1261.

Shi, M., Umbach, D. M., Wise, A. S., and Weinberg, C. R. (2018). Simulating autosomal genotypes with realistic linkage disequilibrium and a spiked-in genetic effect. *BMC Bioinformatics*, 19(1):2.

Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, 27(16):2304–2305.

Trynka, G., Westra, H.-J., Slowikowski, K., Hu, X., Xu, H., Stranger, B., Klein, R., Han, B., and Raychaudhuri, S. (2015). Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *The American Journal of Human Genetics*, 97(1):139–152.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.*, 101(1):5–22.

Wallace, C., Cutler, A. J., Pontikos, N., Pekalski, M. L., Burren, O. S., Cooper, J. D., García, A. R., Ferreira, R. C., Guo, H., Walker, N. M., Smyth, D. J., Rich, S. S., Onengut-Gumuscu, S., Sawcer, S. J., Ban, M., Richardson, S., Todd, J. A., and Wicker, L. S. (2015). Dissection of a Complex Disease Susceptibility Region Using a Bayesian Stochastic Search Approach to Fine Mapping. *PLoS Genet.*, 11(5):e1005272.

Walter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J. R. B., Xu, C., Futema, M., Lawson, D., Iotchkova, V., Schiffels, S., Hendricks, A. E., Danecek, P., Li, R., Floyd, J., Wain, L. V., Barroso, I., Humphries, S. E., Hurles, M. E., Zeggini, E., Barrett, J. C., Plagnol, V., Brent Richards, J., Greenwood, C. M. T., Timpson, N. J., Durbin, R., Soranzo, N., Bala, S., Clapham, P., Coates, G., Cox, T., Daly, A., Du, Y., Edkins, S., Ellis, P., Flicek, P., Guo, X., Guo, X., Huang, L., Jackson, D. K., Joyce, C., Keane, T., Kolb-Kokocinski, A., Langford, C., Li, Y., Liang, J., Lin, H., Liu, R., Maslen, J., Muddyman, D., Quail, M. A., Stalker, J., Sun, J., Tian, J., Wang, G., Wang, J., Wang, Y., Wong, K., Zhang, P., Birney, E., Boustred, C., Chen, L., Clement, G., Cocca, M., Davey Smith, G., Day, I. N. M., Day-Williams, A., Down, T., Dunham, I., Evans, D. M., Gaunt, T. R., Geihs, M., Hart, D., Howie, B., Hubbard, T., Hysi, P., Jamshidi, Y., Karczewski, K. J., Kemp, J. P., Lachance, G., Lek, M., Lopes, M., MacArthur, D. G., Marchini, J., Mangino, M., Mathieson, I., Metrustry, S., Moayyeri, A., Northstone, K., Panoutsopoulou, K., Paternoster, L., Quaye, L., Ring, S., Ritchie, G. R. S., Shihab, H. A., Shin, S.-Y., Small, K. S., Soler Artigas, M., Southam, L., Spector, T. D., St Pourcain, B., Surdulescu, G., Tachmazidou, I., Tobin, M. D., Valdes, A. M., Visscher, P. M., Ward, K., Wilson, S. G., Yang, J., Zhang, F., Zheng, H.-F., Anney, R., Ayub, M., Blackwood, D., Bolton, P. F., Breen, G., Collier, D. A., Craddock, N., Curran, S., Curtis, D., Gallagher, L., Geschwind, D., Gurling, H., Holmans, P., Lee, I., Lönnqvist, J., McGuffin, P., McIntosh, A. M., McKechanie, A. G., McQuillin, A., Morris, J., O'Donovan, M. C., Owen, M. J., Palotie, A., Parr, J. R., Paunio, T., Pietilainen, O., Rehnström, K., Sharp, S. I., Skuse, D., St Clair, D., Suvisaari, J., Walters, J. T. R., Williams, H. J., Bochukova, E., Bounds, R., Dominiczak, A., Farooqi, I. S., Keogh, J., Marenne, G., Morris, A., O'Rahilly, S., Porteous, D. J., Smith, B. H., Wheeler, E., Al Turki, S., Anderson, C. A., Antony, D., Beales, P., Bentham, J., Bhattacharya, S., Calissano, M., Carss, K., Chatterjee, K., Cirak, S., Cosgrove, C., Fitzpatrick, D. R., Reghan Foley, A., Franklin, C. S., Grozeva, D., Mitchison, H. M., Muntoni, F., Onoufriadis, A., Parker, V., Payne, F., Lucy Raymond, F., Roberts, N., Savage, D. B., Scambler, P., Schmidts, M., Schoenmakers, N., Semple, R. K., Serra, E., Spasic-Boskovic, O., Stevens, E., van Kogelenberg, M., Vijayarangakannan, P., Williamson, K. A., Wilson, C., Whyte, T., Ciampi, A., Oualkacha, K., Bobrow, M., Griffin, H., Kaye, J., Kennedy, K., Kent, A., Smee, C., Charlton, R., Ekong, R., Khawaja, F., Lopes, L. R., Migone, N., Payne, S. J., Pollitt, R. C., Povey, S., Ridout, C. K., Robinson, R. L., Scott, R. H., Shaw, A., Syrris, P., Taylor, R., Vandersteen, A. M., Amuzu, A., Pablo Casas,

Scenario	Description
1	Single common causal variant, weak effect (MAF=0.5, odds ratio=1.1)
2	Single low frequency causal variant, strong effect (MAF=0.02, odds ratio=1.5)
3	Three causal variants, unlinked (odds ratios 1.1, 1.2, 1.3)
4	Two causal variants, weakly linked (r=0.15, odds ratios 1.2 and 1/1.2)
5	Two causal variants, strongly linked (r=0.8, odds ratios 1.2 and 1/1.2)

Table 1: Five simulation scenarios considered for validation of results

J., Chambers, J. C., Dedoussis, G., Gambaro, G., Gasparini, P., Isaacs, A., Johnson, J., Kleber, M. E., Kooner, J. S., Langenberg, C., Luan, J., Malerba, G., März, W., Matchan, A., Morris, R., Nordestgaard, B. G., Benn, M., Scott, R. A., Toniolo, D., Traglia, M., Tybjaerg-Hansen, A., van Duijn, C. M., van Leeuwen, E. M., Varbo, A., Whincup, P., Zaza, G., and Zhang, W. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, 526(7571):82–90.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M., and Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.*

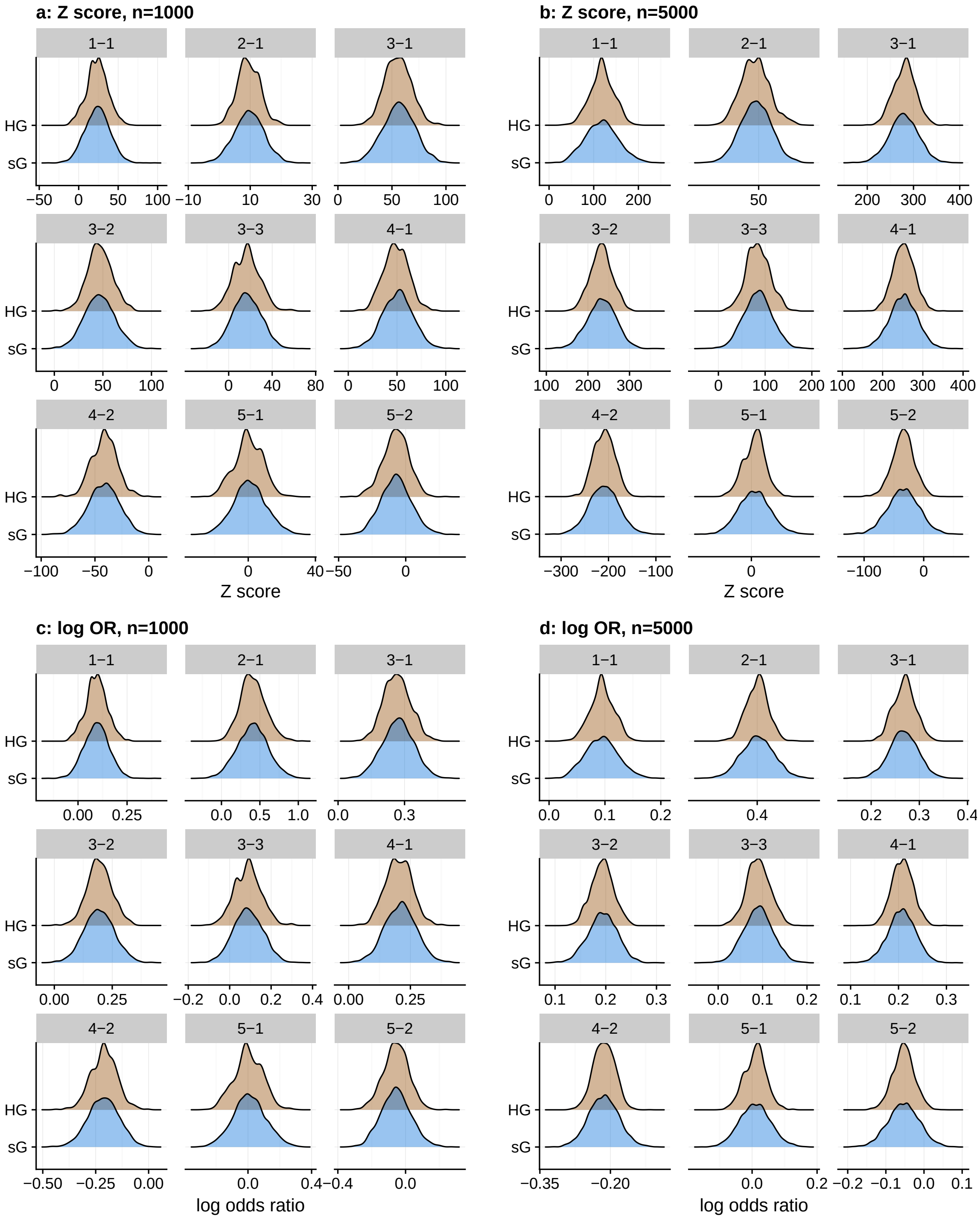

Figure 1: Results from simGWAS are similar to those from HAPGEN+SNPTEST at simulated causal variants. Distributions of simulated Z scores (a-b) and log odds ratios (c-d) from HAPGEN+SNPTEST (top line in each pair, "HG") and simGWAS (bottom line "sG") for single and multi-SNP models under five scenarios, described in Table 1. The sample size (n) is the number of cases and controls - i.e. n=1000 indicates the simulations related to 1000 cases and 1000 controls. The label of each plot gives the corresponding "scenario-snp" pair - i.e. the label 3-1 refers to scenario 3, first causal SNP.

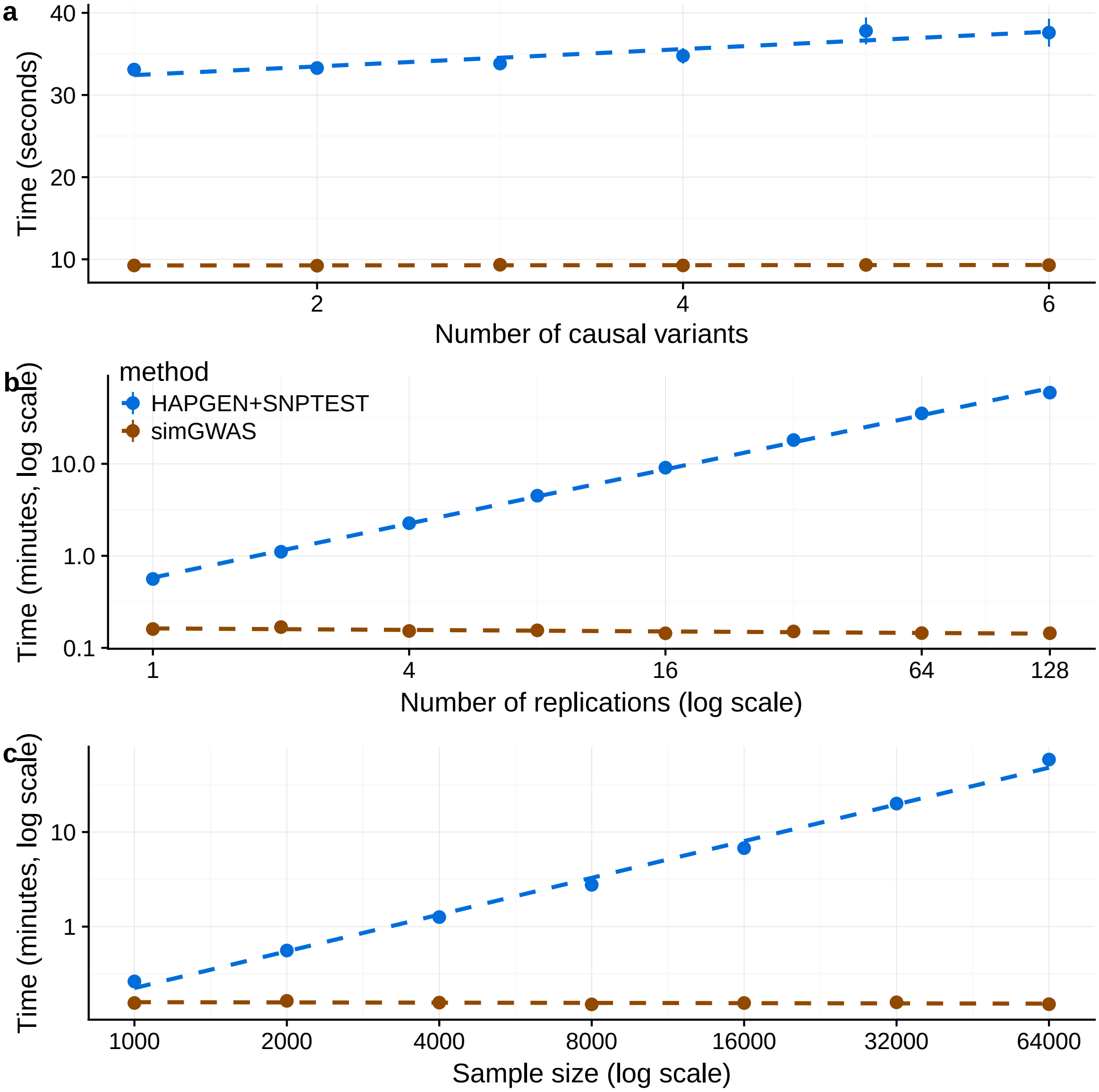

Figure 2: Time taken to perform simulations under simGWAS or HAPGEN2+SNPTEST2 strategies. simGWAS is given in brown and HAPGEN2+SNPTEST2 in blue. For each comparison, 50 simulations were run. The points show mean time taken, with standard deviation about that mean indicated by the vertical bars. **(a)** The effect of number of causal variants on run time. 2000 cases, 2000 controls, single replication, causal variants varying from 1 to 6. **(b)** The effect of number of replications on run time. 2000 cases, 2000 controls, 2 causal variants, number of replications varying from 1 to 100. **(c)** The effect of sample size on run time. single replication, 3 causal variants, number of cases and controls (each) varying from 1000 to 64000.

Appendix

Cochran-Armitage test of association

For a GWAS dataset, let $Y_i \in \{0,1\}$ denote the indicator of disease status at the ith sample. Let there be a total of N samples selected, with N_1 having been chosen from disease cases ($Y_i = 1$) and N_0 having been chosen from disease controls ($Y_i = 0$). Since this sampling is conditional upon case/control status, genotype frequencies may differ between our N samples and the whole population at disease associated SNPs. We therefore need to distinguish between which datasets the genotype probabilities are from; write $\mathbb{P}_{sam}$ for probabilities computed for the samples (i.e. $\mathbb{P}_{sam}(Y_i = 1) = \frac{N_1}{N}$), and $\mathbb{P}$ for probabilities generated with reference to the whole population.

Let n be the total number of SNPs. For any SNP X, write G_i^X for its genotype coding $\in \{0,1,2\}$ at sample i .

For the commonly used Cochran-Armitage test, the Z-Score at SNP X is computed as:

$$Z_X = \frac{U_X}{\sqrt{V}}$$

Where:

$$U_X = \sum_{i=1}^{N}((G_i^X - \overline{G^X})(Y_i - \bar{Y}))$$

$$V = (N-1)V_X V_Y$$

and V_X, V_Y are the variance of G^X and Y respectively:

$$V_X = \frac{N}{N-1}\frac{\sum_{i=1}^{N}(G_i^X - \overline{G^X})^2}{N}$$

$$V_Y = \frac{N_0 N_1}{N(N-1)}$$

i.e.:

$$V = \frac{N_0 N_1}{N(N-1)}\sum_{i=1}^{N}(G_i^X - \overline{G^X})^2$$

Under the null hypothesis of no association at SNP X, Z_X is distributed as a standard normal. Hence the two-sided p-value at X is given by:

$$p_X = 2(1 - \Phi(|Z_X|))$$

where Φ is the cumulative distribution function of the standard normal distribution. Conversely, given the unsigned p-value at X, the absolute value of the Z-Score is:

$$-\Phi^{-1}\left(\frac{p}{2}\right)$$

Allelic frequencies under a Causal Model

Write $\mathbf{W} = W_1, ..., W_m$ for the vector of causal SNPs. From phased publicly available reference datasets such as UK10K (Walter et al., 2015), it is possible to estimate haplotype frequencies across all SNPs in $\mathbf{W}$ at any subset of potential causal SNPs in control datasets. Since they are causal, these frequencies will differ in cases, and it is those frequencies we derive first. Note that, since sampling dependent only upon case/control status, we can assume:

$$\mathbb{P}_{sam}(G^{\mathbf{W}} = \mathbf{w}|Y = 0) = \mathbb{P}(G^{\mathbf{W}} = \mathbf{w}|Y = 0)$$

Write $\gamma_1, ..., \gamma_m$ for the log odds ratios of effect for the causal SNPs in the population. We assume that Y given $G^{\mathbf{W}}$ can be modelled as a binomial logistic regression. Then, from (Prentice and Pyke, 1979), the sample-specific odds ratios are the same as those at the population-level, and we can write:

$$\mathbb{P}_{sam}(Y_i = 1|G_i^{\mathbf{W}} = \mathbf{w}) = \frac{e^{\gamma_0 + \gamma_1 w_1 + ... + \gamma_m w_m}}{1 + e^{\gamma_0 + \gamma_1 w_1 + ... + \gamma_m w_m}}$$

where γ_0 is an intercept parameter. Since GWAS sampling is retrospective, the proportion of cases in the sample is fixed at $\frac{N_1}{N}$, constraining γ_0, which can be computed as follows:

$$\begin{aligned}
\mathbb{P}_{sam}(Y_i = 1) &= \frac{N_1}{N} \\
&= \sum_{\mathbf{w}\in\mathbb{Z}_3^m} \mathbb{P}_{sam}(Y_i = 1|G_i^{\mathbf{W}} = \mathbf{w})\mathbb{P}_{sam}(G_i^{\mathbf{W}} = \mathbf{w}) \\
&= \sum_{\mathbf{w}\in\mathbb{Z}_3^m} \mathbb{P}_{sam}(Y_i = 1|G_i^{\mathbf{W}} = \mathbf{w})\frac{\mathbb{P}_{sam}(Y_i = 0|G_i^{\mathbf{W}} = \mathbf{w})}{\mathbb{P}_{sam}(Y_i = 0|G_i^{\mathbf{W}} = \mathbf{w})}\mathbb{P}_{sam}(G_i^{\mathbf{W}} = \mathbf{w}) \\
&= \sum_{\mathbf{w}\in\mathbb{Z}_3^m} \frac{\mathbb{P}_{sam}(Y_i = 1|G_i^{\mathbf{W}} = \mathbf{w})}{\mathbb{P}_{sam}(Y_i = 0|G_i^{\mathbf{W}} = \mathbf{w})}\mathbb{P}_{sam}(Y_i = 0)\mathbb{P}(G_i^{\mathbf{W}} = \mathbf{w}|Y_i = 0) \\
&= \frac{N_0}{N}\sum_{\mathbf{w}\in\mathbb{Z}_3^m} e^{\gamma_0+\gamma_1 w_1+\ldots+\gamma_m w_m}\mathbb{P}(G_i^{\mathbf{W}} = \mathbf{w}|Y_i = 0)
\end{aligned}$$

$$\gamma_0 = \ln\left(\frac{N_1}{N_0\sum_{\mathbf{w}\in\mathbb{Z}_3^m} e^{\gamma_1 w_1+\ldots+\gamma_m w_m}\mathbb{P}(G_i^{\mathbf{W}} = \mathbf{w}|Y_i = 0)}\right)$$

Hence we can compute:

$$\mathbb{P}_{sam}(G_i^{\mathbf{W}} = \mathbf{w}) = \frac{\mathbb{P}(G_i^{\mathbf{W}} = \mathbf{w}|Y_i = 0)\mathbb{P}_{sam}(Y_i = 0)}{\mathbb{P}_{sam}(Y_i = 0|G_i^{\mathbf{W}} = \mathbf{w})}$$

And also:

$$\mathbb{P}_{sam}(G_i^{\mathbf{W}} = \mathbf{w}|Y_i = 1) = \mathbb{P}(G_i^{\mathbf{W}} = \mathbf{w}|Y_i = 1) = \frac{\mathbb{P}_{sam}(Y_i = 1|G_i^{\mathbf{W}} = \mathbf{w})\mathbb{P}_{sam}(G_i^{\mathbf{W}} = \mathbf{w})}{\mathbb{P}_{sam}(Y_i = 1)}$$

To derive genotype probabilities at SNPs in LD with the causal SNPs, we assume that LD structures do not differ between cases and controls, and hence the correlation between $\mathbf{W}$ and X is independent of both disease status and our sampling. Thus:

$$\mathbb{P}_{sam}(G_i^X = x|G_i^{\mathbf{W}} = \mathbf{w}) = \mathbb{P}(G_i^X = x|G_i^{\mathbf{W}} = \mathbf{w})$$

and we can estimate, for both the whole population, and for our sample:

$$\mathbb{E}((G_i^X)^a|G_i^{\mathbf{W}}=\mathbf{w}) = 2^a\frac{\mathbb{P}(G_i^X=2\cap G_i^{\mathbf{W}}=\mathbf{w}|Y_i=0)}{\mathbb{P}(G_i^{\mathbf{W}}=\mathbf{w}|Y_i=0)}+\frac{\mathbb{P}(G_i^X=1\cap G_i^{\mathbf{W}}=\mathbf{w}|Y_i=0)}{\mathbb{P}(G_i^{\mathbf{W}}=\mathbf{w}|Y_i=0)}$$

from our reference dataset, for any constant a. From this, we compute:

$$\begin{aligned}\mathbb{E}((G_i^X)^a|Y_i=1) &= \sum_{\mathbf{w}\in\mathbb{Z}_3^m}\mathbb{E}((G_i^X)^a|G_i^{\mathbf{W}}=\mathbf{w})\mathbb{P}(G_i^{\mathbf{W}}=\mathbf{w}|Y_i=1)\\ &= \sum_{\mathbf{w}\in\mathbb{Z}_3^m}\frac{\mathbb{P}(G_i^{\mathbf{W}}=\mathbf{w}|Y_i=1)}{\mathbb{P}(G_i^{\mathbf{W}}=\mathbf{w}|Y_i=0)}\left[2^a\mathbb{P}(G_i^X=2\cap G_i^{\mathbf{W}}=\mathbf{w})+\mathbb{P}(G_i^X=1\cap G_i^{\mathbf{W}}=\mathbf{w})\right]\end{aligned}$$

$$\begin{aligned}\mathbb{E}((G_i^X)^a|Y_i=0) &= \sum_{\mathbf{w}\in\mathbb{Z}_3^m}\mathbb{E}((G_i^X)^a|G_i^{\mathbf{W}}=\mathbf{w})\mathbb{P}(G_i^{\mathbf{W}}=\mathbf{w}|Y_i=0)\\ &= \sum_{\mathbf{w}\in\mathbb{Z}_3^m}2^a\mathbb{P}(G_i^X=2\cap G_i^{\mathbf{W}}=\mathbf{w})+\mathbb{P}(G_i^X=1\cap G_i^{\mathbf{W}}=\mathbf{w})\end{aligned}$$

By expanding out the numerator in terms of probabilities within the sample dataset, we see that:

$$\begin{aligned}\frac{\mathbb{P}(G_i^{\mathbf{W}}=\mathbf{w}|Y_i=1)}{\mathbb{P}(G_i^{\mathbf{W}}=\mathbf{w}|Y_i=0)} &= \frac{\mathbb{P}_{sam}(Y_i=1|G_i^{\mathbf{W}}=\mathbf{w})\mathbb{P}_{sam}(G_i^{\mathbf{W}}=\mathbf{w})}{\mathbb{P}(G_i^{\mathbf{W}}=\mathbf{w}|Y_i=0)\mathbb{P}_{sam}(Y_i=1)}\\ &= \frac{\mathbb{P}_{sam}(Y_i=1|G_i^{\mathbf{W}}=\mathbf{w})\mathbb{P}_{sam}(G_i^{\mathbf{W}}=\mathbf{w}|Y_i=0)\mathbb{P}_{sam}(Y_i=0)}{\mathbb{P}(G_i^{\mathbf{W}}=\mathbf{w}|Y_i=0)\mathbb{P}_{sam}(Y_i=1)\mathbb{P}_{sam}(Y_i=0|G_i^{\mathbf{W}}=\mathbf{w})}\\ &= \frac{N_0}{N_1}e^{\gamma_0+\gamma_1 w_1+\ldots+\gamma_m w_m}\end{aligned}$$

And hence:

$$\begin{aligned}
\mathbb{E}_{sam}((G_i^X)^a) &= \frac{N_1}{N}\mathbb{E}((G_i^X)^a|Y_i=1)+\frac{N_0}{N}\mathbb{E}((G_i^X)^a|Y_i=0)\\
&= \frac{N_0}{N}\sum_{\mathbf{w}\in\mathbb{Z}_3^m} e^{\gamma_0+\gamma_1 w_1+\ldots+\gamma_m w_m}\left[2^a\mathbb{P}(G_i^X=2\cap G_i^{\mathbf{W}}=\mathbf{w})+\mathbb{P}(G_i^X=1\cap G_i^{\mathbf{W}}=\mathbf{w})\right]\\
&\quad +\frac{N_0}{N}\sum_{\mathbf{w}\in\mathbb{Z}_3^m}\left[2^a\mathbb{P}(G_i^X=2\cap G_i^{\mathbf{W}}=\mathbf{w})+\mathbb{P}(G_i^X=1\cap G_i^{\mathbf{W}}=\mathbf{w})\right]\\
&= \frac{N_0}{N}\sum_{\mathbf{w}\in\mathbb{Z}_3^m}\left(e^{\gamma_0+\gamma_1 w_1+\ldots+\gamma_m w_m}+1\right)\\
&\qquad \left[2^a\mathbb{P}(G_i^X=2\cap G_i^{\mathbf{W}}=\mathbf{w})+\mathbb{P}(G_i^X=1\cap G_i^{\mathbf{W}}=\mathbf{w})\right]
\end{aligned}$$

Estimation of Z Score for the causal model given by W and γ

Finding the true expectation of $\frac{U_X}{\sqrt{V}}$ is intractable, so instead we compute a first order approximation by assuming independence:

$$\mathbb{E}(Z_X)=\mathbb{E}\left(\frac{U_X}{\sqrt{V}}\right)\approx\mathbb{E}(U_X)\times\mathbb{E}\left(\frac{1}{\sqrt{V}}\right)$$

These terms can be computed as shown in the following sections.

Estimation of U_X, the covariance between G^X and Y, for the causal model given by $\mathbf{W}$ and γ

We compute the expectation of U_X in our sample as follows:

$$\begin{aligned}
\mathbb{E}_{sam}(U_X) &= \mathbb{E}_{sam}\left[\sum_{i=1}^{N}(G_i^X - \overline{G^X})(Y_i - \bar{Y})\right] \\
&= \mathbb{E}_{sam}\left[N\left(\sum_{i=1}^{N} G_i^X Y_i\right) - \frac{1}{N}\left(\sum_{i=1}^{N} G_i^X\right)\left(\sum_{i=1}^{N} Y_i\right)\right] \\
&= N\mathbb{E}_{sam}(G_i^X Y_i) - \frac{1}{N}\left[N\mathbb{E}_{sam}(G_i^X Y_i) + N(N-1)\mathbb{E}_{sam}(G_i^X Y_j)\right] \qquad\qquad i \neq j \\
&= (N-1)\left[\mathbb{E}_{sam}(G_i^X Y_i) - \mathbb{E}_{sam}(G_i^X Y_j)\right] \\
&= (N-1)\left[\mathbb{E}_{sam}(G_i^X | Y_i = 1)\mathbb{P}_{sam}(Y_i = 1)\right] - \\
&\quad -(N-1)\mathbb{E}_{sam}(Y_j)\left[\mathbb{E}_{sam}(G_i^X | Y_i = 1)\mathbb{P}_{sam}(Y_i = 1) + \mathbb{E}_{sam}(G_i^X | Y_i = 0)\mathbb{P}_{sam}(Y_i = 0)\right] \\
&= \frac{(N-1)N_0 N_1}{N^2}\left[\mathbb{E}_{sam}(G_i^X | Y_i = 1) - \mathbb{E}_{sam}(G_i^X | Y_i = 0)\right]
\end{aligned}$$

Using the expressions for $\mathbb{E}_{sam}(G_i^X | Y_i)$ given in Section, this becomes:

$$\begin{aligned}
\mathbb{E}_{sam}(U_X) &= \frac{(N-1)N_0 N_1}{N^2} \sum_{\mathbf{w} \in \mathbb{Z}_3^m} \left[\left(\frac{N_0}{N_1} e^{\gamma_0 + \gamma_1 w_1 + \ldots + \gamma_m w_m} - 1\right)\right. \\
&\qquad\qquad \left.\left[2\mathbb{P}(G_i^X = 2 \cap G_i^{\mathbf{W}} = \mathbf{w}) + \mathbb{P}(G_i^X = 1 \cap G_i^{\mathbf{W}} = \mathbf{w})\right]\right]
\end{aligned}$$

Estimation of V_X, the variance of G^X, for the causal model given by $\mathbf{W}$ and γ

Recall:

$$\begin{aligned}
V_X &= \frac{1}{(N-1)} \sum_{i=1}^{N} (G_i^X - \overline{G^X})^2 \\
&= \frac{1}{(N-1)}\left[\left(\sum_{i=1}^{N} (G_i^X)^2\right) - \frac{1}{N}\left(\sum_{i=1}^{N} G_i^X\right)^2\right]
\end{aligned}$$

This is tractable, however, we need to find $\mathbb{E}\left(\frac{1}{\sqrt{V_X}}\right)$, which is more complex.

V_X is the variance of a normal, and so we model it as an Inverse Gamma (α, β) distribution. Then V_X^{-1} has a $\Gamma(\alpha, \beta^{-1})$ distribution, and $\sqrt{V_X^{-1}}$ has a generalised gamma distribution with parameters $p = 2, d = 2\alpha, a = \sqrt{\beta^{-1}}$. If $V_X \sim$ Inverse Gamma (α, β), then

$$\mathbb{E}(V_X) = \frac{\beta}{\alpha - 1} \qquad\qquad Var(V_X) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$$

Assuming we have computed $\mathbb{E}_{sam}(V_X)$ and $\mathbb{E}_{sam}(V_X^2)$, α and β are completely specified as:

$$\alpha = \frac{2\mathbb{E}(V_X^2) - (\mathbb{E}(V_X))^2}{\mathbb{E}(V_X^2) - (\mathbb{E}(V_X))^2} \qquad\qquad \beta = \frac{\mathbb{E}(V_X)\mathbb{E}(V_X^2)}{\mathbb{E}(V_X^2) - (\mathbb{E}(V_X))^2}$$

and $\mathbb{E}\left(\frac{1}{\sqrt{V_X}}\right)$ may be simply computed using:

$$\mathbb{E}\left(\frac{1}{\sqrt{V_X}}\right) = a\frac{\Gamma(\frac{d+1}{p})}{\Gamma(\frac{d}{p})} = \frac{1}{\sqrt{\beta}}\frac{\Gamma(\frac{2\alpha+1}{2})}{\Gamma(\alpha)}$$

Expectation of V_X

$$\begin{aligned}
\mathbb{E}_{sam}(V_X) &= \frac{1}{(N-1)}\left[N\mathbb{E}_{sam}((G_i^X)^2) - \frac{1}{N}\left(N\mathbb{E}_{sam}((G_i^X)^2) + N(N-1)\mathbb{E}_{sam}(G_i^X G_j^X)\right)\right] \\
&= \frac{1}{(N-1)}\left[(N-1)\mathbb{E}_{sam}((G_i^X)^2) - (N-1)\mathbb{E}_{sam}(G_i^X G_j^X)\right] \\
&= \mathbb{E}_{sam}((G_i^X)^2) - (\mathbb{E}_{sam}(G_i^X))^2
\end{aligned}$$

Expectation of V_X^2

$$\mathbb{E}_{sam}(V_X^2) = \left(\frac{1}{(N-1)}\right)^2 \mathbb{E}_{sam}\left[\left(\sum_{i=1}^{N}(G_i^X)^2\right)^2 - \frac{2}{N}\left(\sum_{i=1}^{N}(G_i^X)^2\right)\left(\sum_{i=1}^{N}G_i^X\right)^2 + \frac{1}{N^2}\left(\sum_{i=1}^{N}G_i^X\right)^4\right]$$

Let $E_n = \mathbb{E}_{sam}((G_i^X)^n)$. Breaking this down into terms, for (i, j, k, l) representing different indices, we have:

$$\begin{aligned}
&\mathbb{E}_{sam}\left[\left(\sum_{i=1}^{N}(G_i^X)^2\right)^2\right] \\
=\ & N\mathbb{E}_{sam}((G_i^X)^4)+N(N-1)\mathbb{E}_{sam}((G_i^X)^2(G_j^X)^2) \\
=\ & NE_4+N(N-1)
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_{sam}\left[\left(\sum_{i=1}^{N}(G_i^X)^2\right)\left(\sum_{i=1}^{N}G_i^X\right)^2\right] \\
=\ & N\mathbb{E}_{sam}((G_i^X)^4)+2N(N-1)\mathbb{E}_{sam}((G_i^X)^3(G_j^X))+N(N-1)\mathbb{E}_{sam}((G_i^X)^2(G_j^X)^2)+ \\
+\ & N(N-1)(N-2)\mathbb{E}_{sam}((G_i^X)^2(G_j^X)(G_k^X)) \\
=\ & NE_4+2N(N-2)E_3E_1+N(N-1)E_2^2+N(N-1)(N-2)E_2E_1^2
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_{sam}\left[\left(\sum_{i=1}^{N}G_i^X\right)^4\right] \\
=\ & N\mathbb{E}_{sam}((G_i^X)^4)+4N(N-1)\mathbb{E}_{sam}((G_i^X)^3(G_j^X))+6N(N-1)\mathbb{E}_{sam}((G_i^X)^2(G_j^X)^2)+ \\
+\ & 6N(N-1)(N-2)\mathbb{E}_{sam}((G_i^X)^2(G_j^X)(G_k^X))+ \\
+\ & N(N-1)(N-2)(N-3)\mathbb{E}_{sam}((G_i^X)(G_j^X)(G_k^X)(G_l^X)) \\
=\ & NE_4+4N(N-1)E_3E_1+6N(N-1)E_2^2+6N(N-1)(N-2)E_2E_1^2+ \\
+\ & N(N-1)(N-2)(N-3)E_1^4
\end{aligned}$$

Giving:

$$\mathbb{E}_{sam}(V_X^2) = \frac{1}{N}E_4-\frac{4}{N}E_3E_1+2\frac{N^2-2N+6}{N(N-1)}E_2^2-2\frac{(N-2)(N-3)}{N(N-1)}E_2E_1^2+\frac{(N-2)(N-3)}{N(N-1)}E_1^4$$

Summary

Thus, given only a choice of which SNPs are causal ($\mathbf{W}$), their effect sizes ($\boldsymbol{\gamma}$), sample sizes (N_0, N_1) and a reference dataset from which we can derive allele frequencies ($\mathbb{E}(G_i^X|Y_i = 0)$) and the relationships between SNPs ($\mathbb{E}(G_i^X|G_i^{\mathbf{W}} = \mathbf{w})$), we can derive an expected Z Score, $\mathbf{Z}^{EXP}$ at any SNP, causal or not.

This can then be used directly. However, most applications require simulated output from such a GWAS. $\mathbf{Z}^{SIM}$ can therefore be computed, which will be distributed:

$$\mathbf{Z}^{SIM} \sim \mathbf{N}(\mathbf{Z}^{EXP}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}$ is the genotype correlation matrix for the SNPs in this regionBurren et al. (2014).

Supplementary Figures

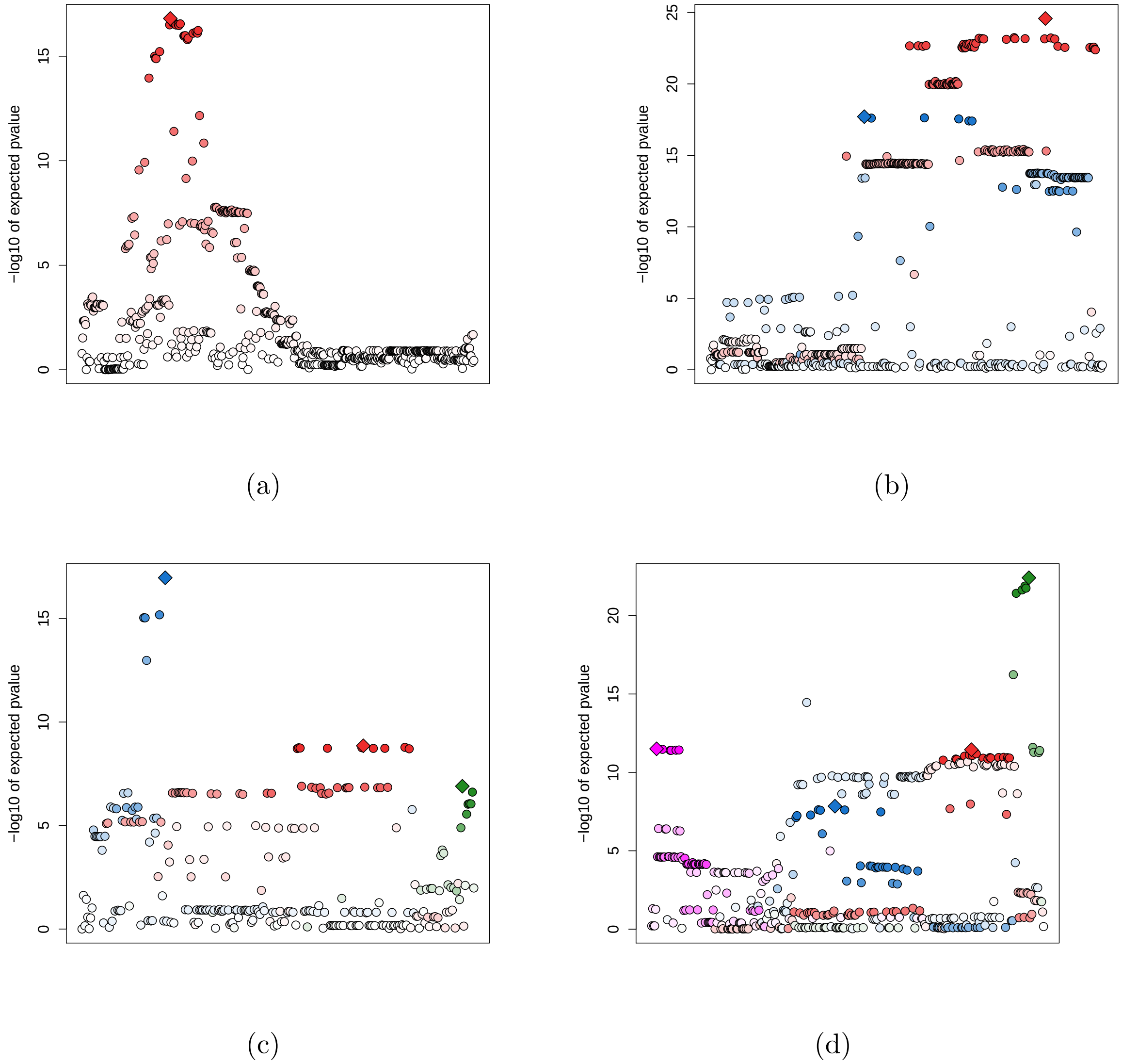

Figure S1: Local Manhattan plots for p values generated from expected Z scores under different scenarios, in order to confirm by visual inspection that the expected Z Scores produced by our algorithm are consistent with the behaviour we would expect from their causal SNPs. In order to easily see the pattern of association, causal variants chosen were common, with a strong effect and (in the case of multiple causal variants) only weakly linked. Causal SNPs are designated by a coloured diamond. Non-causal SNPs are designated by a circle, coloured according to their LD with their most correlated causal SNP. (a) A single causal variant with MAF = 0.34 and Odds Ratio of effect = 1.3 (b) Two causal variants with MAF = $(0.14, 0.30)$ and Odds Ratio of effect = $(1.5, 1.2)$ (c) Three causal variants with MAF = $(0.12, 0.43, 0.17)$ and Odds Ratio of effect = $(1.2, 1.2, 1.2)$ (d) Four causal variants with MAF = $(0.33, 0.44, 0.17, 0.28)$ and Odds Ratio of effect = $(1.5, 1.5, 1.5, 1.5)$

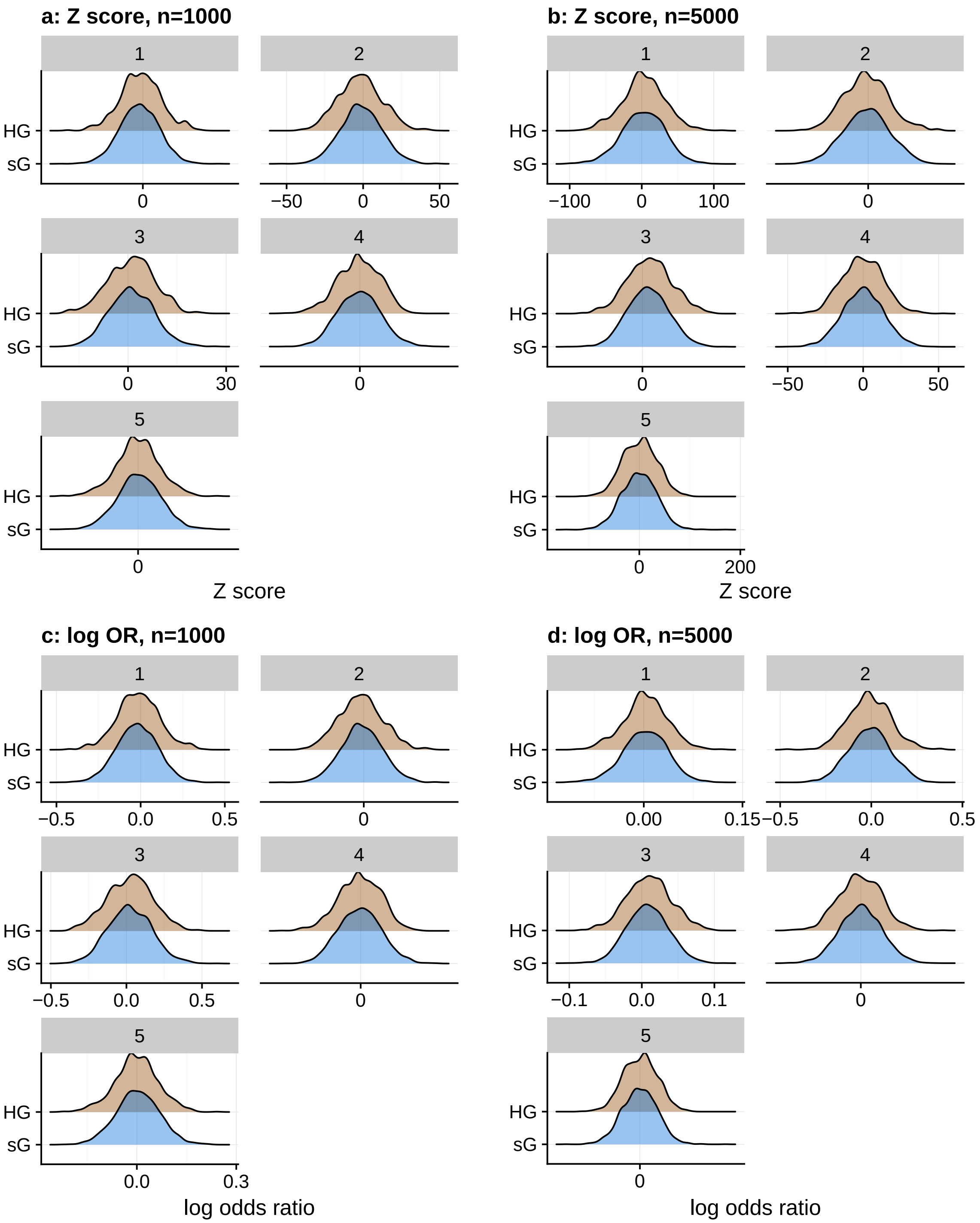

Figure S2: Results from simGWAS are similar to those from HAPGEN+SNPTEST at SNPs unlinked to the simulated causal variants. Distributions of simulated Z scores (a-b) and log odds ratios (c-d) from HAPGEN+SNPTEST (top line in each pair, "HG") and simGWAS (bottom line "sG") for single and multi-SNP models under five scenarios, described in Table 1. The sample size (n) is the number of cases and controls - i.e. n=1000 indicates the simulations related to 1000 cases and 1000 controls. The label of each plot gives the corresponding "scenario-snp" pair - i.e. the label 3-1 refers to scenario 3, first causal SNP. We chose a SNP in $r^2 < 0.1$ with any causal variant to show that simulated results at non-causal are also valid.