BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 23 2014, pages 3432–3434
doi:10.1093/bioinformatics/btu564

Genetics and population analysis

Advance Access publication August 26, 2014

MR_predictor: a simulation engine for Mendelian Randomization studies

Benjamin F. Voight[1,2,*]

[1]Department of Pharmacology and [2]Department of Genetics, University of Pennsylvania - Perelman School of Medicine, Philadelphia, PA 19143, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: I present MR_predictor, a simulation engine designed to guide the development and interpretation of statistical tests of causality between phenotypes using genetic instruments. MR_predictor provides a framework to model either individual traits or complex scenarios where multiple phenotypes are correlated or dependent on each other. Crucially, MR_predictor can incorporate the effects of multiple biallelic loci (linked or unlinked) contributing genotypic variability to one or more simulated phenotypes. The software has a range of options for sample generation, and output files generated by MR_predictor port into commonly used analysis tools (e.g. PLINK, R), facilitating analyses germane for Mendelian Randomization studies. Benchmarks for speed and power calculations for summary statistic-based Mendelian Randomization analyses are presented and compared with analytical expectation.

Availability and implementation: The simulation engine is implemented in PERL, and the associated scripts can be downloaded from github.com, and online documentation, tutorial and example datasets are available at http://coruscant.itmat.upenn.edu/mr_predictor.

Contact: bvoight@upenn.edu

Supplementary information: Supplementary derivations are available at *Bioinformatics* online.

Received on May 20, 2014; revised on August 12, 2014; accepted on August 16, 2014

1 INTRODUCTION

The discovery of genetic polymorphisms contributing to a wide range of human phenotypes has made causal inference between traits or disease endpoints feasible, by using these polymorphisms as instrumental variables (Smith and Ebrahim, 2003). The approach, dubbed *Mendelian Randomization*, can be conceptualized as a natural version of a randomized control trial, as alleles that elevate or lower a biomarker of interest assort randomly at meiosis. One popular application finds a researcher having selected multiple, unlinked polymorphic sites reproducibly associated to a biomarker of interest, which are then assayed for relationship to a disease endpoint (Voight *et al.*, 2012). Causal inference can be assessed using either individual genotypes or the equivalent estimated effects summarized from large-scale association studies; both approaches to inference are statistically valid if assumptions are satisfied (Burgess and Thompson, 2013).

As additional phenotypes and genetic associations continue to be cataloged, Monte Carlo approaches that efficiently model complex phenotypic and genotypic relationships will be useful for continued analytical and methodological progress (Brion *et al.*, 2013; Harbord *et al.*, 2013) and addressing more complex models (VanderWeele *et al.*, 2014). Unfortunately, existing software is optimized for genomics/genetic architecture applications, which is careful to incorporate linkage between sites as well as other population genetic features at computational expense, and does not easily handle complex relationships between multiple traits and genetic variants, simultaneously (Günther *et al.*, 2011; Li and Li, 2008; Peng and Amos, 2010). Thus, I report a Monte Carlo approach designed to generate samples of unrelated individuals presenting multiple phenotypes that also segregate genetic polymorphisms that contribute to them.

2 METHODS

Users begin by specifying the list of phenotypes along with associated trait parameters and covariates, pairwise correlation between intermediate traits and the impact each intermediate trait has to the outcome trait and, finally, the list of genetic variants, alleles, population frequencies and effects of each site for a given trait. This basic framework allows users to generate data from a number of models relevant for Mendelian Randomization (MR) studies.

The program initializes the variance genetic component for each trait, σ^2_G, by generating values from the genotype–phenotype relationship based on the user specification. From this, the non-genetic component, σ^2_E, for the mean, variance and covariance for all intermediate traits is determined. Values are selected such that each intermediate trait is distributed in the population with zero mean, unit variance ($\sigma^2_E + \sigma^2_G = 1$) and covariance matching the user specification. Here, I have implemented a model without interactions between genetic components, non-genetic components, or among different phenotypes/traits.

After initialization, individuals are simulated with associated genotypes and phenotypes from the model by

(1) simulating the contribution from σ^2_E for all intermediate traits from a multivariate normal distribution;

(2) simulating genetic data based on specified frequencies, assuming Hardy–Weinberg Equilibrium;

(3) incorporating the genetic effect on the phenotypes for the individual based on the simulated genotypes; an additive effect is added or

*To whom correspondence should be addressed.

subtracted from the phenotype for homozygotes, and dominance term added to the heteozygote class.

(4) simulating additional covariates, if necessary;

(5) assigning an endpoint state to the individual using contributing variables;

(6) if the number of cases and controls has been specified and has not been met, accept the individual. Otherwise, reject and repeat until the total numbers have been ascertained.

A binary endpoint is assigned to an individual using an approach analogous to the logit model described by Wu *et al.* (2011), where trait or covariates contributes log-additively to the probability of the endpoint, with user-specified background rate (α_0) for the population prevalence. Users can also control how genetic variants relate to the endpoint, i.e. through an intermediate trait or directly on the endpoint. Simulated data are returned as pedigree, map and covariate files that are compatible with the genetic analysis package, PLINK (Purcell *et al.*, 2007).

3 RESULTS

3.1 Benchmarking

I simulated 10 000 individuals for three biomarker traits—high and low density lipoprotein cholesterol levels and triglycerides (HDL-C, LDL-C and TG, respectively)—with between-trait correlation determined from epidemiological observation (Emerging Risk Factors Collaboration *et al.*, 2009), using 139 variants with 175 variant to trait relationships (Global Lipids Genetics Consortium *et al.*, 2013). On an Intel Xeon E7-4870 Core (2.40 GHz), it took 64 s, with time scaling linearly by total sample size (e.g. 20 000 samples took 131 s on average) and variants (n = 70 variants took 31 s). Ascertainment generating 2000 affected individuals (10% population prevalence), with estimates for LDL-C and TG to the endpoint from Voight *et al.* (2012), took 97 s.

3.2 Example: power calculations

In MR studies, one goal is to estimate the effect of the change in a biomarker of interest to risk of a disease endpoint. Define β as the causal effect of the biomarker on endpoint, and the null hypothesis that $\beta = 0$. By selecting a genetic instrument that has a strong, direct, and exclusive effect on the biomarker, β can be estimated via instrumental variable regression or score-based MR, generating a test statistic to evaluate if β is significantly different from zero. A significant test provides support for the hypothesis that the biomarker is causally related to the endpoint of interest (based on the genetic instrument that changes the biomarker trait), and addresses issues of confounding and reverse causality owing to random assortment of alleles at meiosis. The mean of this statistic directly relates to the power to discover a causal effect between a biomarker and the endpoint.

Analytic power determined for a single marker on a binary outcome has been described by Burgess (2014), by which the simulation engine can be compared with. From there, equations that incorporate multiple genetic instrumental variables with differences in variance explained and in samples sizes at each variant can be written (see Supplementary Note). Across L total variants, for n_i^A cases and n_i^U controls and the variance of the biomarker explained by the i-th genetic instrument (σ_i^2), the log-odds causal effect of the biomarker on disease endpoint (β) and the total variance of the biomarker explained by the genetic instrument across all loci (σ^2), the given test statistic under the alternative is distributed with unit variance and a mean of:

$$\mu = \frac{\beta}{\sqrt{\sigma^2}} \cdot \sum_{i=1}^{L} \sqrt{\frac{n_i^A n_i^U}{n_i^A + n_i^U}} \cdot \sigma_i^2 \quad (1)$$

Results presented in Table 1 demonstrate that simulation closely matches analytical expectations (Equation 1) for a single genetic instrument, for multiple genetic instrumental variables assembled into a single score-based MR analysis (Johnson, 2012) and, finally, when the effect of the instrument and/or sample sizes also vary. In all cases, the median estimates for the true, causal effect (β) matched closely to that which was simulated (data not shown).

3.3 Example: unweighted score statistics

I next compared the mean of the score-based MR test statistic, without weighting on the biomarker effect. If the biomarker effect is equal for all single nucleotide polymorphisms (SNPs) contributing to the score, the mean of the weighted and unweighted test is equal and matches analytical expectation (Equation 1). If the biomarker effects are unequal, the unweighted test has slightly lower mean (Table 1, row 8 versus 10), consistent with previous observations (Burgess and Thompson, 2013). If selected variants were subject to winner's curse, selected weights, in some cases, may be upwardly biased.

Table 1. Comparison of the mean of the MR test statistic estimated from simulation ($\hat{\mu}$) to analytical expectation (μ) for a one or more genetic instruments

n^A	n^U	Number of IVs	σ^2	β	μ	$\hat{\mu}$ (95% SEM)
10 000	10 000	1	0.02	0.2	2.00	2.02 (2.00–2.04)
10 000	10 000	1	0.02	0.3	3.00	2.99 (2.97–3.01)
10 000	10 000	1	0.03	0.3	3.67	3.66 (3.64–3.68)
10 000	20 000	1	0.03	0.3	4.24	4.26 (4.24–4.28)
10 000	10 000	5	0.006	0.3	3.67	3.68 (3.66–3.70)
10 000	10 000	50	0.0006	0.3	3.67	3.66 (3.64–3.68)
5, 10 k	5, 10 k	2	0.03	0.3	4.50	4.49 (4.47–4.51)
10 000	10 000	10	0.055[a]	0.3	4.98	4.95 (4.93–4.97)
5, 10k[b]	5, 10 k	10	0.055[a]	0.3	3.92	3.95 (3.93–3.97)
10 000	10 000	10	0.055[a]	0.3	[c]	4.74 (4.72–4.76)
10 000	10 000	50	0.0006	0.3	[d]	3.50 (3.48–3.52)

Note. IVs, number of instrumental variables (SNPs) in the score; σ^2, the variance contributed to the biomarker per variant site; β, the assumed effect of biomarker on the endpoint, SEM, standard error on the mean; k, thousand; n^A, number of cases; n^U, number of controls.

[a]Total variance explained is listed; individual loci range from 0.001, 0.002, …, 0.01.

[b]The strongest five genetic variants were evaluated in 5000 cases and controls; the weakest five in 10 000 cases and controls.

[c]Result based on an unweighted score statistic.

[d]Result where biomarker weights for 20% of IVs are biased upward (winner's curse).

In this case, the unweighted statistic has slightly better performance (3.67 if unweighted vs. 3.51 if weighted). Thus, weighting can improve the power of the test if modeled appropriately, while tests without such weighting in some case provides a degree of model robustness.

Funding: The author is indebted to the Alfred P. Sloan Foundation (BR2012-087), the American Heart Association (13SDG14330006), and the W.W. Smith Charitable Trust (H1201) who provided support for the work.

Conflict of interest: none declared.

REFERENCES

Brion,M.J. *et al.* (2013) Calculating statistical power in mendelian randomization studies. *Int. J. Epidemiol.*, **42**, 1497–1501.

Burgess,S. (2014) Sample size and power calculations in mendelian randomization with a single instrumental variable and a binary outcome. *Int. J. Epidemiol.*, **43**, 922–929.

Burgess,S. *et al.* (2013) Use of allele scores as instrumental variables for mendelian randomization. *Int. J. Epidemiol.*, **42**, 1134–1144.

Emerging Risk Factors Collaboration *et al.* (2009) Major lipids, apolipoproteins, and risk of vascular disease. *JAMA*, **302**, 1993–2000.

Global Lipids Genetics Consortium *et al.* (2013) Discovery and refinement of loci associated with lipid levels. *Nat. Genet.*, **45**, 1274–1283.

Günther,T. *et al.* (2011) Phenosim–a software to simulate phenotypes for testing in genome-wide association studies. *BMC Bioinformatics*, **12**, 265.

Harbord,R.M. *et al.* (2013) Severity of bias of a simple estimator of the causal odds ratio in mendelian randomization studies. *Stat. Med.*, **32**, 1246–1258.

Johnson,T. (2012) Efficient calculation for multi-snp genetic risk scores. *Presented at the American Society of Human Genetics Annual Meeting, San Francisco, November 6-10*. Abstract Number 1400W. Poster available at: http://cran.r-project.org/web/packages/gtx/vignettes/ashg2012.pdf.

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

Peng,B. *et al.* (2010) Forward-time simulation of realistic samples for genome-wide association studies. *BMC Bioinformatics*, **11**, 442–442.

Purcell,S. *et al.* (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Smith,G.D. and Ebrahim,S. (2003) 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.*, **32**, 1–22.

VanderWeele,T.J. *et al.* (2014) Methodological challenges in mendelian randomization. *Epidemiology*, **25**, 427–435.

Voight,B.F. *et al.* (2012) Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet*, **380**, 572–580.

Wu,M.C. *et al.* (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.