Bioinformatics, 36(15), 2020, 4374–4376
doi: 10.1093/bioinformatics/btaa549
Advance Access Publication Date: 29 May 2020
Applications Note

Genetics and population analysis

bGWAS: an R package to perform Bayesian genome wide association studies

Ninon Mounier [1,2] **and Zoltán Kutalik** [1,2,*]

[1]Department of Training, Research and Innovation, University Center for Primary Care and Public Health, Lausanne 1010, Switzerland and [2]Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on December 6, 2019; revised on May 18, 2020; editorial decision on May 19, 2020; accepted on May 25, 2020

Abstract

Summary: Increasing sample size is not the only strategy to improve discovery in Genome Wide Association Studies (GWASs) and we propose here an approach that leverages published studies of related traits to improve inference. Our Bayesian GWAS method derives informative prior effects by leveraging GWASs of related risk factors and their causal effect estimates on the focal trait using multivariable Mendelian randomization. These prior effects are combined with the observed effects to yield Bayes Factors, posterior and direct effects. The approach not only increases power, but also has the potential to dissect direct and indirect biological mechanisms.

Availability and implementation: bGWAS package is freely available under a GPL-2 License, and can be accessed, alongside with user guides and tutorials, from https://github.com/n-mounier/bGWAS.

Contact: zoltan.kutalik@unil.ch

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

1 Introduction

In the last decade, Genome Wide Association Studies (GWASs) have been widely used to identify genetic variants, usually single nucleotide polymorphisms (SNPs), associated with complex traits. These GWASs led to a large number of discoveries, helping to better understand the underlying biology of the studied traits (Visscher *et al.*, 2017). However, large sample sizes (typically > 1 million) are needed to achieve sufficient power to identify SNPs with small to moderate effects.

Besides ever-increasing sample sizes one can borrow strength from studies of related traits or risk factors (RFs). To leverage this information, several methods have already been published, such as MTAG (Turley *et al.*, 2018) or GenomicSEM (Grotzinger *et al.*, 2019) for example and we developed a Bayesian GWAS approach, first described by McDaid *et al.* (2017). The aim of our approach is to increase power by comparing the observed Z-statistics from the focal phenotype (representing association strength) to prior effects using Bayes factors (BFs) and computing the corresponding *P*-values. Prior effects are estimated from publicly available GWASs for RFs showing a significant multivariable causal effect (similar to Sanderson *et al.*, 2019) on the focal phenotype, established by Mendelian randomization (MR). Such approach has previously been used to identify new loci associated with lifespan (McDaid *et al.*, 2017; Timmers *et al.*, 2019).

Here, we present substantial improvements to the method and its implementation in an R package bGWAS. We optimized the causal effect estimation and improved the step-wise selection approach used to identify relevant RFs. We derived and implemented a fast analytical approach to accurately estimate BF *P*-values. Notably, the method now also provides posterior- and direct effect estimates (not acting through the RFs) that can be used for downstream analyses.

2 Materials and methods

The approach consists of five main steps: (i) Identification of relevant RFs, (ii) Out-of-sample estimation of prior effects, (iii) Computation of BFs and their respective *P*-values, (iv) Estimation of posterior and direct effects, (v) Extraction and visualization of the results (Supplementary Fig. S1).

2.1 Identification of relevant RFs

In the first step we identify relevant RFs to build the prior. The package currently includes 38 publicly available GWASs, which can easily be modified to include additional RFs. Using the package, all available RFs can be displayed (list_priorGWASs()) and an arbitrary subset can be selected (select_priorGWASs()). First, RFs with non-significant ($P > 0.05$) univariable causal effects are removed. Then, a step-wise selection approach applied to multivariable MR models identifies RFs that are jointly affecting the phenotype. Since Akaike information criterion (AIC)-based model comparison assumes equal

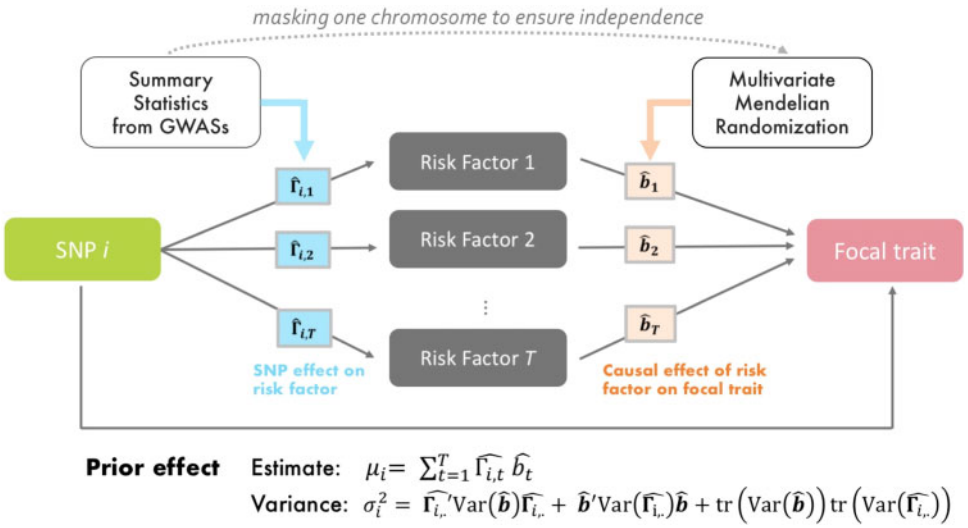

Fig. 1. Prior estimation design. For each SNP i, its prior effect on the focal trait is calculated as the product of the effect of SNP i on the RF t ($\hat{\Gamma}_{i,t}$) and the causal effect of RF t on the focal trait ($\hat{b}_t$, estimated using multivariable MR), summed over all T RFs identified in the step-wise selection approach. In our implementation, we use $\mathbf{I}_\mathbf{T}$, a T×T identity matrix, as an approximation of $\mathrm{Var}(\hat{\Gamma}_{i,\cdot})$ to estimate σ_i^2. Adapted from McDaid *et al.* (2017)

number of observations (instruments) in the compared models (which is not the case for two-sample MR), we rather implemented a *P*-value-based step-wise selection approach to identify all the RFs that have a significant multivariable causal effect on our focal phenotype (see Supplementary Section S1).

2.2 Out-of-sample estimation of prior effects

After identifying the RFs, prior effect estimates (μ) and standard errors σ are calculated for each SNP by multiplying SNP-RF effects with RF-trait causal effect estimates. To ensure that priors are independent of the observed association for SNP i, we estimate multivariable causal effects based on SNPs that do not lie on the same chromosome as SNP i (Fig. 1). Shrinking SNP-RF effects before estimating the prior leads to poorer priors (Supplementary Fig. S2).

2.3 Computation of BFs and their respective *P*-values

We use BFs to quantify the evidence in favor of the prior by comparing two competing hypotheses. Both our null and our alternative hypotheses are assuming that for a SNP i, the observed Z-statistic z_i is following a normal distribution. Under H_0, this distribution is centered on zero and has a variance of 1, whereas under H_1, the distribution is centered on μ_i and has a variance of σ_i^2 (prior parameters). The BFs can be derived in closed form (Equation 1) (Murphy, 2007).

$$\mathrm{BF}_i = \mathrm{BF}(z_i; \mu_i : \sigma_i) = \frac{L(z_i\,;\,\mu_i\,;\,1+\sigma_i^2)}{L(z_i\,;\,0\,;\,1)} \tag{1}$$

with $L(z\,;\,\mu\,;\,\sigma^2)$: the density of z under the corresponding Gaussian distribution.

Since BF alone does not readily control type I error rate, we also compute a corresponding *P*-value. The *P*-value $p_{\mathrm{BF}-i}$ represents the probability of observing any null BF (obtained for standard normal Z statistics and the same genome-wide priors) larger than the observed BF_i. We have now analytically derived the *P*-values and sample only certain percentiles of the null BF distribution (see Supplementary Section S2), yielding highly concordant *P*-value estimates with the ones from the (>8-times slower) gold-standard permutation approach (Supplementary Fig. S3).

2.4 Estimation of posterior and direct effects

The posterior effect μ'_{p-i} and posterior standard error σ_{p-i} can be easily derived for each SNP i (Equation 2) (Murphy, 2007).

$$\mu_{p-i} = \frac{\sigma_i^2}{\sigma_i^2+1}\left(\frac{\mu_i}{\sigma_i^2}+z_i\right) \quad \text{and} \quad \sigma_{p-i} = \sqrt{\frac{\sigma_i^2}{\sigma_i^2+1}} \tag{2}$$

We define the direct effect μ_{d-i} (and its standard error σ_{d-i}) as the part of the observed effect that is not mediated through the RFs and hence cannot be explained by the prior (Equation 3).

$$\mu_{d-i} = z_i - \mu_i \quad \text{and} \quad \sigma_{d-i} = \sqrt{\sigma_i^2+1} \tag{3}$$

Analogous formulae based on observed effect sizes and standard errors (instead of Z statistics) are implemented and provided in Supplementary Section S3.

2.5 Extraction and visualization of the results

We implemented dedicated functions in the bGWAS package to list, visualize and interpret the results (Supplementary Fig. S1). FDR threshold and SNP-pruning stringency can be set in the bGWAS() function to produce a final list of associated markers. Summary statistics (BFs, prior, posterior and direct effect) can be extracted from the returned *bGWAS* object using the extract_results_bGWAS() function. RFs causal effects can be obtained using extract_MRcoeffs_bGWAS() or visualized using the coefficients_plot_bGWAS() function. manhattan_plot_bGWAS() automatically creates a Manhattan plot and heatmap_bGWAS() illustrates through which RFs SNPs are exerting their (prior) effects on the focal phenotype.

3 Application to lifespan

In order to see how the improved method [implemented in the bGWAS R-package (v1.0.2)] compares to the original one (McDaid *et al.*, 2017), we applied both to the summary statistics from a GWAS on lifespan (Timmers *et al.*, 2019), which already included the latter application. A full description of the analysis and the results are available in Supplementary Section S4.

In the new analysis, we identified five RFs with a significant causal effect on lifespan (Supplementary Fig. S5): years of schooling, LDL cholesterol, diastolic blood pressure, coronary artery disease and body mass index. The priors obtained from the improved method are more informative: the squared correlation between prior and observed effects has improved to 0.377 from 0.082. In the new analysis, 28 SNPs reached genome-wide significance ($p_{\mathrm{BF}} < 5e^{-8}$). Among these variants, 15 are not identified by the conventional GWAS at the same threshold and 11 of them have never been

reported in any previous lifespan GWAS (Supplementary Table S1). Four of the seven SNPs identified based on the old BF *P*-value in Timmers *et al.* (2019) are confirmed in the new analysis, and three have low prior effects (Supplementary Table S2) due to the change in RFs. We identified nine additional loci with significant posterior effect ($p_p < 5e^{-8}$) (Supplementary Table S3) and four loci with significant direct effect ($p_d < 5e^{-8}$) (Supplementary Table S4), including the highly pleiotropic APOE locus, which might be acting on lifespan through RFs not included in the analysis (e.g. Alzheimer's disease) (Belloy *et al.*, 2019).

4 Conclusion

Leveraging information from related traits is an efficient approach to increase the power of GWAS of complex traits, which is now fully implemented in the bGWAS R package. Through an application to lifespan GWAS, we have demonstrated that this approach could lead to meaningful new discoveries in lifespan genetics and dissect direct from indirect mechanisms.

Acknowledgements

The authors thank Nicola Pirastu for the helpful discussions, as well as Jonathan Sulc and Sina Rüeger for the helpful comments on the documentation.

Funding

This work was supported by the Swiss National Science Foundation [31003A-143914] [310030-189147].

Conflict of Interest: none declared.

References

Belloy,M.E. *et al.* (2019) A quarter century of APOE and Alzheimer's disease: progress to date and the path forward. *Neuron*, **101**, 820–838.

Grotzinger,A.D. *et al.* (2019) Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.*, **3**, 513–525.

McDaid,A.F. *et al.* (2017) Bayesian association scan reveals loci associated with human lifespan and linked biomarkers. *Nat. Commun.*, **8**, 15842.

Murphy,K.P. (2007) Conjugate Bayesian analysis of the Gaussian distribution. *Technical report*.

Sanderson,E. *et al.* (2019) An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. *Int. J. Epidemiol.*, **48**, 713–727.

Timmers,P.R. *et al.* (2019) Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *eLife*, **8**, e39856.

Turley,P. *et al.* (2018) Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.*, **50**, 229–237.

Visscher,P.M. *et al.* (2017) 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.*, **101**, 5–22.