Leveraging correlated risks to increase power in Genome-Wide Association Studies

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

1: *University Center for Primary Care and Public Health, Lausanne, Switzerland*
2: *Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland*

unisanté
Centre universitaire de médecine générale
et santé publique · Lausanne

SIB
Swiss Institute of
Bioinformatics

ninon.mounier@unil.ch

@Nin0nM

http://wp.unil.ch/sgg/

Genome-Wide Association Studies (GWASs) are nowadays often conducted in more than 1 million samples. Improving discovery by further increasing study sizes is not the only strategy.

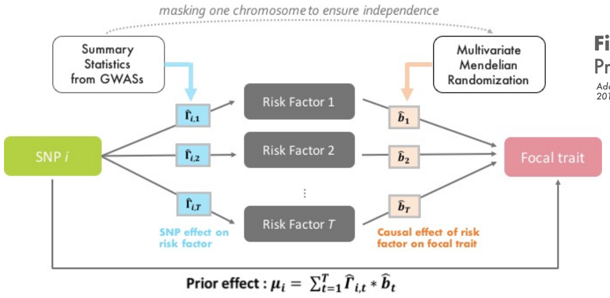
Leveraging information from published studies of related traits can improve inference. To this end, we developed a Bayesian GWAS approach that builds informative priors for each single nucleotide polymorphism (SNP) using GWASs of related risk factors (RFs).

Bioinformatics, *in press*

Method

Estimation of the prior:

Summary statistics of GWASs of RFs are used to estimate causal effects of these RFs on a focal trait and compute prior effects (Fig. 1).


masking one chromosome to ensure independence
Summary Statistics from GWASs
Multivariate Mendelian Randomization
SNP *i*
$\hat{\Gamma}_{i,1}$
$\hat{\Gamma}_{i,2}$
$\hat{\Gamma}_{i,T}$
Risk Factor 1
Risk Factor 2
Risk Factor *T*
$\hat{b}_1$
$\hat{b}_2$
$\hat{b}_T$
Focal trait
SNP effect on risk factor
Causal effect of risk factor on focal trait

Prior effect : $\mu_i = \sum_{t=1}^{T} \hat{\Gamma}_{i,t} * \hat{b}_t$

Figure 1:
Prior estimation design.
Adapted from McDaid et al, 2017

Two different Mendelian Randomisation steps:

1) Identify risk factors significantly affecting the focal trait
Multivariable MR – stepwise selection based on p-values
2) Estimate the prior effect, using the risk factors identified in 1)
Multivariable MR – masking one chromosome to ensure independence

For each SNP *I* :
- observed effect : z_i
- prior effect : μ_i (prior mean) and σ_i^2 (prior variance)

3 different ways of comparing/combining observed and prior effects

✹ Comparison using Bayes Factors (BFs) :

Boost signal to identify SNPs acting through the risk factors

Null Hypothesis : $z_i \sim N(0,1)$
Alternative Hypothesis : $z_i \sim N(\mu_i, \sigma_i^2)$

$$BF_i = \frac{L(z_i; \mu_i; 1 + \sigma_i^2)}{L(z_i; 0; 1)}$$

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

Estimation of the p-values (p_{BF}) corresponding to observed BFs:

« *Probablity of observing a null BF* (*obtained from a GWAS for a permuted outcome with the same priors*) *larger than the observed* BF_i »

Derivation of an analytical formula to estimate $p_{BF,i}$ for each SNP
→ use an approximation to speed up estimation, and re-estimate values near significance threshold using the full formula.

✹ Estimation of posterior effects:

Combine information from prior and observed effects

Can be used for downstream analyses as any other GWAS Summary Statistics

$$\mu_{p-i} = \frac{\sigma_i^2}{\sigma_i^2 + 1}\left(\frac{\mu_i}{\sigma_i^2} + z_i\right) \qquad \sigma_{p-i}^2 = \frac{\sigma_i^2}{\sigma_i^2 + 1}$$

✹ Estimation of direct effects:

Identify SNPs with effects not mediated through the RFs

Can be used for downstream analyses as any other GWAS Summary Statistics

$$\mu_{d-i} = z_i - \mu_i \qquad \sigma_{d-i}^2 = \sigma_i^2 + 1$$

Note: all effects are estimated using Z-statistics z but can be rescaled to be comparable to effect sizes β.

Application to Lifespan (1/2)

Analysis performed using bGWAS R package, version 1.0.2

GWAS Summary Statistics : analysis of more than1 million parental lifespans (data from Timmers *et al*, 2019)

➢ Estimation of the prior:

5 out of 38 RFs were selected in the stepwise selection procedure and used to create the prior (Fig. 2).

We used the correlation between observed and prior effects to assess the quality of the prior estimated, using only moderately associated* SNPs : 0.377

* (observed GWAS p-value < 0.001)

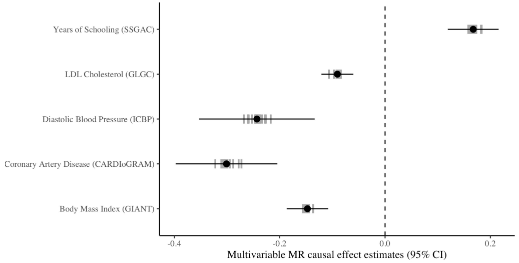


Figure 2: Causal effect estimates of the 5 RFs affecting lifespan

➢ Identification of 28 variants associated with lifespan (p_{BF} < 5e-8)

➢ 15 loci missed by the conventional GWAS (highlighted in red on Fig. 3)

- 4 already significant Timmers *et al* bGWAS results
- 11 new loci

Among these 15 variants, 8 are associated with at least one of the RFs (from the summary statistics used to create the prior) (Fig. 4).
Using the GWAS Catalog to further investigate the remaining variants and their neighbouring regions, we found that in more recent studies variants near IL6R have been associated with coronary artery disease, and variants near SLC4A7, PINX1 and TNKS have been associated with Diastolic Blood Pressure. The other loci (near BNC2, CUX2 and PDE3A) have not been associated with any of the risk factors, and are likely to be acting on lifespan through moderate effects on several risk factors (pleiotropic effects).

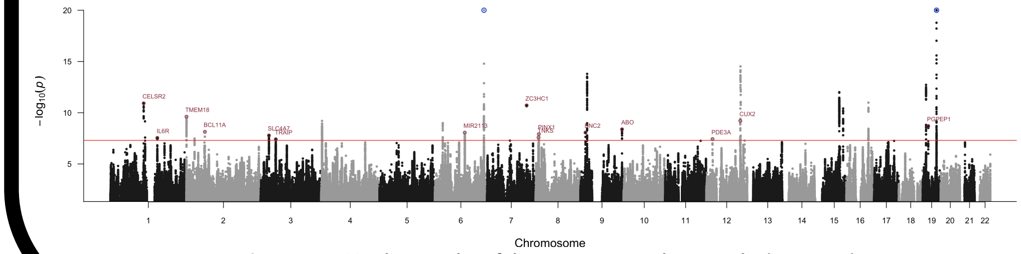
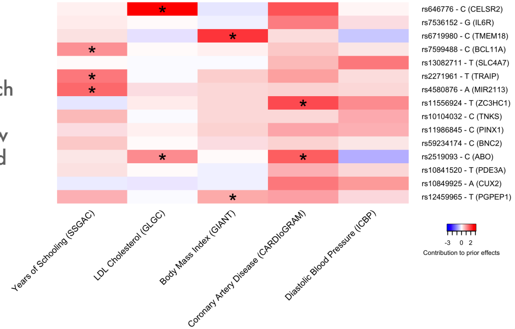


Figure 3: Manhattan Plot of the Bayesian Analysis results (using p_{BF})

Figure 4: Heatmap representing the contribution of each risk factor to the prior effects of new hits (alleles aligned to be life-lengthening)



Application to Lifespan (2/2)

Analysis performed using bGWAS R package, version 1.0.2

➤ Identification of 28 variants associated with lifespan $(p_p < 5^{e-8})$ using posterior effects

➢ 9 loci missed by the conventional GWAS and Bayes Factors results

rsid	at or near	chr	pos	alt/ref	z	μ_p	σ_p	p_p
rs13086611	USP4	3	49385417	A/T	-3.019	-3.043	0.503	1.46e-09
rs13130484	GNPDA2	4	45175691	T/C	-3.161	-3.270	0.487	1.86e-11
rs34809719	MAD1L1	7	2028968	T/G	3.227	2.774	0.483	8.97e-09
rs964184	ZPR1	11	116648917	C/G	3.791	3.165	0.497	1.96e-10
rs1183910	HNF1A	12	121420807	A/G	-3.676	-2.915	0.481	1.33e-09
rs8049439	ATXN2L	16	28837515	T/C	2.809	2.949	0.480	8.05e-10
rs1421085	FTO	16	53800954	T/C	3.550	3.760	0.612	7.89e-10
rs999474	UBE2Z	17	46987665	A/G	3.502	2.674	0.469	1.22e-08
rs303757	RMC1	18	21078716	T/G	2.793	2.983	0.478	4.37e-10

All these variants are associated with at least one of the RFs (from the summary statistics used to create the prior).

➤ Identification of 4 variants having significant direct effects $(p_d < 5^{e-8})$

Amongst these variants, 3 are likely to act through RFs that were not included in our subset of RFs and therefore not used to create the prior :
APOE locus : Alzheimer's disease
HYKK locus : smoking, pulmonary diseases and cancers
LPA : lipoprotein levels for example could also influence lifespan.
The variant near RAD52, however has a quite strong effect on lifespan in the conventional GWAS but a small prior effect in the other direction. There here is no strong association reported for this region, and the discrepancy between observed and prior effects could be due to some direct effect on lifespan.

rsid	at or near	chr	pos	alt/ref	z	μ_d	σ_d	p_d
rs55730499	LPA	6	161005610	T/C	-10.258	-8.295	1.166	1.13e-12
rs7307680	RAD52	12	1052488	A/G	-5.286	-6.196	1.134	4.64e-08
rs8042849	HYKK	15	78817929	T/C	10.659	10.395	1.118	1.41e-20
rs429358	APOE	19	45411941	T/C	19.328	17.473	1.228	5.79e-46

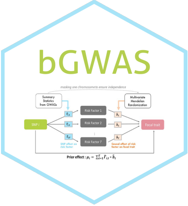
R Package

Available on github: https://github.com/n-mounier/bGWAS

Only input required : GWAS Summary Statistics - Set of 38 RFs available to create the prior

`bGWAS()`: main function, performs the Bayesian GWAS, identifies relevant RFs, estimates prior effect, BFs, p-values, posterior and direct effects

+ functions facilitating results extraction and visualisation



References:

McDaid, A. F. *et al.* (2017).
Bayesian association scan reveals loci associated with human lifespan and linked biomarkers. *Nature Communications*
doi.org/10.1038/ncomms15842

Timmers, P. R. *et al.* (2019).
Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *eLife*
doi.org/10.7554/eLife.39856

FNSNF

SystemsX.ch
The Swiss Initiative in Systems Biology