

Statistical genetics journal club

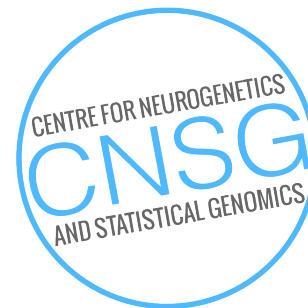


Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis

Eli A Stahl^{1-3*}, Daniel Wegmann⁴, Gosia Trynka⁵, Javier Gutierrez-Achury⁵, Ron Do^{2,6}, Benjamin F Voight⁷, Peter Kraft⁸, Robert Chen¹⁻³, Henrik J Kallberg⁹, Fina A S Kurreeman¹⁻³, Diabetes Genetics Replication and Meta-analysis Consortium¹⁰, Myocardial Infarction Genetics Consortium¹⁰, Sekar Kathiresan^{2,6}, Cisca Wijmenga⁵, Peter K Gregersen¹¹, Lars Alfredsson⁹, Katherine A Siminovitch¹², Jane Worthington¹³, Paul I W de Bakker^{2,3,14,15}, Soumya Raychaudhuri^{1-3,16} & Robert M Plenge^{1-3,16}

10/05/2016

Baptiste Couvy-Duchesne



Chabris et al., 2015

The fourth law of behaviour genetics:
“A typical human behavioral trait is associated with very many genetic variants, each of which accounts for a very small percentage of the behavioral Variability”

Result for SCZ calculated
in Ripke et al., 2013

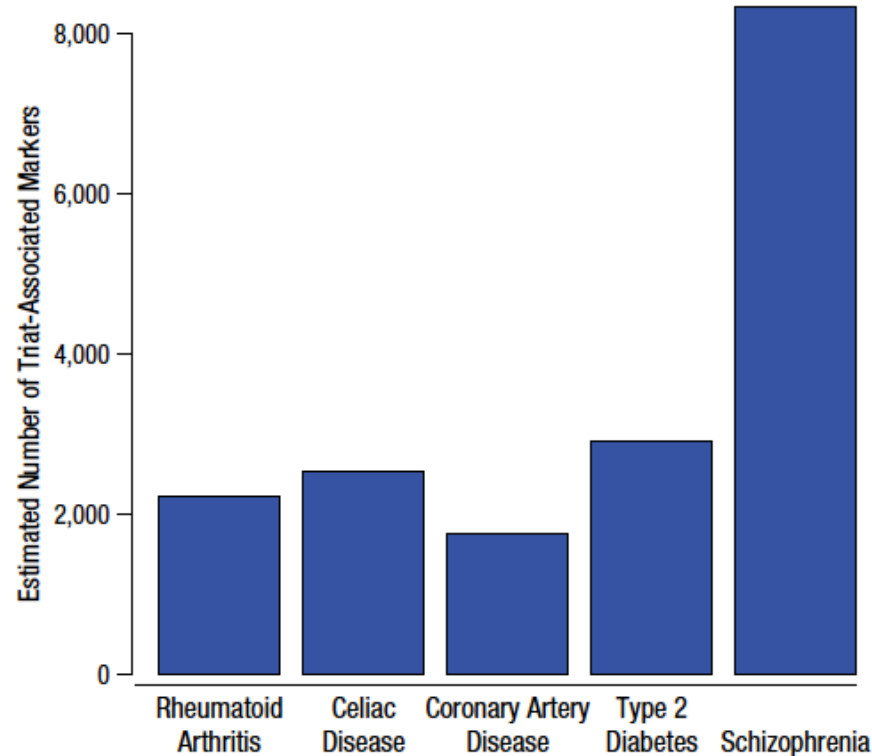


Fig. 2. Estimated total number of single-nucleotide polymorphisms associated with five disease phenotypes based on results of genome-wide association studies (data drawn from Ripke et al., 2013; see that paper and Stahl et al., 2012, for further methodological details).

Abstract (Stahl et al., 2012)

The genetic architectures of common, complex diseases are largely uncharacterized. We modeled the genetic architecture underlying genome-wide association study (GWAS) data for rheumatoid arthritis and developed a new method using polygenic risk-score analyses to infer the total liability-scale variance explained by associated GWAS SNPs. Using this method, we estimated that, together, thousands of SNPs from rheumatoid arthritis GWAS explain an additional 20% of disease risk (excluding known associated loci). We further tested this method on datasets for three additional diseases and obtained comparable estimates for celiac disease (43% excluding the major histocompatibility complex), myocardial infarction and coronary artery disease (48%) and type 2 diabetes (49%). Our results are consistent with simulated genetic models in which hundreds of associated loci harbor common causal variants and a smaller number of loci harbor multiple rare causal variants. These analyses suggest that GWAS will continue to be highly productive for the discovery of additional susceptibility loci for common diseases.

Polygenic Risk score analysis

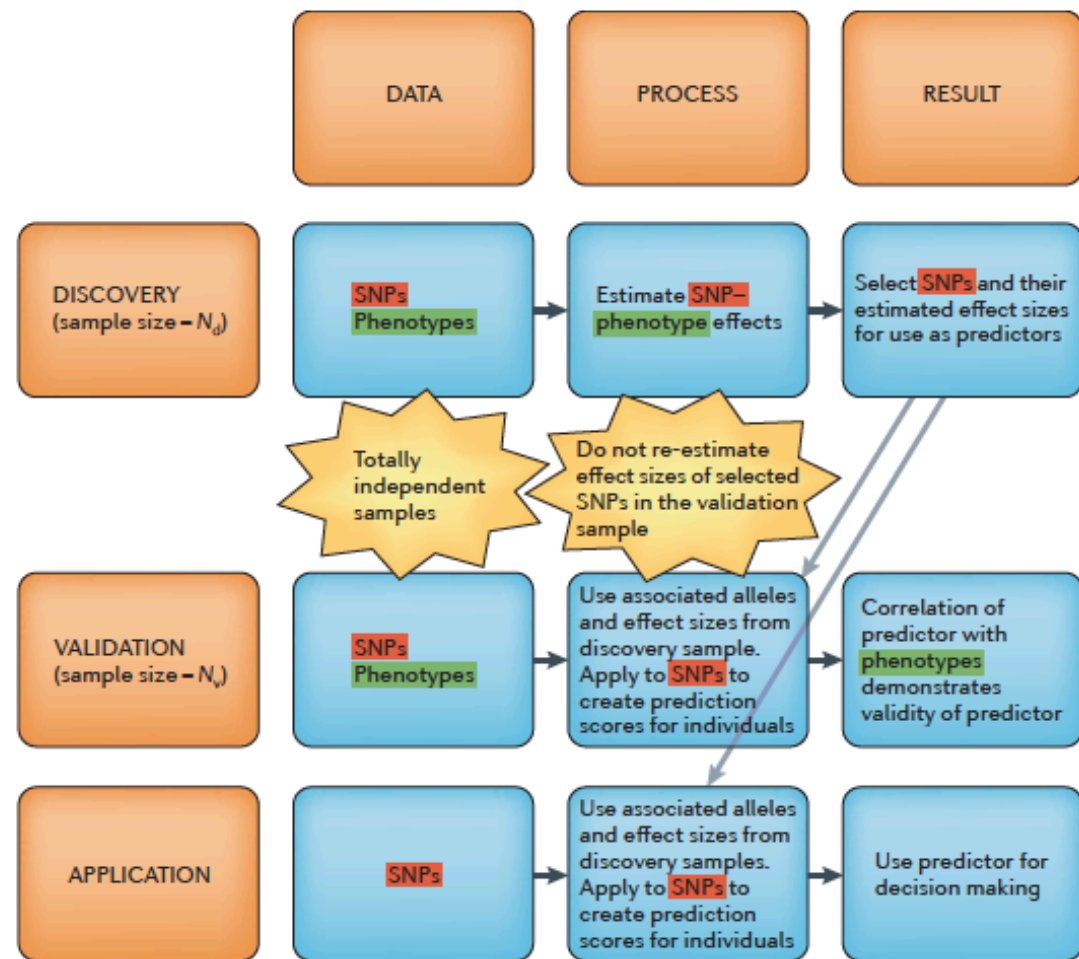


Figure from Wray et al., 2013

Figure 1 | Flowchart of SNP-based prediction analysis. There are three stages for the development of a risk predictor: discovery, validation and application. At each stage, data are needed as an input, and a process is applied to the data and a result is generated. At the application stage, effect sizes estimated from combined discovery and validation samples can be used. SNP, single-nucleotide polymorphism.

Method

1. Identify a Discovery sample with genome-wide association analysis summary statistics.
2. Identify a Target sample with genome-wide genotypes. The Target sample should not include individuals closely related to those in the Discovery sample. Results can be inflated if there is overlap between samples.
3. Determine the list of SNPs in common between Discovery and Target samples.
4. Construct a clumped SNP list: association p -value informed removal of correlated SNPs, e.g. LD threshold of $r^2 < .2$ across 500 kb. (e.g. in the program PLINK (Purcell et al., 2007): `-clump-p1 1 -clump-p2 1 -clump-r2 0.2 -clump-kb 500`).
5. Limit the SNP list to those with association p -value less than a defined threshold (often several thresholds are considered, i.e. $<.00001$, $.0001$, $.001$, $.01$, $.1$, $.2$, $.3$, etc.).
6. Generate genomic profile scores in the target sample: e.g. sum of risk alleles weighted by Discovery sample effect size, for example, $\log(\text{odds ratio})$. (e.g. in PLINK: `-score`).
7. Regression analysis: y = phenotype, x = profile score. Compare variance explained from the full model (with x) compared to a reduced model (covariates only). Check the sign of the regression coefficient to determine if the relationship between y and x is in the expected direction.

Outcomes

1. Measure of association between Discovery and Target sample (R^2 , Nagelkerke's R^2 (NR^2), area under the receiver operating curve, proportion of variance explained on liability scale, see Lee, Goddard, et al., 2012)

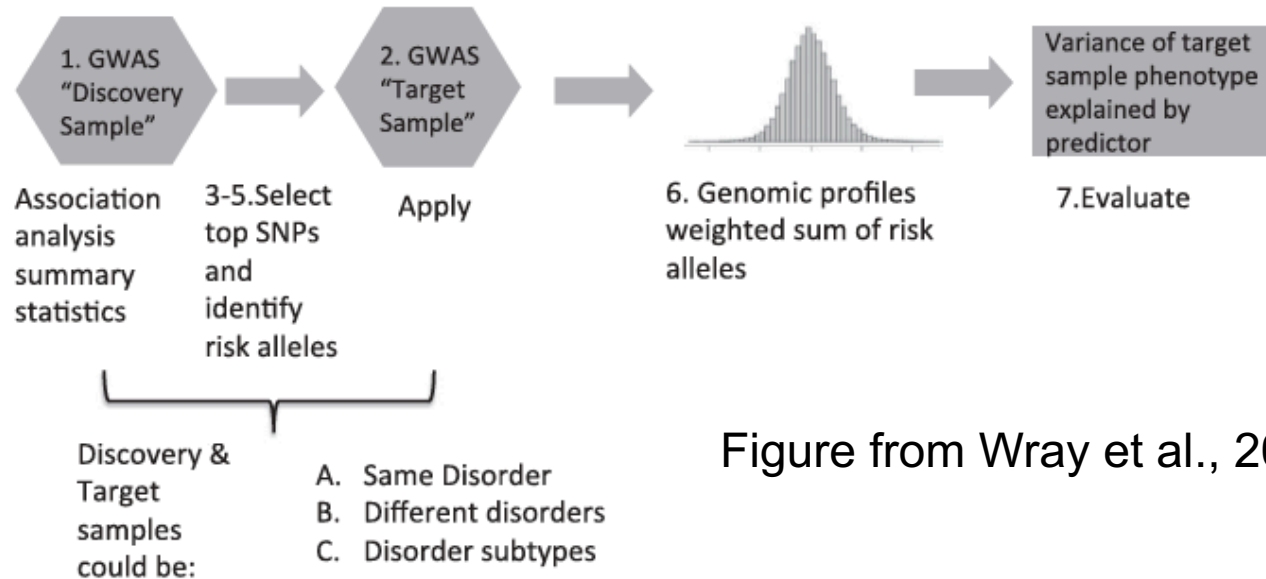


Figure from Wray et al., 2014

Discovery and target samples

Table 1 Common disease GWAS data

Disease	Discovery and test data (cohorts)	Cases	Controls	Total	SNP platform	
					<i>N</i> after QC	<i>N</i> after LD pruning
Rheumatoid arthritis	Discovery (5)	3,964	12,052	10,565	HapMap2	
					2,100,000	84,000
Celiac disease	Test (WTCCC)	1,521	10,557	5,318		
	Discovery (3)	2,091	3,218	4,776	Illumina 550K	
Early onset MI/CAD					503,000	91,000
	Test (UK2)	1,849	4,936	5,380		
T2D mellitus	Discovery (MIGEN)	2,967	3,075	6,040	HapMap2	
					1,800,000	90,000
	Test (WTCCC)	1,926	2,935	4,652		
	Discovery (7)	6,206	8,713	17,427	HapMap2	
					2,000,000	76,000
	Test (WTCCC)	1,924	2,938	4,651		

WTCCC, Wellcome Trust Case Control Consortium; MIGEN, Myocardial Infarction Genetics Consortium; UK2, Stage 1 Collection 2 from reference 19; QC, quality control.

PRS analysis

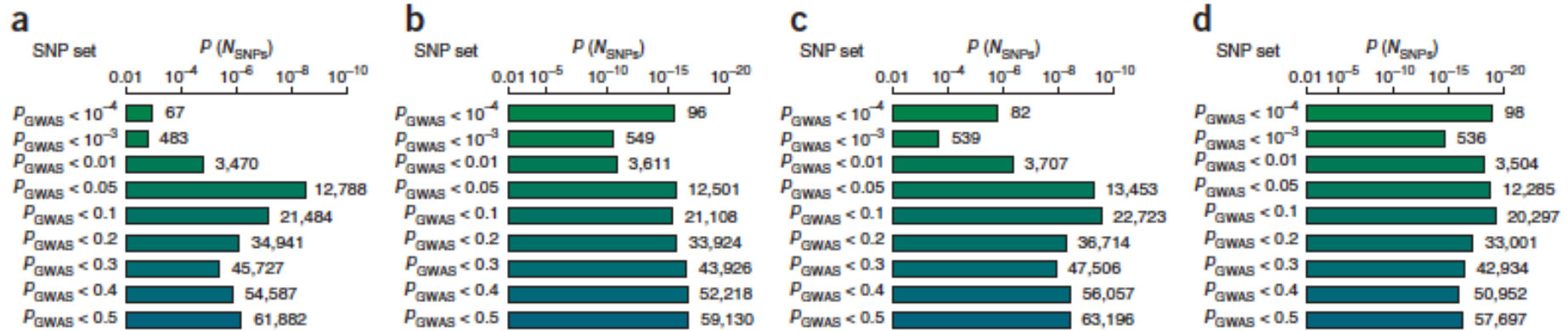


Figure 1 Association of polygenic risk scores with common disease case-control status in independent validation datasets. Association P values (\log_{10} scale) are plotted, with the number of SNPs used for the calculation of the risk scores shown at right, for SNP sets based on P_{GWAS} thresholds ranging from 10^{-4} (top, green) to 0.5 (bottom, blue). (a) Rheumatoid arthritis (all known risk loci removed). (b) Celiac disease (with the extended MHC region removed). (c) Myocardial infarction (discovery data) and coronary artery disease (test data). (d) T2D.

Conclusion: many snps across the genome together explain a significant portion of the trait variance

However, PRS explain at most 0.3, 0.4, 0.5 and 0.7% of variance (Nagelkerke's R^2).
 PRS are made up of unknown number of true-positive SNPs, as well as unassociated SNPs (noise)

Bayesian analysis

Aim: estimate the number of **associated SNPs** and the **total liability scale variance explained (heritability)**

As a side product of the analyses they also estimate the distribution of genotypic relative risk and of MAF

For this they use: “bayesian computation with **rejection sampling** and **general linear model post-sampling adjustment (ABC-GLM)**”

Model

Consider a model M , creating D data determined by parameters θ (from a **bounded** space), with joint prior density denoted $\pi(\theta)$

D: SNP and phenotype data

θ : (N_{snp} , V_{tot} , β_v , α_{raf} , β_{raf})

N_{snp} : number of causal SNPs

V_{tot} : total variance explained

SNP-wise variance explained $\sim \beta(1, \beta_v)$

Risk allele frequency (RAF) $\sim \beta(\alpha_{\text{raf}}, \beta_{\text{raf}})$

Hypotheses/Priors:

N_{snps} between 10-10,000

V_{tot} between 0.01-0.99

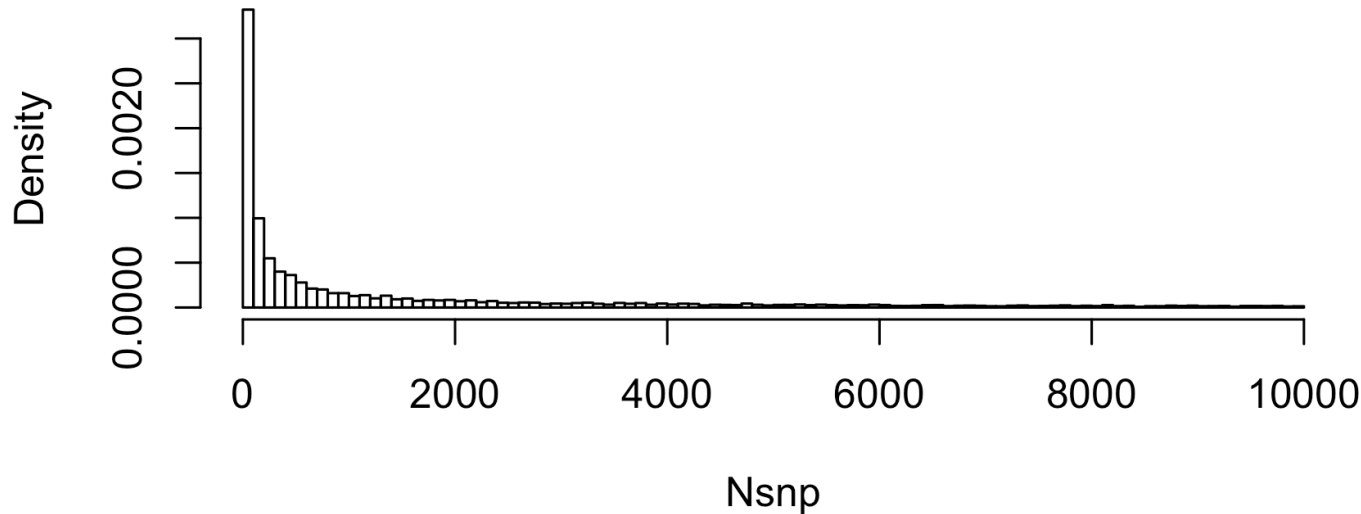
β_v between 1-5

α_{raf} and β_{raf} between 0.5-10

Rejection sampling approach

1) Draw θ from **prior distributions**

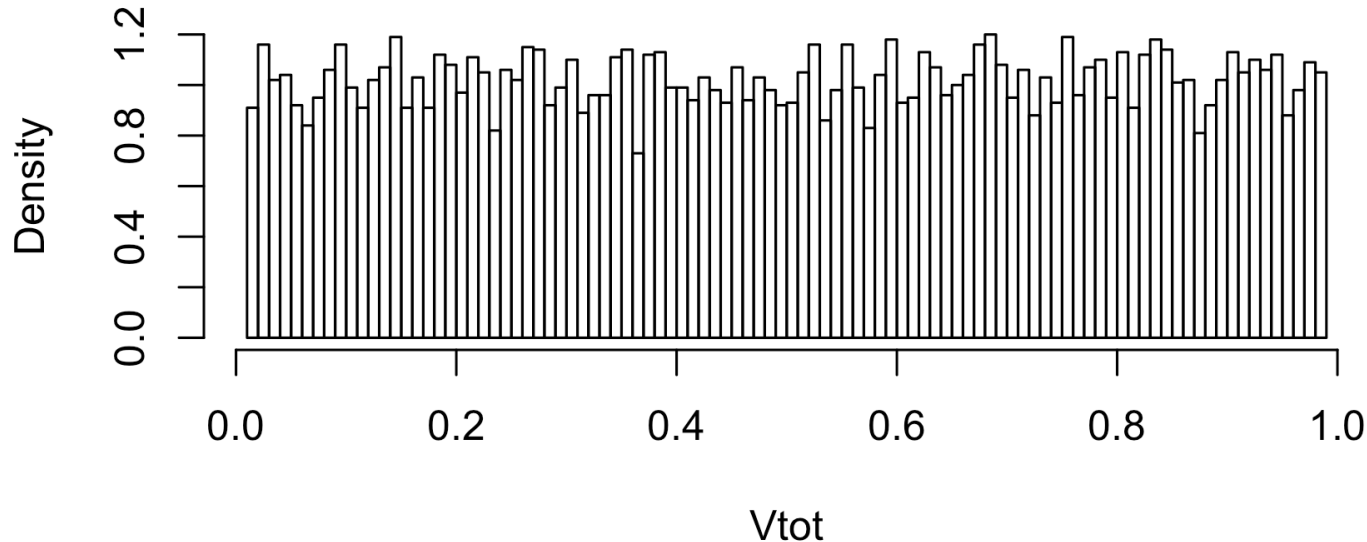
- $\text{Log N}_{\text{snr}} \sim \text{uniform distribution}$
- $V_{\text{tot}}, \beta_v, \alpha_{\text{raf}}$ and $\beta_{\text{raf}} \sim \text{uniform distributions}$



Rejection sampling approach

1) Draw θ from **prior distributions**

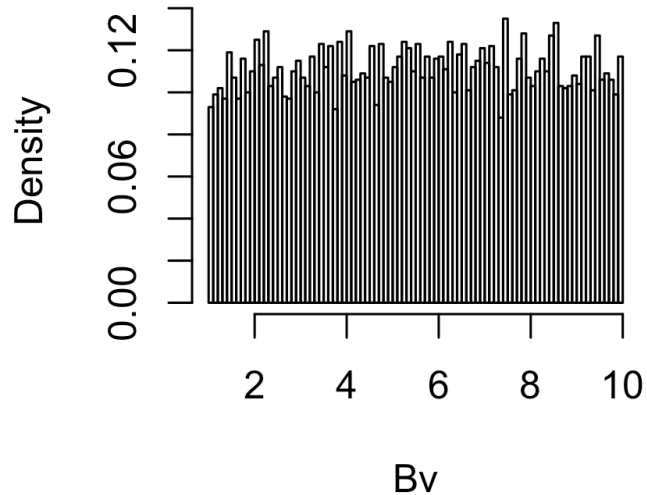
- $\text{Log N}_{\text{sn}} \sim \text{uniform distribution}$
- $V_{\text{tot}}, \beta_v, \alpha_{\text{raf}}$ and $\beta_{\text{raf}} \sim \text{uniform distributions}$



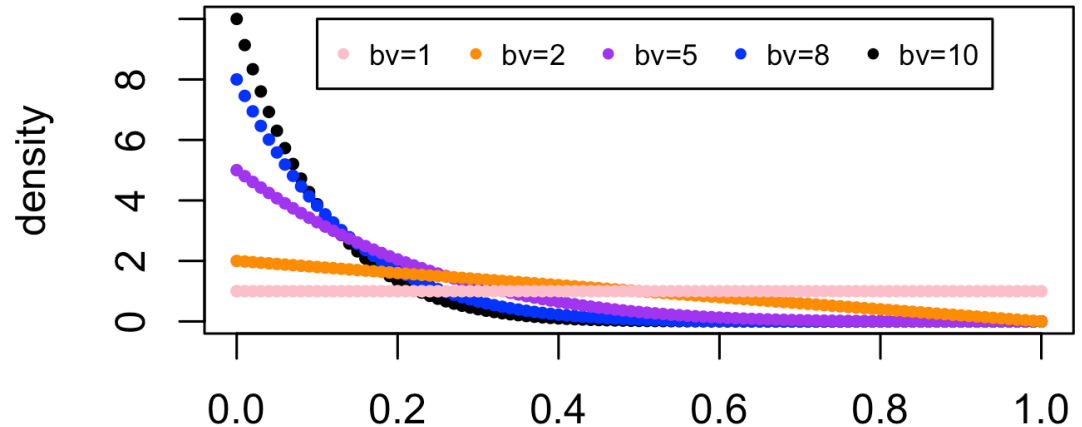
Rejection sampling approach

1) Draw θ from **prior distributions**

- $\text{Log N}_{\text{snr}} \sim \text{uniform distribution}$
- $V_{\text{tot}}, \beta_v, \alpha_{\text{raf}}$ and $\beta_{\text{raf}} \sim \text{uniform distributions}$



distribution of effect sizes

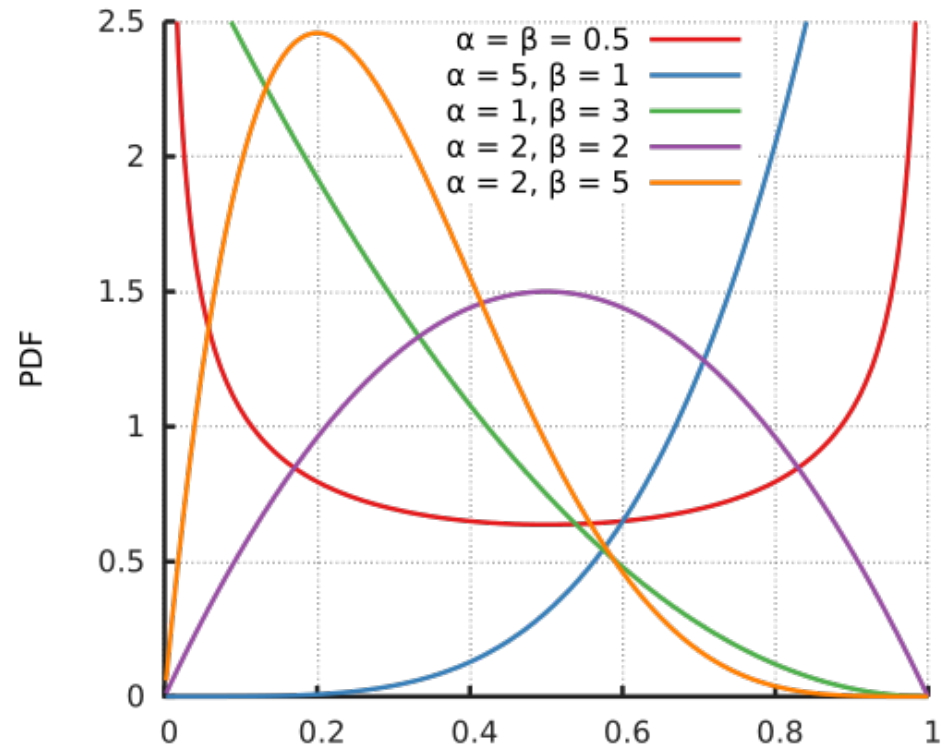


Rejection sampling approach

1) Draw θ from **prior distributions**


- $\text{Log N}_{\text{snp}} \sim \text{uniform distribution}$
- $V_{\text{tot}}, \beta_v, \alpha_{\text{raf}}$ and $\beta_{\text{raf}} \sim \text{uniform distributions}$

α_{raf} and $\beta_{\text{raf}} \sim U(0.5, 10)$ lead to a wide range of Risk Allele Frequency distribution



Rejection sampling approach

- 1) Draw θ from **prior distributions**
 - $\text{Log N}_{\text{snp}} \sim \text{uniform distribution}$
 - $V_{\text{tot}}, \beta_v, \alpha_{\text{raf}}$ and $\beta_{\text{raf}} \sim \text{uniform distributions}$
- 2) Simulate GWAS discovery and target samples from θ
- 3) Calculate R^2 (PRS analysis similar as before)
- 4) Reject the set of θ values if R^2 too different from observed one



Repeat a million time

Objective: identify plausible θ , knowing the data

=> Posterior probability of θ knowing the data

Bayesian Computation and Model Selection in
Population Genetics

Christoph Leuenberger,^{*†} Daniel Wegmann,^{*‡} Laurent Excoffier[‡]

Abstract

Until recently, the use of Bayesian inference in population genetics was limited to a few cases because for many realistic population genetic models the likelihood function cannot be calculated analytically. The situation changed

Rejection sampling approach

Bayes formula:

$$\pi(\theta \mid D) = c \cdot F_M(D \mid \theta) \cdot \pi(\theta)$$

$\pi(\theta \mid D)$: **posterior probability** of θ knowing the data D (**quantity of interest**)

c a constant

$F_M(D \mid \theta)$ **the likelihood of the data** (cannot be calculated analytically for many realistic population genetic models => stochastic simulation or rejection sampling)

$\pi(\theta)$ **the joint prior distribution**

Here, we substitute D by a **set of statistics “s” not necessarily sufficient** (R^2 is not)

And rejection sampling still provides a valid estimation of the posterior probability

$$\pi(\theta \mid s) = c \cdot F_\epsilon(s \mid \theta) \cdot \pi_\epsilon(\theta)$$

with **$\text{dist}(s, s_{\text{obs}}) < \epsilon$ small**, so that we can assume the likelihood to be constant around s_{obs} (on the ϵ ball)

Approximate Bayesian computation

Problem 1:

When s is a vector large dimension (e.g. here 24)
 $\text{dist}(s, \text{sobs}) < \epsilon$ with ϵ small, leads to a prohibitively low acceptance rate which
would require millions of simulations

To overcome this problem, ϵ is chosen rather large and a variant of the Metropolis-Hastings algorithm is used to sample directly from the truncated prior ($\pi_\epsilon(\theta)$)

Approximate Bayesian Computation

ABC – General Linear Model

Problem 2:

Now ε is rather large and the hypothesis that the likelihood is constant over the ε -ball (around s_{obs}) is pretty rough

To overcome this problem: **post sampling adjustment using General Linear Model** (\neq Generalised linear models)

Assume the summary statistics created by the truncated model's likelihood $F_{\varepsilon}(s \mid \theta)$ to satisfy:

$$s \mid \theta = C \theta + c_0 + e$$

Practical validation / Method Improvement

The method was tested on simulated datasets with

Nsnp=10, 100 or 1000

Vtot=0.1 or 0.2

RAF = 0.01 pr 0.5

Small number of iterations

Posterior density credible intervals all contained the true values for Nsnp and Vtot
(results not shown)

The authors extended the ABC-GLM to estimate joint posterior densities of Nsnp and Vtot

Results

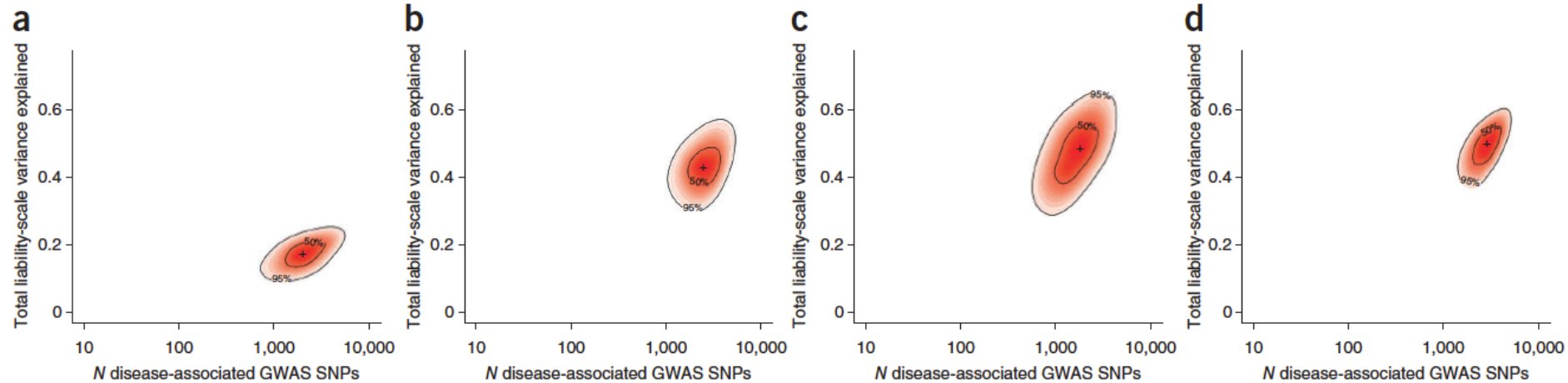


Figure 2 Posterior probability densities of the number of associated SNPs and the total liability-scale variance explained for the Bayesian analysis of the polygenic analysis results. N_{SNPs} are shown on the log₁₀ scale on the x axis, and V_{tot} values are shown on the y axis. The heat map colors represent the probability density height, with darker colors indicating higher density. Contour lines show the highest posterior density and the 50%, 90% and 95% credible regions. (a) Rheumatoid arthritis (with all known risk loci removed). (b) Celiac disease (with the extended MHC region removed). (c) MI/CAD. (d) T2D.

Results

Table 2 Comparison of results of different polygenic methods across diseases

Disease	Prevalence (%)	Family based heritability ^a	Caused by common GWAS SNPs		
			LMM-based heritability (s.e.)	Polygenic modeling and Bayesian inference	
				Total variance explained (50% CI)	<i>N</i> SNPs (50% CI)
Rheumatoid arthritis	1	0.53–0.68 (–0.13 MHC) ^b	0.32 (0.037)	0.18 (0.15–0.20) (+0.04 known non-MHC) ^b	2,231 (1,588–2,740)
Celiac disease	1	0.5–0.87 (–0.35 MHC) ^b	0.33 (0.042)	0.44 (0.40–0.47)	2,550 (1,907–3,061)
MI/CAD	6	0.3–0.63	0.41 (0.067)	0.48 (0.43–0.54)	1,766 (1,215–2,125)
T2D mellitus	8	0.26–0.69	0.51 (0.065)	0.49 (0.46–0.53)	2,919 (2,335–3,442)

^aFamily based heritability estimates were taken from previous data for rheumatoid arthritis^{27,28}, celiac disease^{18,30}, MI/CAD^{31,32} and T2D^{33,34}. ^bWe excluded some loci in certain analyses: although the family based heritability estimates are based on the whole genome, the extended MHC region was removed from the common GWAS SNP analyses for rheumatoid arthritis and celiac disease, and validated non-MHC loci were further removed from the polygenic modeling analysis of the rheumatoid arthritis GWAS data. 50% CI, 50% credible interval; s.e., standard error.

Results

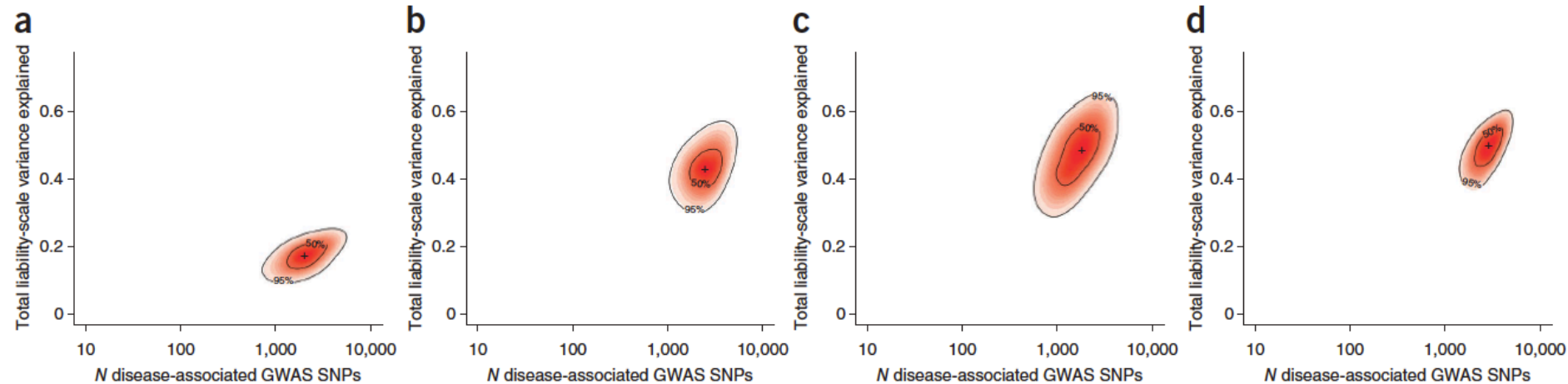


Figure 2 Posterior probability densities of the number of associated SNPs and the total liability-scale variance explained for the Bayesian analysis of the polygenic analysis results. N_{SNPs} are shown on the log₁₀ scale on the x axis, and V_{tot} values are shown on the y axis. The heat map colors represent the probability density height, with darker colors indicating higher density. Contour lines show the highest posterior density and the 50%, 90% and 95% credible regions. (a) Rheumatoid arthritis (with all known risk loci removed). (b) Celiac disease (with the extended MHC region removed). (c) MI/CAD. (d) T2D.

Nsnps RH = 2,231 [800-6,000]

Nsnps CD = 2,550 [1,000-5,000]

Nsnps MI/CAD = 1,766 [500-4,000]

Nsnps T2D = 2,919 [1,050-5,000]

Chabris et al., 2015

The fourth law of behaviour genetics:
“A typical human behavioral trait is associated with very many genetic variants, each of which accounts for a very small percentage of the behavioral Variability”

Result for SCZ calculated
in Ripke et al., 2013

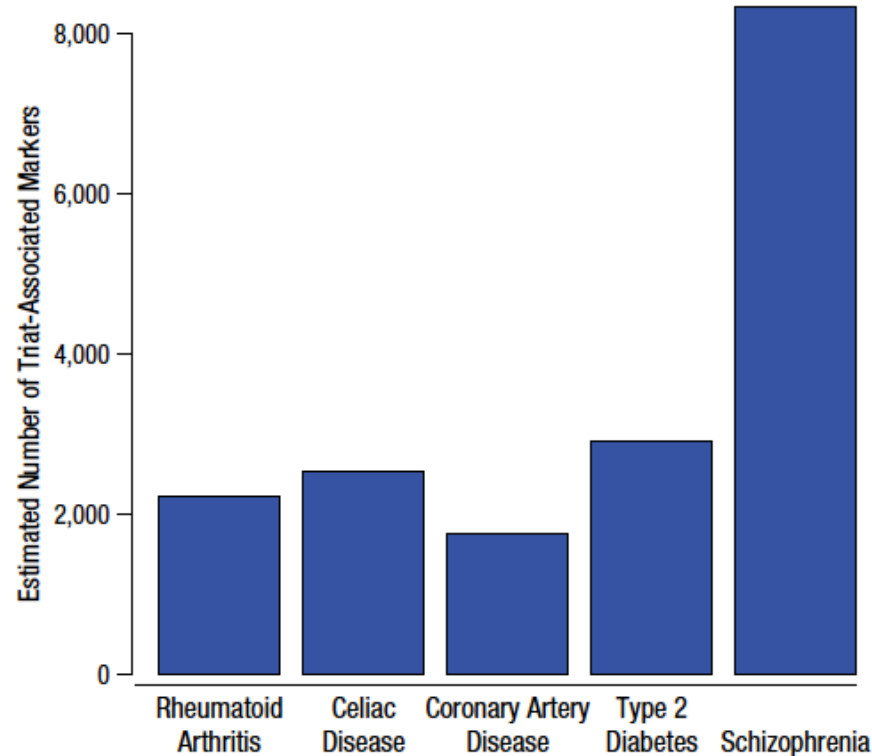
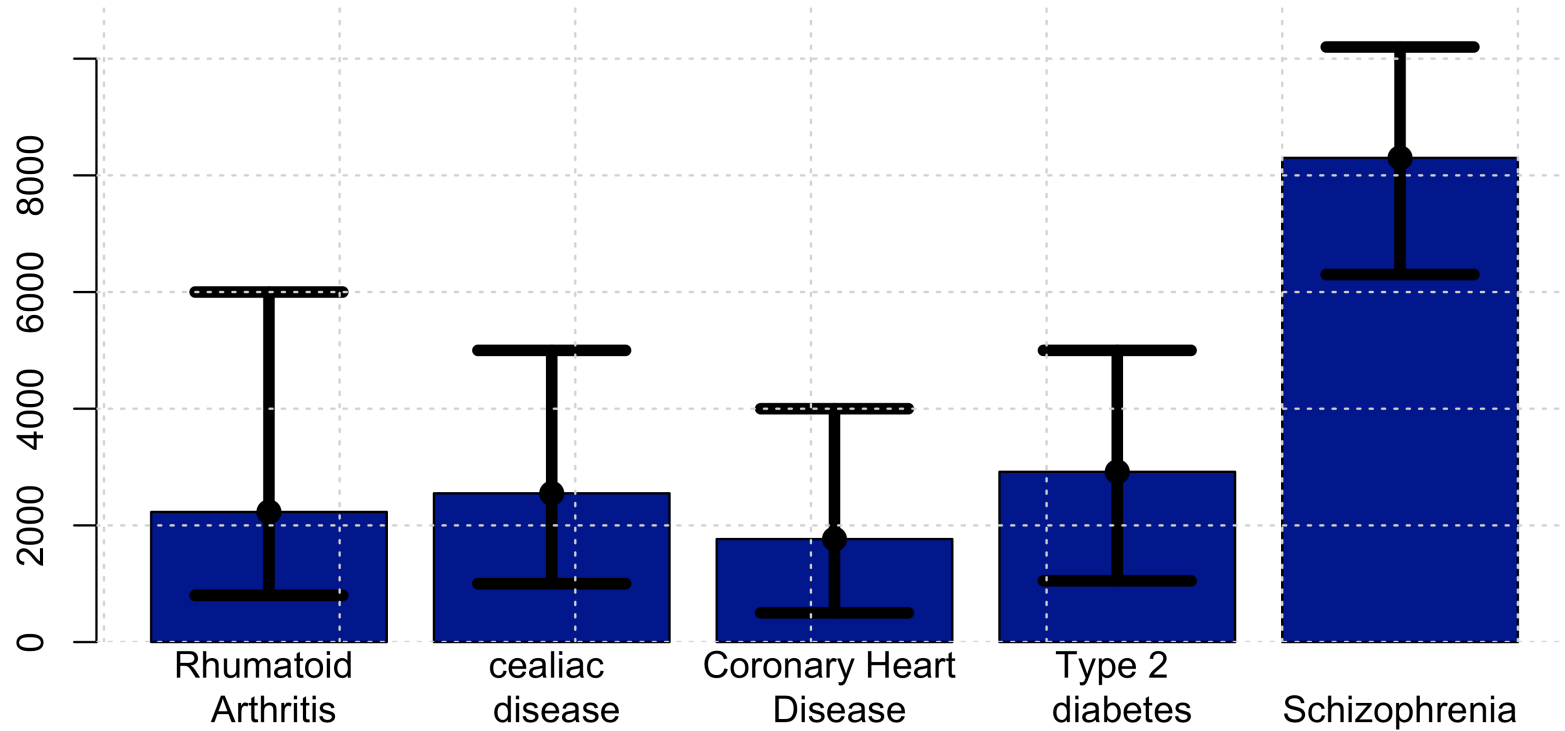


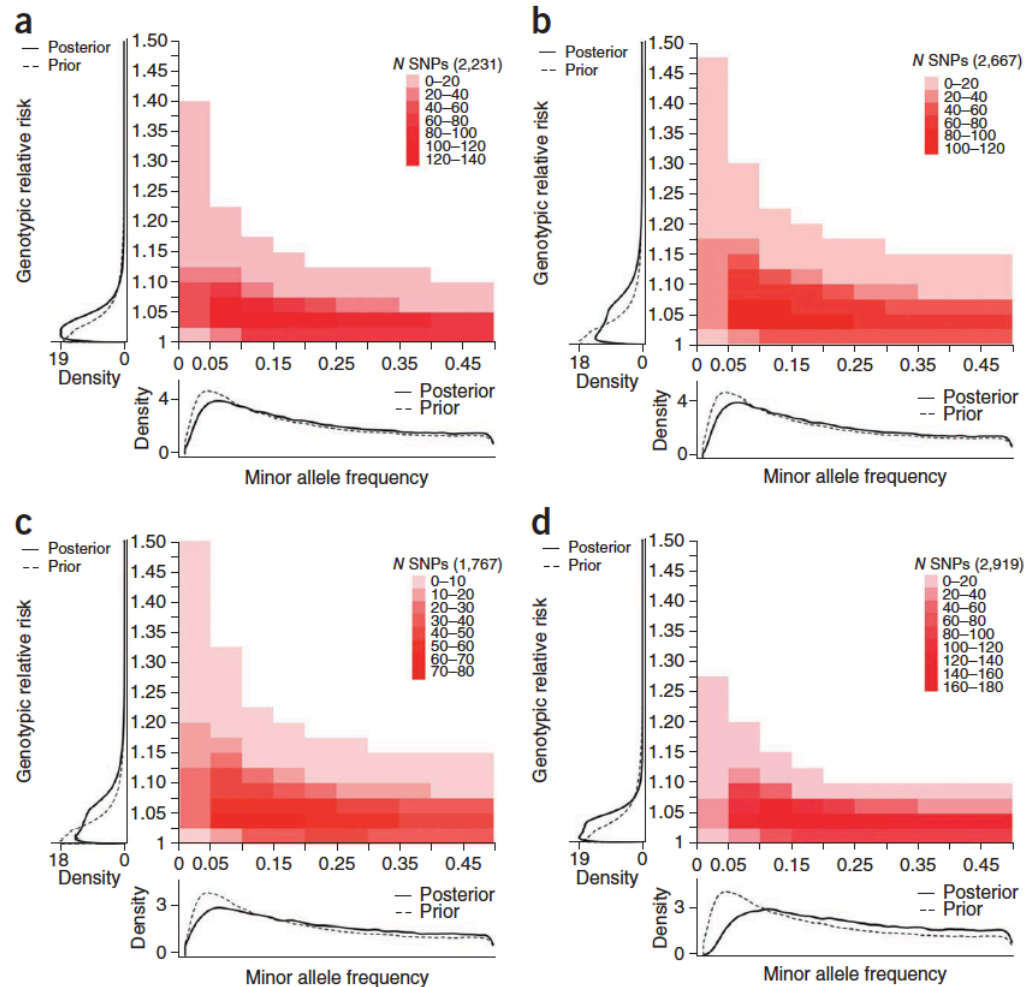
Fig. 2. Estimated total number of single-nucleotide polymorphisms associated with five disease phenotypes based on results of genome-wide association studies (data drawn from Ripke et al., 2013; see that paper and Stahl et al., 2012, for further methodological details).

Nsnp



More Results

Figure 3 Posterior probability distributions of the relative risk and minor allele frequency of the inferred disease-associated SNPs. The GRR is shown on the y axis in the left and middle images, and the MAF is shown on the x axis in the middle and bottom images. Heat map colors indicate the mean posterior numbers of SNPs in risk allele frequency (RAF)-GRR bins scaled to the posterior mean number of disease-associated SNPs (indicated in the legend). The graphs on the left and at the bottom show the marginal posterior (solid line) and prior (dashed line) probability densities. (a) Rheumatoid arthritis (with all known risk loci removed). (b) Celiac disease (with the extended MHC region removed). (c) MI/CAD. (d) T2D.



Excluded SNPs for RA:
HLA-DRB1: OR 2.73

Other range between: OR
1.07-1.75

Causal variant modeling

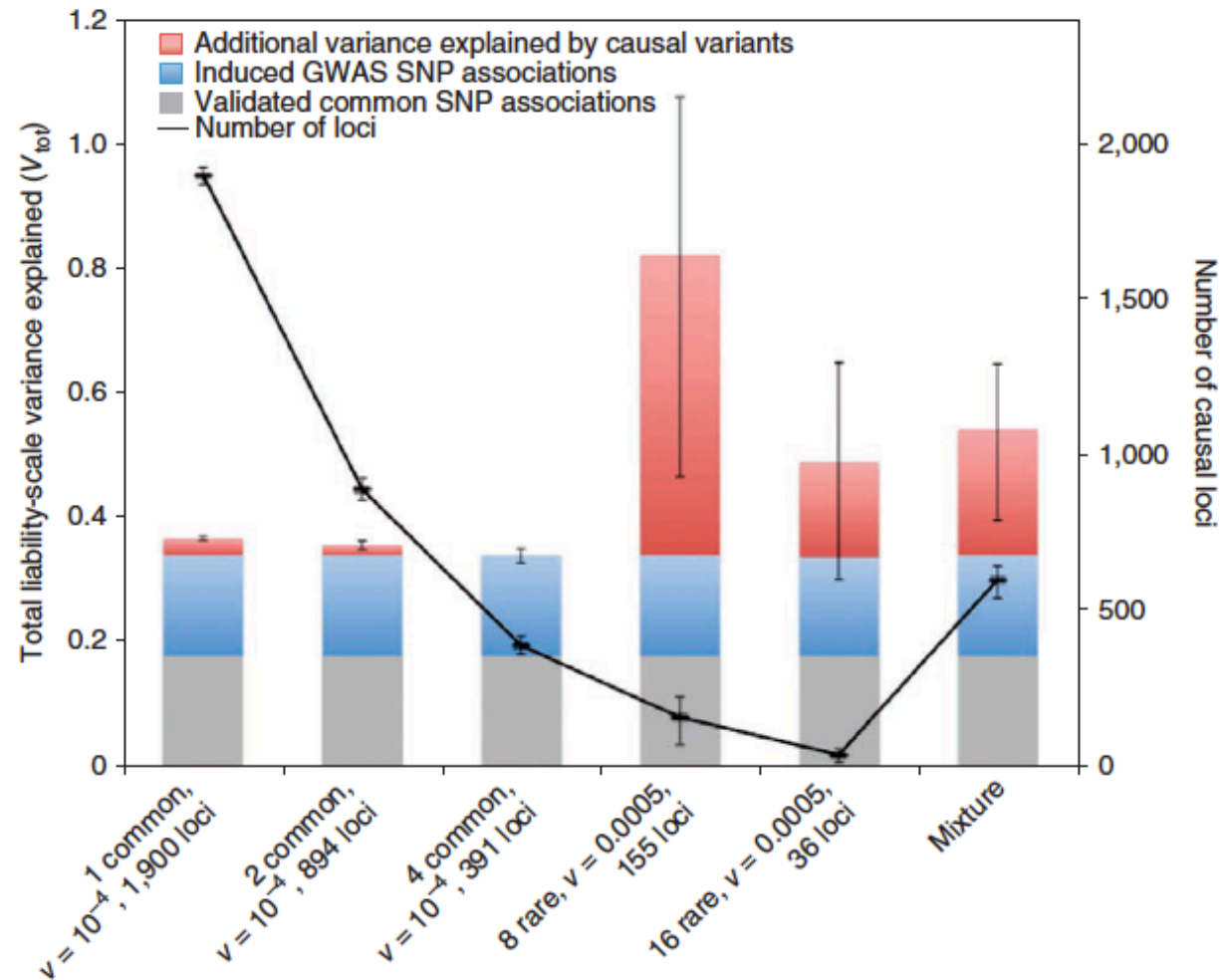


Figure 4 Causal variants underlying the rheumatoid arthritis polygenic disease architecture inferred from the GWAS data. Plotted are the liability-scale variances explained (V_{tot} , bars, left y axes) and the number of loci harboring causal variants (black line, right y axes). The colored sections in the bars partition the V_{tot} values for previously validated common SNP associations (gray), undiscovered GWAS SNP associations induced by causal variants (blue) and causal variants (V_{tot} , in addition to the values for GWAS SNPs, red). Error bars show 95% confidence intervals for causal variant numbers and V_{tot} values based on simulations achieving a GWAS SNP V_{tot} value equal to that inferred from the polygenic modeling. Six plausible causal variant models are plotted (left to right): (i) 1,900 loci each with a single common (MAF > 5%) causal variant, (ii) 894 loci each with 2 common causal variants, (iii) 391 loci each with 4 common causal variants, (iv) 155 loci each with 8 rare (MAF < 1%) causal variants, (v) 16 rare causal variants per locus with $v = 0.0005$ and (vi) a mixture (60:40 ratio of model 2 to model 4 in terms of GWAS SNPs V_{tot} values, implying 536 common causal variant loci and 62 rare causal variant loci). The per-causal-variant liability-scale variances explained (v) for models that are consistent with the polygenic modeling and inference results were $v = 0.0001$ for common causal variants and $v = 0.0005$ for rare causal variants.

Limitations

- **Long computational time** (1,000,000 iterations)
- Wide confidence interval / credibility intervals
 - Is method applicable to very large datasets? Maybe as it was performed for SCZ using PGC2 data. Can be parallelised.
 - Confidence intervals / credibility intervals for SCZ not reported
- Requires raw data
- In this paper, **not excluding known loci for RA** could have been more informative (full distribution of effect sizes and Nsnp)
- Validation of method on simulated data not thorough or realistic
- Not sure about the conclusions about presence of rare variants / synthetic association

Summary

- **Method allows to estimate number of trait associated SNPs (and distribution of effect sizes)**
 - Could be of interest for power calculation in GWAS
 - Differences between psychiatric traits architecture, brain phenotypes?
- Method has been used in another context (Schizophrenia)
- Method conceptually appealing even if probably fastidious to set up and long to run
- Method implemented in ABCtoolbox