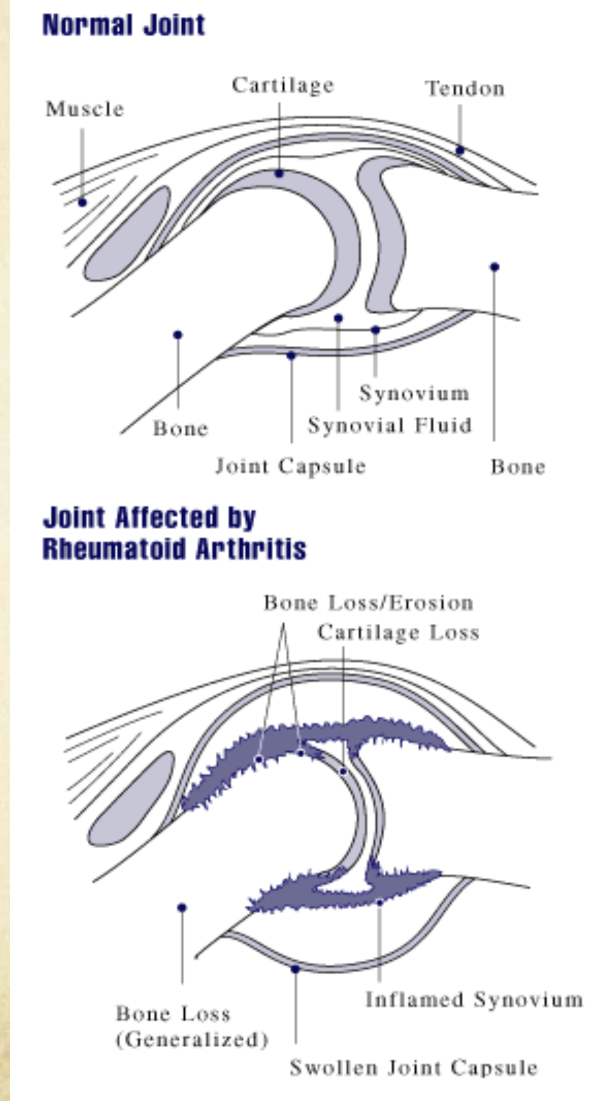
Bayesian inference analysis of the polygenic arhitecture of rheumatoid arthritis

Bioinformatics JC
2.4.12

Rheumatoid arthritis

- Chronic, systemic inflammatory disorder
- Mainly attacks flexible joints
- Cause unknown, considered a systemic autoimmune disease
- 1% of world population affected (women three times more often)
- Onset most often between 40 - 50 y.o.



http://en.wikipedia.org/

Heritability of RA

- Estimated heritability approximately 55%
  - WTCCC paper 2007 - 7 loci
  - Plenge et al. 2007 – 3 loci
  - …
  - Stahl et al 2010 – 7 new (31 altogether)
- Most importantly HLA genes
  - 16% of disease variance explained (12% HLA genes)

Polygenic methods

- Methods for assessing the contribution of SNPs that does not reach the GW significance
  - Polygenic prediction method (Purcell et al. 2009)
    - Schizophrenia (additional 3% of heritability)
  - Mixed linear modeling (Yang et al. 2010)
    - 45% of height genetic variability can be explained

Polygenic architecture of RA

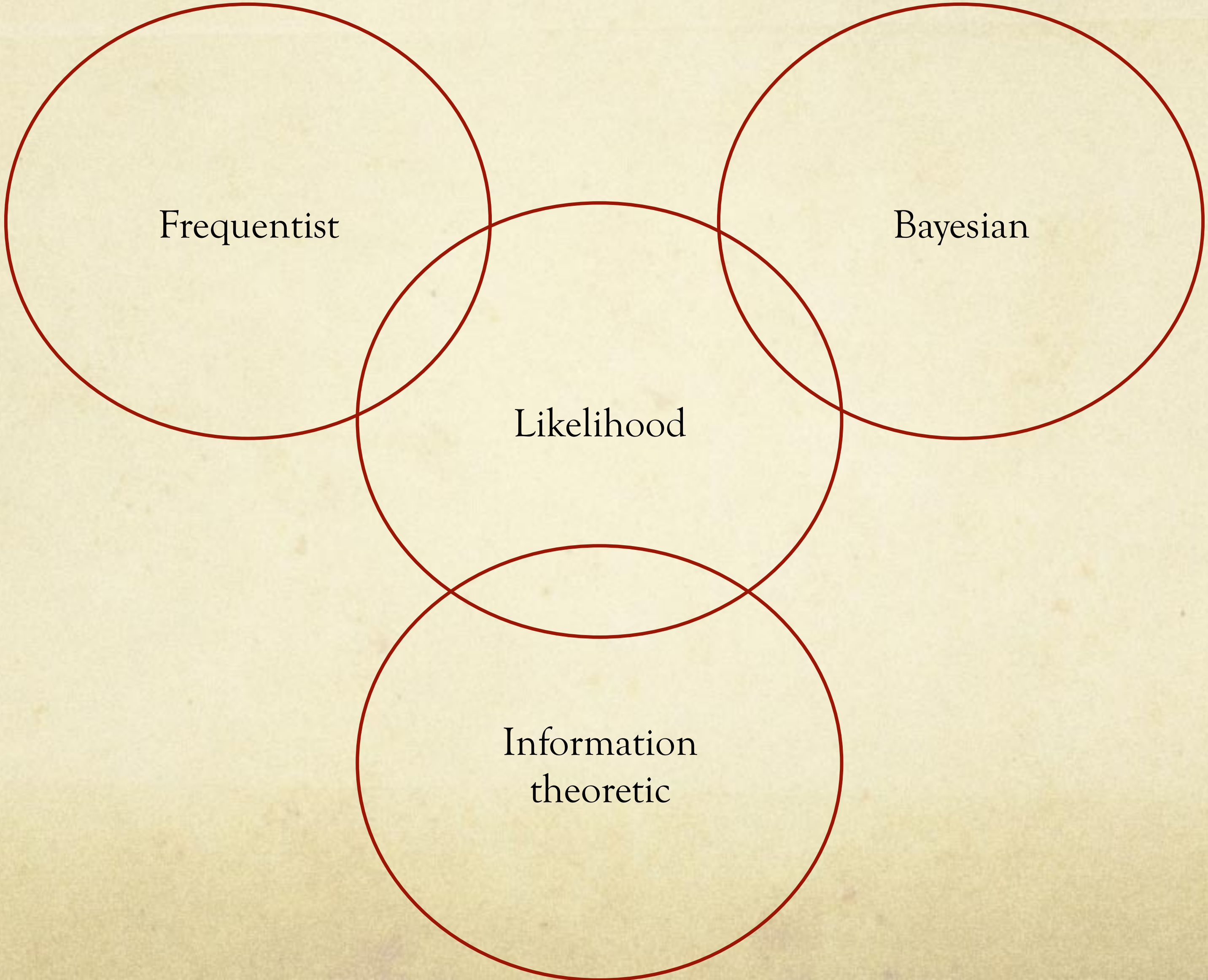
- Polygenic prediction methods explain additional variance, but they do not offer meaningful estimates for the additional numbers and effect sized of associated SNPs

- New method integrates polygenic prediction method with simulation of GWAS data under polygenic disease model using approximate Bayesian computation

Rev. Thomas Bayes

- c. 1701 – 7 April 1761
- Presbyterian minister
- Studied logic and theology in University of Edinburgh
- Author of Bayes' theorem, which was published after his death by Richard Price

Statistical methods



Frequentist inference

- Sir Ronald Fisher – null hypothesis and p-value (evidence against H_0)
- Neyman & Pearson – Type I and type II errors, power, H_1 etc.

Fisher was opposed to the conclusions of Richard Doll and A.B. Hill that smoking caused lung cancer. He compared the correlations in their papers to a correlation between the import of apples and the rise of divorce in order to show that correlation does not imply causation.

Bayesian inference

- Basic idea is that you combine experiment (expressed in terms of likelihood) with some prior information to get posterior probability

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

Material

- GWAS data from 6 independent case/control collections was used
  - 5 sets for discovery
  - WTCCC data for test
- Data was imputed using HapMap2 CEU reference

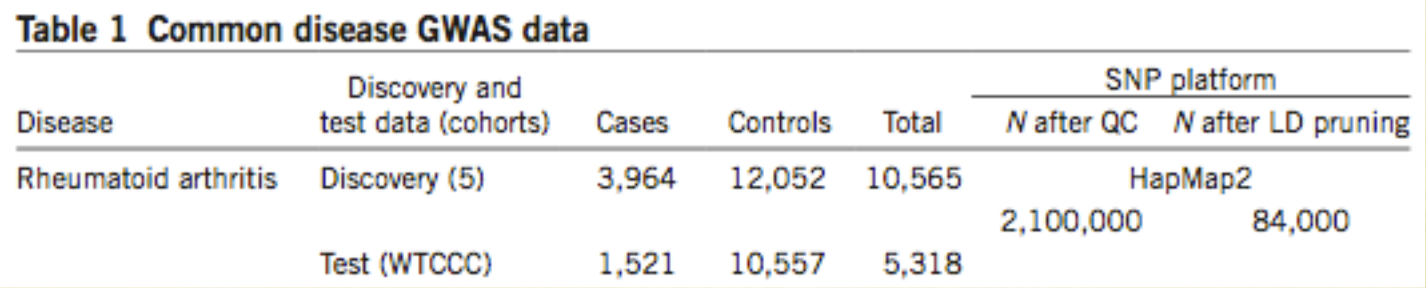


Table 1 Common disease GWAS data

	Discovery and				SNP platform	
Disease	test data (cohorts)	Cases	Controls	Total	*N* after QC	*N* after LD pruning
Rheumatoid arthritis	Discovery (5)	3,964	12,052	10,565	HapMap2	
					2,100,000	84,000
	Test (WTCCC)	1,521	10,557	5,318		

Method

- Logistic regression analysis in each discovery set using five PC as covariates

- Datasets were combined using inverse-variance weighted meta analysis

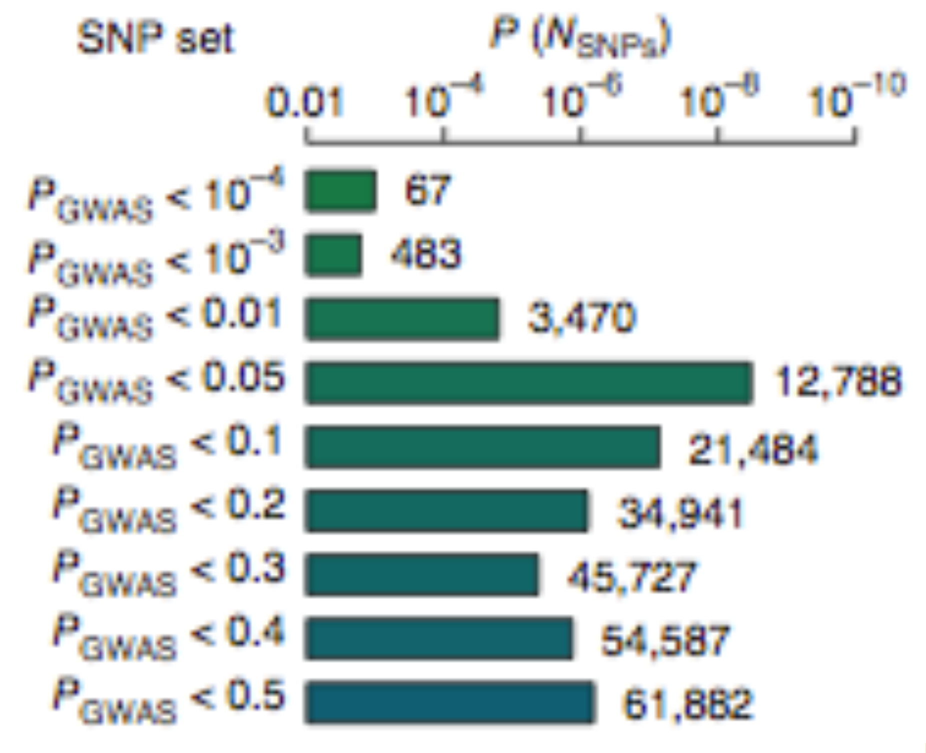
$$B_j = \frac{\Sigma_{i=1}^{N} \beta_{ij} w_{ij}}{\Sigma_{i=1}^{N} w_{ij}},$$

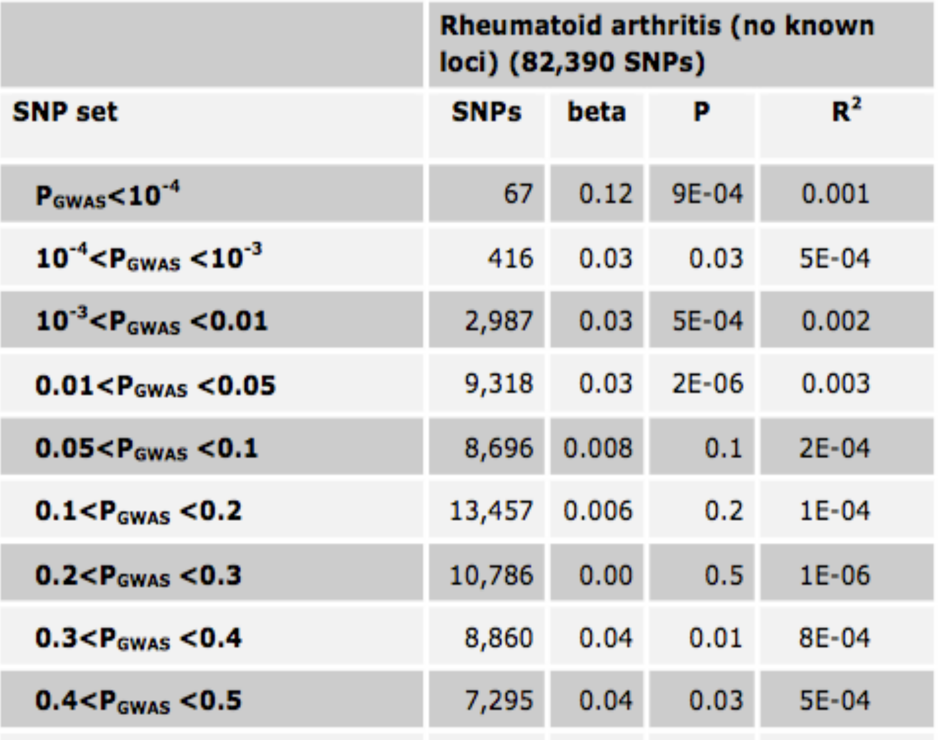
where *wij* = [Var(β *ij*)]$^{-1}$ is the inverse of the variance of the estimated allelic effect in the *i*th study, obtained from the standard error.

Method 2

- Then all known RA risk loci were removed to focus on previously unknown associations
- Rest of markers are pruned by $r2<0.1$ to get a set of independent loci
- Nine different P_{GWAS} thresholds were used for generating SNP sets ($P_{GWAS} < 10^{-4}$, 10^{-3}, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5)
- For each test set (WTCCC) individual log-odds weighted risk allele counts were calculated and summed

Results





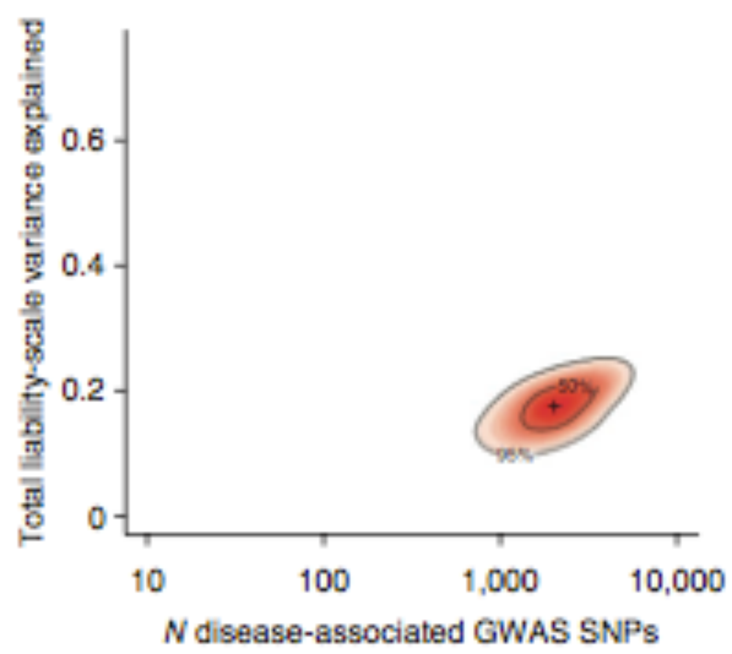
	Rheumatoid arthritis (no known loci) (82,390 SNPs)			
SNP set	SNPs	beta	P	R^2
$P_{GWAS}<10^{-4}$	67	0.12	9E-04	0.001
$10^{-4}<P_{GWAS}<10^{-3}$	416	0.03	0.03	5E-04
$10^{-3}<P_{GWAS}<0.01$	2,987	0.03	5E-04	0.002
$0.01<P_{GWAS}<0.05$	9,318	0.03	2E-06	0.003
$0.05<P_{GWAS}<0.1$	8,696	0.008	0.1	2E-04
$0.1<P_{GWAS}<0.2$	13,457	0.006	0.2	1E-04
$0.2<P_{GWAS}<0.3$	10,786	0.00	0.5	1E-06
$0.3<P_{GWAS}<0.4$	8,860	0.04	0.01	8E-04
$0.4<P_{GWAS}<0.5$	7,295	0.04	0.03	5E-04

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).

Total variance explained

- Polygenic scores are made up of an unknown number of true positive associations and noise
- Bayesian inference analysis were used on polygenic association results to assess:
  - Number of associated SNPs and
  - their total variance explained

Posterior probability densities



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 $\log_{10}$ 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.

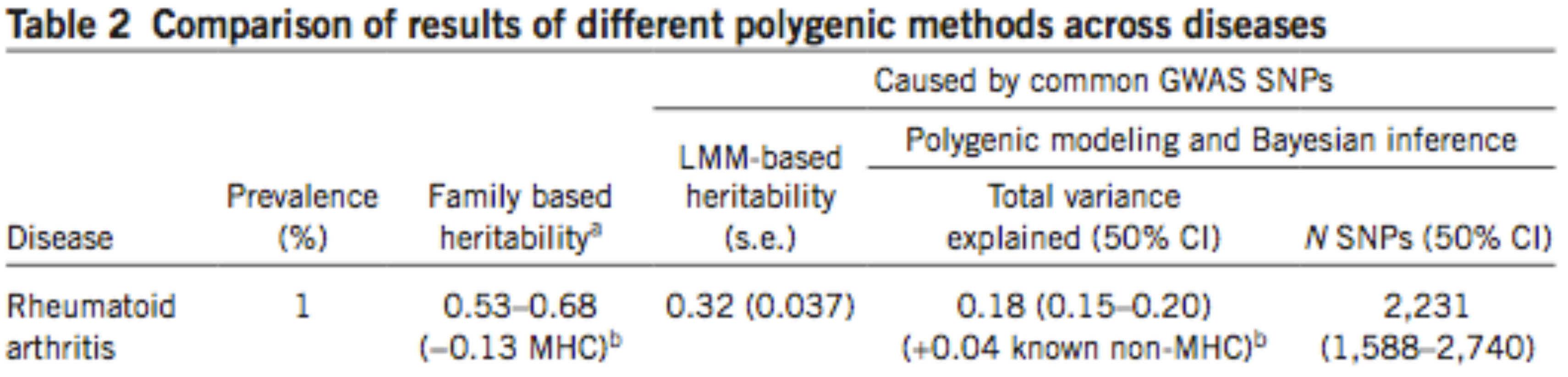
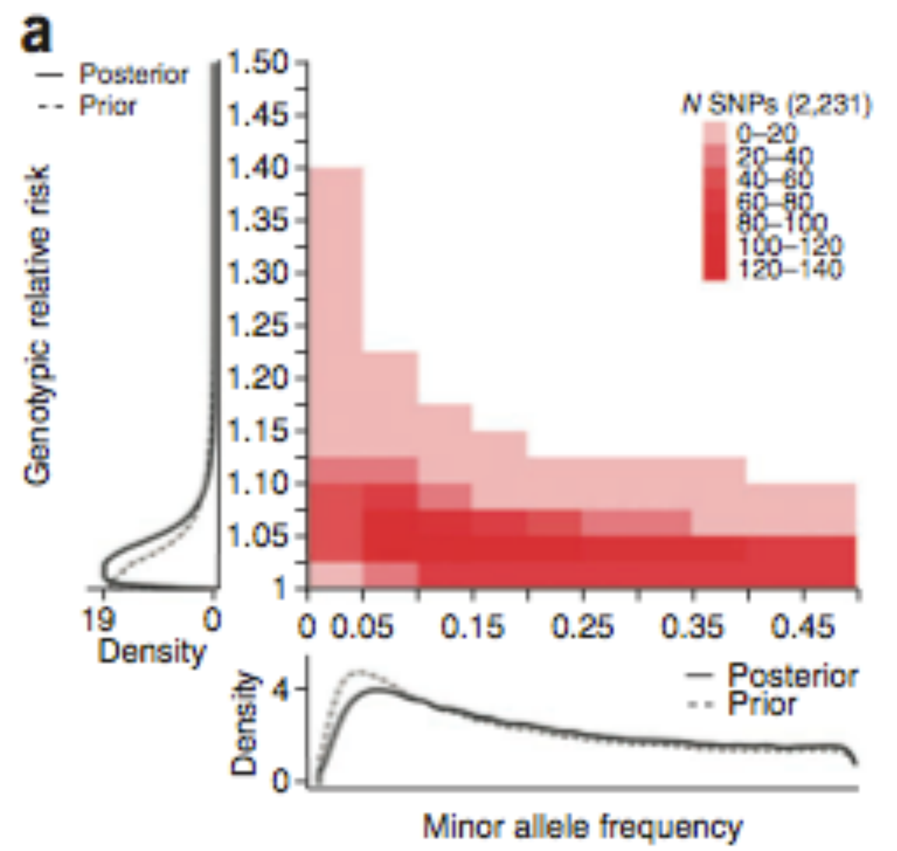


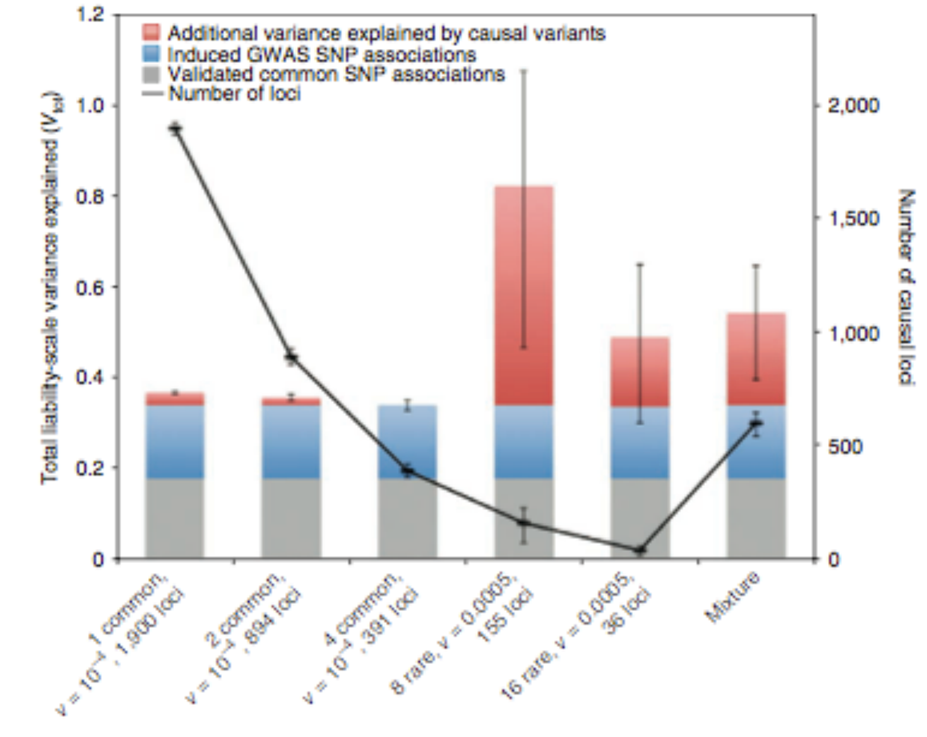
Table 2 Comparison of results of different polygenic methods across diseases

			Caused by common GWAS SNPs		
				Polygenic modeling and Bayesian inference	
Disease	Prevalence (%)	Family based heritability[a]	LMM-based heritability (s.e.)	Total variance explained (50% CI)	*N* 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)

Posterior probability distributions of the relative risk and minor allele frequency



Modeling causal variants



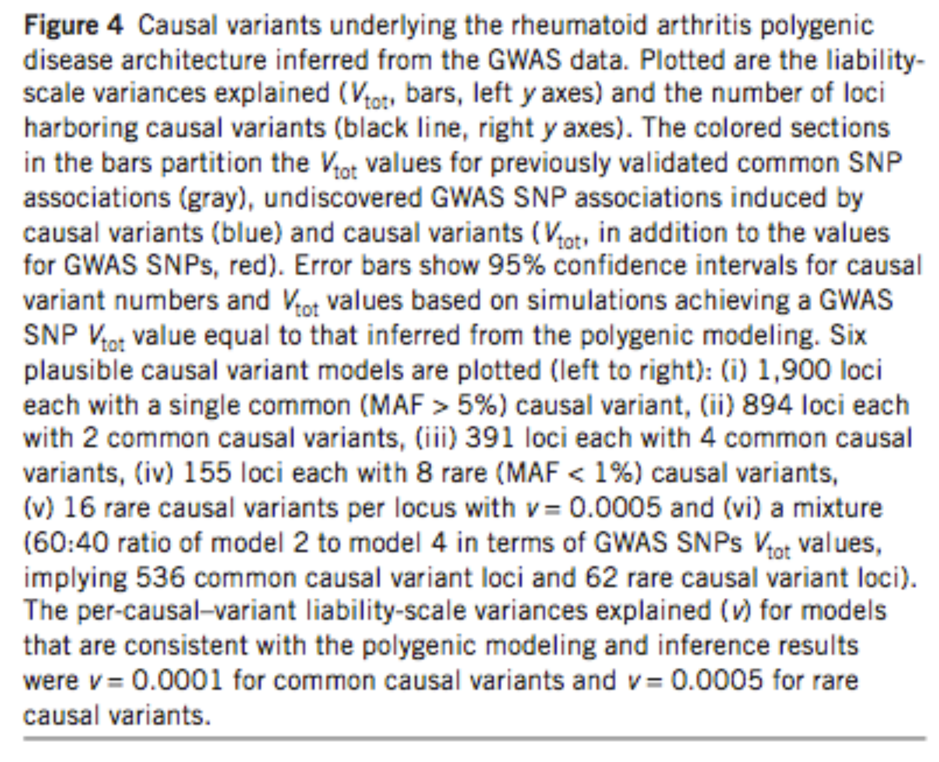


Figure 4 Causal variants underlying the rheumatoid arthritis polygenic disease architecture inferred from the GWAS data. Plotted are the liability-scale variances explained (V_{tot}, bars, left *y* axes) and the number of loci harboring causal variants (black line, right *y* axes). The colored sections 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.

Conclusions

- Bayesian analyses allow for computation of the posterior distribution of polygenic disease model parameters, which can then be used to address questions relating to the genetic architecture of common disease.
- Other potential applications of this type of analysis include performing power calculations to predict the outcomes of future genetic studies, developing future discovery efforts such as Bayesian and pathway-based GWAS

Conclusions 2

- The polygenic model posterior distributions for each of the four diseases examined here give expecta- tions of hundreds of SNPs with moderate effect sizes (GRR > 1.05), especially for celiac disease and MI/CAD.
- Results indicate that the common variant GWAS approach will con- tinue to be a highly productive method of identifying additional risk alleles for common disease.

Questions