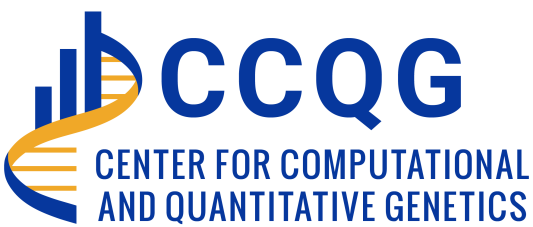
Scalable Bayesian Functional GWAS Method Accounting for Multiple Quantitative Functional Annotations

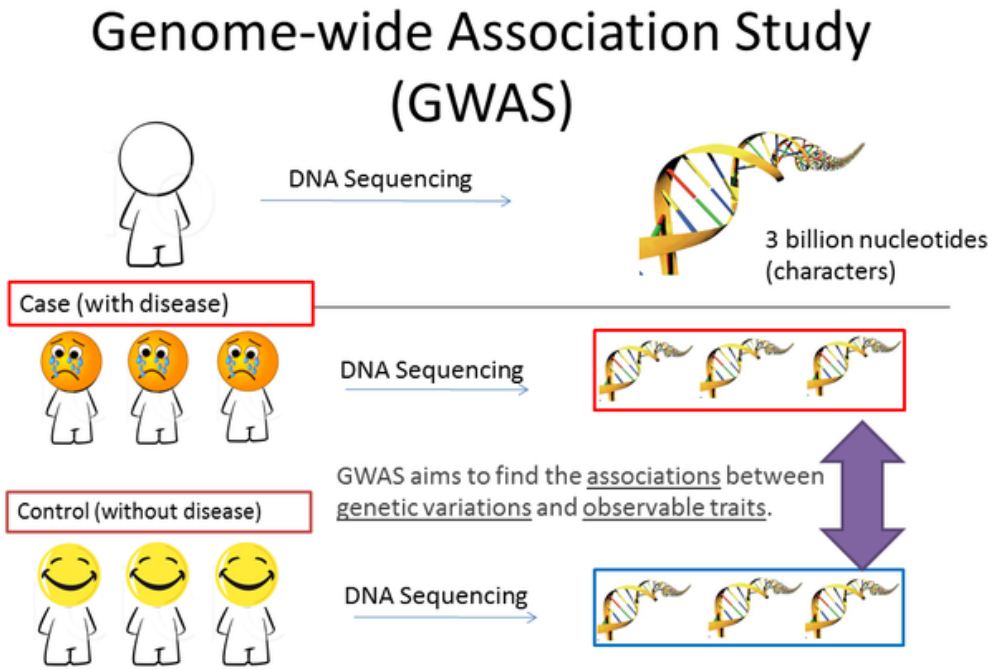
Jingjing Yang, Assistant Professor

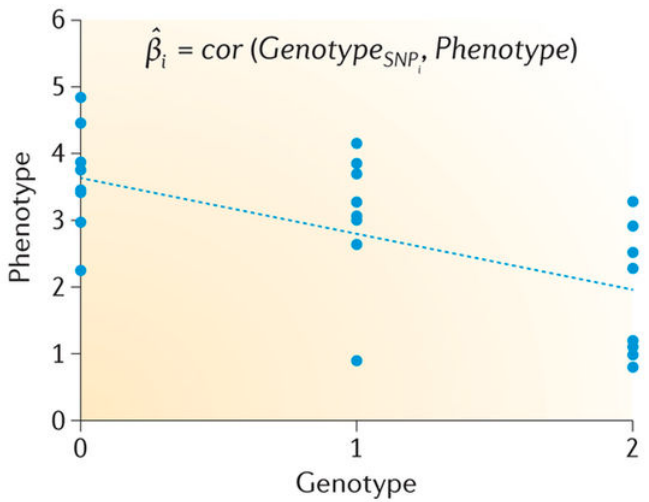


EMORY
UNIVERSITY
SCHOOL OF
MEDICINE

Outline

GWAS





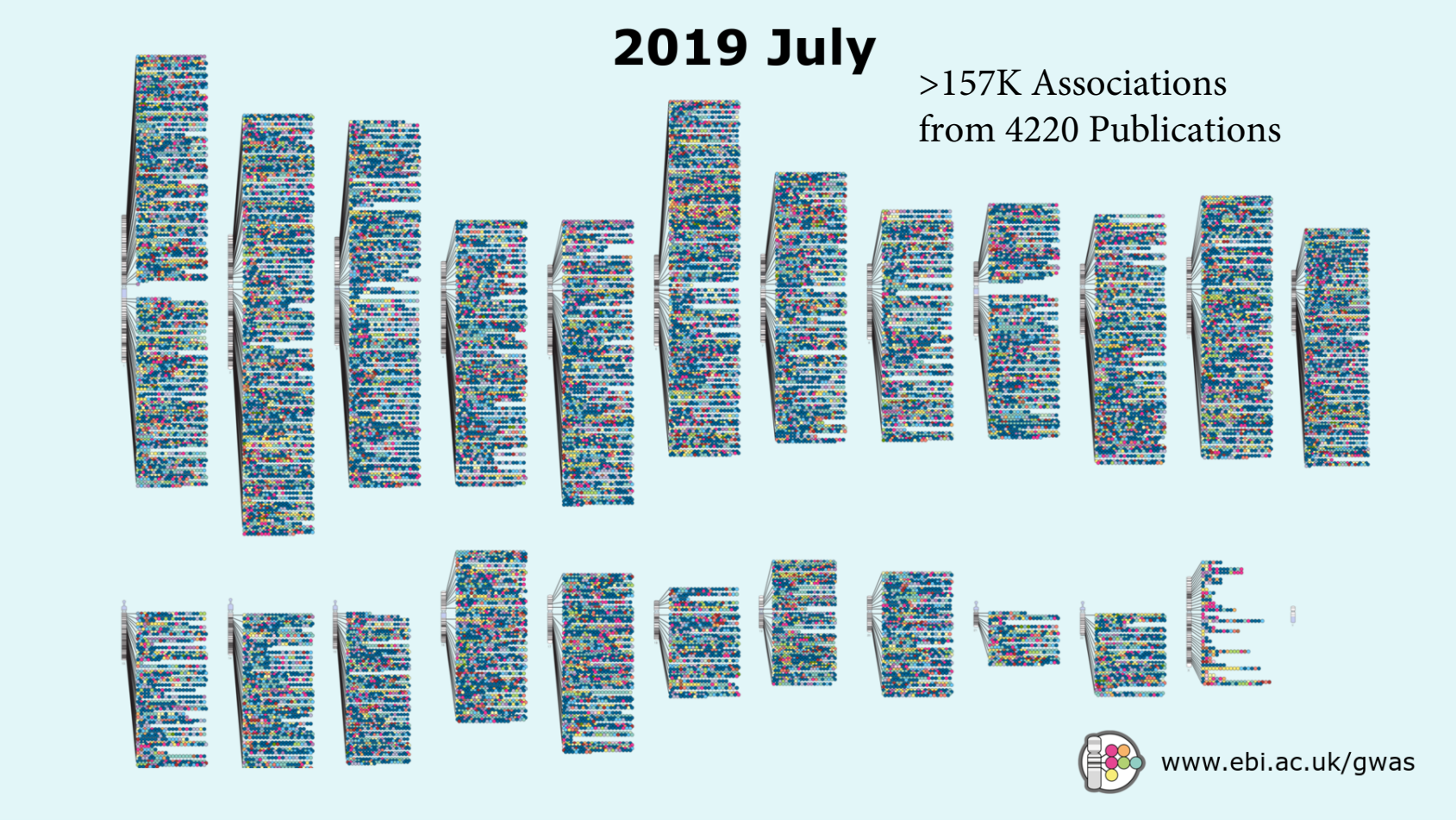
From `Quora.com` and Pasaniuc B & Price AL, Nat. Rev. 2017

Standard GWAS Method

Consider centered phenotype (y) and genetic variant (x_i)

- Logistic regression model $logit(Y) = x_i\beta_i$ for case-control studies
- Linear regression model $Y = x_i\beta_i$ for quantitative phenotypes
- Testing $H_0 : \beta_i = 0$
- Significance threshold **PVALUE** $\leq 5 \times 10^{-8}$, accounting for genome-wide multiple independent tests

GWAS Findings



Standard GWAS Results

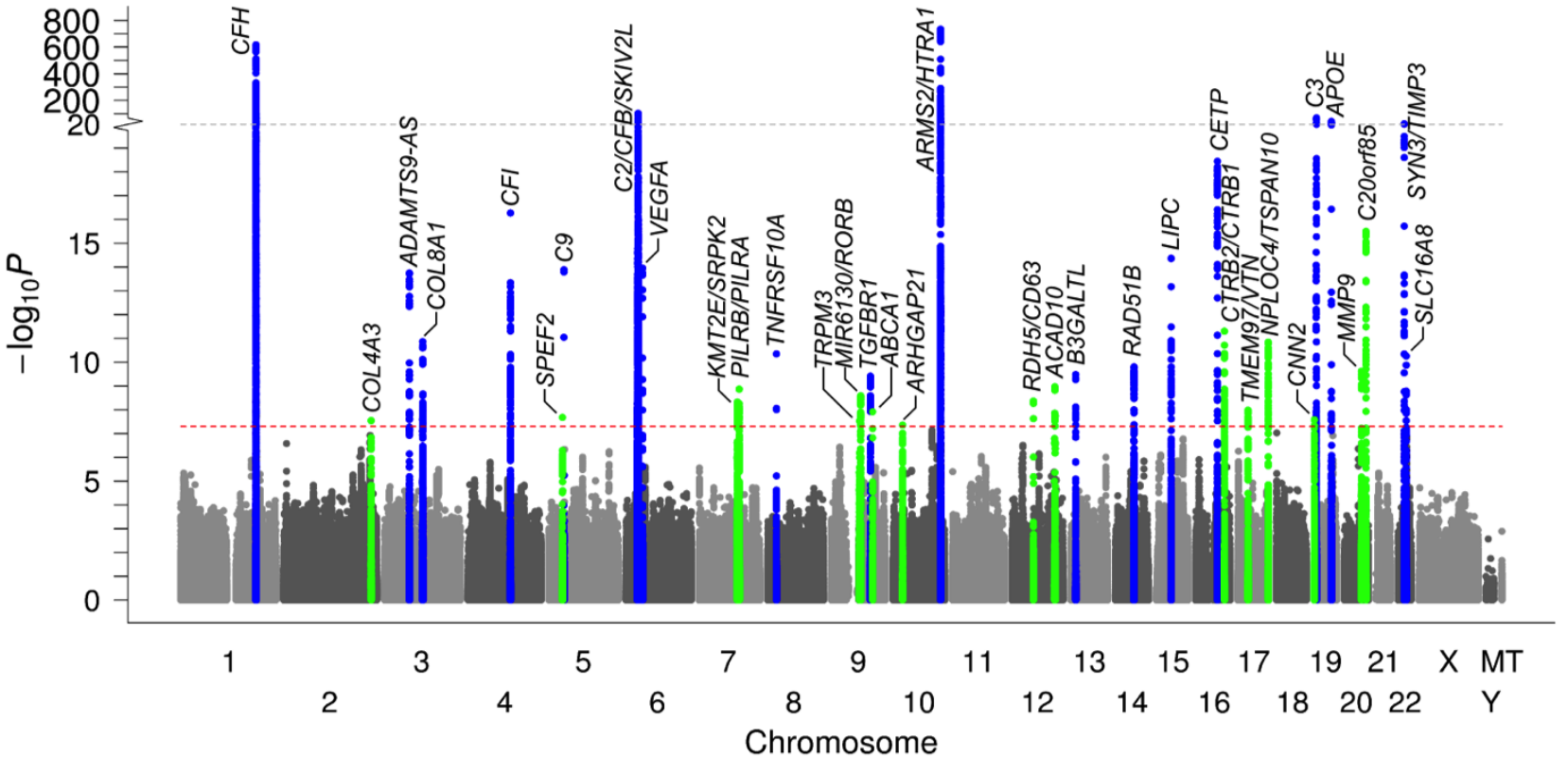
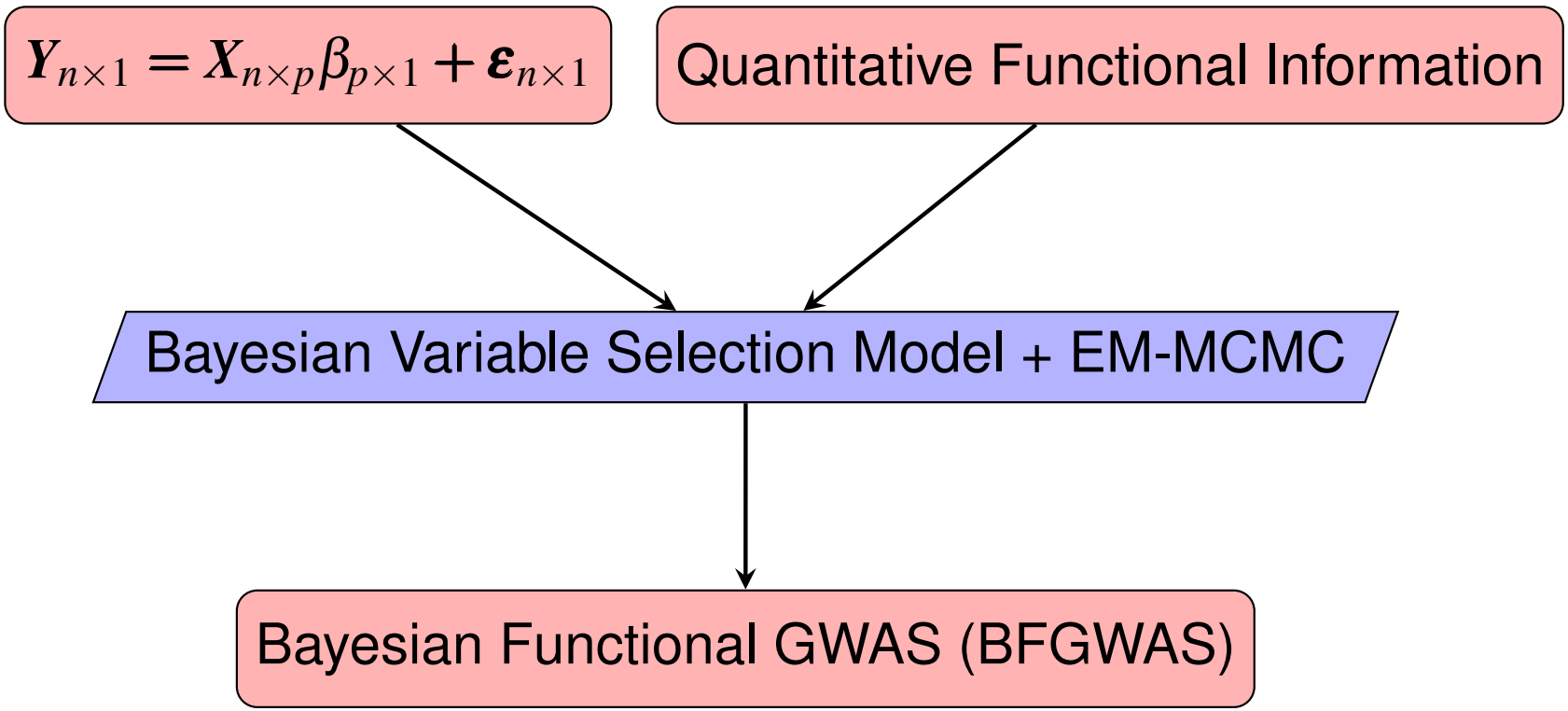


Figure 1: Majority of the associated variants are of unknown functions (Fritsche LG et al., Nature Genetics, 2016).

Motivations

- Account for linkage disequilibrium (LD) to fine-map "causal" variants
- Understand biological mechanisms underlying genetic associations
- Account for multivariate quantitative functional annotations
- Use publicly available summary-level GWAS data of large sample sizes

Method Diagram



Bayesian Hierarchical Model Framework

Multivariable linear regression model with **Standardized** phenotype and genotype vectors:

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times p}\beta_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}, \quad \varepsilon \sim MN(0, I). \tag{1}$$

Prior:

- $\beta_i \sim \pi_i N(0, \frac{1}{n}\tau_\beta^{-1}) + (1-\pi_i)\delta_0(\beta_i);\ i = 1, \cdots, p$
- With augmented quantitative functional annotation data vector $\boldsymbol{A_i} = (1, A_{i,1}, \cdots, A_{i,J})$ for variant $i = 1, \cdots, p$,

$$logit(\pi_i) = \boldsymbol{A_i'\alpha};\ \pi_i = \frac{e^{\boldsymbol{A_i'\alpha}}}{1+e^{\boldsymbol{A_i'\alpha}}};\ \boldsymbol{\alpha} = (\alpha_0, \alpha_1, \cdots, \alpha_J)$$

- Introduce a latent indicator vector $\boldsymbol{\gamma}_{p\times 1}$, equivalently

$$\gamma_i \sim Bernoulli(\pi_i),\ \boldsymbol{\beta}_{-\boldsymbol{\gamma}} \sim \delta_0(\cdot),\ \boldsymbol{\beta}_{\boldsymbol{\gamma}} \sim MVN_{|\boldsymbol{\gamma}|}(0, \frac{1}{n}\tau_\beta^{-1}\boldsymbol{I}_{\boldsymbol{\gamma}})$$

Shared Parameters by Genome-wide Variants

- Fix $\tau_\beta \in (0, 1]$: Inverse of effect size variance in the multivariable regression model.
  - $\tau_\beta = 1$ assumes same prior effect size variance as the marginal effect sizes.
  - $\tau_\beta < 1$ assumes larger magnitude for effect sizes in the multivariable model than the marginal ones.
- $\alpha_j \sim N(0, 1), j = 1, \cdots, J$: Enrichment parameters
- Fix $\alpha_0 \in (-13.8, -9)$ to induce a sparse model.
  - $\alpha_0 = -13.8$ assumes prior causal probability 10^{-6} when $\alpha_j = 0, j = 1, \cdots, J$.

Parameters of Interest

- Enrichment parameters:
  - $(\alpha_1, \cdots, \alpha_J)$: for J annotations
- Variant-specific parameters (association evidence):
  - β_i: Effect-size
  - $\widehat{\pi}_i = E[\gamma_i]$: Causal Posterior Probability (CPP)
- Region-level (Association evidence):
  - `Regional_CPP` = $E[max(\gamma_{i_1}, \cdots, \gamma_{i_k})]$: Regional probability of having at least one causal variant
  - `Sum_CPP` = $\sum_{i=1}^{p} \widehat{\pi}_i I(\pi_i > 0.01)$: Expected number of causal SNPs

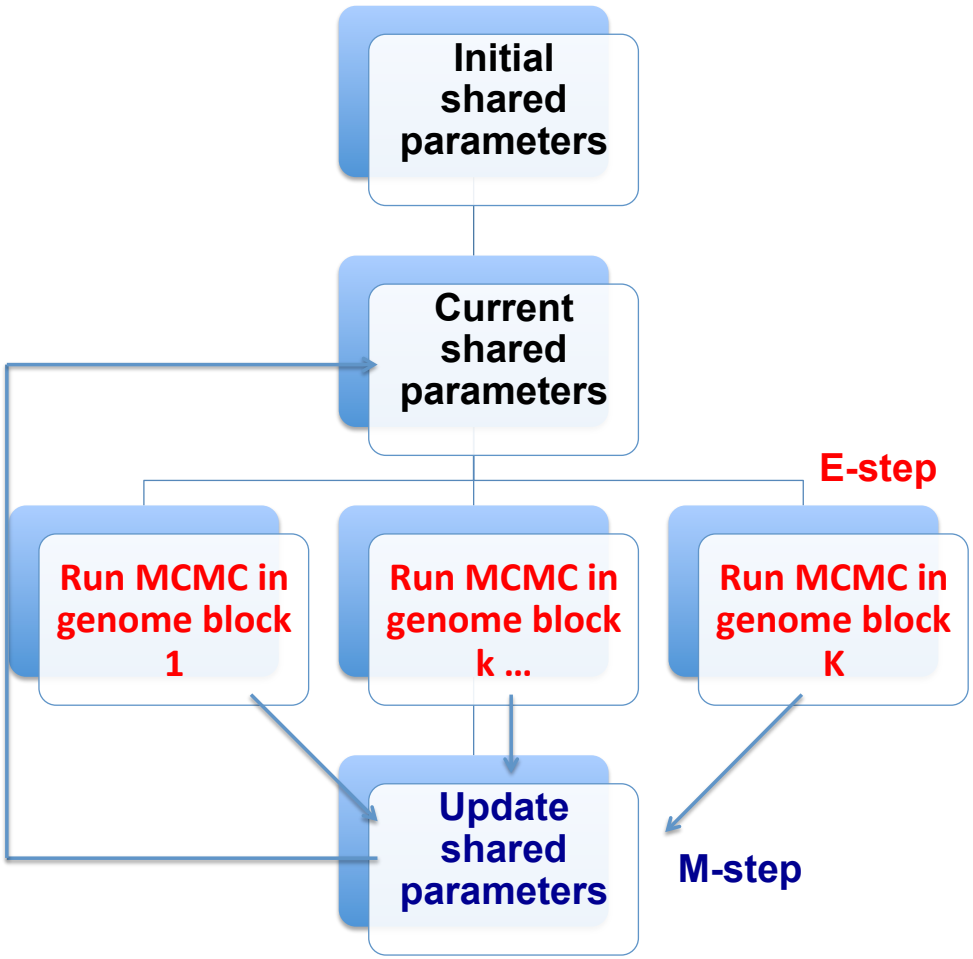
Bayesian Inference

- Joint posterior distribution

$$P(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\pi}(\boldsymbol{\alpha}) | \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{A}) \propto \tag{2}$$
$$P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\gamma}) P(\boldsymbol{\beta}|\boldsymbol{\gamma}, \tau_\beta) P(\boldsymbol{\gamma}|\boldsymbol{\pi}(\boldsymbol{\alpha}), \boldsymbol{A}) P(\boldsymbol{\pi}(\boldsymbol{\alpha}))$$

- Product of Likelihood and Prior
- MCMC by Metropolis-Hastings algorithm with a proposal strategy for $\boldsymbol{\gamma}$
- Challenges of Standard MCMC: Memory Usage and Convergence Rate for $\sim 10M$ genome-wide variants

EM-MCMC Algorithm



Enable Genome-wide Analysis

Improve MCMC Convergence Rate

MCMC Algorithm

Given shared parameters $(\alpha_0, \alpha_1, \cdots, \alpha_J), \tau_\beta$:

- Propose a new indicator vector γ
- Posterior conditional distribution for $\boldsymbol{\beta}_{|\gamma|}$:

$$P(\boldsymbol{\beta}_{|\gamma|}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\gamma}, \tau_\beta) \sim MVN_{|\boldsymbol{\gamma}|}(\boldsymbol{\mu}_{\boldsymbol{\beta}_{|\gamma|}}, \boldsymbol{\Sigma}_{\beta_{|\gamma|}});$$

$$\boldsymbol{\mu}_{\boldsymbol{\beta}_{|\gamma|}} = \boldsymbol{\Sigma}_{\beta_{|\gamma|}} \boldsymbol{X}^T \boldsymbol{Y},\ \boldsymbol{\Sigma}_{\beta_{|\gamma|}} = \frac{1}{n}(\boldsymbol{R} + \tau_\beta \boldsymbol{I}_{m\times m})^{-1},\ \boldsymbol{R} = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$$

- Conditional posterior likelihood:

$$P(\gamma|Y, X, \pi(\alpha), \tau_\beta) \propto$$

$$\sqrt{|\boldsymbol{\Sigma}_{\beta_{|\gamma|}}|} \cdot (n\tau_\beta)^{\frac{m}{2}} \cdot \exp\left\{-\frac{n}{2} + \frac{1}{2}(\boldsymbol{X}^T\boldsymbol{Y})^T \boldsymbol{\Sigma}_{\beta_{|\gamma|}} (\boldsymbol{X}^T\boldsymbol{Y})\right\} \cdot \prod_{i=1}^{P} P(\gamma_i|\pi(\alpha_i))$$

MCMC Algorithm

- Apply Metropolis-Hastings algorithm
- If accepted, update effect-size estimates:

$$\widehat{\boldsymbol{\beta}}_{|\gamma|} = \boldsymbol{\mu}_{\boldsymbol{\beta}_{|\gamma|}} = \boldsymbol{\Sigma}_{\beta_{|\gamma|}} \boldsymbol{X}^T \boldsymbol{Y}$$

- Reference LD and GWAS Summary statistics can be used to derive values for $(\boldsymbol{R}, \boldsymbol{X}^T \boldsymbol{Y})$ in the MCMC algorithm:
  - $\boldsymbol{R}$: Reference LD correlation matrix of the same ancestry
  - $X_i^T Y = \sqrt{n} Z_{score_i}$: Z_{score_i} is the single variant Z-score test statistic using standardized Y and X_i for SNP i

Update Enrichment Parameters by Maximum A Posteriori (MAP)

- The expected log-posterior-likelihood function of $\boldsymbol{\alpha}$:

$$l(\boldsymbol{\alpha}) = E_{\gamma}[ln(P(\boldsymbol{\alpha}|\boldsymbol{\gamma},A))] \propto \sum_{i=1}^{p}\left[\hat{\gamma}_i ln\left(e^{A_i'\boldsymbol{\alpha}}\right) - ln\left(1+e^{A_i'\boldsymbol{\alpha}}\right)\right] - \frac{\boldsymbol{\alpha}'\boldsymbol{\alpha}}{2}$$

- Enrichment parameters $\boldsymbol{\alpha}$ are estimated by using the following gradient and hessian functions:

$$\frac{dl(\boldsymbol{\alpha})}{d\boldsymbol{\alpha}} = \sum_{i=1}^{p}\left[\widehat{\gamma}_i A_i' - \left(1+e^{-A_i'\boldsymbol{\alpha}}\right)^{-1} A_i'\right] - \boldsymbol{\alpha}'$$

$$\frac{d^2 l(\boldsymbol{\alpha})}{d\boldsymbol{\alpha} d\boldsymbol{\alpha}'} = -\sum_{i=1}^{p}\left[\frac{e^{-A_i'\boldsymbol{\alpha}}}{(1+e^{-A_i'\boldsymbol{\alpha}})^2}(A_i A_i')\right] - \boldsymbol{I}$$

Parameters of Interest

- Significant causal SNPs with $CPP > 0.1068$ (equivalent to p-value < 5×10^{-8}), effect size estimates $\widehat{\beta}_i$ and posterior causal probability $\widehat{\pi}_i$
- Estimates of enrichment parameters $(\alpha_1, \cdots, \alpha_J)$
- Sum of CPP estimates for variants with $\widehat{\pi}_i > 0.01$ estimates the number of expected GWAS signals
- Polygenic Risk Score (PRS) with genotype data $\boldsymbol{X}$

$$PRS = \sum_{i=1}^{p} I(\widehat{\pi}_i > 0.01)\widehat{\beta}_i X_i$$

eQTL based Annotations

Derived from frontal cortex brain tissues and Microglia cells:

- AllcisQTL: Binary annotation denoting if a SNP is a significant cis-eQTL
- 95%CredibleSet: Binary annotation denoting if a SNP is in within a fine-mapped 95% credible set of cis-eQTL by CAVIAR
- MaxCPP: Maximum cis-CPP per SNP across all genes
- BGW_MaxCPP: Maximum CPP (cis- or trans-) per SNP across all genes derived by our BGW-TWAS method
- Microglia-eQTL : Binary annotation denoting if a SNP is a significant cis-eQTL of Microglia cell type

Histone Modifications based Annotations

Derived from the epigenomics data in the brain mid frontal gyrus region from the ROADMAP Epigenomics database:

- H3K4me1 (primed enhancers)
- H3K4me3 (promoters)
- H3K36me3 (gene bodies)
- H3K27me3 (polycomb regression)
- H3K9me3 (heterochromatin)

All binary annotations denoting if the SNP is located in the peak regions of the above histone modifications.

Individual-level WGS Data used for Simulation

Whole Genome Sequencing (WGS) Genotype Data

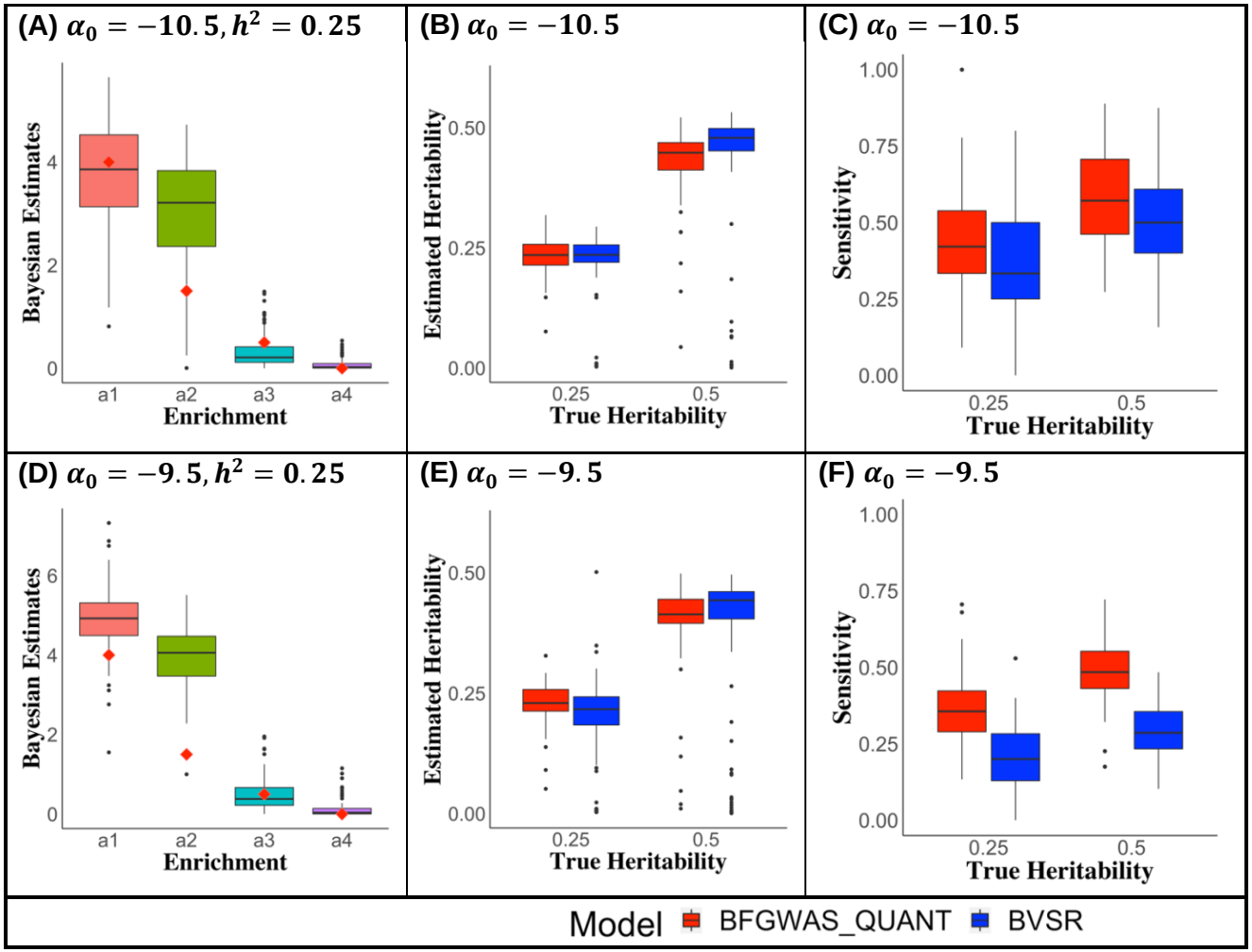
- Religious Orders Study and Rush Memory and Aging Project (ROS/MAP): 1,417 WGS samples
- Mount Sinai Brain Bank (MSBB): 476 WGS samples
- European ancestry

Simulate Phenotype

- WGS of SNPs on Chromosome 19 ($122,745$) with MAF > 1% and HWP $> 10^{-5}$, with sample size $1,893$
- Consider three real cis-eQTL based annotations of Allcis-QTL, 95%CredibleSet, MaxCPP, and a fourth artificial annotation from $N(0,1)$
- Enrichment parameters $(\alpha_0, \alpha_1 = 4, \alpha_2 = 1.5, \alpha_3 = 0.5, \alpha_4 = 0)$, with $\alpha_0 = (-10.5, -9.5)$ to simulate $(5,10)$ or $(15,30)$ true causal SNPs
  - Calculate π_i per SNP
  - Simulate $\gamma_i \sim Bernoulli(\pi_i)$
  - Generate $\beta_i \sim bN(0,1)$ with b selected to ensure target phenotype heritability $h^2 = (0.25, 0.5)$ were equally explained
  - Simulate quantitative gene expression traits:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,\ \varepsilon \sim N(0, (1-h^2)\mathbf{I})$$

Results of Simulation Studies



BVSR: Same Bayesian variable selection regression (BVSR) model but not accounting for annotations.

BFGWAS using individual-level GWAS data

WGS Genotype data

- Religious Orders Study and Rush Memory and Aging Project (ROS/MAP): 1,417 WGS samples of European ancestries

AD Related Phenotypes

- Clinical diagnosis of AD dementia
- AD pathology indices (tangles, β-Amyloid, global AD pathology)
- Cognition decline rate

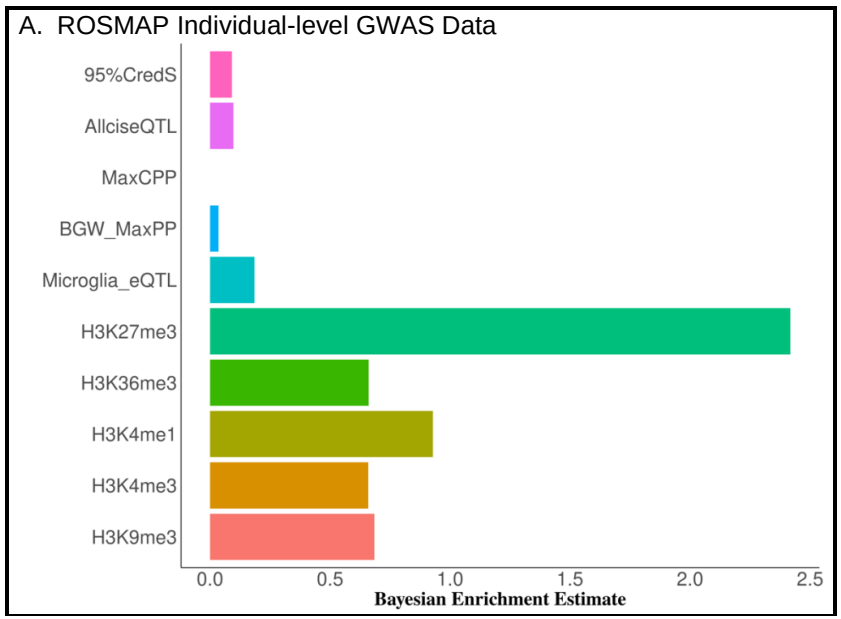
Adjust Phenotypes for Covariates: age, sex, smoking status, study index (ROS or MAP), and top 3 PCs

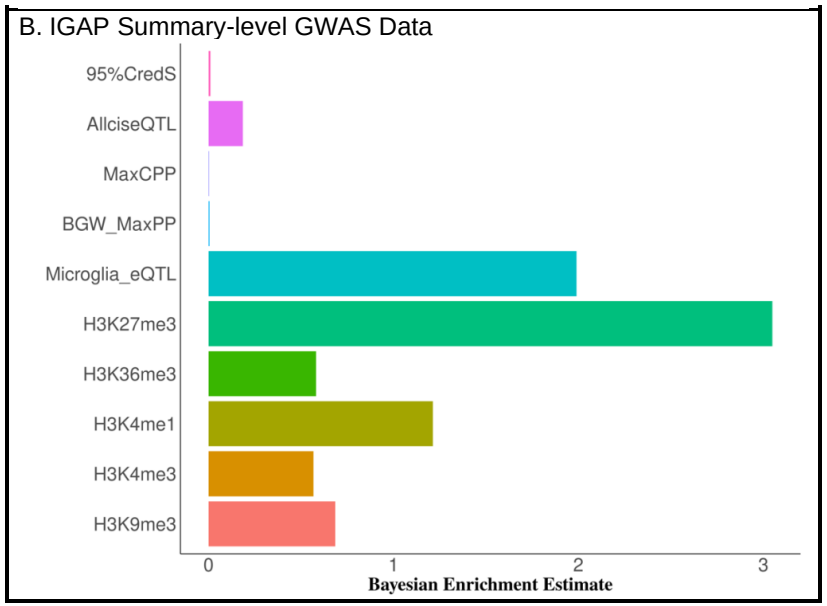
Considered 10 eQTL and histone modification based annotations

BFGWAS using Summary-level GWAS data of IGAP

- GWAS summary data of International Genomics of Alzheimer's Project (IGAP) generated by meta-analysis of four consortia ($\sim$ 17K cases and $\sim$ 37K controls of European ancestries)
  - Alzheimer's Disease Genetic Consortium (ADGC)
  - Cohorts for Heart and Ageing Research in Genomic Epidemiology (CHARGE) Consortium
  - European Alzheimer's Disease Initiative (EADI)
  - Genetic and Environmental Risk in Alzheimer's Disease (GERAD) Consortium
- Reference LD was derived from the ROS/MAP WGS genotype data
- Considered 10 eQTL and histone modification based annotations

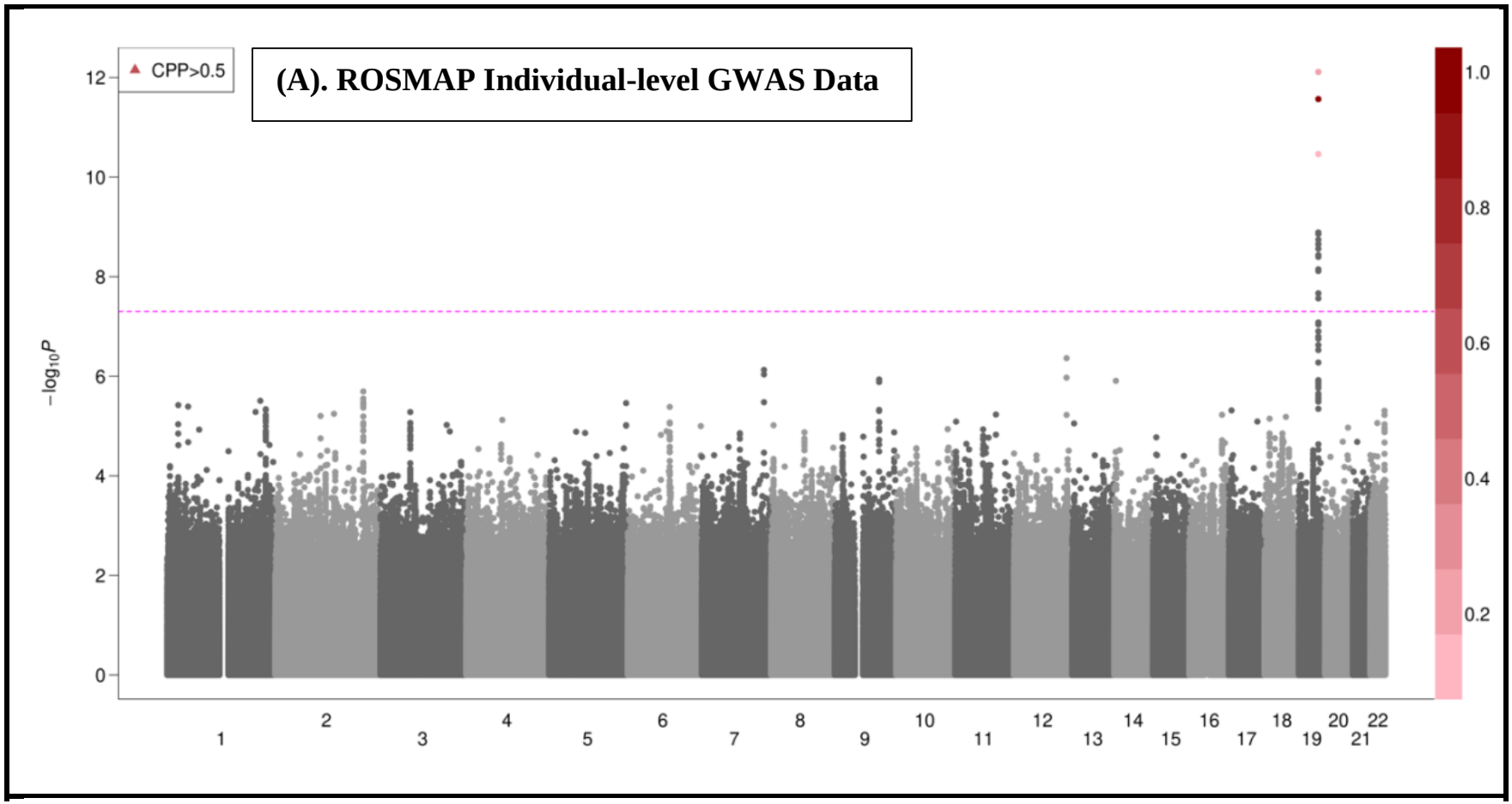
Estimates of Enrichment Parameters in Real Data





- **Microglia is a known related cell type in the brain for Alzheimer's disease.**
- **H3K27me3 is an epigenetic modification to the DNA packaging protein Histone H3, the tri-methylation of lysine 27 on histone H3 protein, which is associated with the downregulation of nearby genes via the formation of heterochromatic regions.**

BFGWAS Results of AD Dementia



BFGWAS Results of AD Dementia

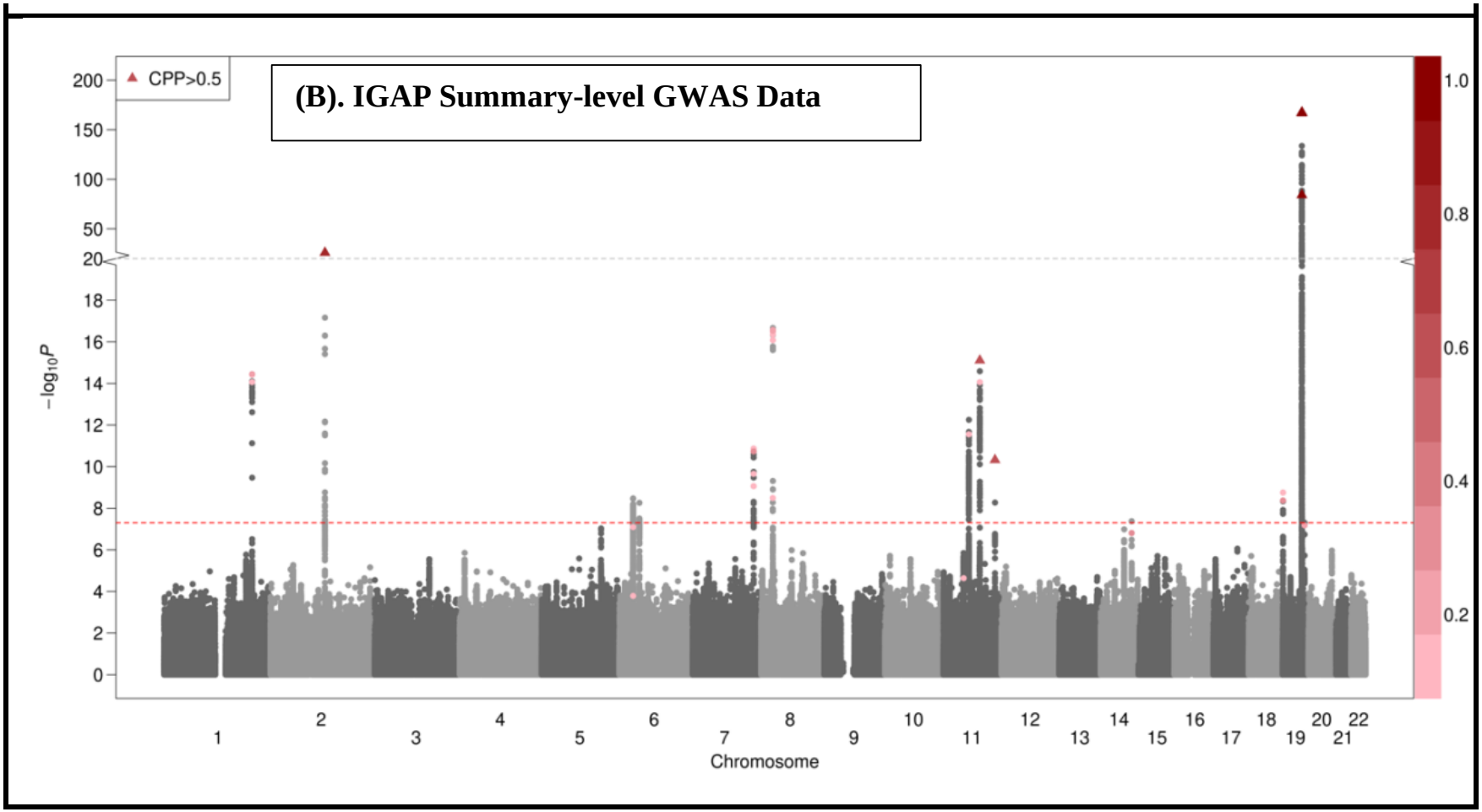


Table 1. Significant SNPs with Bayesian CPP > 0.1068 by BFGWAS_QUANT for studying AD related phenotypes using the ROS/MAP individual-level GWAS data.
SNPs with single variant test P-value $> 5 \times 10^{-8}$ were shaded in gray.

CHR	rsID	Gene	Function	MAF	CPP	Beta	P-value	Phenotype
1	rs148348738	SPATA6	Intron	0.011	0.149	-0.039	4.47E-07	Cognition decline rate
2	rs147749419	CXCR1	Regulatory	0.017	0.154	-0.043	2.94E-08	Cognition decline rate
8	rs11787066	LOC 107986930	Intron	0.148	0.276	0.015	6.93E-08	β-Amyloid
19	rs34134669	ADAMTS10	Regulatory	0.234	0.119	-0.005	8.57E-07	Cognition decline rate
19	rs769449	APOE TOMM40	Regulatory	0.111	0.121	0.076	3.45E-11	Alzheimer's Dementia
				0.112	0.116	0.022	1.51E-16	Tangle density
				0.109	0.475	-0.025	2.09E-15	Cognition decline rate
19	rs429358	APOE	Missense	0.138	0.144	0.037	7.72E-13	Alzheimer's Dementia
				0.138	0.631	0.037	1.17E-20	Tangle density
				0.138	0.999	0.083	6.60E-27	β-Amyloid
				0.139	0.999	0.089	1.19E-33	Global AD pathology
				0.136	0.17	-0.036	1.29E-17	Cognition decline rate
19	rs7412	APOE	Missense	0.077	0.108	-0.027	6.67E-13	Global AD pathology
19	rs1065853	*APOC1*	Intergenic	0.076	0.381	-0.026	8.31E-13	Global AD pathology
19	rs10414043	*APOC1*	Intergenic	0.113	0.111	0.028	2.71E-12	Alzheimer's Dementia
19	rs7256200	*APOC1*	Regulatory	0.113	0.315	0.028	2.71E-12	Alzheimer's Dementia
				0.113	0.228	0.03	3.86E-17	Tangle density
				0.111	0.270	-0.024	3.66E-15	Cognition decline rate
20	rs1131695	APOC1	Stop gained	0.435	0.119	0.039	1.06E-06	Tangle density

Table 2. Significant SNPs with Bayesian CPP > 0.1068 by BFGWAS_QUANT for studying AD using the IGAP summary-level GWAS data. SNPs with single variant test P-value $> 5 \times 10^{-8}$ were shaded in gray.

CHR	rsID	Gene	Function	CPP	Beta	P-value
1	rs6656401	CR1	Intron	0.119	-0.017	8.67E-15
1	rs7515905	CR1	Intron	0.206	-0.019	3.75E-15
1	rs1752684	CR1	Regulatory	0.125	-0.017	3.77E-15
1	rs679515	CR1	Intron	0.220	-0.018	3.60E-15
2	rs4663105	BIN1	Regulatory	0.631	0.050	1.26E-26
2	rs6733839	BIN1	Regulatory	0.796	0.053	1.24E-26
6	rs9270999	HLA-DRB1	Intron	0.181	0.001	8.04E-08
6	rs9273472	HLA-DRB1	Intron	0.110	0.074	1.63E-04
7	rs10808026	EPHA1	Intron	0.123	-0.020	1.36E-11
7	rs11762262	EPHA1	Intron	0.117	-0.011	2.21E-10
7	rs11763230	EPHA1	Intron	0.325	-0.020	1.86E-11
7	rs11771145	EPHA1	Intron	0.173	-0.021	8.69E-10
8	rs28834970	PTK2B	Intron	0.137	0.066	3.22E-09
8	rs2279590	CLU	Intron	0.166	0.021	4.47E-17
8	rs4236673	CLU	Intron	0.123	0.020	3.25E-17
8	rs11787077	CLU	Intron	0.247	0.022	2.94E-17
8	rs9331896	CLU	Intron	0.154	0.022	8.38E-17
8	rs2070926	CLU	Intron	0.278	0.023	2.69E-17
11	rs11039390	NUP160	Downstream	0.145	-0.004	2.31E-05
11	rs4939338	MS4A6E	Upstream	0.139	0.011	2.79E-12
11	rs7110631	PICALM	Intergenic	0.134	0.014	8.77E-15
11	rs10792832	RNU6-560P	Regulatory	0.633	0.027	7.89E-16
11	rs11218343	SORL1	Regulatory	0.643	-0.046	4.77E-11
14	rs10498633	SLC24A4	Intron	0.371	-0.059	1.55E-07
19	rs3752246	ABCA7	Missense	0.361	-0.027	4.27E-09
19	rs4147929	ABCA7	Regulatory	0.111	-0.030	1.77E-09
19	rs41289512	PVRL2	Regulatory	1.000	0.132	1.81E-167
19	rs6857	PVRL2	3' UTR	1.000	0.359	0
19	rs769449	APOE/TOMM40	Regulatory	1.000	0.292	0
19	rs56131196	APOC1	Regulatory	1.000	0.251	0
19	rs78959900	APOC1	Downstream	1.000	-0.096	8.22E-85
19	rs12459419	CD33	Missense	0.245	-0.027	6.66E-08

Sum of Bayesian CPP

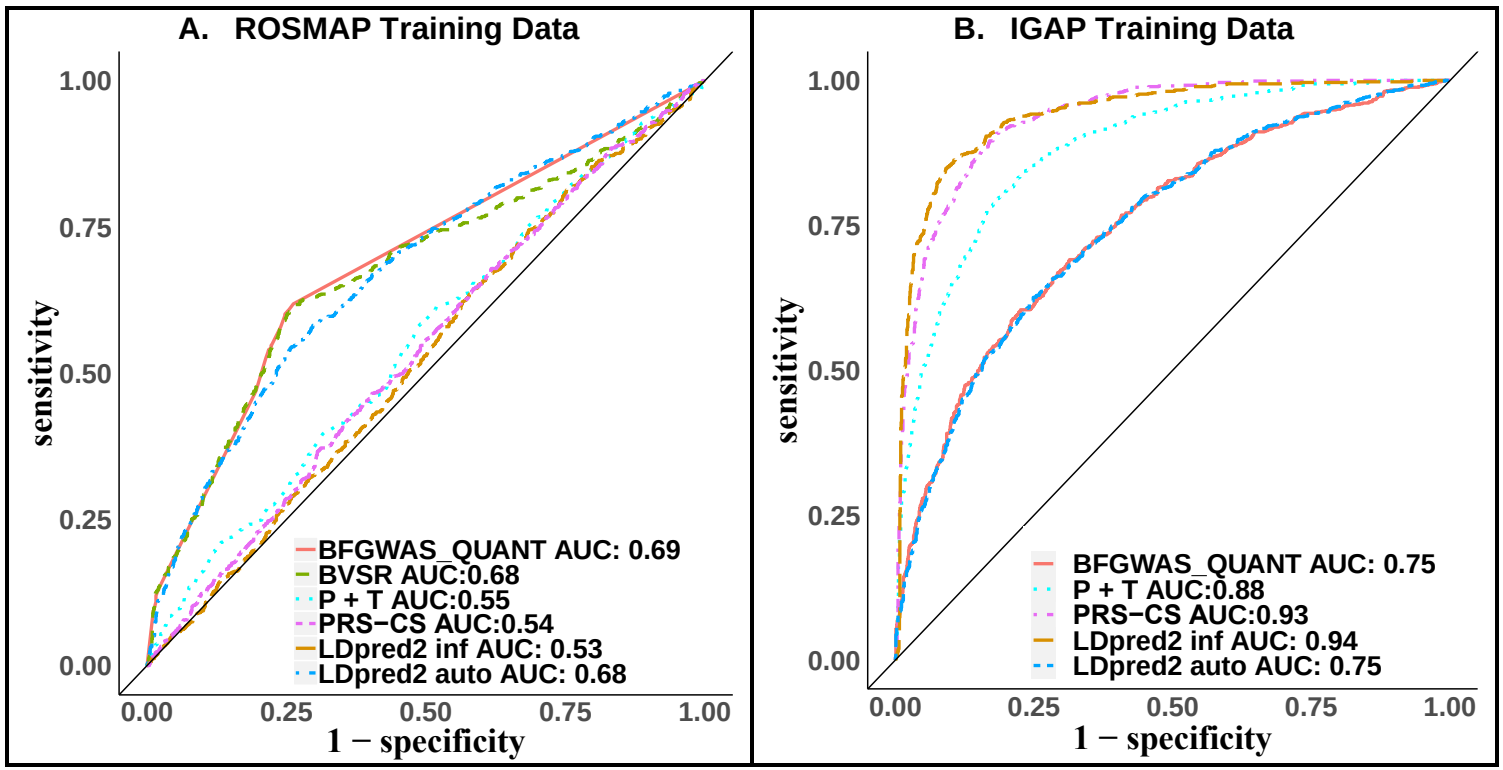
Table 3. Estimates of total causal SNPs. The summations of the Bayesian CPP estimates of SNPs with CPP>0.01 estimate the total number of causal SNPs.

GWAS Data	Phenotype	BFGWAS_QUANT	BVSR[a]
ROS/MAP	Alzheimer's dementia	0.718	6.472
	Tangle density	3.179	6.127
	β-Amyloid	5.375	7.316
	Global AD pathology	5.375	6.174
	Cognition decline rate	6.219	7.136
IGAP	Alzheimer's dementia	54.282	-

[a.] BVSR was not developed for using summary-level GWAS data.

PRS Prediction Accuracy

Predicting the risk of AD in independent Mayo Clinic samples:



Summary

- BFGWAS assumes a sparse causal genetic architecture.
- Ancestry matched reference LD is needed for studying GWAS summary data.
- Quantify enrichment of GWAS signal for multivariate quantitative functional annotations
- Generate fine-mapped GWAS results by accounting for LD and multiple quantitative functional annotations
- Estimate the total number of GWAS signals by `Sum_CPP`

Paper on HGGA

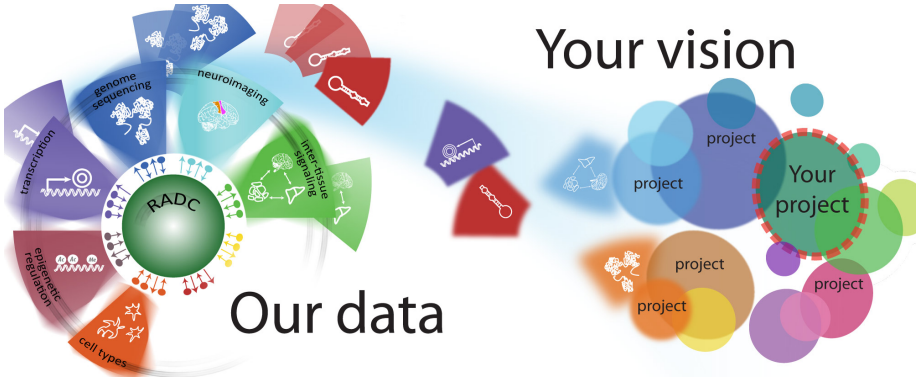
https://doi.org/10.1016/j.xhgg.2022.100143

Tool is available on Github:

https://github.com/yanglab-emory/BFGWAS_QUANT

Acknowledgement

Rush Alzheimer's Disease Center

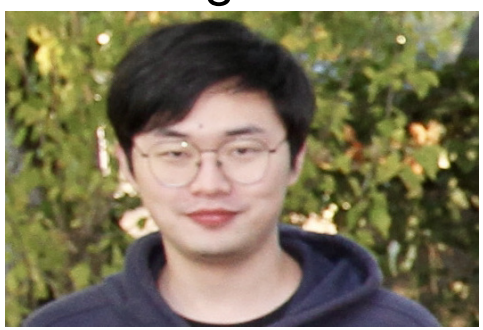


Mayo Clinic LOAD GWAS

AMP-AD Knowledge Portal

Junyu Chen

Lei Wang



Supported by
R35GM138313

NIH National Institute of General Medical Sciences