A scalable Bayesian functional GWAS method accounting for multivariate quantitative functional annotations with applications for studying Alzheimer disease

Junyu Chen,[1,2] Lei Wang,[2,3] Philip L. De Jager,[4] David A. Bennett,[5] Aron S. Buchman,[5] and Jingjing Yang[2,6,*]

Summary

Existing methods for integrating functional annotations in genome-wide association studies (GWASs) to fine-map and prioritize potential causal variants are limited to using non-overlapped categorical annotations or limited by the computation burden of modeling genome-wide variants. To overcome these limitations, we propose a scalable Bayesian functional GWAS method to account for multivariate quantitative functional annotations (BFGWAS_QUANT), accompanied by a scalable computation algorithm enabling joint modeling of genome-wide variants. Simulation studies validated the performance of BFGWAS_QUANT for accurately quantifying annotation enrichment and improving GWAS power. Applying BFGWAS_QUANT to study five Alzheimer disease (AD)-related phenotypes using individual-level GWAS data (n = ~1,000), we found that histone modification annotations have higher enrichment than expression quantitative trait locus (eQTL) annotations for all considered phenotypes, with the highest enrichment in H3K27me3 (polycomb regression). We also found that *cis*-eQTLs in microglia had higher enrichment than eQTLs of bulk brain frontal cortex tissue for all considered phenotypes. A similar enrichment pattern was also identified using the International Genomics of Alzheimer's Project (IGAP) summary-level GWAS data of AD (n = ~54,000). The strongest known *APOE E4* risk allele was identified for all five phenotypes, and the *APOE* locus was validated using the IGAP data. BFGWAS_QUANT fine-mapped 32 significant variants from 1,073 genome-wide significant variants in the IGAP data. We also demonstrated that the polygenic risk scores (PRSs) using effect size estimates by BFGWAS_QUANT had a similar prediction accuracy as other methods assuming a sparse causal model. Overall, BFGWAS_QUANT is a useful GWAS tool for quantifying annotation enrichment and prioritizing potential causal variants.

Introduction

Although thousands of significant associations have been identified by single-variant genome-wide association studies (GWASs) for complex traits and diseases, the majority of GWAS signals reside in the noncoding genome regions and have unknown biological meaning.[1–3] Existing GWAS results based on single-variant tests are still difficult to interpret with respect to the underlying biological mechanisms.[4,5] Promising advancements in sequencing technology have made plenteous multi-omics data available that provide functional annotations of genetic variants available to the scientific community: Combined Annotation-Dependent Depletion (CADD) score;[6] the Roadmap Epigenomics Mapping Consortium,[7] providing DNA methylation, histone modification, and chromatin accessibility information for various human tissues; the Encyclopedia of DNA Elements (ENCODE),[8] providing functional information of human and mouse genomes; and the Genotype-Tissue Expression (GTEx) project,[9] providing expression quantitative trait locus (eQTL) information of 54 human tissues. In particular, molecular QTLs

mapped from profiles of molecular phenotypes (e.g., gene expression from GTEx,[9] chromatin marks from Roadmap,[7,10,11] and protein abundances[12]) and corresponding genomes (genotype data) have been shown to be enriched with GWAS signals and help interprete the underlying biology for studying complex traits and diseases.[13–15] These molecular QTL have been leveraged especially to prioritize GWAS associations of complex traits and diseases.[16–20]

An intuitive but widely used *ad hoc* approach is to fine-map and prioritize potential causal GWAS signals that are also molecular QTLs[21] or located in a region with histone modifications. Recently, advanced statistical methods have been proposed to integrate non-overlapped categorical functional annotation (assigning one function label per variant) with GWAS data to fine-map GWAS results.[22–26] PAINTOR has been proposed to integrate multivariate quantitative functional annotations with GWAS summary statistics to fine-map GWAS loci with thousands of variants[27,28]; FunSPU[29] and STAAR[30] have been proposed to incorporate multiple biological annotations for rare variant association tests. These existing methods have

[1]Department of Epidemiology, Emory University School of Public Health, Atlanta, GA 30322, USA; [2]Center for Computational and Quantitative Genetics, Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA; [3]Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA; [4]Center for Translational and Computational Neuroimmunology, Department of Neurology and Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University Irving Medical Center, New York, NY 10032, USA; [5]Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL 60612, USA
[6]Lead contact
*Correspondence: jingjing.yang@emory.edu
https://doi.org/10.1016/j.xhgg.2022.100143

shown the feasibility and promising results of integrating multivariate quantitative functional annotations with GWAS data to fine-map GWAS results and prioritize potential causal variants. However, these methods were not developed for joint modeling millions of genome-wide variants with multivariate quantitative annotations, which would lead to less accurate quantification of annotation enrichment and reduced power of fine-mapping.

Here, we propose a scalable Bayesian functional GWAS method for integrating multivariate quantitative functional annotations with GWAS data by a Bayesian hierarchical variable selection regression model, referred to as BFGWAS_QUANT. BFGWAS_QUANT assumes a hierarchical logistic prior for the causal probabilities of genetic variants in the standard Bayesian variable selection regression (BVSR)-based GWAS method[31] to jointly model millions of genome-wide genetic variants. BFGWAS_QUANT adapts the scalable expectation maximization Markov chain Monte Carlo (EM-MCMC) algorithm developed by the previous Bayesian functional GWAS (BFGWAS) method, which only models non-overlapped categorical annotations (referred to as BFGWAS_CAT in this paper).[22] BFGWAS_QUANT further improves the computation efficiency by pre-calculating the linkage disequilibrium (LD) correlation matrix and single variant test Z score statistics that are used in the MCMC algorithm or using reference LD correlation matrix and summary-level GWAS data. Bayesian causal posterior probability (CPP) and genetic effect size estimates will be generated by BFGWAS_QUANT, along with enrichment quantification of considered multivariate quantitative functional annotations. Bayesian estimates of genetic effect sizes can be used to derive polygenic risk scores (PRSs) that account for functional annotations.

By simulation studies, we showed that our Bayesian estimates of functional enrichment converged and GWAS power was improved over the standard BVSR method without accounting for functional annotations.[31] We then applied BFGWAS_QUANT to real GWAS data for studying Alzheimer disease (AD) related phenotypes,[32,33] accounting for multivariate quantitative annotations with respect to Roadmap histone modifications[7] of brain mid-frontal gyrus, eQTLs of brain frontal cortex tissue[32,34] and eQTLs of microglia cell type.[35,36] We showed that BFGWAS_QUANT identified interesting enrichment patterns and generated fine-mapped GWAS results using individual-level and summary-level GWAS data. We also showed that PRSs derived from BFGWAS_QUANT effect size estimates led to similar accurate AD risk prediction as other PRS methods assuming a sparse causal model.

Under Material and methods, we provide an overview of the BFGWAS_QUANT method, Roadmap histone modification and eQTL-based quantitative functional annotations, simulation study design, and application studies of AD. Under Results, we describe the results of simulation and application studies. We then end with a Discussion of the advantages and limitations of BFGWAS_QUANT.

Material and methods

Hierarchical BVSR model

BFGWAS_QUANT assumes a hierarchical BVSR[31] model for genome-wide variants,

$$\boldsymbol{y_{n \times 1}} = \boldsymbol{X_{n \times p}} \boldsymbol{\beta_{p \times 1}} + \epsilon_{n \times 1}; \epsilon_{n \times 1} \sim \mathrm{N}(0, \boldsymbol{I})$$

$$\beta_i \sim \pi_i N\left(0, \frac{1}{n}\tau_\beta^{-1}\right) + (1 - \pi_i)\delta_0(\beta_i), \quad i = 1, ..., p; \quad \text{(Equation 1)}$$

where $\boldsymbol{y_{n \times 1}}$ is a vector of standardized phenotype with n subjects, $\boldsymbol{X_{n \times p}}$ is the standardized genotype matrix with p genome-wide genetic variants, and $\boldsymbol{\beta_{p \times 1}}$ is the vector of the genetic effect sizes. Spike-and-slab variable selection prior is assumed per effect size β_i. That is, β_i has probability π_i to be non-zero and follows a normal distribution centered at zero and probability $(1 - \pi_i)$ to be zero with a point-mass density function at 0, where π_i denotes the "casual" probability of the i th variant.

We assume a hierarchical logistic model for the causal probabilities of genetic variants to account for multivariate quantitative functional annotations,

$$logit(\pi_i) = \boldsymbol{A_i'}\boldsymbol{\alpha}, \ i = 1, ..., p, \quad \text{(Equation 2)}$$

where π_i denotes the casual probability of the i th variant as in Equation 1, $\boldsymbol{A_i} = (1, \ A_{i1}, ..., A_{iJ})'$ denotes the augmented annotation vector for the i th variant with an intercept term as the first element, and coefficient vector $\boldsymbol{\alpha} = (\alpha_0, \ \alpha_1, ..., \alpha_J)'$ denotes the intercept term α_0 and enrichment quantification with respect to functional annotation $j = 1, ..., J$.

Further, we assume a fixed value in the domain of (0, 1] for τ_β, a fixed value for α_0 in the domain of $(-13.8, -9)$, and a standard normal prior for enrichment parameters ($\alpha_j \sim N(0, 1); j = 0, 1, ...J$). In particular, $\tau_\beta = 1$ would assume that the prior variance of effect sizes is the same as the marginal effect size estimates in a single-variant regression model, and smaller τ_β values would inflate the magnitude of Bayesian effect size estimates. The lower bound value of $\alpha_0 = -13.8$ would assume that the prior causal probability is 10^{-6} when $\alpha_j = 0, j = 1, ..., J$ (see supplemental information for model details).

Adapted EM-MCMC algorithm

To overcome the heavy computational burden and poor mixing rate of posterior samplings by the standard MCMC algorithm,[31] we adapt the scalable EM-MCMC algorithm developed for the original BFGWAS method.[22] Specifically, we first segment genome-wide variants into approximately independent genome blocks with about 5,000–10,000 variants based on LD structure.[22,37] Second, conditioning on given hyperparameters (α_j, $j = 1, ..., J$), we conduct a standard MCMC algorithm within each genome block to obtain Bayesian posterior estimates for genetic effect size (β_i) and causal probability (π_i) per variant (expectation (E) step). Third, conditioning on the Bayesian estimates of ($\beta_i, \pi_i, i = 1, ..., p$), we update the values of hyperparameters ($\alpha_j, i = 1, ..., J$) by maximizing their conditional posterior likelihood (maximization [M] step). The computational optimization algorithm Broyden-Fletcher-Goldfarb-Shanno (BFGS)[38] is used to obtain the maximum *a posteriori* (MAP) estimates for $\alpha_j, j = 1, ..., J$. The EM steps will be iterated (~5 iterations) until the estimates of hyperparameters converge.

The Bayesian estimate for the causal probability (π_i) per variant is referred to as CPP, and the Bayesian estimates for annotation coefficients ($\alpha_j, j = 1, ..., J$) are referred to as quantified enrichment of

multivariate functional annotations. SNPs with CPP greater than 0.1068 will be considered significantly associated with the phenotype of interest, where the significance threshold has been shown to be equivalent to $p < 5 \times 10^{-8}$ in a previously published BFGWAS_CAT paper.[22] Because a multivariate regression model is fitted per genome block, the LD among all variants per genome block is accounted for during the Bayesian inference of CPP and effect size β. The GWAS results obtained by BFGWAS_QUANT will be fine-mapped, and variants with enriched annotations will be prioritized.

In particular, implementing the MCMC algorithm per genome block can greatly reduce the search space (from genome-wide to a genome block) and facilitate parallel computing (one core per genome block), leading to an efficient convergence rate and improved mixing rate. Computation efficiency is further improved by implementing the MCMC algorithm using a pre-calculated LD correlation matrix per genome block and single-variant test Z score statistics or a reference LD correlation matrix and GWAS Z score statistics, which will save up to 90% computation time compared with using individual-level GWAS data.[37] With 32 computation cores in one node, BFGWAS_QUANT can complete analyzing approximately 10 million SNPs in approximately 4 h for 5 EM iterations.

eQTL-based functional annotations

In this paper, we considered 5 real eQTL-based quantitative functional annotations. Three of these annotations (Allcis-eQTL, 95% CredibleSet, and MaxCPP) were constructed based on standard *cis*-eQTL data of brain frontal cortex tissue from GTEx data[9,34]: (1) binary annotation Allcis-eQTL was constructed by taking all SNPs that were identified as a significant *cis*-eQTL (false discovery rate [FDR] < 5%; 1 Mb from the transcription start site [TSS]) for at least one expression quantitative trait (one gene) as 1 and otherwise as 0. (2) For each gene expression trait that has at least one significant (FDR < 5%) *cis*-eQTL, CAVIAR[16] was used to calculate the CPP (*cis*-CPP) of each *cis*-SNP and identify a 95% credible set. SNPs that do not belong to any 95% credible set were taken as 0 or otherwise 1 for the annotation of 95%CredibleSet. (3) Maximum *cis*-CPPs per SNP across all genes were taken as quantitative values of MaxCPP. We also took the maximum Bayesian genome-wide CPP of being *cis*- or *trans*-eQTL across all genes in brain frontal cortex tissue from the Religious Orders Study and Rush Memory and Aging Project (ROS/MAP)[21,33] as the fourth annotation BGW_MaxCPP, where *cis*- and *trans*-CPPs were estimated by Bayesian genome-wide transcriptome-wide association study (BGW-TWAS) method.[32,37] Last, we derived a fifth Microglia-eQTL annotation from two datasets of recent microglia cell-type specific eQTL summary statistics,[35,36] where 1 indicates being identified as a *cis*-eQTL to any gene in either microglia dataset or otherwise 0.

Histone modification-based functional annotation

We constructed 5 histone modification-based functional annotations using the epigenomics data of core histone modifications in the brain mid-frontal gyrus region from the Roadmap Epigenomics database:[7] H3K4me1 (primed enhancers), H3K4me3 (promoters), H3K36me3 (gene bodies), H3K27me3 (polycomb regression), and H3K9me3 (heterochromatin). For each histone modification, peak regions from replicates of the same sample were first merged and then overlapped with peak regions of other samples by Bedtools (v.2.27.0).[39] If a genetic variant resides in the overlapped peak regions of a histone modification, then 1 would be assigned to the function annotation of such a modification for this variant, or 0 would be assigned otherwise.

Simulation study design

We conducted simulation studies to validate the performance of BFGWAS_QUANT. Continuous phenotypes were simulated using the real whole-genome sequence (WGS) data of chromosomes 19 (122,745 SNPs with minor allele frequency [MAF] > 0.01) for 1,893 samples from the ROS/MAP cohort[33,40] and Mount Sinai Brain Bank (MSBB) study.[41] Phenotypes were simulated based on the multivariate linear additive model (Equation 1) with true genetic effect sizes $\beta_{p \times 1}$ generated based on the hierarchical logistic model with multivariate quantitative functional annotations (Equation 2). Scenarios with various numbers of true causals and heritability were considered.

Besides the real *cis*-eQTL-based functional annotations of Allcis-QTL, 95%CredibleSet, and MaxCPP, we also considered a fourth artificial annotation randomly generated from $N(0, 1)$ as a negative control. With chosen annotation enrichment parameters ($\alpha_0 = (-10.5, -9.5), \alpha_1 = 4, \alpha_2 = 1.5, \alpha_3 = 0.5, \alpha_4 = 0$), we first calculated casual probabilities (π_i) for all considered 122,745 SNPs by Equation 2, where α_0 was chosen to ensure the total number of true causal SNPs fall in (5, 10) with $\alpha_0 = -10.5$ or (15, 30) with $\alpha_0 = -9.5$. Second, a vector of binary indicator (γ_i) of true causal SNPs was generated from the corresponding Bernoulli distribution with probability (π_i) for $i = 1, ..., p$. Third, genetic effect sizes were taken as 0 for SNPs with $\gamma_i = 0$ or generated from a normal distribution for SNPs with $\gamma_i = 1$. Finally, phenotypes were generated from Equation 1 with simulated genetic effect sizes and random errors $\epsilon \sim N(0, (1 - h^2)\boldsymbol{I})$ to ensure that a target total heritability $h^2 = (0.25, 0.5)$ was equally explained by all true causal SNPs. Four scenarios were considered, including one with relatively sparse true causals in the range of (5, 10) and one with the number of true causals in the range of (15, 30) with respect to two different heritability values (0.25, 0.5).

We considered a null enrichment scenario where none of the annotations were enriched. In this scenario, we randomly selected 10 true causal SNPs and assigned them genetic effect sizes generated from a normal distribution. We simulated the phenotype as described above with a targeted $h^2 = 0.5$.

We repeated 100 simulations per scenario to evaluate our Bayesian estimates for annotation enrichment, total heritability, and true causal SNPs with respect to sensitivity (power) and positive predictive values (PPVs).[42] Sensitivity (power) is defined as the proportion of true positive findings among all true casual variants, and PPV is defined as the proportion of true positive findings among all identified significant associations. We took simulated true casual SNPs and those having $R^2 > 0.3$ with true casual SNPs with $\hat{\pi}_i > 0.1$ as true positive findings, following the significance rule used by the BFGWAS_CAT method.[22] The sensitivity and PPV are given by

$$Sensitivity \ (Power) = \frac{\#True \ Positive \ Findings}{\#True \ Causal \ SNPs};$$

$$PPV = \frac{\#True \ Positive \ Findings}{\# \ Positive \ Findings}.$$

We compared this with the standard BVSR method,[31] which does not account for functional annotations. We estimated the total heritability by the squared correlation between the simulated phenotypes and the PRSs based on Bayesian estimates of genetic effect sizes ($\hat{\beta}_i$) of SNPs with $\hat{\pi}_i > 0.01$,
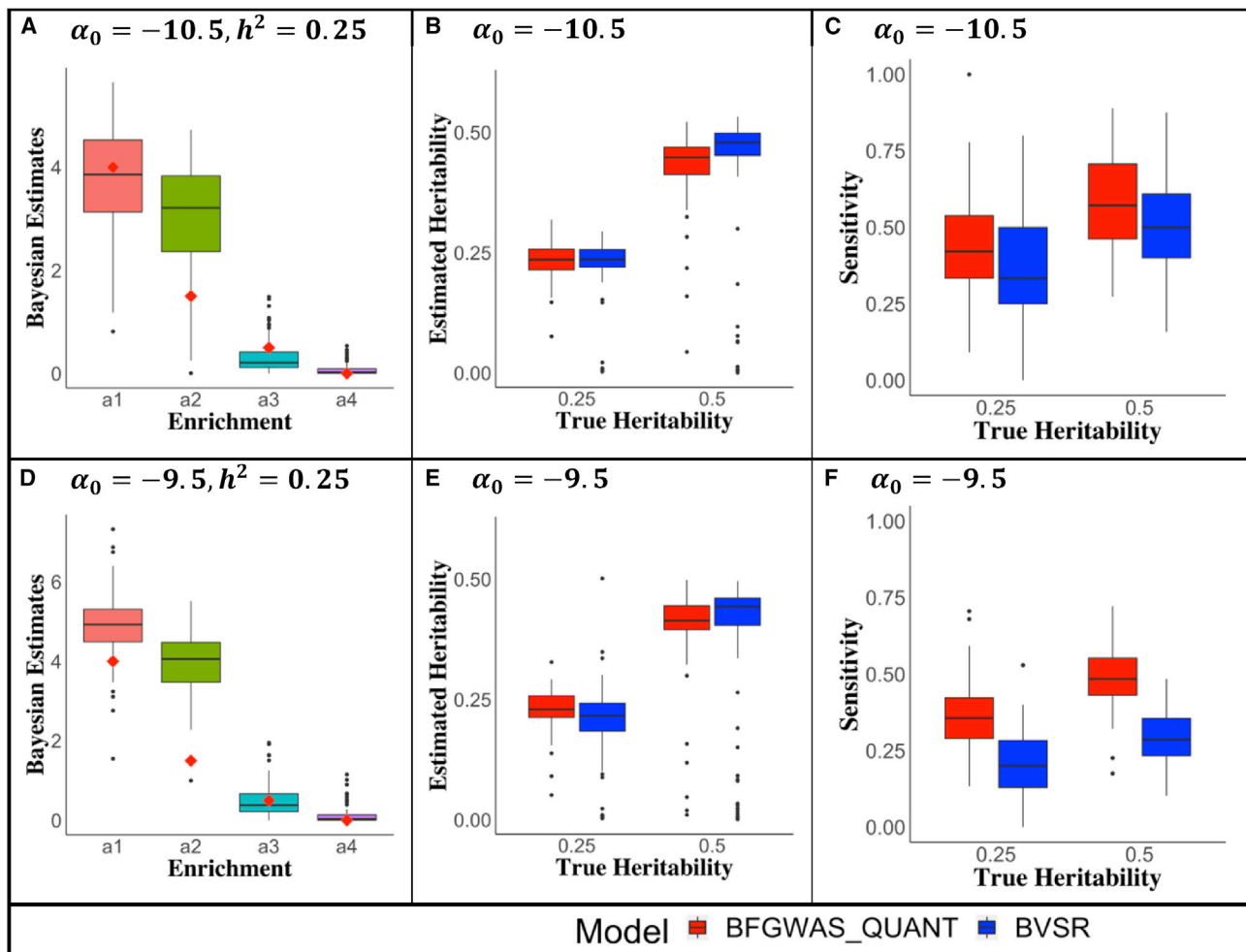
Figure 1. Bayesian enrichment estimates, heritability estimates, and sensitivities of simulation studies.
(A–F) Simulations with $\alpha_0 = -10.5$ (A–C) and simulations with $\alpha_0 = -9.5$ (D–F). Bayesian estimates of annotation enrichment (α_1, α_2, α_3, α_4) of 100 simulations with true heritability $h^2 = 0.25$ are shown in the respective boxplots (A and D), where red dots denote true enrichment values. Comparable heritability estimates (B and E) and higher sensitivities (C and F) were obtained by BFGWAS_QUANT (red) versus BVSR (blue).

$$PRS = \sum_{i=1}^{p} I_{\widehat{\pi}_i > 0.01} \widehat{\beta}_i X_i . \qquad \text{(Equation 3)}$$

Applications to ROS/MAP individual-level GWAS data

ROS/MAP are two prospective longitudinal community-based cohort studies that recruit older adults without known dementia at baseline and follow up with participants annually until the time of death.[33,40,43] Participants agree to annual clinical evaluation and brain autopsy at the time of death, signing an informed consent form and Anatomic Gift Act. All participants in this study also sign a repository consent to allow their data to be re-purposed. WGS data were profiled for 1,200 samples by using the KAPA Hyper Library Preparation Kit and Illumina HiSeq X sequencer.

We applied BFGWAS_QUANT to account for five eQTL-based and five histone modification-based quantitative annotations as described above to study five AD-related phenotypes,[9,32–34] including the binary clinical diagnosis of late-onset Alzheimer dementia (n = 1,087), three quantified postmortem pathology indices of AD (i.e., PHFtau tangle density with n = 1,105, β-amyloid load with n = 1,113, and a global measurement of AD pathol-

ogy burden with n = 1,123), and a quantitative measurement of cognition decline rate with n = 1,049.[33,40] The cognition decline rate was constructed as the random slope per sample from a linear mixed model of annual longitudinal measurements of cognition function. Details about ROS/MAP and phenotypic traits measured can be found in a previously published paper.[44] We also adjusted for the covariates of age, sex, smoking status, study index (ROS or MAP), and first 3 genotype principal components by regressing these covariates out from the phenotypes and taking the corresponding regression residuals as the outcome in the BFGWAS_QUANT method.

Application to IGAP summary-level GWAS data

We applied BFGWAS_QUANT to the stage 1 summary-level GWAS data of the International Genomics of Alzheimer's Project (IGAP),[45] along with the above 10 eQTL- and histone modification-based functional annotations and reference LD generated from ROS/MAP. The stage 1 IGAP summary-level GWAS data were generated by meta-analyses consisting of 17,008 individuals with AD and 37,154 control individuals of European ancestry (n = ~54,000).
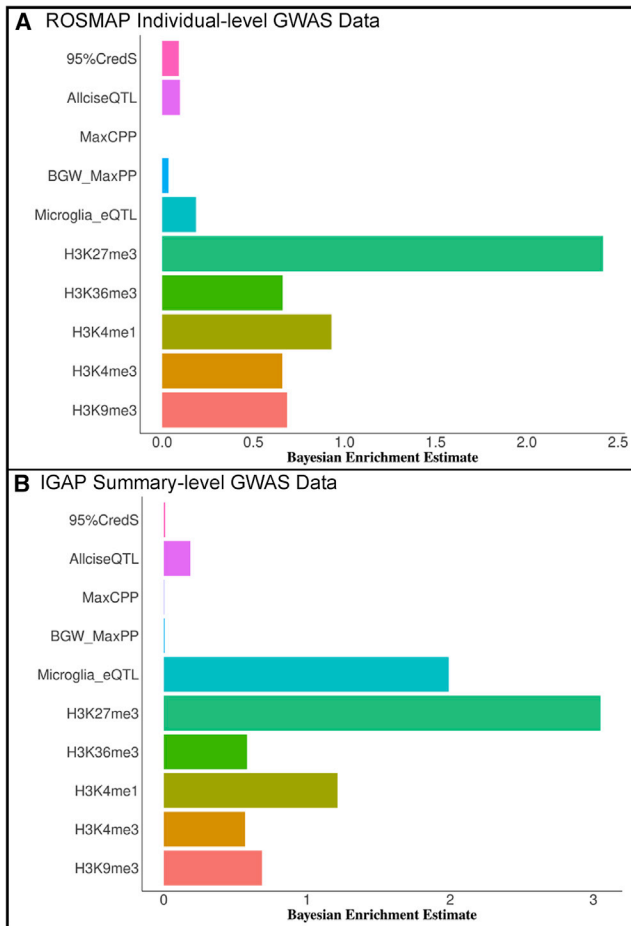
Figure 2. Bayesian estimates of functional annotation enrichment for Alzheimer dementia.
(A): Using ROS/MAP individual-level GWAS data and (B): Using IGAP summary-level GWAS data. Histone modification H3K27me3 (polycomb regression) and microglia *cis*-eQTL annotations were found to be most enriched for association signals of AD.

AD risk prediction by PRS

To show the usefulness of PRSs for risk prediction, we evaluated two sets of PRSs that were, respectively, derived from BFGWAS_QUANT and BVSR summary statistics using ROS/MAP GWAS data. We used the independent test samples from Mayo Clinic Alzheimer's Disease Genetics Studies (MCADGS).[46,47] MCADGS contain 2,099 European-descent samples (844 individuals with AD and 1,255 control individuals) with microarray genotype data profiled that were further imputed to the 1000 Genome Project Phase.[48]

We compared PRSs using Bayesian effect size estimates with three commonly used PRS methods: the standard method using informed LD pruning and p value thresholding (P + T),[49,50] LDpred2,[51] and PRS-CS.[52] p value thresholds of $(10^{-2}, 10^{-3}, 10^{-5}, 5 \times 10^{-8}, 10^{-8})$ and LD thresholds of $(0.1, 0.3, 0.5, 0.7, 0.9)$ were considered by the P + T method. Reference LD derived from 1000 Genome data were used by the LDpred and PRS-CS methods. AD risk prediction accuracy was evaluated using the area under the receiver operating characteristic (ROC)[53] curve (AUC) for MCADGC test samples.

Ethics statement

The ROS/MAP and MCADGS data analyzed in this study were generated with approval of the institutional review board (IRB)

of Rush University Medical Center, Chicago, IL, and Mayo Clinic, respectively. All samples analyzed in this study were de-identified, and all analyses were approved by the IRB of Emory University School of Medicine.

Results

Simulation results

For all considered simulation scenarios, our Bayesian estimates of functional annotation enrichment achieved convergence with 4 EM iterations, as shown in boxplots of 100 simulation replicates (Figures 1A, 1D, and S1). Although BFGWAS_QUANT overestimated α_1 and α_2, the Bayesian estimates still reflect the correct enrichment pattern among all considered annotations. By taking the 2.5th and 97.5th quantiles of these 100 Bayesian estimates to estimate the corresponding 2.5th and 97.5th quantiles of the estimator distributions, the true enrichment values ($\alpha_1 = 4$, $\alpha_2 = 1.5, \alpha_3 = 0.5, \alpha_4 = 0$) indeed fell within this range. For example, in the scenario with $\alpha_0 = -10.5$ and $h^2 = 0.25$ (Figure 1A), the estimated 2.5th and 97.5th quantiles are (1.62, 5.50) for α_1, (0.98, 4.57) for α_2, (0.02, 1.20) for α_3, and (0.00, 0.38) for α_4. We observed precise estimates of 0 enrichment for the artificial annotation (Figures 1 and S1) and all annotations in the scenario with null enrichment (Figure S3A), which demonstrated the ability of BFGWAS_QUANT to identify null enrichment.

By taking PRS as estimated phenotypes and taking the squared correlation between RPS and simulated phenotypes as the estimate of phenotype heritability, we obtained similar heritability estimates by BFGWAS_QUANT and BVSR, which are close to the true heritability (Figures 1B, 1E, and S3B). For scenarios with true enrichment, BFGWAS_QUANT obtained substantially higher sensitivity (power) and similar PPVs compared with BVSR for all scenarios (Figures 1C, 1F, S2, S4, and S5). For the scenario with null enrichment, BFGWAS_QUANT and BVSR performed comparably (Figures S3B–S3D).

These simulation studies validated the usefulness of BFGWAS_QUANT for quantifying multivariate functional annotations, estimating phenotype heritability, and identifying true causal SNPs. By accounting for multivariate quantitative annotations, BFGWAS_QUANT showed improved performance than the standard BVSR method, especially with higher power for identifying true causal SNPs and accurate enrichment estimation.

Application GWAS results for studying AD

Applying BFGWAS_QUANT to the individual-level ROS/MAP and summary-level IGAP GWAS data, we obtained consistent patterns for Bayesian enrichment estimates of 5 eQTL-based and 5 histone modification-based functional annotations (Figures 2, S6, and S7). In particular, the histone modification-based functional annotations had higher enrichment than eQTL-based annotations when studying the individual-level ROS/MAP GWAS data, with the highest enrichment for H3K27me3 and second highest

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

| CHR	rsID	Gene	Function	MAF	CPP	Beta	p Value	Phenotype
1	rs148348738[a]	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[a]	LOC 107,986,930	intron	0.148	0.276	0.015	6.93E−08	β-amyloid
19	rs34134669[a]	ADAMTS10	regulatory	0.234	0.119	−0.005	8.57E−07	cognition decline rate
19	rs769449	APOE TOMM40		0.111	0.121	0.076	3.45E−11	Alzheimer dementia
			regulatory	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		0.138	0.144	0.037	7.72E−13	Alzheimer dementia
				0.138	0.631	0.037	1.17E−20	tangle density
			missense	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 dementia
19	rs7256200	APOC1	regulatory	0.113	0.315	0.028	2.71E−12	Alzheimer 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[a]	APOC1	stop gained	0.435	0.119	0.039	1.06E−06	tangle density

[a]SNPs with a single variant test p value >5 × 10^{-8} that did not reach genome-wide significance by standard GWAS.

for H3K4me1. BFGWAS_QUANT estimated the second highest enrichment for the Microglia_eQTL annotation when studying the summary-level IGAP GWAS data with a larger sample size (n = ~54,000). Even with a small sample size in the individual-level ROS/MAP GWAS data, the Microglia_eQTL annotation was still identified with higher enrichment than other annotations based on eQTL of the bulk brain frontal cortex tissue. There is mounting evidence showing that microglia (composing <10% cells in the bulk brain frontal cortex tissue) play important roles in development and progression of AD pathology,[54] and cell-type-specific differential expression of GWAS risk genes of AD is only present in microglia.[55]

Using the ROS/MAP individual-level GWAS data, four significant SNPs with Bayesian CPP greater than 0.1068 were identified for AD (rs429358, rs10414043, rs769449, and rs7256200) by BFGWAS_QUANT (Table 1; Figures 3A, S8, and S9). In particular, SNP rs429358 (CPP = 0.144, p = 7.72 × 10^{-13}, missense variant) is the famous known APOE E4 risk allele of AD[56] and has a significant Bayesian CPP greater than 0.1068 for all 5 AD related phenotypes. SNPs rs10414043 (CPP = 0.111, p = 2.71 × 10^{-12}) and rs7256200 (CPP = 0.315, p = 2.71 × 10^{-12}, regulatory variant) are upstream of the known risk gene APOC1 of AD and blood protein traits.[37,57,58] Besides the missense APOE E4 risk alleles rs429358 and rs7412, one additional signifi-

cant SNP, rs1065853, is identified upstream of APOC1 for global AD pathology, which is a known GWAS signal for blood protein traits such as low-density lipoprotein.[58,59] Of 11 significant SNPs identified for at least one AD-related phenotype (Table 1), 2 are intergenic, and the other 9 SNPs are intron, regulatory, missense, and stop-gained variants.

Using the IGAP summary-level GWAS data, BFGWAS_QUANT fine-mapped 32 significant SNPs with Bayesian CPP greater than 0.1068 associated with AD (Table 2; Figure 3B). Multiple SNPs located in genes PVRL2, APOE/TOMM40, and APOC1 in chromosome 19 were found to be associated with AD with CPP = 1. Interestingly, 10 of 32 significant SNPs are cis-eQTL of microglia, and all significant SNPs except one intergenic SNP (rs7110631) are intron, regulatory, downstream, upstream, 3′ UTR, and missense variants (Table 2). All significant SNPs except rs78959900 are located in the peak regions of histone modification H3K27me3 (polycomb regression), which has the highest enrichment. Several SNPs that did not pass the genome-wide significance threshold (p < 5 × 10^{-8}) were identified by integrating 10 functional annotations in BFGWAS_QUANT. These SNPs are located in genes that have been found previously to be genetically linked to AD, such as HLA-DRB1, NUP160, SLC24A4, and CD33.[45,56,60–63]
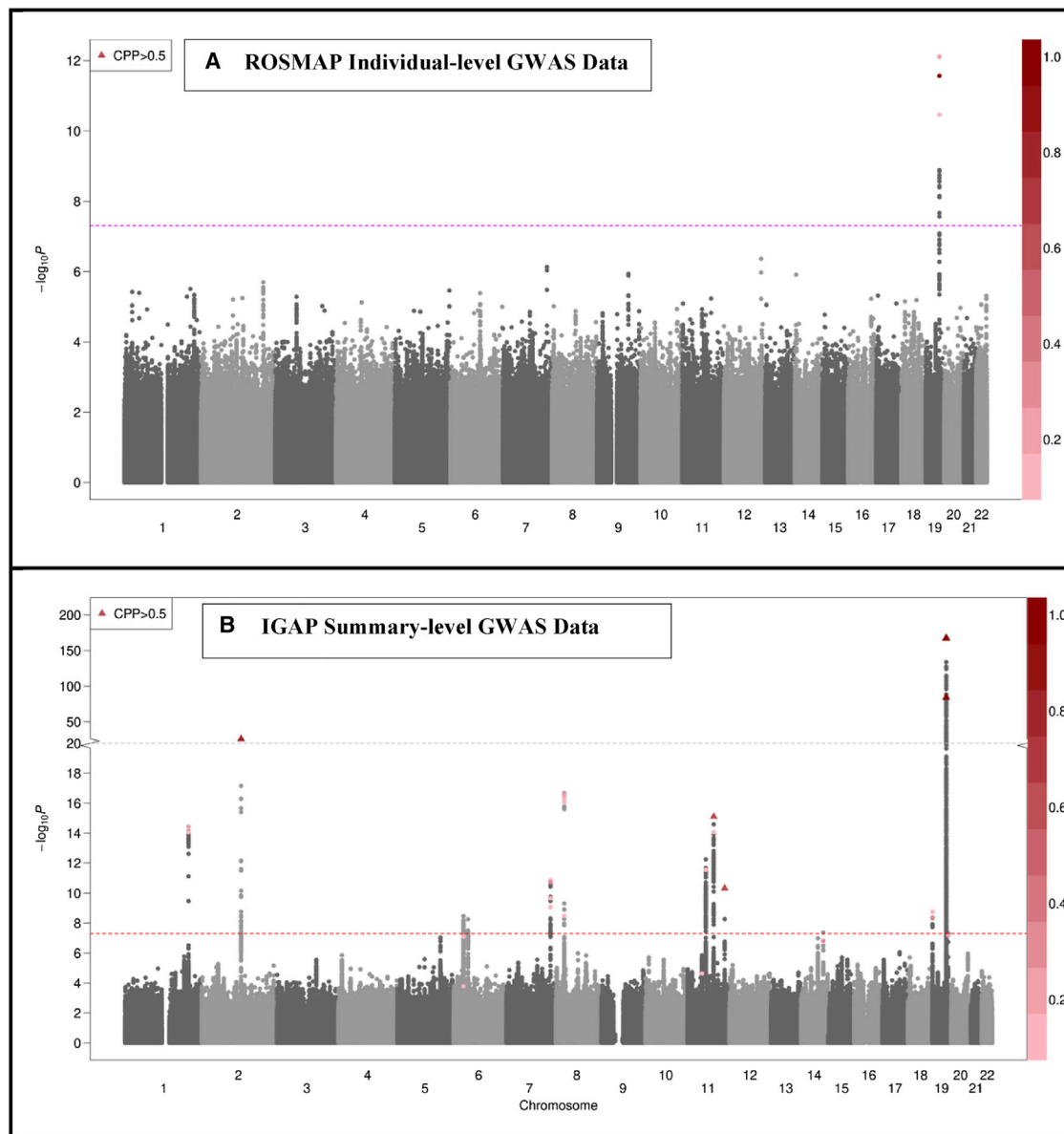
Figure 3. Manhattan plots of BFGWAS_QUANT results for studying Alzheimer dementia.
(A): Using ROS/MAP individual-level GWAS data; (B): Using IGAP summary-level GWAS data. Single-variant test p values were plotted in $-\log_{10}$ scale on the y axis. The dashed horizontal line denotes the genome-wide significant threshold 5×10^{-8}. SNPs with Bayesian CPP greater than 0.1068 were colored according to the color scale of their Bayesian CPP values by BFGWAS_QUANT. SNPs with Bayesian CPP greater than 0.5 were plotted as solid triangles.

The summation of genome-wide Bayesian CPP of SNPs with CPP greater than 0.01 can be used to estimate the number of total causal SNPs for the phenotype of interest. The threshold of CPP >0.01 is used to exclude adding CPP from random MCMC selections. Although the power is limited for analyzing the ROS/MAP individual-level GWAS data with a small sample size, BFGWAS_QUANT estimated a total of 54 potential causal SNPs for AD using the IGAP summary-level GWAS data (Table 3).

AD risk prediction by PRS in MCADGS

To show the usefulness of BFGWAS_QUANT for studying complex traits and diseases, we derived a PRS using the Bayesian effect size estimates by BFGWAS_QUANT for an independent GWAS cohort, MCADGS (n = 2,099), and compared the risk prediction accuracy with the PRS using effect size estimates by BVSR, P + T, LDpred2, and PRS-CS. When the ROS/MAP individual-level GWAS data were used as training data, a comparable AUC was obtained by using Bayesian effect size estimates by BFGWAS_QUANT (0.69) and BVSR (0.68), which was similar to the one obtained by LDpred2-auto (0.68) but significantly higher than the ones obtained by P + T (0.55), LDpred2-inf (0.53), and PRS-CS (0.54) (Figure 4A). This showed the advantage of deriving PRSs using Bayesian effect size estimates by BFGWAS_QUANT when individual-level GWAS

Table 2. Significant SNPs with Bayesian CPP > 0.1068 by BFGWAS_QUANT for studying AD using the IGAP summary-level GWAS data

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[a]	HLA-DRB1	intron	0.181	0.001	8.04E−08
6	rs9273472[a]	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[a]	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[a]	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[a]	CD33	missense	0.245	−0.027	6.66E−08

a: SNPs with single-variant test p value >5 × 10^{-8} that did not reach genome-wide significance by standard GWAS.

data are available, especially when the sample size is small. When the IGAP summary-level GWAS data were used as training data, the AUC by BFGWAS_QUANT (0.75) was the same as by LDpred2-auto (0.75) but much lower than P + T (0.88 with a p value threshold of 10^{-3}), LDpred2-inf (0.94), and PRS-CS (0.93) (Figure 4B). These results show that an infinitesimal model[64] as assumed by PRS-CS and LDpred2-inf is more suitable for PRS development than the sparse model assumed by BFGWAS_QUANT and LDpred2-auto.

Discussion

We developed a scalable BFGWAS method to account for multivariate quantitative functional annotations (BFGWAS_QUANT) for studying complex traits and diseases based on a hierarchical BVSR model and accompanied by a scalable EM-MCMC computation algorithm. BFGWAS_QUANT has the advantages of quantifying enrichment of functional annotations as well as modeling LD to generate fine-mapped GWAS results that are also

Table 3. Estimates of total causal SNPs

GWAS data	Phenotype	BFGWAS_QUANT	BVSR[a]
ROS/MAP	Alzheimer 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 dementia	54.282	–

The summations of the Bayesian CPP estimates of SNPs with CPP >0.01 estimate the total number of causal SNPs.
[a]BVSR was not developed for using summary-level GWAS data.

prioritized based on their functional annotations. BFGWAS_QUANT can be applied to individual-level and summary-level GWAS data. In particular, the Bayesian effect size estimates can be used to derive a PRS that accounts for functional annotations.

Our simulation studies validated the performance of BFGWAS_QUANT with respect to annotation enrichment quantification, GWAS association identification, and heritability estimation. Compared with BVSR, the BFGWAS_QUANT method had higher sensitivity (i.e., power) and comparable PPV and accuracy of phenotype heritability estimation.

In real studies of AD-related phenotypes using ROS/MAP individual-level and IGAP summary-level GWAS data, we showed that interesting enrichment patterns were identified, fine-mapped GWAS signals were identified, and predictive PRSs were derived. In particular, we found that the histone modification H3K27me3 (polycomb regression) and microglia *cis*-eQTL annotations were most en-riched for association signals of AD. We also showed that SNPs with single variant test p values $< 5 \times 10^{-8}$ could be identified for being prioritized because of their functional annotations.

Despite these advantages, BFGWAS_QUANT does have its limitations. First, the BFGWAS_QUANT model was developed for quantitative traits. However, following previous studies,[22,65] GWAS analysis can still be done for dichotomous traits by quantifying cases as 1 and controls as 0, which will have a similar performance as a probit model when samples are independent and population structure can be addressed by top genotype principal components. Extending the BFGWAS_QUANT method for studying dichotomous traits is also part of our ongoing research. Second, BFGWAS_QUANT assumes the summary-level GWAS data, and reference LDs were derived from populations of the same ancestry. BFGWAS_QUANT per ancestry needs to be applied first, and then the results have to be meta-analyzed for studying GWAS cohorts with
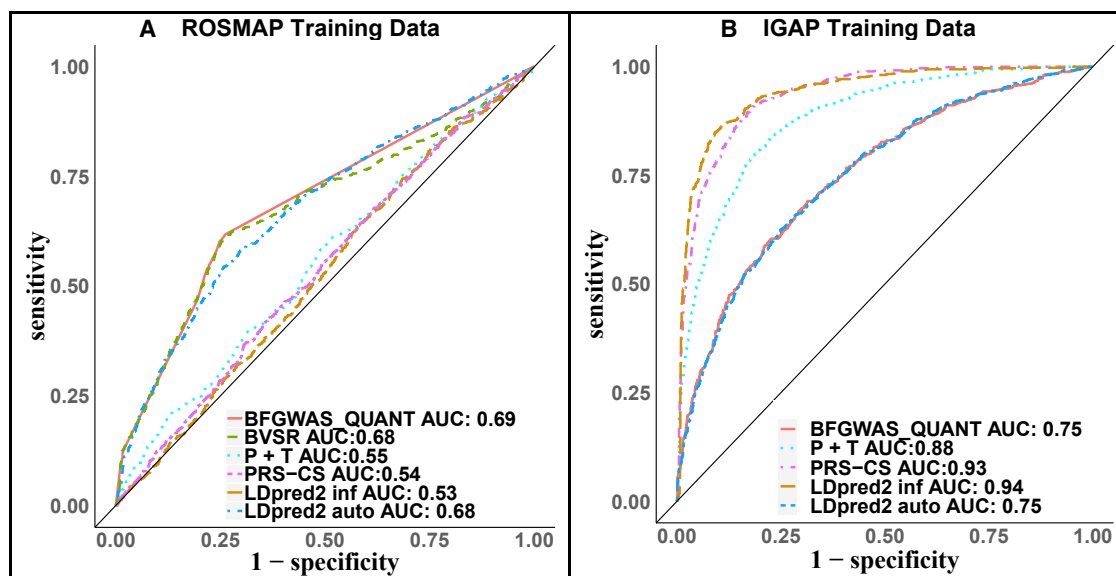
Figure 4. ROC plots comparing prediction accuracy of Alzheimer dementia in the test data of MCDGC.
(A): PRSs derived using the ROS/MAP individual-level GWAS data; (B): PRSs derived using IGAP summary-level GWAS data . The PRS derived using Bayesian effect size estimates by BFGWAS_QUANT has comparable prediction accuracy as the PRSs derived by BVSR and LDpred2 auto, for all assuming a sparse causal model. PRSs derived by PRS-CS and LDpred2-inf using IGAP summary-level GWAS data as training data have the highest prediction accuracy for assuming an infinitesimal causal model.

multiple ancestries. Third, BFGWAS_QUANT assumes a sparse causal genetic architecture that is suitable for generating fine-mapped GWAS results but might lack of power for deriving PRS for complex traits and diseases.

Our work demonstrated the usefulness of integrating multivariate quantitative functional annotations in GWASs for quantifying the enrichment of multiple functional annotations and generating fine-mapped GWAS results with higher power. Specifically, accurate quantification of annotation enrichment would help prioritize GWAS signals (fine-mapping) and then help illustrate the underlying genomic etiology of complex traits and diseases. Because publicly available molecular QTL datasets, epigenomic features, and GWAS summary data continue to grow, BFGWAS_QUANT provides a convenient tool for integrative multi-omics analyses of these datasets.

Data and code availability

ROS/MAP data can be requested through Rush Alzheimer's Disease Center (http://www.radc.rush.edu/) and Synapse:syn3219045 (https://www.synapse.org/#!Synapse:syn3219045). MCADGS data can be requested through Synapse:syn2910256 (https://www.synapse.org/#!Synapse:syn2910256). IGAP summary statistics are available from IGAP:http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php. Annotations derived from *cis*-eQTL of brain frontal cortex tissue are available from LDSC_QTL:https://alkesgroup.broadinstitute.org/LDSCORE/LDSC_QTL/. *cis*-eQTL data of microglia are available from Zenodo:6104982 (https://zenodo.org/record/6104982) and Zenodo:4118605 (https://zenodo.org/record/4118605). Source code of BFGWAS_QUANT is available through Github (https://github.com/yanglab-emory/BFGWAS_QUANT).

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2022.100143.

Declaration of interests

D.A.B. receives a consulting fee from B4X Inc. for advising on neurodegenerative disease.

Web resources

BFGWAS_QUANT Github directory, https://github.com/yanglab-emory/BFGWAS_QUANT

LDpred, https://github.com/bvilhjal/ldpred

MCADGC data, https://www.synapse.org/#!Synapse:syn2910256

PRS-CS, https://github.com/getian107/PRScs

RADC Research Resource Sharing Hub, http://www.radc.rush.edu/

Roadmap, http://www.roadmapepigenomics.org/

ROS/MAP data, https://www.synapse.org/#!Synapse:syn3219045

References

1. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. Am. J. Hum. Genet. *101*, 5–22.
2. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001–D1006.
3. McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A., and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. *9*, 356–369.
4. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA *106*, 9362–9367.
5. Gallagher, M.D., and Chen-Plotkin, A.S. (2018). The post-GWAS era: from association to function. Am. J. Hum. Genet. *102*, 717–730.
6. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. *47*, D886–D894.
7. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.
8. The ENCODE Project Consortium (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. *9*, e1001046.

9. The GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.

10. McVicker, G., van de Geijn, B., Degner, J.F., Cain, C.E., Banovich, N.E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., and Pritchard, J.K. (2013). Identification of genetic variants that affect histone modifications in human cells. Science *342*, 747–749.

11. Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat. Biotechnol. *28*, 817–825.

12. Robins, C., Liu, Y., Fan, W., Duong, D.M., Meigs, J., Hererimana, N.V., Gerasimov, E.S., Dammer, E.B., Cutler, D.J., Beach, T.G., et al. (2021). Genetic control of the human brain proteome. Am. J. Hum. Genet. *108*, 400–410.

13. Banovich, N.E., Lan, X., McVicker, G., van de Geijn, B., Degner, J.F., Blischak, J.D., Roux, J., Pritchard, J.K., and Gilad, Y. (2014). Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. PLoS Genet. *10*, e1004663.

14. Waszak, S.M., Delaneau, O., Gschwind, A.R., Kilpinen, H., Raghav, S.K., Witwicki, R.M., Orioli, A., Wiederkehr, M., Panousis, N.I., Yurovsky, A., et al. (2015). Population variation and genetic control of modular chromatin architecture in humans. Cell *162*, 1039–1050.

15. Heyn, H. (2016). Quantitative trait loci identify functional noncoding variation in cancer. PLoS Genet. *12*, e1005826.

16. Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. Genetics *198*, 497–508.

17. Le, K.T.T., Matzaraki, V., Netea, M.G., Wijmenga, C., Moser, J., and Kumar, V. (2019). Functional annotation of genetic loci associated with sepsis prioritizes immune and endothelial cell pathways. Front. Immunol. *10*, 1949.

18. Matzaraki, V., Gresnigt, M.S., Jaeger, M., Ricaño-Ponce, I., Johnson, M.D., Oosting, M., Franke, L., Withoff, S., Perfect, J.R., Joosten, L.A.B., et al. (2017). An integrative genomics approach identifies novel pathways that influence candidaemia susceptibility. PLoS One *12*, e0180824.

19. Ruffieux, H., Fairfax, B.P., Nassiri, I., Vigorito, E., Wallace, C., Richardson, S., and Bottolo, L. (2021). EPISPOT: an epigenome-driven approach for detecting and interpreting hotspots in molecular QTL studies. Am. J. Hum. Genet. *108*, 983–1000.

20. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. *6*, e1000888.

21. Ng, B., White, C.C., Klein, H.U., Sieberts, S.K., McCabe, C., Patrick, E., Xu, J., Yu, L., Gaiteri, C., Bennett, D.A., et al. (2017). An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. Nat. Neurosci. *20*, 1418–1426.

22. Yang, J., Fritsche, L.G., Zhou, X., Abecasis, G.; and International Age-Related Macular Degeneration Genomics Consortium (2017). A scalable bayesian method for integrating functional information in genome-wide association studies. Am. J. Hum. Genet. *101*, 404–416.

23. Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am. J. Hum. Genet. *94*, 559–573.

24. Trynka, G., Westra, H.-J., Slowikowski, K., Hu, X., Xu, H., Stranger, B.E., Klein, R.J., Han, B., and Raychaudhuri, S. (2015). Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex-trait loci. Am. J. Hum. Genet. *97*, 139–152.

25. Iversen, E.S., Lipton, G., Clyde, M.A., and Monteiro, A.N.A. (2014). Functional annotation signatures of disease susceptibility loci improve SNP association analysis. BMC Genom. *15*, 398.

26. Chen, G.K., and Witte, J.S. (2007). Enriching the analysis of genomewide association studies with hierarchical modeling. Am. J. Hum. Genet. *81*, 397–404.

27. Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. *10*, e1004722.

28. Kichaev, G., Roytman, M., Johnson, R., Eskin, E., Lindström, S., Kraft, P., and Pasaniuc, B. (2017). Improved methods for multi-trait fine mapping of pleiotropic risk loci. Bioinformatics *33*, 248–255.

29. Ma, Y., and Wei, P. (2019). FunSPU: a versatile and adaptive multiple functional annotation-based association test of whole-genome sequencing data. PLoS Genet. *15*, e1008081.

30. Li, X., Li, Z., Zhou, H., Gaynor, S.M., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D.K., Aslibekyan, S., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. Nat. Genet. *52*, 969–983.

31. Guan, Y., and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. Ann. Appl. Stat. *5*, 1780–1815.

32. De Jager, P.L., Ma, Y., McCabe, C., Xu, J., Vardarajan, B.N., Felsky, D., Klein, H.U., White, C.C., Peters, M.A., Lodgson, B., et al. (2018). A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. Sci. Data *5*, 180142.

33. Bennett, D.A., Buchman, A.S., Boyle, P.A., Barnes, L.L., Wilson, R.S., and Schneider, J.A. (2018). Religious orders study and rush memory and aging project. J. Alzheimers Dis. *64*, S161–S189.

34. Hormozdiari, F., Gazal, S., van de Geijn, B., Finucane, H.K., Ju, C.J.T., Loh, P.R., Schoech, A., Reshef, Y., Liu, X., O'Connor, L., et al. (2018). Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. Nat. Genet. *50*, 1041–1047.

35. Bryois, J., Calini, D., Macnair, W., Foo, L., Urich, E., Ortmann, W., Iglesias, V.A., Selvaraj, S., Nutma, E., Marzin, M., et al. (2021). Cell-type specific cis-eQTLs in eight brain cell-types identifies novel risk genes for human brain disorders. Preprint at medRxiv. https://doi.org/10.1101/2021.10.09.21264604.

36. Lopes, K.d.P., Snijders, G.J.L., Humphrey, J., Allan, A., Sneeboer, M.A.M., Navarro, E., Schilder, B.M., Vialle, R.A., Parks, M., Missall, R., et al. (2022). Genetic analysis of the human microglial transcriptome across brain regions, aging and disease pathologies. Nat. Genet. *54*, 4–17.

37. Luningham, J.M., Chen, J., Tang, S., De Jager, P.L., Bennett, D.A., Buchman, A.S., and Yang, J. (2020). Bayesian genome-wide TWAS method to leverage both cis- and trans-eQTL information through summary statistics. Am. J. Hum. Genet. *107*, 714–726.

38. BROYDEN, C.G. (1970). The convergence of a class of double-rank minimization algorithms 1. General considerations. IMA J. Appl. Math. *6*, 76–90.

39. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

40. Bennett, D.A., Schneider, J.A., Arvanitakis, Z., and Wilson, R.S. (2012). Overview and findings from the religious orders study. Curr. Alzheimer Res. *9*, 628–645.

41. Wang, M., Beckmann, N.D., Roussos, P., Wang, E., Zhou, X., Wang, Q., Ming, C., Neff, R., Ma, W., Fullard, J.F., et al. (2018). The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. Sci. Data *5*, 180185.

42. Trevethan, R. (2017). Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. Front. Public Health *5*, 307.

43. Bennett, D.A., Schneider, J.A., Buchman, A.S., Barnes, L.L., Boyle, P.A., and Wilson, R.S. (2012). Overview and findings from the rush memory and aging project. Curr. Alzheimer Res. *9*, 646–663.

44. Bennett, D.A., Wilson, R.S., Boyle, P.A., Buchman, A.S., and Schneider, J.A. (2012). Relation of neuropathology to cognition in persons without cognitive impairment. Ann. Neurol. *72*, 599–609.

45. Lambert, J.-C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., DeStafano, A.L., Bis, J.C., Beecham, G.W., Grenier-Boley, B., et al. (2013). Meta-analysis of 74, 046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat. Genet. *45*, 1452–1458.

46. Carrasquillo, M.M., Zou, F., Pankratz, V.S., Wilcox, S.L., Ma, L., Walker, L.P., Younkin, S.G., Younkin, C.S., Younkin, L.H., Bisceglio, G.D., et al. (2009). Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. Nat. Genet. *41*, 192–198.

47. Allen, M., Carrasquillo, M.M., Funk, C., Heavner, B.D., Zou, F., Younkin, C.S., Burgess, J.D., Chai, H.S., Crook, J., Eddy, J.A., et al. (2016). Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. Sci. Data *3*, 160089.

48. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1, 092 human genomes. Nature *491*, 56–65.

49. Choi, S.W., Mak, T.S.H., and O'Reilly, P.F. (2020). Tutorial: a guide to performing polygenic risk score analyses. Nat. Protoc. *15*, 2759–2772.

50. Euesden, J., Lewis, C.M., and O'Reilly, P.F. (2015). PRSice: polygenic risk score software. Bioinformatics *31*, 1466–1468.

51. Privé, F., Arbel, J., and Vilhjálmsson, B.J. (2020). LDpred2: better, faster, stronger. Bioinformatics *36*, 5424–5431.

52. Ge, T., Chen, C.Y., Ni, Y., Feng, Y.C.A., and Smoller, J.W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat. Commun. *10*, 1776.

53. Linden, A. (2006). Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. J. Eval. Clin. Pract. *12*, 132–139.

54. Hansen, D.V., Hanson, J.E., and Sheng, M. (2018). Microglia in Alzheimer's disease. J. Cell Biol. *217*, 459–472.

55. Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J.Z., Menon, M., He, L., Abdurrob, F., Jiang, X., et al. (2019). Single-cell transcriptomic analysis of Alzheimer's disease. Nature *570*, 332–337.

56. Kunkle, B.W., Grenier-Boley, B., Sims, R., Bis, J.C., Damotte, V., Naj, A.C., Boland, A., Vronskaya, M., van der Lee, S.J., Amlie-Wolf, A., et al. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nat. Genet. *51*, 414–430.

57. Takei, N., Miyashita, A., Tsukie, T., Arai, H., Asada, T., Imagawa, M., Shoji, M., Higuchi, S., Urakami, K., Kimura, H., et al. (2009). Genetic association study on in and around the APOE in late-onset Alzheimer disease in Japanese. Genomics *93*, 441–448.

58. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514–518.

59. Andaleon, A., Mogil, L.S., and Wheeler, H.E. (2018). Gene-based association study for lipid traits in diverse cohorts implicates BACE1 and SIDT2 regulation in triglyceride levels. PeerJ *6*, e4314.

60. Naj, A.C., Jun, G., Beecham, G.W., Wang, L.-S., Vardarajan, B.N., Buros, J., Gallins, P.J., Buxbaum, J.D., Jarvik, G.P., Crane, P.K., et al. (2011). Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat. Genet. *43*, 436–441.

61. Karch, C.M., Jeng, A.T., Nowotny, P., Cady, J., Cruchaga, C., and Goate, A.M. (2012). Expression of novel Alzheimer's disease risk genes in control and Alzheimer's disease brains. PLoS One *7*, e50976.

62. Lu, R.C., Yang, W., Tan, L., Sun, F.R., Tan, M.S., Zhang, W., Wang, H.F., and Tan, L. (2017). Association of HLA-DRB1 polymorphism with Alzheimer's disease: a replication and meta-analysis. Oncotarget *8*, 93219–93226.

63. Novikova, G., Kapoor, M., Tcw, J., Abud, E.M., Efthymiou, A.G., Chen, S.X., Cheng, H., Fullard, J.F., Bendl, J., Liu, Y., et al. (2021). Integration of Alzheimer's disease genetics and myeloid genomics identifies disease risk regulatory elements and genes. Nat. Commun. *12*, 1610.

64. Nelson, R.M., Pettersson, M.E., and Carlborg, Ö. (2013). A century after Fisher: time for a new paradigm in quantitative genetics. Trends Genet. *29*, 669–676.

65. Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. *9*, e1003264.