bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A scalable Bayesian method for integrating functional information in genome-wide association studies

Jingjing Yang[1], Lars G. Fritsche[1,2], Xiang Zhou[1,*], Gonçalo Abecasis[1,*] for the International Age-related Macular Degeneration Genomics Consortium (IAMDGC)

[1]Center for Statistical Genetics, Department of Biostatistics, University of Michigan School of Public Health, 1415 Washington Heights, Ann Arbor, MI 48109, USA.

[2]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health, NTNU, Norwegian University of Science and Technology, Trondheim, Norway.

*Correspondence to: X.Z. (xzhousph@umich.edu) or G.A. (goncalo@umich.edu)

Abstract: (164 words)

Genome-wide association studies (GWASs) have identified many complex trait loci. To understand the biological mechanisms underlying these, we pair a flexible Bayesian method with efficient computational techniques to model functional information in GWASs. We model the effect-size distribution and probability of causality for variants with different annotation, explicitly allowing for multiple causal-variants per locus. In simulations, our method shows higher power to identify true causal-variants than competing methods. In a GWAS of age-related macular degeneration with 33,976 individuals and 9,857,286 variants, we find the strongest enrichment for causality among non-synonymous variants (54x more likely to be causal, 1.4x larger effect-sizes) and among variants in active promoters (7.8x more likely, 1.4x larger effect-sizes). Importantly, when multiple causal-variants reside in the same locus, our approach improves upon the list of candidate variants produced by sequential forward selection or methods only allowing for a single causal-variant per locus. In conclusion, our method is shown to efficiently integrate functional information in GWASs, helping identify causal-variants and underlying biology.

Keywords: functional annotation, genome-wide association study (GWAS), Bayesian variable selection regression (BVSR), expectation-maximization (EM), Markov chain Monte Carlo (MCMC).

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genome-wide association studies (GWASs) have identified thousands of genetic loci for complex traits and diseases, providing new insights into the underlying genetic architecture[1-5]. Each associated locus typically contains hundreds of variants in linkage disequilibrium (LD)[6,7], most of which are of unknown function and located outside protein-coding regions. Unsurprisingly, the biological mechanisms underlying the identified associations are often unclear[8] and pinpointing causal variants is difficult[9].

Recent functional genomic studies help understand and pinpoint causal variants and mechanisms[10-12]. Genetic variants can be annotated based on the genomic location (e.g., coding, intronic, and intergenic), role in determining protein structure and function (e.g., Sorting Intolerant From Tolerant (SIFT)[13] and Polymorphism Phenotyping (PolyPhen)[14] scores), ability to regulate gene expression (e.g., expression quantitative trait loci (eQTL) and allelic specific expression (ASE) evidence[15,16]), biochemical function (e.g., DNase I hypersensitive sites (DHS), metabolomic QTL (mQTL) evidence[17], and chromatin states[18-20]), evolutionary significance (e.g., Genomic Evolutionary Rate Profiling (GERP) annotations[21]), and a combination of different types of annotation (e.g., CADD[22]). Many statistical methods, including stratified LD score regression[23] and MINQUE[24], can now evaluate the role of functional annotations in GWASs through heritability analysis. Preliminary studies also show higher proportions of associated variants in protein-coding exons, regulatory regions, and cell-type-specific DHSs[25-27].

Integrating functional information into GWASs is expected to help identify and prioritize true causal associations. However, accomplishing this goal in practice requires methods to account for both LD and computational cost. Consider two recent methods, Fgwas[26] and PAINTOR[27], as examples: Fgwas assumes that variants are independent and there is at most one causal variant per locus, modeling no LD, which dramatically improves computational speed and allows Fgwas to be applied at genome-wide scale; PAINTOR accounts for LD, assuming the possibility of multiple association signals per locus, but is computationally slow and can only be used to fine-map small regions.

Here, we pair a flexible Bayesian method with an efficient computational algorithm. Together the two represent an attractive means to incorporate functional information into association

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mapping. Our model accounts for genotype correlation due to LD, allows for multiple causal variants per locus and, importantly, shares information genome-wide to increase association-mapping power. Our algorithm takes advantage of the local LD structure in the human genome[28-30] and refines previous Markov chain Monte Carlo (MCMC) algorithms to greatly improve mixing, which is key when searching for causal variants among many associated variants in LD (but less important in other applications such as modeling total genomic heritability). Because of these features, we refer to our method as the Scalable Functional Bayesian Association (SFBA). Below, we illustrate the benefits of SFBA with extensive simulations and real data analyses of a large-scale GWAS on age-related macular degeneration (AMD)[31] with 33,976 individuals and 9,857,286 genotyped or imputed variants. Our method is implemented in the software SFBA, freely available at https://github.com/yjingj/SFBA.

RESULTS

Method overview

Our method is based on the standard Bayesian variable selection regression (BVSR) model (Online Methods and Supplementary Information; Supplementary Figure 1(a)), allowing for annotations that classify variants into K non-overlapping categories. We assume that variants in annotation category q share a "spike-and-slab" prior[32,33] for effect-sizes, $\beta_i \sim \pi_q N\left(0, \tau^{-1}\sigma_q^2\right) + \left(1-\pi_q\right)\delta_0(\beta_i)$. This model implies effect sizes are normally distributed as $\beta_i \sim N\left(0, \tau^{-1}\sigma_q^2\right)$ with probability π_q, or set to zero with probability $(1-\pi_q)$, with $\delta_0(\beta_i)$ denoting the point-mass function at 0. Here, π_q represents the (unknown) causal probability for variants in the qth category and σ_q^2 represents the (unknown) corresponding effect-size variance. An enhancement to previous Bayesian models[33-35] is that we model both the proportion of associated variants and their effect-size distribution in each annotation category.

Our goal is to simultaneously make inference on category specific parameters (π_q, σ_q^2) that represent the importance of each functional category, and on the variant specific parameters —— effect-size β_i and the probability of $\beta_i \neq 0$ (referred as posterior inclusion probability (PP_i),

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

representing association evidence). Our model shares information among genome-wide to estimate category specific parameters, which then inform the variant specific parameters. As a result, variant associations will be prioritized based on the inferred importance of functional categories.

Because standard MCMC algorithms suffer from heavy computational burden and poor mixing of posterior samples for large GWASs, we develop a novel scalable expectation-maximization MCMC (or EM-MCMC) algorithm. Our algorithm is based on the observation that LD decays exponentially with distance and displays local block-wise structure along the human genome[28-30,36,37]. This observation allows us to decompose the complex joint likelihood of our model into a product of block-wise likelihoods (Online Methods and Supplementary Information). Intuitively, conditional on a common set of category specific parameters (π_q, σ_q^2), we can infer (β_i, PP_i) by running the MCMC algorithm per genome-block. A diagram of this EM-MCMC algorithm is shown in Supplementary Figure 1(b).

Running MCMC per genome-block facilitates parallel computing and reduces the search space. Unlike previous MCMC algorithms for GWAS that use proposal distributions based only on marginal association evidence (such as implemented in GEMMA[38]), our MCMC algorithm uses a proposal distribution that favors variants near the “causal” variants being considered in each iteration, and prioritizes among these neighboring variants based on their conditional association evidence (see Supplementary Information). Our strategy dramatically improves the MCMC mixing property, encouraging our method to explore different combinations of potentially causal variants in each locus (Supplementary Figure 2). In addition, we implemented memory reduction techniques that reduce memory usage up to 97%, effectively reducing the required physical memory from 120 GB (usage by GEMMA[38]) to 3.6 GB for a GWAS with ~33K individuals and ~400K genotyped variants (Online Methods and Supplementary Information).

In practice, we segment the whole genome into blocks of 5,000 ~ 10,000 variants, based on marginal association evidence, genomic distance, and LD. We always ensure variants in LD (R^2 >0.1) with significant signals (P-values <5×10^{-8}) are in the same block (Online Methods). We first initialize category specific parameters (π_q, σ_q^2), then run the MCMC algorithm per block (E-step),

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

summarize the MCMC posterior estimates of (β_i, PP_i) across all blocks to update (π_q, σ_q^2) (M-step), and repeat the block-wise EM-MCMC steps until (π_q, σ_q^2) estimates converge (Supplementary Figure 1(b)).

In addition, we calculate the regional posterior inclusion probability (regional-PP) per block that is the proportion of MCMC iterations with at least one "causal" variant (see Supplementary Information). Because Bayesian PP might be split among multiple variants in high LD, the threshold of regional-PP >0.95 (conservatively analogous to false discovery rate 0.05) is used for identifying loci.

Simulation

We simulated phenotypes with the genotype data (chromosomes 20-22) from the AMD GWAS[31], including 33,976 individuals and 241,500 variants with minor allele frequency (MAF) >0.1. We segmented this small genome into 50 x 2.5Mb blocks, each with ~5,000 variants. Within each block, we marked a 25KB continuous region (starting 37.5Kb from the beginning of a block) as the causal locus and randomly selected two causal single nucleotide polymorphisms (SNPs) per locus. We simulated two complementary annotations to classify variants into "coding" and "noncoding" groups, where the coding variants account for ~1% overall variants but ~10% variants within the causal loci (matching the pattern in the real AMD data). We simulated two scenarios: (i) coding variants ~44x enriched among causal variants (30 coding vs. 70 noncoding); (ii) no enrichment (1 coding vs. 99 noncoding). A total of 15% of phenotypic variance was divided equally among causal variants. We compared SFBA with single variant likelihood-ratio test, conditional analysis (CA), and Fgwas. The single variant test P-value (also referred to as P-value), conditioned P-value, Fgwas posterior association probability (PP, see Online Methods), and our Bayesian PP were used as criteria to identify associations.

We first compared power of different methods using average ROC curves[27,33] across 100 simulation replicates. Fgwas was more powerful than P-value at low false-positive rates (FPR), presumably because Fgwas incorporates annotation information (Figure 1(a)). However, with high

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

false-positive rates, Fgwas underperformed P-value, presumably because Fgwas incorrectly assumes one variant per locus. In contrast, SFBA (modeling LD and allowing multiple causal variants per locus) outperformed both Fgwas and P-value for false-positive rates in (0, 0.01). Importantly, the advantage of SFBA became more pronounced with increasing sample size (Supplementary Figure 3). Specifically, the power (based on FPR=0.5%) of SFBA increased from 48% to 64% as the sample size increased from 20K to 33K, while the power of Fgwas only increased from 52% to 56% and the power of P-values only increased from 47% to 52%. In addition, with sample size 33K and the threshold of regional-PP >0.95, SFBA has power 92.3% to identify associated loci, versus Fgwas with 88.6% power. The advantage of SFBA with large sample size suggests that SFBA can better extract the richer information available as sample size increases.

In a typical GWAS, researchers identify a series of associated loci and then examine associated variants within each locus independently. We examined the ability of each method to prioritize the true causal variants in each locus. Since we simulated two causal SNPs per locus (SNP1 and SNP2), we examine the power for identifying each of these separately (Figure 1(b)). All methods have the same median rank for causal SNP1 (typically, ranked 3rd rank among 150 SNPs in the locus by P-value, Fgwas and SFBA), suggesting that the strongest signal in a locus can often be identified without incorporating functional information. The median rank for the second causal SNP2 was the 7th by SFBA, 12th by Fgwas, 17th by P-value, and 18th by conditional analysis — suggesting that incorporating functional information improves power to identify multiple signals in a locus. Stratified results based on the LD between two causal variants further demonstrate that SFBA has the highest power for identifying the weaker signal, especially when both SNPs are in high LD (Supplementary Figure 4).

Both SFBA and Fgwas correctly identified enrichment in scenario (i) and properly controlled for the type I error of enrichment in scenario (ii), despite some numerical issues for Fgwas (Supplement Figure 5). Moreover, SFBA estimated the effect-size variance per annotation. For all 100 simulation replicates under both scenarios, the 95% confidence intervals of the log-ratio of estimated effect-size variances between coding and noncoding overlapped with 0 (Supplementary

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 6), suggesting effect-size variances were similar between two annotations (matching the simulated truth).

In summary, our simulation studies show that, in comparisons with competing methods, SFBA has higher power, especially in loci with multiple associated variants and when the sample size is large. Further, SFBA produces enrichment parameter estimates that can help with interpretation of association results.

GWAS of AMD

Next, we applied our method to a GWAS of age-related macular degeneration (AMD) with 16,144 advanced cases and 17,832 controls, for a total of 33,976 unrelated European individuals. A total of 439,350 variants were genotyped on a customized Exome-Chip, and then imputed up to 12,023,830 variants in 1000 Genomes Project Phase 1[39,40]. We analyzed 9,866,744 (~10M) low-frequency and common variants (MAF >0.5%) with three types of genomic annotations: gene-based functional annotations by SeattleSeq, summarized regulatory annotations[41], and the chromatin states profiled in nine human cell types from chromHMM[42,43].

Coding variation and AMD.

We used SeattleSeq to classify variants according to their impact on coding sequences (Supplementary Table 1) and then applied our method SFBA and Fgwas. SFBA identified 37 loci out of 1,063 considered genome-blocks with regional-PP >0.95 (Supplementary Tables 2, 3, and 5), including 32 among the 34 known AMD loci[31] and 5 potentially novel loci. Using the threshold of Bayesian PP >0.1068 (roughly equivalent to the P-value 5×10^{-8} based on permutations of AMD data; Supplementary Figure 7), we identified 150 associated variants (Supplementary Figure 9(a); Supplementary Table 3), with 47 distributed among 42,005 non-synonymous variants, 4 among 67,165 synonymous coding variants, 54 among 3,679,235 intronic variants, 18 among 5,512,423 intergenic variants (including non-annotated variants), and 27 among 565,916 “other-genomic” variants (UTR, non-coding exons, upstream and downstream of genes). Very roughly, this corresponds to fraction of associated variants of ~1:1,000 among non-synonymous variants,

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1:15,000 among synonymous variants, 1:100,000 among intronic variants, 1:300,000 among intergenic variants and 1:20,000 among "other-genomic" variants.

Similarly, Fgwas identified 46 loci by regional-PP >0.95, including all 34 known loci and 12 potentially novel loci (Supplementary Tables 2, 4, and 6; Supplementary Figure 9(b)). Since Fgwas analyzed the whole genome as 4,934 segments (each with 2,000 variants) and, thus, partitioned the genome somewhat differently than our method. Fgwas identified 178 associated variants with Fgwas PP >0.1068, including 24 non-synonymous, 13 coding-synonymous, 42 intronic, 40 intergenic, and 59 other-genomic signals. Compared with SFBA, the proportion of loci that contain at least one non-synonymous variant with PP >0.1068 is significantly smaller (11 out of 46 by Fgwas vs. 18 of 37 by SFBA; P-value = 0.017). Similarly, the proportion of non-synonymous variants prioritized by Fgwas is also significantly smaller (24 out of 178 by Fgwas vs. 47 of 150 by SFBA; P-value =7.7×10^{-5}), indicating that SFBA places greater weight on coding variants —— which, as a group, appears to have both a higher prior probability of association and larger effect sizes when associated.

Besides replicating the association results within known AMD loci[31], SFBA identified five novel loci (Supplementary Table 5): missense *rs7562391/PPIL3*, *rs61751507/CPN1*, *rs2232613/LBP*, downstream *rs114348558/ZNRD1-AS1*, and splice *rs6496562/ABHD2*. These loci were also identified by Fgwas (Supplementary Table 6) with different top association variants for *CPN1* (coding-synonymous *rs61733667*) and *ZNRD1-AS1* (downstream *rs116112857*). Interestingly, there are several connections between these potentially novel loci and known AMD loci. For example, the protein encoded by *LBP* is part of the lipid transfer protein family (which also includes *CETP* among the known AMD risk loci) that promotes the exchange of neutral lipids and phospholipids between plasma lipoproteins[46]. Similarly, *ZNRD1-AS1* has been associated with lipid metabolisms[47] and *ABHD2* has been associated with coronary artery disease[48], two other traits where the AMD loci encoding *CETP*, *APOE*, and *LIPC* are also involved. The gene *CPN1* has been associated with age-related disease (specifically, hearing impairment[45]).

Multiple signals in a single locus

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We use two examples to illustrate the importance of studying multiple signals in a single locus. Our first example focuses on a 1Mb region around locus *C2/CFB/SKIV2L* on chromosome 6 where 1,862 variants have P-values < 5×10^{-8}. There are an estimated 4 independent signals in the region by conditional analysis[31], 21 variants with Fgwas PP >0.1068, 11 with Bayesian PP >0.1068 by the standard Bayesian variable selection regression (BVSR) method that models no functional information, and 12 with Bayesian PP >0.1068 by SFBA. Interestingly, the alternative methods (P-value, Fgwas, and BVSR) identified intronic SNP *rs116503776/SKIV2L/NELFE* as the top candidates (P-value = 2.1×10^{-114}; Fgwas PP = 0.912; BVSR PP = 1.0), while SFBA identified two missense SNPs *rs4151667/C2/CFB* (P-value = 1.4×10^{-44}; SFBA PP = 0.917) and *rs115270436/SKIV2L/NELFE* (P-value = 2.8×10^{-99}; SBA PP = 0.633) as the top functional candidates (Figure 2; Supplementary Tables 2-4).

A haplotype analysis describing the odds ratios (ORs) for all possible haplotypes for SNPs *rs116503776*, *rs4151667*, and *rs115270436*, helps clarify the region. Intronic SNP *rs116503776* with the smallest P-value appears to be associated with the phenotype by tagging the other two missense SNPs (Supplementary Table 15). In particular, haplotypes with *rs116503776* can either increase or decrease risk, depending on alleles at the other two SNPs. To further confirm the importance of the missense SNPs *rs4151667* and *rs115270436*, we compared the AIC/BIC/loglikelihood between two models: one model with top two independent signals (*rs116503776* and *rs114254831*) identified by single-variant conditional analysis[31], versus the other model with top two signals (*rs4151667* and *rs115270436*) identified by SFBA. As expected, the second model has smaller AIC/BIC and larger loglikelihood than the first one (Supplementary Table 16). Thus, we can see that while alternative methods (P-value, Fgwas, and BVSR) focus on the SNP with the smallest P-value, our SFBA method finds an alternative pairing of missense signals that better accounts for all data.

Our second example focuses on a 1Mb region around gene *C3* on chromosome 19 (Supplementary Figure 10) with 112 genome-wide significant variants with P-value $< 5 \times 10^{-8}$. Fgwas only discovered a single missense signal, *rs2230199* with the most significant P-value=$1.7 \times$

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

10^{-77} (top blue triangle in Supplementary Figure 10(a, c)). However, both BVSR and SFBA identified 2 missense variants with PPs = 1.0, and 5 intronic variants with 0.11< PPs <0.18. The top two missense signals *rs2230199* and *rs147859257* (241 base pairs apart) were confirmed by conditional analysis[31], where the second signal *rs147859257* has conditioned P-value=6.0×10^{-33} (the purple triangle in Supplementary Figure 10(b, d), overlapping with *rs2230199*). These two missense signals match the interpretation of previous studies[49-51]. Because other 5 intronic variants (*rs11569479, rs11569470, rs201063729, rs10408682, rs11569466*) are in high LD with between variant R^2 >0.98, we believe this is the third independent signal whose Bayesian PP was split among 5 variants in high LD by SFBA.

Enrichment analysis

SFBA estimated that non-synonymous variants are 10-100 times more likely to be causal than variants in other categories and that they also have larger effect-sizes (Figure 3(a, b)). To better compare enrichment among multiple categories, we define two new sets of parameters (Supplementary Information). The first set of parameters, (π_q/π_{avg}), is defined to contrast the posterior association probability estimate (π_q) for each category to the genome-wide average (π_{avg}). The second set of parameters $(\sigma_q^2/\sigma_{avg}^2)$ is similarly defined to contrast the effect-size variance from each category to the genome-wide average. Moreover, the square root of the effect-size variance reflects the effect-size magnitude because of the prior assumption for the effect-size in our model.

Compared to the genome-wide average probability of causality π_{avg} = 4.3×10^{-06} (Supplementary Figure 12(a)), we found that non-synonymous category were 54x more likely to be causal (P-value= 7.24×10^{-84}); that coding-synonymous and other variants were 4.3x and 2.2x more likely (P-values = 0.005, 0.003); and that intergenic 0.7x less likely (P-value=4.9×10^{-6}); while the intronic variants matched the genome-wide average (P-value=0.659). In addition, compared to the genome-wide average effect-size variance ($\sigma_{avg}^2 = 0.02$; Supplementary Figure 12(b)), we found that the effect size variance of was 1.9x larger for non-synonymous variants (P-value=0.014; i.e., 1.4x larger effect-size); and 0.4x smaller for variants in the intronic category (P-value=4.5×10^{-06});

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

remaining categories were not significantly different (P-values >0.2). The estimated enrichment parameters by Fgwas show a similar pattern, although the contrast of the estimated enrichment for non-synonymous versus other annotations is not as pronounced as by SFBA (Supplementary Figure 11(a)).

Analysis with regulatory annotations

Second, we analyzed the GWAS data of AMD with the summarized regulatory annotations[41]: coding, UTR, promoter (defined as within 2KB of a transcription starting site), DHS in any of 217 cell types, intronic, intergenic, and “others” (not annotated as any of the previous six categories). Overall GWAS results were similar as the ones described in previous context (Supplementary Tables 7-10). Compared to the genome-wide average association probability (π_{avg}=4.03×10^{-6}; Supplementary Figure 12(c)), we found that the association probability of the coding category was 28x higher (P-value <2.2×10^{-16}); the promoter was 2.6x (P-value=0.028) higher; the intergenic and “others” were 0.5x and 0.9x less (P-values = 5.3×10^{-4}, 0.033); while the DHS and intronic were not significantly different (P-values >0.1). In addition, compared to the genome-wide average effect-size variance ($\sigma^2_{avg} = 0.024$), we found that the effect-size variance of the coding category was 1.9x larger (P-value=0.019; i.e., 1.4x larger effect-size); the DHS and intronic were 0.5x less (P-values = 0.011, 0.007); while the promoter, intergenic, and “others” were not significantly different (P-values >0.1; Supplementary Figure 12(d)). Here, Fgwas identified a slightly different enrichment pattern (Supplementary Figure 11(b)), where UTR was identified as the second most enriched category. This is presumably because Fgwas assumes one causal variant per locus and tends to prioritize the variant with the smallest P-value in each locus, e.g., UTR variants *rs1142/KMT2E/SPRK2* and *rs10422209/CNN2* have the highest Fgwas PP and the smallest P-value in their respective locus (Supplementary Tables 2 and 8).

Analysis with chromatin states

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Last, we considered the annotations of seven chromatin states obtained with ChromHMM in nine human cell types[43]: active promoter (APromoter), poised promoter (PPromoter), strong enhancer (SEnhancer), weak enhancer (WEnhancer), insulator, transcription elongation (TxnElong), repetitive/copy number variation (CNV). Nine human cell types include: embryonic stem cells (H1-hESC), erythrocytic leukaemia cells (K562), B-lymphoblastoid cells (GM12878), hepatocellular carcinoma cells (HepG2), umbilical vein endothelial cells (HUVEC), skeletal muscle myoblasts (HSMM), normal lung fibroblasts (NHLF), normal epidermal keratinocytes (NHEK) and mammary epithelial cells (HMEC).

With each set of chromatin states profiled in one cell type, we applied SFBA on the GWAS data of AMD, and then examined the list of variants that contribute 95% posterior probabilities in the identified loci with regional-PP >95%. We found that the results by accounting for the chromatin states profiled in the erythrocytic leukaemia cells (K562) gave the shortest list (average 14 variants per locus; Supplementary Table 17), and the enrichment analysis results of other cell types were slightly different (Supplementary Figures 13-15).

Here, we present the results of accounting for the chromatin states profiled in the K562 cell type (Figure 3(e, f); Supplementary Tables 11-14). Compared to the genome-wide average association probability ($\pi_{avg} = 4.0 \times 10^{-6}$; Supplementary Figure 12(e)), the association probability was 7.8x higher for the active promoter category (P-value = 7.4×10^{-10}), 3x higher for the strong enhancer category (P-value=0.013), 2.6x higher for the weak enhancer category (P-value = 0.002), 1.8x higher for the transcription elongation category (P-value = 0.002), 0.4x less for the CNV category (P-values = 0.004). In addition, the effect-size variances of associated variants in active promoter and strong enhancer were found 2x larger than the genome-wide average ($\sigma^2_{avg} = 0.022$; P-values = 0.048, 0.073), while the effect-size variances of weak enhancer, transcription elongation, and CNV categories were not significantly different (P-values >0.1; Supplementary Figure 12(f)).

Note that the Bayesian enrichment estimates of the poised promoter and insulator categories are the same as their priors (not plotted in Figure 3(e, f)), suggesting that SFBA identified no

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

associations in these two categories. Again, Fgwas identified a similar enrichment pattern (Supplementary Figure 11(c)).

DISCUSSION

Here, we describe a scalable Bayesian hierarchical method, SFBA, for integrating functional information in GWASs to help prioritize functional associations and understand underlying genetic architecture. SFBA models both association probability and effect-size distribution as a function of annotation categories for improving fine-mapping resolution. Unlike previous methods[26,27], SFBA accounts for LD and allows for the possibility of multiple association signals per locus while remaining capable of genome-wide inference. Further, SFBA employs an improved MCMC sampling strategy to greatly improve the mixing of MCMC samples, which ensures the capability of identifying a list of association candidates.

By simulation studies, we demonstrated that SFBA had higher power than Fgwas and conditioned P-value for identifying multiple signals in a single locus by accounting for both functional information and LD. We also showed that SFBA accurately estimated the enrichment patterns under scenarios with or without enrichment for one annotation in simulations. In the real analysis using the AMD GWAS data and three different types of annotations, by SFBA, we obtained posterior association probabilities and effect-size variances for variants of considered annotation categories, as well as an improved list of fine-mapped association signals. In addition, we replicated the findings of 32 out of 34 known AMD risk loci, as well as identified 5 potentially novel loci by SFBA. Further, we gave two fine-mapped AMD loci *C2/CFB/SKIV2L* and *C3* by SFBA as examples with justifications by haplotype analysis, model comparison, and previous findings. Thus, we believe our method is useful for understanding the underlying genetic architecture of complex traits and diseases, for efficiently integrating functional information into GWASs.

Our flexible framework allows for many further extensions. For example, it can be extended to deal with overlapping or quantitative annotations (Supplementary Information). These extensions will allow us to investigate the importance of a broader class of annotations (e.g. Combined Annotation Dependent Depletion (CADD) scores, MAF, and eQTL evidence). Importantly, as the

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

development of new genomic assays and computational tools enables new variant annotations, simultaneous modeling of available annotations will be critical to identify the set of annotations that are important for a specific trait. Then extending SFBA to select relevant annotations would be useful.

SFBA makes a key assumption that the variant correlation matrix has a block-wise structure, which allows us to segment the genome into approximately independent blocks, analyze variants per block by MCMC, and summarize genome-wide information by an EM algorithm. In parallel to our study, many recent studies have also explored the benefits of dividing the human genome into approximately independent LD blocks to facilitate genome-wide analyses[26,52]. Although the standard segmentation methods (e.g., based on genomic location[52] as we adopted here, or the number of variants per block[26]) are often sufficient in practice, we expect that a better segmentation method[30] based on LD blocks will likely further increase the association mapping power.

The biggest limitation of SFBA is probably computational cost, as we perform MCMC using the complete genotype data. Specifically, SFBA took 5,000 CPU hours (~5 hours with parallel computations on 1,000 CPUs for the 1,063 genome-blocks) to analyze the AMD GWAS data with 33,976 individuals and 9,857,286 variants. Implementing SFBA with summary statistics is expected to reduce the computation cost significantly, which is part of our continuing project. In addition, the variational approximation[53,54] and other approximations[55,56] of MCMC may provide an efficient alternative for posterior inference in large GWAS.

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ONLINE METHODS

Bayesian variable selection regression model

Our method is based on the standard Bayesian variable selection regression (BVSR) model

$$\boldsymbol{y}_{\boldsymbol{n}\times\boldsymbol{1}} = \boldsymbol{X}_{\boldsymbol{n}\times\boldsymbol{p}}\boldsymbol{\beta}_{\boldsymbol{p}\times\boldsymbol{1}} + \boldsymbol{\epsilon}_{\boldsymbol{n}\times\boldsymbol{1}}\,, \quad \beta_i \sim \pi_i N\left(0, \tau^{-1}\sigma_i^2\right) + (1-\pi_i)\delta_0(\beta_i), \qquad \epsilon_i \sim \mathrm{N}(0, \tau^{-1}),$$

where n denotes the number of individuals and p denotes the number of genetic variants; $\boldsymbol{y}_{\boldsymbol{n}\times\boldsymbol{1}}$ is the phenotype vector; $\boldsymbol{X}_{\boldsymbol{n}\times\boldsymbol{p}}$ is the genotype matrix; $\boldsymbol{\beta}_{\boldsymbol{p}\times\boldsymbol{1}}$ is a vector of genetic effect-sizes where each element β_i follows a spike-and-slab prior (known as the point-normal distribution) ---- that is, β_i follows a normal distribution $N\left(0, \tau^{-1}\sigma_i^2\right)$ with probability π_i, or β_i is set as 0 with probability $(1-\pi_i)$ and a point mass density function $\delta_0(\beta_i)$ at 0 ($\delta_0(\beta_i) = 1$ if $\beta_i = 0$, $\delta_0(\beta_i) = 0$ otherwise)[32,33]; and ϵ_i is the residual error that independently and identically follows a normal distribution $\mathrm{N}(0, \tau^{-1})$. We assume that both the phenotype vector $\boldsymbol{y}_{\boldsymbol{n}\times\boldsymbol{1}}$ and columns of the genotype matrix $\boldsymbol{X}_{\boldsymbol{n}\times\boldsymbol{p}}$ are centered, thus dropping the intercept. Although this model is developed for quantitative traits, we can treat binary phenotypes (e.g., cases and controls) as quantitative following previous approaches[33,35].

Bayesian hierarchical model accounting for functional information

For integrating functional information into the above BVSR model, we classify all variants into disjoint categories by assuming one annotation per variant. We further assume that variants in the same functional category have the same spike-and-slab prior for the effect-sizes, i.e., $\pi_i = \pi_q, \sigma_i^2 = \sigma_q^2$ for the qth category. Consequently, π_q denotes the category specific causal probability and σ_q^2 denotes the category specific effect-size variance (the square root of σ_q^2 reflects the magnitude of effect size). Although we focus on discrete non-overlapping annotations in this paper, our method can be extended to overlapping and continuous annotations (Supplementary Information).

We assume a Bayesian hierarchical framework[34] of BVSR with the following independent hyper priors:

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\pi_q \sim Beta\left(a_q, b_q\right), \qquad \sigma_q^2 \sim IG(k_1, k_2), \qquad \pi_q \perp \sigma_q^2,$$

where π_q follows a Beta distribution with positive shape parameters a_q and b_q, σ_q^2 follows an Inverse-Gamma distribution with shape parameter k_1 and scale parameter k_2. In order to adjust for the unbalanced distribution of functional annotations among all variants and enforce a sparse model in our analysis, we choose values for a_q and b_q such that the Beta distribution has mean $\frac{a_q}{a_q+b_q} = 10^{-6}$ with $(a_q + b_q)$ equal to the number of variants in category q. We set $k_1 = k_2 = 0.1$ in our analysis to induce non-informative prior for σ_q^2. Note that τ is fixed at the phenotype variance value in our Bayesian inferences (Supplementary Information).

Bayesian references

We introduce a latent indicator vector $\boldsymbol{\gamma}_{\boldsymbol{p}\times\mathbf{1}}$ to facilitate computation, where each binary element γ_i indicates whether $\beta_i = 0$ by $\gamma_i = 0$, or $\beta_i \sim N(0, \tau^{-1}\sigma_i^2)$ by $\gamma_i = 1$. Equivalently,

$$\gamma_i \sim \text{Bernoulli}(\pi_i), \qquad \boldsymbol{\beta}_{-\boldsymbol{\gamma}} \sim \boldsymbol{\delta}_{\mathbf{0}}, \qquad \boldsymbol{\beta}_{\boldsymbol{\gamma}} \sim \boldsymbol{MVN}_{|\boldsymbol{\gamma}|}\left(\mathbf{0}, \tau^{-1}\boldsymbol{V}_{\boldsymbol{\gamma}}\right),$$

where $|\boldsymbol{\gamma}|$ denotes the number of 1's in $\boldsymbol{\gamma}$; $\boldsymbol{\beta}_{-\boldsymbol{\gamma}}$ denotes the sub-vector of $\boldsymbol{\beta}_{\boldsymbol{p}\times\mathbf{1}}$ corresponding to variants with $\gamma_i = 0$; $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ denotes the sub-vector of $\boldsymbol{\beta}_{\boldsymbol{p}\times\mathbf{1}}$ corresponding to variants with $(\gamma_j = 1; j = 1, \dots, |\boldsymbol{\gamma}|)$; and $\boldsymbol{V}_{\boldsymbol{\gamma}}$ denotes the sub-matrix of the diagonal matrix $\boldsymbol{V}_{\boldsymbol{p}\times\boldsymbol{p}}$ whose ith diagonal element is $V_{ii} = \sigma_i^2$. Consequently, the expectation of γ_i is an estimate of the posterior inclusion probability (PP) for the ith variant, $E[\gamma_i] = Prob(\gamma_i = 1) = PP_i$.

For the described Bayesian hierarchical model above, the posterior joint distribution is proportional to

$$P\left(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\pi}, \boldsymbol{\sigma}^{\mathbf{2}}, \tau \,\middle|\, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{A}\right) \propto P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \tau) P\left(\boldsymbol{\beta}, \middle|\boldsymbol{A}, \boldsymbol{\pi}, \boldsymbol{\sigma}^{\mathbf{2}}, \boldsymbol{\gamma}, \tau\right) P(\boldsymbol{\gamma}|\boldsymbol{\pi}) P(\boldsymbol{\pi}) P(\boldsymbol{\sigma}^{\mathbf{2}}) P(\tau),$$

where $\boldsymbol{\pi} = \left(\pi_1, \dots, \pi_Q\right)^T$, $\boldsymbol{\sigma}^{\mathbf{2}} = \left(\sigma_1^2, \dots, \sigma_Q^2\right)^T$, $\boldsymbol{A}$ is the $p \times Q$ matrix of binary annotations, and Q is the total number of annotations. The goal is to estimate the category specific parameters $(\boldsymbol{\pi}, \boldsymbol{\sigma}^{\mathbf{2}})$ and the variant specific parameters $(\boldsymbol{\beta}, E[\boldsymbol{\gamma}])$ from their posterior distributions, conditioning on the data

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{A})$. Here, the category specific parameters denote the shared characteristics among all variants with the same annotation, which are also called enrichment parameters.

EM-MCMC algorithm

The basic idea of the EM-MCMC algorithm is to segment the whole genome into approximately independent blocks each with 5,000 ~ 10,000 variants; run MCMC algorithm per block with fixed category specific parameter values $(\boldsymbol{\pi}, \boldsymbol{\sigma}^{2})$ to obtain posterior estimates of $(\boldsymbol{\beta}, E[\boldsymbol{\gamma}])$ (E-step); then summarize the genome-wide posterior estimates of $(\boldsymbol{\beta}, E[\boldsymbol{\gamma}])$ and update values of $(\boldsymbol{\pi}, \boldsymbol{\sigma}^{2})$ by maximizing their posterior likelihoods (M-step). Repeat such EM-MCMC iterations for a few times until the estimates of $(\boldsymbol{\pi}, \boldsymbol{\sigma}^{2})$ (maximum a posteriori estimates, i.e., MAPs) converge (Supplementary Figure 1).

We derive the log-posterior-likelihood functions for $(\boldsymbol{\pi}, \boldsymbol{\sigma}^{2})$ and the analytical formulas for their MAPs. In addition, we construct their confidence intervals using Fisher information, whose analytical forms are derived for our Bayesian hierarchical model (Supplement Information). In our practical analyses, we find that, in general, with about 5 EM iterations, the estimates for $(\boldsymbol{\pi}, \boldsymbol{\sigma}^{2})$ would achieve convergence. Our method of conducting GWAS with functional information by using the above Bayesian hierarchical model and EM-MCMC algorithm is referred as “Scalable Functional Bayesian Association” (SFBA).

Convergence diagnosis

Here, the MCMC algorithm is essentially a random walk over all possible linear regression models with combinations of variants, which can start with either a model containing multiple significant variants by sequential conditional analysis or the most significant variant by P-value. In each MCMC iteration, a new model is proposed by including an additional variant, or deleting one variant from the current model, or switching one variant within the current model with one outside; and then up to acceptation or rejection by the Metropolis-Hastings algorithm (Supplementary Information). Importantly, we refine the standard proposal strategy for the switching step, by prioritizing variants in the neighborhood of the switch candidate according to their conditional

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

association evidence (e.g., P-values conditioning on variants, except the switch candidate, in the current model). As a result, this MCMC algorithm encourages our method to explore different combinations of potentially causal variants in each locus, and significantly improves the mixing property.

We used the potential scale reduction factor (PSRF)[57] to quantitatively diagnose MCMC mixing property. PSRF is essentially a ratio between the average within-chain variance of the posterior samples and the overall-chain variance with multiple MCMC chains. From the example plots of the PSRFs of Bayesian PPs (Supplementary Figure 2), for 58 top marginally significant SNPs (with P-values $<5 \times 10^{-8}$) in the WTCCC GWAS data of Crohn's disease[1], we can see that about half of the PSRF values by the standard MCMC algorithm (used in GEMMA[35]) exceed 1.2, suggesting the standard MCMC algorithm has poor mixing property. In contrast, the PSRF values by our MCMC algorithm are within the range of (0.9, 1.2), suggesting that our MCMC algorithm has greatly improved mixing property.

Computational technics

We employ two computational technics to save memory in the SFBA software. One is to save all genotype data as unsigned characters in memory, because unsigned characters are equivalent to unsigned integers in [0, 256] that can be easily converted to genotype values within the range of (0.0, 2.0) by multiplying with 0.01. This technic saves up to 90% memory comparing with saving genotypes in double type. Second, with an option of in-memory compression, SFBA will further save additional 70% memory. As a result, we can decrease the memory usage from ~120 GB (usage by GEMMA[35]) to ~3.6GB, for a typical GWAS dataset with ~33K individuals and ~500K variants.

The SFBA software wraps a C++ executable file for the E-step (MCMC algorithm) and an R script for the M-step together by a Makefile, which is generated by a Perl script and enables parallel computation through submitting jobs. Generally, 50K MCMC iterations with ~5K variants and ~33K individuals take about 300MB memory and 1hr CPU time on a 1.6GHz core, where the computation

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cost is of order $O(nm^2)$ with the sample size (n) and number of variants (m) considered in the linear models during MCMC iterations (usually $m < 10$). The computation cost for M-step is almost negligible because of analytical formulas for the MAPs.

Fgwas

In this paper, the Fgwas results were generated by using summary statistics from single variant likelihood-ratio tests and the same annotation information used by SFBA. Fgwas[26] produces variant-specific posterior association probabilities (PPs), segment-specific PPs, and enrichment estimates for all annotations. To avoid the issue of failing convergence, we used segment size of 2,000 variants for Fgwas in both simulations and real data analyses. As a result, the final Fgwas PP is given by the product of the variant-specific PP and the corresponding segment–specific PP, and the Fgwas regional-PP is given by the highest segment-specific PP in a region or genome block.

Simulation data

We used genotype data on Chromosome 20-22 from the AMD GWAS (33,976 individuals and 241,500 variants with MAF>0.1) to simulate quantitative phenotypes from the standard linear regression model $y_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + \epsilon_i,\ i = 1, \ldots, 33976$, where $\boldsymbol{X}_i$ is the genotype vector of the ith individual and ϵ_i is the noise term generated from $N(0, \sigma_\epsilon^2)$. We segmented the genotype data into 50x2.5Mb blocks each with ~5,000 variants. Within each block, we marked a ~25Kb continuous region (starting 37.5Kb from the beginning of a block) as the causal locus and randomly selected two causal SNPs per locus. Two complementary annotations (“coding” vs. “noncoding”) were simulated, where the coding variants account for ~1% overall variants but ~10% variants within the causal loci (matching the pattern in the real AMD analysis). We selected positive effect-size vector $\boldsymbol{\beta}$ and noise variance σ_ϵ^2 such that a total of 15% phenotypic variance was equally explained by causal SNPs. We controlled the enrichment-fold of coding variants by varying the number of coding variants among these 100 causal SNPs.

We compared SFBA with P-value, conditioned P-value, and Fgwas. In the simulation studies, P-values were obtained from a series of likelihood-ratio tests based on the standard linear regression model. P-values conditioning on the top significant variant per locus were used to identify

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the second signal by conditional analysis. Fgwas was implemented with summary statistics from single variant tests and the segment size of 2,000 variants (selected to avoid convergence issues). We failed to include PAINTOR in the comparison, because PAINTOR cannot complete the analysis for one block in >1,000 CPU hours (on a 2.5GHz, 64-bit CPU) and is thus expected to require >1 million CPU hours for a genome-wide analysis.

GWAS data of AMD

In the GWAS data of AMD, the advanced AMD cases – including wet cases with choroidal neovascularization (CNV, when accompanied by angiogenesis) and dry cases with geographic atrophy (GA, when angiogenesis is absent) – and control subjects were gathered across 26 studies, with DNA samples collected and genotyped centrally[39]. All genotypes were generated by a customized chip that contains (i) the usual genome-wide variant content, (ii) exome content comparable to the Exome chip (protein-altering variants across all exons), (iii) variants in known AMD risk loci (protein-altering variants and previously associated variants), and (iv) previously observed and predicted variation in *TIMP3* and *ABCA4* (two genes implicated in monogenic retinal dystrophies). The genotyped variants (439,350) were then imputed to the 1000 Genomes reference panel (Phase I)[40], resulting a total of 12,023,830 variants.

SFBA used dosage genotype data and standardized phenotypes. Phenotypes were first coded quantitatively with 1's for cases and 0's for controls; second corrected for the first and second principle components, age, gender, and source of DNA samples; and then standardized to have mean 0 and standard deviation 1. In order to make the Bayesian inferences scalable to the AMD GWAS data (33,976 individuals, 9,866,744 variants with MAF >0.5%), we segmented the whole genome into 1,063 non-overlapped blocks, such that each block has length ~2.5Mb (containing ~10,000 variants) and all previously identified loci along with variants in LD (R^2 >0.1) were not split. Then we applied the EM-MCMC algorithm with 5 EM steps and 50,000 MCMC iterations (including 50,000 extra burn-ins).

For comparison, P-values were obtained by a series of likelihood-ratio tests, using the same "quantitative" phenotype vector as used by SFBA; Fgwas was implemented with the summary statistics from single variant tests and the segment size of 2,000 variants (resulting 4,934

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

segments); and a standard Bayesian variable selection regression (BVSR) method that models no functional information was also applied.

Three types of genomic annotations were considered for analyzing the AMD data: gene-based functional annotations of SNPs and small indels from SeattleSeq (http://snp.gs.washington.edu/SeattleSeqAnnotation138/index.jsp), summarized regulatory annotations[41], and the chromatin states profiled respectively in nine human cell types from chromHMM[19,42,43]. For variants annotated with multiple functions, we used the most severe function in the analysis: non-synonymous > coding-synonymous > other-genomic > intronic > intergenic for the gene-based annotations; coding > UTR > promoter > DHS > intronic > intergenic > "others" for the summarized regulatory annotations; active promoter > poised promoter > strong enhancer > weak enhancer > insulator > transcription elongation > CNV for the chromatin states.

Software

Our software SFBA is freely available on Github (https://github.com/yjingj/SFBA).

ACKNOWLEDGMENTS

XZ is supported by startup funds from the University of Michigan and a grant from the Foundation for the National Institutes of Health through the Accelerating Medicines Partnership (BOEH15AMP, co-PIs M. Boehnke and G. Abecasis).

COMPETING FINANCIAL INTERESTS

None.

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

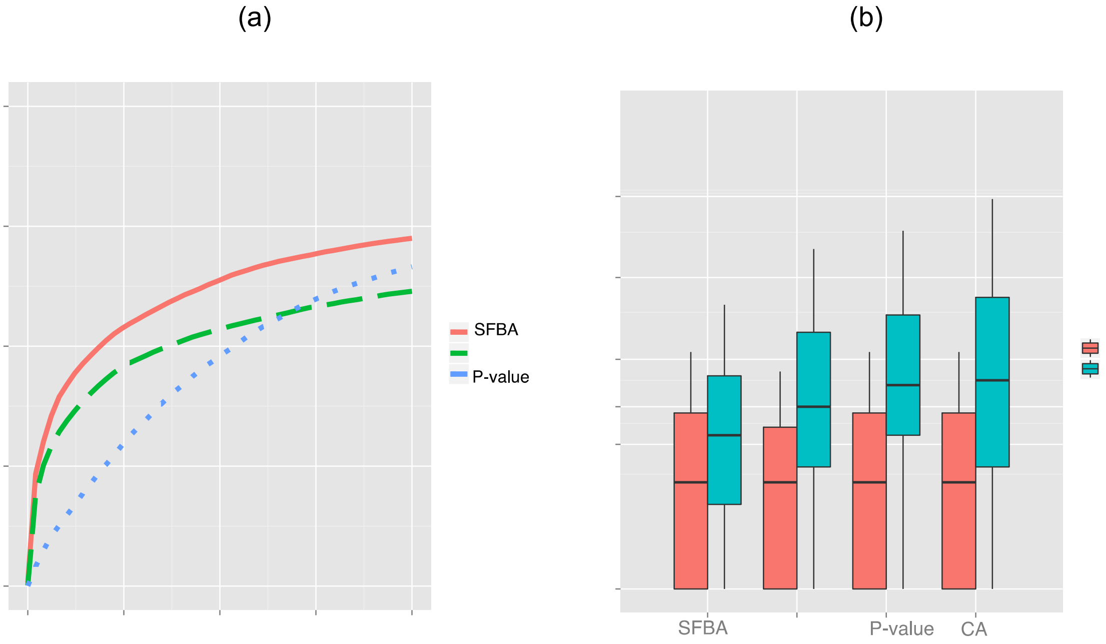

Figure 1: (a) Average ROC curves of Bayesian PP by SFBA, Fgwas PP, and P-value, and (b) boxplot of the ranks of the true causal SNP1 (with smaller P-value) and SNP2 by SFBA, Fgwas, P-value, and conditional analysis (CA), with 100 simulation replicates and the complete sample size 33,976.

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

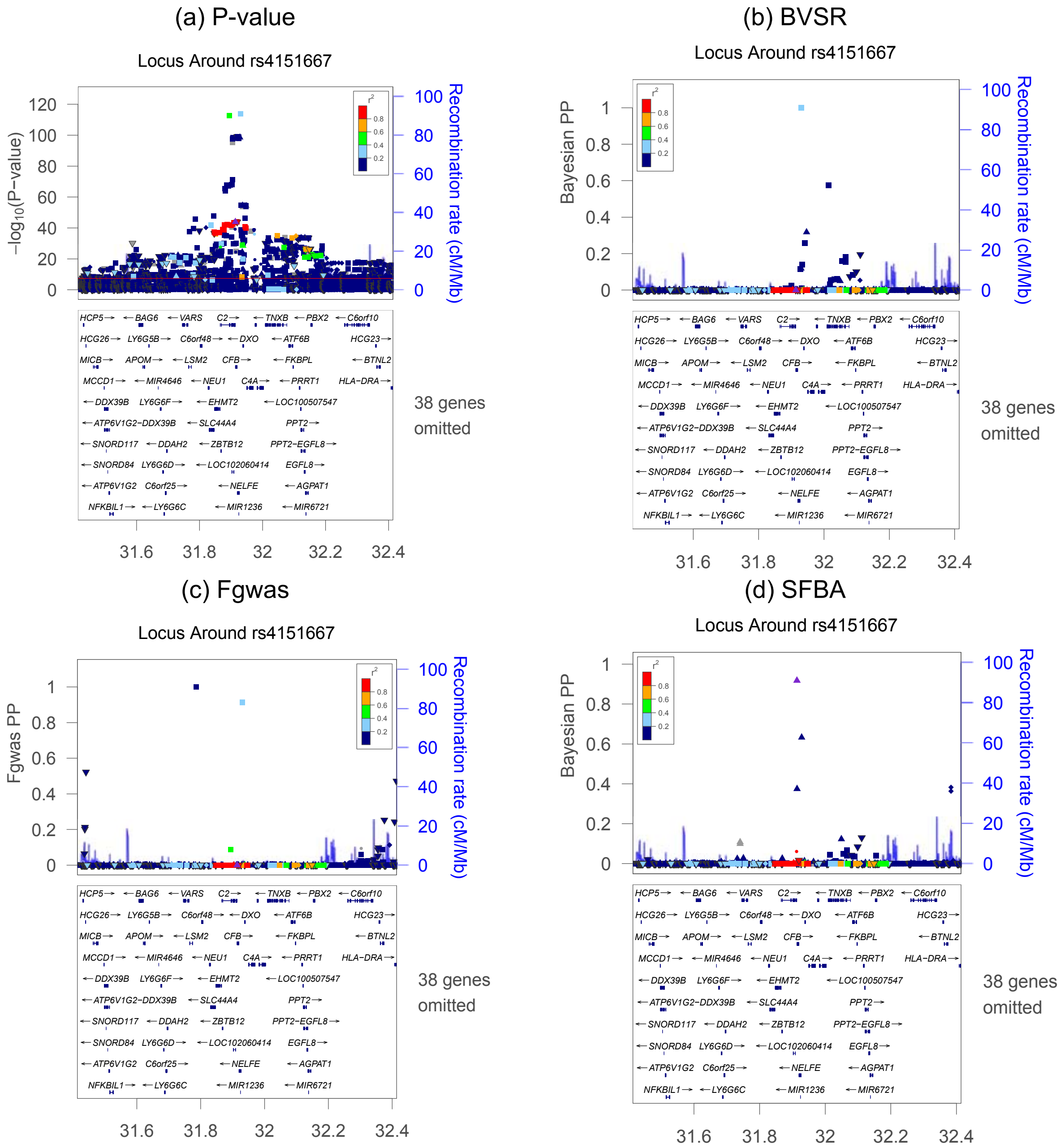

Figure 2: ZoomLocus plots with P-values by single variant tests (a), Bayesian PPs by BVSR (b), Fgwas PPs (c), and Bayesian PPs by SFBA (d); the top cyan squares in panels (a, b, c) denote the intronic variant *rs116503776*; the purple triangle in (d) denotes the non-synonymous variant *rs4151667*; shapes denote different annotations (triangle point up ∆ for non-syn, circle o for coding-syn, square □ for intronic, diamond ◊ for intergenic, and triangle point down ∇ for other-genomic).

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

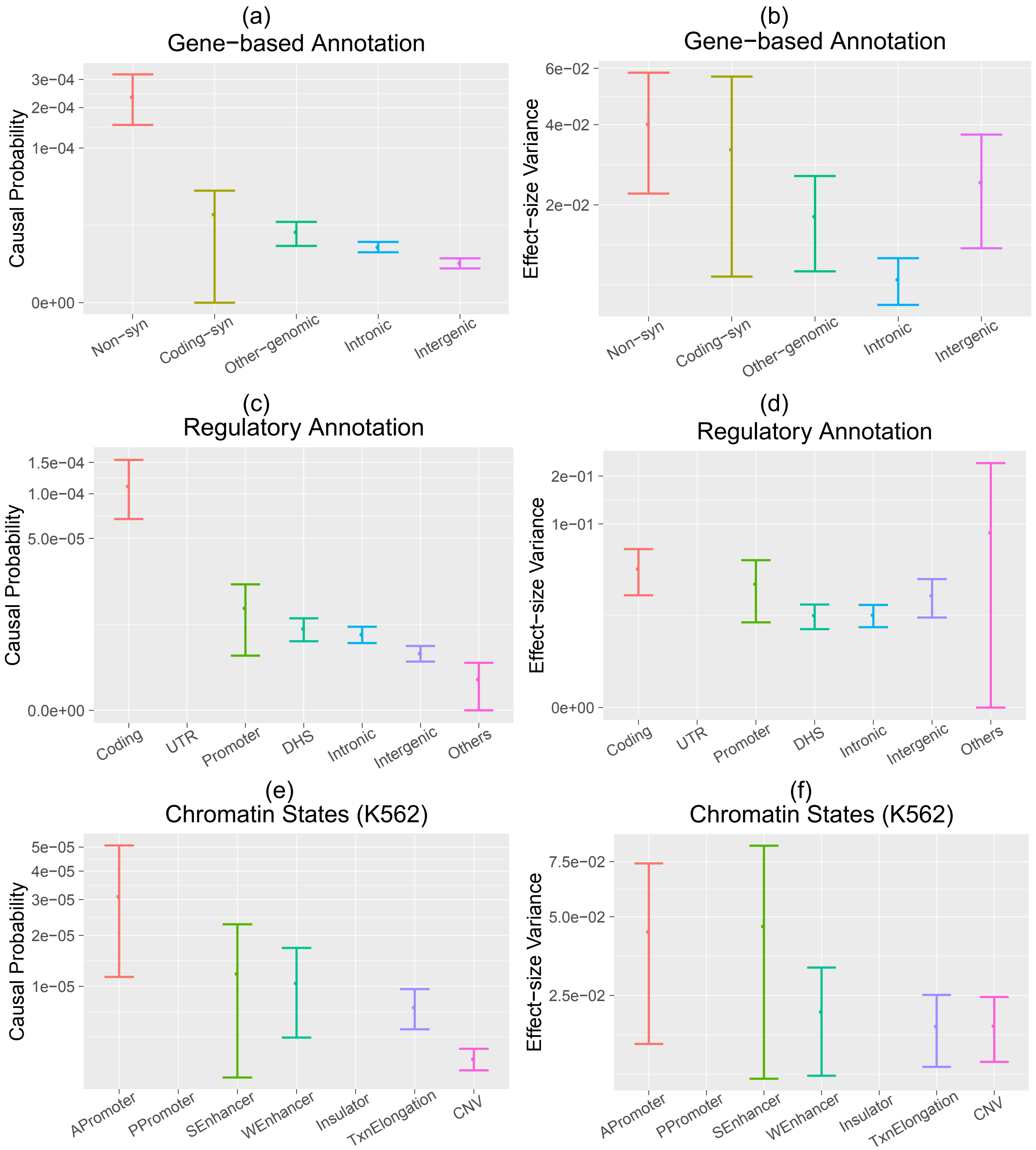

Figure 3: Category specific (enrichment) parameter estimates with 95% error bars by SFBA, panels (a, c, e) for causal probabilities and panels (b, d, f) for effect-size variances, with 3 sets of annotations. The estimates that are the same as their priors are not ploted: estimates of UTR in (c, d), estimates of the active/poised promoter in (e, f). Note that the estimate of the effect-size variance for the "Others" category in (d) is also close to the prior because of low region-association evidence, hence it has a wide 95% error bar.

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES (Limited to 30)

1. Wellcome Trust Case Control C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661-78.
2. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature reviews Genetics 2008;9:356-69.
3. Voight BF, Scott LJ, Steinthorsdottir V, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nature genetics 2010;42:579-89.
4. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. American journal of human genetics 2012;90:7-24.
5. Global Lipids Genetics C, Willer CJ, Schmidt EM, et al. Discovery and refinement of loci associated with lipid levels. Nature genetics 2013;45:1274-83.
6. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nature reviews Genetics 2005;6:95-108.
7. Yu J, Pressoir G, Briggs WH, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics 2006;38:203-8.
8. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America 2009;106:9362-7.
9. Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature genetics 2012;44:369-75, S1-3.
10. Carithers LJ, Moore HM. The Genotype-Tissue Expression (GTEx) Project. Biopreservation and biobanking 2015;13:307-8.
11. Dixon JR, Jung I, Selvaraj S, et al. Chromatin architecture reorganization during stem cell differentiation. Nature 2015;518:331-6.
12. Kellis M, Wold B, Snyder MP, et al. Defining functional DNA elements in the human genome. Proceedings of the National Academy of Sciences of the United States of America 2014;111:6131-8.
13. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols 2009;4:1073-81.
14. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Current protocols in human genetics / editorial board, Jonathan L Haines [et al] 2013;Chapter 7:Unit7 20.
15. Pickrell JK, Marioni JC, Pai AA, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 2010;464:768-72.
16. Tung J, Zhou X, Alberts SC, Stephens M, Gilad Y. The genetic architecture of gene expression levels in wild baboons. eLife 2015;4.
17. Lea AJ, Tung J, Zhou X. A Flexible, Efficient Binomial Mixed Model for Identifying Differential DNA Methylation in Bisulfite Sequencing Data. PLoS genetics 2015;11:e1005650.
18. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome research 2011;21:447-55.
19. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nature methods 2012;9:215-6.
20. McVicker G, van de Geijn B, Degner JF, et al. Identification of Genetic Variants That Affect Histone Modifications in Human Cells. Science 2013;342:747-9.
21. Cooper GM, Stone EA, Asimenos G, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome research 2005;15:901-13.
22. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics 2014;46:310-5.
23. Finucane HKaB-S, Brendan and Gusev, Alexander and Trynka, Gosia and Reshef, Yakir and Loh, Po-Ru and Anttila, Verneri and Xu, Han and Zang, Chongzhi and Farh, Kyle and Ripke, Stephan and Day, Felix R and ReproGen Consortium and Schizophrenia Working Group of the Psychiatric Genomics Consortium and

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The RACI Consortium and Purcell, Shaun and Stahl, Eli and Lindstrom, Sara and Perry, John R B and Okada, Yukinori and Raychaudhuri, Soumya and Daly, Mark J and Patterson, Nick and Neale, Benjamin M and Price, Alkes L. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature genetics 2015;47:1228--35.

24. Zhou X. A Unified Framework for Variance Component Estimation with Summary Statistics in Genome-wide Association Studies. bioRxiv 2016.

25. Schork AJ, Thompson WK, Pham P, et al. All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. PLoS genetics 2013;9:e1003449.

26. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. American journal of human genetics 2014;94:559-73.

27. Kichaev G, Yang WY, Lindstrom S, et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS genetics 2014;10:e1004722.

28. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. Science 2002;296:2225-9.

29. Wall JD, Pritchard JK. Haplotype blocks and linkage disequilibrium in the human genome. Nature reviews Genetics 2003;4:587-97.

30. Berisa T, Pickrell JK. Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics 2016;32:283-5.

31. Fritsche LG, Igl W, Bailey JN, et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nature genetics 2015.

32. Chipman H, George EI, McCulloch RE. The Practical Implementation of Bayesian Model Selection. In: Lahiri P, ed. Model selection. Beachwood, OH: Institute of Mathematical Statistics; 2001:65-116.

33. Guan Y, Stephens M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. 2011:1780-815.

34. Carbonetto P, Stephens M. Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for IL-2 signaling genes in type 1 diabetes, and cytokine signaling genes in Crohn's disease. PLoS genetics 2013;9:e1003770.

35. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with bayesian sparse linear mixed models. PLoS genetics 2013;9:e1003264.

36. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature genetics 2007;39:906-13.

37. Wen X, Stephens M. Bayesian Methods for Genetic Association Analysis with Heterogeneous Subgroups: From Meta-Analyses to Gene-Environment Interactions. The annals of applied statistics 2014;8:176-203.

38. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nature genetics 2012;44:821-4.

39. Fritsche LG, Igl W, Cooke Bailey JN, et al. Insights into Rare and Common Genetic Variation From a Large Study of Age-Related Macular Degeneration. Nature genetics in press.

40. Genomes Project C, Auton A, Brooks LD, et al. A global reference for human genetic variation. Nature 2015;526:68-74.

41. Gusev A, Lee SH, Trynka G, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. American journal of human genetics 2014;95:535-52.

42. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nature biotechnology 2010;28:817-25.

43. Ernst J, Kheradpour P, Mikkelsen TS, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 2011;473:43-9.

44. Chauhan L, Jenkins GD, Bhise N, et al. Genome-wide association analysis identified splicing single nucleotide polymorphism in CFLAR predictive of triptolide chemo-sensitivity. BMC genomics 2015;16:483.

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

45. **Fransen E, Bonneux S, Corneveaux JJ, et al. Genome-wide association analysis demonstrates the highly polygenic character of age-related hearing impairment. European journal of human genetics : EJHG 2015;23:110-5.**

46. **Masson D, Jiang XC, Lagrost L, Tall AR. The role of plasma lipid transfer proteins in lipoprotein metabolism and atherogenesis. Journal of lipid research 2009;50 Suppl:S201-6.**

47. **Kettunen J, Tukiainen T, Sarin AP, et al. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. Nature genetics 2012;44:269-76.**

48. **Nikpay M, Goel A, Won HH, et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nature genetics 2015;47:1121-30.**

49. **Helgason H, Sulem P, Duvvari MR, et al. A rare nonsynonymous sequence variant in C3 is associated with high risk of age-related macular degeneration. Nature genetics 2013;45:1371-4.**

50. **Seddon JM, Yu Y, Miller EC, et al. Rare variants in CFI, C3 and C9 are associated with high risk of advanced age-related macular degeneration. Nature genetics 2013;45:1366-70.**

51. **Zhan X, Larson DE, Wang C, et al. Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. Nature genetics 2013;45:1375-9.**

52. **Loh PR, Bhatia G, Gusev A, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nature genetics 2015;47:1385-92.**

53. **Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. An Introduction to Variational Methods for Graphical Models. Machine Learning 1999;37:183-233.**

54. **Carbonetto P, Stephens M. Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies. 2012:73-108.**

55. **Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2009;71:319-92.**

56. **Singh SaW, Michael and McCallum, Andrew. Monte Carlo MCMC: efficient inference by approximate sampling: Association for Computational Linguistics; 2012.**

57. **Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. Statistical Science 1992;7:457-72.**

Supplementary Table 1: Classification of gene-based functional annotations.

Native gene-based functional annotations	Annotation categories considered in the analysis
frameshift, frameshift-near-splice	Non-synonymous
splice-acceptor, splice-donor,	
stop-gained, stop-gained-near-splice, stop-lost	
missense, missense-near-splice	
synonymous-near-splice, non-coding-exon-near-splice, coding-near-splice, coding-unknown-near-splice, intron-near-splice	
coding, coding-unknown, synonymous, nc-transcript-variant	Coding-synonymous
intronic	Intronic
intergenic, NAs	Intergenic
3-prime-UTR, 5-prime-UTR,	Other-genomic
downstream-gene, upstream-gene, non-coding-exon	

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Table 2: Compare results by P-value (single variant test), Fgwas, and SFBA in the known 34 AMD loci, accounting for gene-based functional annotations.

Known 34 Loci				Top significant variant by P-value					Bayesian Regional-PP	Fgwas Regional-PP
Locus name	**Chr**	**Start**	**End**	**dbSNPID**	**Chr:Position**	**MAF**	**P-value**	**Anno**		
CFH	1	195,679,832	197,768,053	rs10922109	1:196,704,632	0.329	$<9 \times 10^{-321}$	intronic	1.000	1.000
COL4A3	2	227,573,015	228,592,110	rs11884770	2:228,086,920	0.731	5.6×10^{-9}	intronic	0.984	0.986
ADAMTS9-AS	3	64,199,445	65,230,121	rs62247658	3:64,715,155	0.551	1.4×10^{-15}	intronic	0.978	1.000
COL8A1	3	98,551,114	100,381,567	rs140647181	3:99,180,668	0.019	5.4×10^{-13}	intergenic	1.000	0.999
CFI	4	110,126,506	111,185,820	rs10033900	4:110,659,067	0.506	7.1×10^{-19}	downstream	1.000	1.000
C9	5	38,699,134	39,831,894	rs62358361	5:39,327,888	0.012	3.1×10^{-16}	intronic	1.000	1.000
PRLR/SPEF2	5	34,769,332	36,493,378	rs114092250	5:35,494,448	0.018	2.5×10^{-9}	intergenic	0.961	0.987
C2/CFB/SKIV2L	6	30,505,490	33,238,589	rs116503776	6:31,930,462	0.120	2.1×10^{-114}	intronic	1.000	1.000
VEGFA	6	43,305,296	44,329,629	rs943080	6:43,826,627	0.518	2.0×10^{-16}	intergenic	1.000	1.000
KMT2E/SRPK2	7	104,081,402	105,563,372	rs1142	7:104,756,326	0.357	1.5×10^{-10}	downstream	0.999	0.999
PILRB/PILRA	7	99,394,940	100,611,776	rs7803454	7:99,991,548	0.199	3.6×10^{-10}	intronic	0.999	0.999
TNFRSF10B	8	22,582,971	23,588,984	rs79037040	8:23,080,971	0.534	2.9×10^{-12}	nc-transcript	1.000	0.999
MIR6130/RORB	9	75,935,160	77,189,752	rs10781180	9:76,615,662	0.683	3.0×10^{-10}	intergenic	0.997	0.999
TRPM3	9	72,938,605	73,946,180	rs7150714	9:73,438,605	0.584	3.2×10^{-9}	intronic	0.929	0.999
TGFBR1	9	101,358,102	102,431,769	rs1626340	9:101,923,372	0.199	2.3×10^{-11}	intergenic	1.000	0.999
ABCA1	9	107,139,414	108,167,147	rs2740488	9:107,661,742	0.265	1.7×10^{-9}	intronic	0.963	0.985
ARHGAP21	10	24,360,361	25,556,538	rs12357257	10:24,999,593	0.232	4.3×10^{-9}	intronic	0.962	0.986

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Known 34 Loci				Top significant variant by P-value					Bayesian Regional-PP	Fgwas Regional-PP
Locus name	**Chr**	**Start**	**End**	**dbSNPID**	**Chr:Position**	**MAF**	**P-value**	**Anno**		
ARMS2/HTRA1	10	123,702,126	124,735,355	rs3750846	10:124,215,565	0.316	$<9 \times 10^{-321}$	intronic	1.000	1.000
RDH5/CD63	12	55,615,585	56,713,297	rs3138141	12:56,115,778	0.214	4.7×10^{-10}	intronic	0.034	0.999
ACAD10	12	110,919,995	113,502,935	rs73205633	12:112,357,085	0.019	1.2×10^{-10}	intergenic	0.997	0.999
B3GALTL	13	31,242,232	32,339,274	rs9564692	13:31,821,240	0.288	3.2×10^{-11}	splice	1.000	0.999
RAD51B	14	68,227,506	69,550,783	rs1956526	14:68,799,787	0.650	$1..0 \times 10^{-11}$	intronic	1.000	0.999
LIPC	15	58,171,721	59,242,418	rs2414577	15:58,680,638	0.365	4.8×10^{-17}	nc-transcript	1.000	1.000
CETP	16	56,485,514	57,506,829	rs5817082	16:56,997,349	0.248	1.7×10^{-21}	intronic	1.000	1.000
CTRB2/CTRB1	16	74,732,528	76,017,115	rs72802342	16:75,234,872	0.073	2.8×10^{-13}	downstream	1.000	1.000
TMEM97/VTN	17	26,092,946	27,240,139	rs11080055	17:26,649,724	0.524	1.5×10^{-9}	intronic	0.996	0.998
NPLOC4/TSPAN10	17	79,015,509	80,186,552	rs6565597	17:79,526,821	0.390	1.0×10^{-12}	intronic	1.000	0.999
C3	19	5,311,717	7,224,340	rs2230199	19:6,718,387	0.764	1.7×10^{-77}	missense	1.000	1.000
CNN2	19	523,867	1,533,360	rs10422209	19:1,026,318	0.132	5.5×10^{-9}	upstream	0.970	0.993
APOE	19	44,892,254	46,313,830	rs429358	19:45,411,941	0.118	3.3×10^{-46}	missense	1.000	1.000
MMP9	20	44,114,991	45,160,699	rs142450006	20:44,614,991	0.132	1.4×10^{-11}	intergenic	1.000	0.999
C20orf85	20	56,084,276	57,174,034	rs117739907	20:56,652,781	0.062	7.8×10^{-18}	intergenic	1.000	1.000
SYN3/TIMP3	22	32,546,536	33,613,375	rs5754227	22:33,105,817	0.123	2.0×10^{-27}	intronic	1.000	1.000
SLC16A8	22	37,795,271	39,003,972	rs8135665	22:38,476,276	0.205	2.9×10^{-12}	intronic	1.000	0.999

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Table 3: AMD risk variants by SFBA in the known 34 loci, accounting for gene-based functional annotations. Variants with Bayesian PPs >0.5 or the highest Bayesian PPs in the loci are listed. Shown are reside/nearby genes, dbSNPIDs, positions, functional annotations, MAFs (unfolded, corresponding to the direction of effect-sizes), P-values, and Bayesian PPs/effect-sizes.

Signal number	Reside/Nearby Gene	dbSNPID	Chr:Position	Anno	MAF	Bayesian PP	Effect-size	P-value
1.1	*CFH*	rs800292	1:196,642,233	missense	0.183	0.997	-0.312	2.4×10^{-319}
1.2	*CFH*	rs10922094	1:196,661,505	intronic	0.530	1.000	-0.214	$< 9.0 \times 10^{-321}$
1.3	*CFHR1*	rs605082	1:196,801,917	downstream	0.353	0.518	-0.092	7.5×10^{-257}
1.4	*CFHR4*	rs58175074	1:196,820,080	intronic	0.158	0.792	-0.314	$< 9.0 \times 10^{-321}$
1.5	*CFHR4*	rs149032610	1:196,857,150	5'-UTR	0.015	1.000	0.195	6.6×10^{-38}
1.6	*CFHR4*	rs10494745	1:196,887,457	missense	0.134	0.526	0.092	7.4×10^{-137}
1.7	*CFHR2*	rs138579109	1:196,923,955	intronic	0.043	0.893	0.167	8.4×10^{-85}
1.8	*CFHR5*	rs35662416	1:196,967,354	missense	0.022	0.889	-0.122	5.8×10^{-6}
2	*COL4A3*	rs11884770	2:228,086,920	intronic	0.731	0.269	0.052	5.6×10^{-9}
3	*ADAMTS9-AS2*	rs7428936	3:64,710,850	intronic	0.448	0.167	-0.061	1.5×10^{-15}
4	*COL8A1*	rs140647181	3:99,180,668	intergenic	0.019	0.687	0.224	54×10^{-13}
5	*CFI*	rs10033900	4:110,659,067	downstream	0.506	0.999	-0.067	7.2×10^{-19}
6	*C9*	rs34882957	5:39,331,894	missense	0.012	0.998	0.278	4.0×10^{-16}
7	*PRLR/SPEF2*	rs114092250	5:35,494,448	intergenic	0.019	0.403	-0.174	2.5×10^{-9}
8.1	*C2/CFB*	rs4151667	6:31,914,024	missense	0.036	0.917	-0.279	1.4×10^{-44}
8.2	*SKIV2L/NELFE*	rs115270436	6:31,928,306	missense	0.071	0.633	-0.321	2.8×10^{-99}
8.3	*HLA-DQB1*	rs3891176	6:32,634,318	missense	0.159	0.726	0.153	1.2×10^{-11}
9	*VEGFA*	rs943080	6:43,826,627	intergenic	0.518	0.435	0.063	2.0×10^{-16}
10	*KMT2E/SRPK2*	rs1142	7:104,756,326	downstream	0.357	0.125	0.052	1.5×10^{-10}
11	*PILRB*	rs35986051	7:99,956,439	missense	0.139	0.193	0.075	4.0×10^{-10}
12	*TNFRSF10A*	rs79037040	8:23,082,971	nc-transcript	0.534	0.996	0.053	2.9×10^{-12}
13	*MIR6130/RORB*	rs10781182	9:76,617,720	intergenic	0.684	0.070	-0.052	3.0×10^{-10}
14	*TRPM3*	rs71507014	9:73,438,605	intronic	0.584	0.822	-0.046	3.2×10^{-9}
15	*TGFBR1*	rs10819635	9:101,864,510	upstream	0.186	0.137	-0.066	2.4×10^{-11}
16	*ABCA1*	rs2740488	9:107,661,742	intronic	0.266	0.756	-0.053	1.7×10^{-9}
17	*ARHGAP21*	rs12357257	10:24,999,593	intronic	0.232	0.318	0.053	4.3×10^{-9}
18	*ARMS2*	rs10490924	10:124,214,448	missense	0.316	0.996	0.474	$< 9.0 \times 10^{-321}$
19	*RDH5/CD63*	rs3138142	12:56,115,585	coding-syn	0.213	0.706	0.074	6.1×10^{-10}
20	*MAPKAPK5*	rs61941287	12:112,330,305	intronic	0.019	0.309	0.191	1.2×10^{-10}
21	*B3GLCT*	rs9564692	13:31,821,240	splice	0.288	0.942	-0.056	3.2×10^{-11}
22	*RAD51B*	rs2842339	14:68,986,999	intronic	0.899	0.243	-0.082	3.1×10^{-7}
23	*ALDH1A2*	rs2414577	15:58,680,638	intronic	0.366	0.501	-0.067	4.8×10^{-17}

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Signal number	Reside/Nearby Gene	dbSNPID	Chr:Position	Anno	MAF	Bayesian PP	Effect-size	P-value
24	*CETP*	rs1532625	16:57,005,301	splice	0.448	0.358	0.044	7.9×10^{-19}
25	*CTRB2*	rs72802342	16:75,234,872	downstream	0.360	0.297	-0.114	2.8×10^{-13}
26	*CTB-96E2.2/VTN*	rs704	17:26,694,861	missense	0.483	0.325	0.042	3.3×10^{-8}
27	*NPLOC4/TSPAN10*	rs6420484	17:79,612,397	missense	0.622	0.402	-0.055	4.0×10^{-12}
28.1	*FUT6/NRTN*	rs17855739	19:5,831,840	missense	0.044	0.681	-0.159	1.5×10^{-16}
28.2	*C3/CTD-3128G10.7*	rs147859257	19:6,718,146	missense	0.008	1.000	0.501	4.3×10^{-31}
28.3	*C3/CTD-3128G10.7*	rs2230199	19:6,718,387	missense	0.764	1.000	-0.172	1.7×10^{-77}
29.1	*ABCA7*	rs3752237	19:1,047,161	coding-syn	0.644	0.544	-0.065	6.7×10^{-3}
29.2	*ABCA7*	rs12151021	19:1,050,874	intronic	0.708	1.000	0.091	1.9×10^{-5}
30	*APOE/TOMM40/ CTB-129P6.7*	rs429358	19:45,411,941	missense	0.118	1.000	-0.173	3.3×10^{-46}
31	*MMP9/RP11-465L10.10*	rs2274755	20:44,639,692	splice	0.138	0.435	-0.073	5.4×10^{-11}
32	*C20orf85*	rs201459901	20:56,653,724	intergenic	0.063	0.078	-0.135	7.9×10^{-18}
33	*SYN3*	rs5754227	22:33,105,817	intronic	0.124	0.764	-0.128	2.0×10^{-27}
34.1	*SLC16A8/BAIAP2L2*	rs4289289	22:38,477,342	missense	0.485	0.824	0.056	1.1×10^{-09}
34.2	*SLC16A8/BAIAP2L2*	rs77968014	22:38,478,666	splice	0.009	0.973	0.212	3.1×10^{-6}

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Table 4: AMD risk variants by Fgwas in the known 34 loci, accounting for gene-based functional annotations. Variants with Fgwas PPs >0.5 or the highest Fgwas PPs in the loci are listed in this table. Shown are reside/nearby genes, dbSNPIDs, positions, functional annotations, MAFs (unfolded), Fgwas PPs, and P-values.

Signal number	Reside/Nearby Gene	dbSNPID	Chr:Position	Anno	MAF	Fgwas PP	P-value
1.1	*CFH*	rs77498516	1:196,115,300	intergenic	0.048	0.522	8.2×10^{-27}
1.2	*CFH*	rs10922109	1:196,704,632	intronic	0.329	0.802	$< 9.0 \times 10^{-321}$
1.3	*RP4-608O15.3*	rs521631	1:196,813,352	intronic	0.506	0.999	$< 9.0 \times 10^{-321}$
2	*COL4A3*	rs11884770	2:228,086,920	intronic	0.731	0.181	5.7×10^{-9}
3	*ADAMTS9-AS2*	rs62247658	3:64,715,155	intronic	0.551	0.167	1.5×10^{-15}
4	*COL8A1*	rs140647181	3:99,180,668	intergenic	0.019	0.999	5.4×10^{-13}
5	*CFI*	rs10033900	4:110,659,067	downstream	0.506	0.996	7.2×10^{-19}
6.1	*C9*	rs34882957	5:39,331,894	missense	0.012	0.900	4.0×10^{-16}
6.2	*FYB*	rs62358735	5:39,199,134	intronic	0.009	0.999	5.1×10^{-13}
7	*PRLR/SPEF2*	rs114092250	5:35,494,448	intergenic	0.019	0.626	2.5×10^{-9}
8.1	*HCG20/LINC00243*	rs114126524	6:30,763,893	downstream	0.171	0.696	6.5×10^{-12}
8.2	*HCG22*	rs140895602	6:31,024,244	nc-transcript	0.021	0.925	1.2×10^{-12}
8.3	*HLA-B*	rs709055	6:31,324,151	missense	0.440	0.999	1.9×10^{-16}
8.4	*HCP5*	rs116319118	6:31,440,641	nc-transcript	0.017	0.522	5.3×10^{-14}
8.5	*HSPA1L/HSPA1A*	rs62395827	6:31,786,730	upstream	0.073	0.999	1.6×10^{-46}
8.6	*NELFE/SKIV2L*	rs116503776	6:31,930,462	intronic	0.120	0.912	2.1×10^{-114}
8.7	*MTCO3P1*	rs114264172	6:32,672,214	downstream	0.051	0.997	2.1×10^{-14}
8.8	*BRD2*	rs200978040	6:32,945,701	missense	0.036	0.638	7.9×10^{-8}
8.9	*COL11A2*	rs114393147	6:33,125,742	downstream	0.041	0.887	2.1×10^{-10}
9	*VEGFA*	rs943080	6:43,826,627	intergenic	0.518	0.437	2.0×10^{-16}
10	*KMT2E/SRPK2*	rs1142	7:104,756,326	downstream	0.357	0.182	1.5×10^{-10}
11	*ZKSCAN1*	rs72615157	7:99,635,967	3'-UTR	0.178	0.486	4.7×10^{-8}
12	*TNFRSF10A*	rs79037040	8:23,082,971	nc-transcript	0.534	0.996	2.9×10^{-12}
13	*MIR6130/RORB*	rs10781180	9:76,615,662	intergenic	0.683	0.068	3.0×10^{-10}
14	*TRPM3*	rs71507014	9:73,438,605	intronic	0.584	0.860	3.2×10^{-9}
15	*TGFBR1*	rs10819635	9:101,864,510	upstream	0.186	0.188	2.4×10^{-11}
16	*ABCA1*	rs2740488	9:107,661,742	intronic	0.266	0.760	1.7×10^{-9}
17	*ARHGAP21*	rs12357257	10:24,999,593	intronic	0.232	0.280	4.3×10^{-9}
18	*ARMS2/HTRA1*	rs3793917	10:124,219,275	upstream	0.316	1.000	$< 9.0 \times 10^{-321}$
19	*RDH5/CD63*	rs3138142	12:56,115,585	coding-syn	0.213	0.847	6.1×10^{-10}

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

20	*MAPKAPK5*	rs61941287	12:112,330,305	intronic	0.019	0.503	1.2×10^{-10}
21	*B3GALTL*	rs9564692	13:31,821,240	splice	0.288	0.889	3.2×10^{-11}
22	*RAD51B*	rs1956526	14:68,799,787	intronic	0.650	0.039	1.0×10^{-11}
23	*ALDH1A2*	rs2414577	15:58,680,638	intronic	0.366	0.495	4.8×10^{-17}
24	*CETP*	rs5817082	16:56,997,349	intronic	0.248	0.193	1.7×10^{-21}
25	*BCAR1*	rs72802395	16:75,286,484	intronic	0.068	0.605	2.1×10^{-11}
26	*POLDIP2/TNFAIP1*	rs13469	17:26,676,135	coding-syn	0.523	0.168	5.1×10^{-9}
27	*NPLOC4/TSPAN10*	rs6420484	17:79,612,397	missense	0.622	0.351	4.0×10^{-12}
28.1	*FUT6*	rs17855739	19:5,831,840	missense	0.044	0.568	1.5×10^{-16}
28.2	*C3*	rs2230199	19:6,718,387	missense	0.764	0.999	1.7×10^{-77}
29	*CNN2*	rs10422209	19:1,026,318	upstream	0.132	0.229	5.2×10^{-9}
30	*APOE/TOMM40*	rs429358	19:45,411,941	missense	0.118	1.000	3.3×10^{-46}
31	*MMP9*	rs2274755	20:44,639,692	splice	0.138	0.194	5.4×10^{-11}
32	*C20orf85*	rs117739907	20:56,652,781	intergenic	0.063	0.079	7.8×10^{-18}
33	*SYN3*	rs5754227	22:33,105,817	intronic	0.124	0.781	2.0×10^{-27}
34	*SLC16A8/PICK1*	rs8135665	22:38,476,276	intronic	0.205	0.596	2.9×10^{-12}

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Table 5: Novel AMD loci (with Bayesian regional-PP >0.95) identified by SFBA, accounting for gene-based functional annotations. Variants with the highest Bayesian single variant PP in the novel loci are listed in this table. Shown are reside genes, dbSNPIDs, positions, functional annotations, MAFs, P-values, Bayesian regional-PPs, and Bayesian PPs/effect-sizes.

Locus	Reside gene	*dbSNPID*	Chr:Position	Anno	MAF	P-value	Regional-PP	Bayesian PP	Effect-size
1	*PPIL3*	*rs7562391*	2:201,736,166	missense	0.127	4.8×10^{-7}	0.989	0.666	-0.061
2	*ZNRD1-AS1*	*rs114318558*	6:29,966,787	downstream	0.175	2.3×10^{-7}	0.993	0.135	0.058
3	*CPN1*	*rs61751507*	10:101,829,514	missense	0.043	6.7×10^{-8}	0.994	0.598	-0.106
4	*ABHD2*	*rs6496562*	15:89,736,558	splice	0.417	8.4×10^{-8}	0.974	0.517	0.042
5	*LBP*	*rs2232613*	20:36,997,655	missense	0.073	4.3×10^{-7}	0.955	0.881	-0.079

Supplementary Table 6: Novel AMD loci (with Fgwas regional-PP >0.95) identified by Fgwas (Supplementary Table 4), accounting for gene-based functional annotations. Variants with the highest Fgwas single variant PP in the novel loci are listed in this table. Shown are reside genes, dbSNPIDs, positions, functional annotations, MAFs, P-values, Fgwas regional-PPs, Fgwas PPs, and Bayesian effect-sizes.

Locus	Reside gene	dbSNPID	Chr:Position	Anno	MAF	P-value	Regional-PP	Fgwas PP	Effect-size
1	*PPIL3*	*rs7562391*	2:201,736,166	missense	0.127	4.8×10^{-7}	0.981	0.322	-0.061
2	*SERPINE2*	*rs114750941*	2:224,875,718	intronic	0.025	3.2×10^{-5}	0.960	0.001	0.125
3	*Intergenic*	*rs4674883*	2:225,184,903	intergenic	0.573	1.2×10^{-7}	0.960	0.141	0.043
4	*ABI3BP*	*rs182405490*	3:100,545,967	nc-transcript	0.007	3.3×10^{-5}	0.999	0.001	0.247
5	*RPL34-AS1*	*rs185276593*	4:109,513,080	nc-transcript	0.116	1.6×10^{-4}	0.989	0.001	-0.056
6	*ZNRD1-AS1*	*rs116112857*	6:29,951,011	downstream	0.027	1.2×10^{-8}	0.999	0.753	-0.141
7	*PACSIN1*	*rs41312309*	6:34,498,328	missense	0.085	2.4×10^{-5}	0.997	0.017	-0.057
8	*CPN1*	*rs61733667*	10:101,802,262	coding-syn	0.036	1.0×10^{-7}	0.996	0.253	-0.118
9	*Intergenic*	*rs7922823*	10:125,058,372	intergenic	0.991	9.4×10^{-6}	0.969	0.001	-0.210
10	*ABHD2*	*rs6496562*	15:89,736,558	splice	0.417	8.4×10^{-8}	0.978	0.252	0.042
11	*SEMA4B*	*rs908044*	15:90,768,959	missense	0.417	1.0×10^{-4}	0.978	0.001	0.032
12	*LBP*	*rs2232613*	20:36,997,655	missense	0.073	4.3×10^{-7}	0.959	0.647	-0.079

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Table 7: AMD risk variants by SFBA in the known 34 loci, accounting for summarized regulatory annotations. Variants with Bayesian PPs >0.5 or the highest Bayesian PPs in the loci are listed (horizontal lines separate loci). Shown are reside/nearby genes, dbSNPIDs, positions, functional annotations, MAFs (unfolded, corresponding to the direction of effect-sizes), Bayesian PPs/effect-sizes, and P-values.

Signal number	Reside/nearby gene	dbSNPID	Chr:Position	Anno	MAF	Bayesian PP	Effect-size	P-value
1.1	*KCNT2*	rs144520124	1:196,371,908	DHS	0.005	1.000	-0.383	1.9×10^{-23}
1.2	*CFH*	rs74979069	1:196,588,463	intergenic	0.049	1.000	0.181	8.1×10^{-92}
1.3	*CFH*	rs1089033	1:196,666,793	intronic	0.412	1.000	-0.117	$< 9.0 \times 10^{-321}$
1.4	*CFH*	rs2133143	1:196,718,099	intergenic	0.165	0.736	-0.358	5.7×10^{-246}
1.5	*CFH*	esv2672010	1:196,733,401	others	0.157	1.000	-0.283	3.3×10^{-314}
1.6	*CFHR3*	rs188826801	1:196,762,123	intronic	0.014	0.993	0.176	1.2×10^{-39}
1.7	*CFH*	rs79251424	1:196,782,416	intergenic	0.030	0.998	0.144	2.1×10^{-6}
1.8	*RP4-608O15.3*	rs146093852	1:196,811,860	intergenic	0.277	0.994	-0.143	5.7×10^{-254}
2	*COL4A3*	rs11884770	2:228,086,920	intronic	0.731	0.213	0.050	5.6×10^{-9}
3	*ADAMTS9-AS2*	rs11914351	3:64,723,441	intronic	0.240	0.950	-0.064	8.7×10^{-7}
4	*COL8A1*	rs140647181	3:99,180,668	intergenic	0.019	0.575	0.221	5.4×10^{-13}
5	*CFI*	rs10033900	4:110,659,067	intergenic	0.506	0.994	-0.067	7.2×10^{-19}
6	*C9*	rs34882957	5:39,331,894	coding	0.012	0.982	0.278	4.0×10^{-9}
7	*PRLR/SPEF2*	rs114092250	5:35,494,448	intergenic	0.019	0.346	-0.172	2.5×10^{-9}
8.1	*C2/CFB*	rs4151667	6:31,914,024	coding	0.035	0.579	-0.284	1.3×10^{-44}
8.2	*SKIV2/NELFE*	rs115270436	6:31,928,306	coding	0.071	0.566	-0.321	2.8×10^{-99}
9	*VEGFA*	rs943080	6:43,826,627	DHS	0.518	0.678	0.063	2.0×10^{-16}
10	*LINC01004/KMT2E-AS1*	rs6950894	7:104,652,671	promoter	0.511	0.063	-0.047	9.8×10^{-10}
11	*PILRB*	rs7783159	7:100,017,454	coding	0.203	0.115	0.059	5.1×10^{-10}
12	*TNFRSF10A*	rs79037040	8:23,082,971	DHS	0.534	0.995	0.053	2.9×10^{-12}
13	*MIR6130/RORB*	rs10781180	9:76,615,662	intergenic	0.684	0.070	-0.052	3.0×10^{-10}
14	*TRPM3*	rs71507014	9:73,438,605	intronic	0.584	0.763	-0.046	3.2×10^{-9}
15	*TGFBR1*	rs401186	9:101,925,077	promoter	0.200	0.109	-0.063	2.5×10^{-11}
16	*ABCA1*	rs2740488	9:107,661,742	intronic	0.266	0.727	-0.053	1.7×10^{-9}
17	*ARHGAP21*	rs12357257	10:24,999,593	intronic	0.232	0.297	0.053	4.3×10^{-9}
18.1	*ARMS2*	rs7068411	10:124,202,878	intergenic	0.621	1.000	0.252	2.4×10^{-212}
18.2	*ARMS2*	rs7898343	10:124,212,887	promoter	0.083	0.868	-0.311	2.0×10^{-51}
18.3	*ARMS2*	rs10490923	10:124,214,251	coding	0.109	0.962	-0.272	1.7×10^{-53}
18.4	*ARMS2*	rs2736911	10:124,214,355	coding	0.137	0.781	-0.350	1.8×10^{-53}
18.5	*HTRA1*	rs2672601	10:124,220,023	promoter	0.136	0.524	-0.321	4.8×10^{-53}
18.6	*HTRA1*	rs74895474	10:124,230,397	intronic	0.094	1.000	-0.199	1.3×10^{-42}

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Signal number	Reside/nearby gene	dbSNPID	Chr:Position	Anno	MAF	Bayesian PP	Effect-size	P-value
18.7	*HTRA1*	rs12252027	10:124,234,988	intronic	0.099	1.000	-0.189	1.4×10^{-51}
18.8	*HTRA1*	rs2672589	10:124,234988	DHS	0.653	1.000	0.220	8.9×10^{-180}
19	*RDH5/CD63*	rs143673140	12:56,514,414	coding	0.009	0.001	-0.096	1.3×10^{-2}
20	*MAPKAPK5*	rs61941287	12:112,330,305	intronic	0.019	0.318	0.199	1.2×10^{-10}
21	*B3GALTL*	rs9564692	13:31,821,240	DHS	0.288	0.429	-0.056	3.2×10^{-11}
22	*RAD51B*	rs2842344	14:68,976,971	DHS	0.899	0.215	-0.082	3.7×10^{-7}
23	*ALDH1A2*	rs2414577	15:58,680,638	DHS	0.366	0.508	-0.067	1.5×10^{-9}
24	*CETP*	rs5883	16:57,007,353	promoter	0.060	0.415	0.085	1.4×10^{-20}
25	*CTRB2*	rs55993634	16:75,236,763	promoter	0.082	0.321	-0.104	4.6×10^{-5}
26	*POLDIP2/TNFAIP1*	rs13469	17:26,676,135	coding	0.524	0.280	0.044	5.2×10^{-9}
27	*NPLOC4/TSPAN10*	rs9894429	17:79,596,811	coding	0.441	0.261	-0.045	4.0×10^{-12}
28.1	*FUT6/NRTN*	rs17855739	19:5,831,840	coding	0.044	0.549	-0.159	1.5×10^{-16}
28.2	*C3/CTD-3128G10.7*	rs147859257	19:6,718,146	coding	0.008	1.000	0.501	4.3×10^{-31}
28.3	*C3/CTD-3128G10.7*	rs2230199	19:6,718,387	coding	0.764	0.999	-0.173	1.7×10^{-77}
29	*ABCA7*	rs3752241	19:1,053,524	coding	0.160	0.268	0.055	3.2×10^{-7}
30	*APOE(EXOC3L2/MARK4)*	rs429358	19:45,411,941	coding	0.118	1.000	-0.173	3.3×10^{-46}
31	*MMP9/RP11-465L10.10*	rs17577	20:44,643,111	coding	0.138	0.377	-0.072	6.8×10^{-11}
32	*RP13-379L11.1*	rs7266392	20:56,651,542	DHS	0.063	0.115	-0.134	9.2×10^{-18}
33	*SYN3*	rs5754227	22:33,105,817	intronic	0.124	0.524	-0.129	2.0×10^{-27}
34	*SLC16A8/BAIAP2L2*	rs77968014	22:38,478,666	coding	0.009	0.842	0.207	3.1×10^{-6}

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Table 8: AMD risk variants by Fgwas method in the known 34 loci, accounting for summarized regulatory annotations. Variants with Fgwas PPs >0.5 or the highest Fgwas PPs in the loci or are listed (horizontal lines separate loci). Shown are reside/nearby genes, dbSNPIDs, positions, annotations, MAFs (unfolded, corresponding to the direction of effect-sizes), Fgwas PPs, and P-values.

Signal number	Reside/nearby gene	dbSNPID	Chr:Position	Anno	MAF	Fgwas PP	P-value
1	*Intergenic*	rs77498516	1:196,115,300	intergenic	0.048	0.522	8.2×10^{-27}
2	*COL4A3*	rs11884770	2:228,086,920	intronic	0.731	0.146	5.7×10^{-9}
3	*Intergenic*	rs61092465	3:65,149,489	intergenic	0.021	0.001	1.6×10^{-3}
4	*Intergenic*	rs140647181	3:99,180,668	intergenic	0.019	0.999	5.4×10^{-13}
5	*CFI*	rs10033900	4:110,659,067	intergenic	0.506	0.996	7.2×10^{-19}
6.1	*C9*	rs34882957	5:39,331,894	coding	0.012	0.757	4.0×10^{-16}
6.2	*FYB*	rs62358735	5:39,199,134	intronic	0.009	0.999	5.1×10^{-13}
7	*Intergenic*	rs114092250	5:35,494,448	intergenic	0.019	0.617	2.5×10^{-9}
8.1	*HCG20/LINC00243*	rs114126524	6:30,763,893	DHS	0.171	0.785	$6.5. \times 10^{-12}$
8.2	*HCG22*	rs140895602	6:31,024,244	intergenic	0.021	0.553	1.2×10^{-12}
8.3	*HSPA1A*	rs62395827	6:31,786,730	DHS	0.073	1.000	1.6×10^{-46}
8.4	*NELFE/SKIV2L*	rs116503776	6:31,930,462	intronic	0.120	0.789	2.1×10^{-114}
8.5	*MTCO3P1*	rs114264172	6:32,672,214	intergenic	0.051	0.997	2.1×10^{-14}
8.6	*BRD2*	rs200978040	6:32,945,701	coding	0.035	0.522	7.9×10^{-8}
8.7	*COL11A2*	rs114393147	6:33,125,742	intergenic	0.041	0.782	2.1×10^{-10}
9	*Intergenic*	rs943080	6:43,826,627	DHS	0.518	0.557	2.0×10^{-16}
10	*KMT2E/SRPK2*	rs1142	7:104,756,326	UTR	0.357	0.215	1.5×10^{-10}
11	*ZKSCAN1*	rs72615157	7:99,635,967	UTR	0.177	0.561	4.7×10^{-8}
12	*TNFRSF10A*	rs79037040	8:23,082,971	DHS	0.534	0.995	2.9×10^{-12}
13	*Intergenic*	rs10781180	9:76,615,662	intergenic	0.683	0.067	3.0×10^{-10}
14	*TRPM3*	rs71507014	9:73,438,605	intronic	0.584	0.837	3.2×10^{-9}
15	*TGFBR1*	rs10760667	9:101,864,607	DHS	0.105	0.186	2.5×10^{-11}
16	*ABCA1*	rs2740488	9:107,661,742	intronic	0.266	0.667	1.7×10^{-9}
17	*ARHGAP21*	rs12357257	10:24,999,593	intronic	0.232	0.227	4.3×10^{-9}
18	*PSTK*	rs140627984	10:124,723,092	intergenic	0.121	0.003	1.4×10^{-6}
19	*OR6C4*	rs7313899	12:55,945,119	coding	0.985	0.001	3.0×10^{-10}
20	*Intergenic*	rs73205633	12:112,357,085	intergenic	0.019	0.495	1.2×10^{-10}
21	*B3GALTL*	rs9564692	13:31,821,240	DHS	0.288	0.543	3.2×10^{-11}
22	*RAD51B*	rs11158728	14:68,762,205	DHS	0.641	0.040	1.2×10^{-11}
23	*ALDH1A2*	rs2414577	15:58,680,638	DHS	0.366	0.500	4.8×10^{-17}
24	*CETP*	rs5817082	16:56,997,349	intronic	0.248	0.179	1.7×10^{-21}
25	*BCAR1*	rs72802395	16:75,286,484	intronic	0.068	0.623	2.1×10^{-11}
26	*POLDIP2/NFAIP1*	rs13469	17:26,676,135	coding	0.523	0.134	5.1×10^{-12}

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

27	*NPLOC4*	rs8070929	17:79,530,993	intronic	0.378	0.176	1.1×10^{-12}
28	*C3*	rs2230199	19:6,718,387	coding	0.764	0.999	1.7×10^{-77}
29	*CNN2/ABCA7*	rs58369307	19:1,038,290	UTR	0.109	0.207	8.5×10^{-9}
30	*APOE/TOMM40*	rs429358	19:45,411,941	coding	0.118	1.000	3.3×10^{-46}
31	*MMP9*	rs17577	20:44,643,111	coding	0.138	0.131	6.8×10^{-11}
32	*RP13-379L11.1*	rs141945849	20:56,650,604	DHS	0.063	0.092	9.3×10^{-18}
33	*SYN3*	rs5754227	22:33,105,817	intronic	0.124	0.681	2.0×10^{-27}
34	*SLC16A8/PICK1*	rs8135665	22:38,476,276	intronic	0.205	0.607	2.9×10^{-12}

Supplementary Table 9: Novel AMD loci (with Bayesian regional-PP >0.95) identified by SFBA, accounting for summarized regulatory annotations. Variants with the highest Bayesian PP in the novel loci are listed in this table. Shown are reside genes, dbSNPIDs, positions, functional annotations, MAFs, P-values, Bayesian regional-PPs, and Bayesian PPs/effect-sizes.

Locus	Reside gene	*dbSNPID*	Chr:Position	Anno	MAF	P-value	Regional-PP	Bayesian PP	Effect-size
1	*PPIL3*	*rs7562391*	2:201,736,166	coding	0.127	4.8×10^{-7}	0.967	0.475	-0.061
2	*ZNRD1-AS1*	*rs114357644*	6:29,924,728	intergenic	0.669	2.3×10^{-7}	0.999	0.609	0.051
3	*CPN1*	*rs61733667*	10:101,829,514	coding	0.036	1.0×10^{-7}	0.994	0.463	-0.118

Supplementary Table 10: Novel AMD loci (with Bayesian regional-PP >0.95) identified by Fgwas, accounting for summarized regulatory annotations. Variants with the highest Fgwas PP in the novel loci are listed in this table. Shown are reside genes, dbSNPIDs, positions, functional annotations, MAFs, P-values, Fgwas regional-PPs, Fgwas PPs, and Bayesian effect-sizes.

Locus	Reside gene	dbSNPID	Chr:Position	Anno	MAF	P-value	Regional-PP	Fgwas PP	Effect-size
1	*PPIL3*	*rs7562391*	2:201,736,166	coding	0.127	4.8×10^{-7}	0.976	0.322	-0.061
2	*SERPINE2*	*rs7588220*	2:224,873,604	DHS	0.025	3.2×10^{-5}	0.966	0.001	0.129
3	*Intergenic*	*rs4674883*	2:225,184,903	intergenic	0.573	1.2×10^{-7}	0.966	0.141	0.043
4	*ABI3BP*	*rs182405490*	3:100,545,967	others	0.007	3.3×10^{-5}	0.999	0.001	0.247
5	*RPL34-AS1*	*rs151204018*	4:108,847,538	others	0.007	4.8×10^{-4}	0.988	0.001	0.254
6	*ZNRD1-AS1*	*rs75140056*	6:29,608,184	intergenic	0.601	9.6×10^{-9}	0.999	0.261	0.045
7	*PACSIN1*	*rs41312309*	6:34,498,328	coding	0.085	2.4×10^{-5}	0.995	0.017	-0.057
8	*CPN1*	*rs61733667*	10:101,802,262	coding	0.036	1.0×10^{-7}	0.994	0.253	-0.118
9	*Intergenic*	*rs7922823*	10:125,058,372	others	0.991	9.4×10^{-6}	0.961	0.001	-0.210
10	*ABHD2*	*rs8042649*	15:89,740,469	UTR	0.417	1.2×10^{-7}	0.973	0.093	0.049
11	*SEMA4B*	*rs11547962*	15:90,772,005	UTR	0.399	4.3×10^{-5}	0.973	0.001	0.032

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Table 11: AMD risk variants by SFBA in the known 34 loci, accounting for chromatin states profiled in the K562 cell type. Variants with Bayesian PPs >0.5 or the highest Bayesian PPs in the loci are listed in this table. Shown are reside/nearby genes, dbSNPIDs, positions, annotations, MAFs (unfolded, corresponding to the direction of effect-sizes), P-values, and Bayesian PPs/effect-sizes.

Signal number	Reside/nearby gene	dbSNPID	Chr:Position	Anno	MAF	Bayesian PP	Effect-size	P-value
1.1	*KCNT2*	*rs72732259*	1:196,464,113	APromoter	0.266	0.915	-0.064	4.2×10^{-196}
1.2	*Intergenic*	*rs74979069*	1:196,588,463	CNV	0.049	1.000	0.160	8.1×10^{-92}
1.3	*CFH*	*rs72734340*	1:196,681,376	CNV	0.037	1.000	-0.189	1.1×10^{-1}
1.4	*Intergenic*	*rs200467660*	1:196,721,770	CNV	0.161	1.000	-0.405	1.1×10^{-249}
1.5	*Intergenic*	*rs79654026*	1:196,725,939	CNV	0.148	0.935	-0.207	2.2×10^{-310}
1.6	*ZNF675*	*rs146093952*	1:196,811,860	CNV	0.277	1.000	-0.207	2.2×10^{-310}
1.7	*CFHR4*	*rs71631868*	1:196,815,711	CNV	0.149	1.000	-0.172	1.3×10^{-295}
1.8	*CFHR5*	*rs139017763*	1:196,965,193	CNV	0.005	1.000	-0.388	2.8×10^{-25}
2	*COL4A3*	*rs11884770*	2:228,086,920	CNV	0.731	0.161	0.051	5.6×10^{-9}
3	*ADAMTS9-AS2*	*rs11914351*	3:64,723,441	CNV	0.240	0.783	-0.064	8.7×10^{-7}
4	*Intergenic*	*rs140647181*	3:99,180,668	CNV	0.019	0.679	0.222	5.3×10^{-13}
5	*CFI*	*rs10033900*	4:110,659,067	WEnhancer	0.506	0.982	-0.067	7.2×10^{-19}
6	*C9*	*rs62358361*	5:39,327,888	CNV	0.012	0.376	0.271	3.1×10^{-16}
7	*Intergenic*	*rs114092250*	5:35,494,448	WEnhancer	0.019	0.659	-0.171	2.5×10^{-9}
8.1	*C6orf48*	*rs200497397*	6:31,810822	WEnhancer	0.028	0.990	0.160	9.8×10^{-15}
8.2	*PBX2/AGER/GPSM3*	*rs114254831*	6:32,155,581	SEnhancer	0.271	0.999	0.080	8.1×10^{-13}
9	*Intergenic*	*rs943080*	6:43,826,627	CNV	0.518	0.397	0.063	2.0×10^{-16}
10	*KMT2E/SRPK2*	*rs1144*	7:104,756,355	Txn_Elongation	0.362	0.100	0.057	1.6×10^{-10}
11	*TSC22D4*	*rs11559117*	7:100,076,614	APromoter	0.202	0.034	0.059	7.8×10^{-10}
12	*TNFRSF10A*	*rs79037040*	8:23,082,971	APromoter	0.534	0.993	0.053	2.9×10^{-12}
13	*Intergenic*	*rs1078176*	9:76,592,874	APromoter	0.684	0.229	-0.052	3.0×10^{-10}
14	*TRPM3*	*rs71507014*	9:73,438,605	CNV	0.585	0.734	-0.046	3.2×10^{-9}
15	*TGFBR1*	*rs10819635*	9:10,819,635	WEnhancer	0.186	0.117	-0.066	2.5×10^{-11}
16	*ABCA1*	*rs2740488*	9:107,661,742	CNV	0.266	0.736	-0.053	1.7×10^{-9}
17	*ARHGAP21*	*rs12357257*	10:24,999,593	Txn_Elongation	0.232	0.274	0.053	4.3×10^{-9}
18.1	*Intergenic*	*rs7068411*	10:124,202,878	CNV	0.621	1.000	0.198	2.4×10^{-212}
18.2	*HTRA1*	*rs2672595*	10:124,227,288	CNV	0.213	0.844	-0.466	8.7×10^{-111}
18.3	*HTRA1*	*rs4752699*	10:124,234,320	CNV	0.128	1.000	-0.292	2.1×10^{-51}
18.4	*HTRA1*	*rs2672589*	10:124,234,988	CNV	0.653	1.000	0.274	8.9×10^{-180}
19	*SARNP*	*rs77232256*	12:56,170,342	Txn_Elongation	0.024	0.001	0.132	2.5×10^{-4}
20	*NAA25*	*rs56143183*	12:112,545,374	APromoter	0.048	0.541	0.155	4.8×10^{-9}
21	*B3GALTL*	*rs9564692*	13:31,821,240	CNV	0.288	0.379	-0.056	3.2×10^{-11}

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Signal number	Reside/nearby gene	dbSNPID	Chr:Position	Anno	MAF	Bayesian PP	Effect-size	P-value
22	*RAD51B*	*rs2842339*	14:68,986,999	CNV	0.899	0.243	-0.082	3.1×10^{-7}
23	*ALDH1A2*	*rs2414577*	15:58,680,638	Txn_Elongation	0.366	0.483	-0.067	4.8×10^{-17}
24	*CETP*	*rs17231569*	16:56,999,778	WEnhancer	0.172	0.255	-0.072	9.4×10^{-21}
25	*CTRB2*	*rs72802342*	16:75,234,872	CNV	0.074	0.317	-0.114	2.8×10^{-13}
26	*SARM1/SLC46A1*	*rs4795434*	17:26,716,917	WEnhancer	0.524	0.112	0.045	1.8×10^{-9}
27	*NPLOC4*	*rs8070929*	17:79,530,993	Txn_Elongation	0.378	0.188	0.058	1.1×10^{-12}
28.1	*C3*	*rs147859257*	19:6,718,146	SEnhancer	0.008	1.000	0.504	4.3×10^{-31}
28.2	*C3*	*rs2230199*	19:6,718,387	SEnhancer	0.764	0.999	-0.172	1.7×10^{-77}
29	*CNN2/ABCA7*	*rs3087680*	19:1,038,289	SEnhancer	0.109	0.360	0.072	8.6×10^{-9}
30.1	*APOE/TOMM40*	*rs429358*	19:45,411,941	Txn_Elongation	0.118	1.000	-0.186	3.3×10^{-46}
31	*MMP9*	*rs17577*	20:44,643,111	APromoter	0.138	0.181	-0.072	6.8×10^{-11}
32	*Intergenic*	*rs140611615*	20:56,653,111	CNV	0.062	0.080	-0.135	8.2×10^{-18}
33	*SYN3*	*rs5754227*	22:33,105,817	CNV	0.124	0.774	-0.128	2.0×10^{-27}
34	*SLC16A8/PICK1/ BAIAP2L2*	*rs8135665*	22:38,476,276	CNV	0.206	0.652	0.066	2.9×10^{-12}

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Table 12: AMD risk variants by Fgwas method in the known 34 loci, accounting for chromatin states profiled in the K562 cell type. Variants with either the highest Fgwas PP per locus or Fgwas PP > 0.5 are listed (horizontal lines separate loci). Shown are reside/nearby genes, dbSNPIDs, positions, functional annotations, MAFs (unfolded, corresponding to the direction of effect-sizes), Fgwas PPs, and P-values.

Signal number	Reside/Nearby Gene	dbSNPID	Chr:Position	Anno	MAF	Fgwas PP	P-value
1	*CFH*	rs77498516	1:196,115,300	CNV	0.048	0.522	8.2×10^{-27}
2	*COL4A3*	rs11884770	2:228,086,920	CNV	0.731	0.183	5.7×10^{-9}
3	*ADAMTS9-AS2*	rs61092465	3:65,149,489	CNV	0.021	0.001	1.6×10^{-3}
4	*COL8A1*	rs140647181	3:99,180,668	CNV	0.019	0.999	5.4×10^{-13}
5	*CFI*	rs10033900	4:110,659,067	WEnhancer	0.506	0.996	7.2×10^{-19}
6.1	*C9*	rs62358361	5:39,327,888	CNV	0.012	0.559	3.1×10^{-16}
6.2	*FYB*	rs62358735	5:39,199,134	APromoter	0.009	0.999	5.1×10^{-13}
7	*PRLR/SPEF2*	rs114092250	5:35,494,448	WEnhancer	0.019	0.673	2.5×10^{-9}
8.1	*HCG20/LINC00243*	rs114126524	6:30,763,893	SEnhancer	0.171	0.810	6.5×10^{-12}
8.2	*HCG22*	rs140895602	6:31,024,244	CNV	0.021	0.535	1.2×10^{-12}
8.3	*HCP5*	rs116319118	6:31,440,641	CNV	0.017	0.521	5.3×10^{-14}
8.4	*HSPA1L/HSPA1A*	rs62395827	6:31,786,730	SEnhancer	0.073	0.999	1.6×10^{-46}
8.5	*NELFE/SKIV2L*	rs116503776	6:31,930,462	TxnElongation	0.120	0.939	2.1×10^{-114}
8.6	*MTCO3P1*	rs114264172	6:32,672,214	CNV	0.051	0.997	2.1×10^{-14}
8.7	*COL11A2*	rs114393147	6:33,125,742	CNV	0.041	0.784	2.1×10^{-10}
9	*VEGFA*	rs943080	6:43,826,627	CNV	0.518	0.428	2.0×10^{-16}
10	*KMT2E/SRPK2*	rs1142	7:104,756,326	TxnElongation	0.357	0.124	1.5×10^{-10}
11	*ZKSCAN1*	rs1122598	7:99,699,436	APromoter	0.177	0.351	8.9×10^{-8}
12	*TNFRSF10A*	rs79037040	8:23,082,971	APromoter	0.534	0.992	2.9×10^{-12}
13	*Intergenic*	rs10781176	9:76,592,874	APromoter	0.684	0.109	3.0×10^{-10}
14	*TRPM3*	rs71507014	9:73,438,605	CNV	0.584	0.858	3.2×10^{-9}
15	*TGFBR1*	rs10760667	9:101,864,607	SEnhancer	0.186	0.132	2.5×10^{-11}
16	*ABCA1*	rs2740488	9:107,661,742	CNV	0.266	0.761	1.7×10^{-9}
17	*ARHGAP21*	rs12357257	10:24,999,593	TxnElongation	0.232	0.308	4.3×10^{-9}
18	*PSTK*	rs140627984	10:124,723,092	TxnElongation	0.121	0.011	1.4×10^{-6}
19	*OR6C7P*	rs7487174	12:55,738,093	APromoter	0.824	0.001	1.6×10^{-3}
20	*MAPKAPK5*	rs61941287	12:112,330,305	TxnElongation	0.019	0.542	1.2×10^{-10}
21	*B3GALTL*	rs9564692	13:31,821,240	CNV	0.288	0.388	3.2×10^{-11}
22	*RAD51B*	rs11158728	14:68,762,205	SEnhancer	0.640	0.082	1.0×10^{-11}
23	*ALDH1A2*	rs2414577	15:58,680,638	TxnElongation	0.366	0.495	4.8×10^{-17}
24	*CETP*	rs5817082	16:56,997,349	CNV	0.248	0.236	1.7×10^{-21}
25	*BCAR1*	rs72802395	16:75,286,484	TxnElongation	0.068	0.653	2.1×10^{-11}

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

26	*TMEM97/KRT18P55*	rs11080055	17:26,649,724	TxnElongation	0.525	0.103	5.1×10^{-9}
27	*NPLOC4*	rs8070929	17:79,530,993	TxnElongation	0.378	0.186	1.1×10^{-12}
28.1	*FUT6*	rs12019136	19:5,835,677	CNV	0.042	0.614	3.7×10^{-17}
28.2	*C3*	rs2230199	19:6,718,387	APromoter	0.764	0.997	1.7×10^{-77}
29	*CNN2/ABCA7*	rs58369307	19:1,038,290	SEnhancer	0.109	0.151	8.5×10^{-9}
30.1	*APOE/TOMM40*	rs429358	19:45,411,941	TxnElongation	0.118	1.000	3.3×10^{-46}
30.2	*MARK4/AC006126.4*	rs73036519	19:45,748,362	SEnhancer	0.293	0.507	3.6×10^{-8}
31	*MMP9*	rs142450006	20:44,614,991	CNV	0.133	0.132	1.4×10^{-11}
32	*C20orf85*	rs117739907	20:56,652,781	CNV	0.062	0.079	7.8×10^{-18}
33	*SYN3*	rs5754227	22:33,105,817	CNV	0.124	0.781	2.0×10^{-27}
34	*SLC16A8/PICK1*	rs8135665	22:38,476,276	CNV	0.205	0.649	2.9×10^{-12}

Supplementary Table 13: Novel AMD loci (with Bayesian regional-PP>0.95) identified by SFBA, accounting for chromatin states profiled in the K562 cell type. Variants with the highest Bayesian PPs in the novel loci are listed in this table. Shown are reside genes, dbSNPIDs, positions, functional annotations, MAFs, P-values, Bayesian regional-PPs, and Bayesian PPs/effect-sizes.

Locus	Reside gene	*dbSNPID*	Chr:Position	Anno	MAF	P-value	Regional-PP	Bayesian PP	Effect-size
2	*ZNRD1-AS1*	*rs114357644*	6:29,924,728	TxnElongation	0.669	2.3×10^{-7}	0.999	0.669	0.051
3	*CPN1*	*rs111563092*	10:101,808,993	CNV	0.045	7.2×10^{-8}	0.970	0.081	-0.106

Supplementary Table 14: Novel AMD loci (with Bayesian regional-PP>0.95) identified by Fgwas, accounting for chromatin states profiled in the K562 cell type. Variants with the highest Fgwas PPs in the novel loci are listed in this table. Shown are reside genes, dbSNPIDs, positions, functional annotations, MAFs, P-values, Fgwas regional-PPs, Fgwas PPs, and Bayesian effect-sizes.

Locus	Reside gene	dbSNPID	Chr:Position	Anno	MAF	P-value	Regional-PP	Fgwas PP	Effect-size
1	*PPIL3*	*rs3851973*	2:201,732,878	SEnhancer	0.127	1.1×10^{-7}	0.963	0.094	-0.059
2	*SERPINE2*	*rs7588220*	2:224,873,604	WEnhancer	0.025	3.2×10^{-5}	0.966	0.001	0.129
3	*Intergenic*	*rs4674883*	2:225,184,903	CNV	0.573	1.2×10^{-7}	0.965	0.141	0.043
4	*ABI3BP*	*rs182405490*	3:100,545,967	CNV	0.007	3.3×10^{-5}	0.999	0.001	0.247
5	*RPL34-AS1*	*rs151204018*	4:108,847,538	CNV	0.007	4.8×10^{-4}	0.988	0.001	0.254
6	*ZNRD1-AS1*	*rs75140056*	6:29,608,184	TxnElongation	0.601	9.6×10^{-9}	0.999	0.261	0.045
7	*PACSIN1*	*rs6922076*	6:33,807,565	SEnhancer	0.446	9.9×10^{-6}	0.995	0.004	-0.035
8	*CPN1*	*rs111563092*	10:101,808,993	CNV	0.045	7.2×10^{-8}	0.993	0.088	-0.106
9	*ABHD2*	*rs2070780*	15:89,760,997	CNV	0.485	1.6×10^{-7}	0.968	0.075	0.043
10	*SEMA4B*	*rs11547962*	15:90,772,005	TxnElongation	0.399	4.3×10^{-5}	0.973	0.001	0.032

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Table 15: Haplotype analysis in locus C2/CFB/SKIV2L, consisting with the top significant intronic variant found by single variant test P-values (*rs116503776* with p-value=2.1×10^{-114}), the top two significant missense variants (in the $\pm$20KB region around *rs116503776*) found by SFBA (*rs4151667* with Bayesian PP=0.903, *rs115270436* with Bayesian PP= 0.638).

Region	**Haplotype**			**Haplotype Frequency (%)**		**P-value**	**OR (95% CI)**
	SKIV2L intronic (*rs116503776*)	*CFB* missense (*rs4151667*)	*CFB* missense (*rs115270436*)	Cases	Controls		
C2/CFB/SKIV2L	1	1	1	1.5×10^{-3}	4.2×10^{-3}	8.9×10^{-11}	0.364 (0.265, 0.501)
	1	0	1	0.046	0.085	1.5×10^{-86}	0.522 (0.490, 0.557)
	1	1	0	0.023	0.041	5.0×10^{-36}	0.561 (0.513, 0.613)
	0	0	1	8.9×10^{-4}	1.5×10^{-3}	0.024	0.586 (0.375, 0.917)
	1	0	0	0.018	0.017	0.092	1.102 (0.983, 1.236)
	0	0	0	0.909	0.850	1.0×10^{-22}	1.752 (1.670, 1.838)
	0	1	0	6.1×10^{-5}	2.8×10^{-5}	0.306	1.840 (0.243, 13.938)

Supplementary Table 16: Linear regression analysis with a model with the top two independent significant variants (*rs116503776*, *rs114254831*) found by conditional analysis, versus a model with the top two significant variants (*rs4151667*, *rs115270436*) found by SFBA accounting for functional annotations.

Region (*C2/CFB/SKIV2L***)**	*SKIV2L* intronic (*rs116503776*) & *PBX2* intronic (*rs114254831*)	*CFB* missense (*rs4151667*) **&** *SKIV2L* missense (*rs115270436*)	Differences (col2-col3)
Akaike information criterion (AIC)	95857.36	95752.63	104.73
Bayesian information criterion (BIC)	95891.1	95786.36	104.74
Log Likelihood	-47924.68	-47872.31	-52.37

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Table 17: Number of loci (regional-PP>0.95) identified by accounting for chromatin states profiled in 9 human cell types, with the number of variants that contribute 95% posterior probabilities.

Cell types	Number of identified loci	Total number of variants	Average number of variants per locus
H1-hESC	32	481	15.0
K562	**31**	**454**	**14.6**
GM12878	31	481	15.5
HepG2	35	609	18.4
HUVEC	32	595	18.5
HSMM	33	608	18.4
NHLF	33	542	16.4
NHEK	31	524	16.9
HMEC	34	529	15.5

bioRxiv preprint doi: https://doi.org/10.1101/101691; this version posted January 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.