*Bioinformatics*, 32(4), 2016, 542–548
doi: 10.1093/bioinformatics/btv610
Advance Access Publication Date: 25 October 2015
Original Paper

OXFORD

Genetics and population analysis

# GenoWAP: GWAS signal prioritization through integrated analysis of genomic functional annotation

**Qiongshi Lu[1], Xinwei Yao[2], Yiming Hu[1] and Hongyu Zhao[1,3,4,*]**

[1]Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA, [2]Yale College, New Haven, CT, USA, [3]Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA and [4]VA Cooperative Studies Program Coordinating Center, West Haven, CT, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on May 18, 2015; revised on September 15, 2015; accepted on October 16, 2015

## Abstract

**Motivation:** Genome-wide association study (GWAS) has been a great success in the past decade. However, significant challenges still remain in both identifying new risk loci and interpreting results. Bonferroni-corrected significance level is known to be conservative, leading to insufficient statistical power when the effect size is moderate at risk locus. Complex structure of linkage disequilibrium also makes it challenging to separate causal variants from nonfunctional ones in large haplotype blocks. Under such circumstances, a computational approach that may increase signal replication rate and identify potential functional sites among correlated markers is urgently needed.

**Results:** We describe GenoWAP, a GWAS signal prioritization method that integrates genomic functional annotation and GWAS test statistics. The effectiveness of GenoWAP is demonstrated through its applications to Crohn's disease and schizophrenia using the largest studies available, where highly ranked loci show substantially stronger signals in the whole dataset after prioritization based on a subset of samples. At the single nucleotide polymorphism (SNP) level, top ranked SNPs after prioritization have both higher replication rates and consistently stronger enrichment of eQTLs. Within each risk locus, GenoWAP may be able to distinguish functional sites from groups of correlated SNPs.

**Availability and implementation:** GenoWAP is freely available on the web at http://genocanyon.med.yale.edu/GenoWAP

**Contact:** hongyu.zhao@yale.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

In the past ten years, genome-wide association studies (GWAS) have been designed and applied to identify disease genes for almost all complex diseases. As of January 15, 2015, 15 216 single nucleotide polymorphisms (SNP) from over 2000 publications have been documented in the GWAS Catalog (Hindorff *et al.*, 2009). Despite its great success in identifying disease-associated loci, scientists have noted several limitations of current GWAS approaches. First, although linkage disequilibrium (LD) is the basis of GWAS, it also hinders the interpretation of association results. Due to the complex LD structure among SNPs, it is the disease-associated haplotype blocks varying in size from a few kb to more than 100 kb (Wall and Pritchard, 2003) that are identified in GWASs. Therefore, the resolution of GWAS is not sufficient for distinguishing causal variants from a large group of correlated SNPs, especially in non-coding regions where the mechanism of genomic function is still largely

unknown (Cooper and Shendure, 2011; Visscher *et al.*, 2012; Ward and Kellis, 2012). Second, although Bonferroni-corrected significance threshold (i.e. $5 \times 10^{-8}$) is widely accepted as the standard cutoff in GWAS analysis, it is well known that Bonferroni correction, as an approach that controls family-wise error rate, is conservative when the number of hypotheses is large and there are many weak to moderate signals (Efron, 2010). In fact, for most complex diseases, numerous genomic loci are involved in disease etiology while each locus only has a moderate effect size. Therefore, studies based on high-throughput genomic scan may be underpowered if the sample size is not large enough. This has led to so-called missing heritability which refers to the gap between the narrow-sense heritability estimated from twin/pedigree analysis and the proportion of the variance explained by significant SNPs identified from GWAS, that has been reported for many diseases (Manolio *et al.*, 2009; Witte *et al.*, 2014). One explanation of missing heritability is the insufficient statistical power to identify all the disease-associated SNPs (Eichler *et al.*, 2010).

Variant prioritization techniques are crucial for post-GWAS analysis on different scales. Locally, it may be able to reveal truly functional sites within each significant locus. Globally, signals at some loci can be enhanced if proper prior information is used. Many variant prioritization methods have been proposed (Hou and Zhao, 2013). Supervised-learning-based statistical tools for predicting deleterious variants are probably the richest among available approaches. So far, most of the existing deleteriousness prediction tools only focus on protein-coding genes in the human genome. However, coding-region-based tools are not sufficient for GWAS signal prioritization because nearly 90% of the significant SNPs identified in GWAS are intronic or intergenic (Eicher *et al.*, 2015; Hindorff *et al.*, 2009). A few tools targeting non-coding variants have been proposed (Fu *et al.*, 2014; Kircher *et al.*, 2014; Ritchie *et al.*, 2014; Shihab *et al.*, 2015). Detailed comparisons of these methods were reviewed elsewhere (Cooper and Shendure, 2011; Wang *et al.*, 2015). Unlike the extensively studied protein-altering variants, very few non-coding pathogenic variants have been revealed so far (Ward and Kellis, 2012). Moreover, non-coding variants span a much wider functional spectrum. Varied and complex mechanisms from cell-specific Transcription Factor Binding Sites (TFBS), to enhancers, insulators, short and long range epigenetic and structural effects on DNA, make it challenging to understand non-coding variants in the human genome. Therefore, existing non-coding variant prioritization tools based on supervised-learning may suffer from the potentially biased training data. Their performance in GWAS signal prioritization remains to be further investigated. Finally, although deleteriousness of a single SNP is crucial for identifying causal variants, it does not provide all the information needed in GWAS signal prioritization, where each SNP in GWAS also carries information of nearby variants that are not genotyped. A better-informed method should be able to measure the functional potential for the surrounding region of each genotyped marker.

Recently, Lu et al. developed GenoCanyon, a statistical framework to predict functional non-coding regions in the human genome through integrated analysis of multiple biochemical signals and genomic conservation measures (Lu *et al.*, 2015). Its unsupervised-learning framework makes GenoCanyon suffer less from our limited knowledge of non-coding genome. Moreover, since the resolution of its functional prediction is at the nucleotide level, it is possible to use GenoCanyon scores to evaluate the surrounding region of each genotyped SNP. In this paper, we propose Genome Wide Association Prioritizer (GenoWAP), a GWAS signal prioritization approach that integrates GenoCanyon functional prediction and GWAS *P*-values. We apply the method on two smaller GWASs of Crohn's disease and schizophrenia, respectively, to prioritize SNPs. The performance is evaluated using the results from large GWAS meta-analyses of these two diseases. Compared to the top loci ranked on *P*-values only, top ranked loci after prioritization tend to show substantially stronger signals in large GWAS studies. Within each locus, GenoWAP may be able to distinguish true signals among highly correlated SNPs. The method has the potential to reduce noises caused by LD and rescue marginal signals in GWASs with insufficient sample sizes.

## 2 Methods

### 2.1 Statistical model

For each SNP, we define $Z$ to be the indicator of general functionality, and define $Z_D$ to be the indicator of disease-specific functionality. More specifically, if a SNP or its surrounding region is active in any genomic functional pathway, then $Z$ equals to 1. If this SNP or the surrounding region is involved in the disease pathway, then $Z_D$ equals to 1. For each SNP, we use $p$ to denote its *P*-value obtained from the standard GWAS analysis.

The goal of GWAS signal prioritization is to assign each SNP a new score that measures its importance. A reasonable quantity would be the conditional probability of being disease-specific functional given the *P*-value, i.e. $P(Z_D = 1|p)$. Using Bayes formula, we can rewrite the conditional probability as below.

$$P(Z_D = 1|p) = \frac{f(p|Z_D = 1) \times P(Z_D = 1)}{f(p|Z_D = 1) \times P(Z_D = 1) + f(p|Z_D = 0) \times P(Z_D = 0)} \quad (1)$$

Based on the definitions of $Z$ and $Z_D$, SNPs satisfying $Z_D = 1$ must be a subset of the SNPs satisfying $Z = 1$. This is because if a SNP is disease-specific functional, it has to be functional in the general sense as well. Therefore, we get the following formula.

$$\begin{aligned} P(Z_D = 1) &= P(Z = 1, Z_D = 1) \\ &= P(Z_D = 1|Z = 1) \times P(Z = 1) \end{aligned} \quad (2)$$

In order to calculate the conditional probability $P(Z_D = 1|p)$ for a marker, we need its prior probability of being functional, i.e. $P(Z = 1)$, the *P*-value density for disease-specific functional markers, i.e. $f(p|Z_D = 1)$, the *P*-value density for markers that are not related to the disease, i.e. $f(p|Z_D = 0)$, and finally, the conditional probability of being disease-specific functional given the marker is functional in the general sense, i.e. $P(Z_D = 1|Z = 1)$.

### 2.2 Estimation

Recently, Lu et al. developed GenoCanyon, a statistical framework predicting functional genomic regions (Lu *et al.*, 2015). Through integrating diverse types of annotation data, including genomic conservation measures, DNase hypersensitivity, FAIRE, histone modifications and transcription factor binding activities using unsupervised learning techniques, GenoCanyon measures the functional potential for each nucleotide in the human genome. For each SNP in a GWAS dataset, the mean GenoCanyon functional score of its surrounding 10 000 base pairs is used as the prior probability $P(Z = 1)$. Different from using variant-based annotation tools as the prior knowledge, this prior information not only measures the importance of the genotyped marker, but also evaluates its surrounding region where ungenotyped causal variants may reside.

Next, we partition all the SNPs into functional ($Z = 1$) and non-functional ($Z = 0$) subgroups based on the calculated mean

GenoCanyon scores with cutoff 0.1. Since GenoCanyon functional score has a bimodal pattern, this partition is not sensitive to the cutoff choice. The partition step is necessary because of two major reasons. First, the signal pattern in the functional subgroup is amplified after noise reduction (Fig. 1). The increase in the proportion of disease-related markers in the functional subgroup leads to more stable estimates in the following steps. Second, the *P*-value density for non-functional markers, i.e. $f(p|Z=0)$, can now be estimated empirically. Since the *P*-values are acquired from a disease-specific study, we assume that the *P*-values for markers not related to the disease behave just like the *P*-values for markers that are not functional entirely. Mathematically, this assumption is characterized as the equation below.

$$f(p|Z_D=0)=f(p|Z=0) \tag{3}$$

Based on this assumption, we can estimate $f(p|Z_D=0)$ using the *P*-values for SNPs in the non-functional subgroup. Notably, it may seem natural to assume $(p|Z_D=0)$ follows a uniform distribution. However, the *P*-value of a marker with $Z_D=0$ can actually be driven by a nearby disease-related marker due to LD. The empirically estimated density can capture a certain amount of LD information, which is complex and non-trivial to model. Moreover, it is common to see some variants with low minor allele frequencies in GWAS samples. The *P*-values for these markers will form a spike near 1 in the *P*-value density. The empirically estimated density is also able to account for this artifact. We propose to use histogram for density estimation, because it has stable performance near the boundary. In fact, the *P*-value boundary near 0 is where the real signals reside, and the boundary near 1 occasionally has the artifact issue caused by rare variants. Histogram is able to capture both issues. Moreover, the sample size in this empirical Bayes framework is the total number of markers, which is usually large in GWAS.

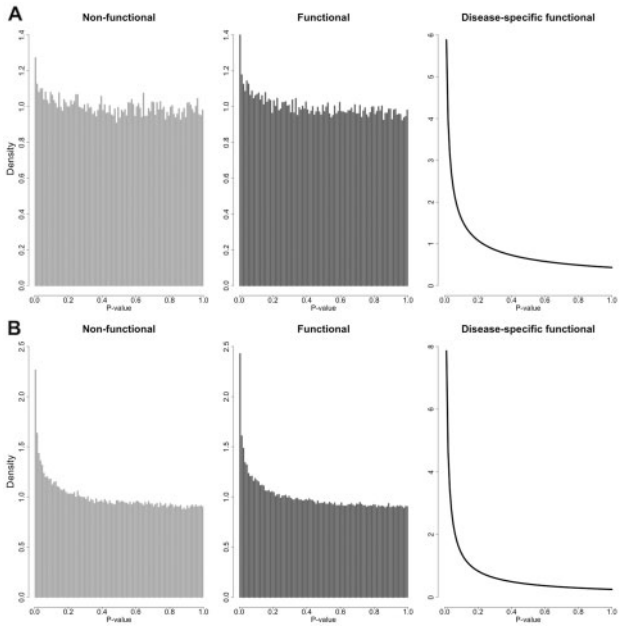


**Fig. 1.** *P*-value densities of different subgroups of SNPs. (A) *P*-value histogram of non-functional SNPs ($Z=0$), *P*-value histogram of functional SNPs ($Z=1$), and estimated *P*-value density of disease-specific functional SNPs ($Z_D=1$) in the NIDDK GWAS of Crohn's disease. (B) *P*-value histogram of non-functional SNPs ($Z=0$), *P*-value histogram of functional SNPs ($Z=1$) and estimated *P*-value density of disease-specific functional SNPs ($Z_D=1$) in the PGC2011 GWAS of schizophrenia

Therefore, histogram is a reasonable choice for density estimation. The number of bins can be chosen based on cross-validation.

It still remains to estimate the *P*-value density for disease-related markers $f(p|Z_D=1)$, and the conditional probability $P(Z_D=1|Z=1)$. Now, we partition the functional subgroup $(Z=1)$ into finer subgroups. First, based on Eq. (3), it is straightforward to show that

$$f(p|Z=1,\ Z_D=0)=f(p|Z_D=0)=f(p|Z=0) \tag{4}$$

Therefore, the *P*-value density for functional markers is the following mixture.

$$\begin{aligned} f(p|Z=1) &= P(Z_D=1|Z=1)\times f(p|Z=1,Z_D=1) \\ &\quad + P(Z_D=0|Z=1)\times f(p|Z=1,Z_D=0) \\ &= P(Z_D=1|Z=1)\times f(p|Z_D=1) \\ &\quad + P(Z_D=0|Z=1)\times f(p|\ Z_D=0) \end{aligned} \tag{5}$$

In formula (5), $f(p|Z_D=0)$ has already been estimated in previous steps. We further assume a parametric form of $f(p|Z_D=1)$. In a recent work of Chung *et al.* (2014), they showed that beta distribution is a robust approximation of *P*-value distribution under some general assumptions of SNP effect size. We adopt the same assumption.

$$(p|Z_D=1)\sim Beta(\alpha,1),\ \ 0<\alpha<1 \tag{6}$$

The constraint $0<\alpha<1$ guarantees that a smaller *P*-value is more likely to occur than a larger *P*-value. Then, we apply the EM algorithm on all the *P*-values in the functional subgroup. One advantage of beta distribution assumption is that a closed-form expression is available at each iteration in the EM algorithm. In this way, the estimates for both $P(Z_D=1|Z=1)$ and $P(p|Z_D=1)$ can be acquired. Finally, since all missing pieces in formula (1) have been estimated, we calculate the conditional probability $P(Z_D=1|p)$ for all the SNPs. This quantity is referred to as the posterior score in this paper.

## 2.3 Data resource and preprocessing

Test statistics of the NIDDK study (Rioux *et al.*, 2007) were downloaded from dbGap (Supplementary Table S1). Among the 298 391 SNPs, 70 were deleted due to unavailable hg19 genomic locations. We calculated the posterior scores for the remaining 298 321 SNPs (Supplementary Fig. S1). Test statistics of the IIBDGC meta-analysis (Franke *et al.*, 2010) were downloaded from the IIBDGC website (http://www.ibdgenetics.org). The dataset contains 953 241 SNPs, including 262 621 SNPs overlapping with the NIDDK dataset.

Test statistics for studies of schizophrenia (Ripke *et al.*, 2011, 2014) were downloaded from the PGC website (Supplementary Table S2). Coordinates were converted to hg19 using UCSC liftover tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Among the 1 252 901 SNPs in PGC2011 study, 264 were removed due to unavailable hg19 locations. Posterior scores were calculated for all the remaining 1 252 637 SNPs (Supplementary Fig. S2). PGC2014 study contains 9 444 230 SNPs, including 1 179 913 SNPs overlapping with the PGC2011 dataset.

Finally, eQTL data used in the enrichment analysis include single-tissue eQTLs from GTEx Analysis Release V4 (Ardlie *et al.*, 2015), cis and trans eQTLs downloaded from Blood eQTL Browser (Westra *et al.*, 2013), and quantitative trait loci for DNA methylation and gene expression in human brain (Gibbs *et al.*, 2010) downloaded from NCBI eQTL Browser (http://www.ncbi.nlm.nih.gov/projects/gap/eqtl/index.cgi).

## 3 Results

### 3.1 Application to Crohn's disease

Several GWASs of different scales have been performed for Crohn's disease. The largest GWAS meta-analysis, which identified 71 disease-associated loci, is among the studies identifying most significant hits to date (Franke *et al.*, 2010). We applied GenoWAP on a smaller Crohn's disease GWAS conducted by the North American National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) IBD Genetics Consortium, and tested the results using the large meta-analysis done by the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC). Cohort information is listed in Supplementary Table S1. Details of both studies have been reported previously (Franke *et al.*, 2010; Rioux *et al.*, 2007). It is worth noting that the samples in these two studies overlap with each other. However, the goal for this paper is not to replicate the detected signals in an independent cohort. Instead, we seek to better prioritize signals using only a small sample size. In order to test the performance, the results from the largest study available are used as the gold standard.

A total of 71 loci passed genome-wide significance level in the validation stage of IIBDGC meta-analysis, including 32 previously reported risk loci and 39 newly confirmed risk loci (Franke *et al.*, 2010). We ranked the 298 321 SNPs in the NIDDK study based on their *P*-values and posterior scores, respectively. Then, within each of the 71 loci, we compared the rank of the lowest *P*-value to the rank of the largest posterior score. 56 out of 71 loci (79%) had an improved rank, 3 loci (4%) had an equal rank, while only 12 loci (17%) had a reduced rank (Supplementary Table S3). The probability of having an increased rank is significantly higher than that of having a decreased rank (*P*-value $= 3.11 \times 10^{-8}$, one-sided binomial test).

Next, we compared the top 20 loci with the smallest *P*-values to the top 20 loci with the largest posterior scores in the NIDDK study. The locus information and the lowest meta-analysis *P*-value at each locus are listed in Supplementary Table S4. 14 out of 20 loci are shared between the two lists. Interestingly, the posterior-specific loci, i.e. the loci that show up only in the list based on posterior score, showed substantially stronger signals in the IIBDGC meta-analysis compared to the *P*-value-specific loci (Supplementary Table S4, Fig. 2A). For example, the risk locus on chromosome 10q22 was a genome-wide significant locus in the meta-analysis (rs1250550, $P_{meta} = 2.00 \times 10^{-10}$). Although the same SNP, rs1250550, had the lowest *P*-value at this locus in the NIDDK dataset ($P_{NIDDK} = 5.95 \times 10^{-5}$), the signal was not strong enough to make this locus surpass other loci such as the one on chromosome 2q24 (rs6733000, $P_{NIDDK} = 2.01 \times 10^{-5}$). However, with posterior scores, locus 10q22 was ranked as the 17th top locus, while the highest posterior score at locus 2q24 was only 0.0142, which agrees with its weak signal in the meta-analysis result ($P_{meta} = 0.019$). Overall, two posterior-specific loci were genome-wide significant in the meta-analysis, while the lowest $P_{meta}$ among the six *P*-value-specific loci was only $1.10 \times 10^{-4}$. These results show that our method can effectively reduce noises likely due to LD and chance and enhance true signals at disease risk loci.

Next, we check if SNPs with high posterior scores are more enriched of eQTLs. The top 1000 SNPs based on *P*-values are moderately enriched for GTEx whole-blood eQTLs (*P*-value = 0.013; hypergeometric test; fold enrichment = 1.60), while the enrichment for the top 1000 SNPs based on the posterior scores is highly significant (*P*-value $= 8.58 \times 10^{-7}$; fold enrichment = 2.47). The difference becomes even more drastic when using the top 2000 SNPs, with *P*-values 0.001 and $9.25 \times 10^{-15}$ (fold enrichment 1.60 and 2.73),

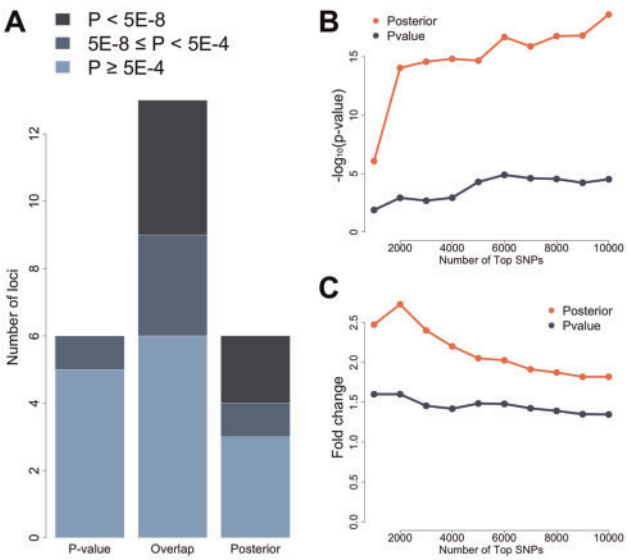


**Fig. 2.** Global performance in studies of Crohn's disease. (**A**) Signals at *P*-value-specific, overlapped, and posterior-specific loci in the IIBDGC meta-analysis. The top 20 loci based on *P*-values in the NIDDK study are compared with the top 20 loci based on posterior scores. Each locus is evaluated using the maximum regional signal strength in the IIBDGC meta-analysis. Darker color indicates stronger signals in the meta-analysis. (**B**) Enrichment of whole-blood eQTLs in the top SNPs selected based on *P*-value and posterior score. The vertical axis shows the transformed *P*-value of hypergeometric test. (**C**) Fold enrichment of whole-blood eQTLs in the top SNPs selected based on *P*-value and posterior score. The vertical axis shows the ratio of observed and expected overlaps between eQTLs and highly ranked SNPs

respectively. When the number of top SNPs increases, the posterior-based approach dominates the *P*-value-based approach in both enrichment *P*-value and fold change (Fig. 2B, C). The same enrichment pattern can be observed when using blood eQTLs from (Westra *et al.*, 2013; Supplementary Fig. S3).

In order to show how our method performs locally, we chose two genome-wide significant loci from the IIBDGC meta-analysis. First, within the risk locus on chromosome 1q23, two SNPs had substantially stronger signals than others, i.e. rs2274910 ($P_{NIDDK} = 4.40 \times 10^{-4}$) and rs955371 ($P_{NIDDK} = 4.84 \times 10^{-4}$). According to the *P*-values, these two SNPs are indistinguishable, because the signal at rs2274910 is only slightly stronger. However, the results from the meta-analysis clearly show the existence of two SNP clusters with strong signals at this locus (Fig. 3A). The cluster closer to gene CD244, in which rs955371 resides, actually has stronger signals than the cluster where rs2274910 is located. Interestingly, the posterior scores capture this difference between two SNPs very well. In fact, the posterior scores for rs955371 and rs2274910 are 0.272 and 0.208, suggesting rs955371 is more likely to be functional even though its *P*-value is larger. The second example is the risk locus on chromosome 14q35, which is one of the 12 loci with a reduced rank under the posterior scores (Supplementary Table S3). Signals at this locus were not strong in the NIDDK study, with the smallest *P*-value only at $4.70 \times 10^{-3}$ (rs1959715). Moreover, the signal peak in the NIDDK study (near 88.2M) was quite far from that in the meta-analysis, which resides in genes GALC and GPR65 (Fig. 3B). However, the posterior scores once again capture the signal pattern in the meta-analysis. Signals near 88.2M on chromosome 14 are shrunk substantially, while the SNPs in GALC and GPR65 are pushed up as the strongest signal (rs4904410). Since these SNPs have very weak signals in their *P*-values, the posterior score is still low (see Section 2). This explains

the reduced rank, because the *P*-value-based rank of rs1959715 was compared with the posterior-based rank of rs4904410. It is worth noting that the SNPs with the strongest signals in the meta-analysis, e.g. rs8005161, were either not genotyped or dropped in the quality control steps in the NIDDK study. It is reasonable to believe that the posterior scores would have had an even better performance if imputations had been done for the NIDDK dataset.

### 3.2 Application to Schizophrenia

In addition to Crohn's disease, we also applied GenoWAP to schizophrenia, a major psychiatric disorder. Psychiatric Genomics Consortium (PGC), the largest international consortium in psychiatry, focuses on genetic studies of many psychiatric disorders including schizophrenia. Two large-scale GWAS mega-analyses of schizophrenia have been published. We applied GenoWAP to the earlier and smaller PGC2011 study (Ripke *et al.*, 2011), and evaluated the performance using results from the larger mega-analysis published in 2014 (Ripke *et al.*, 2014).

PGC2014 study identified 108 schizophrenia-associated loci, from which we removed three loci on chromosome X because the PGC2011 dataset did not contain any SNP on sex chromosomes. We ranked the 1 252 637 SNPs in PGC2011 study based on their *P*-values and posterior scores, respectively. Within each locus, the rank of the lowest *P*-value was compared to the rank of the largest posterior score. Across the 105 loci, 68 (65%) had an improved rank, 1 locus (1%) had an equal rank, and the other 36 loci (34%) had a reduced rank (Supplementary Table S5). The probability of having an increased rank is significantly higher than that of having a reduced rank (*P*-value = 0.001, one-sided binomial test). Interestingly, among the 10 loci with the strongest signals in the PGC2014 study, 8 had an increased rank (80%). The proportion of increased or equal ranks gradually drops when more top loci in the PGC2014 study were considered, showing less confidence in weaker signals (Supplementary Fig. S4).

Next, we compared the top 20 loci with the smallest *P*-values to the top 20 loci with the largest posterior scores in the PGC2011 study. In order to identify 20 independent loci, 582 SNPs were needed when using *P*-value as the criterion. When posterior scores were used to choose top signals, 548 SNPs were sufficient to identify 20 loci, showing better efficiency (Fig. 4A). A total of 14 loci could be identified using both *P*-values and posterior scores. As for the comparisons between the 6 posterior-specific loci and the 6 *P*-value-specific loci, the posterior-specific loci showed better signals than the *P*-value-specific loci (Supplementary Table S6, Fig. 4B) in the PGC2014 study. Four of the 6 posterior-specific loci were genome-wide significant in the PGC2014 study, whereas 2 *P*-value-specific loci passed the genome-wide significance level. Among the 6 *P*-value-specific loci, the locus on chromosome 3q26 had the strongest signal in the PGC2014 study ($P_{2014} = 5.35 \times 10^{-11}$). This locus will be discussed in detail later.

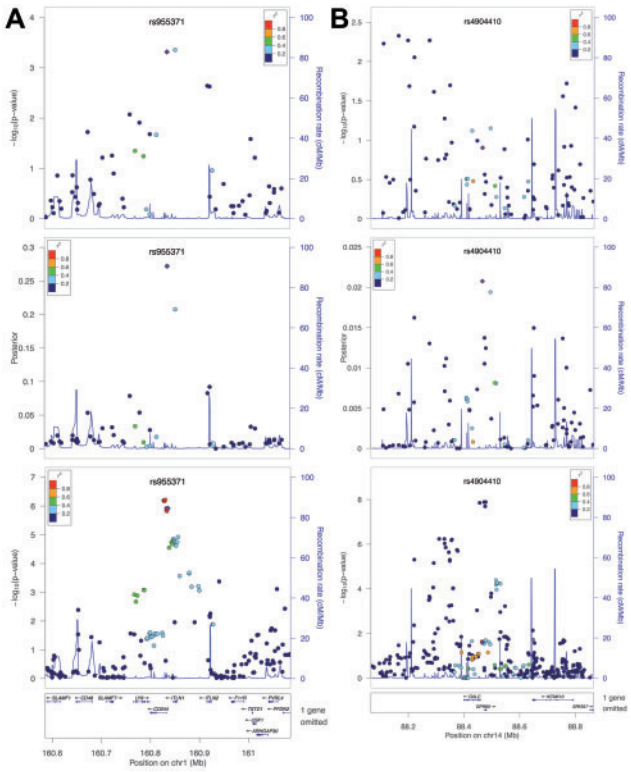


**Fig. 3.** Local performance in studies of Crohn's disease. From top to bottom, the three panels show the *P*-values from the NIDDK study, the posterior scores, and the *P*-values from the IIBDGC meta-analysis, respectively. (**A**) Local performance at the risk locus on chromosome 1q23. The top two SNPs at this locus in the NIDDK study are indistinguishable based on their *P*-values. The posterior scores suggest the importance of the SNP on the left, which is in agreement with the results from the meta-analysis. (**B**) Local performance at the risk locus on chromosome 14q35. Signals at this locus are weak in the NIDDK study, and the signal peak is different from that in the meta-analysis. The posterior score is able to reduce the noises caused by LD, and reveal real signals at genes GALC and GPR65. Figures are generated using LocusZoom (Pruim *et al.*, 2010)

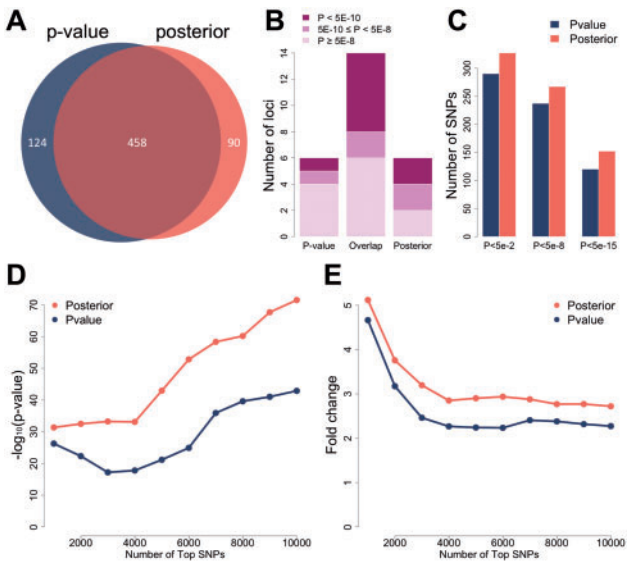


**Fig. 4.** Global performance in studies of schizophrenia. (**A**) SNPs needed for identifying 20 loci. 582 top SNPs are needed when using *P*-value as the criterion. 548 SNPs are sufficient when using posterior score as the criterion. (**B**) Signals at *P*-value-specific, overlapped, and posterior-specific loci in the PGC2014 study. The top 20 loci based on *P*-values in the PGC2011 study are compared with the top 20 loci based on posterior scores. Each locus is evaluated using the maximum regional signal strength in the PGC2014 study. Darker color indicates stronger signals in the large study. (**C**) Replication rates of SNPs before and after prioritization. The top 500 SNPs under posterior scores have substantially higher replication rates than the top 500 SNPs under *P*-values. (**D**) Enrichment of whole-blood eQTLs in the top SNPs selected based on *P*-value and posterior score. The vertical axis shows the transformed *P*-value of hypergeometric test. (**E**) Fold enrichment of whole-blood eQTLs in the top SNPs selected based on *P*-value and posterior score. The vertical axis shows the ratio of observed and expected overlaps between eQTLs and highly ranked SNPs

Since imputation was done for both PGC2011 and PGC2014 studies, and the total number of SNPs is large, it is possible to compare the SNP-level replication rates when the SNPs were ranked based on *P*-values and posterior scores. Among the top 500 SNPs with the largest posterior scores, 327, 267 and 152 had a *P*-value lower than $5 \times 10^{-2}$, $5 \times 10^{-8}$ and $5 \times 10^{-15}$ in the PGC2014 study, respectively. When choosing the top 500 SNPs based on their *P*-values, the corresponding numbers were 290, 237 and 120 (Fig. 4C), respectively. A similar pattern can be observed for the top 200 SNPs (Supplementary Fig. S5). We further performed enrichment analysis using GTEx whole-blood eQTLs. The top 1000 SNPs based on the *P*-values were significantly enriched for eQTLs (*P*-value $= 5.58 \times 10^{-27}$, fold enrichment $= 4.66$), but the enrichment for the top 1000 SNPs based on the posterior scores was even stronger (*P*-value $= 4.48 \times 10^{-32}$, fold enrichment $= 5.12$). As the number of top SNPs increased, the posterior-based top SNPs always had stronger enrichment of eQTL than the *P*-value-based list (Fig 4D, E). Similar results can be observed when using another set of blood eQTLs (Westra *et al.*, 2013). The enrichment results for a set of quantitative trait loci in human brain (Gibbs *et al.*, 2010) also favor posterior scores as the number of top SNPs increase (Supplementary Fig. S3).

Finally, we compared PGC2011 *P*-values, PGC2011 posterior scores, and PGC2014 *P*-values at two loci to further illustrate the performance of our method. The first locus is on chromosome 3q26. It had the strongest signal in PGC2014 among the *P*-value-specific top 20 loci (Supplementary Table S6, $P_{2014} = 5.35 \times 10^{-11}$). Based on the *P*-values in the PGC2011 study, the strongest signals reside in the intergenic region upstream of FXR1. But the posterior scores brought down those intergenic SNPs, and enhanced the signals in FXR1 instead, which is in agreement with the results from PGC2014 (Fig. 5A). In fact, from the PGC2014 *P*-values, we can clearly see that the strongest signals reside in FXR1 while the significant results for the SNPs upstream or downstream of FXR1 are most likely due to LD. The second example is on chromosome 8q21 (Fig. 5B). In the PGC2011 study, the strongest signal at this locus resides in the intergenic region between 89.7M and 89.8M. However, posterior scores removed most of the correlated SNPs at this locus, leaving three separate peaks as candidate functional spots. The first peak lies right upstream of MMP16. The second peak is more upstream (~89.6M), and is suggested to be the strongest signal source. The SNPs with the lowest *P*-values in PGC2011 remained as a signal peak, but their posterior scores were not as strong as the peak in the middle. Most interestingly, the results from the posterior scores perfectly matched the signal patterns in the PGC2014 study. From the lowest panel in Figure 5B, we can clearly see two separate peaks at the same locations suggested by the posterior scores, with the one near 89.6M being the strongest signal source. Also, the SNPs between 89.7M and 89.8M had weaker signals than the peak in the middle. Notably, this entire risk locus resides in an intergenic region. This example shows that our method can effectively prioritize SNPs in the non-coding genome.

### 3.3 Several remarks on gene centricity

We annotated all the SNPs in the NIDDK GWAS for Crohn's disease using the RefGene database. Among the 298 321 SNPs in this dataset, 158 028 (53.0%) are intergenic. Among the top 1000 and the top 2000 SNPs with small *P*-values, the proportion of intergenic SNPs is relatively stable (549 out of 1000, 54.9%; and 1064 out of 2000, 53.2%). However, the proportion of intergenic SNPs substantially decreased among the top SNPs with higher posterior scores

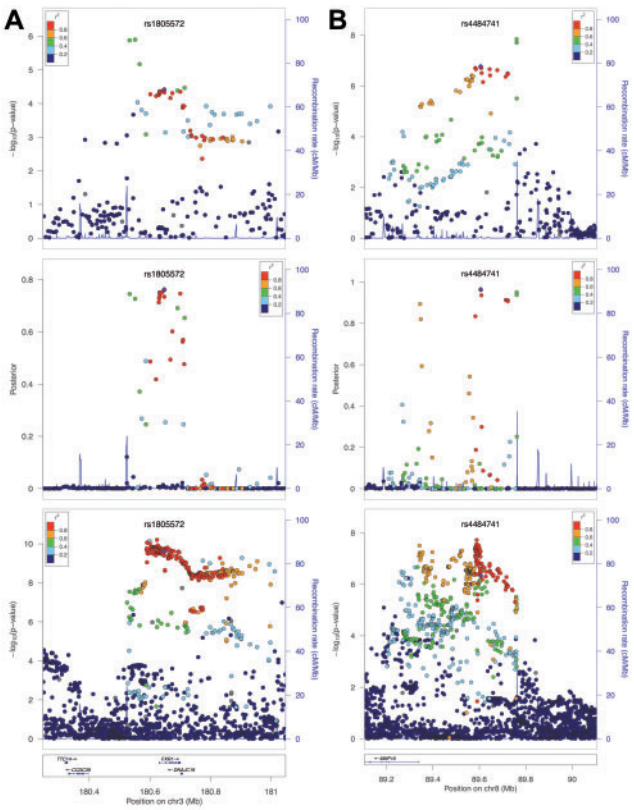


**Fig. 5.** Local performance in studies of schizophrenia. From top to bottom, the three panels show the *P*-values from the PGC2011 study, the posterior scores, and the *P*-values from the PGC2014 study, respectively. (**A**) Local performance at the risk locus on chromosome 3q26. The top signals at this locus in the PGC2011 study reside upstream of gene FXR1, while the posterior scores pull down those signals and suggest the importance of SNPs in FXR1. This agrees with the signal pattern in the PGC2014 study. (**B**) Local performance at the risk locus on chromosome 8q21. Posterior scores diminish most of the correlated SNPs at this locus, leaving three separate signal peaks. The peak near 89.6M is suggested to be the strongest signal source, which cannot be seen using *P*-values from the PGC2011 study. The signal peaks suggested by posterior scores perfectly match the strongest signals in the PGC2014 study. Figures are generated using LocusZoom (Pruim *et al.*, 2010)

(403 out of top 1000, 40.3%; and 769 out of top 2000, 38.5%). We repeated the analysis on PGC2011 schizophrenia GWAS. A similar pattern could be observed (Supplementary Fig. S6). These results show that GenoWAP does favor protein-coding regions. However, a large proportion of non-coding signals still remain after prioritization, indicating the necessity of considering non-coding and even intergenic regions.

## 4 Discussion

In this study, we developed and applied GenoWAP to two sets of GWAS data to illustrate its performance in GWAS signal prioritization. Compared to *P*-values, GenoWAP posterior scores can better prioritize SNPs in many different ways. At the locus level, posterior score is more efficient in the sense that fewer SNPs are needed to identify the same number of top loci. Moreover, noises due to chance are effectively reduced, and the highly ranked loci using posterior scores may be more likely to contain functional elements than the top loci selected purely based on *P*-values. At the SNP level, markers with high posterior scores have both better replication rates

and consistently stronger enrichment of eQTLs than the top SNPs based on *P*-values. More importantly, within each risk locus identified in GWAS, posterior scores can effectively suggest potential functional sites among a large number of correlated SNPs.

The performance of GenoWAP depends on the accuracy of functional annotation. Due to our limited understanding of non-coding genome, it is challenging to provide accurate genomic functional annotation. GenoCanyon is a convenient tool that provides functional prediction at the nucleotide level, yet its predictive ability can still be improved. Large consortia such as ENCODE (Bernstein *et al.*, 2012) and Roadmap project (Kundaje *et al.*, 2015) are continuously generating diverse types of epigenetic annotation data from a variety of cell types. The performance of GenoWAP may be further enhanced when these data become available in the future. In our implemented software, we allow users to use their own annotation file. GenoWAP also depends on the quality of GWAS data. If no information is contained in the GWAS dataset, then GenoWAP can only provide limited insight. Finally, we emphasize that GenoWAP's ability to identify precise functional factors is limited. GenoWAP is a region-based tool, and is powerful in identifying regions that are more likely to have a functional impact within LD blocks. However, definitive proof for functionality of any given SNP still requires thorough allele-specific experimentation.

More than 2000 GWASs have been published in the past decade, and the number continues to grow. It is well known that our ability to identify new risk loci for complex diseases has surpassed our ability to interpret the results. However, although we are overwhelmed by the large amount of information detected in GWASs, evidence such as missing heritability still suggests that many risk loci remain to be discovered. Therefore, there is pressing need for GWAS signal prioritization tools, and our method has great potential for future application. Since GenoWAP uses only *P*-values as the input, it is convenient to apply our method on published results, which may help reveal potential functional sites within large haplotype blocks, and ultimately help understand disease etiology. Moreover, for multi-stage GWASs, GenoWAP can be used to better prioritize SNPs from the discovery stage to the validation planning and increase the replication rates. Finally, next-generation sequencing is widely recognized as the future of genomic epidemiology. However, the high cost of sequencing usually leads to insufficient sample sizes and many other challenging issues (Sboner *et al.*, 2011). The combination of GenoWAP and the rich collection of publicly available GWAS data has the potential to provide functional candidates and guide sequencing analysis in the future.

## Acknowledgements

We thank Dr. Katerina Kechris and all the members in the Data Integration – COPD working group at SAMSI for their advice and useful discussions on this work.

## Funding

This study was supported by the National Institutes of Health grants R01 GM59507 and U01 HG005718, the VA Cooperative Studies Program of the Department of Veterans Affairs, Office of Research and Development, and the Yale World Scholars Program sponsored by the China Scholarship Council.

*Conflict of Interest*: none declared.

## References

Ardlie,K.G. *et al.* (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.

Bernstein,B.E. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*,**489**, 57–74.

Chung,D. *et al.* (2014) GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLoS Genet.*, **10**, e1004787.

Cooper,G.M. and Shendure,J. (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.*, **12**, 628–640.

Efron,B. (2010) *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, New York.

Eicher,J.D. *et al.* (2015) GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes. *Nucleic Acids Res.*, **43**, D799–D804.

Eichler,E.E. *et al.* (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, **11**, 446–450.

Franke,A. *et al.* (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.*, **42**, 1118–1125.

Fu,Y. *et al.* (2014) FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.*, **15**, 480.

Gibbs,J.R. *et al.* (2010) Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.*, **6**, e1000952.

Hindorff,L.A. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA*, **106**, 9362–9367.

Hou,L. and Zhao,H. (2013) A review of post-GWAS prioritization approaches. *Front. Genet.*, **4**, 280.

Kircher,M. *et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.

Kundaje,A. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.

Lu,Q. *et al.* (2015) A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci. Rep.*, **5**, 10576.

Manolio,T.A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

Pruim,R.J. *et al.* (2010) LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, **26**, 2336–2337.

Rioux,J.D. *et al.* (2007) Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat. Genet.*, **39**, 596–604.

Ripke,S. *et al.* (2011) Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.*, **43**, 969–976.

Ripke,S. *et al.* (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**, 421–427.

Ritchie,G.R. *et al.* (2014) Functional annotation of noncoding sequence variants. *Nat. Methods*, **11**, 294–296.

Sboner,A. *et al.* (2011) The real cost of sequencing: higher than you think!. *Genome Biol.*, **12**, 125.

Shihab,H.A. *et al.* (2015) An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, **31**, 1536–1543.

Visscher,P.M. *et al.* (2012) Five years of GWAS discovery. *Am. J. Hum. Genet.*, **90**, 7–24.

Wall,J.D. and Pritchard,J.K. (2003) Haplotype blocks and linkage disequilibrium in the human genome. *Nature Rev. Genet.*, **4**, 587–597.

Wang,Q. *et al.* (2015) A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. *Front. Genet.*, **6**, 149.

Ward,L.D. and Kellis,M. (2012) Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.*, **30**, 1095–1106.

Westra,H.J. *et al.* (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.*, **45**, 1238–1243.

Witte,J.S. *et al.* (2014)The contribution of genetic variants to disease depends on the ruler. *Nat. Rev. Genet.*, **15**, 765–776.