Post-GWAS Prioritization through Integrated Analysis of Genomic Functional Annotation

Qiongshi Lu[1], Xinwei Yao[2], Yiming Hu[1], Hongyu Zhao[1]

[1]Department of Biostatistics, Yale School of Public Health, New Haven, CT; [2]Yale College, New Haven, CT

Introduction

Genome-wide association study (GWAS) has been a great success in the past decade, with tens of thousands of loci identified associated with many complex diseases. However, challenges still remain in both identifying new risk loci and interpreting results. Bonferroni-corrected significance level is very conservative for large-scale hypothesis testing, leading to insufficient statistical power when the effect size is moderate at each risk locus. Complex dependence structure among markers, known as linkage disequilibrium, also makes it challenging to distinguish causal variants from large haplotype blocks. We propose GenoWAP (Genome Wide Association Prioritizer), a post-GWAS prioritization method that integrates genomic functional annotation and GWAS test statistics.

Question 1: How to identify functional regions in the human genome?

We developed GenoCanyon [1], a statistical framework to predict functional non-coding regions in the human genome. GenoCanyon is based on unsupervised learning. 22 diverse types of annotations (2 conservation measures, 2 open-chromatin indicators, 8 histone modifications, and 10 transcription factors) downloaded from ENCODE [2] were integrated to infer the functional potential of each of 3 billion nucleotides in the human genome. GenoCanyon was found to have exceptional performance when it was evaluated using well-studied non-coding regulatory regions [1].

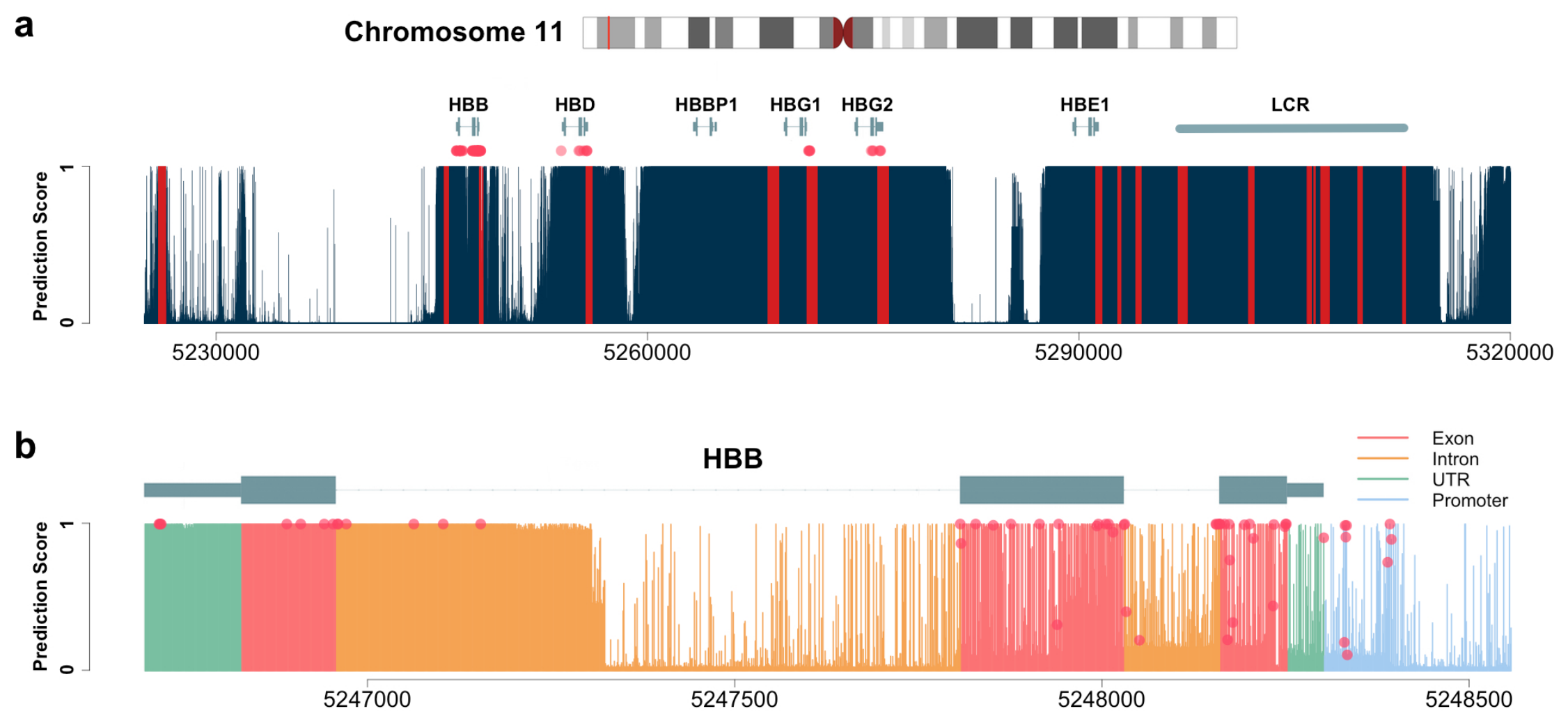


Figure 1. Functional prediction for the HBB gene complex. (a) Dark blue bars indicate the prediction scores. All the 23 known cis-regulatory modules are marked in red [3]. Red dots indicate the locations of known pathogenic variants in this region. (b) Prediction results for the HBB gene and its promoter. The promoter, UTRs, introns and exons are marked with different colors. Red dots show the prediction scores of known pathogenic variants.

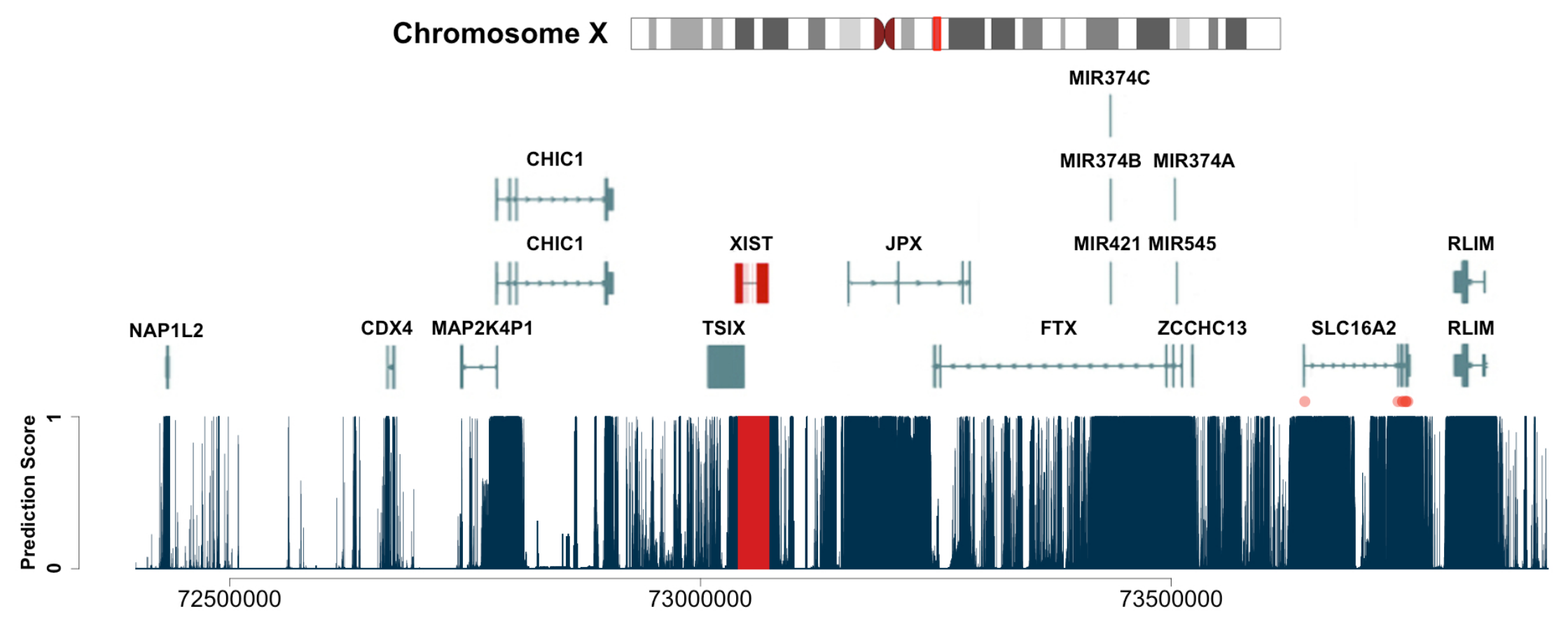


Figure 2. Prediction results for the human X-inactivation Center. All the RefSeq transcripts in this region are plotted. The master lncRNA XIST is highlighted in red. Red dots show the locations of known pathogenic variants.

Question 2: How to use functional annotation to help prioritize SNPs in GWAS?

For each SNP, we introduce the following notations.

Z: *indicator of general functionality;*
Z_D: *indicator of disease-specific functionality;*
p: *p-value obtained in GWAS.*

We use the following posterior probability to prioritize SNPs.

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

Based on the definitions of **Z** and **Z_D**, we transform $P(Z_D=1)$ into the following form.

$$P(Z_D = 1) = P(Z = 1, Z_D = 1) = P(Z_D = 1|Z = 1)\times P(Z = 1)$$

SNPs are divided into functional and non-functional subgroups according to their GenoCanyon functional scores. Then, f(p|Z=0) can be estimated empirically. We further assume $f(p|Z_D=0) = f(p|Z=0)$, and $(p|Z_D=1)$ follows a beta distribution.

$$(p|Z_D = 1) \sim Beta(\alpha, 1), \qquad 0 < \alpha < 1$$

Finally, all the unknown parameters in the posterior probability can be estimated using the EM algorithm.

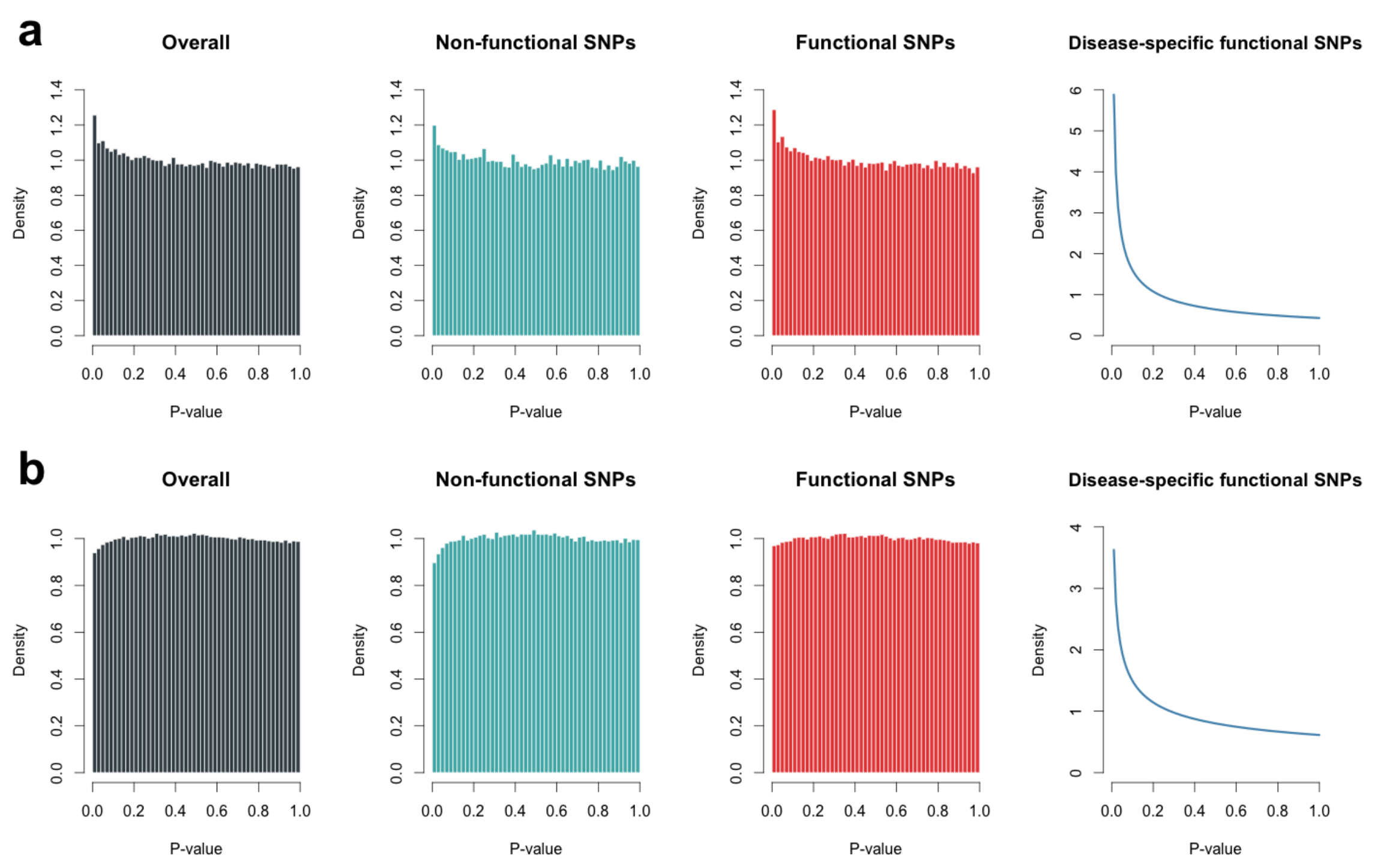


Figure 3. P-value distributions of different functional groups. (a) NIDDK GWAS of Crohn's Disease [4]. (b) COPDGene GWAS of Chronic Obstructive Pulmonary Disease [5] (non-hispanic white population). From these distributions, we can see that it is crucial to use the empirically estimated null distribution instead of the uniform null. The uniform null not only overestimates the difference between functional and non-functional groups when the signal is strong, it also fails to detect the signal in some studies.

Application to Crohn's disease

We applied GenoWAP on a smaller GWAS conducted by the North American National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK, n=1,963, [4]) IBD Genetics Consortium, and tested the results using the largest meta-analysis for Crohn's disease done by the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC, n=21,389, [6]). The top loci based on posterior scores show substantially stronger signals in the IIBDGC meta-analysis. Top SNPs based on posterior scores also show much stronger enrichment of eQTL.

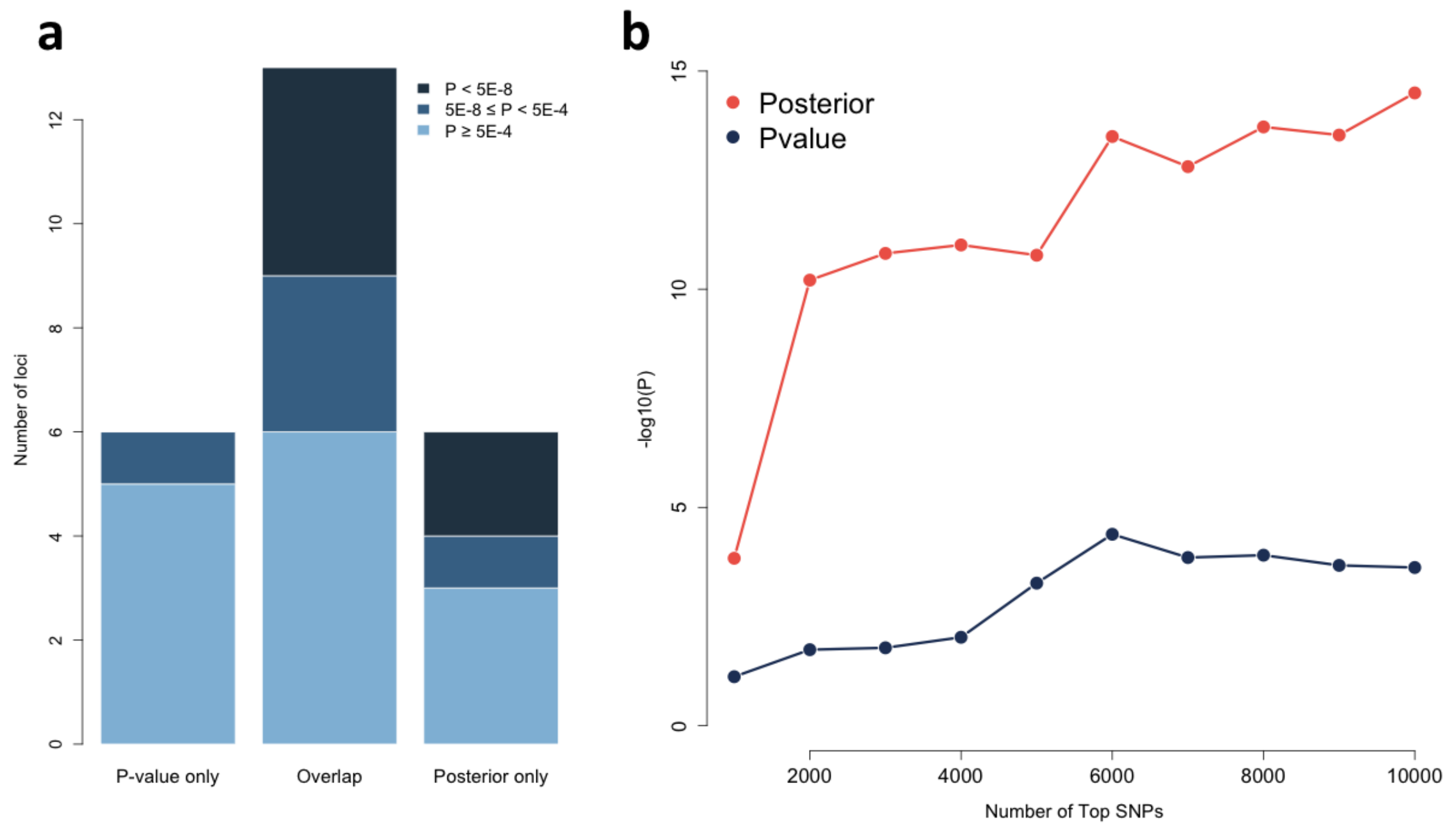


Figure 4. Global performance in studies of Crohn's disease. (a) Darker color indicates stronger signals in the meta-analysis. (b) Enrichment of GTEx whole-blood eQTLs in the top SNPs.

Application to Schizophrenia

We applied GenoWAP to the PGC2011 study (n=21,856, [7]), and evaluated the performance using the 2014 study (n=79,845, [8]).

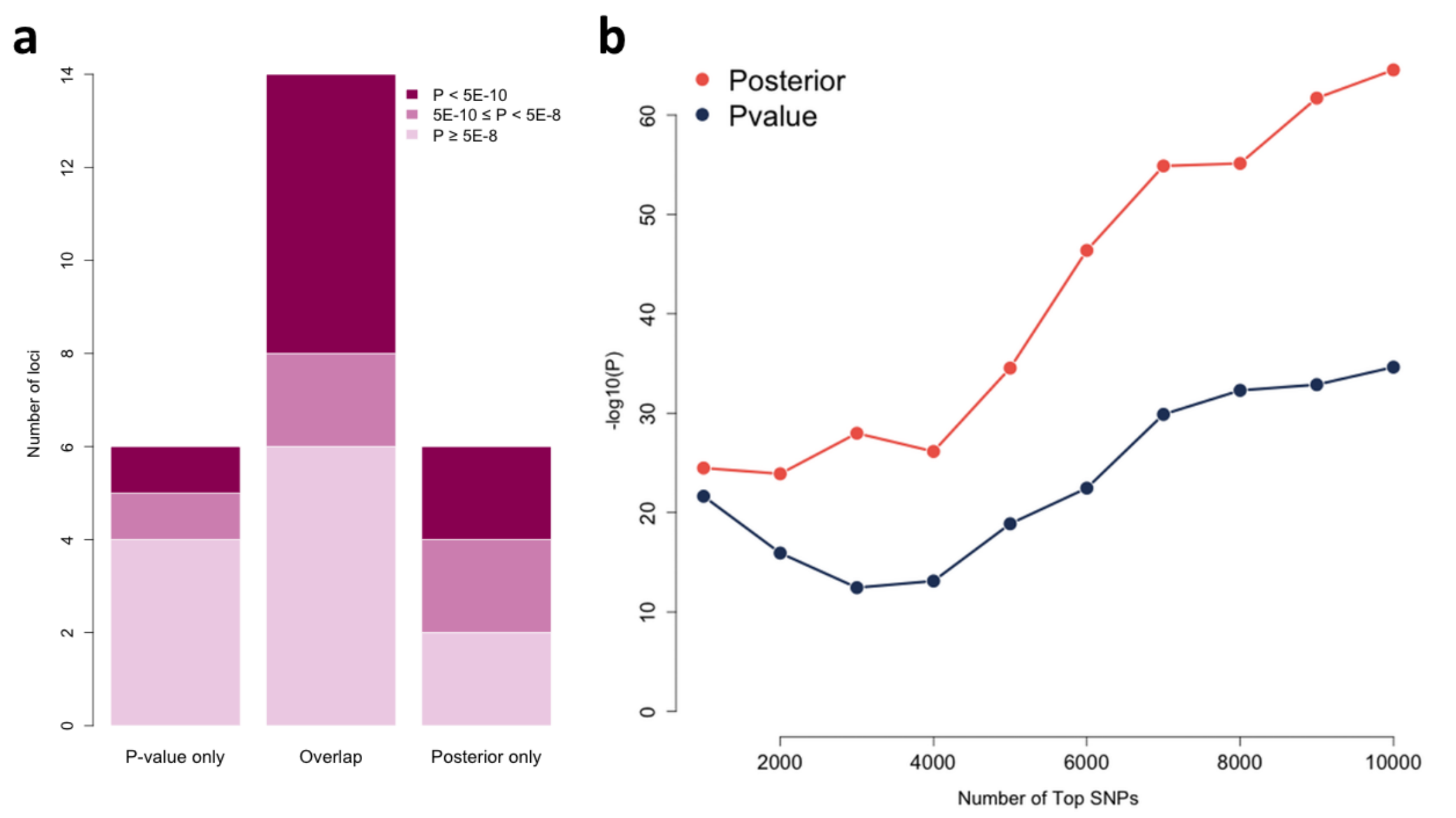


Figure 5. Global performance in studies of Schizophrenia. (a) Darker color indicates stronger signals in the PGC2014 study. (b) Enrichment of GTEx whole-blood eQTLs in the top SNPs.

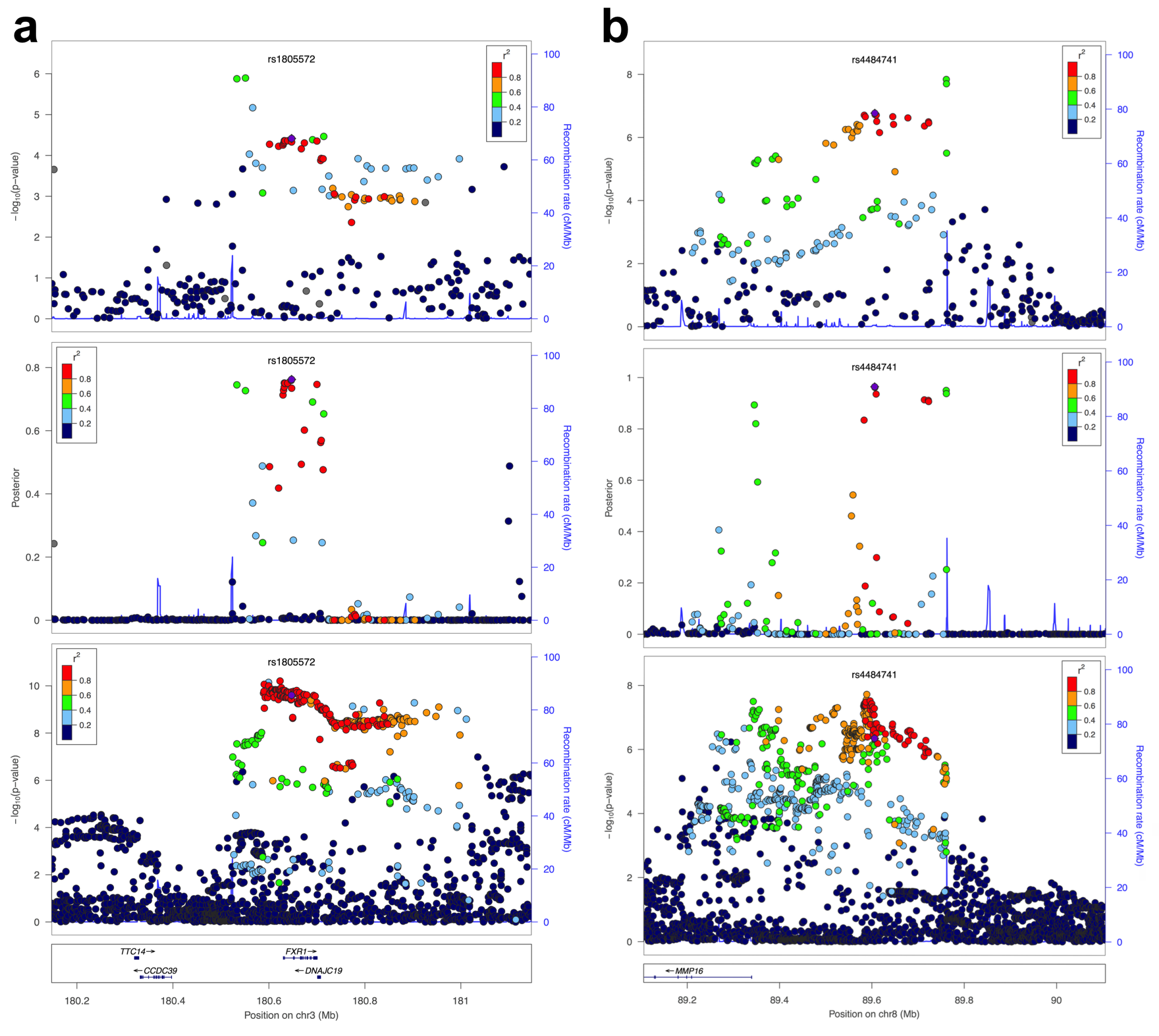
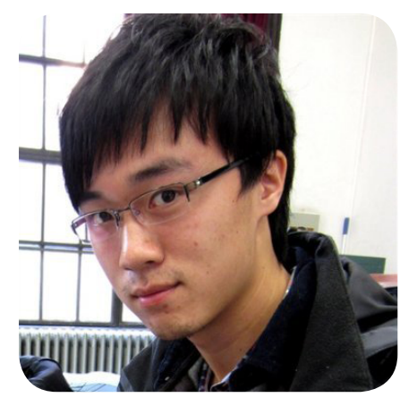


Figure 6. Local performance in studies of Schizophrenia. (a) A risk locus on chromosome 3q26. (b) A risk locus on chromosome 8q21.

Conclusion

Our prioritization method is much more powerful than the traditional approach solely based on p-values. Within each risk locus, GenoWAP is able to distinguish real signals from groups of correlated SNPs. It has the potential to be widely used to reveal functional variants at disease-associated risk loci and guide future studies such as resequencing and functional analysis as well as development of treatments.

Developers



Qiongshi Lu is a doctoral student in Biostatistics at Yale University. His research interest is in predicting disease-specific and tissue-specific functional non-coding regions in the human genome. He is also interested in the application of next generation sequencing and statistical graphical models in genomic epidemiology.

Xinwei (David) Yao is an undergraduate majoring in Intensive Mathematics at Yale College. He is interested in combining his knowledge in math and computer science in interesting applications and is passionate about technology and software development.

Yiming Hu is a doctoral student in Biostatistics at Yale University. His research interest is in developing statistical and computational method in genetics. Specifically, he is interested in developing similarity measure for clustering using gene expression data, and studying tumor heterogeneity using DNA sequencing and single-cell RNA-seq.

Hongyu Zhao is Ira V. Hiscock Professor of Public Health (Biostatistics) and Professor of Genetics and of Statistics at Yale University.

Email: hongyu.zhao@yale.edu

References

[1] Lu et al. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. Scientific Reports, in press.
[2] ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414), 57-74.
[3] King et al. (2005). Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome research, 15(8), 1051-1060.
[4] Rioux et al. (2007). Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nature genetics, 39(5), 596-604.
[5] Cho et al. (2014). Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. The Lancet Respiratory Medicine, 2(3), 214-225.
[6] Franke et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nature genetics, 42(12), 1118-1125.
[7] Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. (2011). Genome-wide association study identifies five new schizophrenia loci. Nature genetics, 43(10), 969-976.
[8] Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature, 511(7510), 421-427.

Web Servers

GenoCanyon
Non-coding functional annotation
http://genocanyon.med.yale.edu

GenoWAP
Genome-wide association prioritizer
http://genocanyon.med.yale.edu/GenoWAP