Integrative Tissue-specific Functional Annotations in the Human Genome Provide Novel Insights on Complex Traits

Qiongshi Lu[1,*], Ryan Powles[2,*], Qian Wang[2], Julie He[3], Hongyu Zhao[1,2]

[1]Department of Biostatistics, Yale School of Public Health; [2]Program of Computational Biology and Bioinformatics, Yale University; [3]Division of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine; *Equal Contribution

Introduction

Genome-wide association study (GWAS) has been a productive approach to study human complex diseases, yet challenges still remain in identifying and interpreting risk loci. In this work, we introduce GenoSkyline, a statistical framework to predict tissue-specific functional regions in the human genome, and illustrate a variety of ways that GenoSkyline could benefit post-GWAS analysis. GWAS signals can be better prioritized when integrating annotations of disease-related tissue types. Combining GenoSkyline with GWAS results also allow us to partition heritability by tissue types and generate new hypotheses regarding the disease etiology of many complex diseases. We believe that GenoSkyline can guide genetics research and greatly benefit the broader scientific community.

Identify tissue-specific functional regions in the human genome

We developed GenoCanyon [1], an unsupervised-learning framework to predict functional non-coding regions in the human genome. In this work, we extend the framework using epigenomic data from Roadmap Epigenomics Project [2] to infer tissue-specific functionality [3]. Eight epigenomic marks were integrated to quantify the functional potential of each nucleotide. GenoSkyline was found to have exceptional performance when it was evaluated using well-studied non-coding functional regions. GenoSkyline scores for seven tissue types (brain, GI, lung, heart, Blood, muscle, epithelium) are currently available.

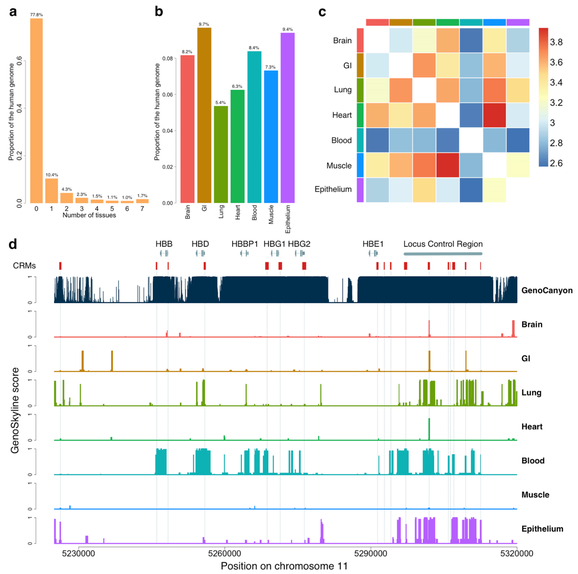

Figure 1. GenoSkyline annotations. (a) Number of tissues in which nucleotides are functional. (b) Proportion of functional genome for each tissue type. (c) Overlap of functional regions across seven tissue types. The scale is log odds ratio. (d) Comparison of GenoCanyon prediction and GenoSkyline scores for seven tissues in HBB gene complex region. Red boxes mark the locations of known cis-regulatory modules (CRM).

Partition heritability by tissue types

Next, we focus on how GenoSkyline could help us understand human complex traits. We applied LD score regression [4] on 15 human complex diseases and traits, and identified tissue types enriched for GWAS signals. Stratified LD scores were estimated using LDSC software and GenoSkyline annotations. For each tissue type, partitioned heritability was estimated, and signal enrichment was then calculated as follows.

$$Enrichment = \frac{\%\ Heritability\ explained}{\%\ Genome\ covered}$$

Standard errors of annotation-stratified heritability estimates were assessed using a resampling-based approach [4]. Some p-values for tissue-specific enrichment are shown in Figure 2.

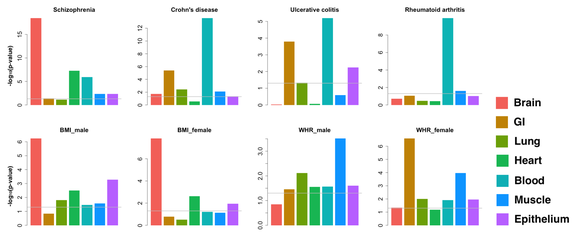

Figure 2. Tissue-specific enrichment of GWAS signals. Enrichment p-values were calculated using LD score regression. The grey line is the 0.05 cutoff for p-value.

Prioritize GWAS signals through integrating summary data and annotations

We developed GenoWAP (Genome-Wide Association Prioritizer), a GWAS signal prioritization method based on integrated analysis of GWAS summary statistics and GenoCanyon annotation [5]. In this work, we further extend our method to make it compatible with tissue-specific annotations. For each SNP, we introduce the following notations.

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

We use the following posterior probability to re-prioritize SNPs.

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

SNPs can be devided into two categories, i.e. (Z_T=1) and (Z_T=0), based on their GenoSkyline scores. Then, f(p|Z_T=1) can be written as the following mixture.

$$f(p|Z_T=1)=P(Z_D=1|Z_T=1)\times f(p|Z_D=1,Z_T=1)+P(Z_D=0|Z_T=1)\times f(p|Z_D=0,Z_T=1)$$

We further assume f(p|Z_D=0, Z_T=1) = f(p|Z_D=0) = f(p|Z=0), and (p|Z_D=1, Z_T=1) follows a beta distribution.

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

The first assumption essentially assumes that p-values of SNPs irrelevant to the disease but in the functional region should behave similar to p-values of non-functional SNPs.

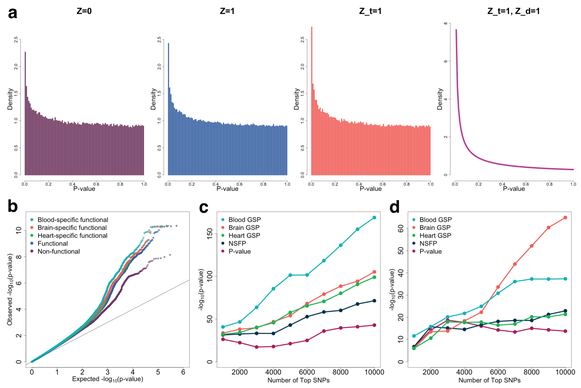

Figure 3. Reprioritization of schizophrenia GWAS signals. (a) Histograms for p-value distributions. (b) Tissue-specific functional regions are more enriched for schizophrenia associations than generally functional regions and non-functional regions. (c) Enrichment of GTEx whole-blood eQTLs in top SNPs. (d) Enrichment of human brain quantitative trait loci in top SNPs.

The beta assumption has also been justified through extensive simulations [6]. Finally, all the remaining parameters in the posterior probability formula can be estimated either directly or using the EM algorithm.

Understand each disease-associated locus

GenoWAP also predicts the most relevant tissue type for each risk locus, which is illustrated in Figure 4. This could help us understand disease etiology at the locus level.

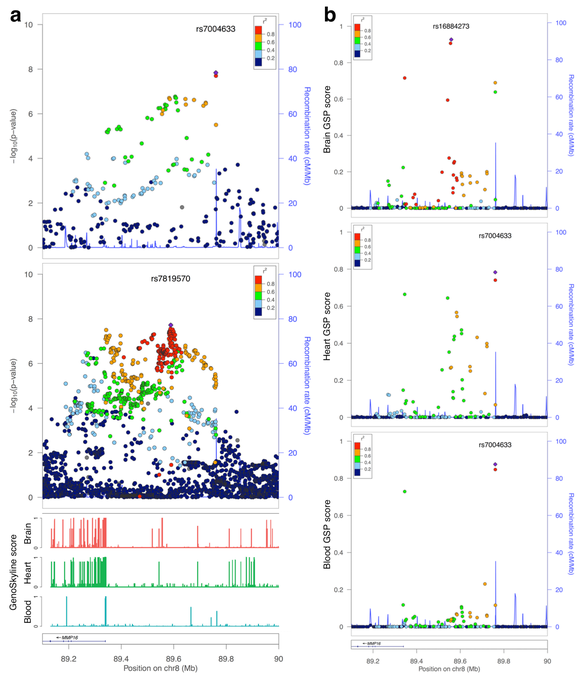

Figure 4. Local performance of signal prioritization. (a) Results at a schizophrenia risk locus on chromosome 8q21 near MMP16 gene. The top and middle panel show p-values from PGC 2011 and 2014 studies [7,8], respectively. The bottom panel shows GenoSkyline annotations at this locus. (b) Locus plots for tissue-specific posterior scores.

Conclusion

Through integrating GenoSkyline annotations with GWAS summary statistics, we illustrated a variety of ways that GenoSkyline could help researchers understand human complex diseases. As epigenomic annotation data become available for an increasing number of cell types in the future, GenoSkyline's ability to facilitate complex disease studies will be further enhanced.

Developers

Qiongshi Lu is a doctoral student in Biostatistics at Yale School of Public Health. His research focuses on genomic functional annotations and their applications in human genetics. He is interested in developing statistical methods to leverage functional annotations in GWAS signal prioritization, variant fine-mapping, and genetic risk prediction.

Ryan Powles is a doctoral student in CBB Program at Yale University. He is interested in the use of statistical methods to effectively characterize genetic variation through functional genomics data. He hopes to apply these techniques in a variety of contexts across the non-coding regions of the human genome.

Qian Wang is a doctoral student in CBB Program at Yale University. Her research interest is in post-GWAS analysis, especially in exploring the interaction effects of genetic and environmental factors on diseases, as well as in studying the shared genetic factors of multiple traits. She is also interested in various applications of NGS.

Julie He is a clinical fellow in Cardiovascular Medicine at the Yale University School of Medicine. She has research interests in studying genetic predisposition to cardiac diseases.

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. (2015). A statistical framework to predict functional non-Coding regions in the human genome through integrated analysis of annotation data. *Sci. Rep.*, 5, 10576.
[2] Kundaje et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature,* 518: 317–330.
[3] Lu et al. (2016). Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. *PLOS Genetics*, 12(4): e1005947.
[4] Finucane et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics,* 47(11), 1228-1235.
[5] Lu et al. (2016). GenoWAP: GWAS signal prioritization through integrated analysis of genomic functional annotation. *Bioinformatics*, 32(4), 542-548.
[6] Chung et al. (2014). GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLoS Genetics*, 10(11): e1004787.
[7] Ripke et al. (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature genetics*, 43(10), 969-976.
[8] Ripke et al. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510), 421-427.

Web Servers

GenoCanyon **GenoWAP** **GenoSkyline**

Stay tuned!
@QiongshiLu