Integrative Analysis of GWAS Summary Data and Functional Annotations Identifies Additional Loci for Late-onset Alzheimer's Disease

Qiongshi Lu[1], Shubhabrata Mukherjee[2], Brian Kunkle[3], Paul K. Crane[2], Hongyu Zhao[1], for the Alzheimer's Disease Genetics Consortium

[1]Department of Biostatistics, Yale School of Public Health; [2]Division of General Internal Medicine, Department of Medicine, University of Washington; [3]Hussman Institute for Human Genomics, University of Miami

Introduction

Genome-wide association studies (GWAS) have identified ~20 loci associated with late-onset Alzheimer's Disease (LOAD). However, only a modest proportion of the phenotypic variance could be explained by significant SNPs, suggesting the existence of many more LOAD-associated loci with small to moderate effect sizes. Here we apply state-of-the-art methods to reprioritize LOAD GWAS signals through integrative analysis of functional annotations and publicly available IGAP stage-I GWAS summary data [1].

Identify tissue and cell type-specific functional regions in the human genome

We developed GenoSkyline [2], a statistical framework to predict tissue and cell type-specific functional regions in the human genome. RNA-seq, DNA methylation, and histone marks from Roadmap Epigenomics Project [3] were integrated to quantify the functional potential of each nucleotide. GenoSkyline scores for seven tissue types (brain, GI, lung, heart, blood, muscle, epithelium) are readily available. We are currently wrapping up GenoSkyline2 annotations for 127 various tissue and cell types.

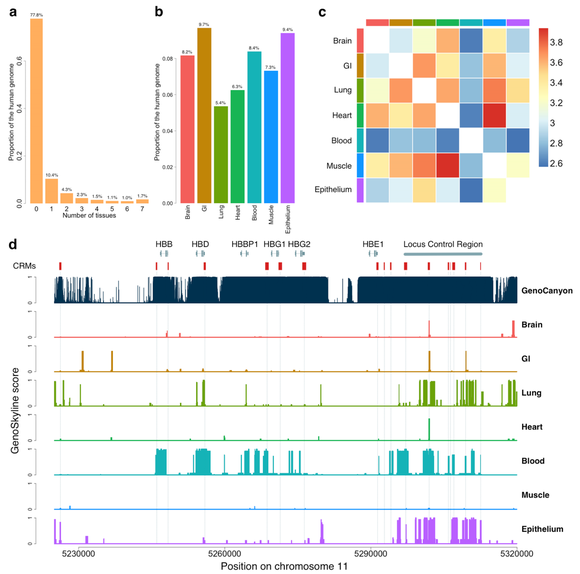

Figure 1. GenoSkyline annotations. (a) Number of tissues in which nucleotides are functional. (b) Proportion of functional genome for each tissue type. (c) Overlap of functional regions across seven tissue types. The scale is log odds ratio. (d) Comparison of GenoCanyon prediction and GenoSkyline scores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

Partition heritability by tissue and cell type

We applied LD score regression [4] on LOAD GWAS, and identified tissue and cell types enriched for GWAS signals. Signal enrichment was calculated as follows.

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

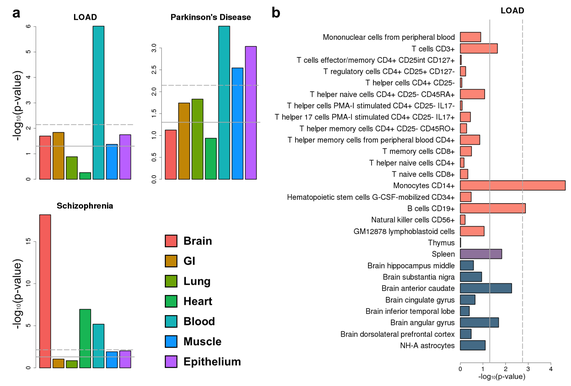

Figure 2. Tissue and cell type-specific enrichment of GWAS signals. (a) Enrichment results based on GenoSkyline annotations. For comparison, results for Parkinson's disease and schizophrenia are also shown. (b) Enrichment results based on 28 immune and brain cell types in GenoSkyline2. Enrichment p-values were calculated using LD score regression. The grey lines mark the p-value cutoffs of 0.05 and Bonferroni correction.

Prioritize GWAS signals through integrating summary data and annotations

We have developed GenoWAP (Genome-Wide Association Prioritizer), a GWAS signal prioritization method based on integrated analysis of GWAS summary statistics and functional annotations [5]. For each SNP, we introduce the following notation.

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

We use the following quantity 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 says that p-values of SNPs in the functional region but irrelevant to the disease should behave similarly to p-values of non-functional SNPs.

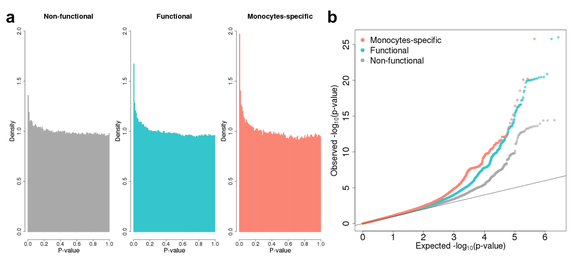

Figure 3. Comparing LOAD GWAS signals in different annotation categories. (a) Histograms for p-value distributions. (b) Monocytes-specific functional regions are more enriched for LOAD associations than generally functional regions and non-functional regions.

The beta assumption has been justified through extensive simulations [6]. Finally, all the remaining parameters can be estimated using the EM algorithm.

Identify novel risk loci for LOAD

We identified 11 novel risk loci for LOAD using GenoWAP.

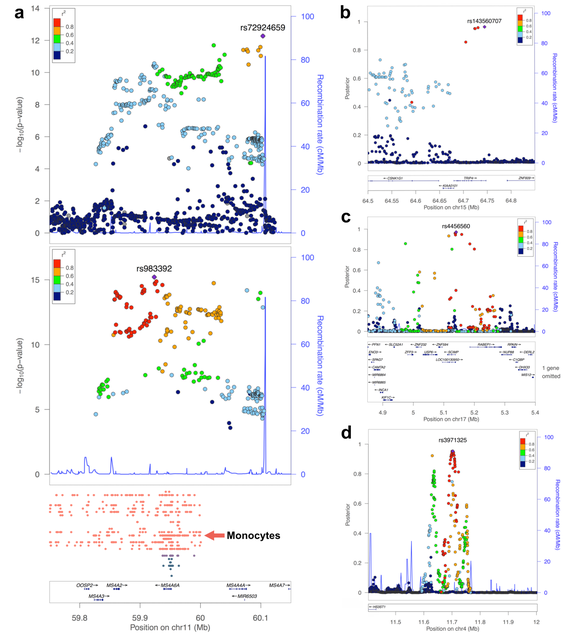

Figure 4. Local performance of signal prioritization. (a) Annotations could suggest functional SNPs within LD blocks. (b-d) Three novel loci successfully replicated in IGAP stage-II data. Chr15, *TRIP4*; Chr17, *SCIMP*; Chr4, *HS3ST1*.

Chr	Gene	SNP	A1	A2	Posterior	Beta_IGAP1	P_IGAP1	Beta_IGAP2	P_IGAP2	P_ADGC
3	*RPN1*	rs62273237	T	C	0.955	0.075	2.77E-06	0.019	0.412	0.101
4	*HS3ST1*	rs6848440	G	A	0.952	0.079	4.04E-06	0.050	**0.045**	0.818
5	*HBEGF*	rs2878896	G	A	0.997	0.083	9.22E-08	-0.003	0.913	0.610
10	*USP6NL*	rs12358692	C	T	0.964	-0.080	2.34E-06	-0.029	0.226	**0.031**
14	*RPS6KL1*	rs76378521	T	C	0.988	0.167	4.47E-07	0.057	0.212	0.095
15	*TRIP4*	rs74615166	C	T	0.950	0.336	1.97E-06	0.169	**0.017**	0.681
15	*EFTUD1*	rs905450	A	G	0.962	-0.087	2.82E-06	-0.028	0.310	0.808
17	*MINK1*	rs8078173	C	T	0.958	0.128	2.87E-06	0.065	0.089	0.120
17	*BZRAP1*	rs2632516	C	G	0.983	-0.078	9.52E-07	-0.036	0.117	0.060
15	*FAM96A*	rs77171973	C	T	0.955	0.182	9.55E-06	0.007	0.889	0.947
17	*SCIMP*	rs4456560	T	G	0.971	0.107	5.20E-06	0.077	**0.022**	0.130

Table 1. Eleven novel loci identified through integrative analysis. Loci with posterior scores greater than 0.95 are listed. 9 loci were identified using GenoCanyon non-tissue-specific annotation. Two additional loci were identified using GenoSkyline2-monocytes annotation.

IGAP1: IGAP stage-I data
IGAP2: IGAP stage-II data
ADGC: ADGC phase-II data

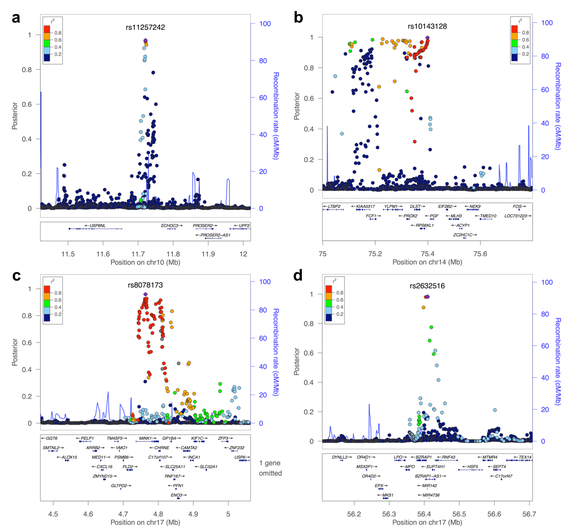

Figure 5. Four additional loci with suggestively significant signals in both IGAP stage-II and ADGC phase-II data. (a) Chr10, *USP6NL*; (b) Chr14, *RPS6KL1*; (c) Chr17, *MINK1*; Chr17, *BZRAP1*.

Conclusion

Integrative analysis using functional annotations increased statistical power and identified 11 novel loci associated with LOAD. Several of these novel genes have known functions directly or indirectly related to LOAD etiology. The validity of these loci requires further replication using independent and larger LOAD cohorts. Nevertheless, these results provide novel insights into LOAD etiology.

About me

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.

References

[1] Lambert et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature genetics*, 45(12), 1452-1458.
[2] 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.
[3] Kundaje et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature,* 518: 317–330.
[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.

Web Servers

GenoCanyon **GenoWAP** **GenoSkyline**

Stay tuned!
@QiongshiLu