The regulatory landscape of genetic variants associated with psychiatric disorders and neurodegenerative diseases

Alexandre Amlie-Wolf[1,2], Liming Qu[2], Elisabeth E. Mlynarski[2], Christopher D. Brown[1,2,3], Gerard D. Schellenberg[1,2,3], Li-San Wang[1,2,3]

1) Genomics and Computational Biology Graduate Group; 2) Penn Neurodegeneration Genomics Center, Department of Pathology and Laboratory Medicine; 3) Department of Genetics; **Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA**

alexaml@pennmedicine.upenn.edu – http://tesla.pcbi.upenn.edu/∼alexaml/

PNGC Penn Neurodegeneration Genomics Center

Perelman School of Medicine University of Pennsylvania

Motivation

- GWAS-identified variants tag linkage disequilibrium (LD) blocks of potentially functional variants, many of which are not causal
- Most GWAS variants are noncoding and may affect transcriptional regulatory elements
- Transcriptional enhancers are context-specific and annotations are incomplete, so information must be integrated across tissue contexts and data sources to identify affected regulatory mechanisms
- To translate GWAS findings into therapeutics, the target gene expression changes underlying disease risk must be identified
- The field lacks an integrative tool to identify functional variants, the specific regulatory elements they affect, the relevant tissue context, and the affected target genes

INFERNO - INFERring the molecular mechanisms of NOncoding genetic variants

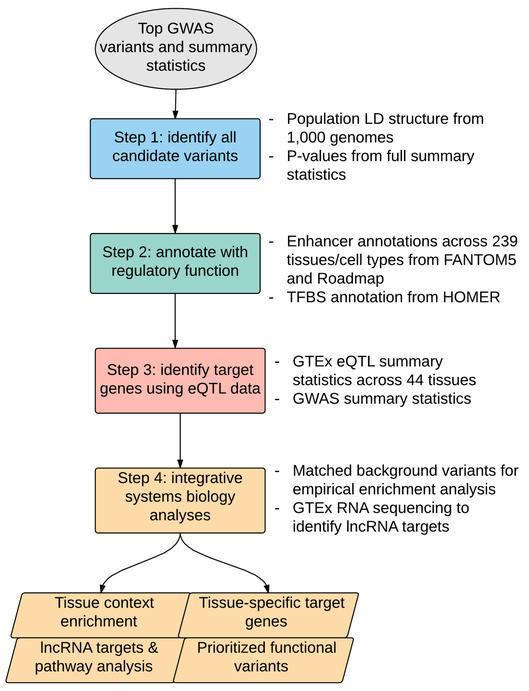

Figure 1: Schematic of INFERNO

- Sets of all putatively causal variants are generated by p-value and LD expansion
- Tissues and cell types from each functional genomics data source are grouped into 32 broad tissue categories for cross-data source comparison
- Empirical p-values for the enrichment of functional overlaps in each tissue category are obtained by sampling control variants matched on LD block size, distance to nearest gene, and MAF
- To improve on direct eQTL overlap, which is biased by LD structure, INFERNO applies the COLOC Bayesian method [1] for co-localization analysis of GWAS and GTEx eQTL signals
- Co-localization analysis often identifies lncRNA eQTL targets, so GTEx RNA-seq data across 11,439 tissue samples is used to compute lncRNA - mRNA expression correlations and identify targeted mRNAs
- INFERNO is available as an open source software tool and Docker image (https://bitbucket.org/alexamlie/inferno)
- Top variant expansion and annotation overlap analysis is provided on a web server (http://inferno.lisanwanglab.org)

Data sources for landscape analysis application of INFERNO

Phenotype	Acronym	Number of input variants	Citation
Frontotemporal dementia	FTD	7 significant	[2]
Corticobasal degeneration	CBD	7 significant	[3]
Amyotrophic lateral sclerosis	ALS	7 significant	[4]
Parkinson's disease	PD	23 significant	[5]
Alzheimer's disease	IGAP	19 significant	[6]
Progressive supranuclear palsy	PSP	7 significant, 12 suggestive	Wang et al., unpublished
Attention deficit hyperactivity disorder	ADHD	7 suggestive	[7]
Autism spectrum disorder	ASD	51 suggestive	Preliminary results, 2015
Bipolar disorder	BIP	43 significant	[8]
Major depressive disorder	MDD	11 suggestive	[9]
Schizophrenia	SCZ	111 significant	[10]

All psychiatric phenotype data come from the Psychiatric Genomics Consortium (PGC)

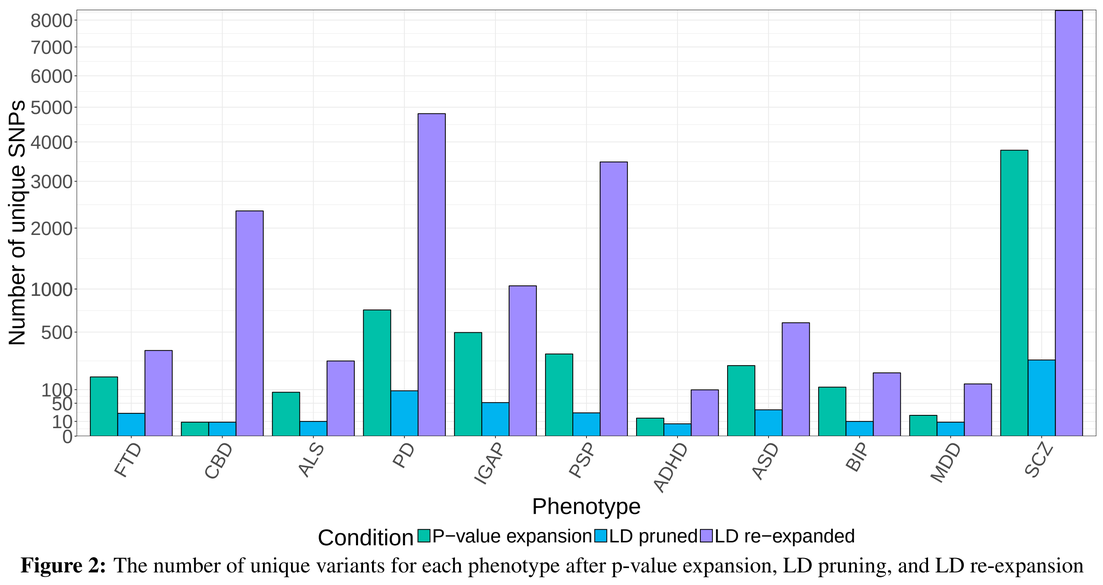

Figure 2: The number of unique variants for each phenotype after p-value expansion, LD pruning, and LD re-expansion

Cross-phenotype analysis results

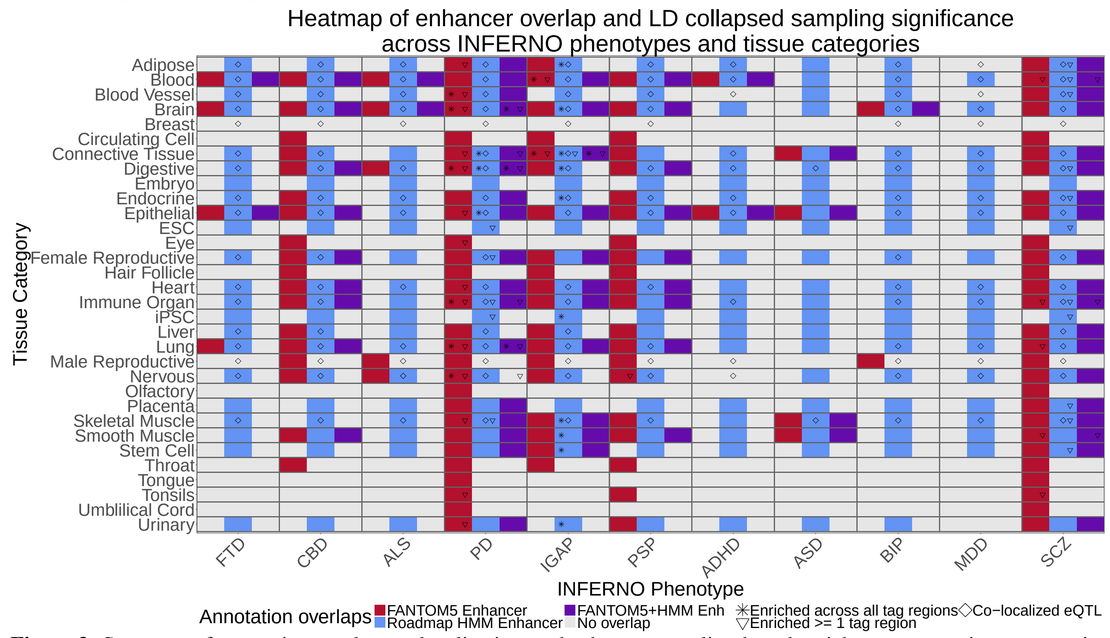

Figure 3: Summary of annotation overlap, co-localization, and enhancer sampling-based enrichment across tissue categories and phenotypes. Enrichments defined as adjusted p-value ≤ 0.05, across all tag regions ($*$) or in at least one ($\bigtriangledown$). Co-localized eQTL ($\Diamond$) defined as at least one strongly co-localized ($P(H_4) \geq 0.5$) eQTL - GWAS signal in a tissue class

Tissue-specific pathway enrichments of SCZ lncRNA targets

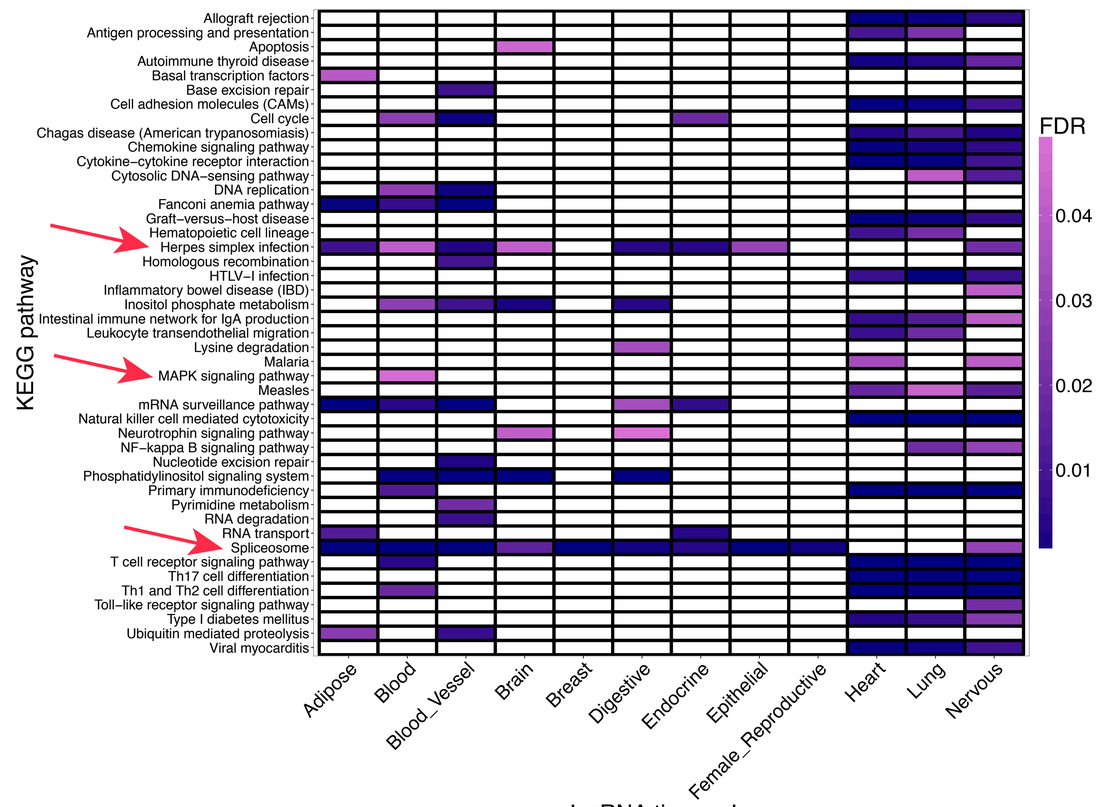

Figure 4: KEGG pathway enrichments for 6,005 total strongly correlated mRNA targets of 46 lncRNAs targeted by eQTLs co-localized with SCZ GWAS signal, split by tissue class of lncRNA eQTL signal

Conclusions

The INFERNO tool provides an easy and powerful approach for inferring the molecular mechanisms of noncoding genetic variants. We have implemented INFERNO in an efficient open source pipeline and Docker image (https://bitbucket.org/alexamlie/inferno). A web server that takes in top variants and expands them by LD and performs annotation overlap and summary is available at http://inferno.lisanwanglab.org.

The application of INFERNO to a range of psychiatric and neurodegenerative phenotypes identified putatively causal variants and the regulatory elements they disrupt, characterized the relevant tissue contexts in a hypothesis free manner, and identified the affected target genes in a sensitive manner. Application to schizophrenia recovered known schizophrenia-associated pathways including MAPK signaling, splicing, and Herpes infection.

Methods

INFERNO is implemented using Python, R, and bash. Datasets from each functional genomics consortium were grouped into tissue categories based on the categorization provided by Roadmap and the CL ontology. For enhancer sampling-based enrichment analysis, variants were matched on minor allele frequency (bin size 0.01), distance to the nearest TSS (rounded to 1kb), and the number of LD partners. Strong lncRNA correlation is defined as $|C_p| \geq 0.5$ & $|C_s| \geq 0.5$ where C_p is Pearson correlation and C_s is Spearman correlation. Multiple testing correction was performed using the Benjamini-Hochberg procedure. All summary statistics were obtained directly from the authors except for the IGAP and PGC datasets, obtained from the IGAP consortium and from the PGC downloads page (https://www.med.unc.edu/pgc/results-and-downloads/downloads).

References

[1] Claudia Giambartolomei, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genetics*, 10(5), 2014.

[2] Raffaele Ferrari, et al. Frontotemporal dementia and its subtypes: A genome-wide association study. *The Lancet Neurology*, 13(7):686–699, 2014.

[3] Naomi Kouri, et al. Genome-wide association study of corticobasal degeneration identifies risk variants shared with progressive supranuclear palsy. *Nature Communications*, 6:7247, 2015.

[4] Wouter van Rheenen, et al. Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nature Genetics*, 48(9), 2016.

[5] Mike a Nalls, et al. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nature Genetics*, 056(9):1–7, 2014.

[6] J C Lambert, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, 45(12):1452–8, 2013.

[7] Benjamin M. Neale, et al. Meta-Analysis of Genome-Wide Association Studies of Attention-Deficit/Hyperactivity Disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, 49(9):884–897, 2010.

[8] Pamela Sklar, et al. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature Genetics*, 43(10):977–983, 2011.

[9] Stephan Ripke, et al. A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular Psychiatry*, 18(4):497–511, 2013.

[10] Stephan Ripke, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427, 2014.

Acknowledgements

This work is supported by NIA T32-AG00255, U01-AG032984 (ADGC), UF1-AG047133 (CASA), U54-AG052427 (GCAD), U24-AG041689 (NIAGADS), R01-GM099962, P30-AG010124 (Penn ADC), RF1-AG055477, R01-AG054060, and U54-NS100693. We thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses. We thank the Psychiatric Genomics Consortium (PGC) for freely providing their summary statistics including for studies in progress.