INFERNO - INFERring the molecular mechanisms of NOncoding genetic variants

Alexandre Amlie-Wolf[1,2,3], Mitchell Tang[2,3], Pavel P. Kuksa[2,3], Yuk Yee Leung[2,3], Barry Slaff[4], Jessica King[3], Beth Dombroski[3], Gerard D. Schellenberg[3], Li-San Wang[1,2,3,4]

1) Genomics and Computational Biology Graduate Group, Perelman School of Medicine; 2) Institute for Biomedical Informatics, Perelman School of Medicine; 3) Department of Pathology and Laboratory Medicine, Perelman School of Medicine; 4) Department of Computer and Information Science, School of Engineering and Applied Science, **University of Pennsylvania**, Philadelphia, PA, USA

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

PerelmanMedicine_logo_blue.pdf

Institute for Biomedical Informatics

Introduction

- 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 like enhancers
- 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
- We propose INFERNO, a new tool that addresses these limitations by integrating hundreds of diverse data sets to identify the causal variants within LD blocks, the enhancers they affect, the tissue context of the enhancers, the affected target genes, and the direction of these effects

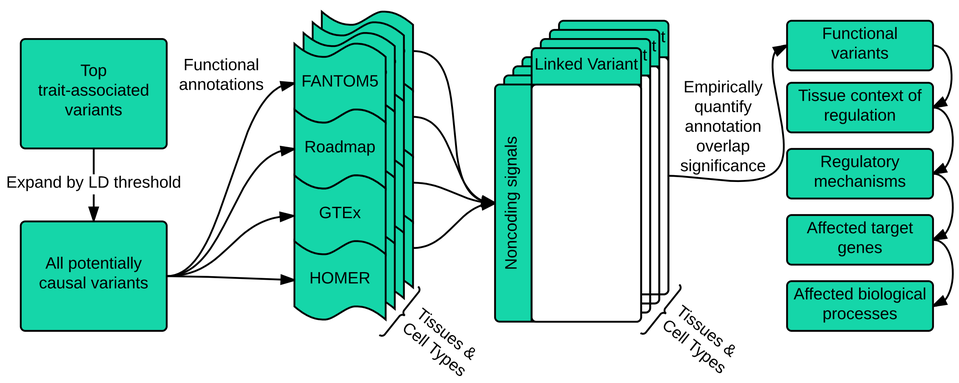

Figure 1: Schematic of INFERNO

- INFERNO uses 1,000 Genomes Project [1] data to define LD blocks
- Variants are annotated with:
  - Sites of enhancer RNA (eRNA) transcription across 112 tissue facets from FANTOM5 [2]
  - ChromHMM-defined epigenetic enhancer states across 127 tissues and cell types from Roadmap Epigenomics [3]
  - Expression quantitative trait loci (eQTL) across 44 tissues from GTEx [4]
  - Transcription factor binding sites (TFBSs) for 332 transcription factors predicted by HOMER [5]
- Tissues and cell types from each data source are grouped into 32 broad tissue categories for cross-data source comparison, and empirical p-values for the enrichment of functional overlaps in each tissue category and tag region are obtained by bootstrapping

Application of INFERNO to Alzheimer's Disease Genetic Signals

We applied INFERNO to 19 noncoding variants associated with late-onset Alzheimer's Disease (LOAD) identified in phase 1 of the International Genomics of Alzheimer's Project (IGAP) meta-analysis [6] excluding the variant in the DSG2 region, which did not replicate, and the variant in the HLA region, which is notoriously hard to analyze. We defined an expanded set of 706 variants by identifying all SNPs within 500 kb of any tag SNP with p-value within one order of magnitude of the top p-value. Then we subjected this set to LD pruning, yielding 67 variants, which were submitted as input to the INFERNO tool. After LD expansion, 1,333 unique variants were analyzed for regulatory potential.

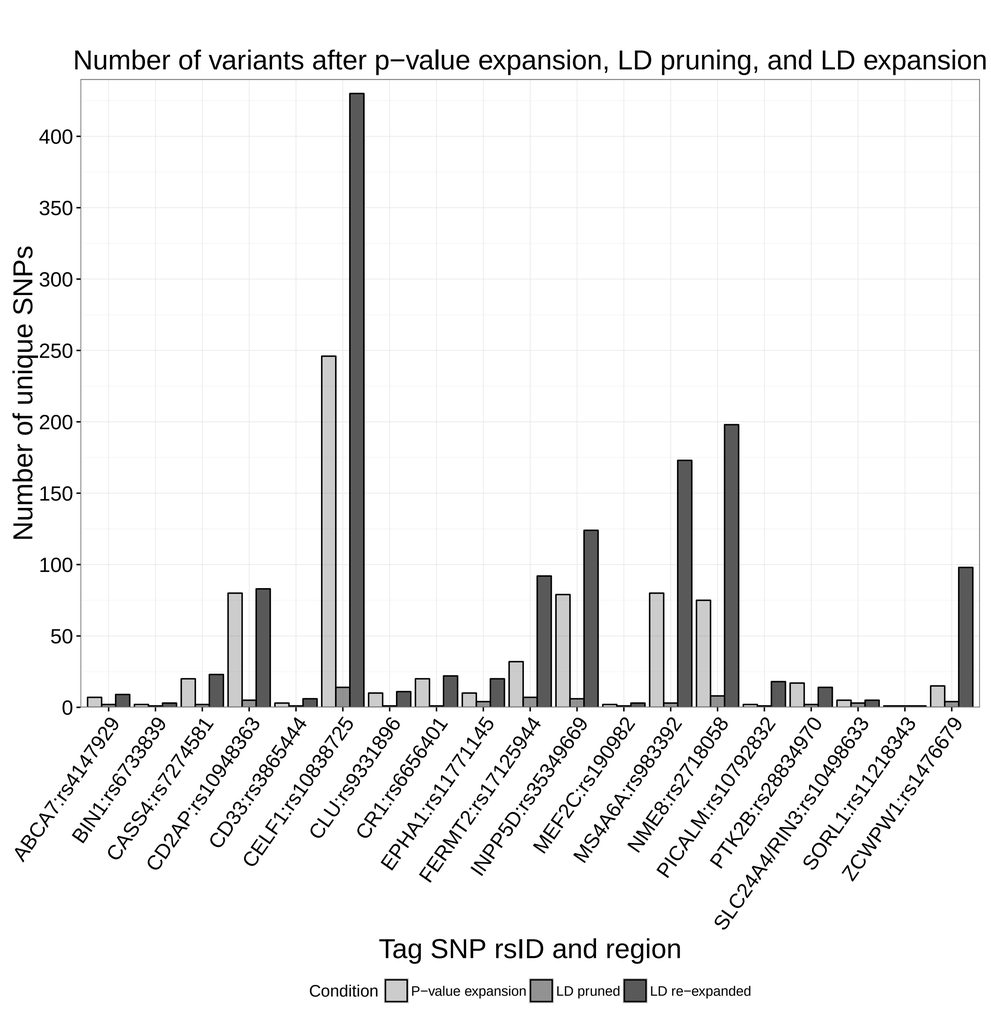

Figure 2: The number of unique variants in each tag region after p-value expansion, LD pruning, and LD re-expansion

Integrative cross-tissue functional analysis

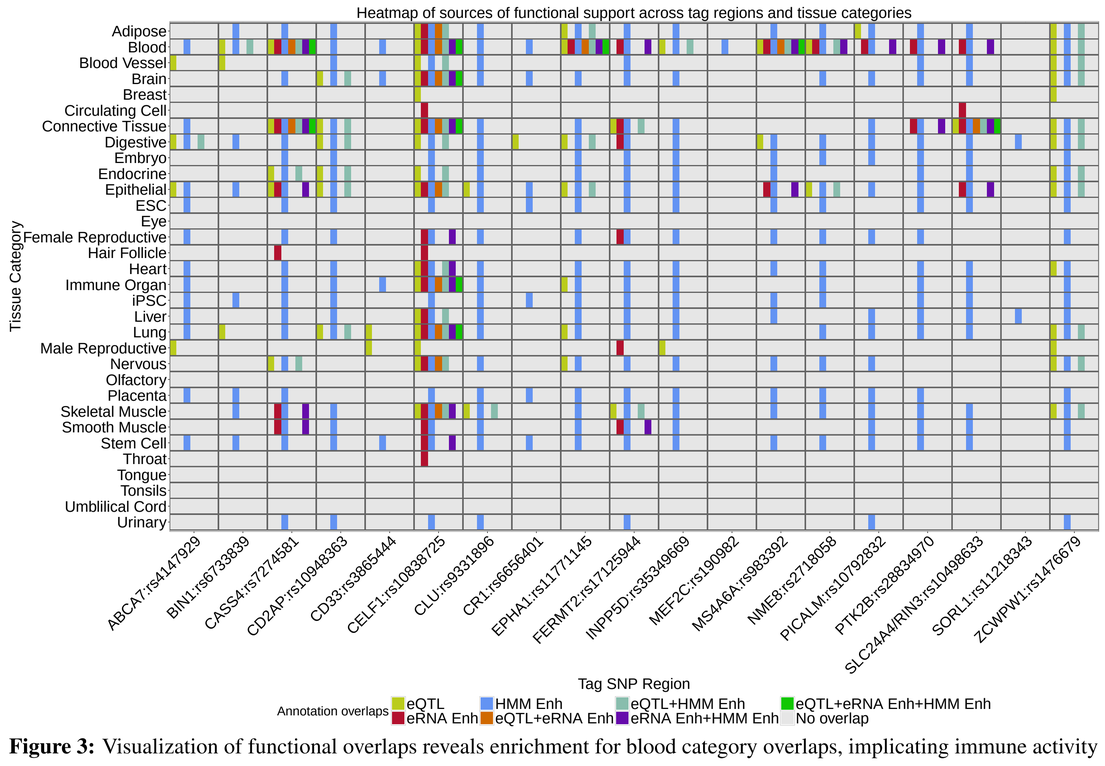

Figure 3: Visualization of functional overlaps reveals enrichment for blood category overlaps, implicating immune activity

Bootstrapping for functional enrichments

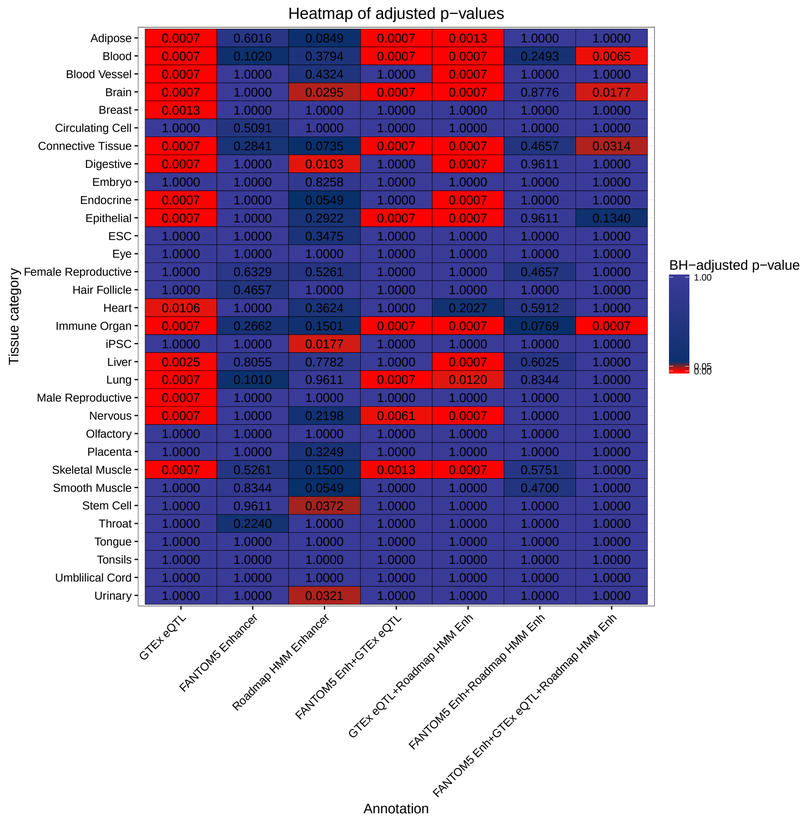

Figure 4: Empirical p-values for tissue and annotation overlap combinations based on 10,000 bootstrapped samples supports blood, brain, connective tissue (fibroblast) and immune organ

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 pipeline with source code and access to a web server version that will be available at http://lisanwanglab.org/INFERNO.

The application of INFERNO to the analysis of LOAD-associated noncoding genetic signals identified a small number of putatively causal SNPs with strong functional evidence, and the significant enrichment of functional overlaps in the blood and immune organ categories supports the hypothesis of immune activity as an important aspect of LOAD pathology.

Methods

INFERNO is implemented using Python, R, and bash. Datasets from each consortium were grouped into tissue categories based on the categorization provided by Roadmap and the CL ontology. For bootstrapping, 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. Multiple testing correction was performed using the Benjamini-Hochberg procedure.

References

[1] Adam Auton, et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, sep 2015.

[2] Robin Andersson, et al. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–61, mar 2014.

[3] Roadmap Epigenomics Consortium, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518:317–330, 2015.

[4] K. G. Ardlie, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, may 2015.

[5] Sven Heinz, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, 38(4):576–589, 2010.

[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.

Acknowledgements

This work is supported by NIA T32-AG00255, NIGMS R01-GM099962, NIA U24-AG041689, U01-AG032984, and P30-AG10124. We thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses.