Inferring the Molecular Mechanisms of Noncoding AD-Associated Genetic Variants

Alexandre Amlie-Wolf, Mitchell Tang, Beth Dombroski, Jessica Way, Ming Jiang, Nikolaos Vrettos, Yi-Fan Chou, Elizabeth E. Mlynarski, Christopher D. Brown, Gerard D. Schellenberg, Li-San Wang

Genomics and Computational Biology Graduate Group, Penn Neurodegeneration Genomics Center, Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine

Introduction

Genome-wide association studies (GWAS) have identified more than 20 genetic variants associated with late-onset Alzheimer's disease (LOAD). However, most of these signals only tag markers in linkage disequilibrium (LD) with nearby causative alleles. Moreover, all of the significant variants identified in phase 1 of the International Genomics of Alzheimer's Project (IGAP) meta-analysis [1] are in non-protein-coding regions, implicating gene regulatory mechanisms as underlying the association signal. Thus, functional annotation of expanded sets of variants spanning the LD regions tagged by the IGAP signals is needed to identify causal variants, underlying regulatory mechanisms, and their corresponding target genes.

INFERNO

INFERNO: INFERring the molecular mechanisms of NOncoding genetic variants

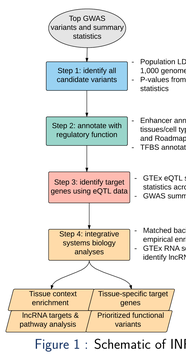

Figure 1 : Schematic of INFERNO

We applied our INFERNO (INFERring the molecular mechanisms of NOncoding genetic variants) [2] algorithm to identify causal noncoding variants and their regulatory effects. INFERNO expands variants by LD using population data from the 1,000 Genomes Project [3]. These variants are tested for overlap with enhancer annotations across 239 tissues and cell types from FANTOM5 [4] and Roadmap [5] and HOMER transcription factor binding sites (TFBSs) for 332 TFs [6]. The COLOC Bayesian statistical model [7] is used to identify GWAS signals sharing co-localized causal variants (posterior probability of shared causal variant $P(H_4) \geq 0.5$) with GTEx expression quantitative trait loci (eQTL) across 44 tissues [8]. GTEx RNA-seq data are used to predict co-regulatory networks of long noncoding RNAs (lncRNAs). Tissue types from individual functional genomic data sources are harmonized into 32 categories for unbiased identification of variants underlying co-localized eQTL signals that also overlap enhancers from the matching tissue context. A statistical sampling approach is used to identify significant enrichments of tissue-specific enhancer overlaps.

INFERNO is available as an open source pipeline and as a web server at http://inferno.lisanwanglab.org/.

Application of INFERNO to AD

We applied INFERNO to 19 IGAP phase 1 top hits 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 first defined an expanded set of 496 variants by identifying all SNPs within 500 kb of any tag SNP with a p-value within one order of magnitude of each tag SNP and the same effect direction. Then, to account for LD structure, we subjected this set to LD pruning, yielding 52 variants, which were submitted as input to the INFERNO tool. After LD expansion, 1,044 unique variants were analyzed for regulatory potential. Note that each region is referred to by the closest gene, but this is not necessarily the causal gene for each association signal.

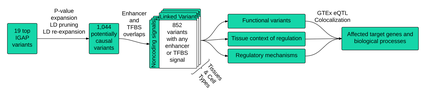

Figure 2 : Flowchart of IGAP INFERNO analysis

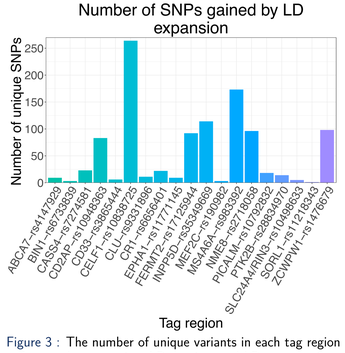

Figure 3 : The number of unique variants in each tag region after p-value expansion, LD pruning, and LD re-expansion

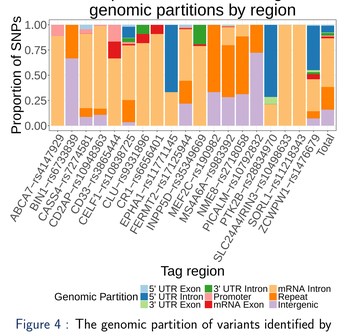

Figure 4 : The genomic partition of variants identified by p-value and LD expansion for each tag region. Only 17 out of 1,044 variants are in mRNA exons

Functional annotation of potentially causal variants

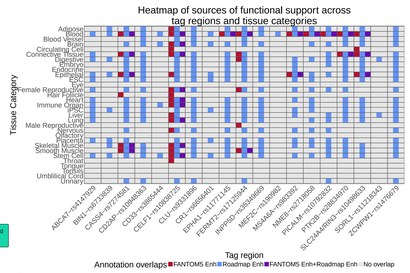

Figure 5 : Heatmap of FANTOM5 and/or Roadmap enhancer overlaps for each tag region and tissue category

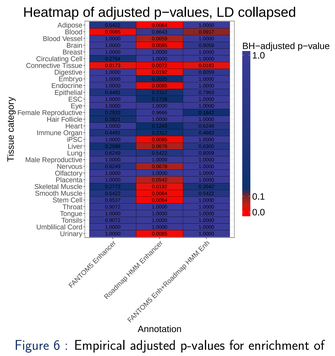

Figure 6 : Empirical adjusted p-values for enrichment of enhancer overlaps in each tissue category. Note that blood contains all the immunity-related cell lines

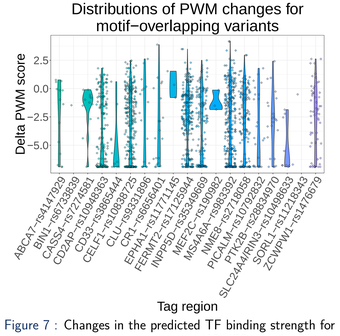

Figure 7 : Changes in the predicted TF binding strength for minor alleles of variants overlapping HOMER motifs. Positive values reflect increased binding strength and negative values reflect TFBS disruptions

eQTL co-localization and integrative analysis

We applied COLOC to eQTL signals for 888 unique genes across all 19 tag regions (median number of genes within each region = 34) for 25,601 tests of GWAS - tissue-specific eQTL co-localization. COLOC identified 155 tissue-specific eQTL signals co-localized with GWAS signals representing 16 tag regions, 37 tissues and 71 target genes where $P(H_4)$ was greater than 0.5, representing strong support for a shared causal signal. We prioritize variants underlying co-localized eQTL signals by filtering for enhancer overlap in the tissue category matching the eQTL signal. Further prioritization is achieved by identifying variants with high causal probability (ABF) and by TFBS overlap.

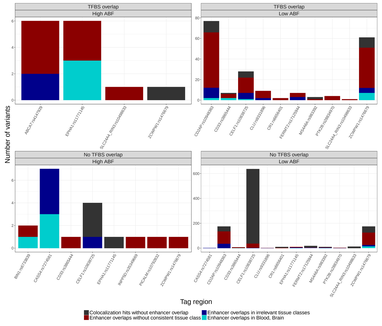

Figure 8 : Summary of variant prioritization

Tag Region	Affected mechanism and evidence	Direction of effect
ABCA7-rs4147929	Digestive system eQTL for **ABCA7**, high ABF variant rs4147929	Risk allele increases ABCA7 expression
BIN1-rs6733839	Lymphocyte eQTL for **BIN1**, high ABF variant rs4663105	Tag is risk allele, eQTL lowers BIN1 expression
CASS4-rs7274581	Whole blood and fibroblast eQTLs for **CASS4**, high ABF variants rs6014724 and rs927174	Tag is protective allele, lowered expression in blood, increase in fibroblasts
CD2AP-rs10948363	rs9367279 has TF disruption in enhancer for **RP11-385F7.1** in brain regulating GTPase signaling, **CD2AP** in fibroblasts	Tag is risk allele, slightly increased lncRNA expression in brain, lowered CD2AP expression in fibroblasts
CD33-rs3865444	Whole blood eQTL for **CD33**, tag variant colocalized with eQTL, high ABF variant	Protective tag variant decreases CD33 expression
CELF1-rs10838725	Brain cerebellar hemisphere eQTL for **RP11-750H9.5**, lncRNA regulating leukocyte activation and immune response, rs7947450 has enhancer and TF disruption	Risk allele, lowered lncRNA expression
CLU-rs9331896	Epithelial and digestive signals for **ZNF395** and **FZD3**, both supported by same variant rs2070926 with moderate ABF but TFBS overlap and many enhancer overlaps	Tag is protective, Variant lowers target gene expression
EPHA1-rs11771145	Strong ABF for rs11765305 affecting **EPHA1-AS1**, lncRNA regulating JAK2 signaling axis, in blood (monocytes) with strengthened CEBP motif	Protective tag variant, strong increase in EPHA1-AS1 expression
FERMT2-rs17125944	Skeletal muscle support for **FERMT2**, several variants (rs11626419, rs12586707, and rs7151474) with enhancer + motif support	Risk tag variant, lowered FERMT2 expression
ZCWPW1-rs1476679	One variant rs1727138 strongly disrupts several motifs and colocalizes with GTEx brain eQTLs for **PVRIG** and **STAG3**	Protective allele, inconsistent effects on expression levels across brain regions (mainly slight decreases)

Table 1 : Summary of co-localization results in 10 regions prioritized by INFERNO

Validation of allelic effects on enhancers

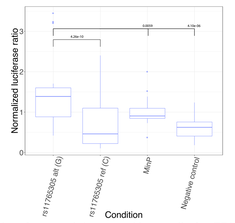

Figure 9 : Luciferase validation in K562 cells for EPHA1 region

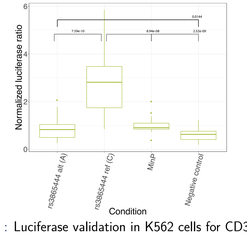

Figure 10 : Luciferase validation in K562 cells for CD33 region

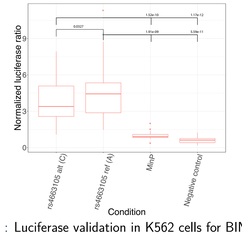

Figure 11 : Luciferase validation in K562 cells for BIN1 region

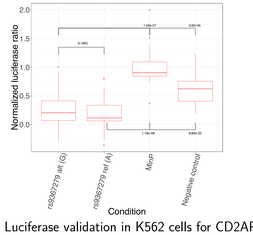

Figure 12 : Luciferase validation in K562 cells for CD2AP region (with no significant allelic difference)

For all plots, numbers refer to p-values from linear mixed effects model. Lack of p-value reflects nonsignificant effect. n = 5 biological replicates, with 4 technical replicates per condition per experimental day

Conclusions

Our approach of using INFERNO to overlap all potential causal variants with functional genomics annotation, perform eQTL co-localization analysis, and prioritize variants using an integrative tissue categorization scheme identified perturbations of tissue-specific regulatory mechanisms in 10 IGAP tag regions. In the EPHA1, CD2AP, and CASS4 regions, the target genes of the strongly colocalized eQTL signals included lncRNA transcripts, suggesting that identifying affected enhancers and target genes is only the first step to understanding genetically mediated dysfunction of regulatory networks. INFERNO includes an approach to identify co-regulatory networks of lncRNAs (Methods), which showed that these lncRNAs were involved in larger regulatory networks including GTPase receptor pathways, the JAK2/STAT3 signaling axis, and widespread leukocyte activation and immune response. Luciferase assays in K562 cells of the macrophage lineage validated several but not all of the predicted effects, including the correct effect directions. The CD2AP signal did not validate, possibly due to the main eQTL signal being in brain and not in blood or an immune-related cell line. Assays in other relevant cell lines are ongoing. INFERNO characterized the relevant causal variants, tissue contexts, regulatory mechanisms, and target genes underlying noncoding AD loci, providing novel post-GWAS insights into the role of immunity and lncRNAs in genetic susceptibility to AD and identifying potential therapeutic targets for future investigation.

References

[1] 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.

[2] Alexandre Amlie-Wolf, et al. INFERNO - INFERring the molecular mechanisms of NOncoding genetic variants. *bioRxiv*, page 211509, 2017.

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

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

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

[6] 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.

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

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

Acknowledgements

We thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses. The investigators within IGAP contributed to the design and implementation of IGAP and/or provided data but did not participate in analysis or writing of this report. IGAP was made possible by the generous participation of the control subjects, the patients, and their families. The i-Select chips was funded by the French National Foundation on Alzheimer's disease and related disorders. EADI was supported by the LABEX (laboratory of excellence program investment for the future) DISTALZ grant, Inserm, Institut Pasteur de Lille, Université de Lille 2 and the Lille University Hospital. GERAD was supported by the Medical Research Council (Grant no 503480), Alzheimer's Research UK (Grant no 503176), the Wellcome Trust (Grant no 082604/2/07/Z) and German Federal Ministry of Education and Research (BMBF): Competence Network Dementia (CND) grant no 01GI0102, 01GI0711, 01GI0420. CHARGE was partly supported by the NIH/NIA grant R01 AG033193 and the NIA AG081220 and AGES contract N01-AG-12100, the NHLBI grant R01 HL105756, the Icelandic Heart Association, and the Erasmus Medical Center and Erasmus University. ADGC was supported by the NIH/NIA grants: U01 AG032984, U24 AG021886, U01 AG016976, and the Alzheimer's Association grant ADGC-10-196728. This work was funded by NIA T32 AG00255.

Methods

INFERNO is implemented using Python, R, and bash and is available at http://bitbucket.org/wanglab-upenn/INFERNO/ and at http://inferno.lisanwanglab.org/. lncRNA co-regulatory networks are identified by thresholds of 0.5 on Spearman and Pearson expression correlation with all other genes in the genome using GTEx RNA-sequencing datasets. For validation experiments, molecular cloning techniques were used to generate vectors spanning the INFERNO-predicted enhancer loci that differed only by the allele of each prioritized variant. These vectors were transfected into K562 cells. The negative control is a random heterochromatic region. Mock transfections were used for background subtraction, and the ratio of luciferase to renilla was calculated for each vector to control for transfection efficiency. These ratios were normalized against the average ratio for the minimal promoter. Statistical analysis was performed using a linear mixed model treating experimental days as random effects and alleles as fixed effect, and p-values for the allelic effects were generated using analysis of variance (ANOVA) using Satterthwaite's approximation for degrees of freedom.

Contact information:
Email: alexaml@upenn.edu
Website: http://tesla.pcbi.upenn.edu/~alexaml/