Computational Identification of Regulatory Mechanisms Affected By Noncoding Variants Associated with Late-Onset Alzheimer's Disease

Perelman School of Medicine, University of Pennsylvania
Institute for Biomedical Informatics
Alexandre Amlie-Wolf, Mitchell Tang, Jessica King, Beth Dombroski, Li-San Wang, Gerard D. Schellenberg
Genomics and Computational Biology Graduate Group, Department of Pathology and Laboratory Medicine, Institute on Aging, Institute for Biomedical Informatics, University of Pennsylvania Perelman School of Medicine

Introduction

Dozens of single nucleotide polymorphisms (SNPs) associated with late-onset Alzheimer's disease (LOAD) have been identified by genome-wide association studies (GWAS). However, these are 'tagging SNPs' for nearby genetic variants in linkage disequilibrium (LD) and may not be actually functional. Moreover, all 21 of the significant SNPs 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. These considerations suggest a need for functional annotation of all the variants in LD with the IGAP tag SNPs in order to prioritize the truly causal variants and identify the affected regulatory mechanisms, tissue contexts, target genes, and affected biological processes.

To address this need, we developed a novel tool, called INFERNO (INFERring the molecular mechanisms of NOncoding genetic variants). Given a list of tagging variants, INFERNO uses 1,000 Genomes Project [2] data to define LD blocks. Each linked variant is compared with: sites of enhancer RNA (eRNA) transcription across 112 tissue facets from the FANTOM5 consortium [3]; ChromHMM-defined epigenetic enhancer states across 127 tissues and cell types from the Roadmap Epigenomics consortium [4]; expression quantitative trait loci (eQTL) across 44 tissues from the GTEx consortium [5]; and predicted transcription factor binding sites (TFBSs) for 332 transcription factors from HOMER [6]. Tissues and cell types from each data source are grouped into 32 broad tissue categories for comparison, and empirical p-values for the enrichment of functional overlaps in each tissue category are obtained by bootstrapping.

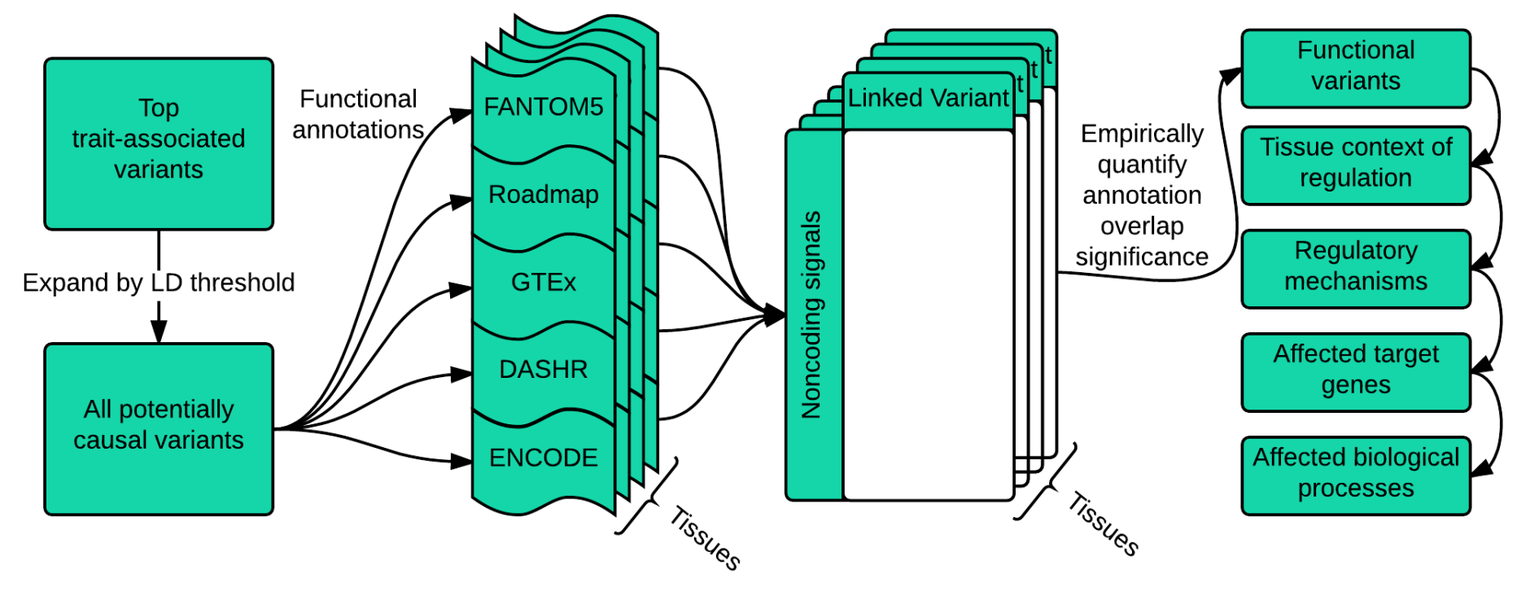

Figure 1 : Schematic of INFERNO

Application of INFERNO to IGAP

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 706 variants by identifying all SNPs within 500 kb of any tag SNP with p-value within one order of magnitude of the tag SNP. Then, to account for LD structure, 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.

Annotation	SNP Count
FANTOM5 eRNA transcription	53
GTEx eQTL	976
Roadmap epigenetic enhancer state (HMM)	852
HOMER Motif	585
eRNA and eQTL	41
eRNA and HMM enhancer	51
eQTL and HMM enhancer	613
eRNA, eQTL, and HMM enhancer	39

P-value expansion, LD pruning, and LD expansion of IGAP top hits

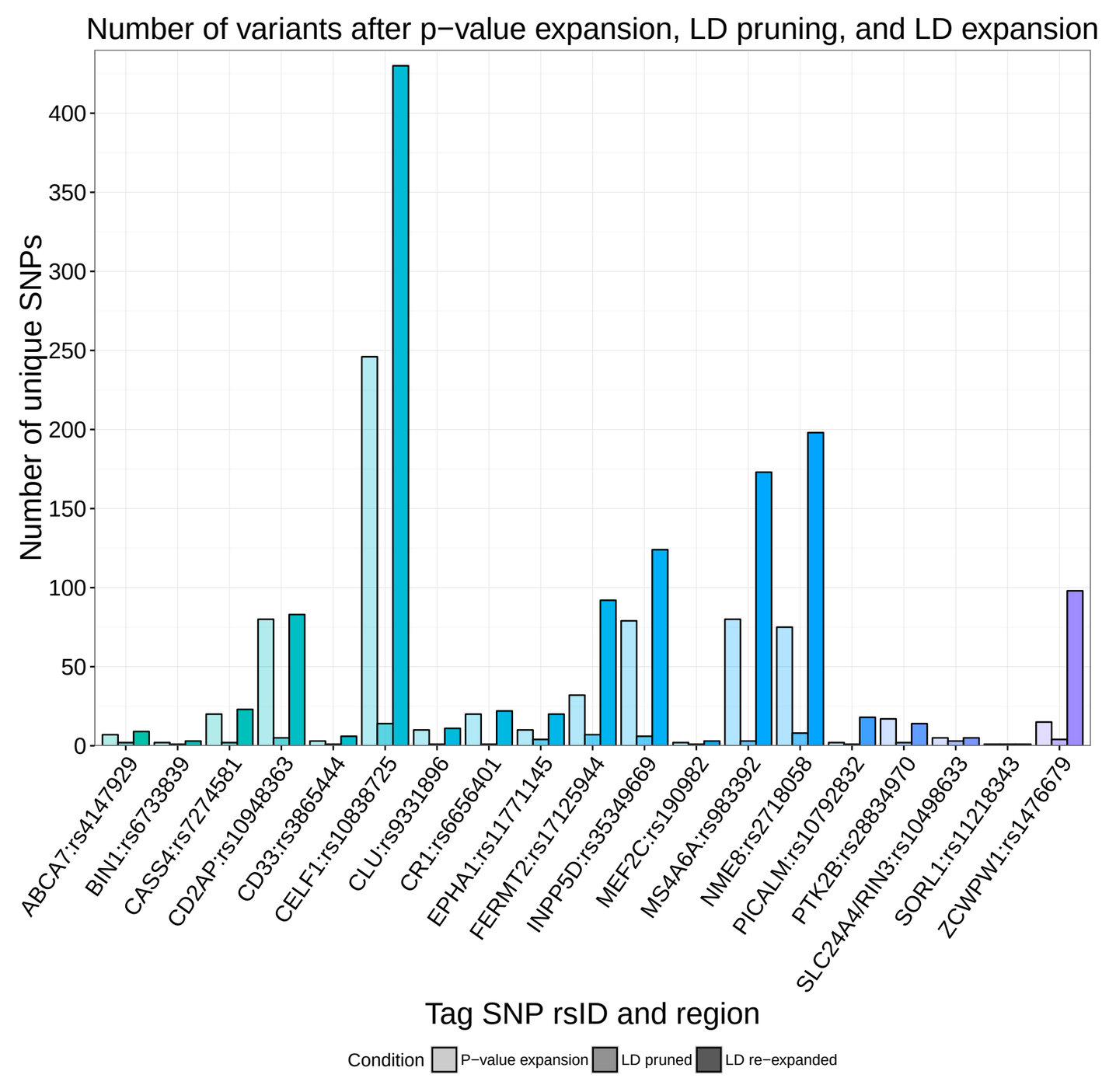

Figure 2 : The number of unique variants in each tag region after p-value expansion, LD pruning, and LD re-expansion

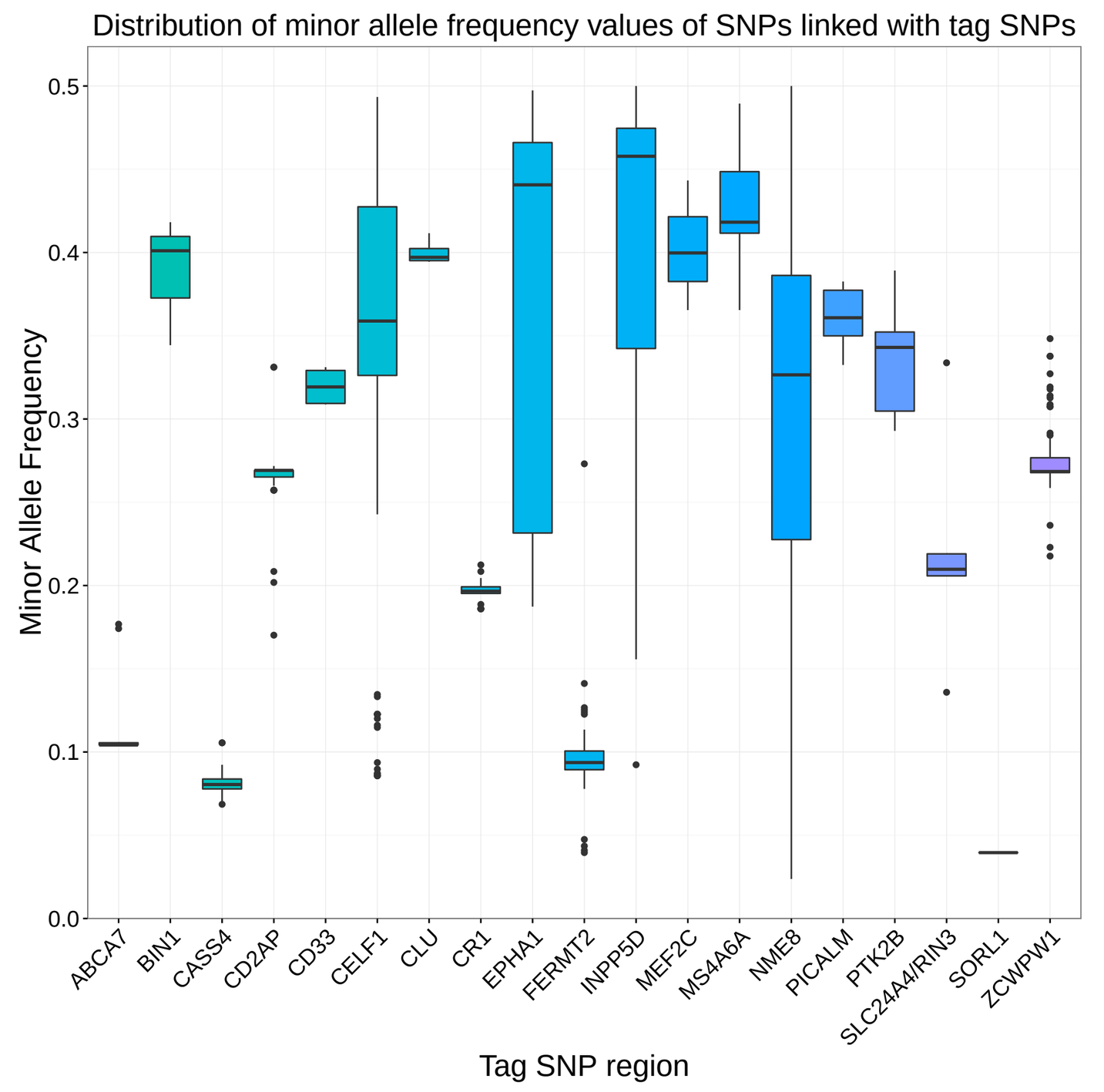

Figure 3 : The distributions of minor allele frequencies for variants in each tag region

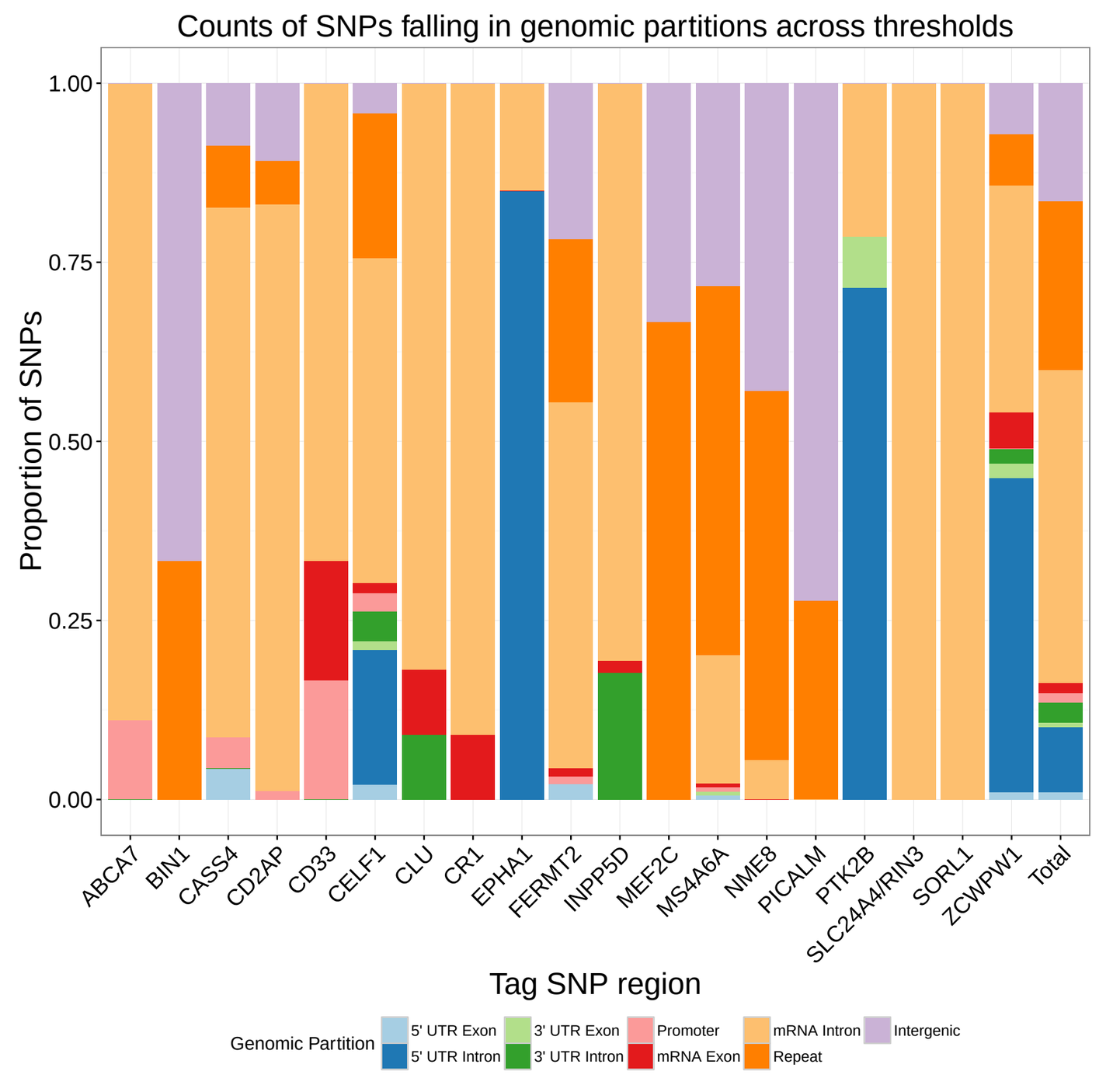

Figure 4 : The genomic partition of variants identified by p-value and LD expansion for each tag region

Functional annotation of expanded variant sets

Figure 5 : Heatmap of number of SNPs overlapping FANTOM5 enhancer RNA transcription

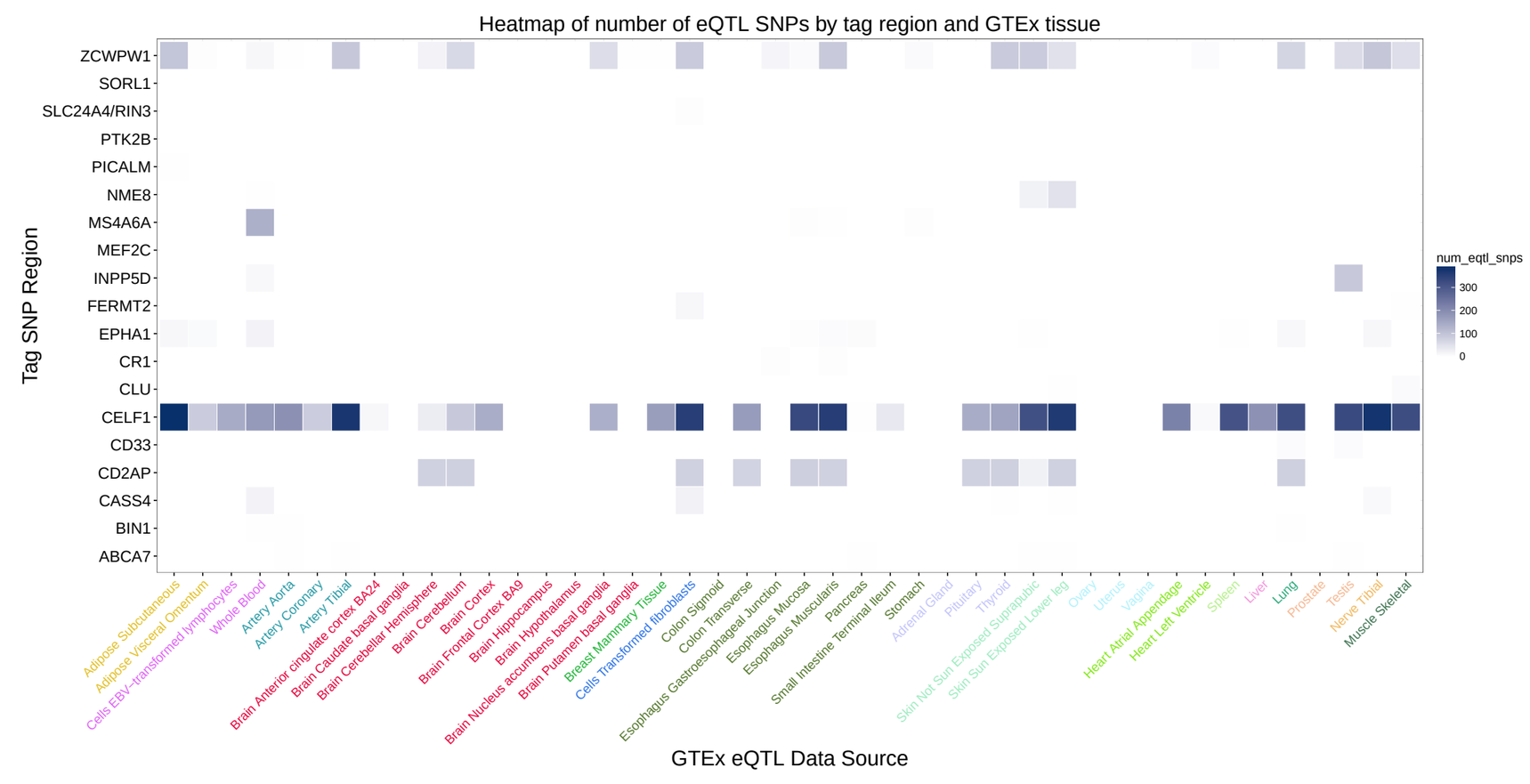

Figure 6 : Heatmap of number of SNPs overlapping GTEx eQTLs

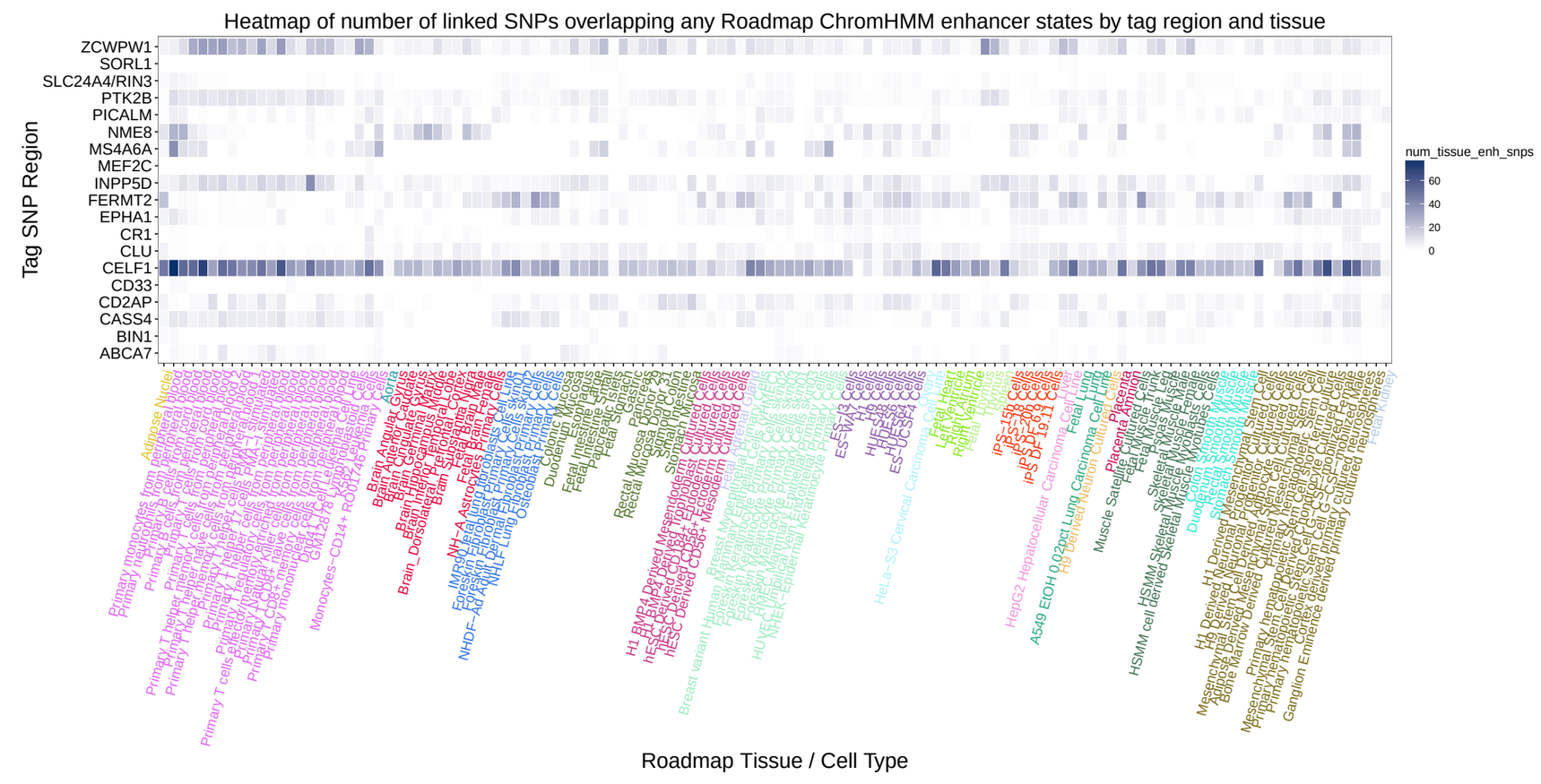

Figure 7 : Heatmap of number of SNPs overlapping ChromHMM-defined enhancer states

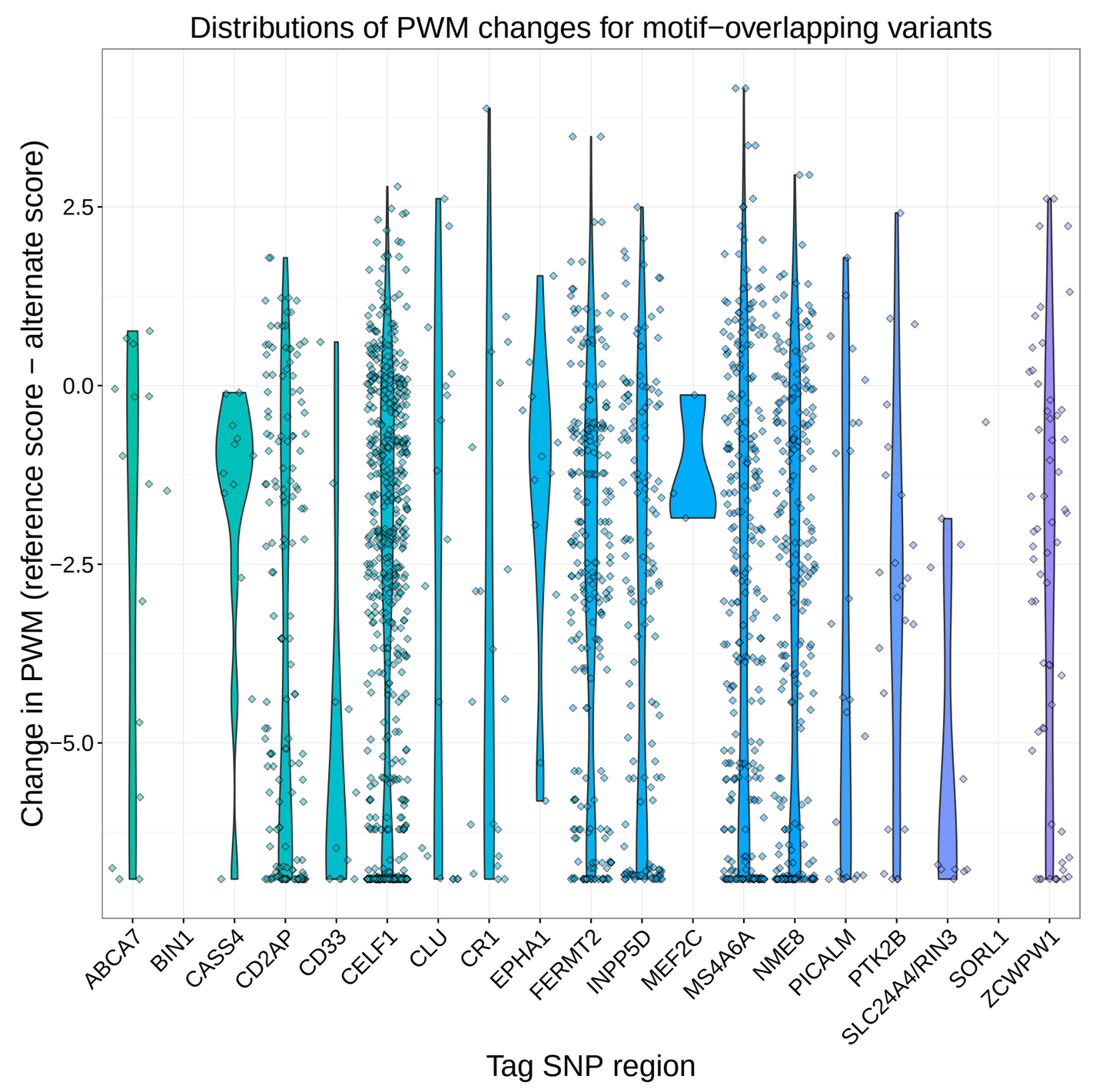

Figure 8 : Distributions of changes in the PWM score for variants overlapping HOMER motifs

Integrative cross-tissue functional overlap analysis

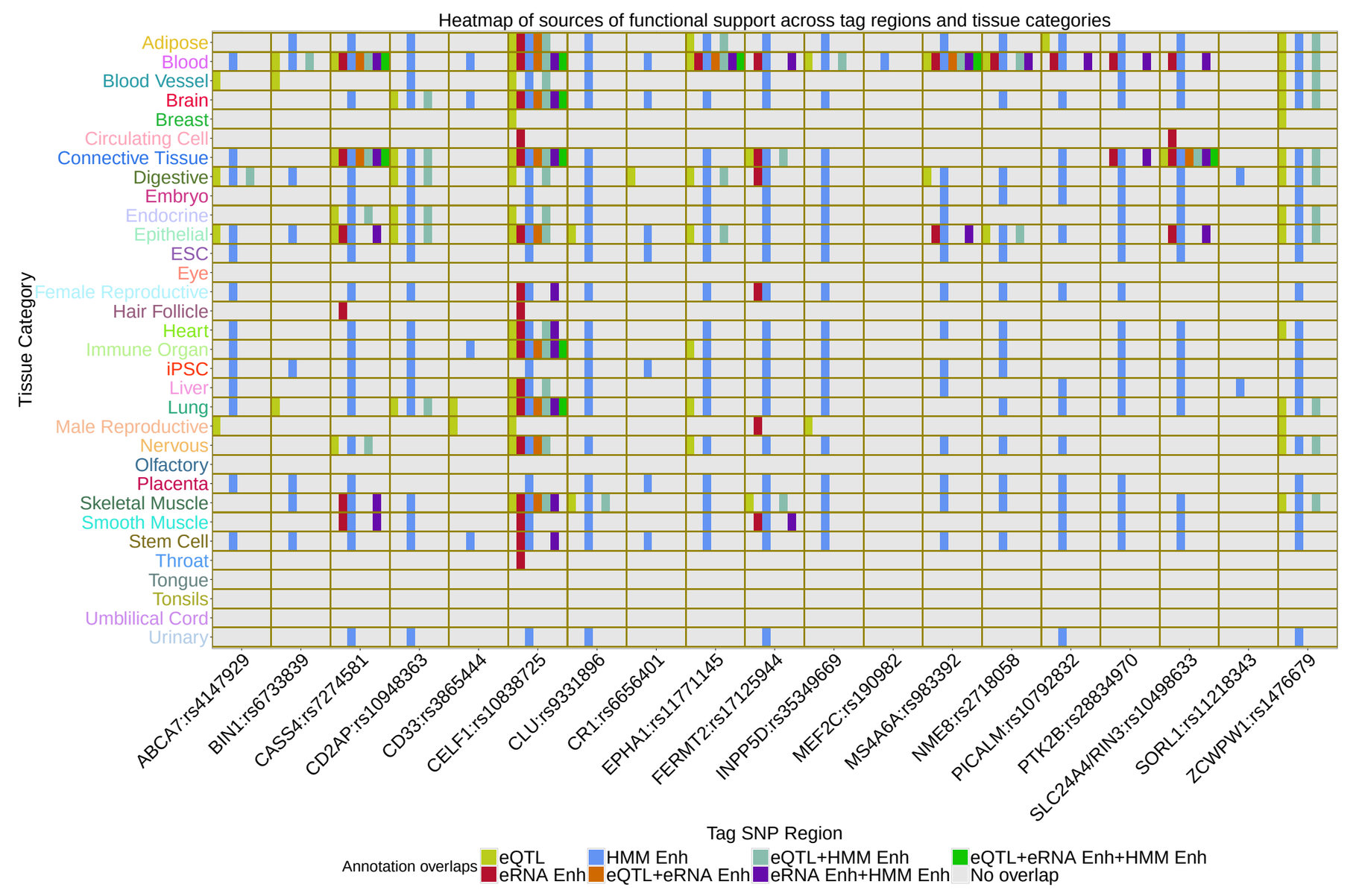

Figure 9 : Visualization of functional overlaps and combination of overlaps for each tag region and tissue category

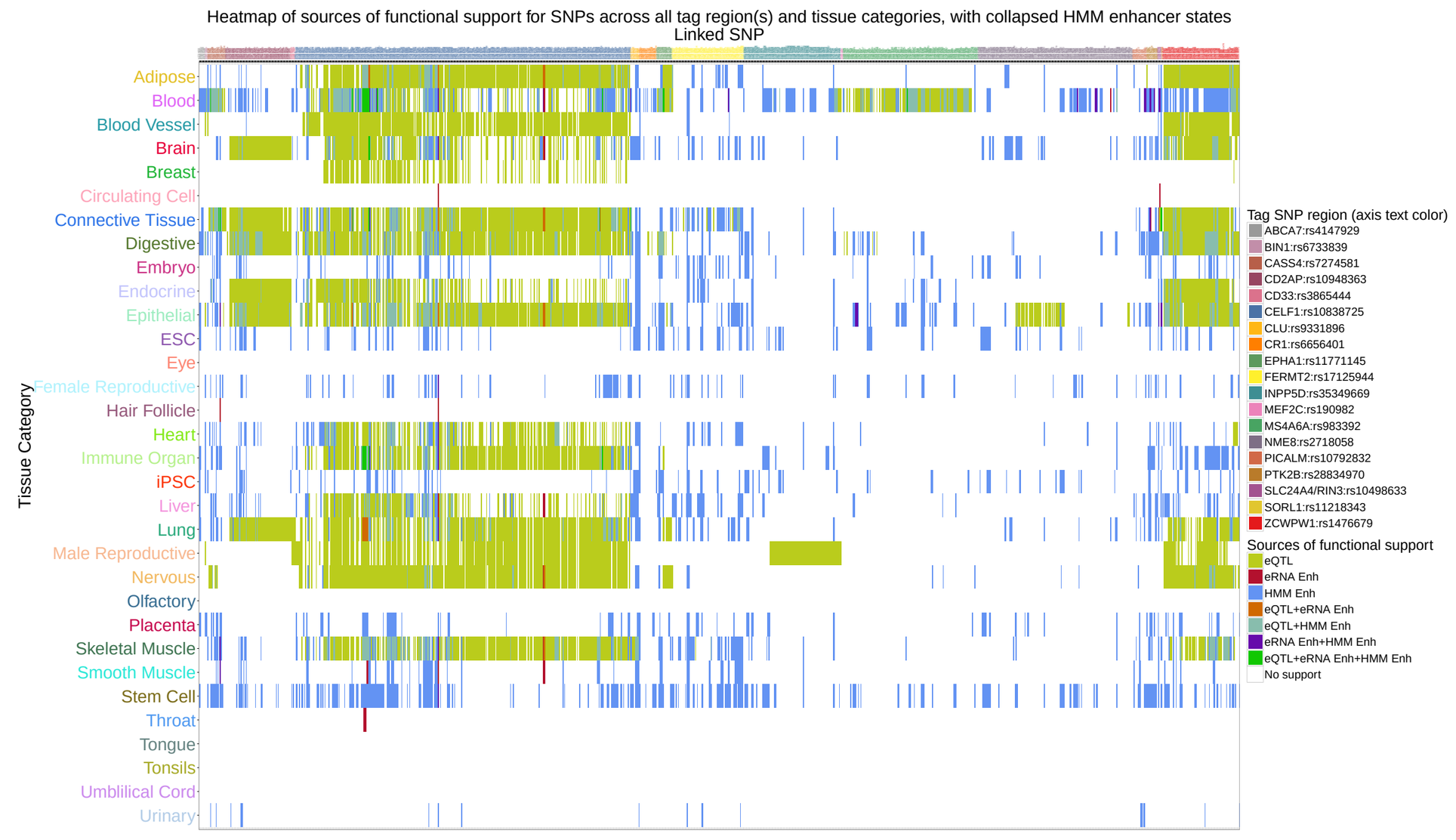

Figure 10 : Visualization of functional overlaps and combination of overlaps across individual SNPs and tissue categories

Bootstrapping for significance of annotation overlap

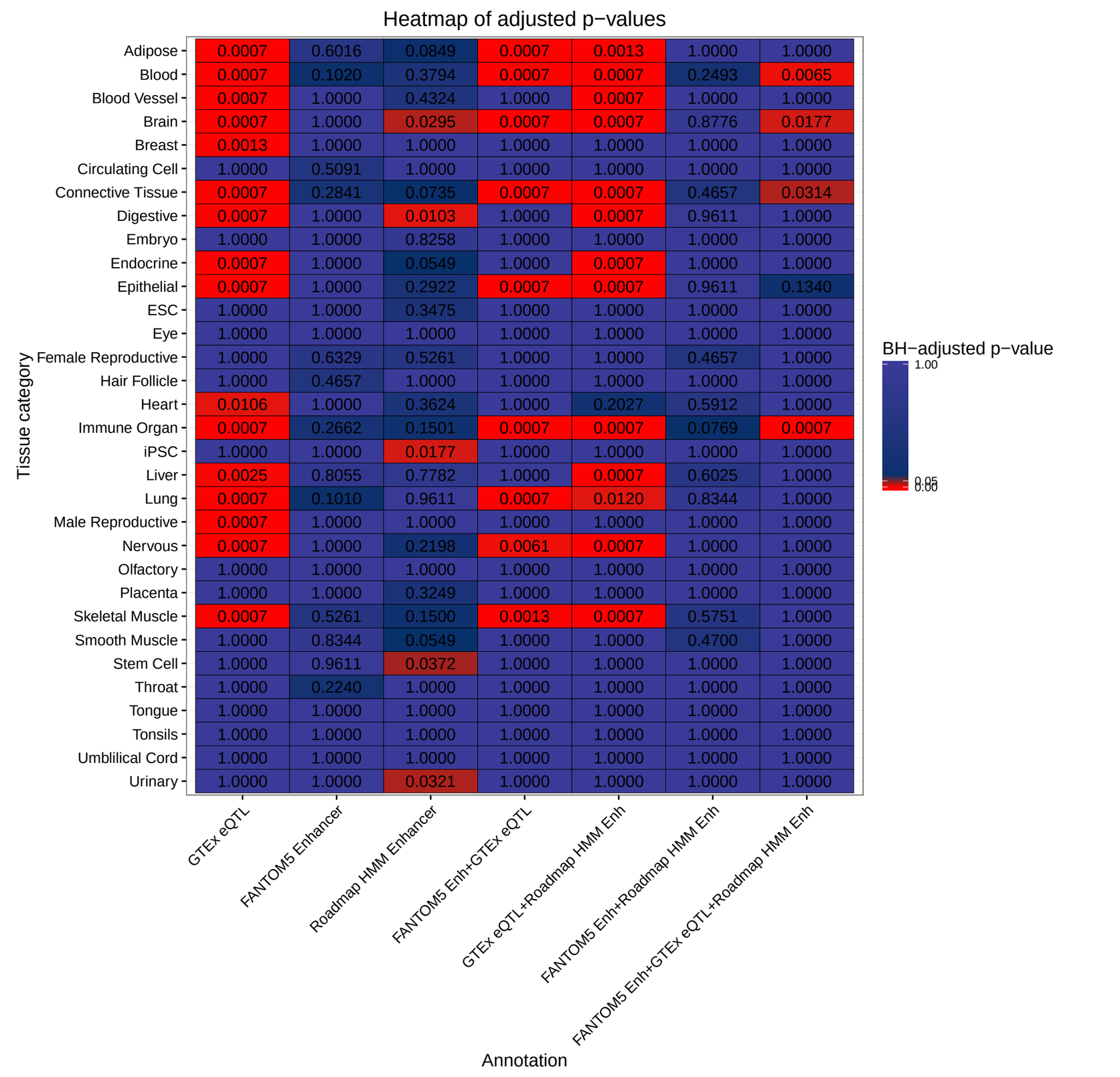

Figure 11 : Empirical p-values for tissue and annotation overlap combinations based on 10,000 bootstrapped samples

EPHA1 tag region analysis

Figure 12 : Annotation overlaps for individual SNPs in the EPHA1 region across tissue categories

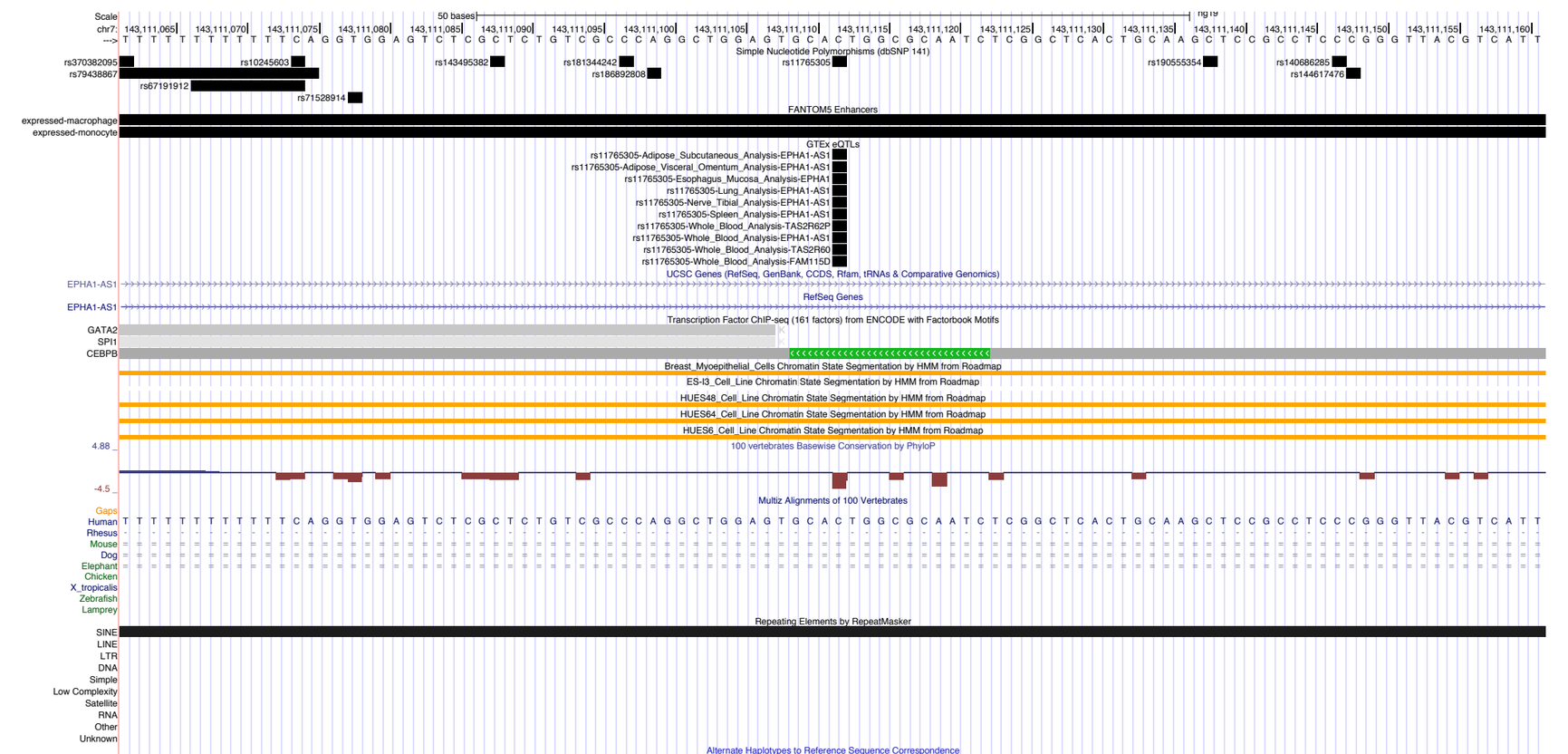

Figure 13 : Genome browser view of rs11765305 locus [chr7:143,111,062-143,111,161]

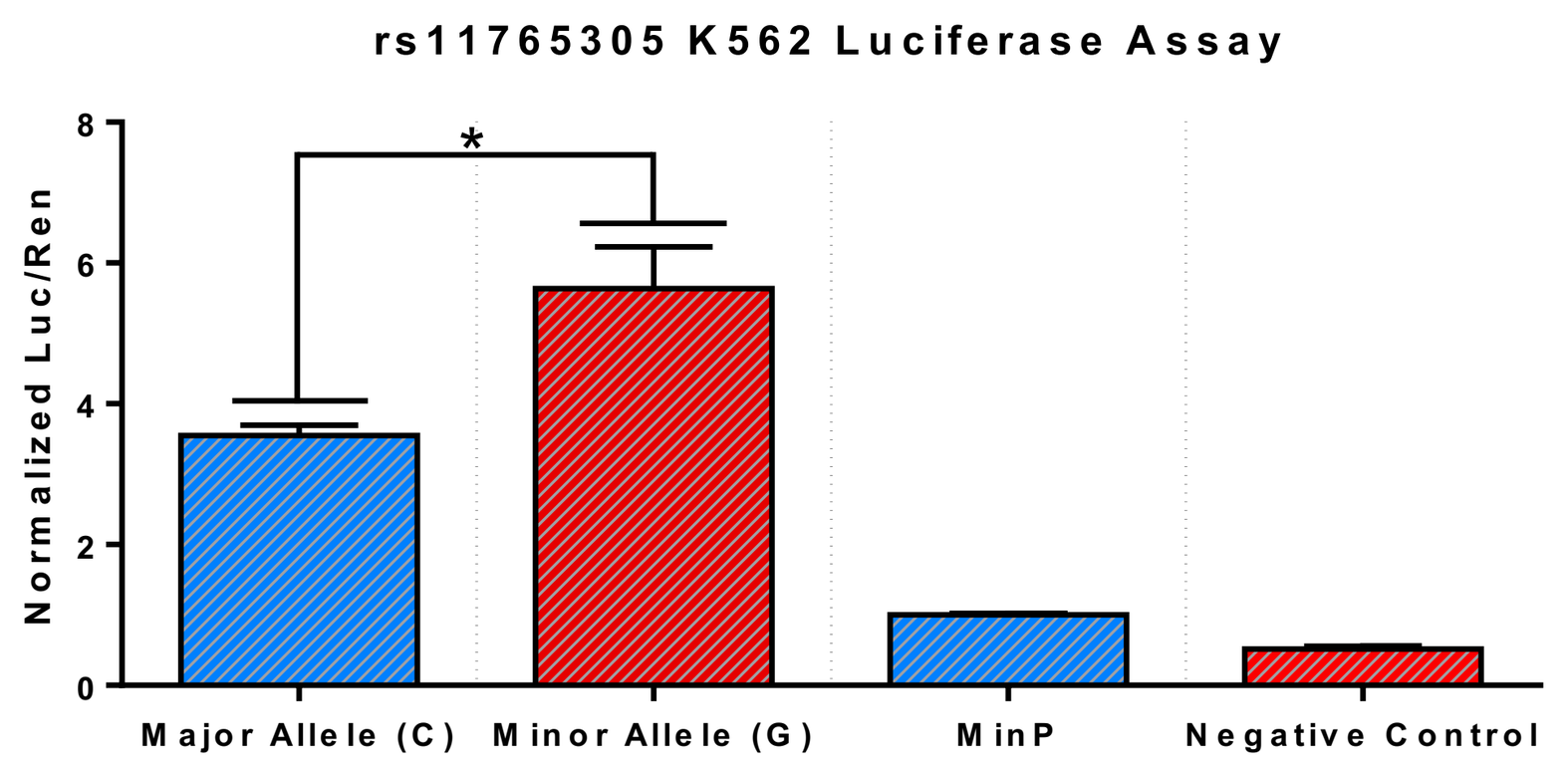

Figure 14 : Luciferase assay results for rs11765305 in K562 cells, a leukemia of monocyte precursor cell line. Asterisk represents statistically significant luciferase expression at the 0.05 level

IGAP suggestive regions

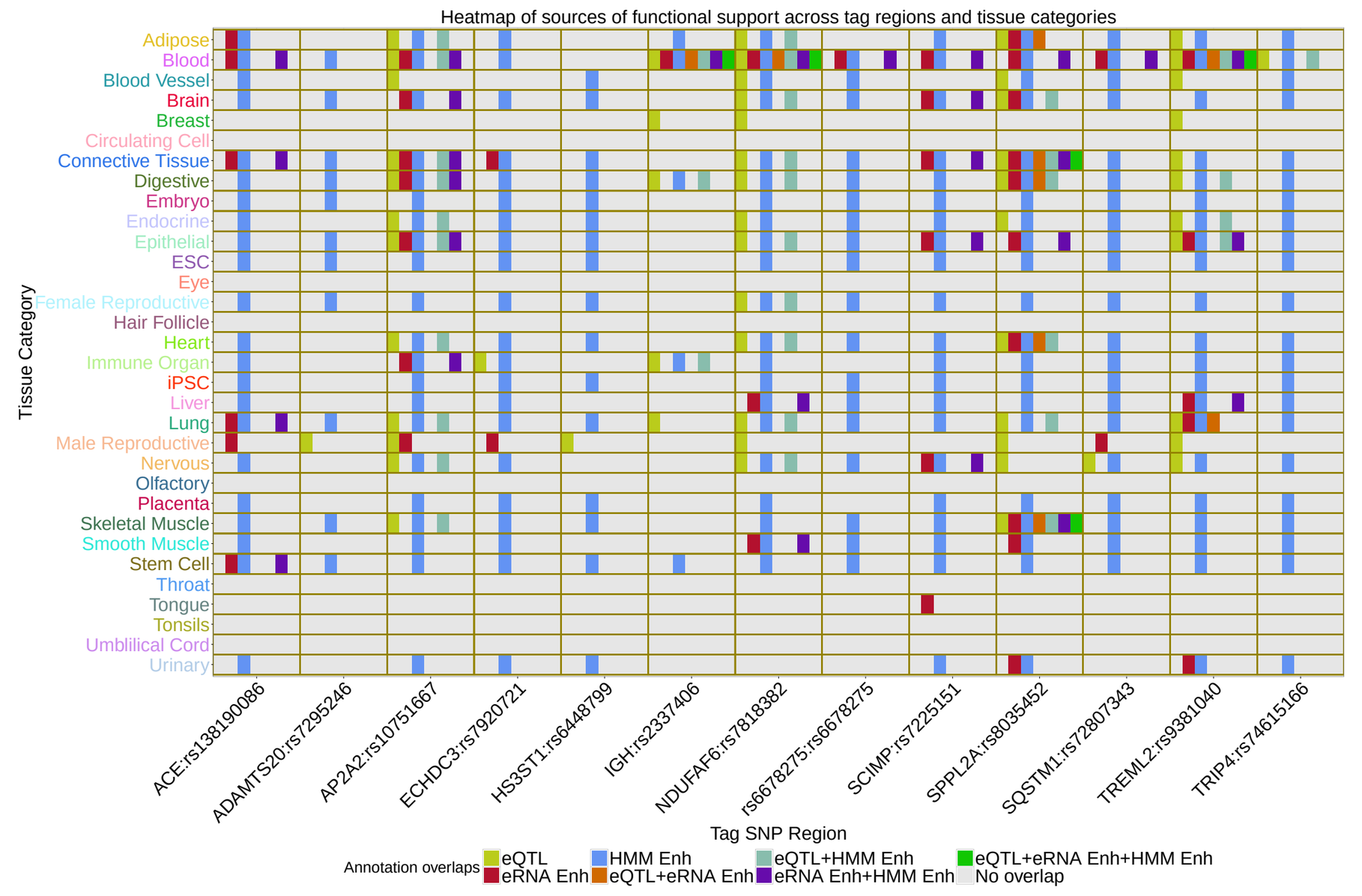

Figure 15 : Visualization of functional overlaps at 13 IGAP suggestive regions

Conclusions

Computational analysis of diverse types of functional genomic data across hundreds of tissues and cell types identified a small number of putatively causal SNPs for LOAD with strong functional evidence. The significant enrichment of functional overlaps in the blood and immune organ categories supports the hypothesis of immune activity forming important aspect of LOAD pathology, and we were able to identify a specific variant in the EPHA1 with strong evidence of enhancer activity and interaction with a long noncoding RNA transcript (EPHA1-AS1), which is currently being investigated.

In addition, 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 available at lisanwanglab.org/INFERNO.

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

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

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

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

Acknowledgements

We gratefully acknowledge the helpful feedback of Pavel Kuksa, Fanny Leung, and Barry Slaff.

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.

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. For the luciferase assay, luciferase expression was normalized against Renilla expression in the same well, and the negative control is a randomly sampled heterochromatin insert.

Contact Information

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