



Yale University
School of Medicine



Sefayet KARACA, Ph.D.

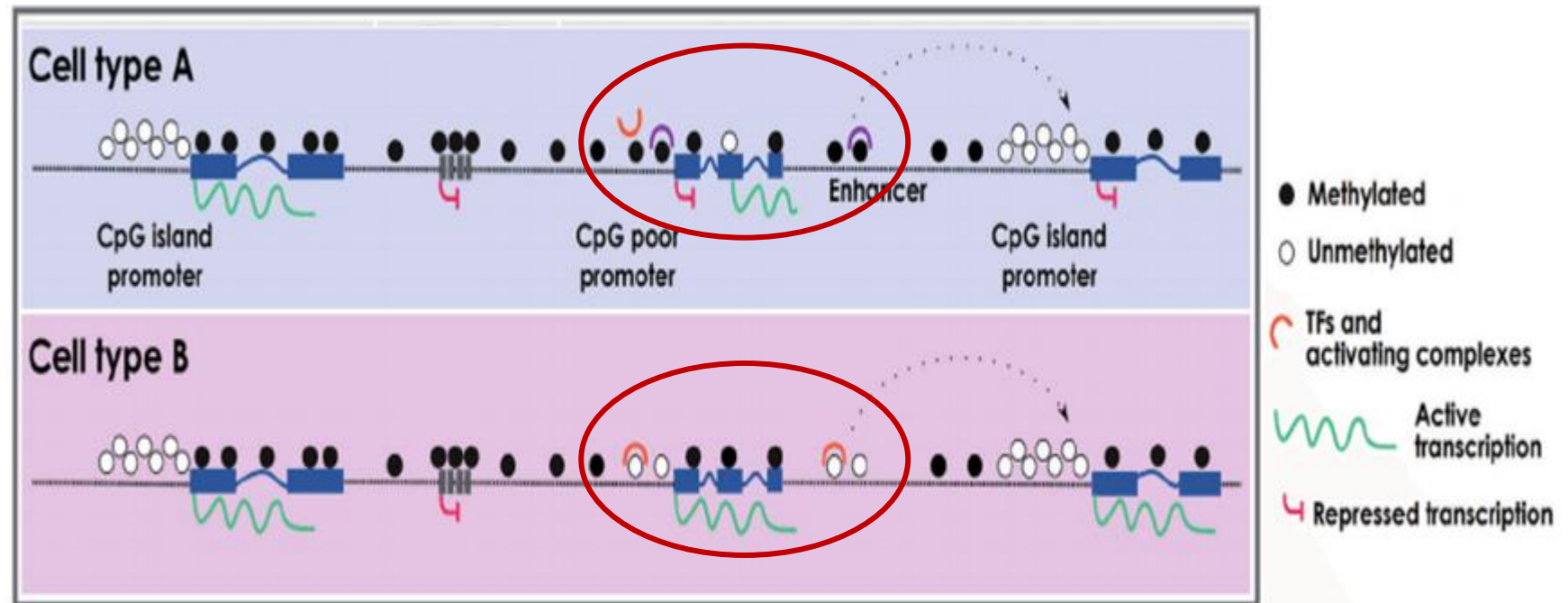
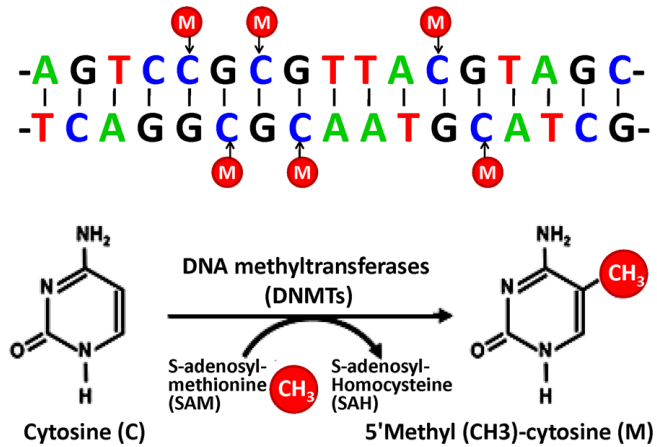
Division of Human Genetics | Polimanti Lab

**DEEP LEARNING
PREDICTS DNA METHYLATION REGULATORY VARIANTS IN
SPECIFIC BRAIN CELL TYPES AND ENHANCES FINE MAPPING
FOR BRAIN DISORDERS**

DNA methylation?

Elvira Carrió and Mònica Suelves., *Frontiers in Aging Neuroscience* (2015)

Wu and Yin ., *Diabetology & Metabolic Syndrome* (2022)



Background

Cell Specific approaches in methylation studies:

Expensive

Required larger sample sizes for reliable results

Linkages across the genome present a significant challenge

Deep learning models:

Does not need large samples

Prediction based on the interaction of variants with DNA motifs.

Isolates the impact of each variant (no LD confounding)

Deep Learning for Functional Insights

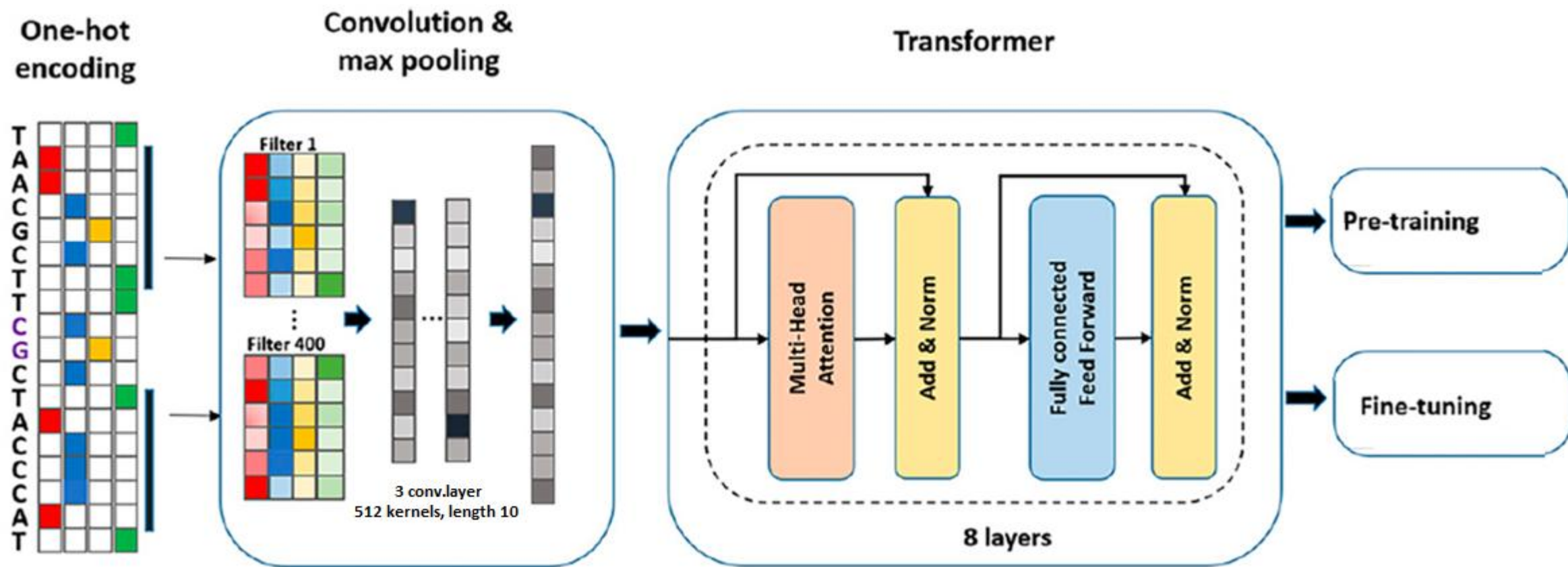
- GWAS reveals statistically significant variants associated with a specific trait but does not explain their functional impact.
- Experimental data for DNAm levels in all cell types for every GWAS variant is often unavailable.
- **Prediction models add a functional layer by estimating the tissue- or cell-specific effects of regulatory variants.**

INTERACT as a DEEP LEARNING MODEL

The INTERACT model combines two types of deep learning modules:

- **CNN (Convolutional Neural Network):**
 - Detects specific sequence features near each CpG site.
- **Transformer:**
 - Detect **distant relationships** in the DNA sequence.

Deep Learning : INTERACT Architecture



Two-Step Training Strategy

Pre-Training Phase

Dataset from the human prefrontal cortex was generated (13 brain cell types, 4,137 nuclei and 25 million CpG sites)

13 brain cell types included:

- **Excitatory neuron subtypes:** L2/3, L4, L5, and L6
- **Inhibitory neuron subtypes:** Ndnf, Vip, Pvalb, and Sst
- **Non-neuronal cell types:** Astrocytes, oligodendrocytes (ODC), oligodendrocyte progenitor cells (OPC), microglia, and endothelial cells.

*mDNA level= methylated reads/total reads

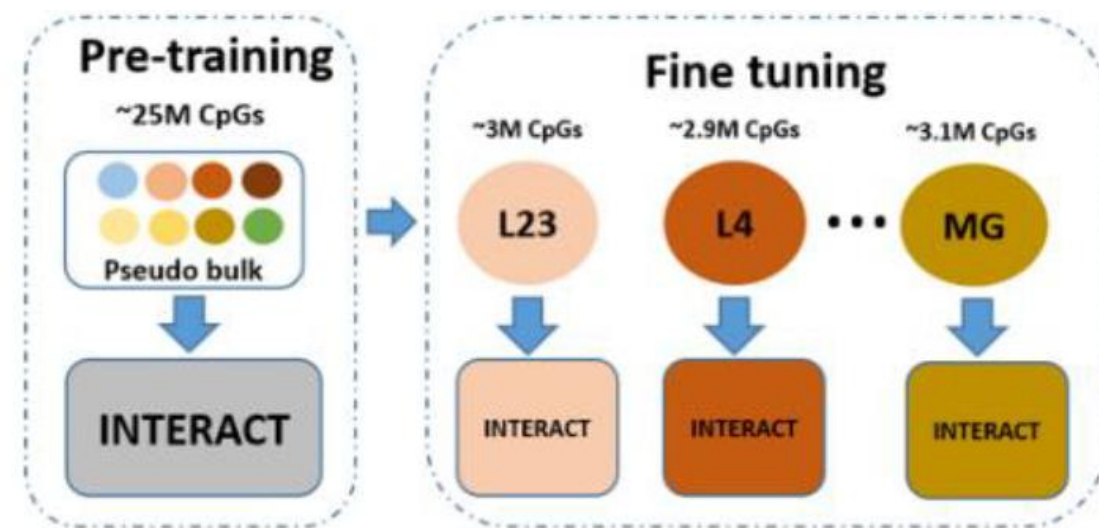
Cell type	# of nuclei	# of CpG sites (>30 reads)	# of CpG sites (>50 reads)
L23	551	20322957	9323696
L4	131	60199	21342
L5	180	273751	31984
L6	86	24812	12099
Ndnf	144	86476	23823
Pvalb	134	65577	21774
Sst	217	2096153	53075
Vip	171	407726	34907
Astro	449	19986213	8766345
MG	422	19861138	7908566
ODC	1244	25632294	24914907
OPC	203	2068845	53075
Endo	205	1786383	50399

Training Phase of INTERACT

Fine-tuning Phase

Cell type	# of nuclei	Coverage cutoff methylated	Coverage cutoff unmethylated	# of CpGs methylated	# of CpGs unmethylated	Total # of CpGs
L23	551	30	30	1879748	1134274	3,014,022
L4	131	10	8	1596875	1301346	2,898,221
L5	180	12	10	1839440	1296925	3,136,365
L6	86	7	5	1659919	1302156	2,962,075
Ndnf	144	11	8	1435657	938355	2,374,012
Pvalb	134	10	8	1829959	1158937	2,988,896
Sst	217	16	13	1493491	1223687	2,717,178
Vip	171	13	11	1640193	1022991	2,663,184
Astro	449	30	30	1639246	1143454	2,782,700
MG	422	25	28	1981637	1106193	3,087,830
ODC	1244	60	60	1637972	1495899	3,133,871
OPC	203	15	12	1610724	1343233	2,953,957
Endo	205	14	13	1587350	1135149	2,722,499

2.9 to 3.1 million CpG sites unique to each cell type are used.



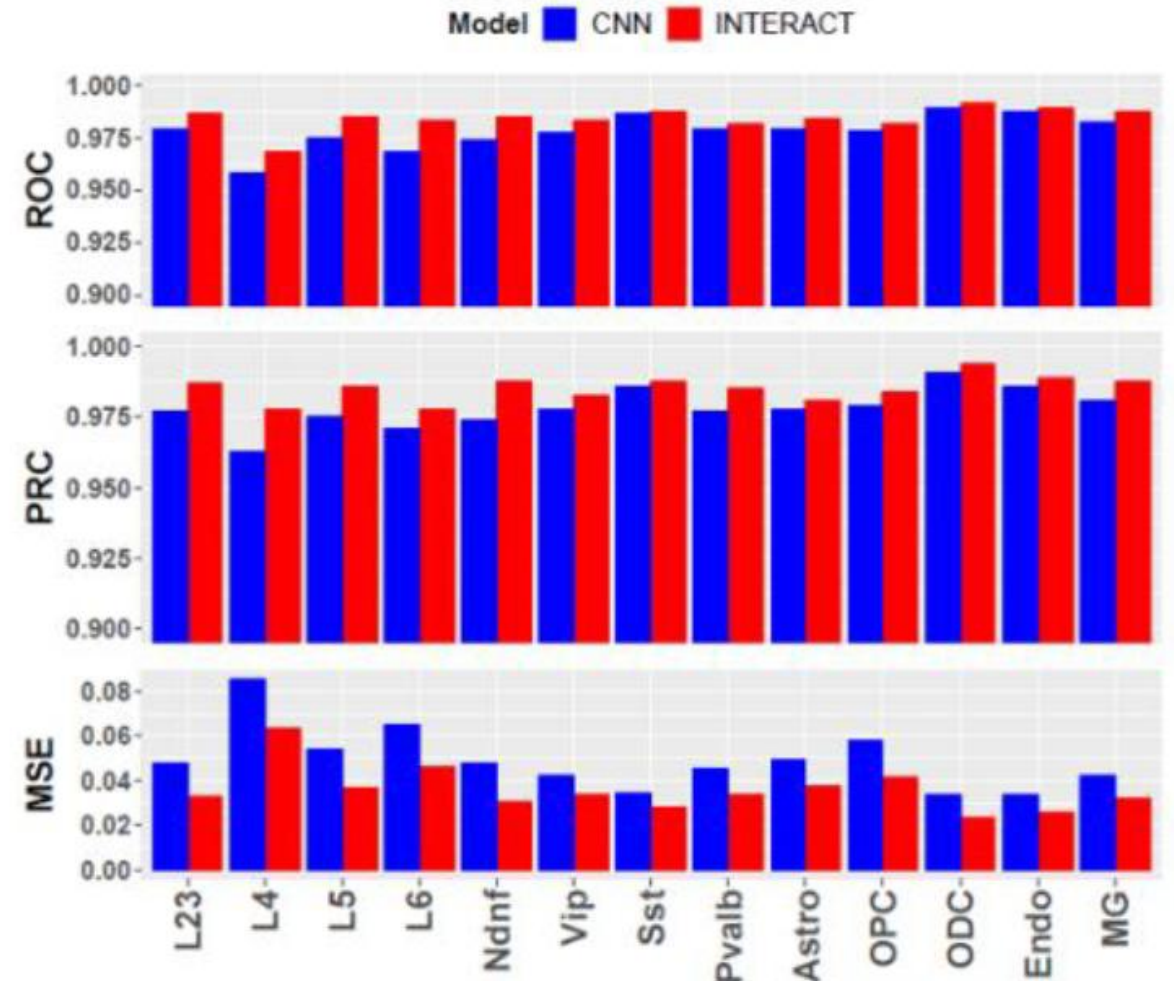
Cell-specific training samples. Coverage refers to the number of reads

Performance TEST for each cell type-specific INTERACT

Average Values: INTERACT vs. CNN
ROC: 0.984 vs. 0.978
PRC : 0.985 vs. 0.978
MSE: 0.0358 vs. 0.0491

Key Metrics:

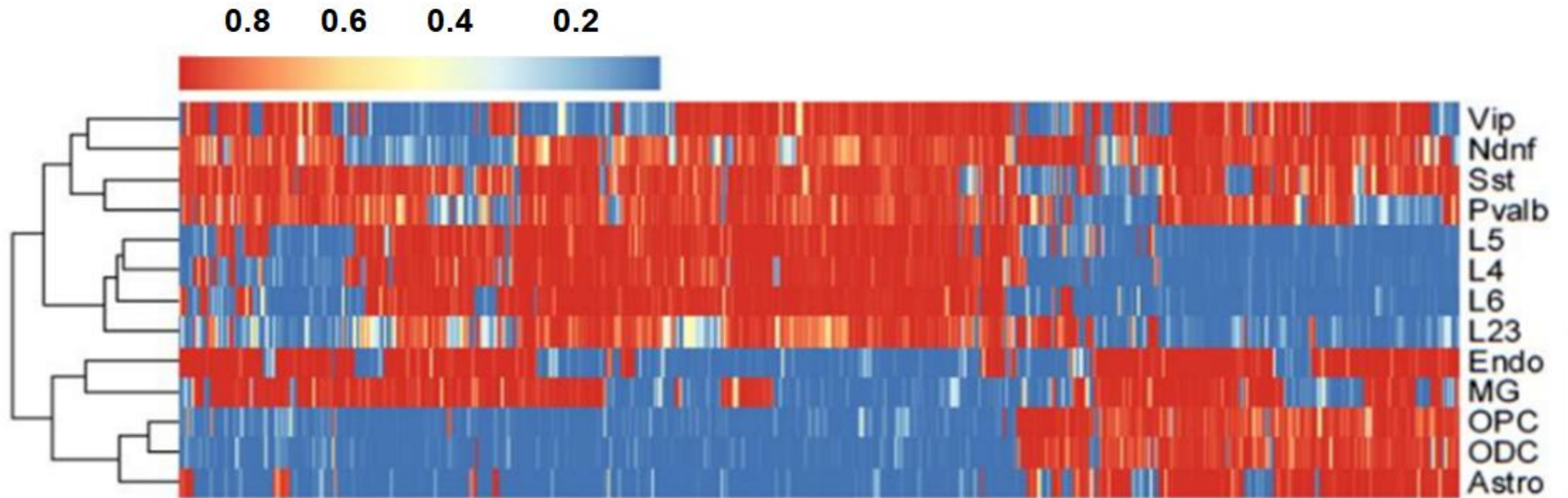
- ✓ **ROC (Receiver Operating Characteristic) Curve:** Measures the model's ability to distinguish between methylated and unmethylated CpG sites
- ✓ **PRC (Precision-Recall Curve):** Assesses the model's precision and recall actual methylated sites .
- ✓ **MSE (Mean Squared Error):** measure model accuracy in predicting DNAm levels.



Cell Type-Specificity of INTERACT

INTERACT model was tested for the most variable CpG sites on chromosome 22

Brain cells were clustered based on predicted DNAm levels, demonstrating a clear alignment with their biological context



*The color scale indicates the level of methylation
Red: high methylation, Blue: low methylation

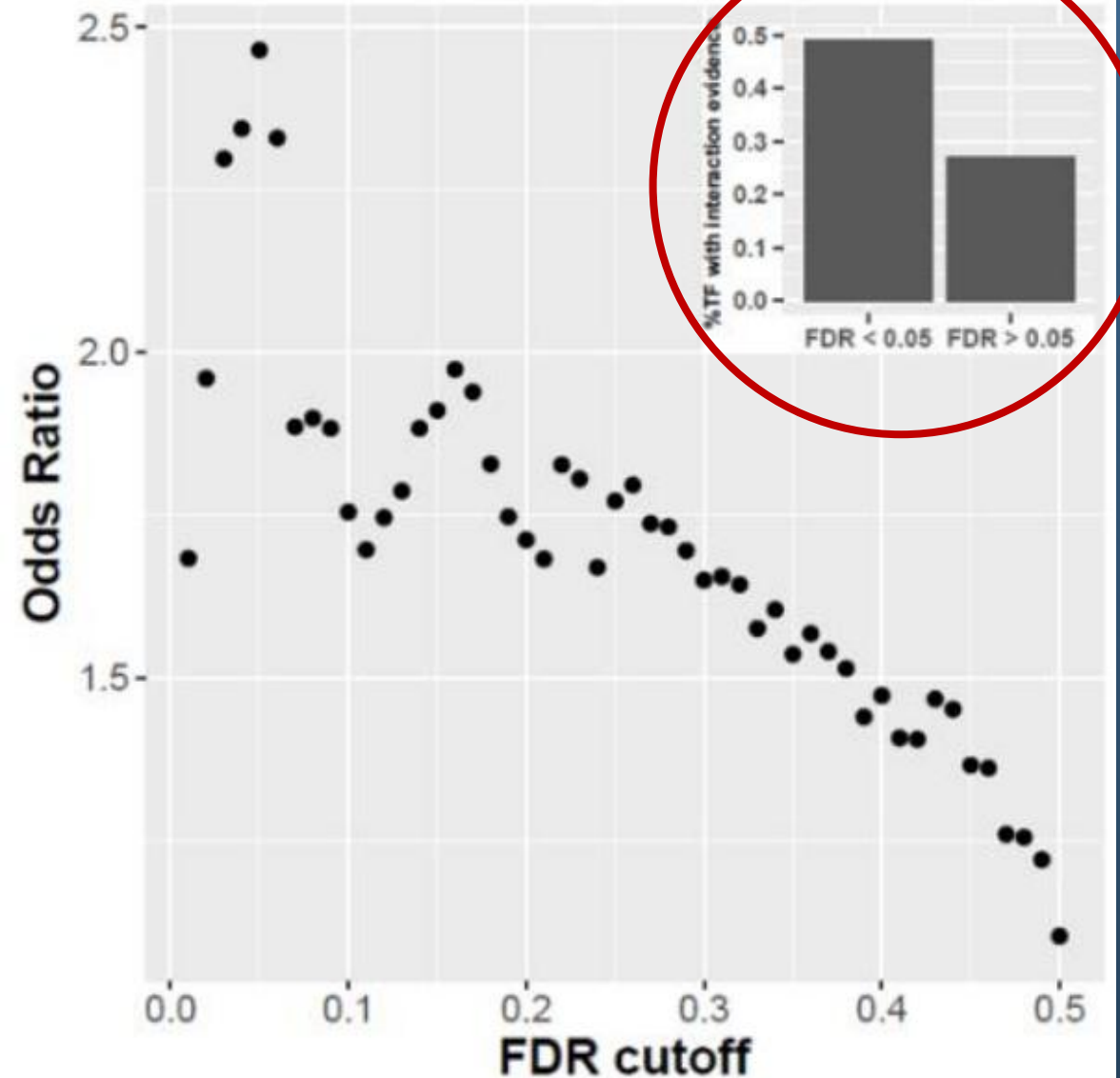
Functionality Assessment

Interact predicted DNA regulatory patterns matched with :

- 143 known TFs binding motifs (FDR < 0.05)
 - 45 match for the excitatory neuron (L23) and microglia
 - 73 for the inhibitory neuron (Sst),
 - an average of 58 TFs matches for each cell type.
- **69 TFs (48%)** were found to have **physical interaction evidence** (OR = 2.5, $p = 1.8 \times 10^{-6}$)

*Database and tools used:

Position Weight Matrix (PWM)
Homo sapiens CIS-BP database
Tomtom comparison tool
STRING database



Prediction of Brain Cell Type-Specific Regulatory Variants

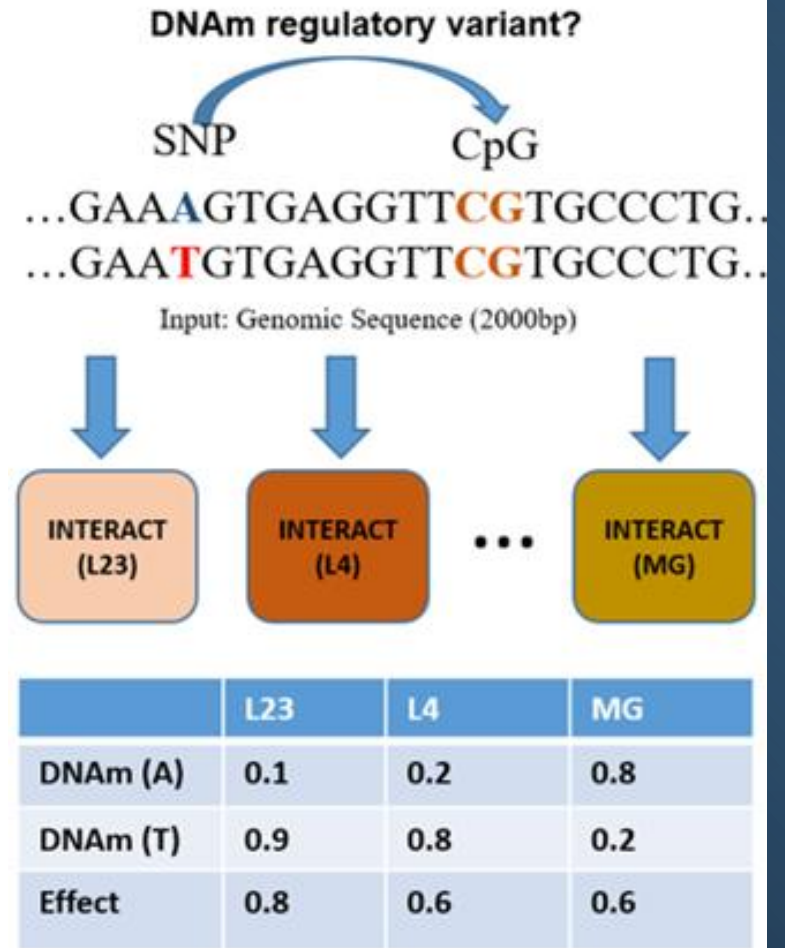
Data Source:

- 1000 Genomes Project, European ancestry

Included Variants:

- 9,042,066 variants (MAF > 0.005) located within 1 kb window of 25,564,506 autosomal CpG sites
- Predictions for **182,211,784 SNP–CpG pairs** across all cell types

In silico mutagenesis



Heritability enrichment analysis

18 Brain related traits
2 non-brain traits (height and type 2 diabetes) were included

DATA SOURCES

HapMap Project Phase 3 SNPs: Used as the reference set for regression SNPs

1000 Genomes Project EUR: Utilized as the reference panel

GWAS Summary Statistics: The traits and their respective data sources

Trait	Summary Statistics Download website
Schizophrenia	https://www.med.unc.edu/pgc/download-results/
Depression	https://datashare.ed.ac.uk/handle/10283/3203
Bipolar Disorder	https://www.med.unc.edu/pgc/download-results/
Autism	https://www.med.unc.edu/pgc/download-results/
ADHD	https://www.med.unc.edu/pgc/download-results/
Anorexia	https://www.med.unc.edu/pgc/download-results/
Smoking initiation	https://conservancy.umn.edu/handle/11299/241912
Smoking cessation	https://conservancy.umn.edu/handle/11299/241912
Cigarettes per day	https://conservancy.umn.edu/handle/11299/241912
Age of smoking	https://conservancy.umn.edu/handle/11299/241912
Drinks per week	https://conservancy.umn.edu/handle/11299/241912
Neuroticism	https://alkesgroup.broadinstitute.org/UKBB
Education Years	https://alkesgroup.broadinstitute.org/UKBB
Intelligence	https://ctg.cncr.nl/software/summary_statistics
BMI	https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
Parkinson Disease	https://drive.google.com/file/d/1FZ9UL99LAqyWnyNBxxIx6qOUIfAnubIN/view?usp=sharing
Epilepsy (genetic gen	https://www.ebi.ac.uk/gwas/
Alzheimer Disease	https://www.ebi.ac.uk/gwas/
Type 2 Diabetes	https://alkesgroup.broadinstitute.org/UKBB
Height	https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files

METHOD

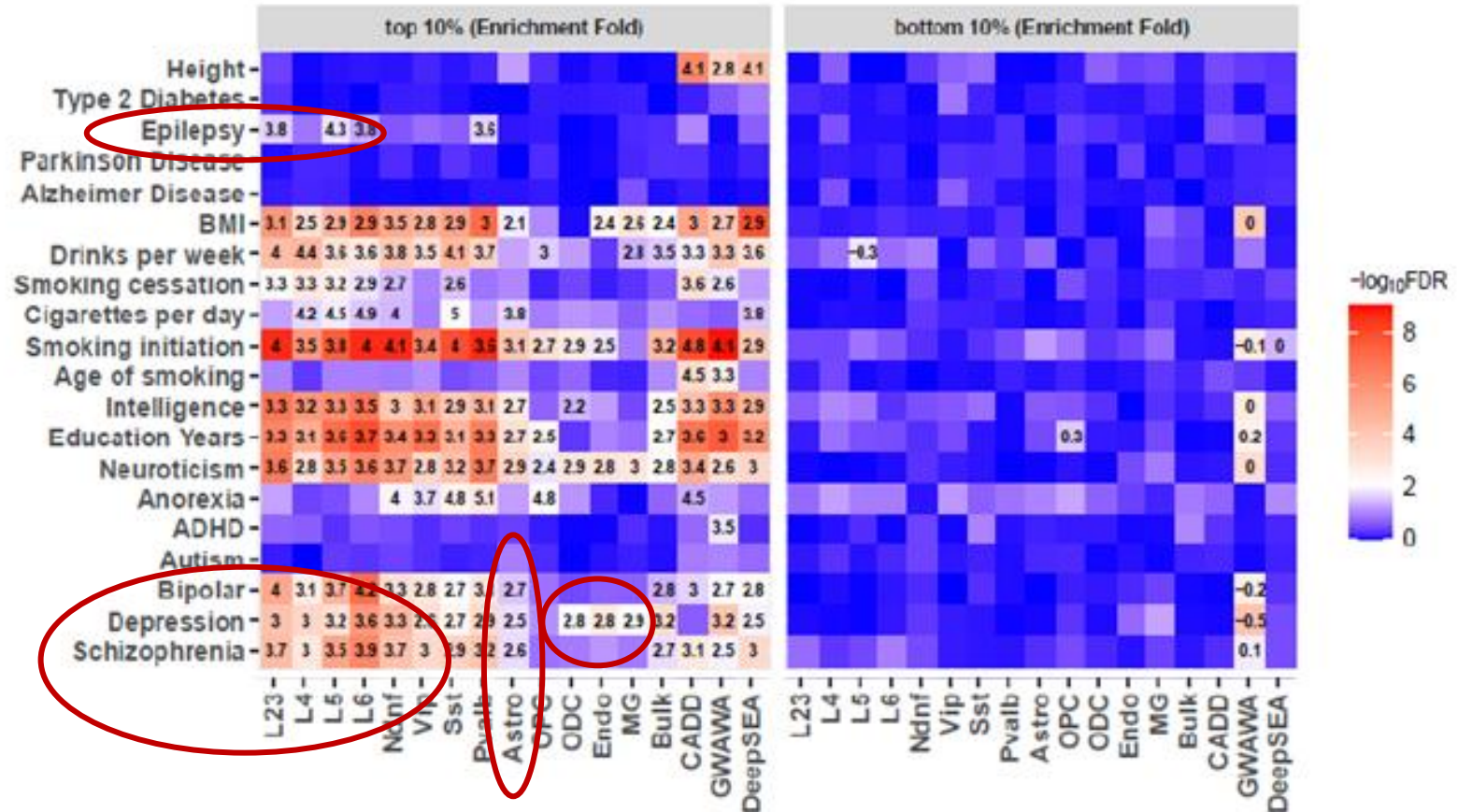
LDSC resource: website (<https://alkesgroup.broadinstitute.org/LDSCORE>)

Baseline LD model v2.2 with 97 pre-defined annotations

Heritability Enrichment Analysis

Strong enrichment was observed for brain-related **disorders**,

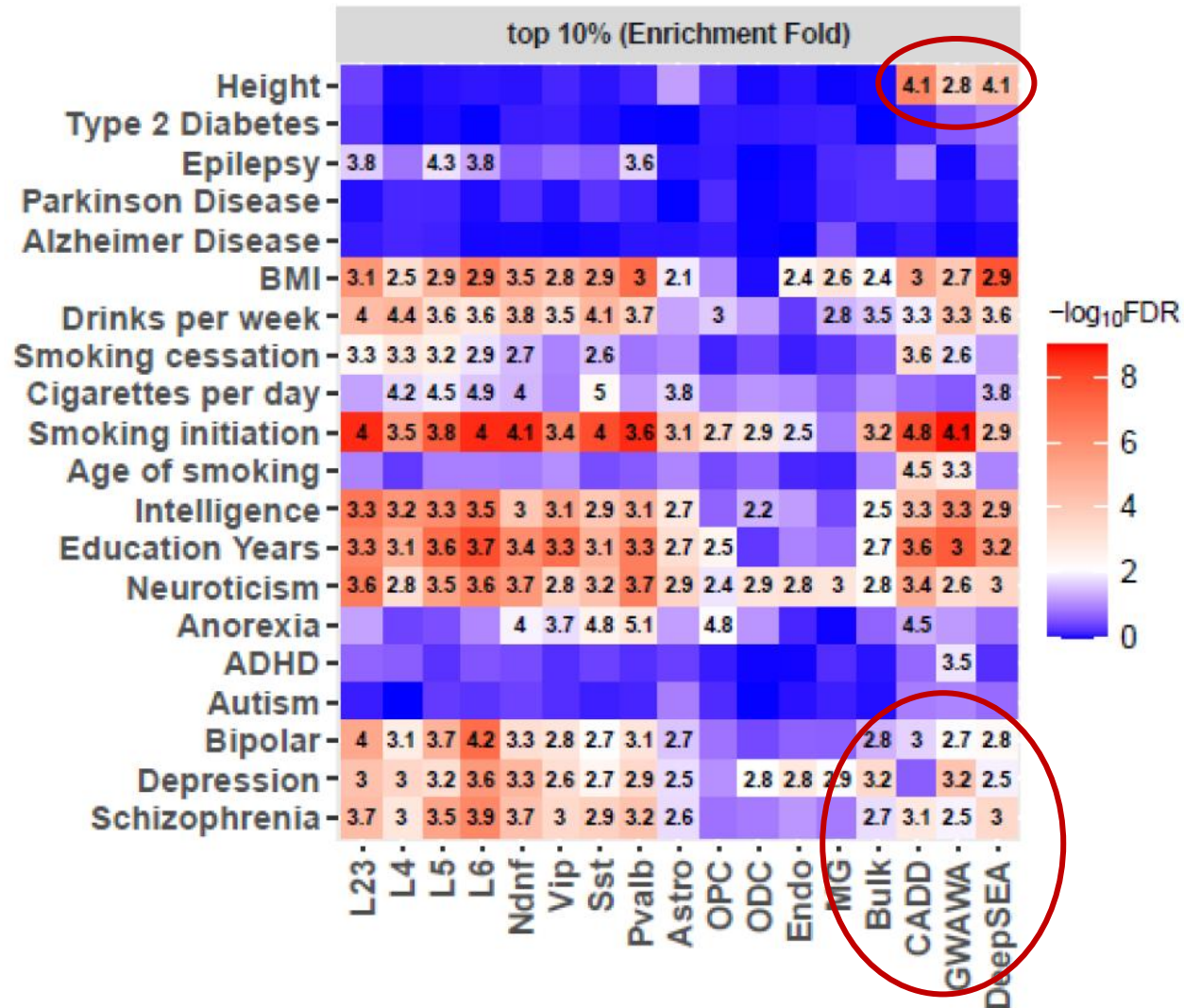
- Schizophrenia
- Bipolar disorder
- Depression
- Epilepsy



Heritability Enrichment Analysis

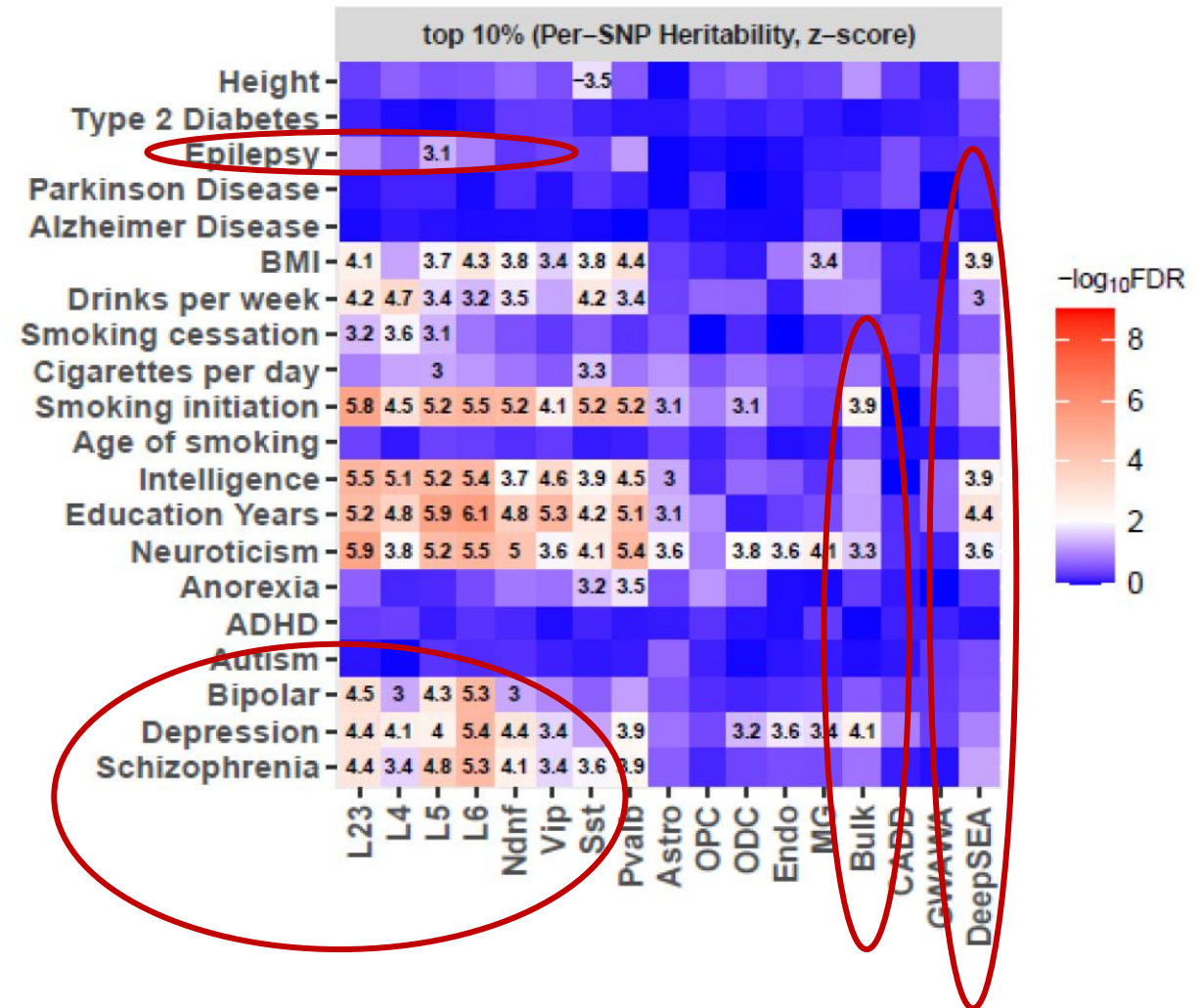
Other Scoring Systems

- CADD (Combined Annotation Dependent Depletion)
- GWAWA (Genome-Wide Annotation of Variants)
- DeepSEA (Deep Learning for Sequence-based Prediction of Variant Effects)
- **BulkBrain-tissue-specific deep learning model**



Heritability Enrichment Analysis

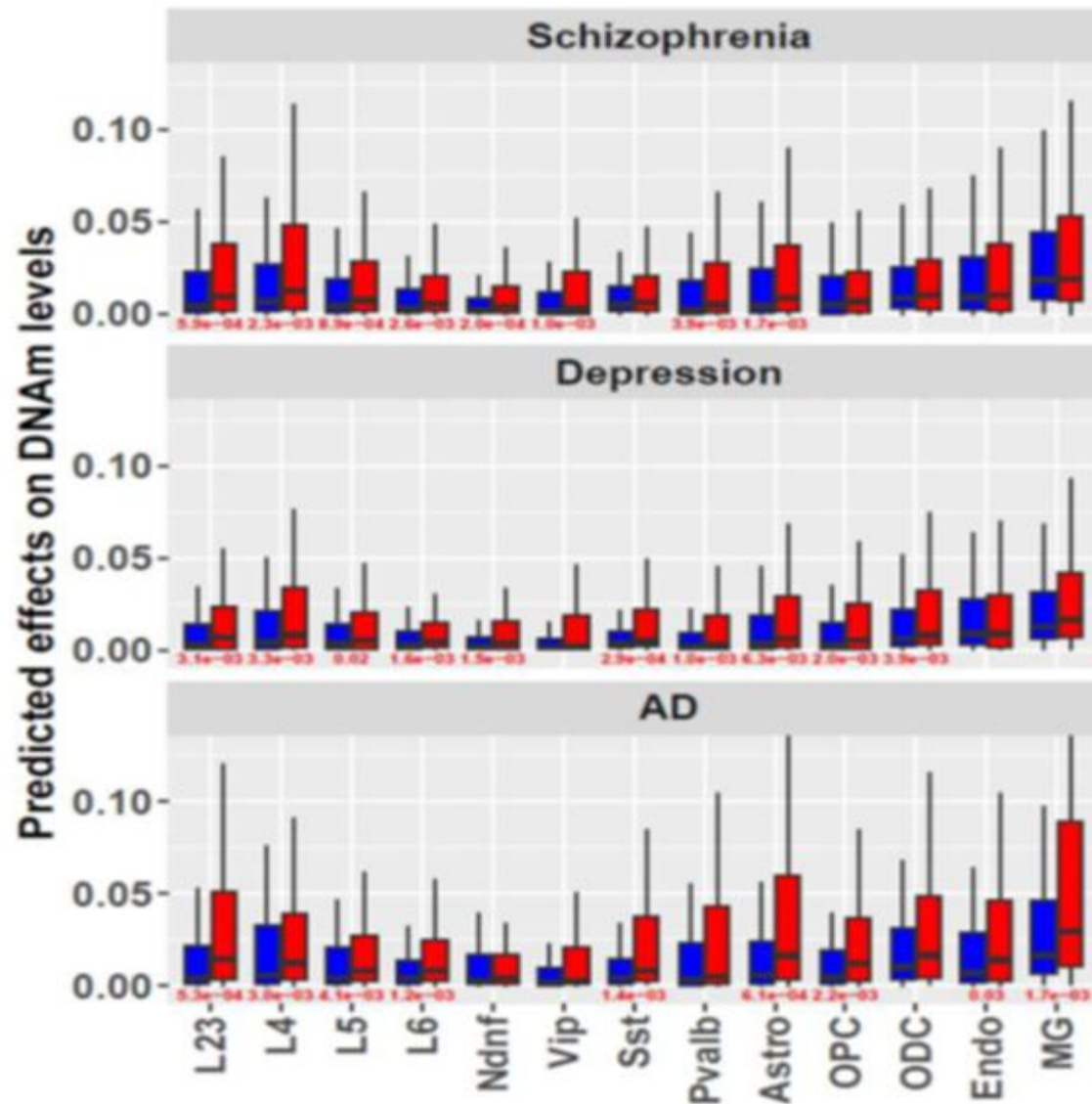
- Bulk Brain tissue-specific INTERACT predictions missed important cell-type-specific signals
- Significant contributions to depression, neuroticism, and smoking initiation were observed



S-LDSC:z-score of per-SNP heritability (FDR < 0.05)

INTERACT Enhance Fine Mapping GWAS risk Loci

- GWAS summary statistics of European ancestry
LD clumping:
 - Index SNPs were identified ($p < 5 \times 10^{-8}$, $r^2 < 0.1$, 3-Mb)
 - Risk loci for each index SNP were defined as regions with $r^2 > 0.2$, extending 50-kb on each sides within the 3-Mb window.
 - Overlapping regions were combined into a single risk locus.
- A total of 177 loci for schizophrenia, 100 loci for depression, and 54 loci for Alzheimer's disease (AD) were included.
- **Fine mapping GWAS loci include:**
 - Causal variants with high posterior inclusion probability (PIP > 0.1) and GWAS significance ($p < 1 \times 10^{-6}$)
 - Control variants with low PIP (PIP $< 1 \times 10^{-4}$) and weak GWAS evidence ($p > 0.99$)

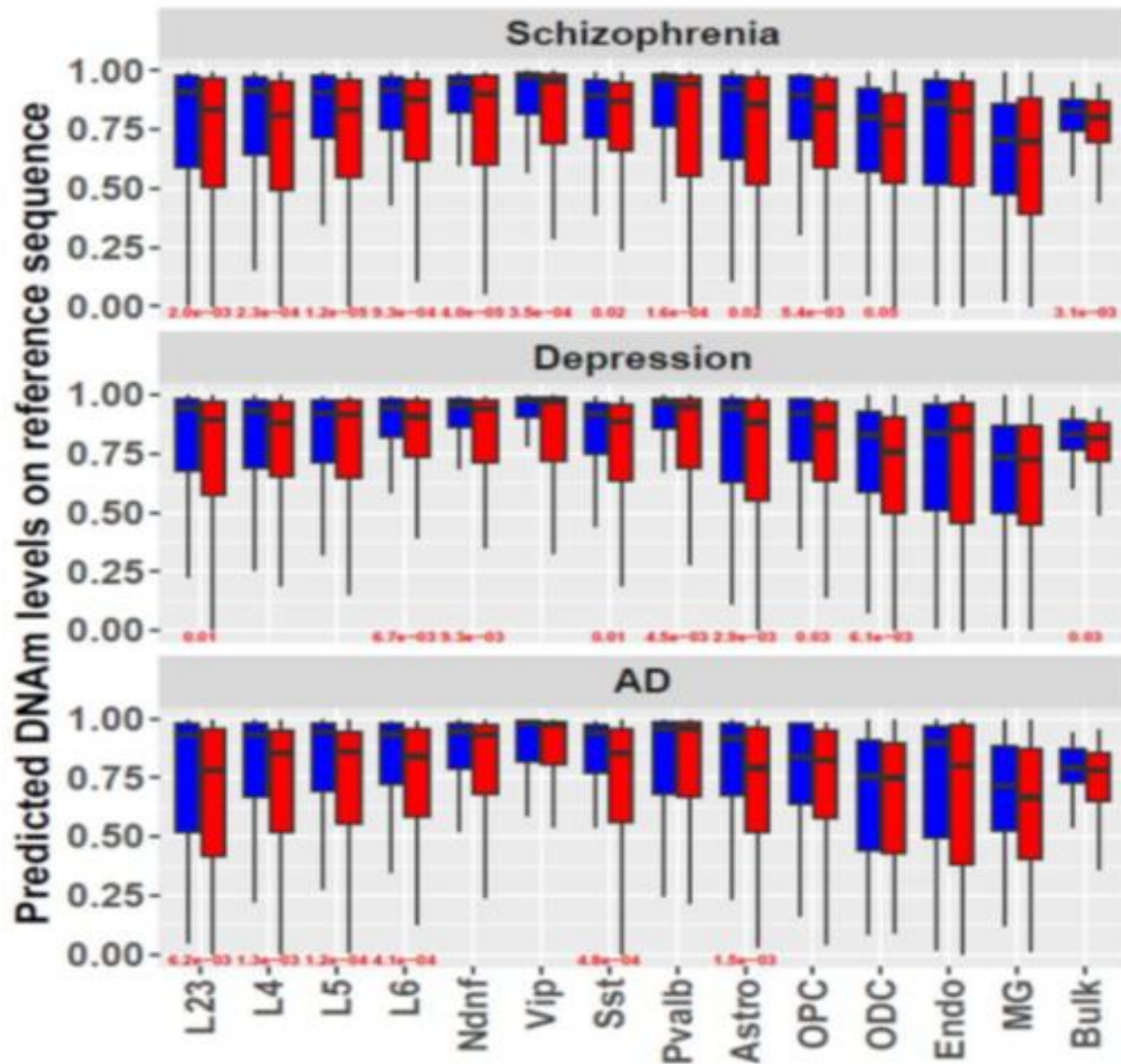


Causal variants have larger effects on DNAm levels:

For schizophrenia: in neuronal cells (except **Sst**) and astrocytes

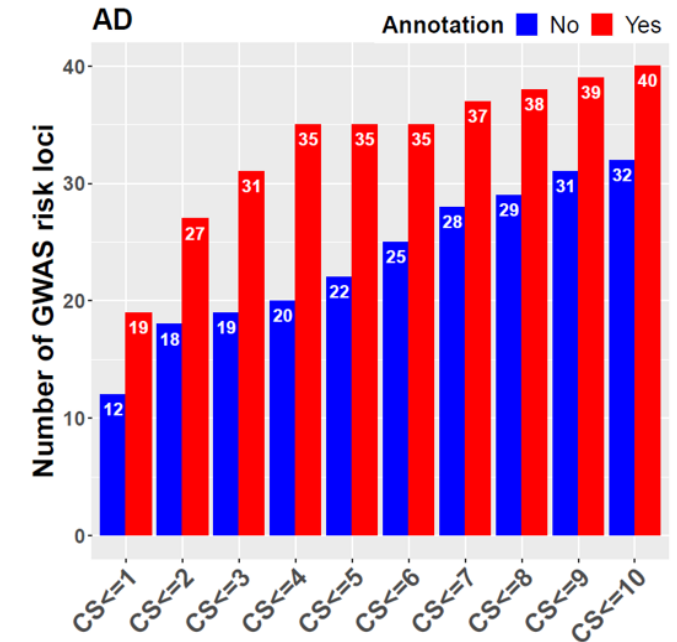
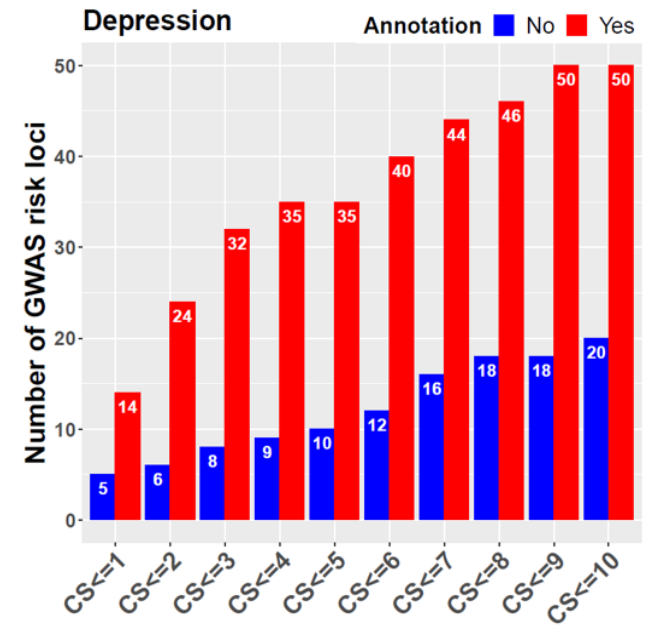
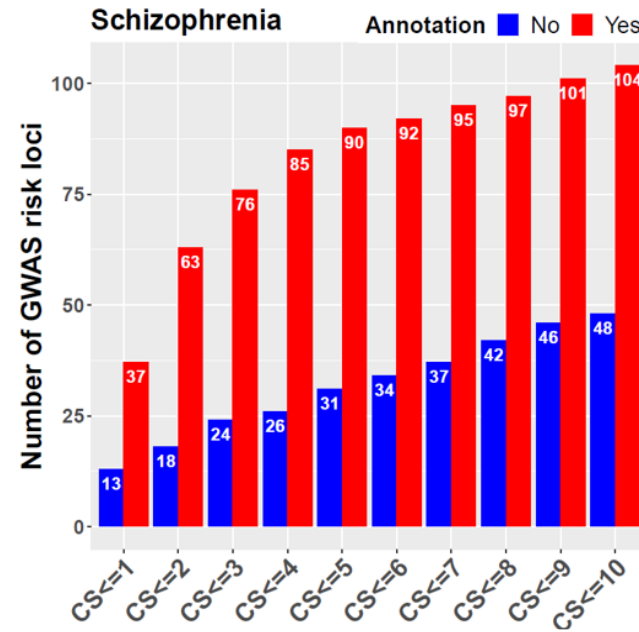
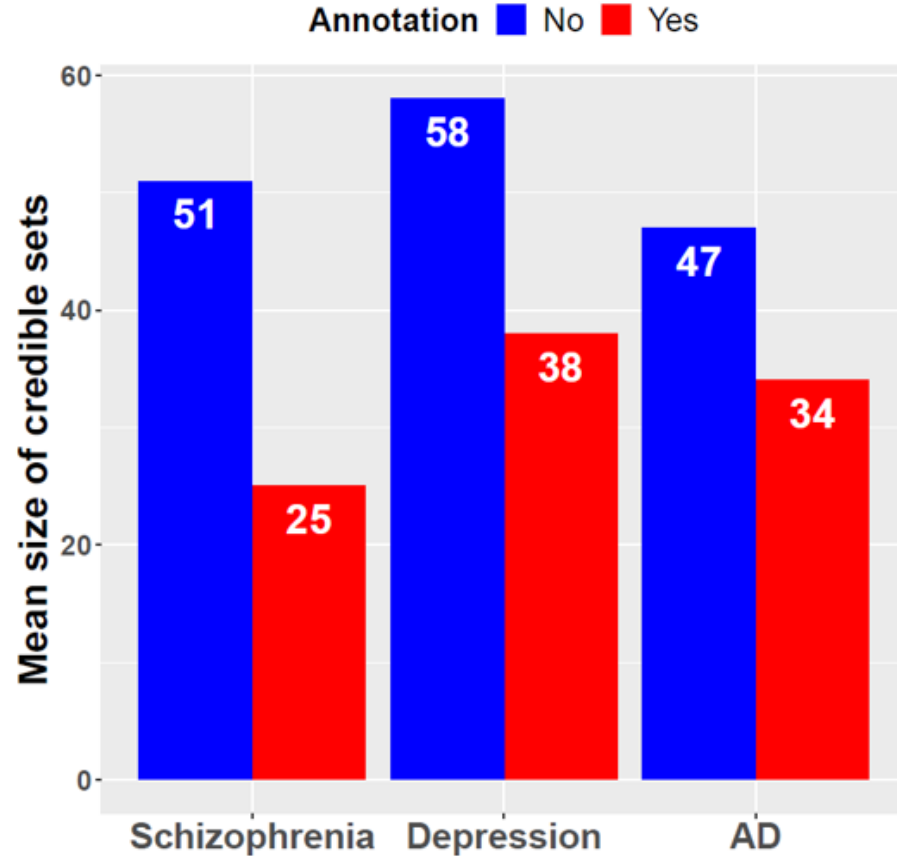
For depression: in **most neuronal cells** (except **Vip**) and glial cell (astrocytes, OPCs and ODCs)

For Alzheimer's Disease: in neuronal cells (except **Vip** and **Ndnf**), and in glial cells (except ODCs). The most pronounced impact is observed in microglia, consistent with their recognized role in AD.



CpG sites impacted by causal variants are located in regions with **low baseline DNAm levels**, likely to function as active regulatory elements.

Fine-Mapping with Annotation



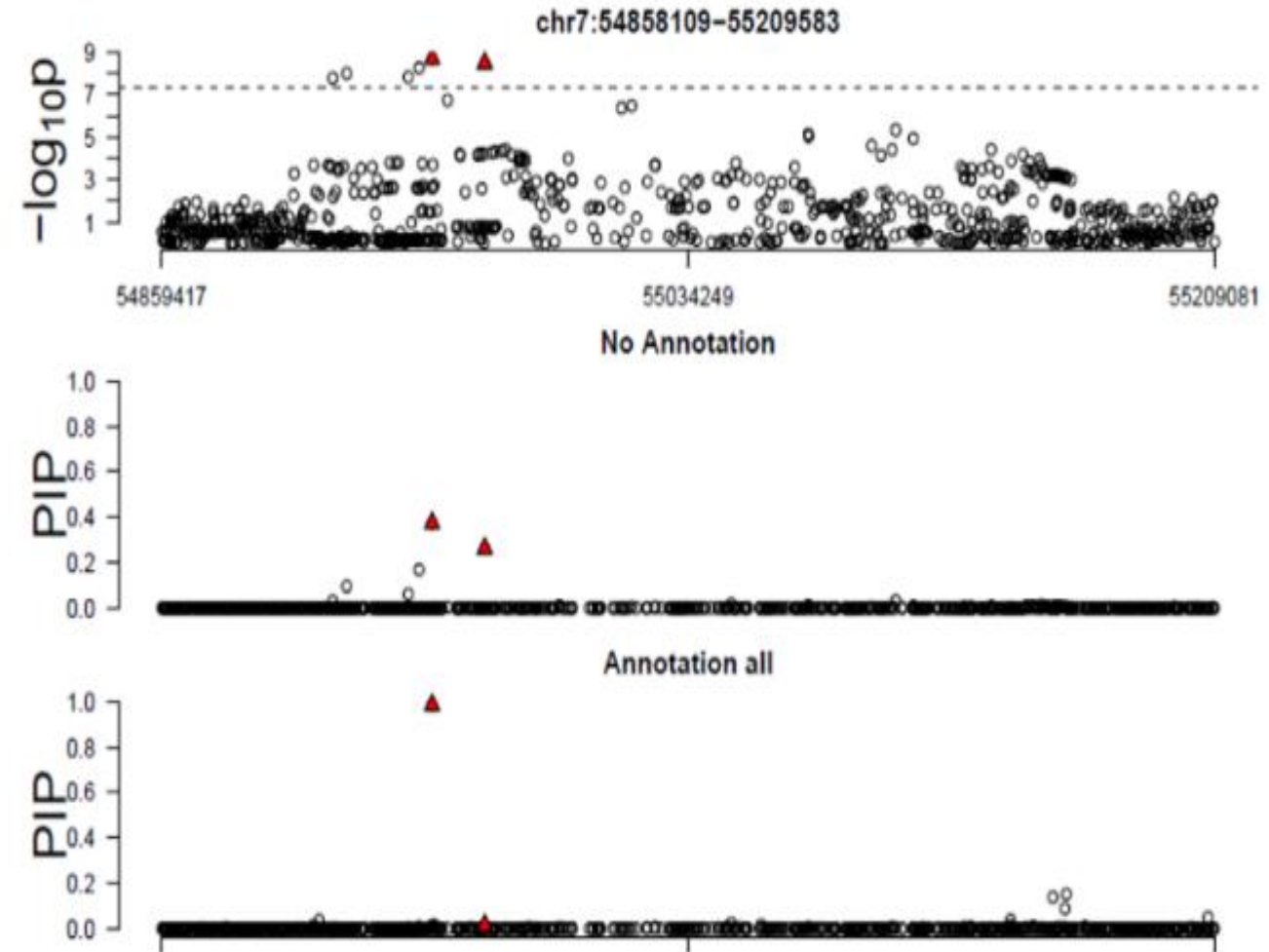
Red Bars: Fine mapping **with** functional annotations
Blue Bars: Fine mapping **without** functional annotations

Fine Mapping AD Locus

GWAS Signal ($-\log_{10}$ P Values), where multiple SNPs show significant signals

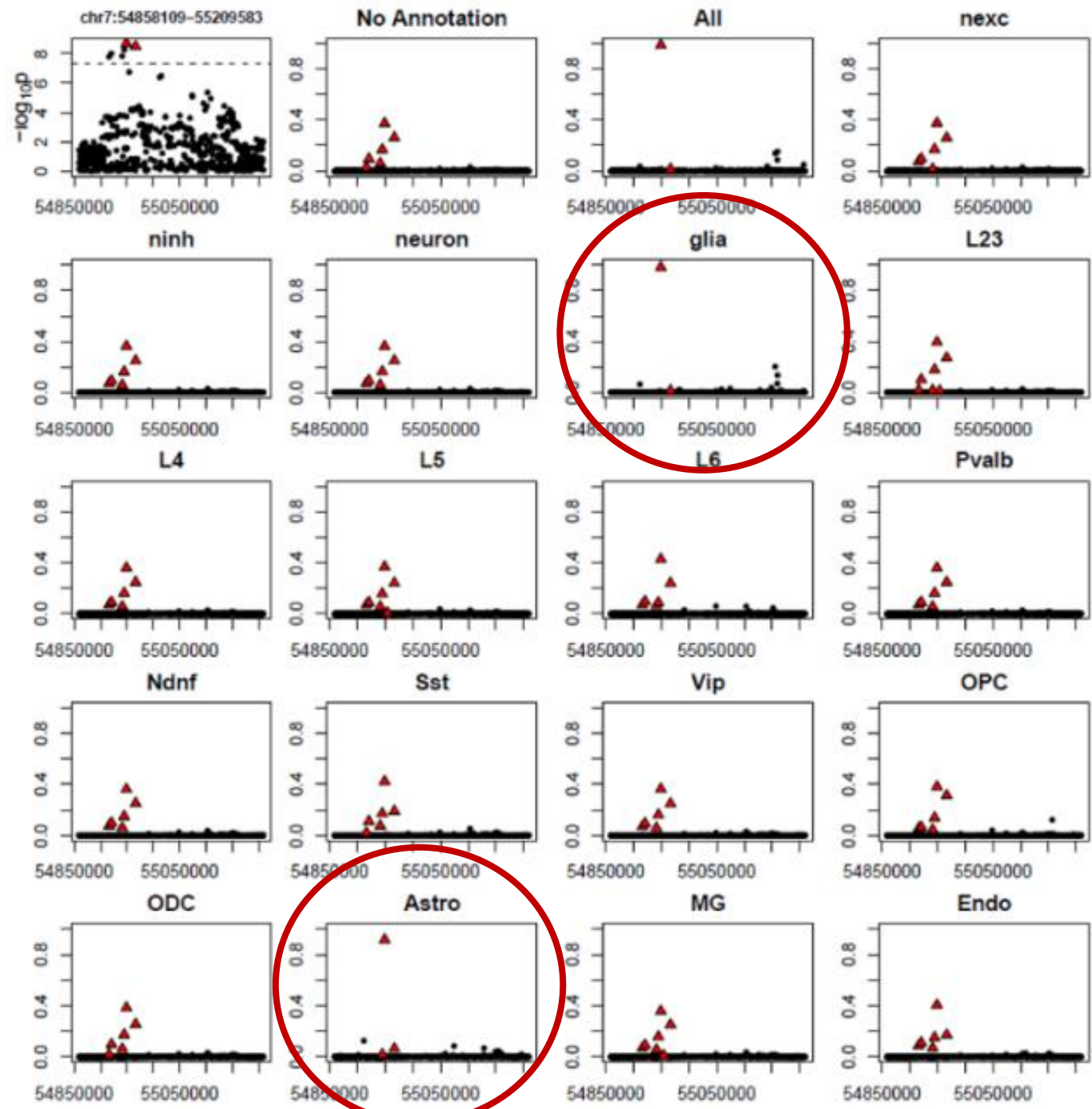
No annotations: a credible set of 6 SNPs, with rs74504435 having the highest PIP (0.37)

With annotations: a credible set of 2 SNPs, with rs74504435's PIP increasing to 0.986

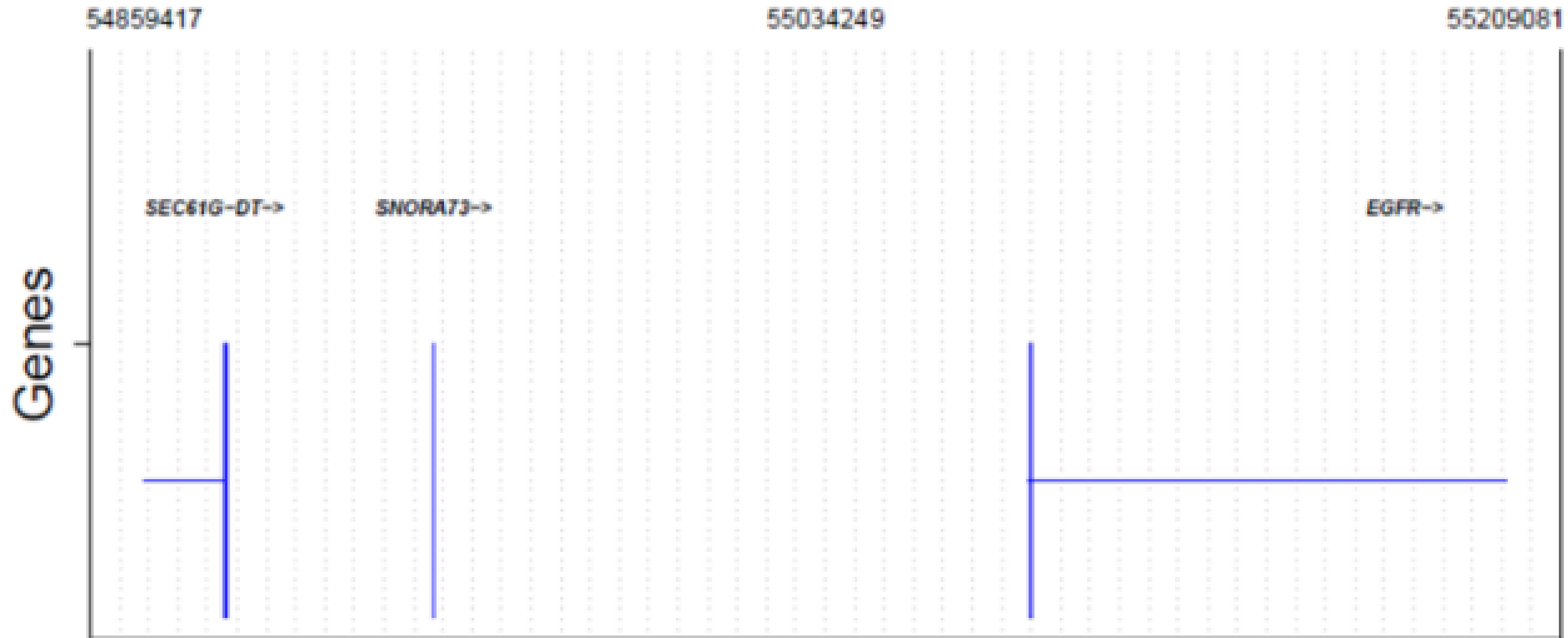


Cell Specific Annotation

Rs74504435 shows strongest signal: in **glia** (PIP = 0.981), particularly in **astrocytes** (PIP = 0.917).



Assign fine-mapped SNPs to target genes



Hi-C chromatin contact data reveals the **3D interactions** between regulatory SNP (rs74504435) and target genes (**EGFR**) through physical proximity in the nucleus.

Key Discussion Points

The **overall findings** highlight the power of deep learning in identifying cell-specific regulatory variants, enhancing our understanding of the biological mechanisms underlying disease risk and gene regulation.

Datasets Utilized in the Study

- **Pseudo-Bulk DNAm Dataset:** 25 million CpG sites aggregated from all 4,137 nuclei, used for Pre-training of INTERACT
- **Single-Nucleus DNAm Dataset (sn-m3C-seq):** Used to train cell type-specific INTERACT models. Includes 4,137 nuclei from the human prefrontal cortex, clustered into 13 cell types.
- **UKBiobank (UKBB):** Reference LD was estimated from 50,000 randomly selected samples of European ancestry in the UKBB dataset.
- **1000 Genomes Project:** Assessing in silico effects of SNPs on DNAm levels. Reference SNPs for calculating LD in S-LDSC.
- **GWAS Summary Statistics:** Publicly available datasets for brain-related traits: Heritability enrichment analysis and Fine mapping of risk loci for brain disorders.
- **6. SNP Functional Scoring Systems:** CADD, GWAWA, DeepSEA used for trait heritability enrichment.
- **7. GENCODE Version 42 Basic Gene Annotations:** Identifies transcript promoters to link fine-mapped SNPs to target genes.
- **8. Hi-C Chromatin Contact Data:** Link enhancers and promoters for regulatory activity analysis, assign SNPs to genes in fine mapping.

Key Discussion Points

Major Findings

- INTERACT successfully predicts DNA methylation (DNAm) regulatory variants in specific brain cell types with high accuracy (AUC = 0.98).
- Strong enrichment of predicted variants for brain-related traits, particularly in relevant cell types like neurons and microglia.
- Incorporation of predicted variants enhances fine mapping for brain disorders

Key Discussion Points

Biological Insights

- Cell type-specific predictions align with known biological relationships
- Identified DNA motifs and transcription factors (TFs) underlying cell-specific DNAm patterns provide insight into molecular mechanisms.
- Observed stronger enrichment of causal variants in regulatory regions with lower DNAm levels, confirming their regulatory role.

Key Discussion Points

Advantages

- Outperforms alternative scoring systems (CADD, GWAWA, DeepSEA) in detecting variants enriched for brain trait heritability.
- A context-specific approach captures cell type-specific contributions to complex traits.
- Applicable to large datasets and potentially extensible to other molecular traits.

Key Discussion Points

Study Limitations

- **Small Sample Size:** The dataset included only 4,137 nuclei
- **Limited Cell Types:** The study focused on 13 brain cell types
- **Focus on DNA Methylation:** Missed other important mechanisms like chromatin changes
- **Weak GWAS Data for Some Traits:** For some traits like Alzheimer's, the available genetic data wasn't strong
- **Dependency on Existing Annotations:** The study relied on known enhancer and promoter annotations, which might not cover all important regulatory regions.