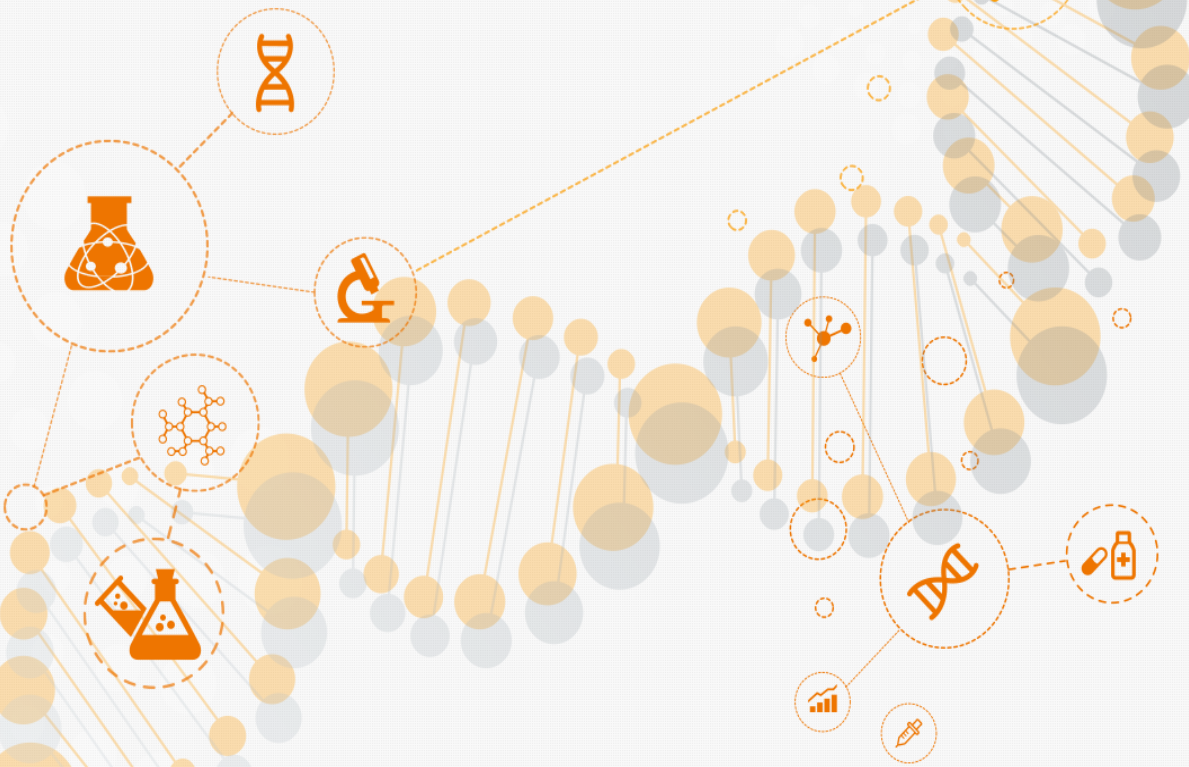
Convolutional neural networks predict causal regulatory variants

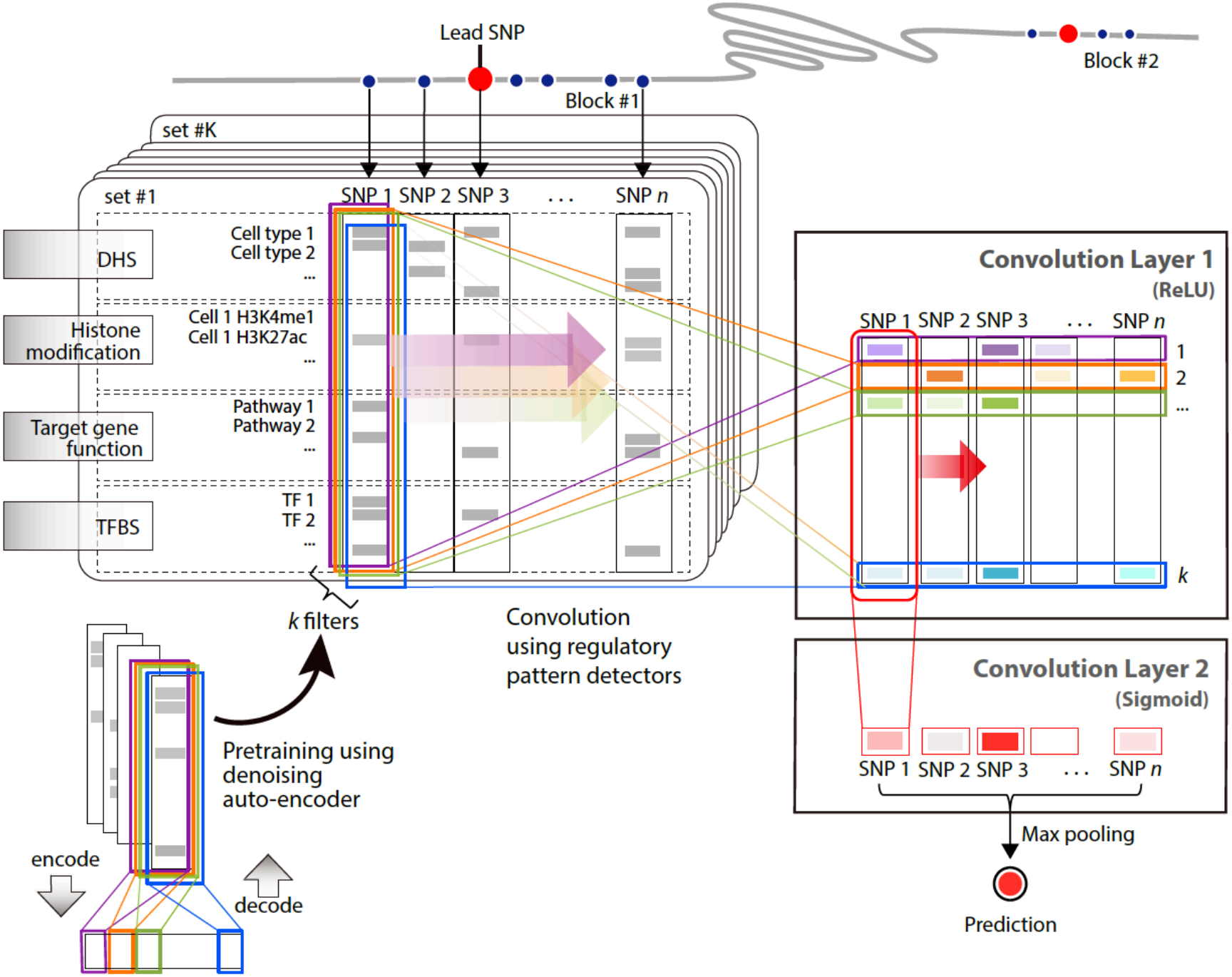
Jung Kyoon Choi, PhD



KAIST Dept. of Bio and Brain Engineering

Major challenges in GWAS

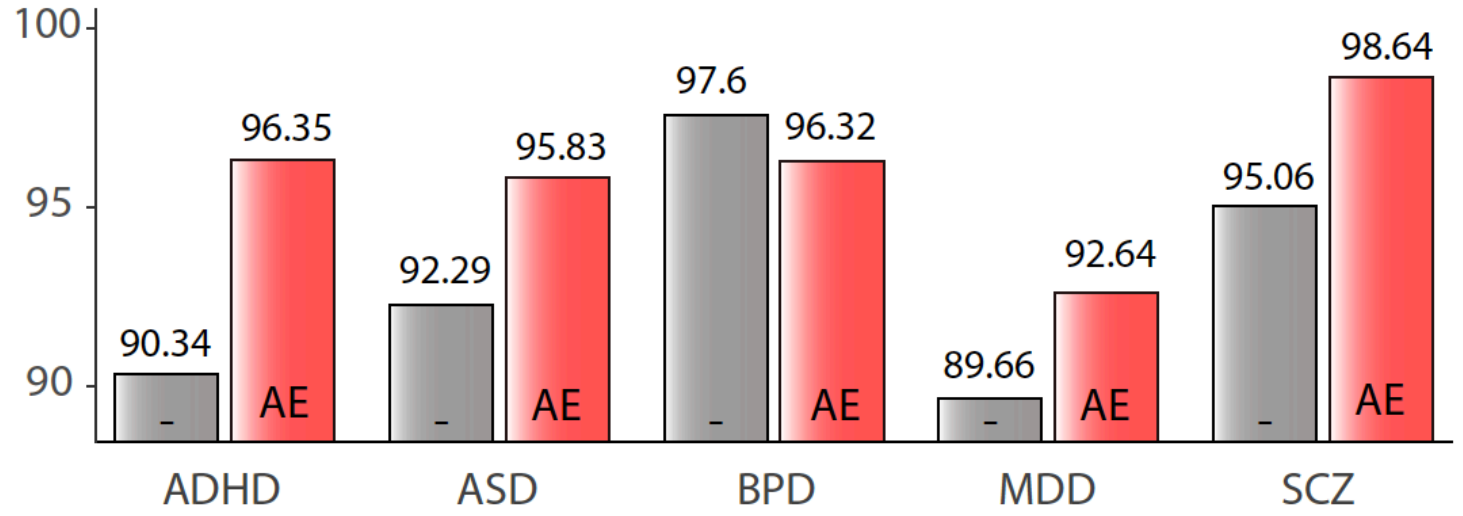
- GWASs can only report large clusters of SNPs (LD blocks) including not only causal variants but also many liked neutral SNPs
- Statistical approaches for fine mapping are not applicable for rare variants
- Majority of disease-associated DNA variants are thought to alter not the gene itself but the regulatory elements
- Our incomplete knowledge of noncoding regions limits the functional interpretation of underlying variants
- The wealth of cell-type-specific human epigenomes help with the identification of functional noncoding variants


Lead SNP
Block #1
Block #2
set #K
set #1
SNP 1 SNP 2 SNP 3 . . . SNP n
DHS
Cell type 1
Cell type 2
...
Histone modification
Cell 1 H3K4me1
Cell 1 H3K27ac
...
Target gene function
Pathway 1
Pathway 2
...
TFBS
TF 1
TF 2
...
k filters
Convolution using regulatory pattern detectors
Convolution Layer 1
(ReLU)
SNP 1 SNP 2 SNP 3 . . . SNP n
1
2
...
k
Convolution Layer 2
(Sigmoid)
SNP 1 SNP 2 SNP 3 . . . SNP n
Max pooling
Prediction
Pretraining using denoising auto-encoder
encode
decode

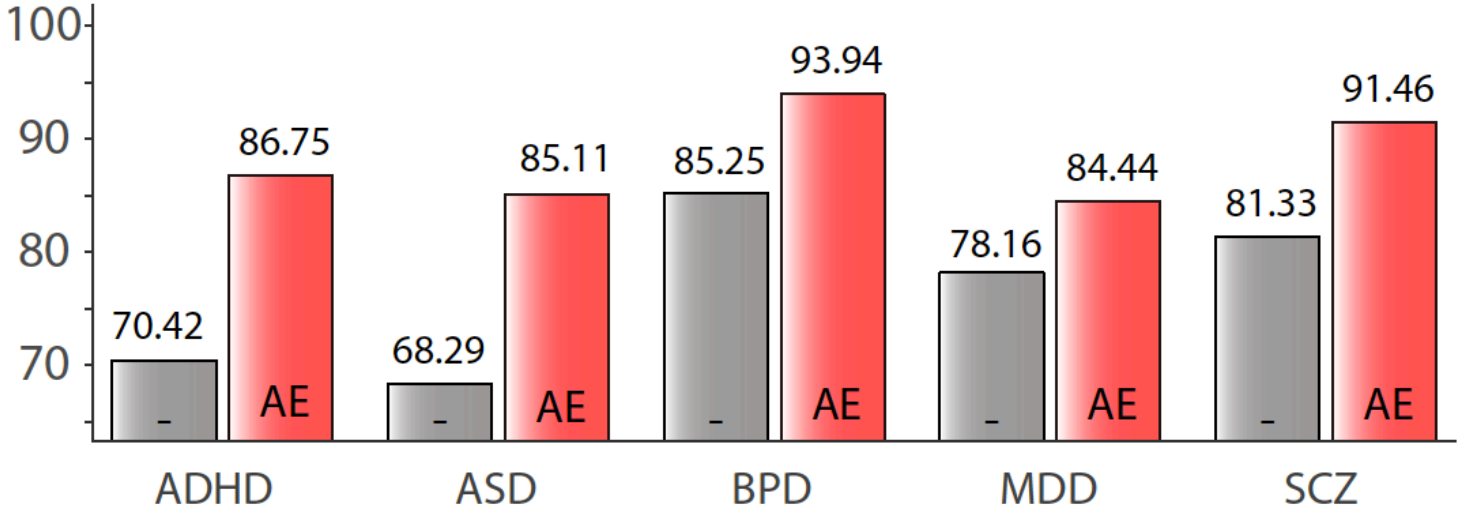
GWAS data by the Psychiatric Genomics Consortium

	GWAS cohort size				
	ADHD	**ASD**	**BPD**	**MDD**	**SCZ**
Case	2,787	4,949	6,990	9,227	9,379
Control	2,635	5,314	4,820	7,383	7,736

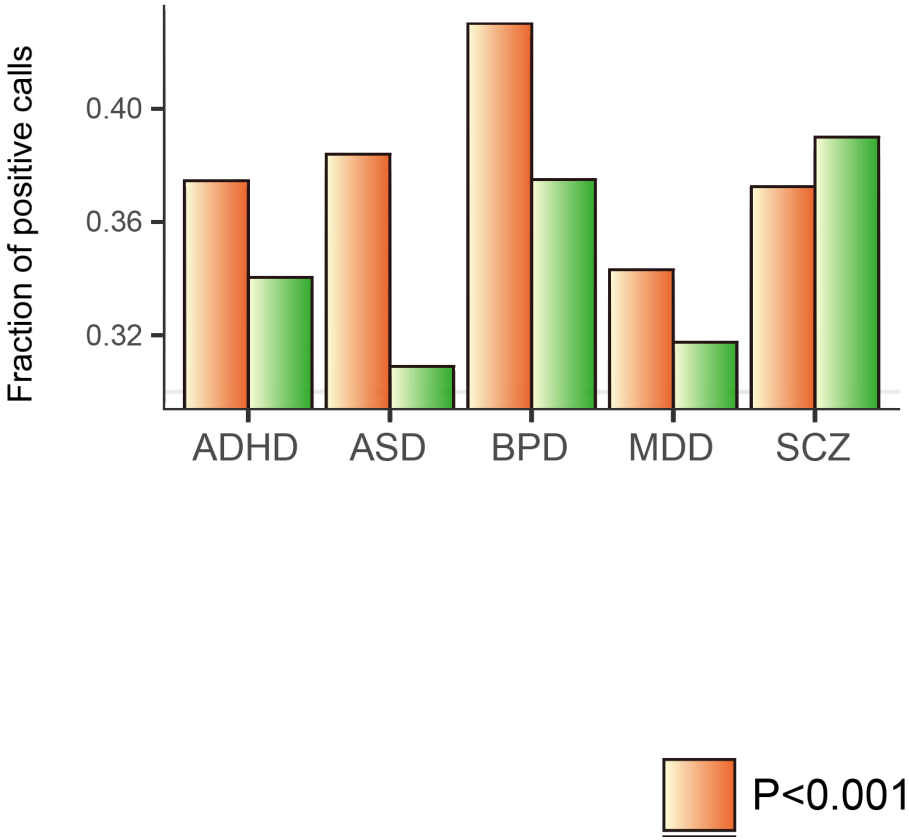
	Number of association blocks				
	ADHD	**ASD**	**BPD**	**MDD**	**SCZ**
Training set (chr 1-10)	227	253	292	267	384
Validation set (chr 15-22)	67	84	113	87	129
Test set (chr 11-14)	46	54	69	51	88
At least one positive call	285	333	430	347	536

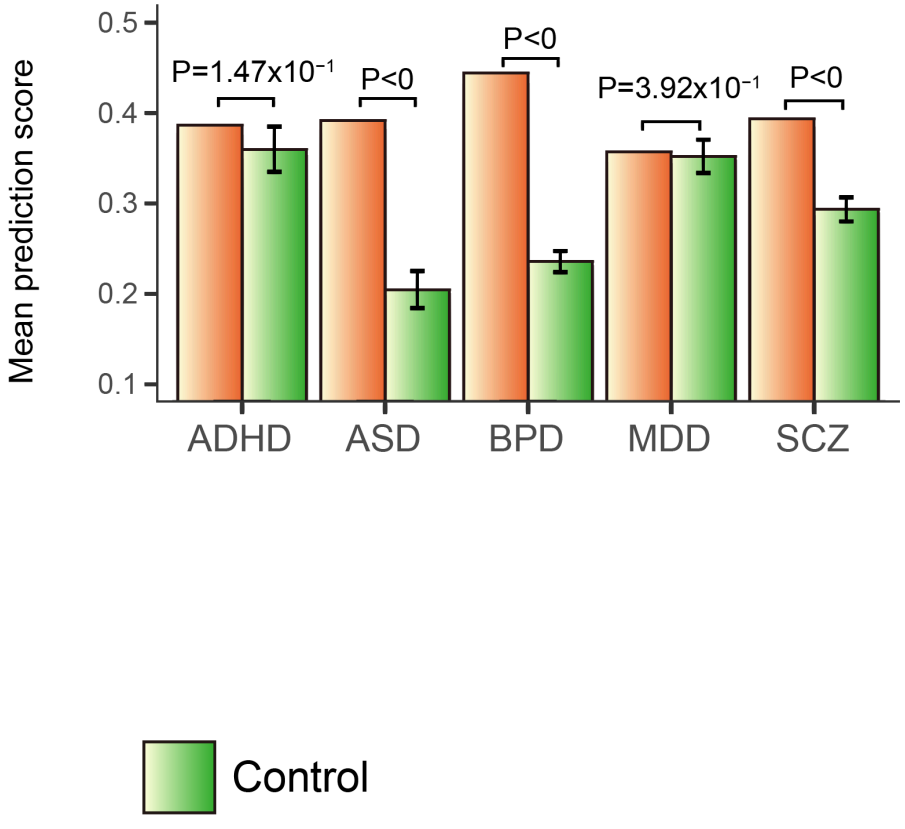

100
95
90
90.34
96.35
92.29
95.83
97.6
96.32
89.66
92.64
95.06
98.64
-
AE
ADHD
ASD
BPD
MDD
SCZ

AUC


100
90
80
70
70.42
86.75
68.29
85.11
85.25
93.94
78.16
84.44
81.33
91.46
-
AE
ADHD
ASD
BPD
MDD
SCZ

F1

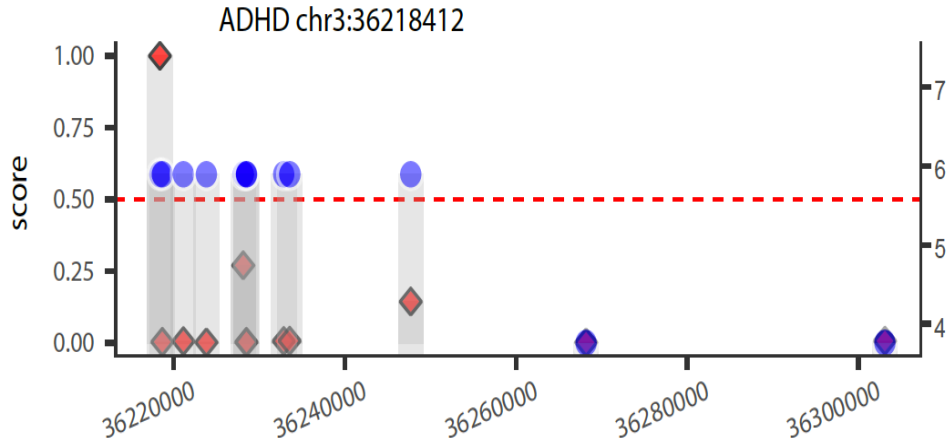

Fraction of positive calls
0.40
0.36
0.32
ADHD
ASD
BPD
MDD
SCZ

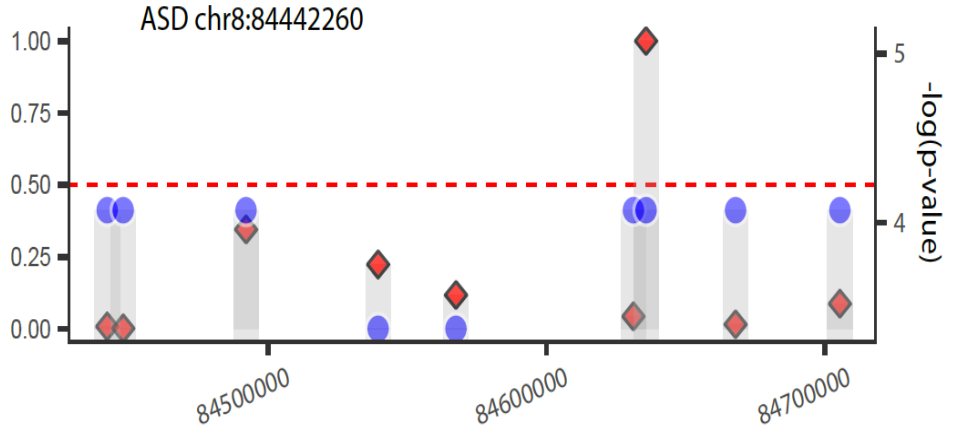

Mean prediction score
0.5
0.4
0.3
0.2
0.1
P=1.47x10⁻¹
P<0
P<0
P=3.92x10⁻¹
P<0
ADHD
ASD
BPD
MDD
SCZ

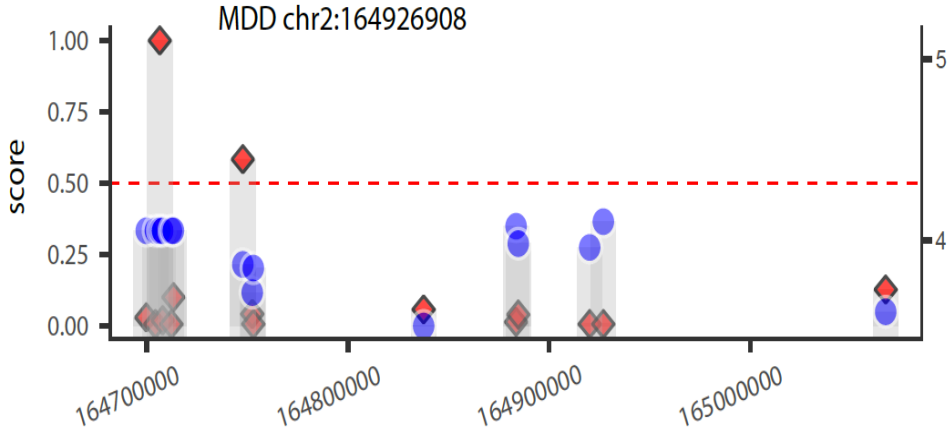
P<0.001
Control

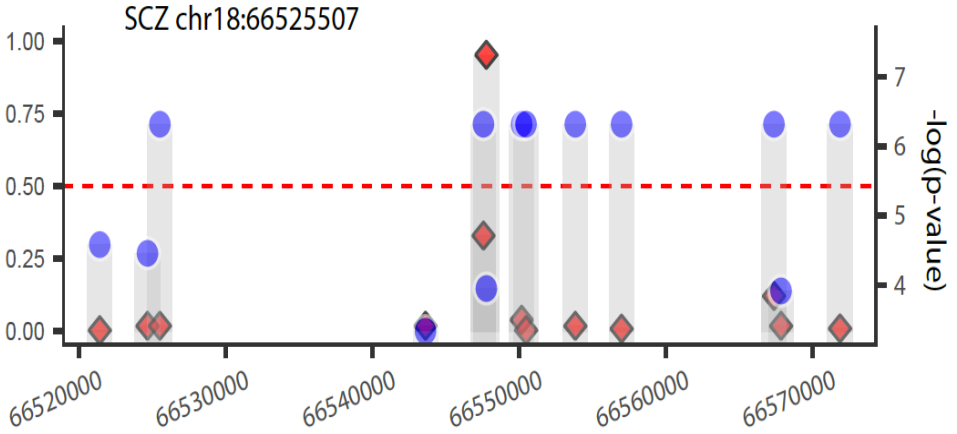
	GWAS cohort size				
	ADHD	**ASD**	**BPD**	**MDD**	**SCZ**
Case	2,787	4,949	6,990	9,227	9,379
Control	2,635	5,314	4,820	7,383	7,736

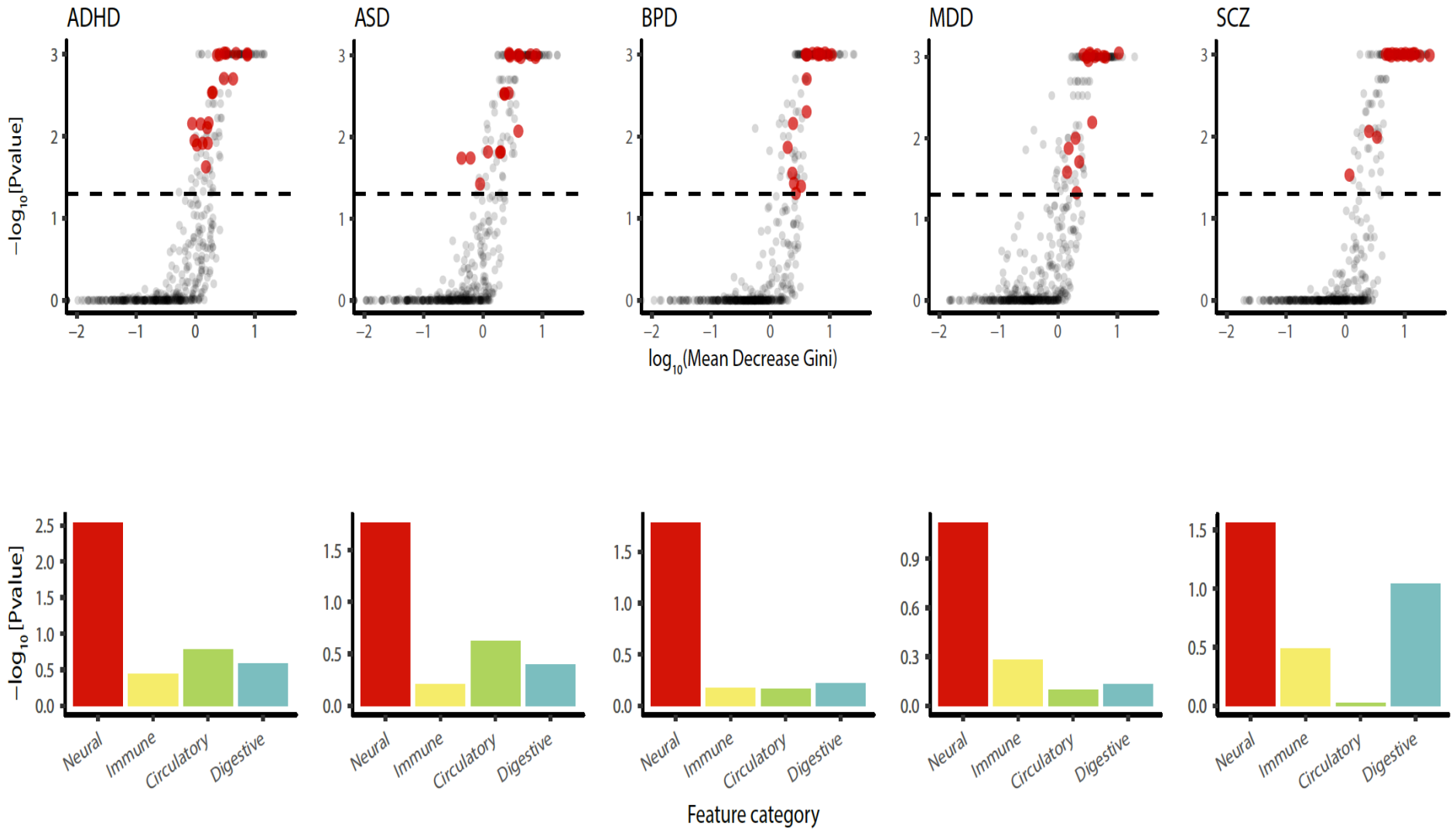
	Number of association blocks				
	ADHD	**ASD**	**BPD**	**MDD**	**SCZ**
Training set (chr 1-10)	227	253	292	267	384
Validation set (chr 15-22)	67	84	113	87	129
Test set (chr 11-14)	46	54	69	51	88
At least one positive call	285	333	430	347	536
Positive call for the most significant SNP	57.5%	52.3%	54.2%	56.5%	50.9%

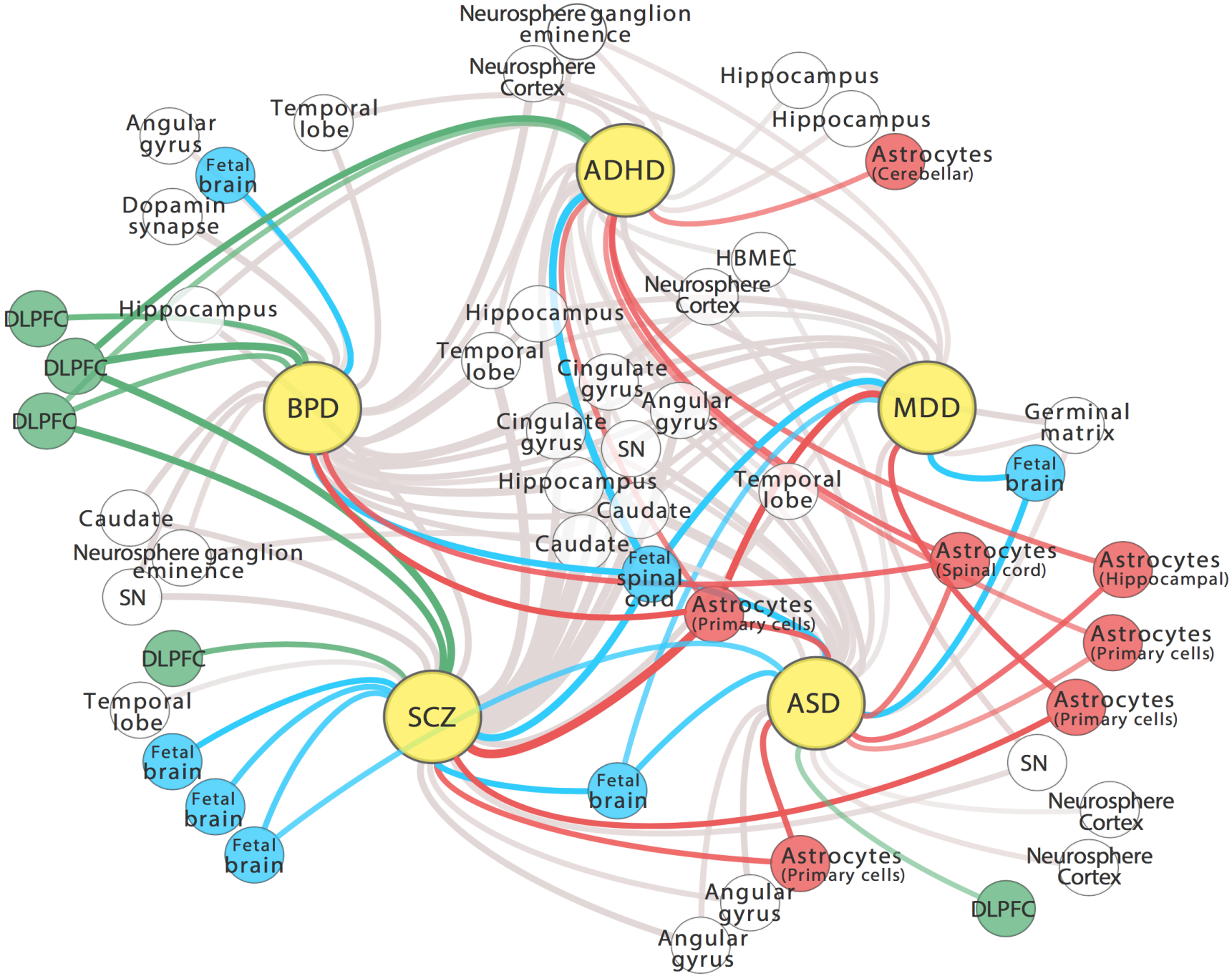

ADHD chr3:36218412
score
1.00
0.75
0.50
0.25
0.00
7
6
5
4
36220000
36240000
36260000
36280000
36300000

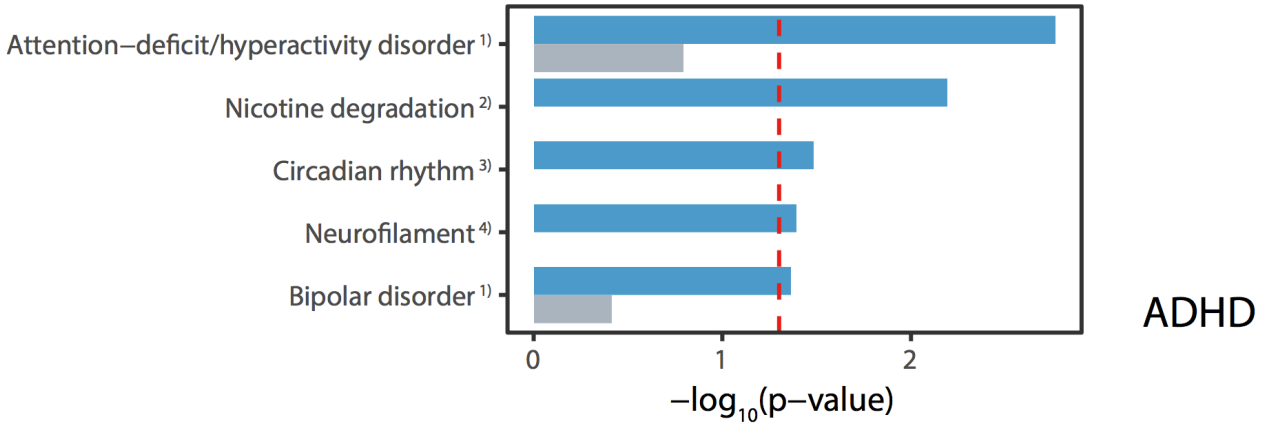

ASD chr8:84442260
1.00
0.75
0.50
0.25
0.00
5
4
-log(p-value)
84500000
84600000
84700000


MDD chr2:164926908
score
1.00
0.75
0.50
0.25
0.00
5
4
164700000
164800000
164900000
165000000

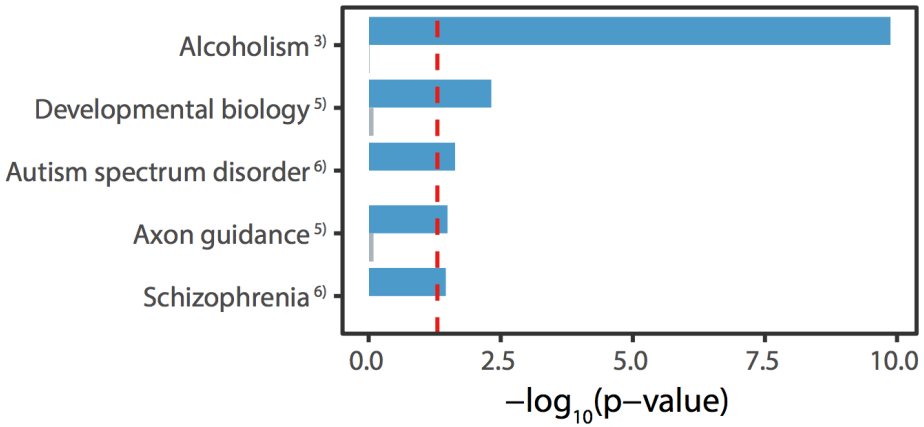

SCZ chr18:66525507
1.00
0.75
0.50
0.25
0.00
7
6
5
4
-log(p-value)
66520000
66530000
66540000
66550000
66560000
66570000


ADHD
ASD
BPD
MDD
SCZ
−log10[Pvalue]
log10(Mean Decrease Gini)
Neural
Immune
Circulatory
Digestive
Feature category

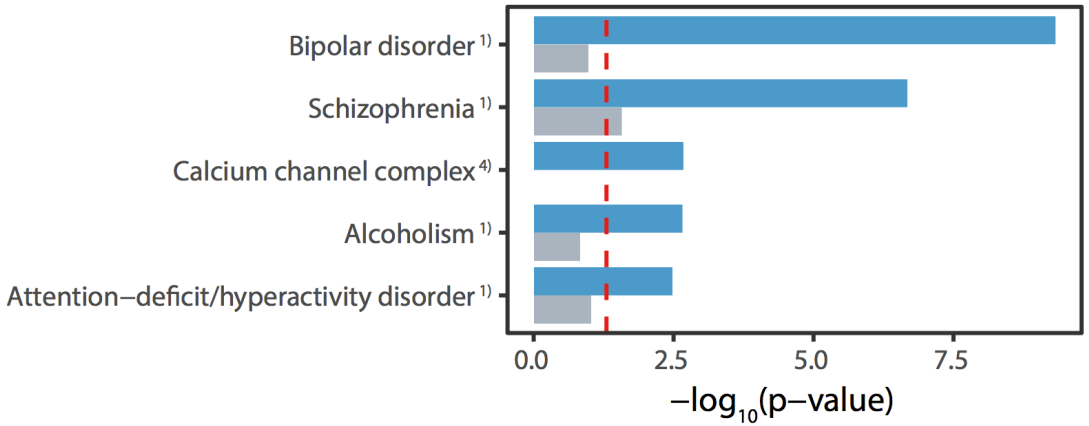

Neurosphere ganglion eminence
Neurosphere Cortex
Hippocampus
Hippocampus
Astrocytes (Cerebellar)
Temporal lobe
Angular gyrus
Fetal brain
Dopamin synapse
ADHD
HBMEC
Neurosphere Cortex
DLPFC
Hippocampus
DLPFC
DLPFC
BPD
Hippocampus
Temporal lobe
Cingulate gyrus
Angular gyrus
Cingulate gyrus
SN
Hippocampus
Caudate
Caudate
MDD
Germinal matrix
Fetal brain
Temporal lobe
Caudate
Neurosphere ganglion eminence
SN
Fetal spinal cord
Astrocytes (Spinal cord)
Astrocytes (Hippocampal)
Astrocytes (Primary cells)
Astrocytes (Primary cells)
DLPFC
Temporal lobe
SCZ
ASD
Astrocytes (Primary cells)
Fetal brain
Fetal brain
Fetal brain
Fetal brain
SN
Neurosphere Cortex
Astrocytes (Primary cells)
Neurosphere Cortex
Angular gyrus
DLPFC
Angular gyrus


Attention−deficit/hyperactivity disorder 1)
Nicotine degradation 2)
Circadian rhythm 3)
Neurofilament 4)
Bipolar disorder 1)
0
1
2
−log10(p−value)

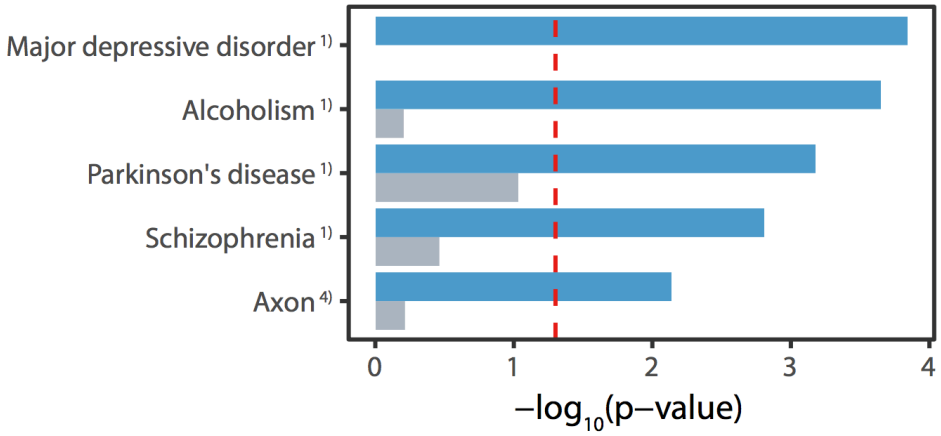
ADHD


Alcoholism 3)
Developmental biology 5)
Autism spectrum disorder 6)
Axon guidance 5)
Schizophrenia 6)
0.0
2.5
5.0
7.5
10.0
−log10(p−value)

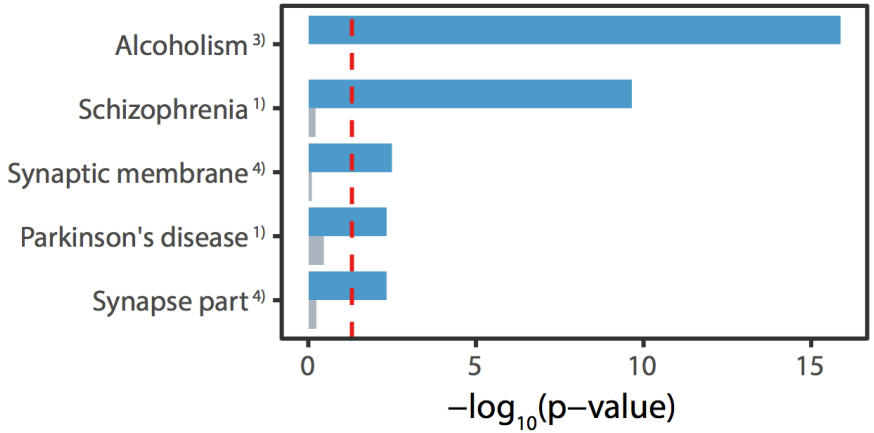
ASD


Bipolar disorder 1)
Schizophrenia 1)
Calcium channel complex 4)
Alcoholism 1)
Attention−deficit/hyperactivity disorder 1)
0.0
2.5
5.0
7.5
−log10(p−value)

BPD

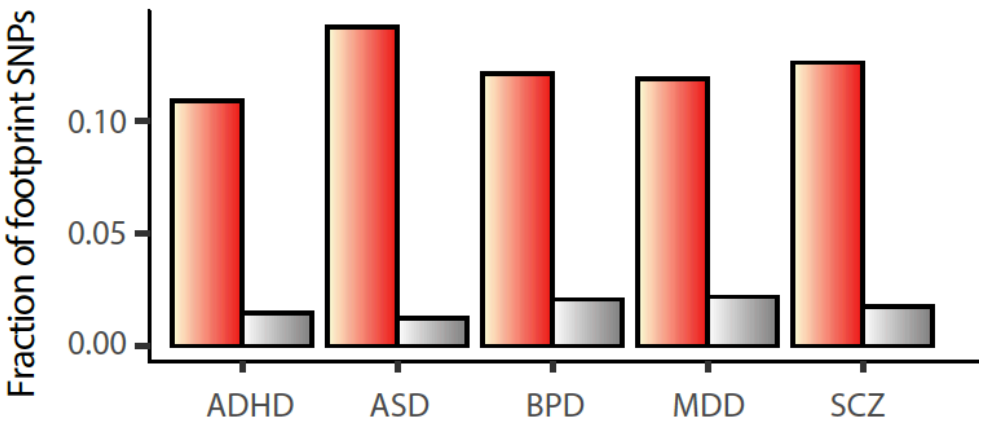


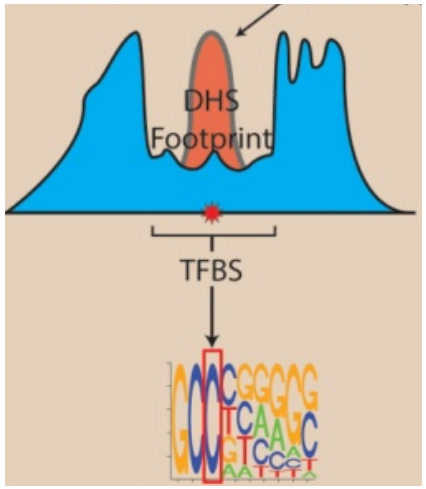
MDD

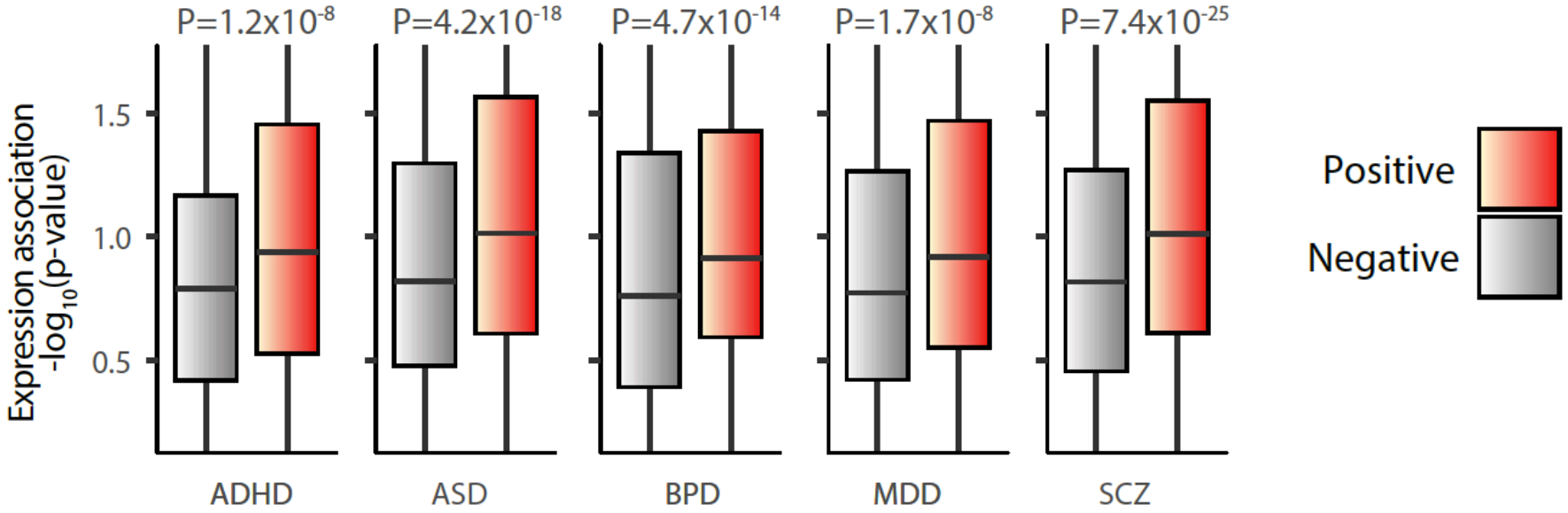


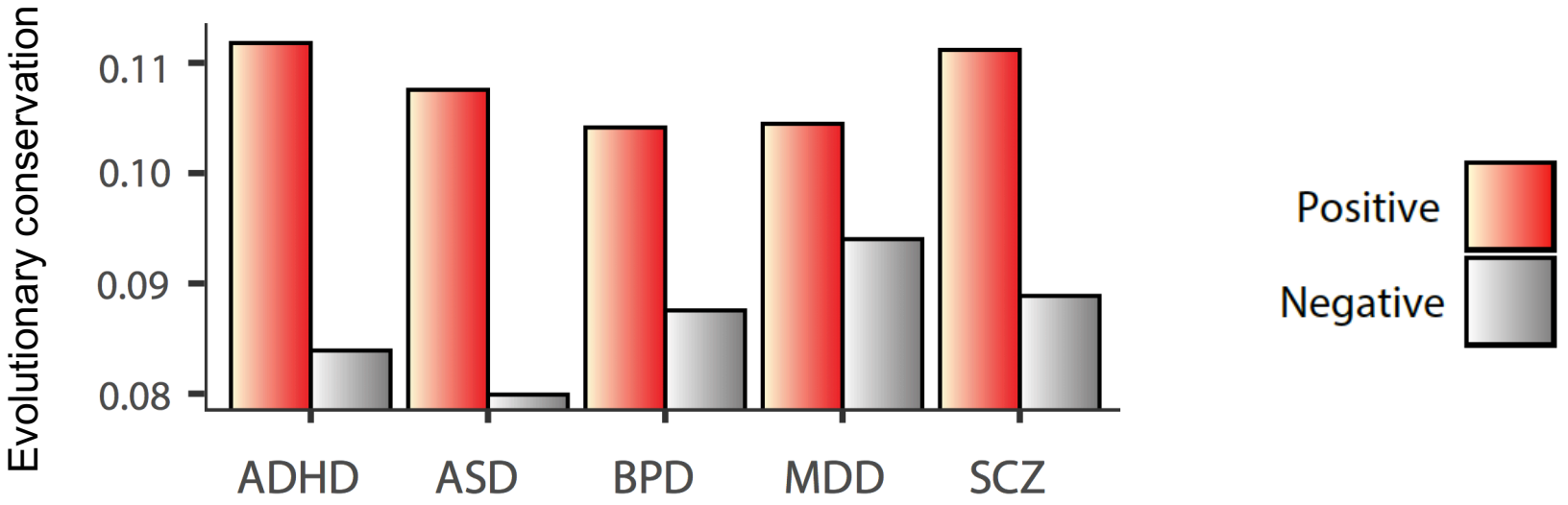
SCZ

1) dbGaP
2) Panther
3) KEGG
4) GO cellular component
5) Reactome
6) Disease perturbations from GEO

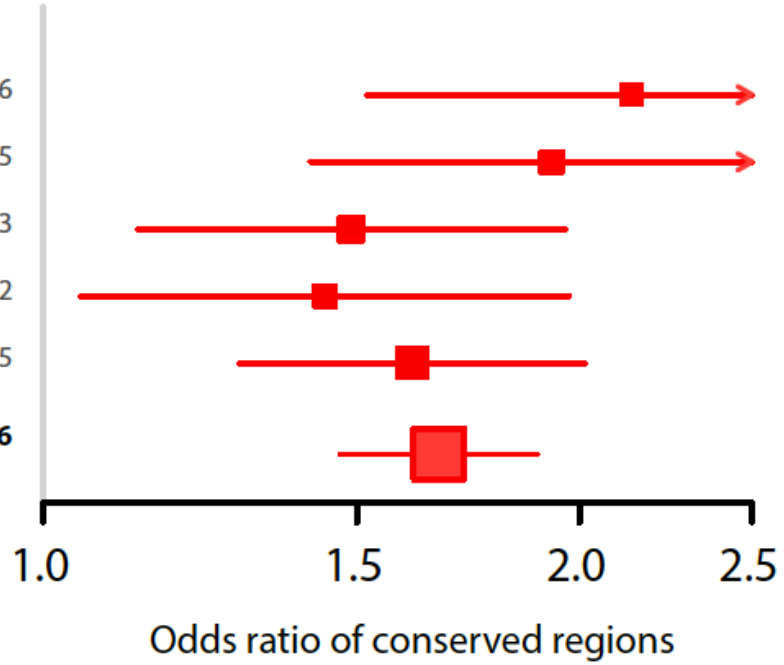

Fraction of footprint SNPs
0.10
0.05
0.00
ADHD
ASD
BPD
MDD
SCZ


DHS
Footprint
TFBS

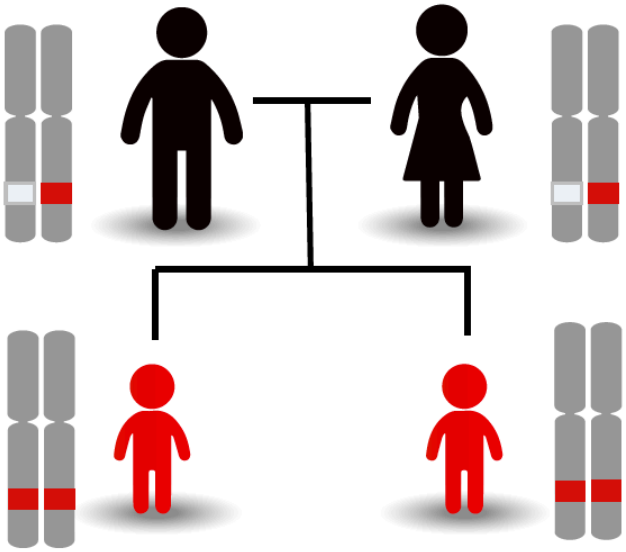

P=1.2x10-8
P=4.2x10-18
P=4.7x10-14
P=1.7x10-8
P=7.4x10-25
Expression association
-log10(p-value)
1.5
1.0
0.5
ADHD
ASD
BPD
MDD
SCZ
Positive
Negative



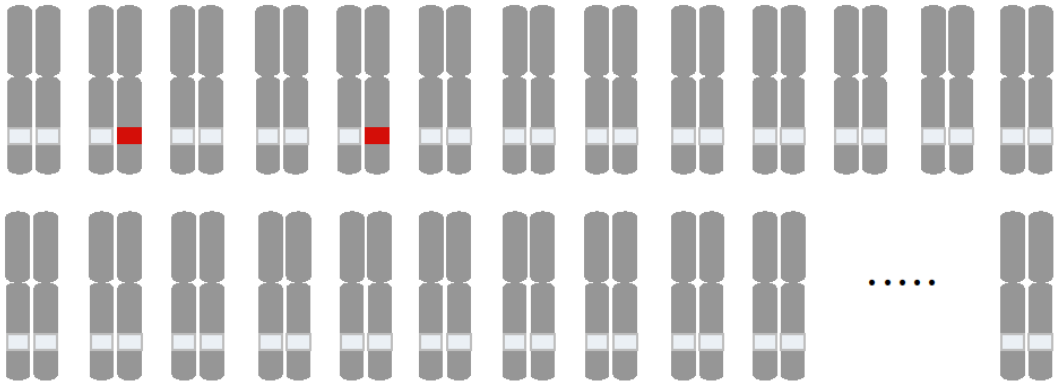
Disease	OR	P value
ADHD	2.140	4.53×10^{-6}
ASD	1.929	1.78×10^{-5}
BPD	1.490	2.34×10^{-3}
MDD	1.439	1.29×10^{-2}
SCZ	1.611	1.29×10^{-5}
Summary	**1.668**	**7.6×10^{-16}**



158 ASD multiplex familes
(350 affected children)



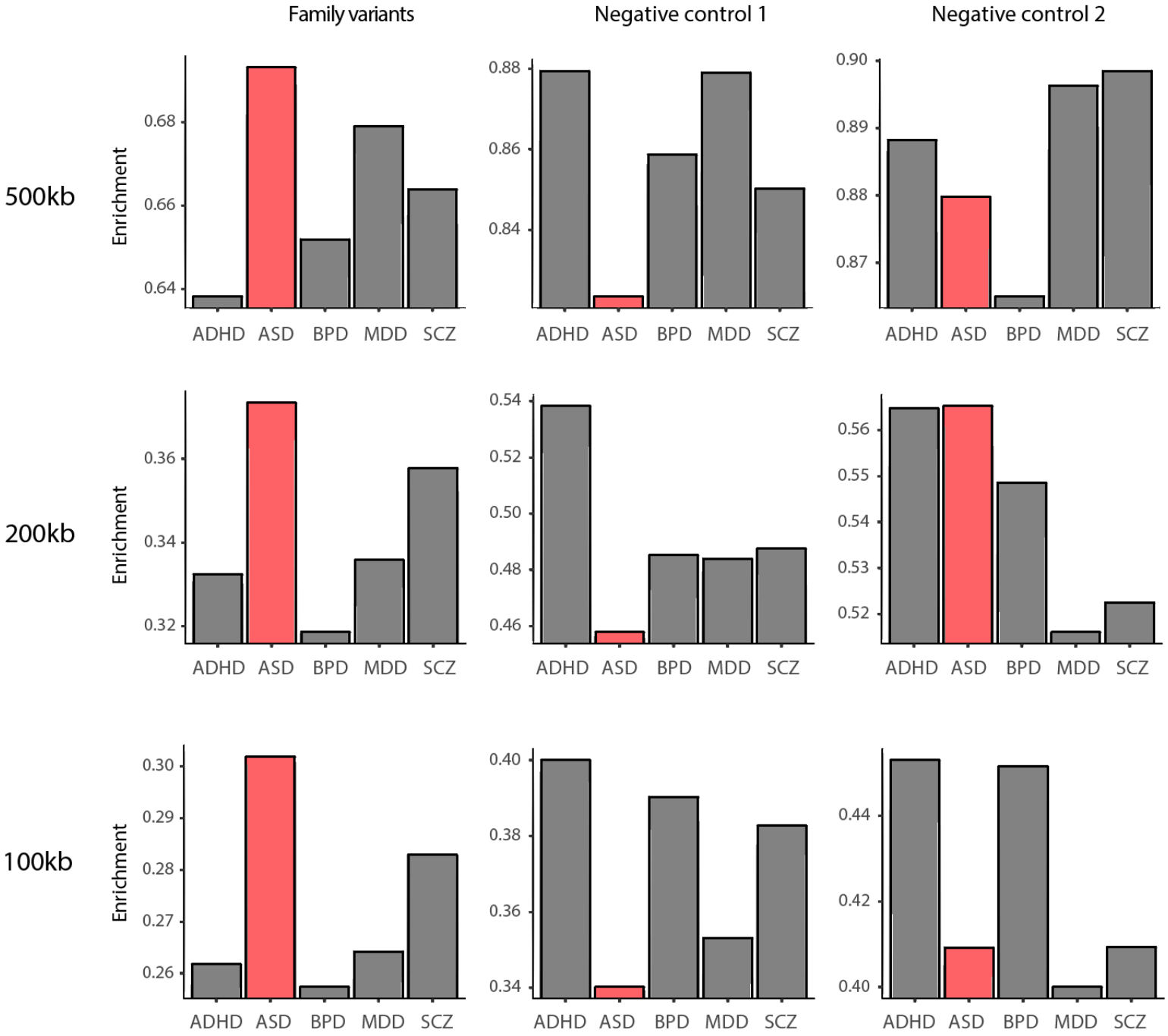
2,504 healthy individuals
from the 1000 Genomes project (MAF < 0.05)

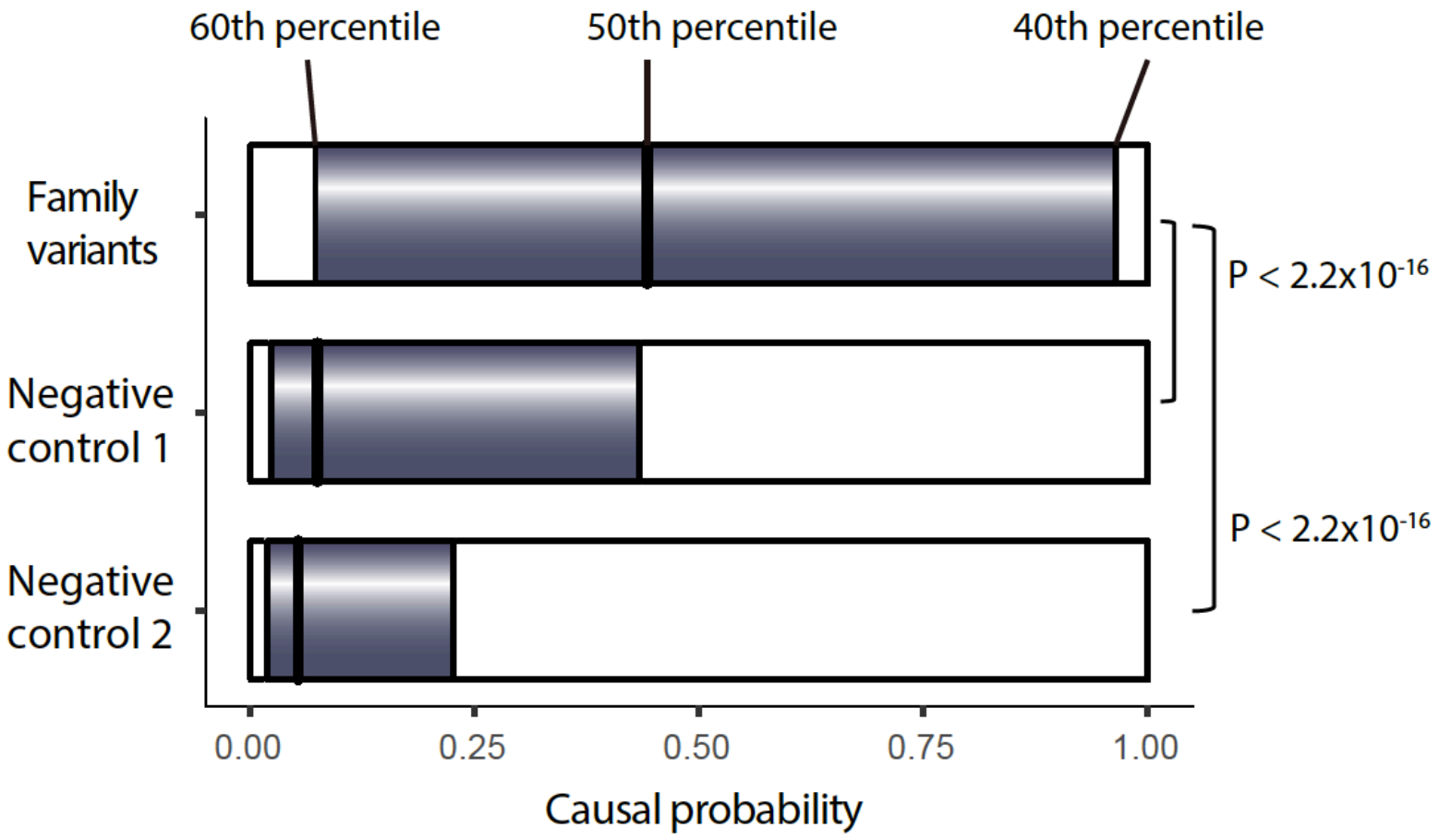


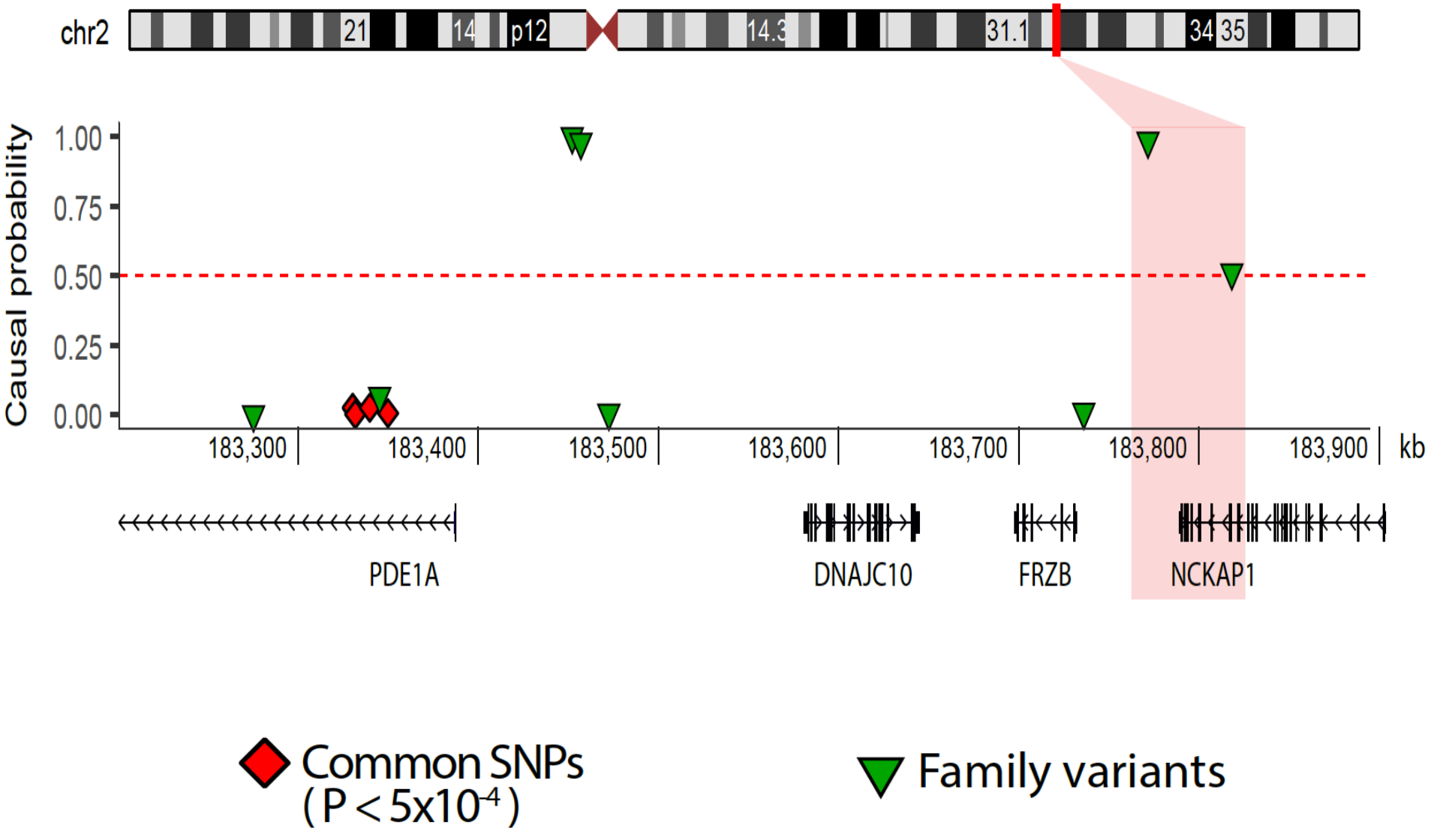
Two negative control sets

: variants that are heterozygous in any affected individual in each family

: variants that are homozygous in any of the 1000 Genomes samples


Family variants
Negative control 1
Negative control 2
500kb
200kb
100kb
Enrichment
0.68
0.66
0.64
0.88
0.86
0.84
0.90
0.89
0.88
0.87
0.36
0.34
0.32
0.54
0.52
0.50
0.48
0.46
0.56
0.55
0.54
0.53
0.52
0.30
0.29
0.28
0.27
0.26
0.40
0.38
0.36
0.34
0.44
0.42
0.40
ADHD
ASD
BPD
MDD
SCZ


60th percentile
50th percentile
40th percentile
Family variants
Negative control 1
Negative control 2
P < 2.2x10-16
P < 2.2x10-16
0.00
0.25
0.50
0.75
1.00
Causal probability

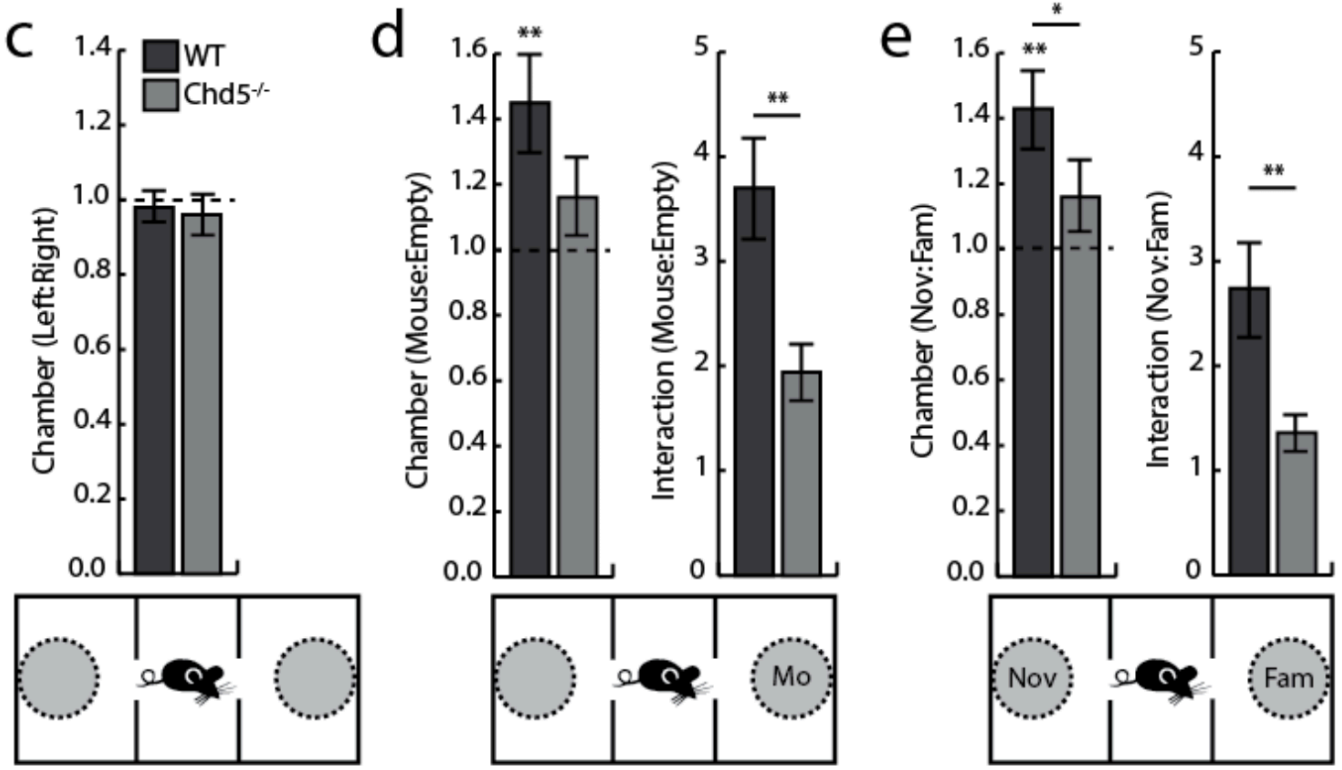

chr2
21
14
p12
14.3
31.1
34
35
Causal probability
1.00
0.75
0.50
0.25
0.00
183,300
183,400
183,500
183,600
183,700
183,800
183,900
kb
PDE1A
DNAJC10
FRZB
NCKAP1
Common SNPs
(P < 5x10-4)
Family variants

Identification of ASD genes

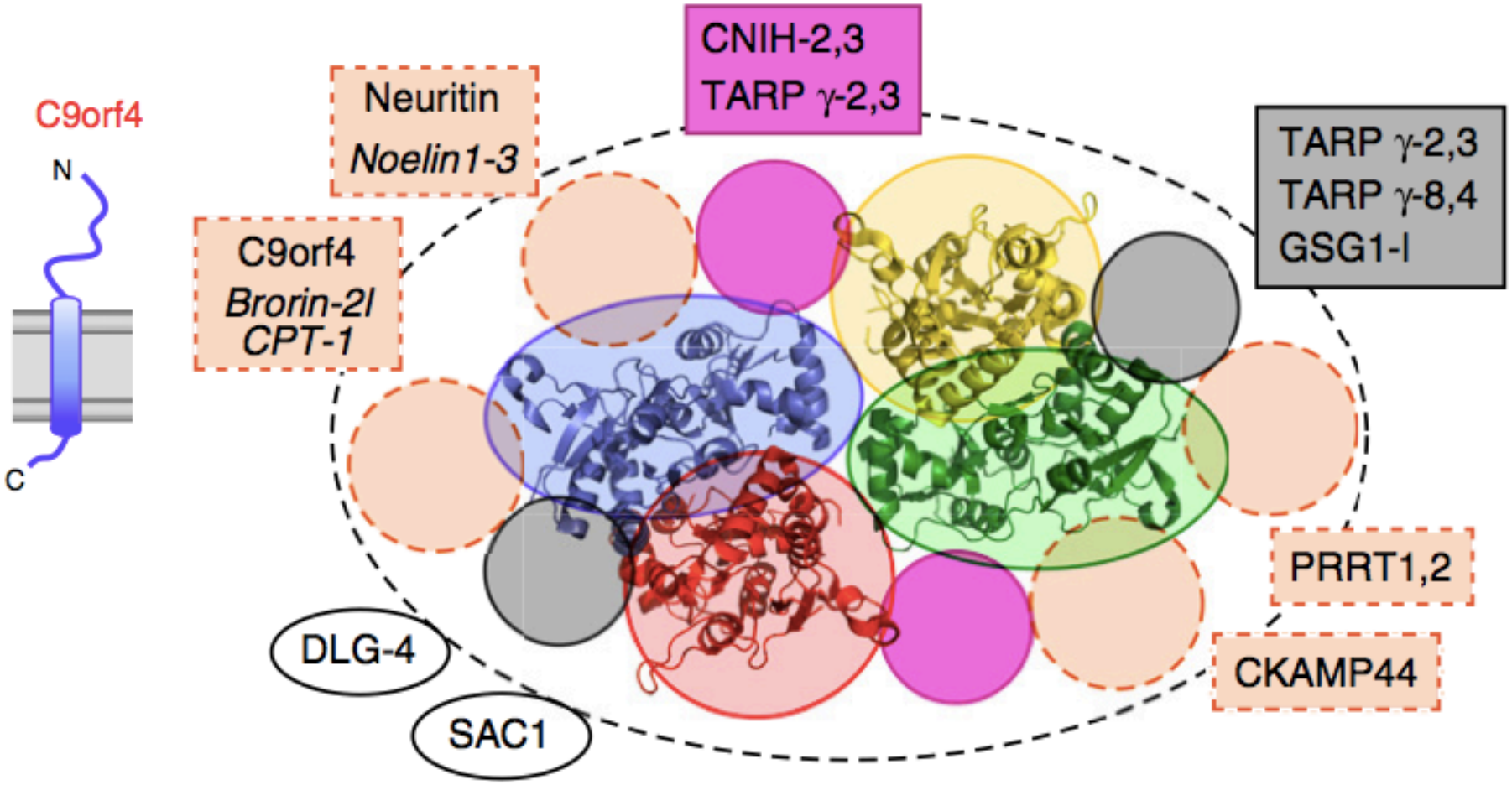
- The 15,874 variants formed clusters consisting of a series of rare alleles derived from the same family.
- Runs of homozygosity (ROH), long stretches of consecutive homozygous genotypes, reflect chromosomal segments shared identically by descent. ROH resulted from consanguinity may allow rare deleterious variants to exist in homozygous form.
- We detected 447 ROH consisting of the rare variants. The prediction scores were significantly higher for the ASD family ROH than the 1000 Genomes ROH.
- We identified three ROH genes, namely CHD5, FRRS1L, and APP, with 10, 8, and 5 rare variants with positive call near the transcription start site.

PhD thesis (University of Minnesota)

Socio-communicative abnormalities of Chd5$^{-/-}$ mice



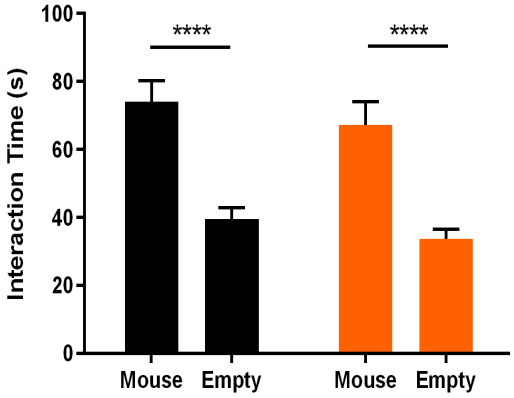
AMPAR complex

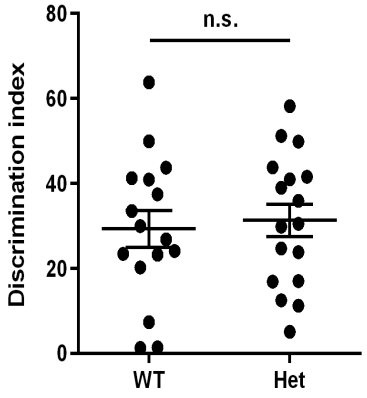


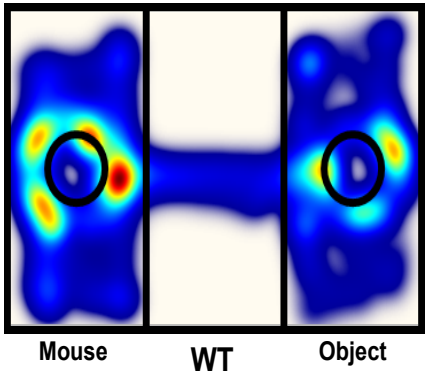
Neuron 74, 621–633 (2012)

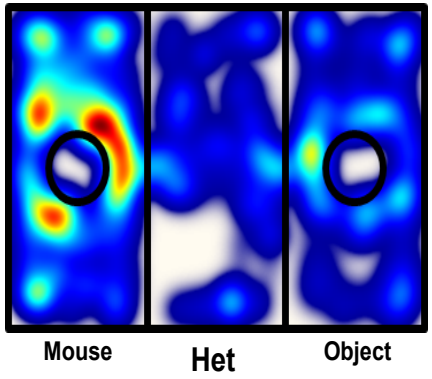
FRRS1L knockout mice

Sociability

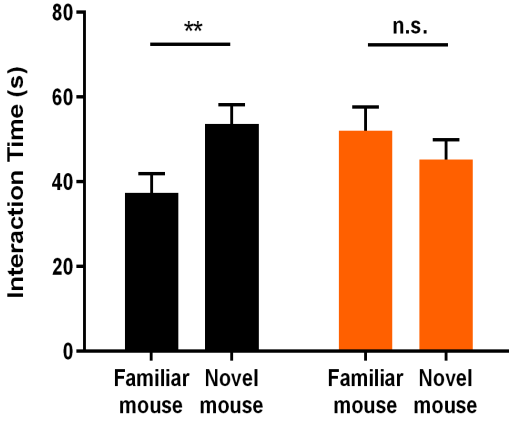


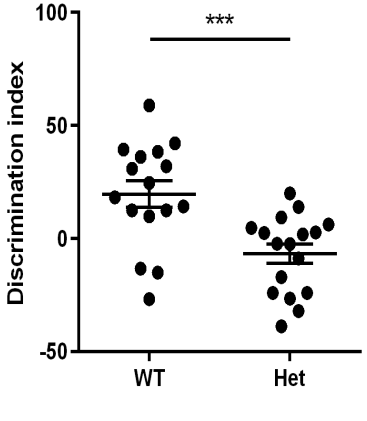


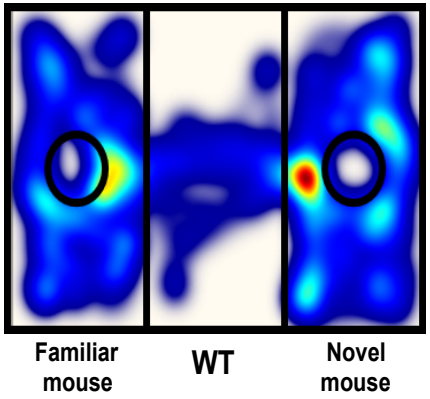


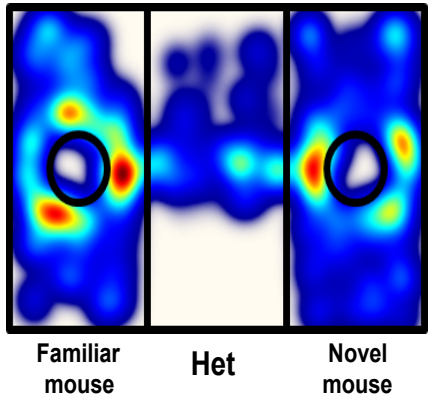


Social Novelty Recognition









Conclusion

- A major challenge facing GWAS is how to pinpoint DNA variants that actually contribute to associated phenotypes
- A majority of disease-associated DNA variations are thought to alter regulatory elements in noncoding regions
- Convolutional neural networks trained on epigenetic and regulatory features accurately predicted causal variants associated with psychiatric disorders
- Combined with family genome data, our deep learning method identified a novel autism gene

Jung Kyoon Choi, PhD
Principal Investigator

Kwoneel Kim, PhD
Post Doctoral Fellow

Hyo-Jeong Ban, PhD
Post Doctoral Fellow

Woojin Yang
PhD Candidate

Kangseon Lee
PhD Candidate

Hyoeun Bang
PhD Candidate

Min Kyung Sung
PhD Candidate

Kiwon Jang
PhD Candidate

Seulkee Lee, MD
PhD Candidate

Hyunchul Jung
PhD Candidate

Taeyeop Lee, MD
PhD Candidate

Jaeho Oh
PhD Candidate

Min Gyun Bae
MS Candidate

Jun Hyeong Lee
MS Candidate

Joonha Kwon
Researcher

Jeong Yeon Kim
Intern

Hyo Jung Yoon
Research Assistant

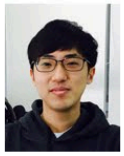
Jin-Hee Han, Ph.D.

Associate Professor of Biological Sciences

Tel. 042-350-2649 (office)

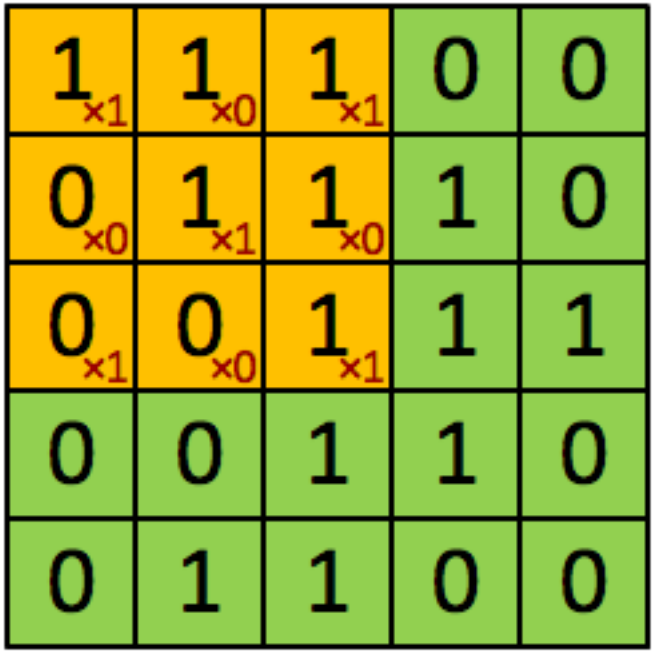
Fax. 042-350-2610

E-mail : han.jinhee@kaist.ac.kr

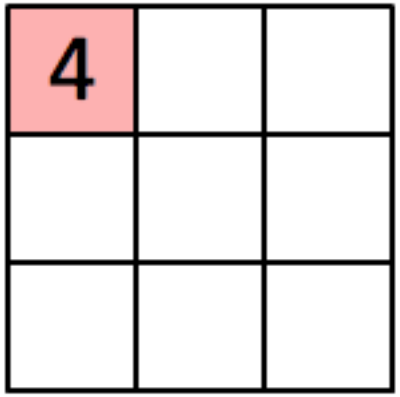
	**Ji-eun Kim, Ph.D.** - Research fellow (2014-) - Post-doc (2009-2013) - Molecular basis of learning and memory		**Jinho Jhang** - MS-Ph.D course (2012-) - Prefrontal fear cotrol
	**Yire Jeong** - MS-Ph.D course (2012-) - Optogenetic control of fear memory circuit		**Miran Yoo** - MS-Ph.D course (2013-) - Epigenetic molecular mechanism of memory
	**Hye-yeon Cho** - MS-Ph.D course (2013-) - Molecular basis of learning and memory		**Jungpyo Oh** - MS-Ph.D course (2014-) - Molecular basis of learning and memory ** Member Representative
	**Mujun Kim** - MS-Ph.D course (2014-) - Molecular basis of learning and memory		**Min Soo Kang** - MS-Ph.D course (2015-)
	**Yeji Lee** - MS-Ph.D course (2015-)		**Hansol Lee** - MS-Ph.D course (2016-)
	**Hyeonju Lee** - MS-Ph.D course (2016-)		**Minju Kang** - Lab manager (2012-) - ncbmanager@gmail.com

Thank you

Convolutional neural network

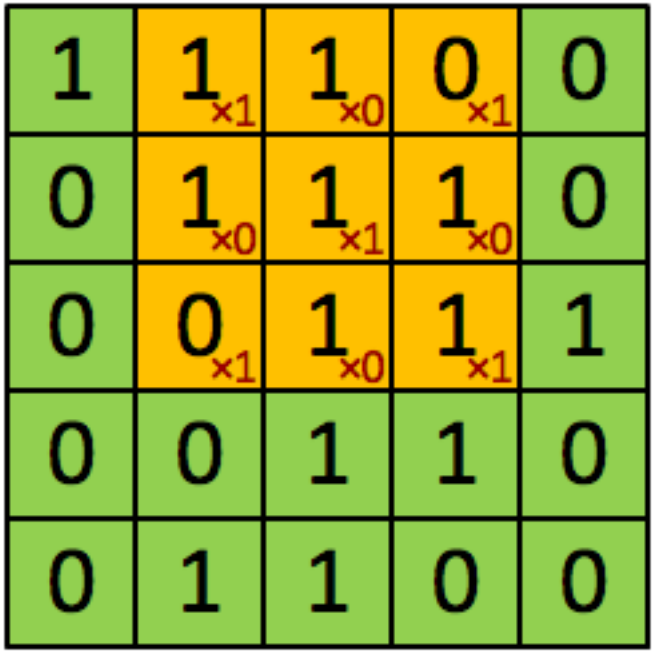


Image

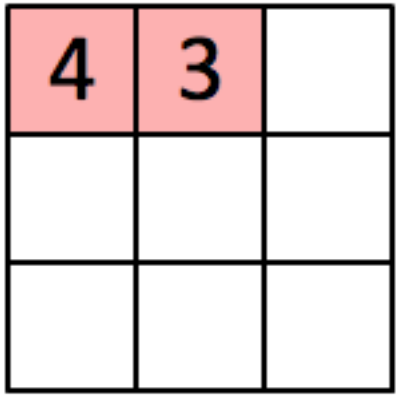


Convolved Feature

Convolutional neural network

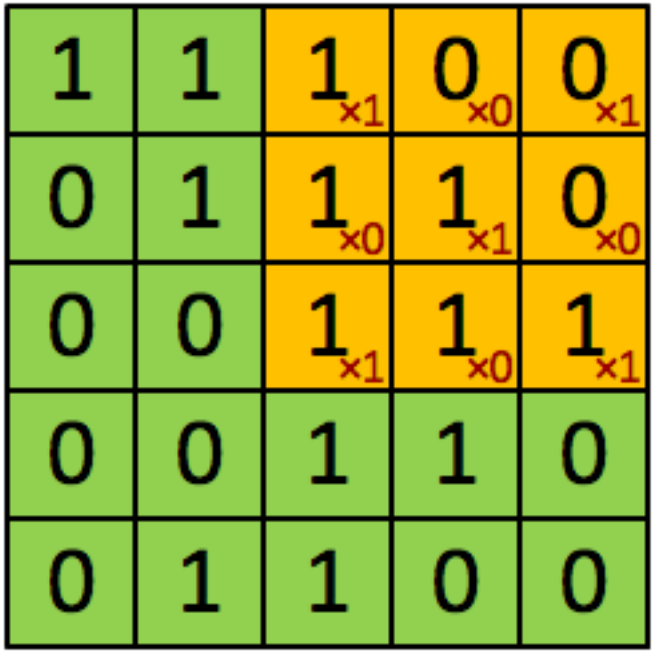


Image

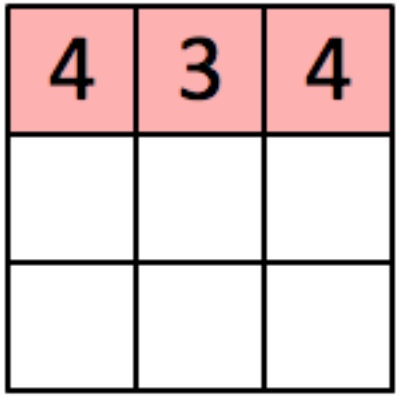


Convolved Feature

Convolutional neural network

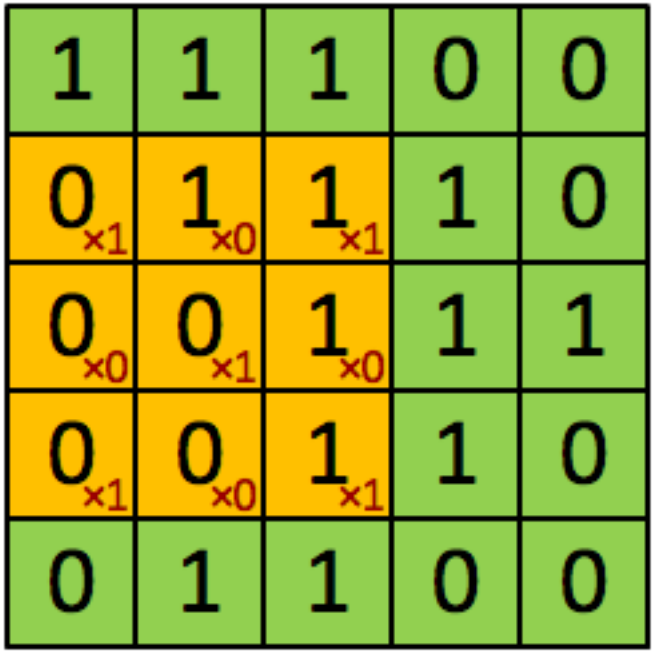


Image

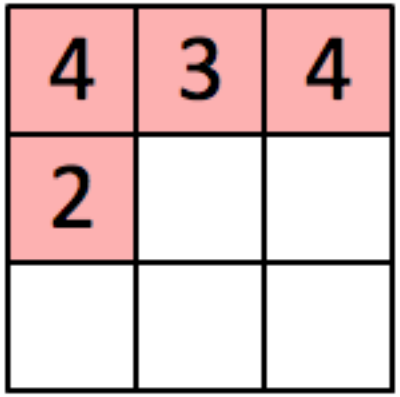


Convolved Feature

Convolutional neural network

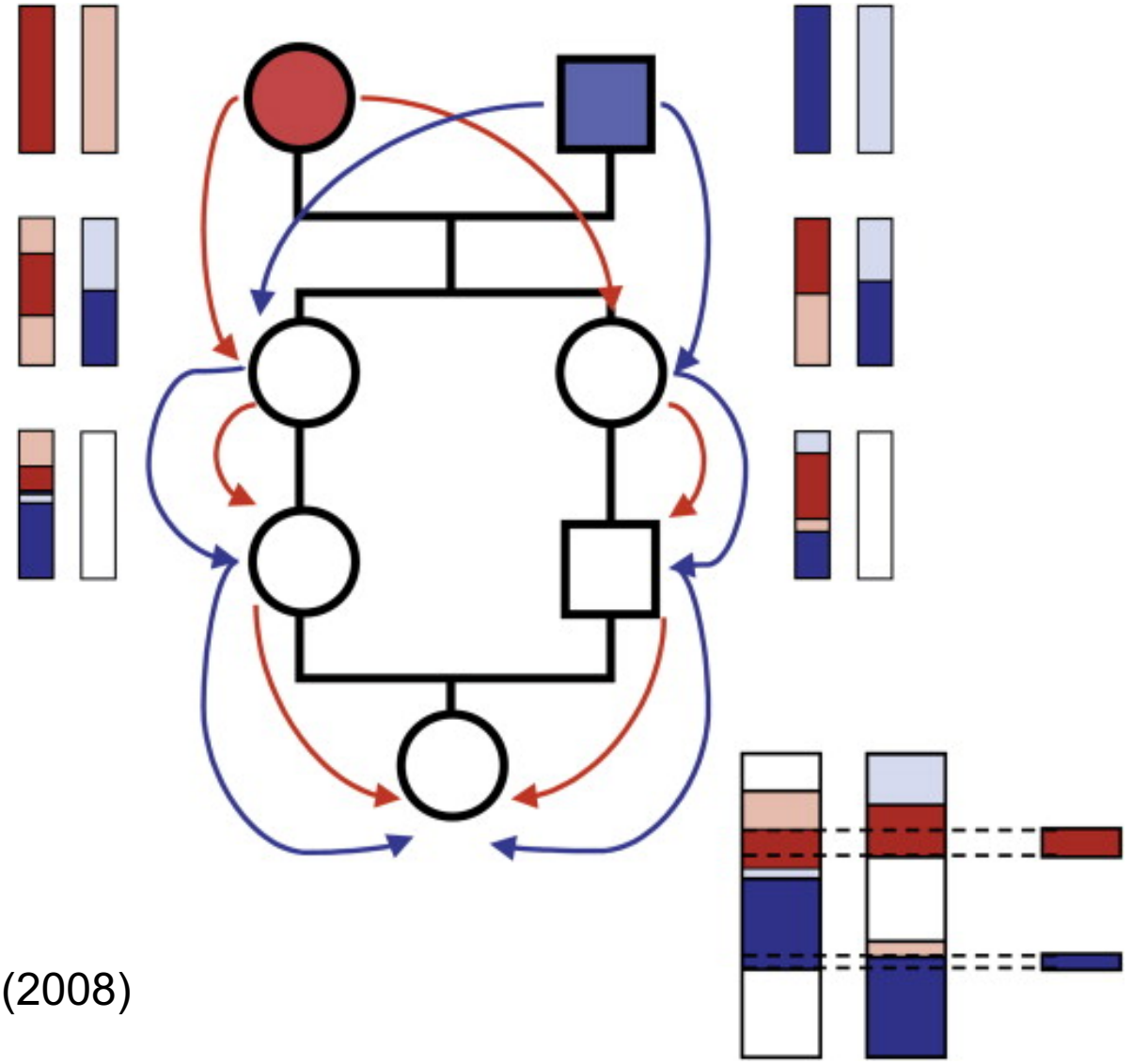


Image

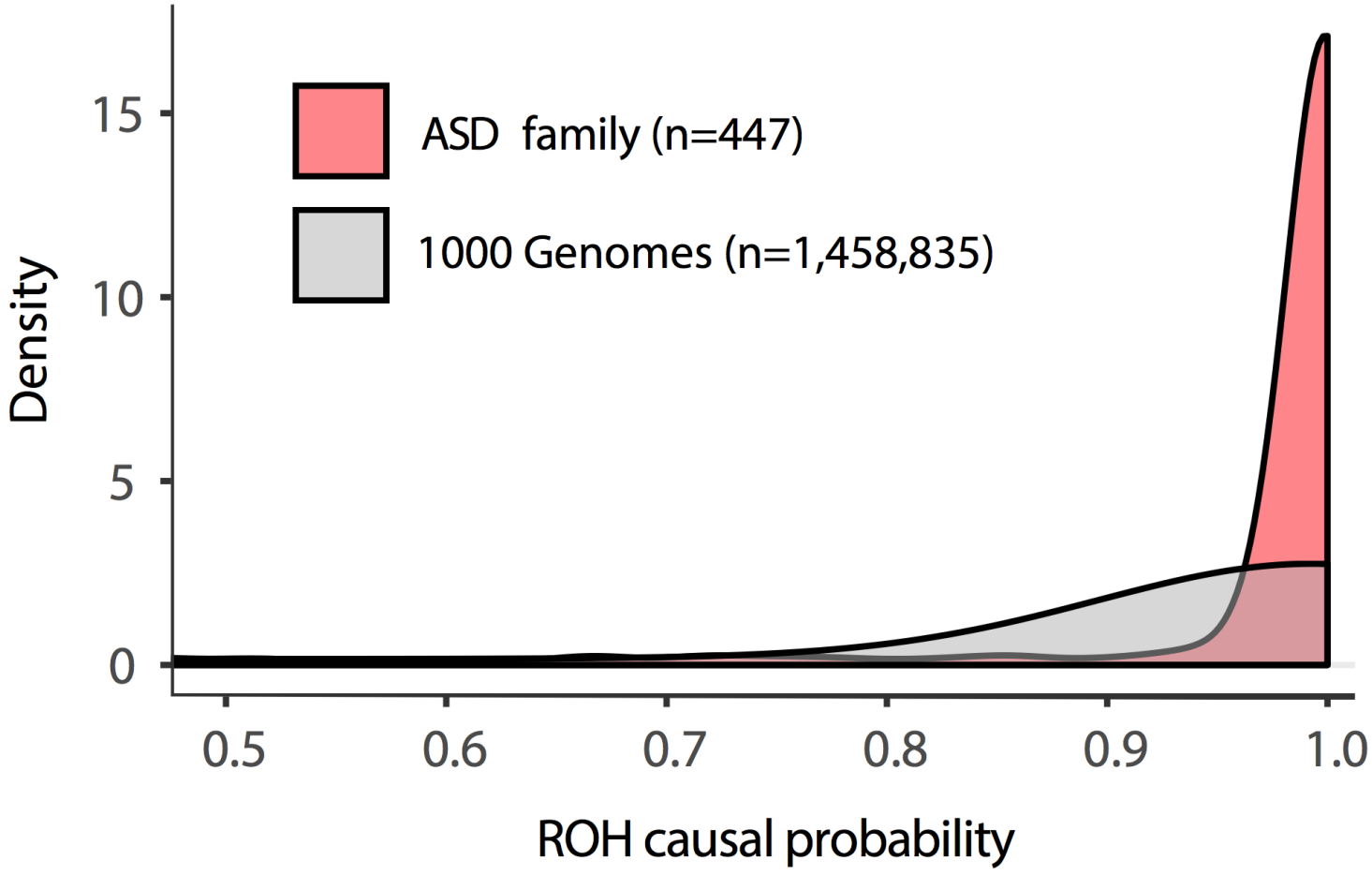


Convolved Feature

Runs of homozygosity (ROH)



AJHG 83:359-372 (2008)


ASD family (n=447)
1000 Genomes (n=1,458,835)
Density
15
10
5
0
0.5
0.6
0.7
0.8
0.9
1.0
ROH causal probability