Cell type-specific interrogation of neurological disease-associated variants

Major: Biological Science

Student: LI Bing-kun Supervisor: Dr. SHEN Yin

[Abstract] For the past decade, Genome Wide Association Studies (GWAS) have identified thousands of trait-associated variants. However, identifying the causal variants and characterizing their underlying biological mechanisms often require linkage disequilibrium annotations as well as context-specific, multi-dimensional genomic datasets. Here, we 1) generated and analyzed cell type-specific, genome-wide chromatin interactions usings promoter capture Hi-C (pcHi-C), open chromatin landscapes using ATAC-Seq and transcriptome using RNA-Seq dataset, 2) built a customized pipeline for running pre-built FINEMAP software and most importantly 3) built a customized data-processing pipeline for integrating fine-mapping results with multiple genomic annotation datasets, including the ones we generated ourselves, in order to comprehensively interrogate the functional roles of both coding and non-coding variants . Finally, we made an attempt to explore the implications of eQTL SNPs, resulting from association studies different from GWAS, as a beginning of future work on integrating eQTL and GWAS results to increase the power of predicting causal genes. To validate the feasibility of our approaches, we applied our pipelines to five neurological disease-associated GWAS results (Alzheimer's disease, AD; Bipolar disorder, BD; Autism spectrum syndrome, ASD; Attention-Deficit/Hyperactivity Disorder, ADHD and Schizophrenia, SCZ) and four distinct cell types (iPSC-derived excitatory neurons and lower motor neurons, iPSC-derived hippocampal dentate gyrus (GD)-like neurons, and primary astrocytes). We illustrated that we provided an approachable pipeline for identifying and charactering putative causal variants, which may serve to prioritize candidate variants for experimental validation using single-cell CRISPR screening and provide basis for potential applications in therapeutic settings.

[Key words] GWAS SNPs; Fine-mapping; Neurological disorders; Functional interpretation

TABLE OF CONTENTS

Introduction

For over a decade, genome-wide association studies (GWAS) have contributed to the identification of reproducible genomic regions on chromosomes that harbor genetic determinants of complex traits. Particularly, large-scale consortiums and biobanks such as UK Biobank (http://www.ukbio-bank.ac.uk/) have made possible the interrogation of larger sample sizes, which drastically increased the power of recent GWAS. To date, GWAS have successfully identified over 70,000 variant-trait association[1]. However, the prioritization of variants within GWAS-associated regions and the functional interpretation of which have fall largely behind, which can be mostly ascribed to two complications.

First, in earlier studies it was common to consider only SNPs passing a certain p-value threshold and to assume that SNPs with the smallest (that is, most significant) P values in distinct regions, sometimes called the lead or index SNPs, are more likely to be causal. However, this method often has serious limitations, given that P-values are influenced by study-specific factors such as power (determined by sample size) and locus-specific factors such as minor allele frequency and effect size, often times the true associations are not likely to result in the smallest P values [2-3]. Furthermore, as a result of the large amounts of linkage disequilibrium (LD) blocks in the human genome[4], there are often many co-inherited variants in strong linkage with the sentinel variants, and variants in strong LD are often statically indistinguishable in terms of associations with disease risks. Therefore, filtering the GWAS SNPs only with a certain p-value threshold (e.g. $5x10^{-8}$) or considering index SNPs to be causal SNPs are often arbitrary and may miss out the true functional associations, and it is important to use more intricate statistical models to prioritize the likely causal variants.

Second, more than 90% of the GWAS SNPs are in the non-coding regions of the genome, and many are far away from their nearest gene[5], therefore their roles in affecting phenotypes often cannot be inferred directly. Possible mechanisms of long-distance regulations may involve regulatory elements that modulate gene expression level such as enhancers, which are often highly cell-type specific. For example, two independent obesity-associated SNPs in the FTO gene have been shown not to regulate FTO, but IRX3 in the brain and both IRX3 and IRX5 in adipocytes, respectively. The FTO locus in obesity illustrates the potentially intricate and cell type-specific manner in which noncoding variants contribute to disease. Also, in the past it had been typical to assign the nearest genes of non-coding variants as their target genes. However, a large number of studies have indicated

that this model does not always reflect the real case scenarios and often times genes are regulated by distal elements thousands base pairs away.

The first challenge can be addressed by the recent advancement of novel statistical fine-mapping frameworks[6]. The general strategy of fine-mapping is to use information provided in the GWAS summary statistics to identify regions of interest, taking LD structures into account. Several approaches have been used to perform fine-mapping, including penalized regression and stepwise conditional analysis; the former tends to result in sparse models which reduce the chance of selecting the causal variants[7], while the latter, although being the standard approach for refining association signal and informative about the number of complementary sources of association signals within the designated region, does fail to determine the probability of being causal for each variant[8]. To overcome the limitations mentioned above, many recent fine-mapping pipelines have turned to adopt a Bayesian framework instead. Examples of Bayesian model-based methods include PAINTOR[9], CARVIAR[10], CAVIARBF and FINEMAP. Although each of those methods are robust for performing fine-mapping, the first three of which implemented an exhaustive search algorithm to go through every possible causal configuration. And while this ensures the software to find the most optimal configurations, the drawback would be heavy computational burdens, making it less favorable under circumstances where each loci harbors hundreds of SNPs. FINEMAP implemented a shotgun search algorithm which made it thousand times faster while still maintaining similar accuracy. We therefore used FINEMAP for performing fine-mapping in this study. Notably, although such methods could pinpoint the most probable configurations, it is important to note that statistical methods alone cannot determine causality and further downstream analyses are often needed.

To address the second challenge, the use of genomic dataset and experimental approaches are needed to dissect the functional roles of putative causal variants identified by statistical methods. Importantly, as mentioned above, many regulatory elements, such as enhancers, function in a tissue or cell type-specific manner[11] and may play different roles in the pathogenesis of different diseases, necessitating the integration of tissue or cell type-specific genomic annotations. However, the annotations for enhancers are very much incomplete and often lack cell type-specificity, hindering the efforts to systematically assign likely target genes to non-coding variants. Recent large-scale efforts in predicting and interpreting the putative causal variants such as INFERNO[12] and POSTGAP[13], although having shown promising results in identifying putative mechanisms underlying non-coding variants[14], use pre-built publicly available genomic annotations as references, making it hard to

customize input and tailor to individual studies. Furthermore, most existing chromatin organization dataset are derived from Hi-C experiment, which captures all interactions in the genome and therefore contains a large number of unwanted noises, making it difficult to pinpoint regulatory elements. To overcome those limitations, we took advantage of the recent-developed promoter capture Hi-C technology[15] which enables a promoter-centered analysis and a more precise linking from genes to regulatory element. We also performed ATAC-Seq and RNA-Seq which delineate open chromatin region and gene expression level respectively, providing comprehensive annotations of potential cis-regulatory elements and likely cognate genes.

Finally, studies have shown that integrative modelling of data from expression quantitative trait locus (eQTL) studies and summary-level data from GWAS effectively promote the transformation from SNP- phenotype associations into functionally informative gene-phenotype association profiles[16-17]. Therefore, as a start of future works, we sought to integrate the eQTL dataset with our cell type-specific epigenomic and chromatin interaction annotation to better predict and interpret the functional variants.

Results

Genomic and epigenomic landscape of distinct neural cell types

In order to interrogate the disease-associated variants in a cell type-specific manner, we characterized the general features of genomic and epigenomic landscape of four specific neural cell types (Fig.1a), three of which are currently impractical to isolate from primary tissues and were therefore induced from WTC 11 iPSC lines (iPSC-induced excitatory neurons, iPSC-derived hippocampal dentate gyrus (DG)-like neurons, and iPSC-induced lower motor neurons). The identity of all three cell types were confirmed by transcriptional signatures for dozens of marker genes. In addition, all three neuronal subtypes showed high expression of synaptic genes SYN1 and SYN2, the NMDA receptor genes GRIN1 and GRIN2A, and the AMPA receptor genes GRIA1 and GRIA2, evidencing mature synaptic functions (Fig 1b). The remaining one cell type was isolated from 19 week gastrulating male fetal brain samples and based on the age of donors and gene expression profile, it was determined to most likely to be astrocyte progenitor cells (APCs) (Fig.1b).

We constructed promoter capture Hi-C, ATAC-Seq and RNA-Seq libraries using two to four biological replicates for each cell types (Fig 1a). In addition, to show that cell types can reliably be grouped according to lineage-specific features and that the distinct genomic and epigenomic patterns of each cell type can indeed be ascribed to innate differences instead of biases from batch effect or sequencing depths, we confirmed the reproducibility of contact frequency and saturation of inter-replicate correlation for our pcHi-C libraries using HiCRep[18] (Fig 2a, d). Hierarchical clustering of ATAC-Seq read density and gene expression similarly group the replicates by cell type (Fig 2b, c), evidencing minimal variations during the cell derivation process.

In addition, to further test the correlation between our iPSC-derived excitatory neurons and primary excitatory neurons, we took advantage of previously published single-cell RNA-Seq data of human developing brain[30]. We found that our iPSC-induced excitatory neurons are slightly more correlated with excitatory neuron populations identified by single-cell RNA sequencing (Supplementary fig 5a, b). This can be partly explained by the fact that 1) tSNE plot has a certain level of randomness, and WGCNA clustering relies on the expression value of a set of signature genes which can be somewhat arbitrary, so the cell types nominated in the tSNE plot may not be very accurate; 2) Our excitatory neurons were more mature than the ones used to generate the single cell RNA-Seq data, so the gene expression values might be biased.

Fine-mapping locates hundreds of possible disease-specific causal variants

Next, to narrow down the list of likely causal variants and to prioritize candidates for further investigation, we built a reproducible and highly-scalable customized pipeline for running FINEMAP in a genome-wide, unbiased manner (Supplementary Fig 1a). We consequently performed fine-mapping analyses on five neurological disorders: Alzheimer's disease, AD; Bipolar disorder, BD; Autism spectrum syndrome, ASD; Attention-Deficit/Hyperactivity Disorder, ADHD and Schizophrenia, SCZ.

For each disease-variant association dataset, we constructed an average of 136 loci (AD, 64; BP, 127; ASD, 58; ADHD, 70; SCZ, 364) for running FINEMAP. The block size distribution of AD and ADHD fall within the range of LD block sizes estimated by previous study[19] while for SCZ, ASD and BP we observed loci whose sizes greatly exceeds the upper boundary of estimated LD block sizes (Supplementary Fig 1b). This may indicate that loci defined by our approach may cross different LD blocks. However, it may also suggest that defining loci by arbitrary distance cutoffs which were

thought to be able to capture all linkage structures may fail to do so. From a total of 683 regions generated, our fine-mapping pipeline pinpoints hundreds of (AD, 566; BP, 839; ASD, 390; ADHD, 607; SCZ, 2961) likely causal variants (pp > 0.1, MAF>0.1%) for each disease type (Fig 3a). The numbers of variants predicted to be causal are proportional to the number of SNPs reported in the original summary statistics.

To gain a general understanding of fine-mapped SNPs, we first overlapped all variants with an accumulative posterior probability larger than 0.1% from FINEMAP output with several genetic and epigenetic annotation categories. Interestingly, in contrast to earlier findings[20], we found a minimum localization in UTR region, a very limited enrichment in coding region, but a very strong localization in intronic and intergenic regions, and the same localization pattern persists across all five diseases (Fig 3b). Moreover, an average of 39 fine-mapped non-coding variants (AD, 27; BP, 27; ASD, 16; ADHD, 16; SCZ,109) falling into chromatin accessible regions identified by ATAC-Seq. Next, we were interested to see whether variants overlapping open chromatin regions are more likely to be causal (e.g. have a higher posterior probability predicted by FINEMAP). However, we did not observe any significant differences between variants within open chromatin regions or not (Supplementary fig 2). This might suggest there is no correlation between the two, or, which is more likely to be the case, that our sample size is too small to derive any significant observations.

To validate our approach, we next examined whether the fine-mapped variants recapitulate previously known risk genes. We took all the genes whose exons are targeted by at least one fine-mapped variants and performed gene ontology and pathway enrichment analyses for each disease. We successfully recapitulated a series of repeatedly reported risk genes in Alzheimer's disease from previous studies, including APOE, TREM2, CD33, CLU, BIN1 and CASS4, which are enriched in biological processes and pathways related to known pathogenesis processes such as amyloid precursor protein catabolic process and amyloid-beta formation (Fig 4a), evidencing the robustness of our fine-mapping methods.

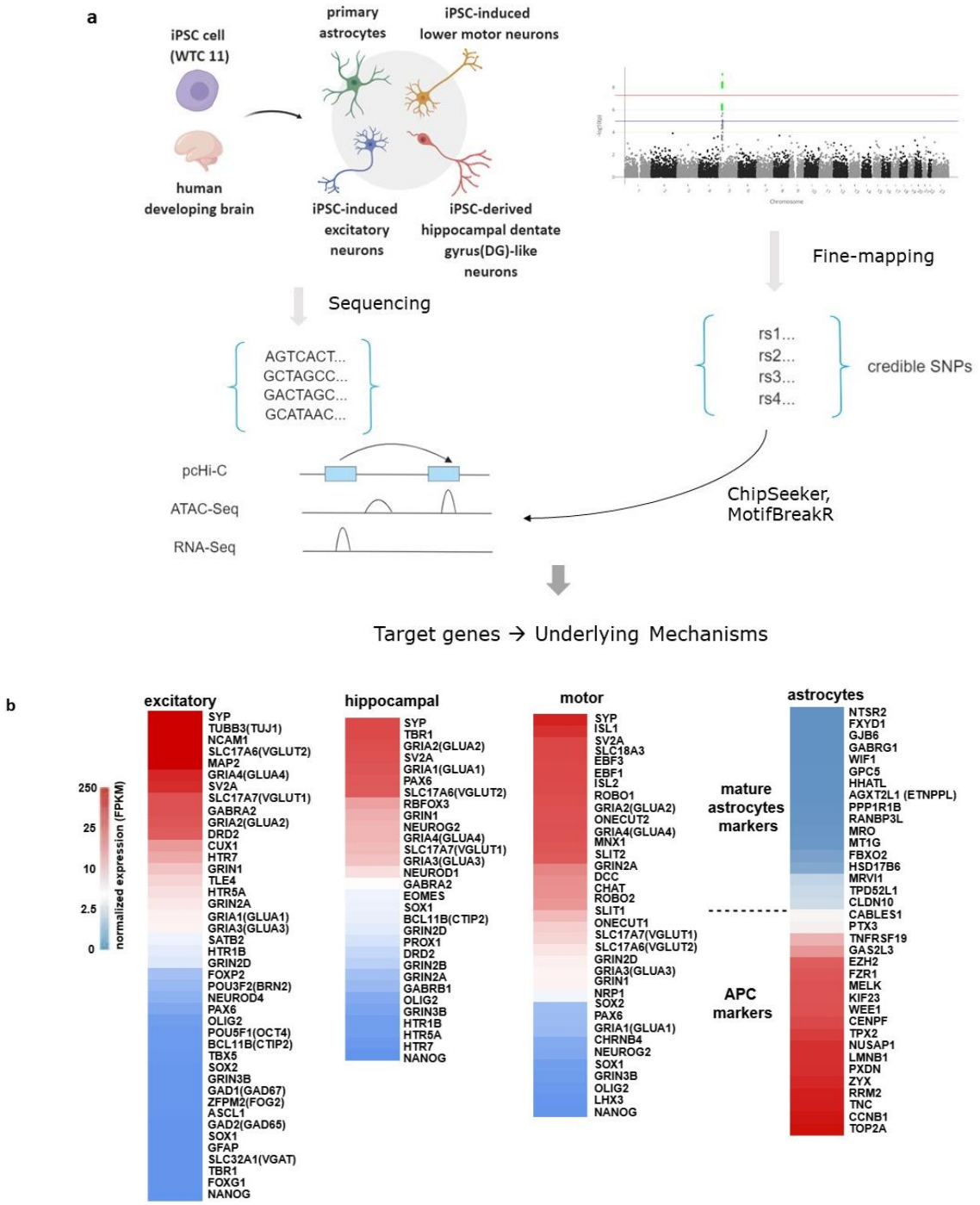


Figure 1.

(a). Schematic of the study design for generating four functionally distinct neural cell types in the CNS and performing integrative analysis of genetic and epigenomic annotations (promoter capture Hi-C, ATAC-seq, and transcriptomes RNA-seq) and fine-mapping.

(b). Heatmaps displaying the expression of key marker genes for the neural cell types. Astrocytes used in this study exhibit an expression profile consistent with APC identity

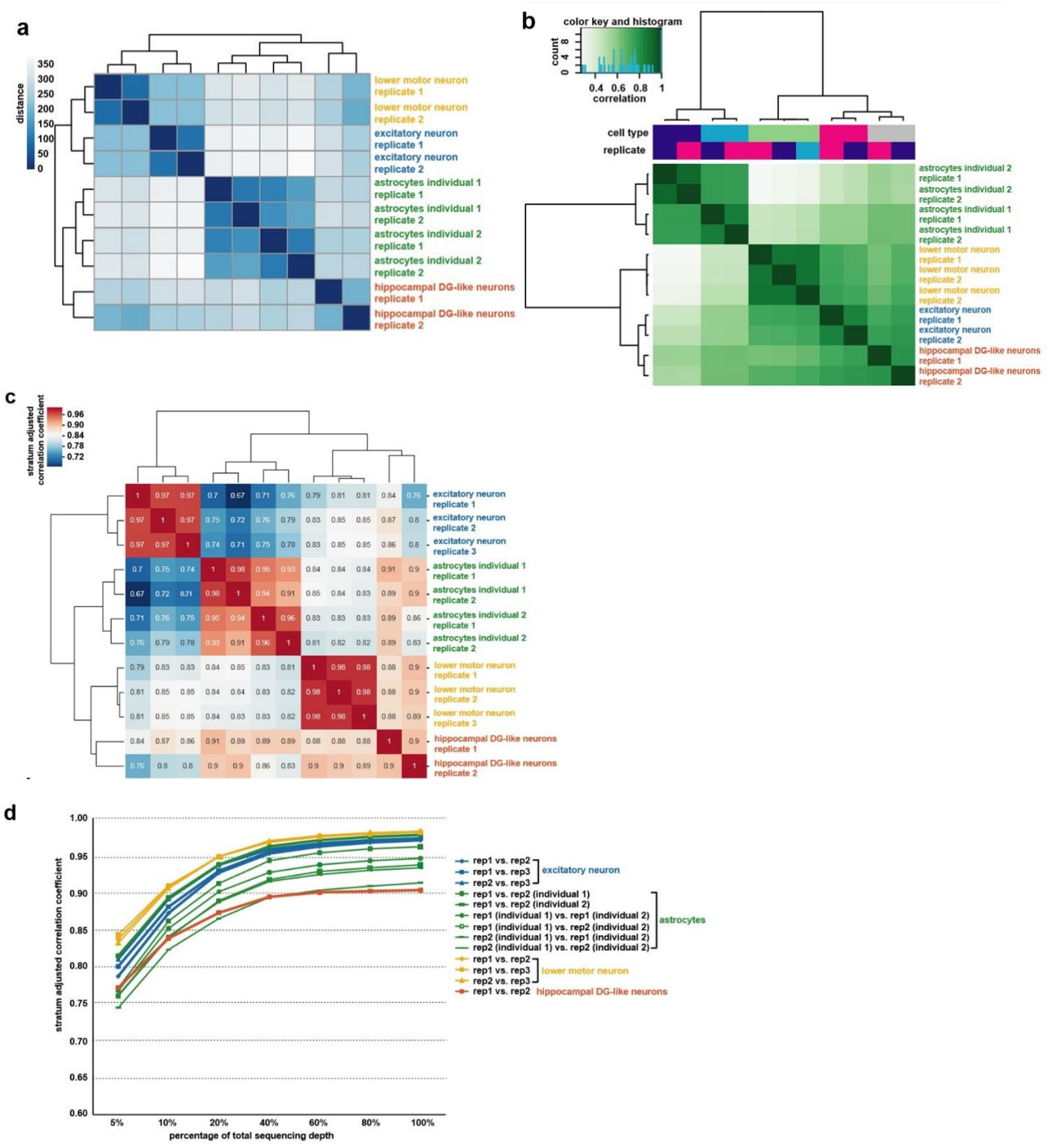


Figure 2. Correlation between pcHi-C, ATAC-seq, and RNA-seq replicates.

(a) Gene expression values for each RNA-seq replicate were hierarchically clustered according to sample distances using DESeq2. (b) Heatmap with pairwise correlations and hierarchical clustering of read densities at the set of unified open chromatin peaks for the ATAC-seq replicates. (c) Heatmap with pairwise correlations based on the stratum-adjusted correlation coefficient (SCC) from HiC-Rep (evaluated at a resolution of 10 kb) for the pcHi-C replicates. (d) Saturation of the SCC between biological replicates for the pcHi-C libraries as a function of total sequencing depth.

Limited by the inherent perplexity of psychiatric diseases, studies on ADHD, ASD, BP and SCZ have not been able to experimentally or clinically validate any clear-cut risk genes or biological processes underlying the pathophysiologic of these diseases, therefore hinder the efforts to determine the accuracy of our methods by directly comparing out list of genes disturbed by putative causal coding variants to published results. Nevertheless, the implicated genes did show enrichment in some pathways reported to be possible to be involved in these diseases. For example, Response to Nicotine and Nicotine Activity on Dopaminergic Neurons pathway is enriched in putative schizophrenia associated genes (Fig 4b), and dopaminergic neurons are known to play a role in the pathology of schizophrenia and studies have linked daily tobacco use to increased risk of psychosis and an earlier age at onset of psychotic illness[21]. Also, the enrichment in regulation of amyloid-beta clearance suggests that the neurodegeneration in patients with schizophrenia may partly share similar mechanisms with Alzheimer's disease. We found strong enrichment in alpha-linolenic acid and linoleic acid metabolic process in Bipolar disorder associated genes (Fig 4c), which are in line with previous findings that lower erythrocyte membrane levels of eicosatetraenoic acid (EPA) and docosahexaenoic acid (DHA), two omega-3 fatty acids derived from alpha-linolenic acid (ALA), were observed in patients with a diagnosis of bipolar disorder[22]. We also found an enrichment in Serotonin Receptor signaling related pathways for fine-mapped coding variants of Bipolar disorder, in line with the previous findings that loss of prefrontal cortex 5-HT1A receptors results in a depression-like phenotype[23].

In summary, we built a customized pipeline for performing genome-wide fine-mapping using FINEMAP and have shown that putative credible coding variants identified by this approach successfully recapitulate well-studied risk genes and possible pathways involved in the pathology of several neurological disorders, evidencing the feasibility of using our fine-mapping results for doing further annotations and investigations.

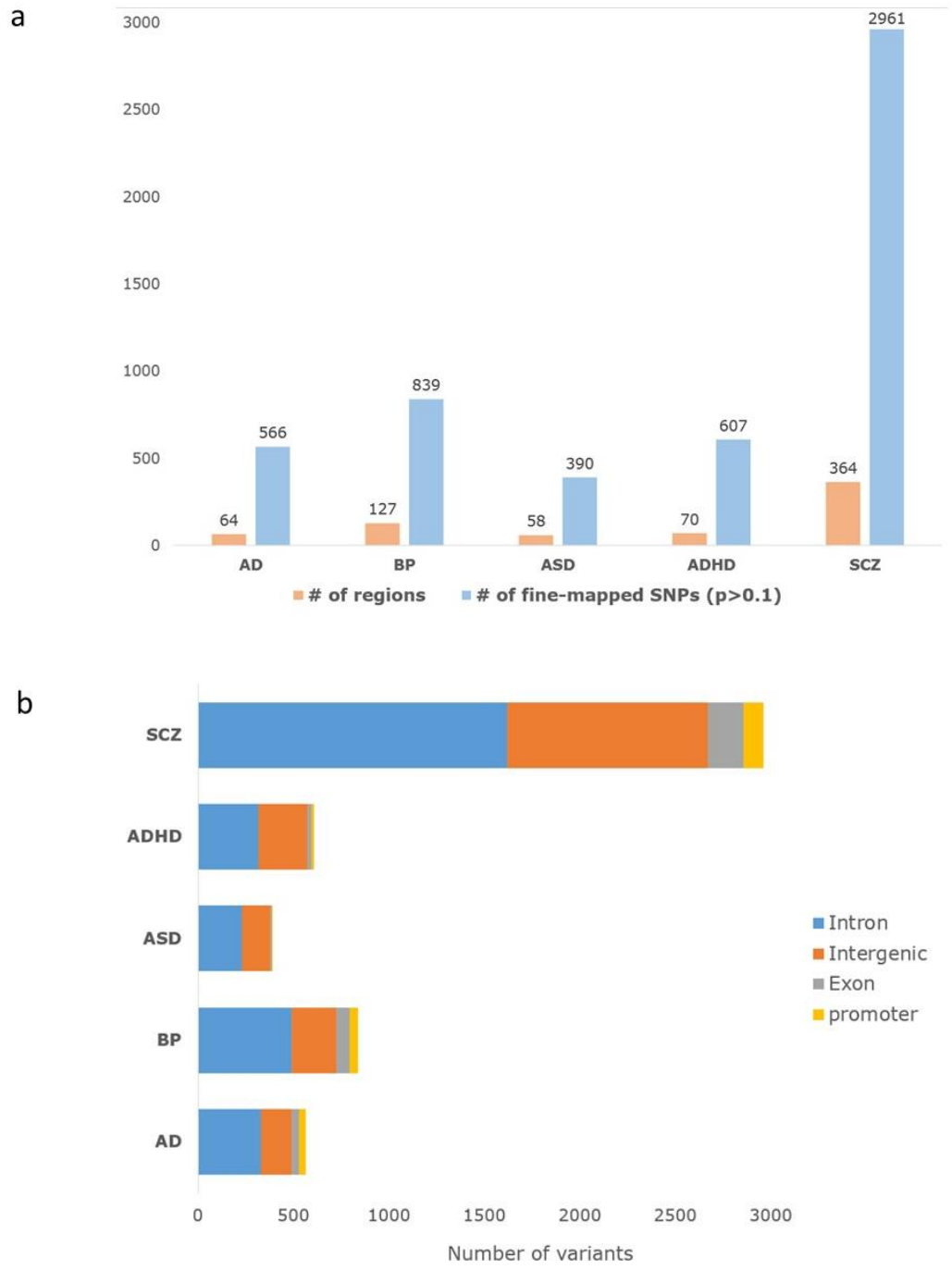


Figure 3.

(a). Number of regions (loci) constructed and number of fine-mapped SNPs (p>0.1) generated.

(b). Localization of fine-mapped SNPs in four genetic catagories.

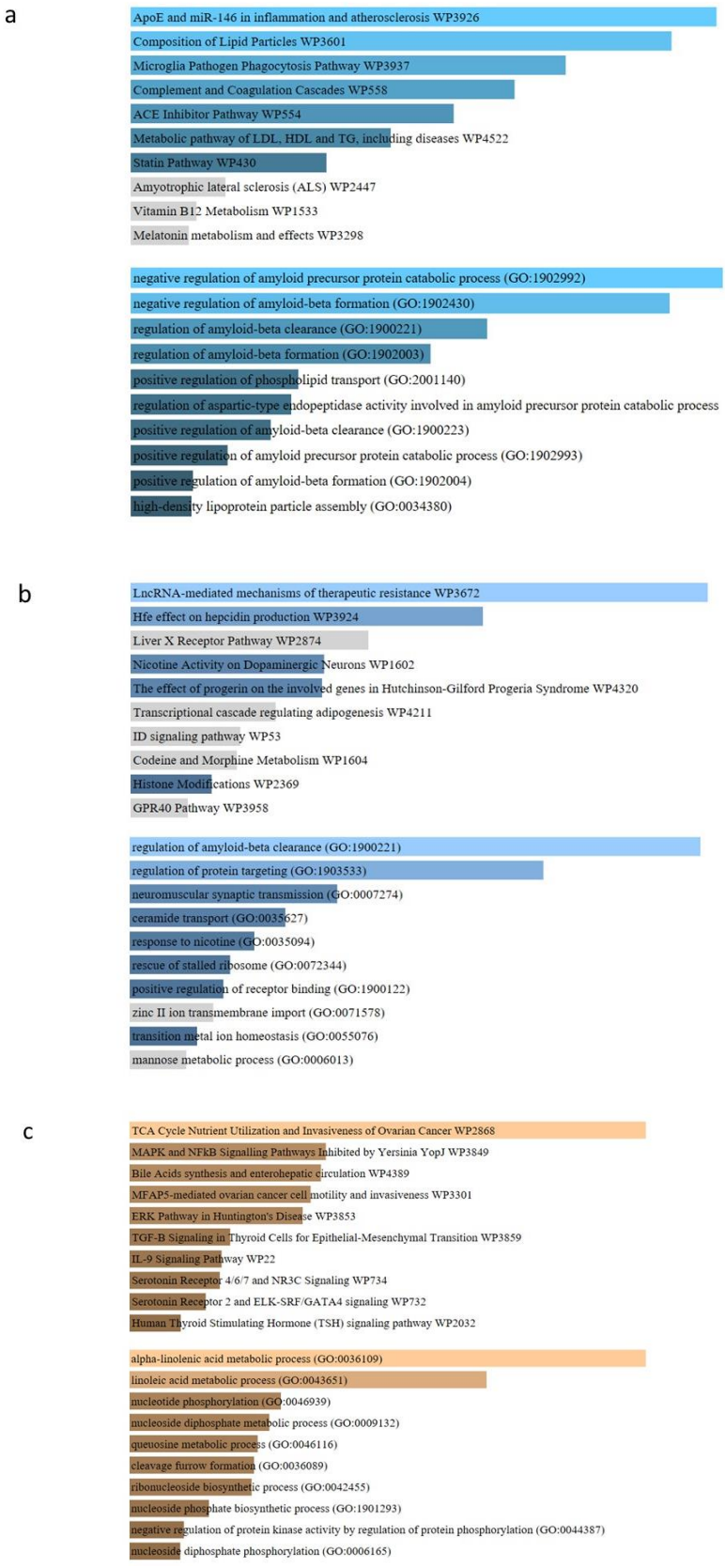


Figure 4. Go enrichment for coding variants (Up, Pathway enrichment; Down, Biological Process enrichment), ranked by combined score. Darkness of color reflects P value (Darker color indicates smaller P value)

(a). Alzheimer's disease

(b). Schizoprenia

(c). Bipolar disorder

Non-coding variants may affect gene expression by perturbing TF binding

To dissect the functional mechanisms underlying putative non-coding causal variants, we first investigated whether the variants may play a role in disrupting at least one of 426 motifs corresponding to known binding preferences for human transcription factors. We identified an average of 526 (AD, 272; BP, 417; ASD, 176; ADHD, 277; SCZ, 1489) non-coding variants that may function by perturbing the transcription factor binding sequences (Fig 5a).

We observed the enrichment of binding sites for FOXJ3, IRF5, ARID3A, HMGA2 in fine-mapped SNPs of all five diseases and FOXJ3 is the most enriched among all not (Supplementary fig 3). However, the biological relevance of the significant enrichment of FOXJ1 need further investigation.

After removing motifs shared by all five diseases, we obtained a list of relatively "disease-specific" disturbed transcription factors whose binding sites are likely to be disturbed (Fig 6) and identified several transcription factors that may play a role in the pathogenesis of the diseases. For example, in Alzheimer's disease, we found strong enrichment of ZBTB4, which is a transcriptional repressor that regulates the cell cycle, including the apoptotic response to amyloid beta and has been associated with the age of onset of Alzheimer's disease[24]. We observed a cluster of interferon regulatory factors (IRF) which are known to be important for immune responses enriched in both ASD and BP, which agree with results of earlier studies that immuno-inflammatory in the brain and periphery in the etiopathogenesis of this illness[25-26] . We also found an enrichment of FOXO1 in SCZ, and FOXO1 in dopaminergic neurons has been shown to regulate energy and glucose homeostasis[27].

We next investigated whether the motifs identified are enriched in known pathways or biological processes. Interestingly, we found the enrichment of prior disease pathway in top enriched transcription factors in Alzheimer's disease, which may provide a support for a recently debated hypothesis that amyloid beta protein's behavior in patients with Alzheimer's disease resemble that of prions. However, we did not observe any seemingly related pathways being enriched in other four diseases, presumably due to the fact that most transcriptional factors have very general functions and therefore thorough context-specific analyses are needed to pinpoint genes they are regulating in order to understand of how they function in terms of pathogenesis of these diseases.

In summary, we identified a few motifs whose normal functions might be disturbed by the change of binding site sequence because of the variants, which may serve to re-emphasize the increasingly

accepted notion of the importance of variation in TF-DNA binding in mediating phenotypic diversity which has been extensively discussed in previous works[28].

a

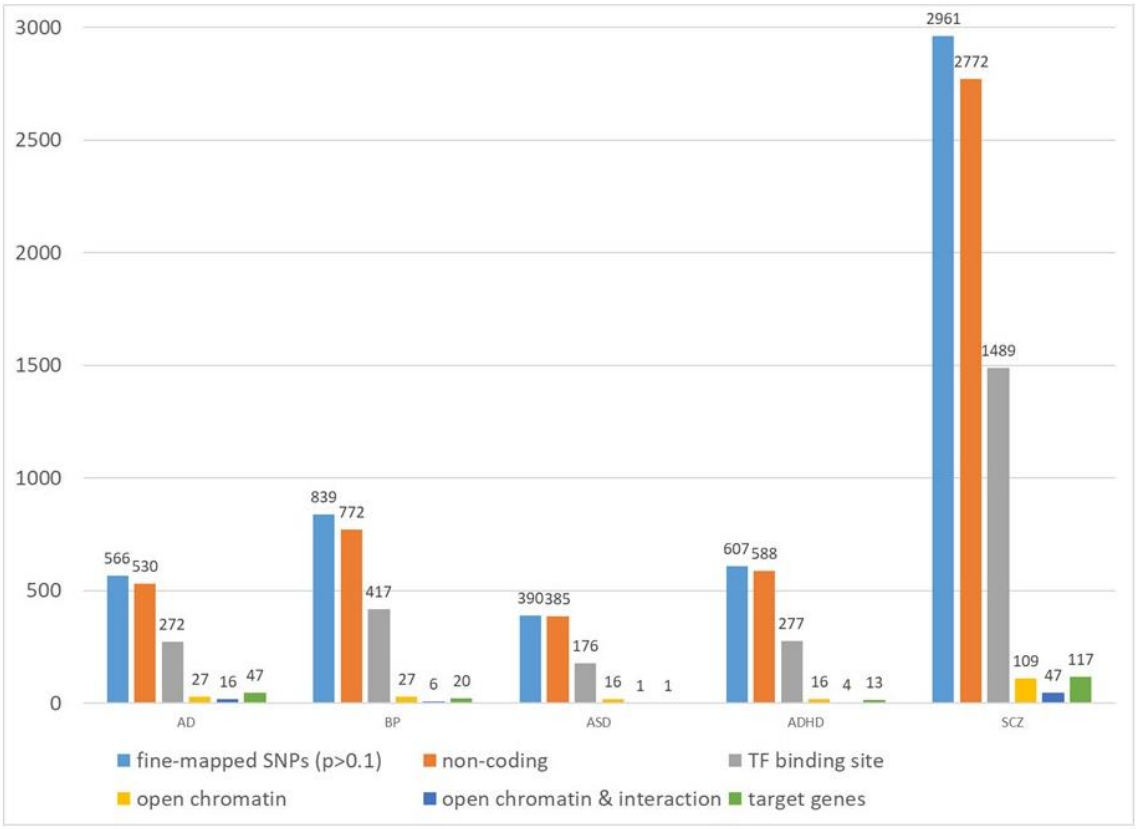


Figure 5. Annotation of non-coding variants

fine-mapped SNPs: number of SNPs with a posterior probability larger than 0.1 (calclulated by FINEMAP)

non-coding: number of SNPs that do not overlap with any known exon regions

TF binding site: number of SNPs which overlap with predicted transcription factor binding site

open-chromatin: number of SNPs which overlap with ATAC-Seq peaks

open chromatin & interaction: number of SNPs which overlap with ATAC-Seq peaks and in the same time overlap with physical interaction bins

target genes: number of target genes nominated by our integrated approach

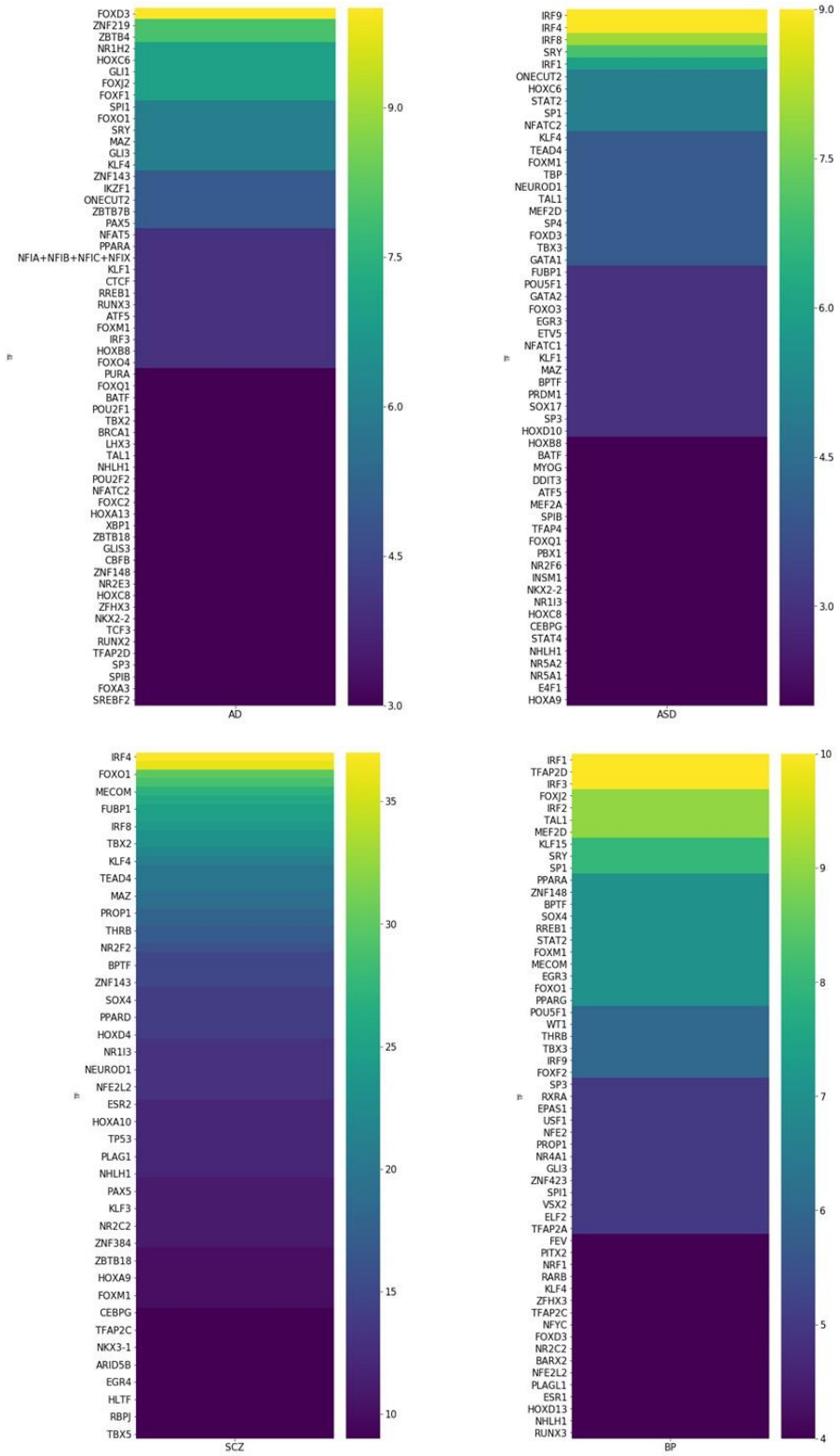


Figure 6. Transcription factors whose binding site are likely to be disturbed by disease-associated SNPs. Color shows the number of SNPs predicted to disturb the binding of each transcription factor.

Non-coding variants regulate distal target genes through physical chromatin interactions

We next investigated whether fine-mapped non-coding variants may influence the gene expression by affecting function of cis-regulatory elements. We identified an average of 14 variants (AD, 16; BP, 6; ASD, 1; ADHD, 4; SCZ, 47) which may have implications in gene regulation through physical chromatin interactions, and an average of 39 genes (AD, 47; BP, 20; ASD, 1; ADHD, 13; SCZ, 117) for each disease predicted to be regulated. Notably, we successfully recapitulated the previously experimentally validated[29] connection between putative SCZ causal variant rs1191549, which falls into intergenic region, and FOXG1, which is a well-known risk gene for schizophrenia, in hippocampal DG-like neurons (Fig 7a). This is a clear-cut example of how non-coding variant may affect risk gene expression by interrupting distal regulatory element and again proves the usefulness and robustness of our approach.

We also identified another candidate AD variant rs405509 which falls into the intron region of APOE and is connected to both TOMM40 and APOC2 (Fig 7b). The TOMM40-APOE-APOC region has been reported to be correlated to AD-related biomarkers. However, whether this variant play a role in orchestrating TOMM40 and APOC2 needs further validations.

Variants can influence gene expression through chromatin interactions

As we have shown that integrative analyses of fine-mapping and chromatin interaction data was successful at capturing cell type-specific regulatory mechanisms of likely causal variants, we were interested to see if physical chromatin interactions are also able to mediate the effects of other types of cis-acting regulatory variants such as expression quantitative trait loci (eQTLs) on gene expression.

To test this hypothesis, we demonstrate that the mean scores for interactions overlapping significant eQTL-TSS pairs are significantly higher than the scores for interactions overlapping randomly shuffled eQTL-TSS pairs (Kolmogorov-Smirnov test, p=2.28x10-4 for excitatory neurons and p=1.76x10-6 for hippocampal DG-like neurons) (Fig. 8). This indicates that significant promoter-PIR interactions recapitulating regulatory relationships between eQTL-TSS pairs are called with increased levels of confidence. Our results present orthogonal lines of evidence that variants can influence gene expression through the formation of chromatin interactions.

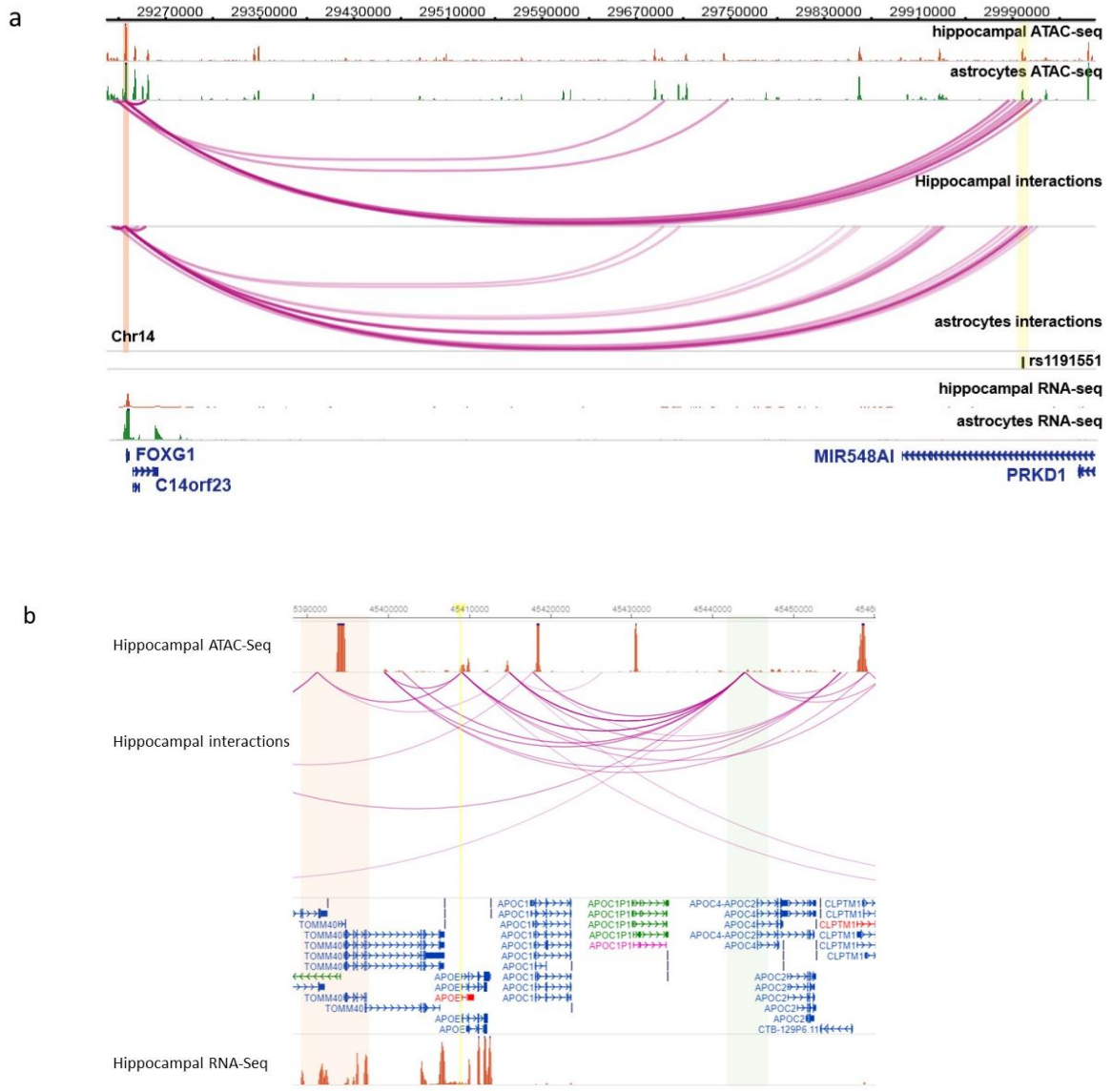


Figure 7. Using chromatin interactions to elucidate the functions of GWAS variants.

(a) Significant promoter-PIR interactions in hippocampal DG-like neurons and astrocytes recapitulate a previously reported interaction between the FOXG1 promoter and a distal open chromatin peak containing rs1191551, a schizophrenia-associated variant

(b) Significant promoter-PIR interactions in hippocampal DG-like neurons identified a potential regulatory element containing r s405509, a Alzheimer's disease-associated variant.

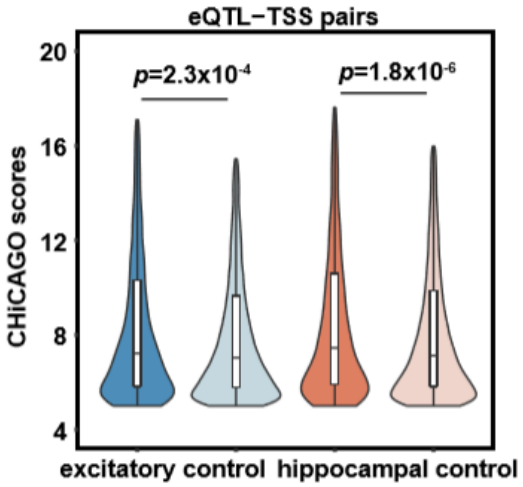


Figure.8

Distributions of interaction scores for chromatin interactions overlapping significant eQTL-TSS pairs versus randomly sampled nonsignificant eQTL-TSS pairs in excitatory neurons and hippocampal DG-like neurons. Interaction scores are significantly enriched for significant eQTL-TSS pairs (two sample t-test, one-sided, p=2.3x10-4 for excitatory neurons and p=1.8x10-6 for hippocampal DG-like neurons). (

Methods and Materials

Promoter capture Hi-C(pcHi-C) and data processing

The library preparations of promoter capture Hi-C were carried out by other members in the lab. Briefly, *in situ* Hi-C libraries for the excitatory neurons, hippocampal DG-like neurons, lower motor neurons and astrocytes were constructed from 1 to 2 million cells using HindIII as a restriction enzyme. pcHi-C was performed using biotinylated RNA probes prepared according to an established protocol.

The final library was eluted, amplified, then sent for paired-end sequencing on the HiSeq 4000 (50 bp reads), the HiSeq X Ten (150 bp reads), or the NovaSeq 6000 (150 bp reads).

Paired-end sequencing reads were first trimmed using fastp running the default settings before being mapped, filtered, and deduplicated using HiCUP v0.71 with bowtie2 and filtering for ditags between 100 and 1200 bp. In addition, the sequencing depths of all libraries was normalized so that each replicate had the same number of usable reads, or the number of on-target cis pairs interacting over a distance of 10 kb. Significant promoter-PIR interactions were called using CHiCAGO running the default settings, retaining baited fragments that are supported by at least 250 reads

(minNPerBaits=250). Promoter-PIR interactions between HindIII fragments with a score (negative log p-value) of 5 or greater in each cell type were determined to be significant.

ATAC-Seq and data-processing

The library preparations of ATAC-Seq were carried out as previously described using the Nextera DNA Library Prep Kit (Illumina #FC-121-1030) by other members in the lab.

Libraries were sent for single-end sequencing on the HiSeq 4000 (50 bp reads) or paired-end sequencing on the NovaSeq 6000 (150 bp reads). ATAC-Seq sequencing reads were mapped to hg19 and processes using the ENCODE ATAC-Seq pipeline (https://github.com/kundajelab-/atac_dnase_pipelines) running the default settings. Briefly, reads were mapped by bowtie2 and peaks were called by MACS2 running default settings. Only the first read was used, and all sequencing reads were trimmed to 50 bp prior to mapping. Open chromatin peaks called by the pipeline were expanded to a minimum width of 500 bp for all downstream analyses.

RNA-Seq and data-processing

The library preparations of RNA-Seq were carried out by other members in the lab using RNeasy Mini Kit (Qiagen #74104).

Libraries were sent for single-end sequencing on the HiSeq 4000 (50 bp reads) or paired-end sequencing on the NovaSeq 6000 (150 bp reads). Raw sequencing reads were aligned to hg19/GRCh37 using STAR running the standard ENCODE parameters, and transcript quantification was performed in a strand-specific manner using RSEM with the annotation from GENCODE 19. Only the first read was used, and all sequencing reads were trimmed using TrimGalore 0.4.5 running the following options: -q 20 --length 20 --stringency 3 --trim-n. The mean gene expression across all replicates was used for each cell type.

Gene expression profiles for each cell type were plotted using R package heatmap2 and the list of reference signature genes were extracted from original papers of cell line differentiation for each cell type.

Reproducibility and saturation analysis

We took pcHi-C contact matrices generated at 10 kb resolution using HiC-Pro 2.11.0 with the following settings (MIN_MAPQ=20, MIN_FRAG_SIZE=100, MAX_FRAG_SIZE=5000000, MIN_INSERT_SIZE=100, MAX_INSERT_SIZE=1200, and reporting only bin pairs that are baited on at least one end with our pcHi-C probes, with all other settings set to their default values) and calculated the pairwise stratum adjusted correlation coefficient (SCC) between replicates across all cell types using HiCRep 1.4.0 on chromosome 1 (h=20 and only considering contacts with distances below 5 Mb, the optimal smoothing parameter, h, was estimated by 'htrain' function implemented in HiCRep using the same contact matrices resolution as calculating SCC. SCCs evaluated on the other chromosomes and with contact matrices generated at other resolutions closely resembled the results for chromosome 1, 10kb contact matrices (data not shown). Hierarchical clustering for the pairwise SCC values was performed using the Seaborn clustermap function in Python.

Pairwise correlation heatmaps and clustering dendrograms for ATAC-seq replicates were generated by counting reads overlapping a set of consensus peaks using the DiffBind package in R, with the set of consensus peaks defined as peaks occurring in at least two replicates across all cell types (minOverlap=2).

Pairwise distance estimates and clustering dendrograms for RNA-seq replicates were generated using the DESeq2 package in R running the default settings.

For saturation analysis, we first down-sampled all promoter capture Hi-C libraries (raw BAM files) to 5%, 10%, 20%, 40%, 60%, 80%, and 100% of the final sequencing depths used in the study using samtools. Next, we generated 10kb contact matrices on chromosome 1 using the same settings running HiC-Pro mentioned above for each down-sampled library. We then computed pairwise SCCs between all pairs of biological replicates using HiCRep at these down-sampled sequencing depths.

Calling significant promoter-PIR interactions

Paired-end sequencing reads were first trimmed using fastp running the default settings before being mapped, filtered, and deduplicated using HiCUP v0.71 with bowtie2. Significant promoter-PIR interactions were called using CHiCAGO running the default settings, retaining baited fragments that are supported by at least 250 reads (minNPerBaits=250). Promoter-PIR interactions between HindIII

fragments with a score (negative log p-value) of 5 or greater in each cell type were determined to be significant.

Fine-mapping

GWAS summary statistics were downloaded from PGC or from links provided in the paper of the original study.

Fine-mapping is often performed at a per-locus scale due to high computational complexity and convolutional linkage structures in *trans*. To split the genome-wide summary statistics into regions readily available for performing fine-mapping, we applied a split-by-distance strategy. For each chromosome, we first filtered the SNPs based on p-values with a cutoff of $p<10^{-6}$, and the significant SNPs retained after filtering were used to construct initial distance-defined loci: given a total of n SNPs after filtering on a chromosome, for any two consecutive SNPs SNP(i) and SNP(i+1) (i < i+1 <=n), if the distance between SNP(i) and SNP(i+1) was larger than 1,000,000 base pair, then the two SNPs were categorized into two different loci. Next, to take into consideration the LD structures, we included the LD buddies ($r^2 > 0.6$) of each SNPs in each locus. Pairwise LD annotations were retrieved from Trans-Omics for Precision Medicine (TOPMed) European sub-population and was initially trimmed with a cutoff of $r^2 > 0.1$, so any pairwise r^2 that was smaller than 0.1 in original TOPMed data was denoted as 0. Each resulting locus, containing both significant SNPs and their LD buddies, was used to generate input files for FINEMAP.

We next calculated minor allele frequency (MAF) for every SNPs include in the loci. 1000 genome phase 3 European sub-population genotyping data was extracted with “bedtools view” and MAF was calculated with “plink –freq”. SNPs with a MAF of 0 in European sub-population were excluded for further analyses. For summary statistics that included MAF in the original report, the original MAF were removed and MAF calculated with 1000 genome was used instead. Since FINEMAP can only be applied to regions containing a minimum of 3 SNPs, summary statistics of SNPs belonging to loci containing only one or two SNPs were saved for later analyses. Z-scores of all the eligible SNPs were calculated with custom script. Pairwise LD matrix, another input file required by FINEMAP, was generated with custom script using the same TOPMed LD annotations.

We then applied FINEMAP v1.3 software to all the eligible loci with default settings. For regions with a size smaller than two hundred SNPs, the maximum number of SNPs allowed to be causal equals

to the size of the region. For those containing more than two hundred SNPs, the number of allowed causal SNPs are 0.2 times the number of SNPs in total. Shotgun Stochastic search was used.

The output of FINEMAP included (i) a configuration file which listed potential causal configurations together with their posterior probabilities and Bayes factors, (ii) the posterior probability that there is a specific number (between 1 and maximum number allowed) of statistical independent associations (causal variants) in the fine-mapped region, and (iii) a file containing the marginal Posterior Inclusion Probability (PIP) for SNP(i) (0<i<n) to be causal. PIPs were computed by summing up the posterior probability of all causal configurations in the configuration file. We took the configuration file for each locus and extracted SNPs included in the top configuration, together with other statistics of those SNPs included in the original summary statistics, to generate a SNP-set file for each locus. For downstream analyses, we retained all the variants in the SNP-file that has a posterior probability larger than 0.1% and has a minor allele frequency larger than 0.1%. We also included all the SNPs in regions that were not eligible for performing fine-mapping. Finally, duplications were removed before further analyses.

Functional variants annotation

All the variants retained as mentioned above (p > 0.1%) were annotated with genomic and epigenomic datasets. Variants were first lifted over using rtracklayer package and UCSC chain file. We used ChIPseeker[31] to partition fine-mapped variants into bins of non-overlapping annotations using GENCODE V19 gene annotations as reference. As in many cases the same variant may fall into different categories (Supplementary Fig 4), we classified each variant on the basis of the following hierarchy: (i) exon (ii) promoter (iii) 5'UTR (iv) 3'UTR (v) intron (vi) intergenic, with promoter defined as regions within 500bp upstream/downstream of transcription start site. For example, for a variant falling in a promoter region and an intron region at the same time, that variant was assigned to the "promoter" class.

For chromatin accessibility annotations, we took all the chromatin accessible regions (ATAC-Seq peaks expanded to 500bp) of all four cell types and intersected with each variants-set using "bedtools intersect".

We took all the significant promoter-PIR interactions (CHiCAGO score >5) in four cell types and intersected with each variants-set using "bedtools intersect". Variants may overlap with either promoter region or PIR region (overlapping at least one of the two interacting bins if an interactions)

Gene ontology and pathway enrichment analysis

For genes whose exons are targeted by fine-mapped variants, the Gene ontology enrichment analysis was performed by Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA)[32] GENE2FUNC function. For GO analysis, we took all the fine-mapped high-confidence (p> 0.1) coding variants (classified as "exon" in previous annotation step). All genes were used as the background model, a minimum overlap of two genes and FDR adjusted P<0.05 was required for each gene set. GTEx v7 53 tissue types and 30 general tissue types were both used in differential gene expression tissue specificity enrichment analysis.

For genes that are putative to be target genes regulated by non-coding causal variants, the Gene Ontology and pathway enrichment were carried out by Enrichr [33]. Only biological processes from "GO Biological Process 2018" ontology are reported according to their combined scores (calculated by multiplying the log of the p-value by the z-score of the deviation from the expected rank) and "WikiPathway 2019 HUMAN" was used as the reference for pathway enrichment.

Transcription factor motif analysis

We investigated whether the con-coding fine-mapped variants (annotated as intronic, intergenic or promoters) may play a role in disturbing transcription factor binding using R package MotifbreakR with the default settings, using HOCOMOCO V10 [34] as the motif database. All the fine-mapped variants (p > 0.1) was used to investigate the perturbation of transcription binding motifs. We randomly sampled variants from all non-coding variants with a pp > 0.1% as the control to show differentially perturbed transcription factor binding motifs. The significantly sets of motifs were selected using the following criteria: (i) The number of variants disturbing the motif must exceed the average number of variants disturbing each motif and (ii) The normalized difference ((number of fine-mapped variants disturbing the motif - number of control variants disturbing the motif) / (number of fine-mapped variants disturbing the motif + number of control variants disturbing the motif)) must be larger than the average difference. The heatmap was generated by python seaborn package.

Identification of putative target genes

All the significant interactions (CHiCAGO score > 5) were split into left-hand interacting bins and right-hand interacting bins and were intersected with all promoter regions (defined as regions within 500 bp upstream/downstream of transcription start site, using GENCODE V19 as gene annotation file) respectively.

We consider a SNP to be potentially interacting with a target gene if it (i) falls into chromatin accessible regions in at least one cell type, (ii) in the meantime overlaps with at least one interacting bins in at least one cell type and (ii) the overlapped interacting bins overlap with promoter regions. The putative loci of interested were visualized using WashU Genome Browser.

eQTL enrichment analysis

The tissue-specific SNP-gene association datasets were downloaded from GTEx V7 website (https://gtex-portal.org/home/). We used Brain_Cortex dataset as the match to our excitatory neurons and Brain_Hippocampus dataset as the match to our hippocampal DG-like neurons. Both the significant pairs, which was filtered based on permutations, and the full association datasets were used in this analysis.

To determine the 2D enrichment of eQTL-TSS pairs in our significant interaction sets, we first filtered out eQTL-TSS pairs that were within 10 kb of each other or on the same HindIII fragment as this would be below the minimum detectable resolution by pcHi-C. Next, we sampled a set of nonsignificant eQTL-TSS pairs with a matching distance distribution as the set of significant eQTL-TSS pairs for each cell type as the control set. We also controlled for the number of genes around which the eQTL-TSS pairs were centered. Finally, we compared the distributions of scores for significant interactions supporting the significant and nonsignificant sets of eQTL-TSS pairs by overlapping the eQTL-TSS pairs with our significant interactions.

Discussion

In the past decade, thousands of genetic variants have been identified by GWAS. Meanwhile, only a very small proportion of them are coding variants and whose functions are relatively well understood. In the past few years, there has been a rapidly increased recognition and appreciation for the term “post-GWAS era”, which refers to the research efforts aiming to improve the functional annotations of regulatory variants and to link putative causal variants to their cognate genes. However, several factors make it difficult to bridge the gap between statistical associations and biological mechanisms, including the fact that the true causal variants may often not be the ones with the most significant associations, and the observations that most disease-associated variants fall into non-coding part of the genome.

In this study, we used Bayesian frameworks implemented in the software FINEMAP for the initial prioritization of causal variants to explain association signals. Next, we leverage pcHi-C, ATAC-seq, and RNA-seq to derive annotations for previously uncharted regulatory relationships between promoters and distal regulatory elements in cell types that are relevant to neurological disease. We then compiled functional annotations including transcription factor binding site and aforementioned cis-regulatory elements to nominate potential underlying mechanisms of putative causal variants.

Using this pipeline, we pinpointed hundreds of likely causal variants using fine-mapping and nominated several mechanisms through which the variants function to influence gene expression and in the end traits and phenotypes.

However, there are also several limitations in this approach that need to be addressed in future works. First, three of the cell types we used in this study were derived from iPSC cells and may not comprehensively reflect the true biological features of those cell in vivo. Given that pcHi-C experiments require a large amount of cells to perform and it is currently very difficult to isolate those cell types from primary brain tissue due to heterogeneity of brain cell types, using iPSC-derive cell models might be the best solution for this challenge at the moment and this situation may improve upon the development of more advanced capture-C technology, the more comprehensive understanding and characterization of neural cell types and better understanding and modeling of differences between iPSC-derived cells and primary cells.

Second, as LD structures and minor allele frequencies differ across populations, for fine-mapping analyses in this study, we only selected studies which performed GWAS on European populations, and we used 1000 Genome European sub-population to calculate minor allele frequency and TOPMed

European sub-population to derive LD annotations of SNPs in summary statistics for all five diseases. However, even after controlling for the sub-population, bias is not completely eliminated. For example, the reference panel we used may be different from the ones used in original studies for genotyping imputation. Also, the LD annotations were not derived from original GWAS populations but from reference panel instead. Both of those may lead to biases and decreased accuracy in terms of fine-mapping results[35].

Moreover, we only used FINEMAP to perform fine-mapping, while there are many other fine-mapping methods are being developed such as susieR (https://stephenslab.github.io/susieR), and it would be interesting to try different pipelines in the next steps and compare results.

We also noticed that only a very limited proportion of SNPs were nominated to play a role in regulating gene expression through chromatin physical interactions. This observation can be partly explained by the fact that we used cell-type specific annotations, and many SNPs may not function in the cell types available in this study. For instance, microglia may be a more relevant cell type in terms of Alzheimer's disease and it has been reported that ASD is more associated with inhibitory neurons[36]. This again emphasizes the importance of interrogating variants in a cell type-specific manner.

It is also worth noting that we proposed in this paper several mechanisms by which causal SNPs may function, such as disturbing transcription factor binding or affecting chromatin interactions, and it is intuitive to assume that the genes whose expression are directly influenced are most responsible for traits and phenotypes observed. However, the gene regulation network can often be very convoluted, and the seemingly most likely causal genes might only affect the phenotypes in a rather indirect way. Therefore, it is important to take cautions when interpreting the findings and applying them to clinical settings.

At the end of this paper, we introduced our preliminary results for integrating eQTL datasets. We demonstrated that chromatin interactions contribute to the influence of eQTL SNPs on gene expression. The next interesting and natural step to take is to incorporate eQTL SNPs and fine-mapped GWAS SNPs and investigate whether it would increase the power of identifying causal variants and cognate genes, and this is indeed one of our next steps planned.

Finally, the methods used in this study are all based on computational approaches, and in the end experimental validations are needed to confirm the findings. In fact, the putative causal variants of Alzheimer's disease identified in this study have been used to design sgRNA for single cell CRISPR

screening in an effort to identify regulatory elements such as enhancers of AD associated risk genes. This benchmarking process will offer great feedback needed for improving the computational pipeline in order for it to reach better performances when guiding experimental designs.

Acknowledgement

First, I would like to thank my thesis mentor Dr. Yin Shen for generously providing me with the opportunity to study at this amazing lab and to contribute to exciting cutting-edge researches, and for creating this warm, supportive working environment which I had been longing for for a very long time. For a very introverted person like me, it surprises me how much I have enjoyed spending time with all the lab members. I feel extremely lucky to have met an incredibly wonderful mentor at the very beginning of my science career. I would also like to thank everyone in our group, with a special thank you to the very talented graduate student Michael Song, for giving me the best trainings and for always being extremely patient with me even when I had lost faith in myself. I have learned from him more things than I had ever expected. I cannot believe less than a year ago I was someone who would get panic at literally every computational task and now I have grown into someone who knows how to find her towels. Also, the Hyperion series is indeed one of the very best.

I would like to thank my mentor at SCU, Dr. Zhenxiang Xi, for providing me with the platform to study bioinformatics and for helping me along my way of pursuing further studies.

I would also like to thank all our collaborators, especially Dr. Yun Li from UNC who offered great guidance for building fine-mapping pipeline. It would have been impossible for me to finish this project without her help.

I would like to thank all my friends, especially Diansheng Sun, Xu Han, Xuetong Qu, Wei Lin and Zhuoxi Liang, for accepting me as who I am, which I believe is not that easy, and for being by my side during my ups and downs. My four years would have been miserable without them.

I would also like to thank Wolfgang Amadeus Mozart, Vincent Van Gogh, Pierre-Auguste Renoir and other impressionists, and Joan Miró, for whose works offered me great spiritual supports.

I would like to thank Bayern Munich, the football team that has accompanied me for almost ten years. We will stand strong, be it in good or bad times.

Finally, I would like to thank my parents, for always supporting me unconditionally in every way they can. I love you mom and dad.

I hope this work marks the beginning of my science journey, and I would like to dedicate it to my grandfather, who always had the deepest faith in me, and believed I would realize my dreams and fight for the improvement of human health. I will.

Reference

1. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
2. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).
3. Evangelou, E. & Ioannidis, J. P. A. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* **14**, 379–389 (2013).
4. Pritchard, J. K. & Przeworski, M. Linkage Disequilibrium in Humans: Models and Data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
5. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190–1195 (2012).
6. Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. *Hum. Mol. Genet.* **24**, R111–R119 (2015).
7. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
8. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).

9. Kichaev, G. *et al.* Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLOS Genet.* **10**, e1004722 (2014).

10. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying Causal Variants at Loci with Multiple Signals of Association. *Genetics* **198**, 497–508 (2014).

11. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.* **16**, 144–154 (2015).

12. Amlie-Wolf, A. *et al.* INFERNO: inferring the molecular mechanisms of noncoding genetic variants. *Nucleic Acids Res.* **46**, 8740–8753 (2018).

13. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).

14. Amlie-Wolf, A. *et al.* Inferring the molecular mechanisms of noncoding Alzheimer's disease-associated genetic variants. (2018). doi:10.1101/401471

15. Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25**, 582–597 (2015).

16. He, X. *et al.* Sherlock: Detecting Gene-Disease Associations by Matching Patterns of Expression QTL and GWAS. *Am. J. Hum. Genet.* **92**, 667–680 (2013).

17. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).

18. Yang, T. *et al.* HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. (2017). doi:10.1101/101386

19. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2016).

20. Ulirsch, J. C. *et al.* Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* 1 (2019). doi:10.1038/s41588-019-0362-6

21. Gurillo, P., Jauhar, S., Murray, R. M. & MacCabe, J. H. Does tobacco use cause psychosis? Systematic review and meta-analysis. *Lancet Psychiatry* **2**, 718–725 (2015).

22. Bozzatello, P., Brignolo, E., De Grandi, E. & Bellino, S. Supplementation with Omega-3 Fatty Acids in Psychiatric Disorders: A Review of Literature Data. *J. Clin. Med.* **5**, (2016).

23. Garcia-Garcia, A. L. *et al.* Serotonin Signaling through Prefrontal Cortex 5-HT1A Receptors during Adolescence Can Determine Baseline Mood-Related Behaviors. *Cell Rep.* **18**, 1144–1156 (2017).

24. Blue, E. E. *et al.* Variants regulating ZBTB4 are associated with age-at-onset of Alzheimer's disease. *Genes Brain Behav.* **17**, e12429 (2018).

25. Muneer, A. Bipolar Disorder: Role of Inflammation and the Development of Disease Biomarkers. *Psychiatry Investig.* **13**, 18–33 (2016).

26. Kern, J. K., Geier, D. A., Sykes, L. K. & Geier, M. R. Relevance of Neuroinflammation and Encephalitis in Autism. *Front. Cell. Neurosci.* **9**, (2016).

27. Doan, K. V. *et al.* FoxO1 in dopaminergic neurons regulates energy homeostasis and targets tyrosine hydroxylase. *Nat. Commun.* **7**, 12733 (2016).

28. Deplancke, B., Alpern, D. & Gardeux, V. The Genetics of Transcription Factor DNA Binding Variation. *Cell* **166**, 538–554 (2016).

29. Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523–527 (2016).

30. Nowakowski, T. J. *et al.* Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**, 1318–1323 (2017).

31. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).

32. Watanabe, K., Taskesen, E., Bochoven, A. van & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).

33. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).

34. Kulakovskiy, I. V. *et al.* HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **46**, D252–D259 (2018).

35. Benner, C. *et al.* Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies. *Am. J. Hum. Genet.* **101**, 539–551 (2017).

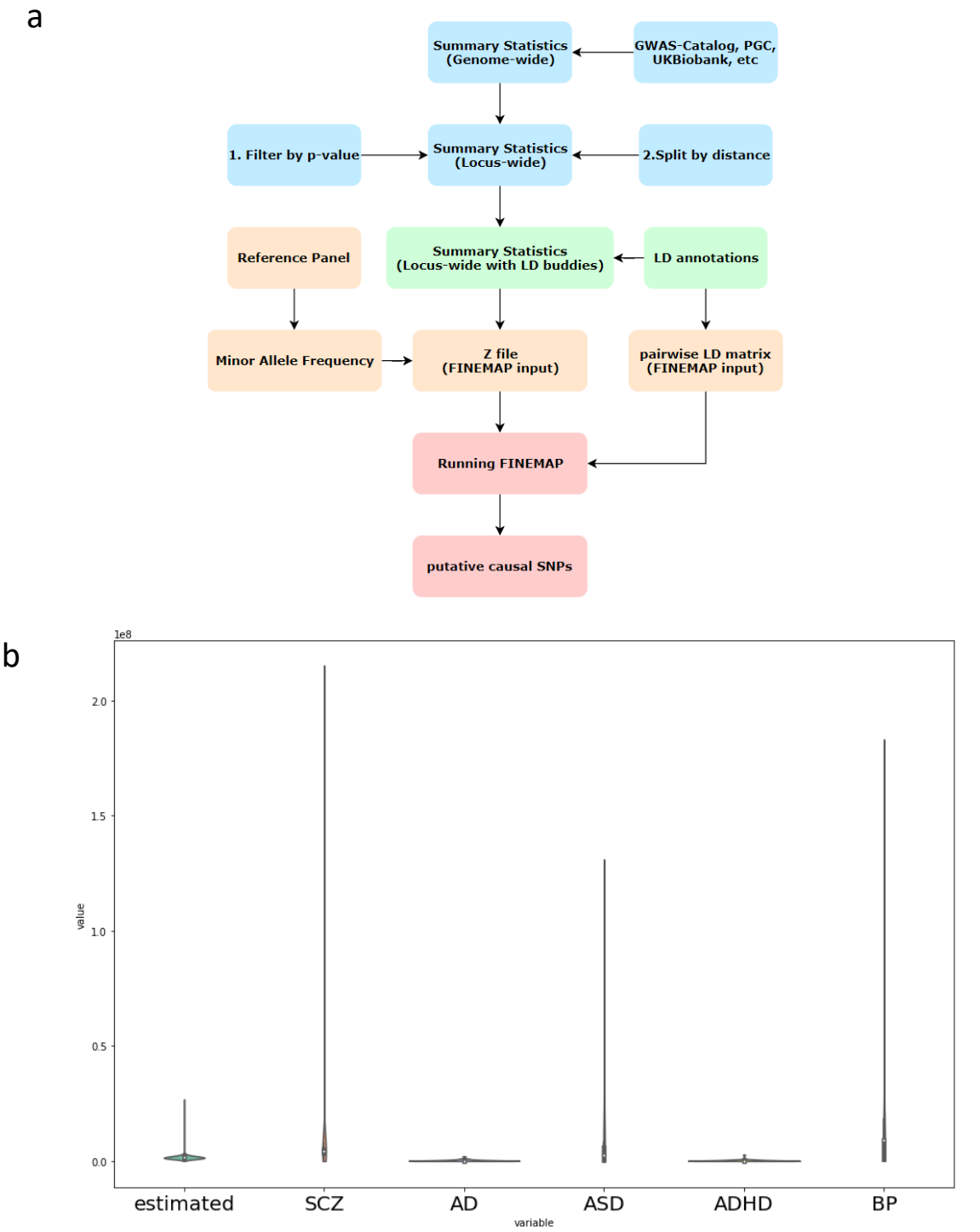
36. Wang, P., Zhao, D., Lachman, H. M. & Zheng, D. Enriched expression of genes associated with autism spectrum disorders in human inhibitory neurons. *Transl. Psychiatry* **8**, 13 (2018).

Supplementary Information

Code Availability

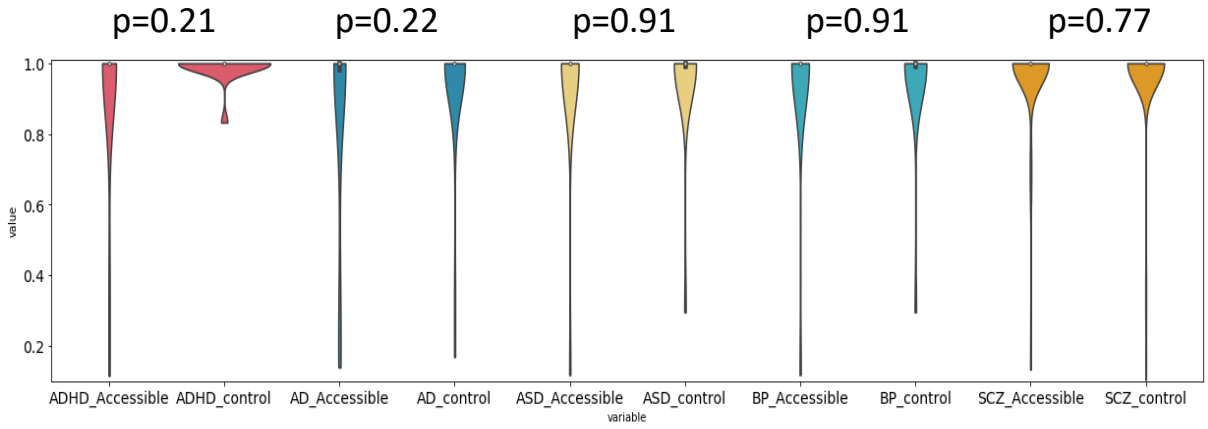
A copy of the custom code used for all the data analysis and figure generation in this study can be viewed and downloaded at the following BitBucket repository: https://bitbucket.org/BKLi/

Supplementary figures



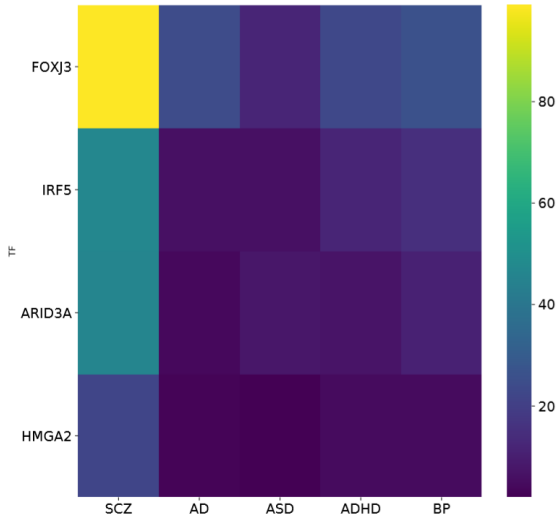
Supplementary Figure 1. Overview of fine-mapping

(a) Workflow of fine-mapping pipeline using FINEMAP
(b) Size distribution of constructed loci for performing fine-mapping and of previously estimated LD blocks



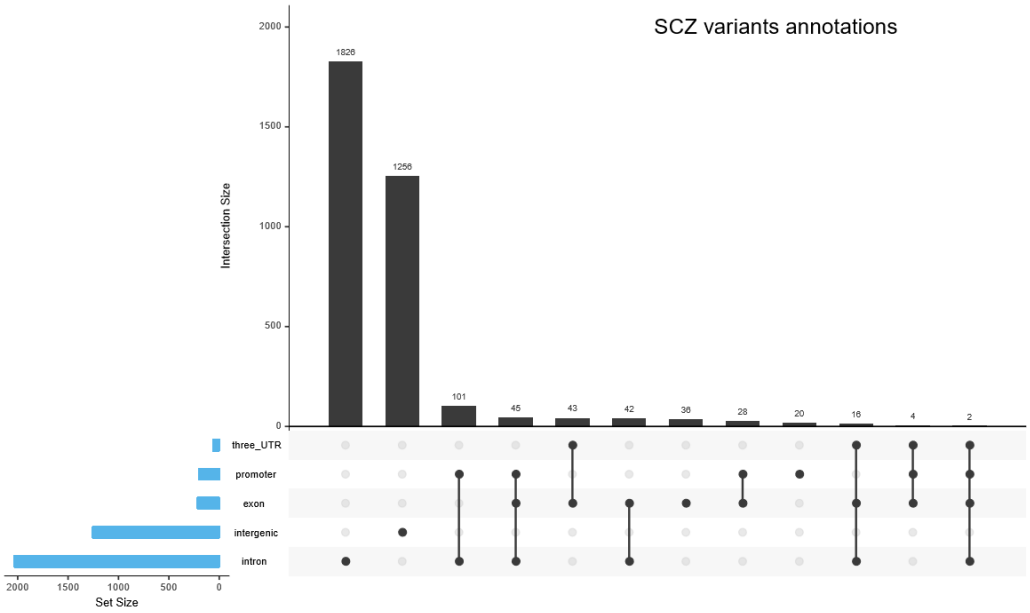
Supplementary Figure 2.

Distribution of posterior probability of variants overlapping open chromatin regions or not. No significant differences were observed (t-test)



Supplementary Figure 3.

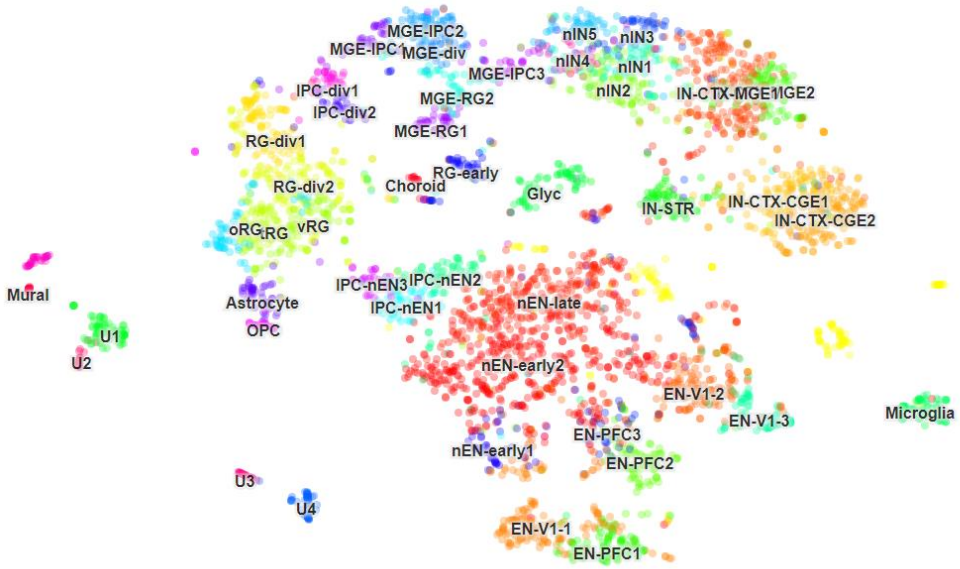
Heatmap showing enrichment of shared disturbed transcription factors. FOXJ3 is the most enriched one across all five diseases.



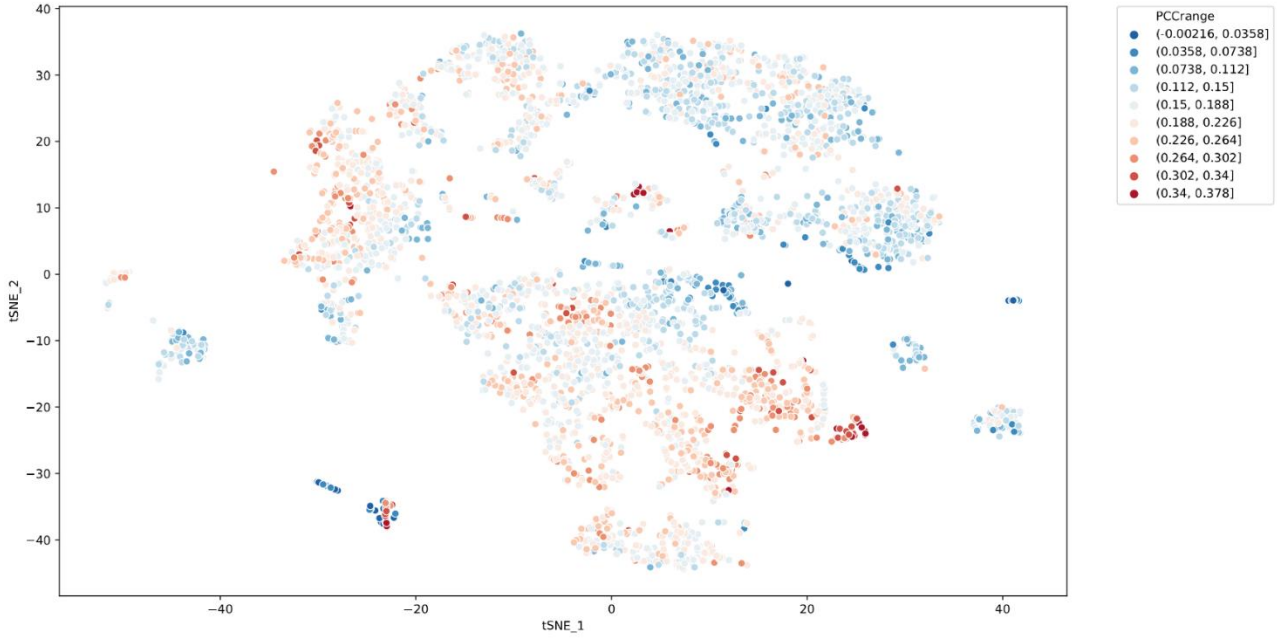
Supplementary Figure 4.

Example showing that variants may fit the criteria of two or more annotation categories.

a



b



Supplementary Figure 5.

(a) tSNE plot colored by WGCNAcluster

RG1-6 = oRG, tRG, vRG, Dividing RG Cluster 1, OPC, Astrocyte, Dividing RG Cluster 2
IPC1-5: IPC-nEN1, IPC-nEN2, Dividing IPC cluster 1, Dividing IPC cluster 2, IPC-nEN5
eN1-9: EN-PFC1, nEN-early2, nEN-late, EN-V1-1, EN-V1-2, EN-PFC2, nEN-early1, EN-PFC3, EN-V1-3
VP1-6: MGE-RG1, MGE-RG2, MGE-div, MGE-IPC1, MGE-IPC2, MGE-IPC3
NiN1-5: NiN1-5 (same)
iN1-5: IN-STR, IN-CTX-CGE1, IN-CTX-CGE2, IN-CTX-MGE1, IN-CTX-MGE2

STR = Striatum
CGE= Caudal Ganglionic Eminence (source of some Interneurons)
MGE = Medial Ganglionic Eminence (source of some interneurons)
CTX = Cortex (mature interneurons in the (dorsal) cortex vs. newborn interneurons that are still in the ventral forebrain).
nIN = newborn interneuron
nEN = newborn excitatory neuron
PFC = prefrontal cortex
V1 = visual cortex (occipital)

(b) tSNE plot colored by Pearson correlation between gene expression value of each single cell and bulk gene expression value of our excitatory neurons.