Journal of Visualized Experiments

Mapping Alzheimer's Disease Variants to Their Target Genes Using Computational Analysis of Chromatin Configuration

--Manuscript Draft--

Article Type:	Invited Methods Article - JoVE Produced Video
Manuscript Number:	JoVE60428R1
Full Title:	Mapping Alzheimer's Disease Variants to Their Target Genes Using Computational Analysis of Chromatin Configuration
Section/Category:	JoVE Genetics
Keywords:	Hi-C, GWAS, non-coding variants, gene mapping, functional genomics, Alzheimer's disease
Corresponding Author:	Hyejung Won University of North Carolina at Chapel Hill Chapel Hill , NC UNITED STATES
Corresponding Author's Institution:	University of North Carolina at Chapel Hill
Corresponding Author E-Mail:	hyejung_won@med.unc.edu
Order of Authors:	Hyejung Won
	Nana Matoba
	Doug H Phanstiel
	Yoseli Quiroga
Additional Information:	
Question	**Response**
Please indicate whether this article will be Standard Access or Open Access.	Standard Access (US$2,400)
Please indicate the **city, state/province, and country** where this article will be **filmed**. Please do not use abbreviations.	Chapel Hill, NC, USA

TITLE:

Mapping Alzheimer's Disease Variants to Their Target Genes Using Computational Analysis of Chromatin Configuration

AUTHORS AND AFFILIATIONS:

Nana Matoba[1,2], Ivana Y. Quiroga[3], Douglas H. Phanstiel[3,4,*], Hyejung Won[1,2,*]

[1]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA
[2]Neuroscience Center, University of North Carolina, Chapel Hill, NC, USA
[3]Thurston Arthritis Research Center, University of North Carolina, Chapel Hill, NC, USA
[4]Department of Cell Biology and Physiology, University of North Carolina, Chapel Hill, NC, USA

*These authors contributed equally.

Email addresses of co-authors:

Nana Matoba	(nana_matoba@med.unc.edu)
Ivana Y. Quiroga	(yoseli@live.unc.edu)

Corresponding author:

Hyejung Won	(hyejung_won@med.unc.edu)
Douglas H. Phanstiel	(douglas_phanstiel@med.unc.edu)

KEYWORDS:

Hi-C, GWAS, non-coding variants, gene mapping, functional genomics, Alzheimer's disease

SUMMARY:

We present a protocol to identify functional implications of non-coding variants identified by genome-wide association studies (GWAS) using three-dimensional chromatin interactions.

ABSTRACT:

Genome-wide association studies (GWAS) have successfully identified hundreds of genomic loci that are associated with human traits and disease. However, because the majority of the genome-wide significant (GWS) loci fall onto the non-coding genome, the functional impact of many remain unknown. Three-dimensional chromatin interactions identified by Hi-C or its derivatives can provide useful tools to annotate these loci by linking non-coding variants to their actionable genes. Here, we outline a protocol to map GWAS non-coding variants to their putative genes using Alzheimer's disease (AD) GWAS and Hi-C datasets from human adult brain tissue. Putative causal single-nucleotide polymorphisms (SNPs) are identified by application of fine-mapping algorithms. SNPs are then mapped to their putative target genes using enhancer-promoter interactions based on Hi-C. The resulting gene set represents AD risk genes, as they are potentially regulated by AD risk variants. To garner further biological insights into molecular mechanisms underlying AD, we characterize AD risk genes using developmental brain expression data and brain single-cell expression profiles. This protocol can be expanded to any GWAS and

Hi-C datasets to identify putative target genes and molecular mechanisms underlying various human traits and diseases.

INTRODUCTION:

Genome-wide association studies (GWAS) have played a pivotal role in revealing the genetic basis of a range of human traits and diseases. This large-scale genotyping has uncovered thousands of genomic variants associated with phenotypes ranging from height to schizophrenia risk. However, despite the enormous success of GWAS in identifying disease and trait associated loci, a mechanistic understanding of how these variants contribute to phenotype has been challenging because most phenotype associated variants reside in the non-coding fraction of the human genome. Since these variants often overlap with predicted regulatory elements, they are likely to alter transcriptional control of a nearby gene. However, non-coding loci can influence transcription of genes at linear distances exceeding one megabase, making the genes affected by each variant difficult to identify. Three-dimensional (3D) chromatin structure plays an important role in mediating connections between distant regulatory loci and gene promoters and can be used to identify genes affected by phenotype associated single-nucleotide polymorphisms (SNPs).

Gene regulation is mediated by a complex process, which involves enhancer activation and chromatin loop formation that physically connect enhancers to gene promoters to which the transcriptional machinery can be directed[1–3]. Because chromatin loops often span several hundred kilobases (kb), detailed maps of 3D chromatin architecture are required to decipher gene regulatory mechanisms. Multiple chromatin conformation capture technologies have been invented to identify the 3D chromatin architecture[4]. Among these technologies, Hi-C provides the most comprehensive architecture, as it captures genome-wide 3D chromatin interaction profiles. Hi-C datasets have been quickly adapted to interpret non-coding genome-wide significant (GWS) loci[5–13], as it can link non-coding variants to their putative target genes based on chromatin interaction profiles.

In this article, we outline a protocol to computationally predict putative target genes of GWAS risk variants using chromatin interaction profiles. We apply this protocol to map AD GWS loci[14] to their target genes using Hi-C datasets in the adult human brain[9]. The resulting AD risk genes are characterized by other functional genomic datasets that include single cell transcriptomic and developmental expression profiles.

PROTOCOL:

1. Workstation setup

1.1. Install R (version 3.5.0) and RStudio Desktop. Open RStudio.

1.2. Install the following libraries in R by typing the following code into the console window in RStudio.

```
if (!"BiocManager" %in% rownames(installed.packages()))
```

```
install.packages("BiocManager", repos="https://cran.r-project.org")
BiocManager::install("GenomicRanges")
BiocManager::install("biomaRt")
BiocManager::install("WGCNA")
install.packages("reshape")
install.packages("ggplot2")
install.packages("corrplot")
install.packages("gProfileR")
install.packages("tidyverse")
install.packages("ggpubr")
```

1.3. Download files.

NOTE: In this protocol, all files are required to be downloaded to ~/work directory.

1.3.1. Download the following files by clicking the links provided in **Table of Materials**.

1.3.1.1. Download fine-mapped credible SNPs for AD (Supplementary Table 8 from Jansen et al.[14]).

NOTE: Before analysis, open sheet eight in 41588_2018_311_MOESM3_ESM.xlsx, remove the first three rows and save the sheet as Supplementary_Table_8_Jansen.txt with tab separated format.

1.3.1.2. Download 10 kb resolution Hi-C interaction profiles in the adult brain from psychencode (described as *Promoter-anchored_chromatin_loops.bed* below).

NOTE: This file has the following format: chromosome, TSS_start, TSS_end, Enhancer_start, and Enhancer_end. In case other Hi-C datasets are used, this protocol requires Hi-C datasets processed at high resolution (5−20 kb).

1.3.1.3. Download single cell expression datasets from the PsychENCODE (described as *singlecell.rda* below).

NOTE: These are from neurotypical control samples.

1.3.1.4. Download developmental expression datasets from the BrainSpan (described as *devExpr.rda* below).

NOTE: 267666527 is a zipped file, so unzip the 267666527 to extract "columns_metadata.csv", "expression_matrix.csv", and "rows_metadata.csv" to generate devExpr.rda (see section 3).

1.3.2. Download exonic coordinates (see **Supplementary Files**, described as *Gencode19_exon.bed* and *Gencode19_promoter.bed* below) from Gencode version 19.

NOTE: Promoters are defined as 2 kb upstream of transcription start site (TSS). These files have the following format: chromosome, start, end, and gene.

1.3.3. Download gene annotation file (see **Supplementary Files**, described as *geneAnno.rda* below) from biomart.

NOTE: This file can be used to match genes based on Ensembl gene IDs and HUGO Gene Nomenclature Committee (HGNC) symbol.

2. Generation of a GRanges object for credible SNPs

2.1. Set up in R by typing the following code into the console window in RStudio.

```
library(GenomicRanges)
options(stringsAsFactors = F)
setwd("~/work") # This is the path to the working directory.
credSNP = read.delim("Supplementary_Table_8_Jansen.txt", header=T)
credSNP = credSNP[credSNP$Credible.Causal=="Yes", ]
```

2.2. Make a GRanges object by typing the following code into the console window in RStudio.

```
credranges = GRanges(credSNP$Chr, IRanges(credSNP$bp, credSNP$bp), rsid=credSNP$SNP,
P=credSNP$P)
save(credranges, file="AD_credibleSNP.rda")
```

3. Positional mapping

NOTE: For each step, type the corresponding code into the console window in RStudio.

3.1. Set up in R.

```
options(stringsAsFactors=F)
library(GenomicRanges)
load("AD_credibleSNP.rda") # (see 2)
```

3.2. Positional mapping of promoter/exonic SNPs to genes

3.2.1. Load promoter and exonic region and generate a GRange object.

```
exon = read.table("Gencode19_exon.bed")
exonranges = GRanges(exon[,1],IRanges(exon[,2],exon[,3]),gene=exon[,4])
promoter = read.table("Gencode19_promoter.bed")
promoterranges = GRanges(promoter[,1], IRanges(promoter[,2], promoter[,3]),
gene=promoter[,4])
```

3.2.2. Overlap credible SNPs with exonic regions.

```
olap = findOverlaps(credranges, exonranges)
credexon = credranges[queryHits(olap)]
mcols(credexon) = cbind(mcols(credexon), mcols(exonranges[subjectHits(olap)]))
```

3.2.3. Overlap credible SNPs with promoter regions.

```
olap = findOverlaps(credranges, promoterranges)
credpromoter = credranges[queryHits(olap)]
mcols(credpromoter) = cbind(mcols(credpromoter), mcols(promoterranges[subjectHits(olap)]))
```

3.3. Link SNPs to their putative target genes using chromatin interactions.

3.3.1. Load Hi-C dataset and generate a GRange object.

```
hic = read.table("Promoter-anchored_chromatin_loops.bed ", skip=1)
colnames(hic) = c("chr", "TSS_start", "TSS_end", "Enhancer_start", "Enhancer_end")
hicranges = GRanges(hic$chr, IRanges(hic$TSS_start, hic$TSS_end),
enhancer=hic$Enhancer_start)
olap = findOverlaps(hicranges, promoterranges)
hicpromoter = hicranges[queryHits(olap)]
mcols(hicpromoter) = cbind(mcols(hicpromoter), mcols(promoterranges[subjectHits(olap)]))
hicenhancer = GRanges(seqnames(hicpromoter), IRanges(hicpromoter$enhancer,
hicpromoter$enhancer+10000), gene=hicpromoter$gene)
```

3.3.2. Overlap credible SNPs with Hi-C GRange object.

```
olap = findOverlaps(credranges, hicenhancer)
credhic = credranges[queryHits(olap)]
mcols(credhic) = cbind(mcols(credhic), mcols(hicenhancer[subjectHits(olap)]))
```

3.4. Compile AD candidate genes defined by positional mapping and chromatin interaction profiles.

```
### The resulting candidate genes for AD:
ADgenes = Reduce(union, list(credhic$gene, credexon$gene, credpromoter$gene))
### to convert Ensembl Gene ID to HGNC symbol
load("geneAnno.rda")
ADhgnc = geneAnno1[match(ADgenes, geneAnno1$ensembl_gene_id), "hgnc_symbol"]
ADhgnc = ADhgnc[ADhgnc!=""]
save(ADgenes, ADhgnc, file="ADgenes.rda")
write.table(ADhgnc, file="ADgenes.txt", row.names=F, col.names=F, quote=F, sep="\t")
```

4. Developmental expression trajectories

NOTE: For each step, type the corresponding code into the console window in RStudio.

4.1. Set up in R.

```
library(reshape); library(ggplot2); library(GenomicRanges); library(biomaRt)
library("WGCNA")
options(stringsAsFactors=F)
```

4.2. Process expression and meta data.

```
datExpr = read.csv("expression_matrix.csv", head = FALSE)
datExpr = datExpr[,-1]
datMeta = read.csv("columns_metadata.csv")
datProbes = read.csv("rows_metadata.csv")
datExpr = datExpr[datProbes$ensembl_gene_id!="",]
datProbes = datProbes[datProbes$ensembl_gene_id!="",]
datExpr.cr= collapseRows(datExpr, rowGroup = datProbes$ensembl_gene_id, rowID=
rownames(datExpr))
datExpr = datExpr.cr$datETcollapsed
gename = data.frame(datExpr.cr$group2row)
rownames(datExpr) = gename$group
```

4.2.1. Specify developmental stages.

```
datMeta$Unit = "Postnatal"
idx = grep("pcw", datMeta$age)
datMeta$Unit[idx] = "Prenatal"
idx = grep("yrs", datMeta$age)
datMeta$Unit[idx] = "Postnatal"
datMeta$Unit = factor(datMeta$Unit, levels=c("Prenatal", "Postnatal"))
```

4.2.2. Select cortical regions.

```
datMeta$Unit = "Postnatal"
datMeta$Region = "SubCTX"
r = c("A1C", "STC", "ITC", "TCx", "OFC", "DFC", "VFC", "MFC", "M1C", "S1C", "IPC", "M1C-S1C",
"PCx", "V1C", "Ocx")
datMeta$Region[datMeta$structure_acronym %in% r] = "CTX"
datExpr = datExpr[,which(datMeta$Region=="CTX")]
datMeta = datMeta[which(datMeta$Region=="CTX"),]
save(datExpr, datMeta, file="devExpr.rda")
```

4.3. Extract developmental expression profiles of AD risk genes.

```
load("ADgenes.rda")
exprdat = apply(datExpr[match(ADgenes, rownames(datExpr)),],2,mean,na.rm=T)
dat = data.frame( Region=datMeta$Region, Unit=datMeta$Unit, Expr=exprdat)
```

4.4. Compare prenatal versus postnatal expression levels of AD risk genes.

```
pdf(file="developmental_expression.pdf")
ggplot(dat,aes(x=Unit, y=Expr, fill=Unit, alpha=Unit)) + ylab("Normalized expression") +
geom_boxplot(outlier.size = NA) + ggtitle("Brain Expression") + xlab("") +
scale_alpha_manual(values=c(0.2, 1)) + theme_classic() + theme(legend.position="na")
dev.off()
```

5. Cell-type expression profiles

NOTE: For each step, type the corresponding code into the console window in RStudio.

5.1. Set up in R.

```
load("ADgenes.rda")
load("singlecell.rda")
load("geneAnno.rda")
targetname = "AD"
targetgene = ADhgnc
datExpr = scale(cellexp,center=T, scale=F)
```

5.2. Extract cellular expression profiles of AD risk genes.

```
exprdat = apply(datExpr[match(targetgene, rownames(datExpr)),],2,mean,na.rm=T)
dat = data.frame(Group=targetname, cell=names(exprdat), Expr=exprdat)

dat$celltype = unlist(lapply(strsplit(dat$cell, split="[.]"),'[[',1))
dat = dat[-grep("Ex|In",dat$celltype),]
dat$celltype = factor(dat$celltype, levels=c("Neurons","Astrocytes","Microglia","Endothelial",
"Oligodendrocytes","OPC","Fetal"))

pdf(file="singlecell_expression_ADgenes.pdf")
ggplot(dat,aes(x=celltype, y=Expr, fill=celltype)) +
ylab("Normalized expression") + xlab("") + geom_violin() +
theme(axis.text.x=element_text(angle = 90, hjust=1)) + theme(legend.position="none") +
ggtitle(paste0("Cellular expression profiles of AD risk genes"))
dev.off()
```

6. Gene annotation enrichment analysis of AD risk genes

6.1. Download and configure HOMER by typing the commands below in terminal.

```
mkdir homer
cd homer
wget http://homer.ucsd.edu/homer/configureHomer.pl
perl ./configureHomer.pl -install
perl ./configureHomer.pl -install human-p
perl ./configureHomer.pl -install human-o
```

6.2. Run HOMER by typing the commands below in terminal.

```
export PATH=$PATH:~/work/homer/bin
findMotifs.pl ~/work/ADgenes.txt human ~/work/
```

6.3. Plot the enriched terms by typing the following code into the console window in RStudio.

```
library(ggpubr)
options(stringsAsFactors=F)
pdf("GO_enrichment.pdf",width=15,height=8)
plot_barplot = function(dbname,name,color){
input = read.delim(paste0(dbname,".txt"),header=T)
input = input[,c(-1,-10,-11)]
input = unique(input)
input$FDR = p.adjust(exp(input$logP))
input_sig = input[input$FDR < 0.1,]
input_sig$FDR = -log10(input_sig$FDR)
input_sig = input_sig[order(input_sig$FDR),]

p = ggbarplot(input_sig, x = "Term", y = "FDR", fill = color, color = "white", sort.val = "asc", ylab =
expression(-log[10](italic(FDR))), xlab = paste0(name," Terms"), rotate = TRUE, label =
paste0(input_sig$Target.Genes.in.Term,"/",input_sig$Genes.in.Term), font.label = list(color =
"white", size = 9), lab.vjust = 0.5, lab.hjust = 1)
p = p+geom_hline(yintercept = -log10(0.05), linetype = 2, color = "lightgray")
return(p)
}

p1 = plot_barplot("biological_process","GO Biological Process","#00AFBB")
p2 = plot_barplot("kegg","KEGG","#E7B800")
p3 = plot_barplot("reactome","Reactome","#FC4E07")

ggarrange(p1, p2, p3, labels = c("A", "B", "C"), ncol = 2, nrow = 2)
```

352 dev.off()
353
354 **REPRESENTATIVE RESULTS:**
355 The process described here was applied to a set of 800 credible SNPs that were defined by the
356 original study[14]. Positional mapping revealed that 103 SNPs overlapped with promoters (43
357 unique genes) and 42 SNPs overlapped with exons (27 unique genes). After positional mapping,
358 84% (669) SNPs remained unannotated. Using Hi-C datasets in the adult brain, we were able to
359 link an additional 208 SNPs to 64 genes based on physical proximity. In total, we mapped 284 AD
360 credible SNPs to 112 AD risk genes (**Figure 1A**). AD risk genes were associated with amyloid
361 precursor proteins, amyloid-beta formation, and immune response, reflecting the known biology
362 of AD[15–18] (**Figure 1B-D**). Developmental expression profiles of AD risk genes showed marked
363 postnatal enrichment, indicative of the age-associated elevated risk of AD (**Figure 2A**). Finally, AD
364 risk genes were highly expressed in microglia, primary immune cells in the brain (**Figure 2B**). This
365 is in agreement with the recurrent findings that AD has a strong immune basis and microglia are
366 the central player in AD pathogenesis[14,19,20].
367
368 **FIGURES LEGENDS:**
369 **Figure 1: Defining putative target genes of AD GWS loci.** (**A**) Credible SNPs derived from the top
370 29 AD loci were categorized into promoter SNPs, exonic SNPs, and unannotated non-coding SNPs.
371 Promoter and exonic SNPs were directly assigned to their target genes by positional mapping,
372 while chromatin interaction profiles in the adult brain were additionally used to map SNPs based
373 on physical interactions. (**B-D**) Enrichment of GO (**B**), KEGG (**C**), and Reactome (**D**) terms in AD
374 risk genes was performed using HOMER as described in protocol section 6. The x axis represents
375 the false discovery rate (FDR) corrected -log10 (P-value). Enriched terms with FDR < 0.1 were
376 plotted. Grey vertical lines represent FDR = 0.05. APP amyloid precursor protein. Numerator, the
377 number of AD risk genes represented in each term; denominator, the number of genes in each
378 term.
379
380 **Figure 2: Characterization of AD risk genes.** (**A**) AD risk genes are highly expressed in the
381 postnatal cortex compared to the prenatal cortex. (**B**) Violin plots depicting distributions of gene
382 expression values (normalized expression) in different cell types from the cortex. These results
383 show that AD risk genes are highly expressed in microglia, consistent with previous studies[14].
384
385 **DISCUSSION**:
386 Here we describe an analytic framework that can be used to functionally annotate GWS loci
387 based on positional mapping and chromatin interactions. This process involves multiple steps (for
388 more details see this review[13]). First, given that chromatin interaction profiles are highly cell-type
389 specific, Hi-C data obtained from the appropriate cell/tissue types that best capture underlying
390 biology of the disorder needs to be used. Given that AD is a neurodegenerative disorder, we used
391 adult brain Hi-C data[9] to annotate GWS loci. Second, each GWS locus often has up to hundreds
392 of SNPs that are associated with the trait because of linkage disequilibrium (LD), so it is important
393 to obtain putative causal ('credible') SNPs by computationally predicting the causality through
394 the use of fine-mapping algorithms[21,22] or experimentally testing regulatory activities using high-
395 throughput approaches such as massively parallel reporter assays (MPRA)[23] or self-transcribing

active regulatory region sequencing (STARR-seq)[24]. For the work described here, we used credible SNPs reported in Jansen et al.[14]. Third, promoter and exonic SNPs are annotated based on positional mapping. We used a simple positional mapping strategy in which SNPs were mapped to the genes when they overlapped with promoters (defined as 2 kb upstream of transcription start site) or exons. However, this approach can be further elaborated by assessing the functional consequences of exonic SNPs, such as whether the SNP induces nonsense mediated decay, missense variation, or nonsense variation. Fourth, chromatin interaction profiles from the appropriate tissue/cell type can be used to assign SNPs to their putative target genes based on physical proximity. We used interaction profiles anchored to promoters, but we can further refine or expand the interaction profiles by taking enhancer activities (guided by histone H3 K27 acetylation or chromatin accessibility) or exonic interactions into account. One important consideration in this process is to use consistent human genome build. For example, if genomic positions of summary statistics are not based on hg19 (i.e., hg18 or hg38), an appropriate version of the reference genome should be obtained or the summary statistics need to be converted to hg19 using liftover[25].

We applied this framework to identify putative target genes for AD GWAS, assigning 284 SNPs to 112 AD risk genes. Using developmental expression profiles[26] and cell-type specific expression profiles[9], we then demonstrated that this gene set was consistent with what is known about AD pathology, revealing the cell types (microglia), biological functions (immune response and amyloid beta), and elevated risk upon age.

While we presented a framework that delineates potential target genes of AD and its underlying biology, it is of note that Hi-C based annotation can be expanded to annotate any non-coding variation. As more whole-genome sequencing data becomes available and our understanding about the non-coding rare variation grows, Hi-C will provide a key resource for interpretation of disease-associated genetic variants. A compendium of Hi-C resources obtained from multiple tissue and cell types will be therefore critical to facilitating a wide application of this framework to garner biological insights into various human traits and disease.

ACKNOWLEDGMENTS:
This work was supported by the NIH grant R00MH113823 (to H.W.) and R35GM128645 (to D.H.P.), NARSAD Young Investigator Award (to H.W.), and SPARK grant from the Simons Foundation Autism Research Initiative (SFARI, to N.M. and H.W.).

DISCLOSURES:
The authors have nothing to disclose.

REFERENCES:
1. Dekker, J., Misteli, T. Long-Range Chromatin Interactions. *Cold Spring Harbor Perspectives in Biology*. **7** (10), a019356 (2015).
2. Sanyal, A., Lajoie, B.R., Jain, G., Dekker, J. The long-range interaction landscape of gene promoters. *Nature*. **489** (7414), 109–113 (2012).
3. Plank, J.L., Dean, A. Enhancer function: mechanistic and genome-wide insights come together.

Molecular Cell. **55** (1), 5–14 (2014).
4. Dekker, J., Marti-Renom, M.A., Mirny, L.A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*. **14** (6), 390–403 (2013).
5. Martin, P. et al. Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nature Communications*. **6**, 10069 (2015).
6. Won, H. et al. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature*. **538** (7626), 523–527 (2016).
7. Jäger, R. et al. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nature Communications*. **6**, 6178 (2015).
8. Chen, J.A.A. et al. Joint genome-wide association study of progressive supranuclear palsy identifies novel susceptibility loci and genetic correlation to neurodegenerative diseases. *Molecular Neurodegeneration*. **13** (1), 41–41 (2018).
9. Wang, D. et al. Comprehensive functional genomic resource and integrative model for the adult brain. *Science*. **362** (6420) eaat8464 (2018).
10. Demontis, D. et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature Genetics*. **51** (1), 63–75 (2019).
11. Grove, J. et al. Identification of common genetic risk variants for autism spectrum disorder. *Nature Genetics*. **51** (3), 431–444 (2019).
12. Lee, P.H. et al. Genome wide meta-analysis identifies genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *bioRxiv*. 528117–528117 (2019).
13. Mah, W., Won, H. The three-dimensional landscape of the genome in human brain tissue unveils regulatory mechanisms leading to schizophrenia risk. *Schizophrenia Research*. In press (2019).
14. Jansen, I.E. et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nature Genetics*. **51** (3), 404–413 (2019).
15. Viola, K.L., Klein, W.L. Amyloid β oligomers in Alzheimer's disease pathogenesis, treatment, and diagnosis. *Acta Neuropathologica*. **129** (2), 183–206 (2015).
16. Mroczko, B., Groblewska, M., Litman-Zawadzka, A., Kornhuber, J., Lewczuk, P. Amyloid β oligomers (AβOs) in Alzheimer's disease. *Journal of Neural Transmission*. **125** (2), 177–191 (2018).
17. Heneka, M.T. et al. Neuroinflammation in Alzheimer's disease. *Lancet Neurology*. **14** (4), 388–405 (2015).
18. Minter, M.R., Taylor, J.M., Crack, P.J. The contribution of neuroinflammation to amyloid toxicity in Alzheimer's disease. *Journal of Neurochemistry*. **136** (3), 457–474 (2016).
19. Hansen, D.V., Hanson, J.E., Sheng, M. Microglia in Alzheimer's disease. *The Journal of Cell Biology*. **217** (2), 459–472 (2018).
20. Gjoneska, E. et al. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature*. **518** (7539), 365–369 (2015).
21. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*. **32** (10), 1493–1501 (2016).
22. Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics*. **198** (2), 497–508 (2014).
23. Tewhey, R. et al. Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell*. **165** (6), 1519–1529 (2016).

24. Arnold, C.D. et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*. **339** (6123), 1074–1077 (2013).

25. Kent, W.J. et al. The human genome browser at UCSC. *Genome Research*. **12** (6), 996–1006 (2002).

26. Kang, H.J. et al. Spatio-temporal transcriptome of the human brain. *Nature*. **478** (7370), 483–489 (2011).

Figure1

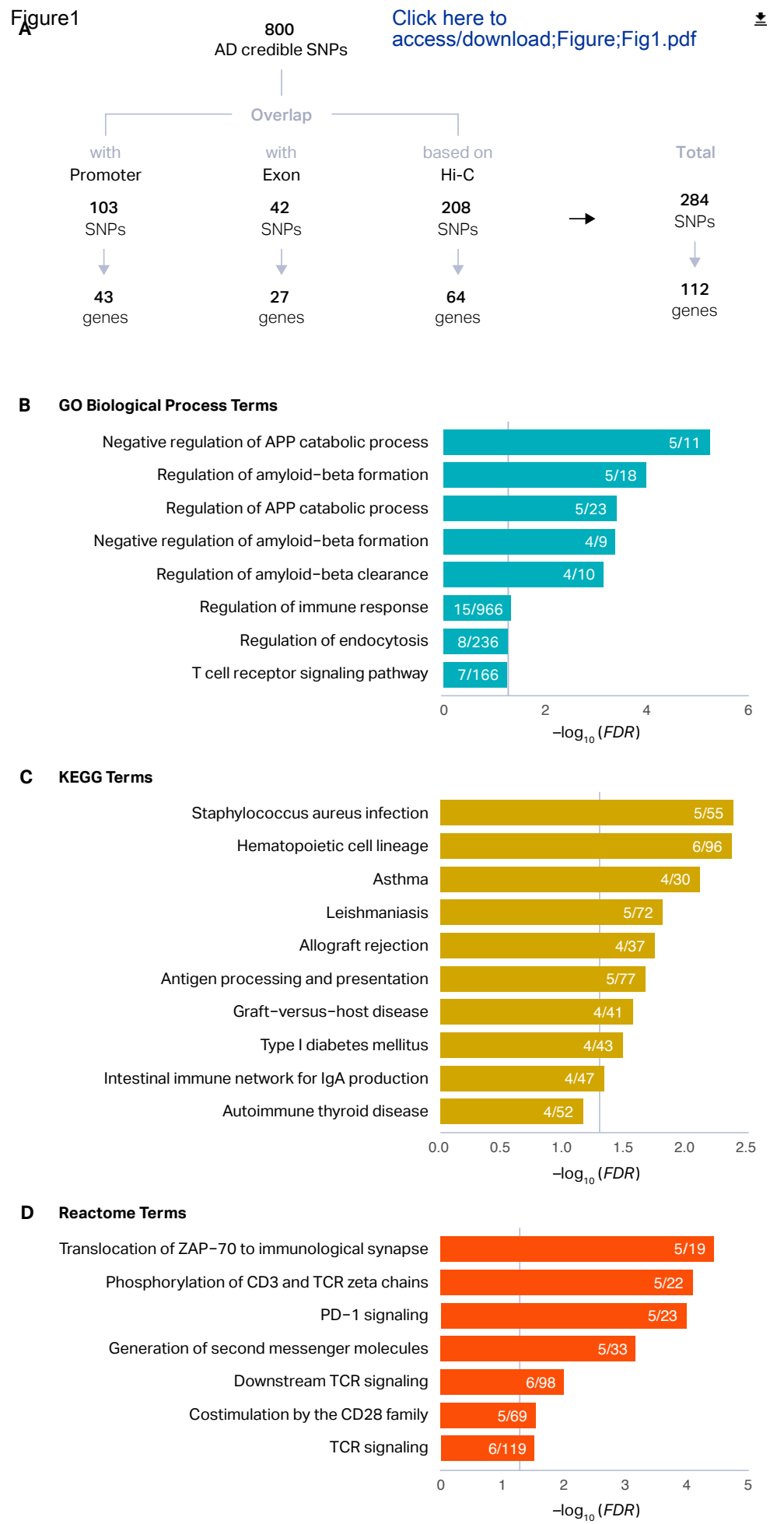
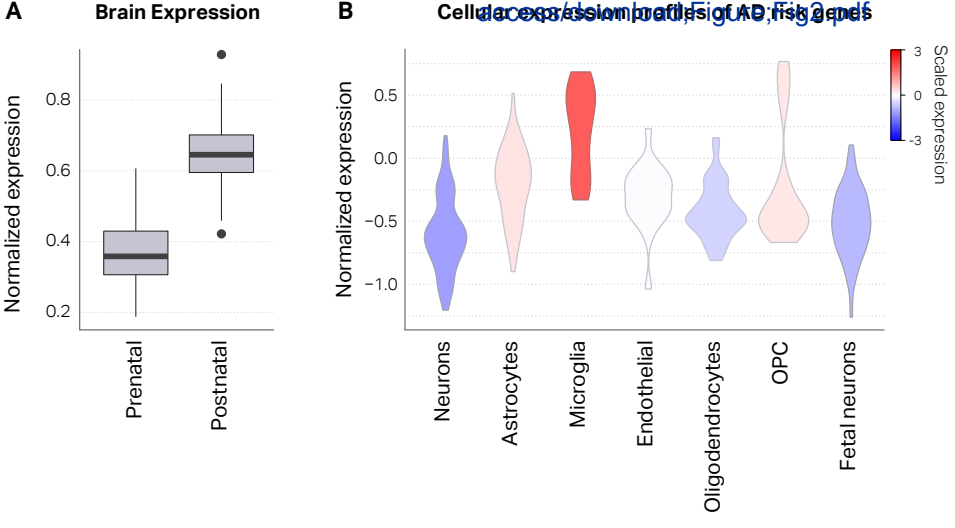


Figure2



Name of Material/Equipment/Files	Company	Catalog Number
10 kb resolution Hi-C interaction profiles in the adult brain from psychencode		
Developmental expression datasets		
Fine-mapped credible SNPs for AD (Supplementary Table 8 from Jansen et al.[14])		
Single cell expression datasets		
R (version 3.5.0)		
RStudio Desktop		
HOMER		

Comments/Description

http://adult.psychencode.org/Datasets/Integrative/Promoter-anchored_chromatin_loops.bed
http://www.brainspan.org/api/v2/well_known_file_download/267666527

https://static-content.springer.com/esm/art%3A10.1038%2Fs41588-018-0311-9/MediaObjects/41588_2018_311_MOESM3_ESM.xlsx
http://adult.psychencode.org/Datasets/Derived/SC_Decomp/DER-19_Single_cell_markergenes_TPM.xlsx
https://www.r-project.org/
https://www.rstudio.com/products/rstudio/download/
http://homer.ucsd.edu/homer/configureHomer.pl

Editorial comments:

1. Please take this opportunity to thoroughly proofread the manuscript to ensure that there are no spelling or grammar issues. The JoVE editor will not copy-edit your manuscript and any errors in the submitted revision may be present in the published version.

We proofread the manuscript as suggested.

2. Title: Please revise to avoid the use of punctuation (colon, dash, etc.).

We now changed the title into “Mapping Alzheimer’s disease variants to their target genes using computational analysis of chromatin configuration”

3. Please revise the Protocol to contain only action items that direct the reader to do something (e.g., “Do this,” “Ensure that,” etc.). The actions should be described in the imperative tense in complete sentences wherever possible. Avoid usage of phrases such as “could be,” “should be,” and “would be” throughout the Protocol. Any text that cannot be written in the imperative tense may be added as a “NOTE.” Please include all safety procedures and use of hoods, etc. However, notes should be used sparingly and actions should be described in the imperative tense wherever possible. Please move the discussion about the protocol to the Discussion.

We ensured that the protocol does not have any phrases such as “could be”, “should be”, and “would be.”

4. Please add more details to your protocol steps. There should be enough detail in each step to supplement the actions seen in the video so that viewers can easily replicate the protocol. Please ensure you answer the “how” question, i.e., how is the step performed?

This request is similar to request #5. Please see our response to request #5.

5. For actions involving software usage, please provide all specific details (e.g., button clicks, software commands, any user inputs, etc.) needed to execute the actions. Please include a step-wise description of software usage; mention what button is clicked on in the software, or which menu items need to be selected, and provide user input commands, etc.

We made the following changes in response to the request #4 and #5.

1. We directed the readers to install RStudio (Line 93: Install RStudio Desktop: https://www.rstudio.com/products/rstudio/download/), which will help them run the code provided.

2. We added a sentence “Type the following code into the console window in RStudio (e.g. Line 96)” in front of each section of code for added clarity.
3. We have revised the Download files section (section **1.3**) to include links to all files and explicit directions for downloading data.

6. Line 283: Figure 2C does not exist. Please revise.

Thanks for pointing this out. We changed it to **Figure 2B**. We also found out that Figures were not labeled properly within the manuscript, which is now corrected.

7. Please remove the embedded figure(s) from the manuscript.

We removed the figures from the manuscript, and submitted separately.

Reviewers' comments:

Reviewer #1:

The authors in this paper developed a computational pipeline for linking GWAS loci to genes using Hi-C data, and applied to Alzheimer's disease for discovering AD risk genes. The pipeline mapped credible SNPs from AD GWAS to various regions including enhancers, gene bodies and promoters, and then linked SNPs to genes if the mapped regions have potential interactions from Hi-C data. Overall, the paper was well organized and provided a complete set of R codes for pipeline implementation. Before recommending for publication, I have the following minor concerns that authors need to address:

We thank the reviewer for his/her positive and constructive comments.

1. Hi-C data description such as protocol, resolution & tissue/cell type is missing. Can the pipeline be scalable to the Hi-C datasets with different resolutions, which authors discussed at the end?

We thank the reviewer for this comment. We already described that Hi-C interaction profiles were generated from the adult brain. Thanks to the comment, we also added Line 125: “NOTE: In case other Hi-C datasets are used, this protocol requires Hi-C datasets processed at high resolution (5-20kb).”

2. Fig. 2A seems mix multiple regions together. Is there any particular developmental expression pattern for particular regions? Authors didn't introduce how to normalize gene expression either.

We used cortical expression data from brainspan. We now updated **Figure 2A legend** to describe the brain region: “AD risk genes are highly expressed in the postnatal cortex compared to the prenatal cortex”. We used the expression data provided by BrainSpan (https://www.brainspan.org/) and did not perform any additional normalization or processing. We now updated section **1.3.1** with detailed instructions for how to download the expression data file and process it.

3. Was the single cell data from healthy or AD brain? If healthy, the cell type specific expression might not represent AD cells. Will cell type Hi-C improve gene linking over tissue Hi-C? Also, details on gene expression normalization is missing.

Single cell data was from healthy (neurotypical) brain. We used the expression data provided by the original paper (Wang et al., Science *2018*) which has been already normalized. We also updated section **1.3.1** to describe that the single cell data comes from healthy brains: Line 127: “Single cell expression datasets from the PsychENCODE (Described as *singlecell.rda* below). This is from neurotypical control samples.”

Reviewer #2:

Manuscript Summary:

The manuscript describes a method to annotate non-coding variations to the candidate genes for GWAS SNPs. Furthermore, the candidate genes have been subjected to enrichment analysis and cross-cell expression comparisons. Addressing the following comments will improve the quality of the manuscript.

We thank the reviewer for his/her critical and insightful comments.

Major Concerns:

1. The explanation about figure 1B is completely missing in the manuscript. It's important to mention and explain what resource was being used to get the figure and what purpose does it serve as the results of the work done. Additionally, the information about the total number of genes involved in each of the GO terms should also be written. Do all the genes take part in each of these GO terms? I don't think so. Also, what does the -log10(FDR) mean? This needs explanation as well. Overall, the results have been represented in a very abstract way. They need to be explained in detail.

We are sorry that **Figure 1B** was labelled incorrectly. We corrected the figure captions for Figures 1 and 2: “AD risk genes were associated with amyloid precursor proteins, amyloid-beta formation, and immune response, reflecting the known biology of AD[15–18] (**Figure 1B**).”

Based on the reviewer’s suggestion, we

(1) Updated **Figure 1B-D** with the # of genes in each term and # of genes represented in our list.
(2) Updated **Figure 1B-D legend** as below:

(B-D) Enrichment of GO (**B**), KEGG (**C**), and Reactome (**D**) terms in AD risk genes was performed using HOMER as described in step **5**. The x-axis represents the FDR corrected -log10 (P-value). Enriched terms with FDR<0.1 were plotted. Grey vertical lines represent FDR=0.05. APP amyloid precursor protein. Numerator, the number of AD risk genes represented in each term; denominator, the number of genes in each term.

2. This is a follow up to the 1st comment: The authors have done a sort of enrichment analysis depicting top 10 GO terms. It's highly recommended to perform and include results from Gene Set Enrichment Analysis (GSEA) using Broad Institute's Molecular signature database. This will give a better overview of associated pathways from KEGG and Reactome for the genes the authors have shortlisted in their work. In addition to this, it is also recommended to use NeuroMMsig (https://neurommsig.scai.fraunhofer.de/) to find out what Alzheimer Disease (AD) mechanisms are represented by the list of genes. NeuroMMsig is a mechanism enrichment analysis platform developed for neurodegenerative diseases, especially for AD and Parkinson disease (PD).

Thanks to this insightful suggestion, we used HOMER to analyze KEGG and Reactome for AD risk genes and reported the result in Figure 1. We think this greatly improved the overall findings of the manuscript, and gave more detailed descriptions about the potential function of AD risk genes.

3. It's difficult to comprehend what do the different shapes of different cellular expression profiles refer to.

Sorry for the confusion. We now updated **Figure 2B legend** as below:

“(B) Violin plots depicting distributions of gene expression values (normalized expression) in different cell types from the cortex. These results show that AD risk genes are highly expressed in microglia, consistent with previous studies[14].”

4. The authors have reported about highly expressed genes in microglia (Figure 2B) as a whole. It is good to make overall comparison of gene expressions across different cell types but when it comes to gene expression analysis, the down-regulated genes are also of interests. Therefore, the author might have to report about individual genes that are either over-expressed or under-expressed to improve the quality of the manuscript.

While it is often of interest to investigate genes with different patterns of change, the intent of this plot is to show that the genes implicated by our analysis are expressed in cell types with a known role in AD. This offers support to the validity of our analyses.

While the reviewer raised an important point, we believe that the analysis of other genes that are differentially expressed across these cell types is outside the scope of this manuscript.

5. The R data files for devExpr.rda, ADgenes.rda and singlecell.rda should be made available.

Because *devExpr.rda* and *singlecell.rda* use datasets that have been made publicly available by the original paper, we do not feel comfortable providing these files. Instead, we updated section **1.3.1** in which we described the download procedure in detail so that readers can easily follow and retrieve the same data. We are providing ADgenes.rda with this paper.

Minor Concerns:

Line 60: There are double full stops.

Line 70: First mention of GWS, therefore needs to be written with the full name and the abbreviation.

Line 75: Double "the".

Line 326-328: Needs to be checked again.

We thank the reviewer for catching these minor errors. We now corrected all the points brought up by the reviewer.

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins Project: (None)

You can open a code editor by clicking here

Console Jobs

R Console

```
R version 3.5.0 (2018-04-23) -- "Joy in Playing"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

>
```

Type codes here

Environment History Connections

Import Dataset List

Global Environment

Environment is empty

Data and Command History

Files Plots Packages Help Viewer

Zoom Export

Once you generate plot, the plot will be appeared here

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Untitled1* | Source on Save | Run | Source

Code Editor

```
install.packages("gProfileR")
install.packages("tidyverse")
install.packages("ggpubr")
library(GenomicRanges)
options(stringsAsFactors = F)
setwd("~/work") # This is the path to the working directory.
credSNP = read.delim("Supplementary_Table_8_Jansen.txt", header=T)
credSNP = credSNP[credSNP$Credible.Causal=="Yes", ]

credranges = GRanges(credSNP$Chr, IRanges(credSNP$bp, credSNP$bp), rsid=cr
save(credranges, file="AD_credibleSNP.rda")
```

20:1 (Top Level) R Script

Environment | History | Connections

Import Dataset | List

Global Environment

Data

datMeta	345 obs. of 10 variables
datProbes	17085 obs. of 5
exon	324422 obs. of 4 variables
exonranges	Large GRanges (324422 ele...
gename	16767 obs. of 2 variables
geneAnno1	57317 obs. of 8 variables
hic	149097 obs. of 5 variables

Console | Jobs

~/work/

R Console

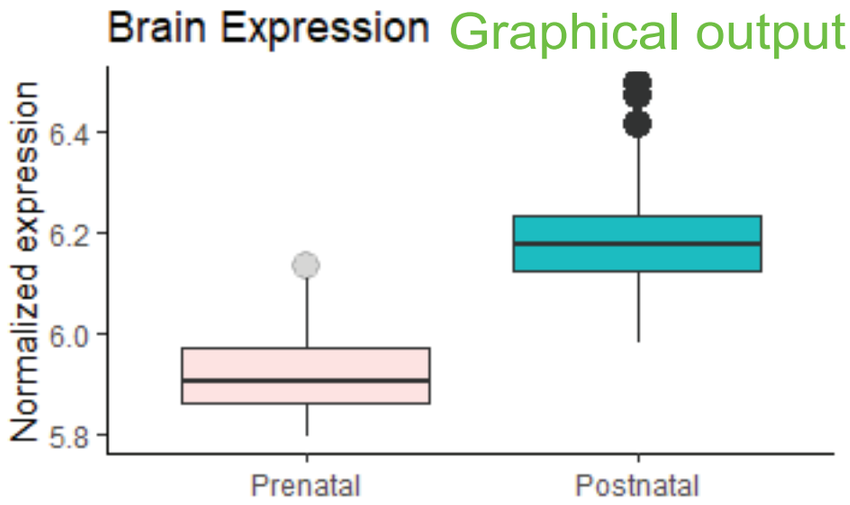
```
V7      CTX Prenatal 5.810975
V9      CTX Prenatal 5.827192
V11     CTX Prenatal 5.883260
V12     CTX Prenatal 5.980518
V13     CTX Prenatal 5.827781
> ggplot(dat,aes(x=Unit, y=Expr, fill=Unit, alpha=Unit)) + ylab("Normalized expr
ession") + geom_boxplot(outlier.size = NA) + ggtitle("Brain Expression") + xlab
("") + scale_alpha_manual(values=c(0.2, 1)) + theme_classic() + theme(legend.pos
ition="na")
>
>
```

Type codes here

Files | Plots | Packages | Help | Viewer

Zoom | Export

Graphical output



Place the code here

homer — -bash — 122×24

```
phanstiel3@                :~/work/homer$ export PATH=$PATH:~/work/homer/bin
phanstiel3@                :~/work/homer$ findMotifs.pl ~/work/ADgenes.txt human ~/work/
```

geneAnno.rda