RESEARCH Open Access

Deciphering the tissue-specific functional effect of Alzheimer risk SNPs with deep genome annotation

Pradeep Varathan Pugalenthi[1], Bing He[1], Linhui Xie[2], Kwangsik Nho[3], Andrew J. Saykin[3] and Jingwen Yan[1,3*]

*Correspondence:
Jingwen Yan
jingyan@iu.edu
[1] Department of Biomedical Engineering and Informatics, Indiana University Indianapolis, 420 University Blvd, Indianapolis, IN 46202, USA
[2] Department of Electrical and Computer Engineering, Purdue University Indianapolis, 420 University Blvd, Indianapolis, IN 46202, USA
[3] Department of Radiology and Imaging Sciences, Indiana University School of Medicine, 550 University Blvd, Indianapolis, IN 46202, USA

Abstract

Alzheimer's disease (AD) is a highly heritable brain dementia, along with substantial failure of cognitive function. Large-scale genome-wide association studies (GWASs) have led to a set of SNPs significantly associated with AD and related traits. GWAS hits usually emerge as clusters where a lead SNP with the highest significance is surrounded by other less significant neighboring SNPs. Although functionality is not guaranteed even with the strongest associations in GWASs, lead SNPs have historically been the focus of the field, with the remaining associations inferred to be redundant. Recent deep genome annotation tools enable the prediction of function from a segment of a DNA sequence with significantly improved precision, which allows in-silico mutagenesis to interrogate the functional effect of SNP alleles. In this project, we explored the impact of top AD GWAS hits around *APOE* region on chromatin functions and whether it will be altered by the genetic context (i.e., alleles of neighboring SNPs). Our results showed that highly correlated SNPs in the same LD block could have distinct impacts on downstream functions. Although some GWAS lead SNPs showed dominant functional effects regardless of the neighborhood SNP alleles, several other SNPs did exhibit enhanced loss or gain of function under certain genetic contexts, suggesting potential additional information hidden in the LD blocks.

Keywords Alzheimer's disease, GWAS annotation, Chromatin feature

Introduction

Alzheimer's disease (AD) is one of the most common forms of dementia, and it is associated with substantial failure of organs and mental health issues. In the United States, nearly 10% of the population aged 65 and older has been diagnosed as AD and projections the total number of cases is projected to indicate that there will reach 13.8 million cases by 2060 [1]. The heritability of AD is estimated to be between 60% and 80% [2]. Therefore, much work has been done in genetic association studies seeking to determine the genetic architecture of AD since the early 1990s, followed by several large-scale genome-wide association studies (GWASs) and meta-analyses [3–5]. It is expected that these increasing findings will better delineate the pathways underlying disease. Yet, there remains a substantial gap in estimated heritability. Only 3–17% of heritability can be explained by current large-scale GWAS findings [6, 7].

GWAS hits usually emerge as clusters where a lead SNP with the highest significance is surrounded by other less significant neighboring SNPs. This observation of hits in clusters aligns with the model of "haplotype blocks." That is, genomic regions are inherited together as sets (i.e., haplotype blocks) and nearby variants within the blocks can be highly correlated, known as linkage disequilibrium (LD) [8–10]. Although functionality is not guaranteed even with the strongest associations detected in GWASs, lead SNPs have been historically the focus of the field, treating the remaining associations as redundant [11]. In polygenic risk analysis where GWAS summary statistics are used to estimate the personal genetic risk of AD, the risk effect of neighboring SNPs is commonly excluded through pruning or clumping [12, 13]. Lead SNPs have also been widely used to assist with drug discovery since drug targets with genetic evidence of disease association are more likely to succeed [14]. Yet, lead SNPs identified from GWASs have not been consistent but rather nearby the same neighborhood [15]. The susceptibility locus in AD, reported as the nearest genes to lead SNPs are sometimes different even for the same SNP [15]. Taken together, information harbored in the neighborhood of lead SNPs may not necessarily be redundant. Focusing only on the lead SNPs will likely limit our understanding of genetic factors in AD [11]. Recent advancements in fine-mapping methods have recognized this limitation and strive to refine GWAS peak regions for causal variants linked to observed associations. Nevertheless, these methods continue to operate under the assumption of individual variant effects rather than considering their interactions [16, 17].

Advances in deep learning models have led to significant improvement in predicting the function of DNA sequence segment, such as transcription factor binding sites (TFBS) and histone marks [18, 19]. These models attained high accuracy in predicting the underlying chromatin marks in a tissue-specific manner [20]. Through in-silico mutagenesis, one can also examine how each individual allele affects the predicted function of the input DNA sequence. In this paper, we will utilize the recent deep learning model called ExPecto to investigate the downstream functional changes associated with the top GWAS hits in AD [18, 19]. In particular, we aim to explore: (1) What are the functional changes associated with AD lead SNPs? (2) Is there any difference in functional effect between lead SNPs and others in the same LD block? and (3) Will the functional effect of AD lead SNPs will be affected by the genetic context (i.e., alleles of neighboring SNPs)? Given that the genetic context (i.e., allele combination of each subject) is largely unknown, we employed a synthetic analysis of genetic context, examining all possible allele combinations within a defined window. To reduce the computational burden, we concentrated on top hits in the *APOE* region, which carry the highest genetic risk and are expected to have substantial downstream functional implications.

Methods

GWAS candidate loci

AD risk SNPs were extracted from a large-scale genome wide association study (GWAS), the International Genomics of Alzheimer's Project (IGAP). We chose IGAP over more recent larger-scale GWAS studies because the latter often include a substantial number of proxy dementia cases, potentially diluting the specificity of identified variants for AD and thereby reducing the explained heritability [6]. IGAP GWAS was performed with the imputed genotype of 11,480,632 single nucleotide poly- morphisms (SNPs) from 21,982 Alzheimer's disease patients and 41,944 cognitively normal controls. It is a combination of four consortia, namely, the Alzheimer Disease Genetics Consortium (ADGC), the European Alzheimer's disease Initiative (EADI), the Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium (CHARGE), and Genetic and Environmental Risk in AD Consortium Genetic and Environmental Risk in AD/Defining Genetic, and the Polygenic and Environmental Risk for Alzheimer's Disease Consortium (GERAD/PERADES) [21]. In this study, we focused on the top 100 significant SNPs with the smallest p-value in IGAP. In addition, neighboring SNPs located within the same linkage disequilibrium (LD) block of those top hits were also included, totaling to 238 variants. LD block information was estimated from the 1000 Genome Project using European population [22].

Deep genome annotation for allele-specific function

AD risk variants from GWASs are located predominantly in non-coding regions of the genome [23–25]. Only 7 out of the top 100 GWAS hits SNPs present in the coding region, with the rest in UTR, intronic regions, and other genetic components as detailed in Appendix B. Therefore, gene regulation is speculated to be one driving factor for Alzheimer's disease. Over the last decade, there has been significant progress in predicting regulatory marks from raw DNA sequences using deep learning models [18, 19, 26, 27]. More specifically, these models can generate the likelihood of functions (e.g., DNase peak or binding of a specific transcription factor) with a given DNA sequence segment. Allele-specific effect can be estimated by comparing the functional likelihood of two input sequences carrying major and minor allele respectively. For example, for DNase peak, if the likelihood generated from a sequence with major allele is much higher than that from a sequence with minor allele, this suggests a potential loss of DNase peak in minor allele carriers.

ExPecto is a pre-trained deep genome annotation model built on the data from the ENCODE and Roadmap Epigenomics projects [18, 19, 28, 29]. As input, a short DNA sequence centering the allele of interest is used to predict chromatin profiles, including transcription factor binding sites, histone marks, and DNase peaks across various tissues and cell types. In other words, it predicts whether any of those chromatin features exist in the input sequence. Given that the majority of GWAS findings are from non-coding regions, these chromatin profiles could reveal the critical role of gene regulation in complex diseases. ExPecto was trained to predict the likelihood of 2002 chromatin features and outperformed gkm-SVM, which was then the best method for chromatin immunoprecipitation–based TF binding prediction, with median AUC ≥ 0.95 across all chromatin features. Source code for the entire pipeline is freely accessible on GitHub website (https://github.com/PradoVarathan/Multi_Specto).

Allele-specific functions without genetic context

We first applied ExPecto to evaluate the allele-specific function of candidate SNPs without considering the genetic context, i.e., all the neighboring SNPs in the input sequence set to major allele. The input for ExPecto is a 2000 bp DNA sequence, centering around the SNP of interest. It was generated using the Hg38 Genome assembly as the reference genome, which was used to train the ExPecto model. For each candidate SNP, two input sequences were generated: (1) one 2000 bp reference sequence directly extracted from the reference genome, 999 bp upstream and 1000 bp down- stream. All the SNPs in the reference sequence were set to major alleles taken from IGAP dataset. (2) Another alternate sequence was generated by replacing the center SNP with minor allele taken from IGAP dataset. For both reference and alternate sequences, ExPecto predicted the functional likelihood of all chromatin features (Fig. 1 (a)). Log odds were derived from the predicted functional likelihood, and the log odds change between reference and alternate sequences reflected the predicted functional effect of the center SNP [18].

$$Log\ (OR) = \left|\log \frac{P\,(reference)}{1 - P\,(reference)} - \log \frac{P\,(alternate)}{1 - P\,(alternate)}\right| \tag{1}$$

Reversed reference and alternate sequences were also examined but the predicted chromatin profiles were almost identical, so the results were not included (Supplementary Fig. 1). In addition, with a focus on Alzheimer's disease, we manually screened all 2002 chromatin features in ExPecto and included only 128 features highly relevant to brain (Appendix A). That is, these chromatin features are either measured from brain tissues or purified brain specific cell types like neuron, microglial and astrocyte. Monocyte is also included due to its close relationship with brain [30]. In addition, we chose to include glioblastoma cell line since previous studies showed a significant overlap of genetic pathways between AD and glioblastoma/cancer [31, 32]. Although we do not fully understand the relationship between the top AD risk variants and glioblastoma yet, we seek to understand how these variants could possibly impact shared pathogenesis.

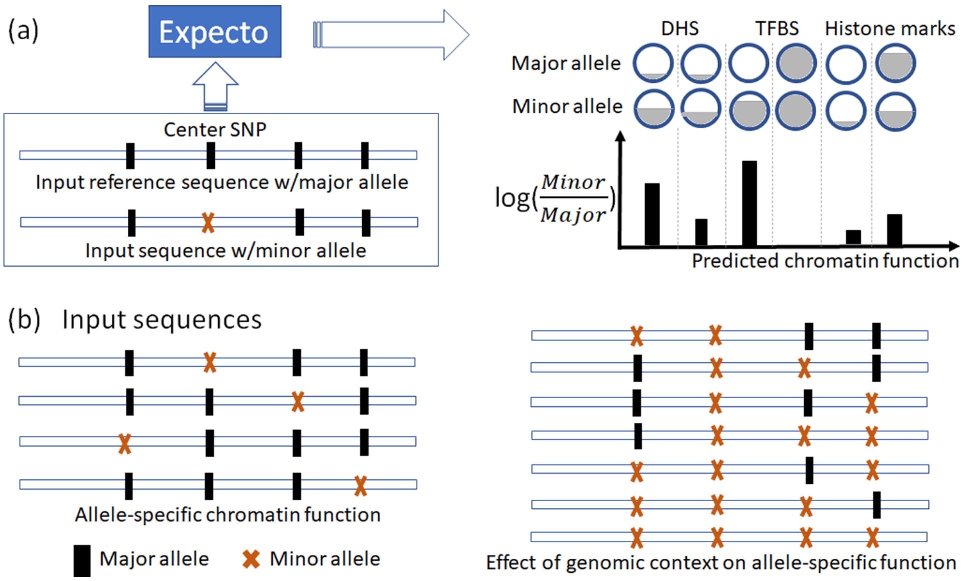
(a)
Expecto
Center SNP
Input reference sequence w/major allele
Input sequence w/minor allele
DHS
TFBS
Histone marks
Major allele
Minor allele
log($\frac{Minor}{Major}$)
Predicted chromatin function
(b) Input sequences
Allele-specific chromatin function
Major allele
Minor allele
Effect of genomic context on allele-specific function

Fig. 1 **(a)** Brief steps of ExPecto to estimate the functional impact of the allele of interest (center SNP). The shadow inside the circle indicates the likelihood of one chromatin function given a specific input sequence. DHS: DNase I hypersensitive site, TFBS: Transcription factor binding site. **(b)** Input sequences used to estimate the allele-specific chromatin effect without genetic context (left) and with genetic context (right)

Allele-specific effect with genetic context

Next, we tested the influence of the neighboring SNPs on the allele-specific functional effect. That is, the alleles of neighboring SNPs within 2000 bp flanking region will change the functional effect of the center SNP, e.g., enhancing or weakening the binding activity of a specific transcription factor. Toward this, we generated a set of alternate sequences with in-silico mutagenesis, where the center SNP remained minor allele, but neighboring SNPs selectively took minor alleles. We tested all possible combinations of major and minor alleles for N neighboring SNPs (2^N combinations in total) and examined whether any of the combinations would cause significant changes in the functional effect of the central SNP (Fig. 1 (b)). Since the input sequence of ExPecto model is only 2000 bp, there are only a limited number of SNPs within this window and thereby the computational feasibility of this step was ensured. Finally, we used the ADNI genotype dataset to validate the epistasis effects of those combinations in AD, which was downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu).

E value generation

To evaluate the significance of our findings, we randomly selected 1,000,000 SNPs from chromosomes 1–22 and examined their allele-specific chromatin effect on all 128 brain-related chromatin features. As such, for each chromatin feature, we obtained a distribution of log odds ratio changes. On top of that, we estimated the empirical p-values of all log odds ratios obtained using sequences around AD risk SNPs. Following [19], E-value was determined by the product of the log odds change (relative change) and the absolute change, and was formulated as follows:

$$\left|\log\frac{P\left(reference\right)}{1-P\left(reference\right)}-\log\frac{P\left(alternate\right)}{1-P\left(alternate\right)}\right|*\left|P\left(reference\right)-P\left(alternate\right)\right| \quad (2)$$

Results

Allele-specific effect without genetic context

After examining each individual AD risk SNP and its neighborhood SNPs, we found 8 of them with noteworthy log odds ratio changes in brain-related chromatin features. Six of those are among the top 100 AD GWAS candidate SNPs and two are in the 2000 bp neighborhood of those top SNPs, with one as GWAS significant but not the other. Among the top AD GWAS SNPs, sequences with minor allele of rs157585 was predicted to be associated with the prominent loss of function for DNase I hypersensitive sites (DHS) in glioblastoma cells, normal human astrocyte (NHA) and monocyte cells, and also histone marks in normal human astrocytes (Fig. 2 (a)). Another top AD GWAS variant, rs74579864, was predicted to provide strong gain of function in acetylation of histones 2 and 3 at various positions in H1-Derived Neuronal Progenitor cells (NPC). rs35396326 is also strongly associated with acetylation of histones 2 and 3 in H1-derived Neuronal Progenitor cells, but at different positions and in a negative way. Another prominent feature predicted to have a spike in the log odds ratio was from glioblastoma CTCF factor in rs75765623, located in the first intron of NECTIN2 gene. This variant is neither among the top GWAS SNPs nor a significant variant, but with the highest log odds ratio change (e-value=0.0458). It is located only 70 bp downstream of a significant AD GWAS hit rs12462573. Yet, the predicted chromatin effect associated with

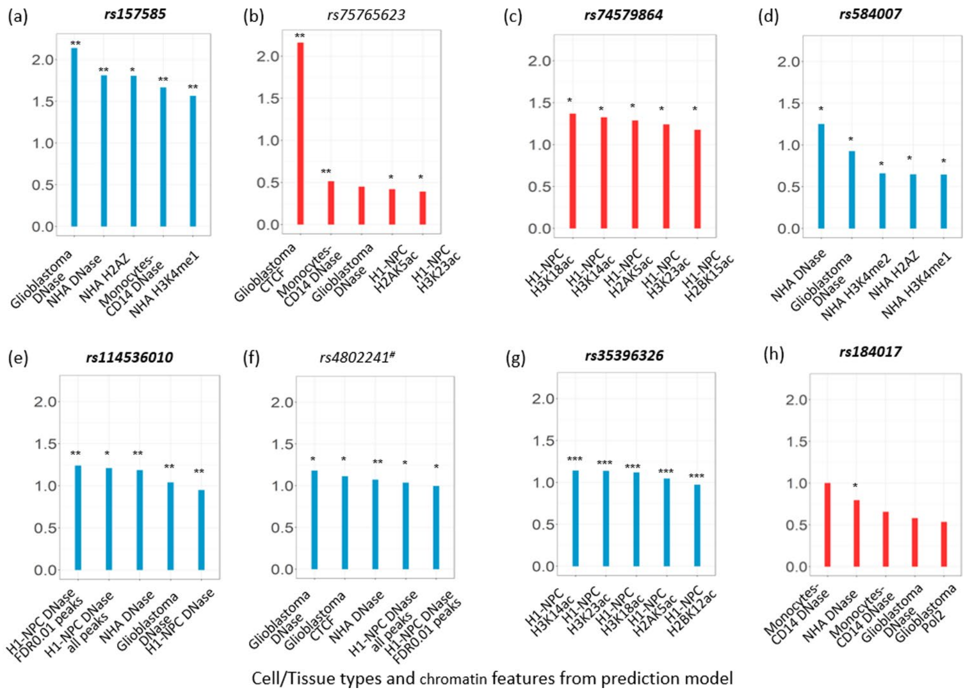

Fig. 2 Top chromatin features affected by GWAS SNPs and their neighboring SNPs without genetic context. That is, only the center SNP in the input sequence takes minor allele. SNPs with log odds change greater than 1 in at least one chromatin features were included in the figure. Red indicates a positive log odds ratio changes and blue indicates a negative log odds ratio change, suggesting gain and loss of function respectively. The variants in bold are among the top 100 AD GWAS SNPs. #: significant variants in the neighborhood of the top GWAS SNPs. *: e-value < 5e-2, **: e-value < 5e-3. ***: e-value < 1e-6. NHA: normal human astrocytes, H1-NPC: H1-derived neural progenitor cells

rs12462573 was minimal and negligible. When examined in the European population, we did not observe strong correlation between these two SNPs despite their closeness in physical location (R^2=0.004). It is therefore worth noticing that SNPs with significant p-value (i.e., lead SNPs) do not necessarily have the strongest downstream functional effect. Actual functional effect could come from less significant variants located in the neighborhood of top hits. Finally, we compared our discoveries with potential causal variants identified through fine mapping [33]. However, there is no overlap of our SNPs and causal SNPs prioritized through fine mapping of IGAP GWAS results, indicating possible additional functional implications due to genetic interactions.

Highly correlated SNPs within LD block showed distinct allele-specific effect

For those 8 SNPs predicted with significant allele-specific functional effect, we identified SNPs in the same LD block using LDLink [34] (Supplementary Fig. 2) and compared their functional effects predicted by ExPecto Among 8 SNPs, 5 of them have highly correlated SNPs (corr ≥ 0.8) located in the same LD block. Shown in Fig. 3 is the comparison of predicted allele-specific functional effect across highly correlated SNP groups. Each panel is a group of correlated SNPs in the same LD block and the first row is the SNP predicted with significant allele-specific effect in the above section. Interestingly, these highly correlated variants seldom had similar predicted chromatin profiles. For example, rs74579864, rs4803761 and rs4803762 are highly correlated (R^2=0.956 and R^2=0.978 respectively). However, minor allele in rs74579864 was predicted to increase the likelihood of histone marks in H1-derived neuronal progenitor cultured cells, but not those

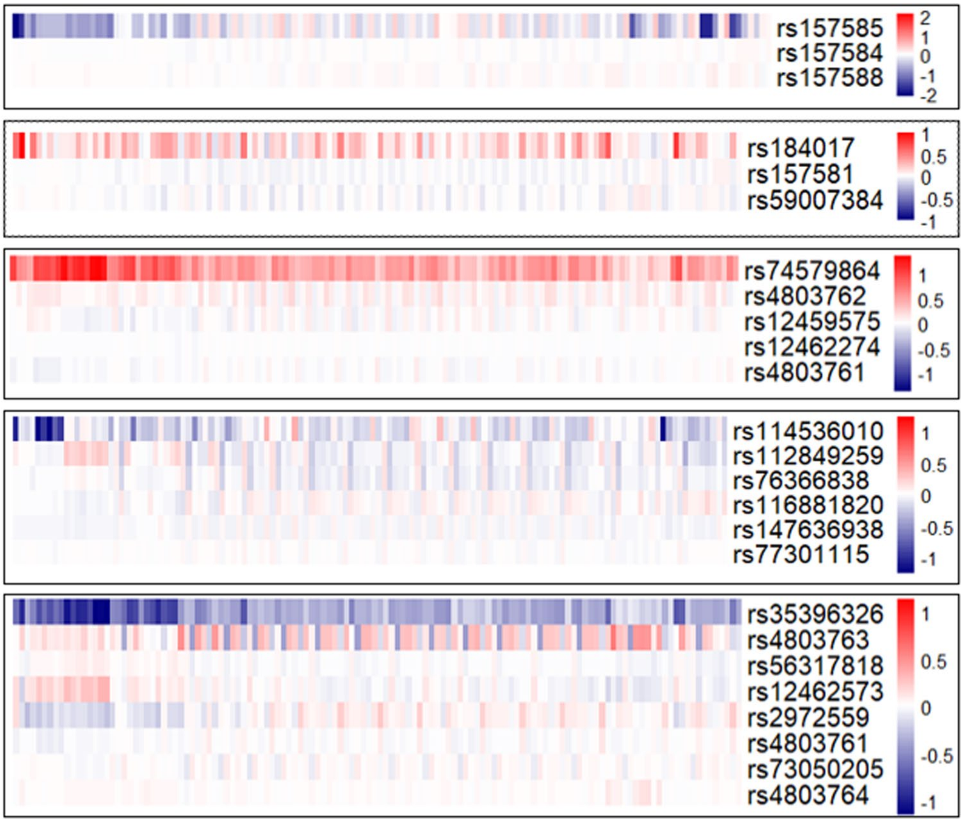

Fig. 3 Highly correlated variants ($R^2 \geq 0.8$) of rs157585, rs184017, rs74579864, rs114536010 and rs35396326 from the same LD block were predicted to have different functional effects

of rs4803761 and rs4803762. Similarly, sequences with minor allele in rs157585 were predicted with strong negative effect in DNase hypersensitive sites in monocytes CD14 and glioblastoma cells, neither of which was observed for its highly correlated neighbor rs157584 (R^2=0.98). Taken together, our findings suggest that caution should be exercised when pruning LD blocks to narrow down the number of SNPs, which will likely result in a loss of information and significantly biased results.

Allele-specific effect with genetic context

For all the top 100 AD GWAS SNPs, we additionally examined the influence of genetic context on the allele-specific functional effect. That is, whether the alleles of neighboring SNPs within the 2000 bp window will affect the predicted functional effect of the center SNP. As shown in Fig. 1 (b) on the right, input sequences centered around each SNP were modified by varying the allele of neighboring SNPs within the 2000 bp window (i.e., major to minor allele). As such, we were able to identify 21 SNPs predicted with strong effect on chromatin features (log odds ratio change≥1, $e \leq 0.05$). Predicted chromatin effects of these input sequences were compared with allele-specific effects without considering the genetic context. Ultimately, four variants were observed with dominant effect, including rs157585, rs184017, rs114536010, and rs74579864. For each of these SNPs, input sequences carrying their minor allele were observed to have very similar functional effects on chromatin features regardless of the genetic context in the 2000 bp window. Three SNPs showed notable and significant log odds ratio changes (≥1, $e \leq 0.05$), indicating the importance of the genetic context for SNP annotation.

We also observed notable and significant log odds ratio changes (≥1, $e \leq 0.05$) for variants rs1305062, rs2972559 and rs584007, which suggests that their predicted functional effect could be dependent on the allele of neighboring SNPs. For sequences carrying the minor allele in rs1305062, a significant loss of function was predicted for the CTCF binding sites in glioblastoma cells, which became even worse when the input sequence also carried the minor allele of rs141864196 (Fig. 4 (a)). A similar effect was observed for rs2972559, which in combination with rs4802241 led to a more significant loss of function in the predicted CTCF binding sites in glioblastoma cells. Interestingly, the chromatin effect observed for rs2972559 (as a top GWAS hit) alone was very weak (log odds

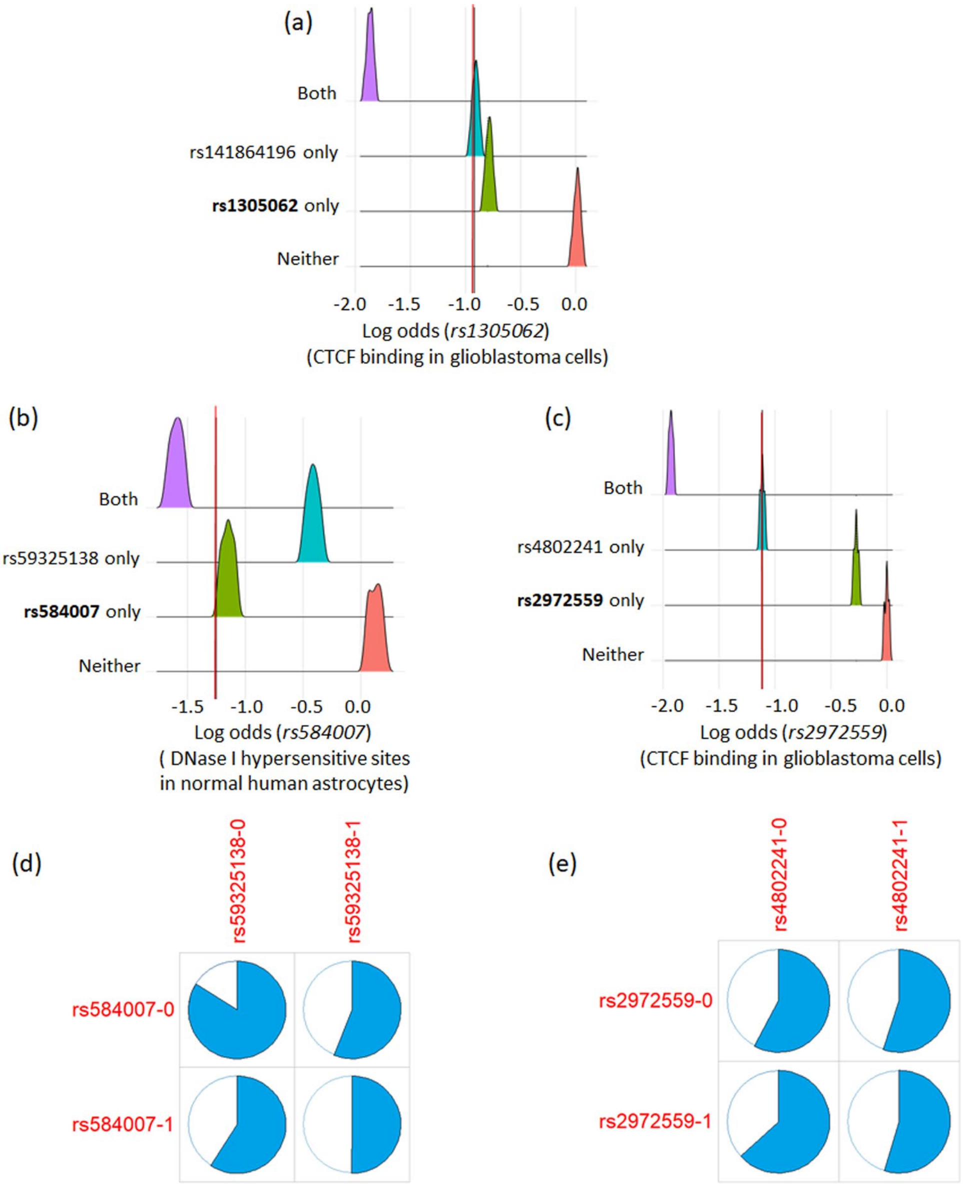

Fig. 4 **(a)** Distribution of the log odds ratio change in CTCF binding in glioblastoma cells for input sequences centered around rs1305062. **(b)** Distribution of log odds change of DNase I hyper- sensitive sites in Normal Human Astrocytes for input sequences centered around rs584007. **(c)** Distribution of the log odds ratio changes of CTCF binding in glioblastoma cells for input sequences centered around rs2972559. **(d)** Proportion of subjects developing AD in the ADNI cohort grouped by rs584007/rs59325138 genotype. **(e)** Proportion of subjects developing AD in the ADNI cohort grouped by rs2972559/rs4802241 genotype

ratio change *le* 0.5) (Fig. 4 (c)). In Fig. 4 (b), minor allele in rs584007 was predicted associated with the loss of DNase I hypersensitive sites in normal human astrocyte derived cells. This loss of function could become even more evident with the presence of minor allele in another SNP rs59325138, which had little chromatin effect as predicted by ExPecto.

We further investigated the interaction effect of these three pairs of SNPs in the ADNI cohort [35]. Shown in Fig. 4 (d) and (e) are the proportion of subjects that ultimately developed AD in each genotype group (0 for no minor allele and 1 for presence of minor allele). rs141864196 was not reported due to its missing genotype in the ADNI. For the first pair of SNPs rs584007 (GWAS p=1.056e-82, beta=-0.37) and rs59325138 (GWAS p=6.945e-89, beta = -0.38), ratio of subjects developing AD decreases with the presence of minor allele, indicating their potential protective effect. Both SNPs are located within the LD block of lead SNP rs1081105 (i.e., top 100 GWAS hits). Combined with their predicted chromatin effect, it is speculated that deactivated DNase hypersensitive sites in astrocyte cells may have a protective role in AD development. When tested in PLINK using genotype data, these two variants also showed significant epistasis effect ($p < 1e-4$). For the second pair of SNPs, only rs2972559 is a significant AD risk SNP in GWASs but not rs4802241. Despite no significant epistasis effect detected in PLINK, we observed that the presence of minor allele in rs4802241 to some extent mitigates the risk of developing AD introduced by the minor allele in rs2972559 (Fig. 4 (e)). Taken together, co-presence of minor alleles in rs4802241 and rs2972559 are likely associated with decreased CTCF binding in glioblastoma cells and reduced risk of AD, but their connections are yet to be further investigated.

Discussion

This study investigated the chromatin effect of sequences surrounding AD risk SNPs by leveraging deep genome annotation tools. Among the top predicted histone marks, the most significant log odds ratio changes were primarily observed in acetylation of histone H3, like H3K18ac and H3K23ac, all associated with GWAS SNP rs35396326 (beta=0.38, p=1.35e-86 in GWAS). Acetylation levels of histones H3 and H4 have been previously reported to be lower overall in postmortem AD brains than in control brains. Among those, H3K18ac and H3K23ac were further validated as the most significantly hypoacetylated histone marks, along with H3K9ac, H3K27ac and H4K16ac [36]. In line with that, elevated levels of H3K14ac within the calpastatin promoter region was observed together with significantly decreased neuronal toxicity in neuroblastoma cells that underwent treatment to inhibit calcium-induced neuronal cell death [37]. While calcium-induced neuronal cell death is found strongly associated with the pathophysiology of AD, this evidence together suggests a potential neuroprotective role of histone acetylation. Our results provide extra support for the hypothesis that the decrease in histone acetylation is associated with the minor allele of the AD risk SNP rs35396326 with positive beta coefficient (β=0.38, p=1.35$e-86$ in GWASs). In other words, our findings suggest that minor allele of rs35396326 is associated with greater risk of developing AD and greater likelihood of decreased histone acetylation.

Another group of chromatin features predicted to be strong associates with the top AD GWAS hits are DNases in normal human astrocytes and monocytes. The most significant log odds ratio changes in DNase activity came from *rs157585* and was specific to monocytes, astrocytes and glioblastoma cells. In a DNase I footprinting analysis, mutations inside two DNase hypersensitive sites within recombinant *AP-2* were found associated with the regulation of the *APOE* promoter region, thereby implicating their role in the pathogenesis of AD [38]. Specifically, multiple DNase-I hypersensitive sites were reported to be significantly associated with AD risk transcriptional factors in monocytes and macrophages [39]. The role of glial cells such as microglia, monocytes and astrocytes in neuroinflammation and AD have been widely studied, wherein the Aβ- activated glial cells produce cytokines and chemokines which in turn activate pathways leading to demyelination, oxidative stress and eventually cell death [40]. Although DNase I has been recently speculated to be a potential therapeutic intervention for AD, cell-type specific DNase I activity is overall under explored in AD [41, 42].

Another crucial finding of this investigation is that SNPs in the same LD block with extremely high correlation (≥0.9) were predicted to have a very distinct effect on chromatin functions, and 2) variants that are not significant but in the neighborhood of GWAS hits could still have an impact on the downstream function (rs75762623 in Fig. 2 (b)). These results provided further proof that treating LD-block as redundant information and having one variant to represent the entire LD block could possibly bias the functional annotation of GWAS findings and our interpretation of disease mechanism.

In addition, we also observed a significant genetic context effect on the predicted functional effect of risk alleles. Among the top 100 AD GWAS SNPs, co-presence of minor alleles in two sets of neighboring SNPs were predicted with greater loss of function in CTCF binding activities in glioblastoma cells and DNase hypersensitivity sites in astrocytes. This provides evidence to support the importance of the genetic context surrounding GWAS hits, which should not be simply treated as redundant information. Similar findings have only been recently reported for other diseases such as Brugada syndrome [11]. While GWAS findings have been increasingly leveraged for many important downstream applications such as polygenic risk estimation and discovery of drug targets, caution should be exercised when utilizing the GWAS findings, especially considering the limited replicability of polygenic risk scores and failure of many clinical trials.

Conclusion

Taken together, our results suggest the need for reanalysis of published AD GWAS data and reconsideration of future application plans for GWAS findings. This work has several limitations that merit further consideration. First, given that a long input sequence and large number of variants could lead to an exponentially high number of genotype combinations as genomic context, we constrained this project to the top 100 AD GWAS SNPs, which are mostly located around the *APOE* region and employed ExPecto with a 2000 bp input sequence. Due to the lack of genome sequence data for real patients, we performed a synthetic analysis of genetic context in which we examined all possible combinations of minor alleles across the SNPs within the 2000 bp window, some of which may not exist in population data. Applying this approach to genomic sequences obtained from actual patients could be especially beneficial, as it would capture a more precise genomic context in which AD-associated risk variants exert their effects. In addition, we restricted our investigation to cell lines and tissues related to the brain, given that Alzheimer's disease primarily affects the brain. However, recent research indicates potentially important, yet unexplored, associations of AD with other organs. Therefore, the functional effects of AD variants in these tis- sues also merit further investigation. Some of these limitations could be addressed with further haplotype estimation from phased genotypes in large cohorts such as the UK Biobank or ADNI. Finally, our analysis is constrained to the 2000 bp window around GWAS lead SNPs near the *APOE* region. With individual subject data containing actual allele combinations, we could enhance our approach by integrating fine-mapping methods with Mendelian randomization. This would allow us to refine genetic regions more precisely and explore the impact of genetic context in greater depth. Overall, this study provides a new perspective for interpreting GWAS findings and new evidence to support the non-redundancy hypothesis of neighborhood variants surrounding GWAS hits. More in-depth work is needed to further investigate the functional effect of GWAS hits as clusters.

Appendix A Top AD GWAS SNPs

The top 100 AD GWAS SNPs included in this study are listed in Appendix A.csv.

Appendix B brain related chromatin features in ExPecto

Brain-related chromatin features related to brain were manually extracted from ExPecto, and are listed in Appendix B.csv.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13040-024-00400-1.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3: Fig. 1- Predicted chromatin profiles of reversed reference and alternate sequences were almost identical, with rs157585 as an example.

Supplementary Material 4: Fig. 2- LD block of 8 SNPs that showed significant effect on chromatin features when genetic context was not considered.

Acknowledgements

Data collection and sharing for this project was partly funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (http://www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Author contributions

P.V and J.Y contributed to the design, the implementation of the research, the analysis of the results and the writing of the manuscript. B.H, and L. X, helped with the data curation for this project. K. N and A. S helped with the result interpretation.

Funding

This research was supported by National Institute of Health grants R01AG081951, R21 AG066135, U01AG068057, U19AG074879 and National Science Foundation CAREER award 1942394.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Competing interests

Dr. Yan is a guest editor of this special collection. Dr. Saykin has received support from Avid Radiopharmaceuticals, a subsidiary of Eli Lilly (in kind contribution of PET tracer precursor) and holds advisory roles with Siemens Medical Solutions USA, Inc., NIH NHLBI, and Eisai. His editorial commitments include serving as Editor-in-Chief for the journal `Brain Imaging and Behavior', and he participates in various NIH/NIA advisory committees.

Received: 17 January 2024 / Accepted: 14 October 2024

Published online: 13 November 2024

References

1. Gaugler J, James B, Johnson T, Reimer J, Solis M, Weuve J, Buckley RF, Hohman TJ. 2022 Alzheimer's disease facts and figures. ALZHEIMERS Dement. 2022;18(4):700–89.
2. Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, Fiske A, Pedersen NL. Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry. 2006;63(2):168–74.
3. Coon KD, Myers AJ, Craig DW, Webster JA, Pearson JV, Lince DH, Zismann VL, Beach TG, Leung D, Bryden L, et al. Focus on Alzheimer's disease and related disorders-a high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. J Clin Psychiatry. 2007;68(4):613.
4. Marioni RE, Harris SE, Zhang Q, McRae AF, Hagenaars SP, Hill WD, Davies G, Ritchie CW, Gale CR, Starr JM, et al. Gwas on family history of Alzheimer's disease. Translational Psychiatry. 2018;8(1):99.
5. Wightman DP, Jansen IE, Savage JE, Shadrin AA, Bahrami S, Holland D, Rongve A, Børte S, Winsvold BS, Drange OK, et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. Nat Genet. 2021;53(9):1276–82.
6. Escott-Price V, Hardy J. Genome-wide association studies for Alzheimer's disease: bigger is not always better. Brain Commun. 2022;4(3):125.
7. Wang H, Bennett DA, De Jager PL, Zhang Q-Y, Zhang H-Y. Genome- wide epistasis analysis for Alzheimer's disease and implications for genetic risk prediction. Alzheimer's Res Therapy. 2021;13(1):1–13.
8. Crawford DC, Bhangale T, Li N, Hellenthal G, Rieder MJ, Nickerson DA, Stephens M. Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat Genet. 2004;36(7):700–6.
9. Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science. 2001;294(5547):1719–23.
10. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. Science. 2004;304(5670):581–4.
11. Olmo B, P´erez-Agustin A, Mates J, Allegue C, Iglesias A, Ma Q, Merkurjev D, Konovalov S, Zhang J, Sheikh F, et al. Analysis of brugada syndrome loci reveals that fine-mapping clustered gwas hits enhances the annotation of disease-relevant variants. Cell Rep Med. 2021;2(4):100250.
12. Priv´e F, Vilhj´almsson BJ, Aschard H, Blum MG. Making the most of clumping and thresholding for polygenic scores. Am J Hum Genet. 2019;105(6):1213–21.
13. Chasioti D, Yan J, Nho K, Saykin AJ. Progress in polygenic composite scores in Alzheimer's and other complex diseases. Trends Genet. 2019;35(5):371–82.
14. King EA, Davis JW, Degner JF. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. PLoS Genet. 2019;15(12):1008489.
15. Andrews SJ, Fulton-Howard B, Goate A. Interpretation of risk loci from genome-wide association studies of Alzheimer's disease. Lancet Neurol. 2020;19(4):326–35.
16. Schwartzentruber J, Cooper S, Liu JZ, Barrio-Hernandez I, Bello E, Kumasaka N, Young AM, Franklin RJ, Johnson T, Estrada K, et al. Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. Nat Genet. 2021;53(3):392–402.
17. Lake J, Warly Solsberg C, Kim JJ, Acosta-Uribe J, Makarious MB, Li Z, Levine K, Heutink P, Alvarado CX, Vitale D, et al. Multi-ancestry meta- analysis and fine-mapping in Alzheimer's disease. Mol Psychiatry. 2023;28(7):3121–32.
18. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat Genet. 2018;50(8):1171–9.
19. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nat Methods. 2015;12(10):931–4.
20. Dey KK, Geijn B, Kim SS, Hormozdiari F, Kelley DR, Price AL. Evaluating the informativeness of deep learning annotations for human complex diseases. Nat Commun. 2020;11(1):4703.
21. Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, Boland A, Vronskaya M, Van Der Lee SJ, Amlie-Wolf A, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates aβ, tau, immunity and lipid processing. Nat Genet. 2019;51(3):414–30.
22. Consortium GP, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68.
23. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. Systematic localization of common disease-associated variation in regulatory dna. Science. 2012;337(6099):1190–5.
24. Pantelis C, Papadimitriou GN, Papiol S, Parkhomenko E, Pato MT, Paunio T, Pejovic-Milovancevic M, Perkins DO, Pietil¨ainen O, et al. Bio- logical insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511(7510):421–7.
25. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 2017;101(1):5–22.
26. Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res. 2018;28(5):739–50.

27. Avsec Zˇ, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Tay- lor KR, Assael Y, Jumper J, Kohli P, Kelley DR. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18(10):1196–203.
28. Souza N. The encode project. Nat Methods. 2012;9(11):1046–1046.
29. Roadmap EC, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317–30.
30. Feng Y, Li L, Sun X-H. Monocytes and Alzheimer's disease. Neurosci Bull. 2011;27(2):115.
31. Lanni C, Masi M, Racchi M, Govoni S. Cancer and Alzheimer's disease inverse relationship: an age-associated diverging derailment of shared pathways. Mol Psychiatry. 2021;26(1):280–95.
32. S´anchez-Valle J, Tejero H, Ib´an˜ez K, Portero JL, Krallinger M, Al- Shahrour F, Tabar´es-Seisdedos R, Baudot A, Valencia A. A molecular hypothesis to explain direct and inverse co-morbidities between Alzheimer's disease, glioblastoma and lung cancer. Sci Rep. 2017;7(1):1–12.
33. Mountjoy E, Schmidt EM, Carmona M, Schwartzentruber J, Peat G, Miranda A, Fumis L, Hayhurst J, Buniello A, Karim MA, et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. Nat Genet. 2021;53(11):1527–33.
34. Machiela MJ, Chanock SJ. Ldlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics. 2015;31(21):3555–7.
35. Weber CJ, Carrillo MC, Jagust W, Jack CR Jr, Shaw LM, Trojanowski JQ, Saykin AJ, Beckett LA, Sur C, Rao NP, et al. The worldwide Alzheimer's disease neuroimaging initiative: Adni-3 updates and global perspectives. Alzheimer's Dementia: Translational Res Clin Interventions. 2021;71:12226.
36. Zhang K, Schrag M, Crofton A, Trivedi R, Vinters H, Kirsch W. Targeted proteomics for quantification of histone acetylation in Alzheimer's disease. Proteomics. 2012;12(8):1261–8.
37. Seo J, Jo SA, Hwang S, Byun CJ, Lee H-J, Cho D-H, Kim D, Koh YH, Jo I. Trichostatin a epigenetically increases calpastatin expression and inhibits calpain activity and calcium-induced sh-sy 5 y neuronal cell toxicity. FEBS J. 2013;280(24):6691–701.
38. Garcia MA, V´azquez J, Gim´enez C, Valdivieso F, Zafra F. Transcription factor ap-2 regulates human apolipoprotein e gene expression in astrocytoma cells. J Neurosci. 1996;16(23):7550–6.
39. Tansey KE, Cameron D, Hill MJ. Genetic risk for Alzheimer's disease is concentrated in specific macrophage and microglial transcriptional networks. Genome Med. 2018;10:1–10.
40. Fakhoury M. Microglia and astrocytes in Alzheimer's disease: implications for therapy. Curr Neuropharmacol. 2018;16(5):508–18.
41. Smalheiser NR. Mining clinical case reports to identify new lines of investigation in Alzheimer's disease: the curious case of dnase I. J Alzheimer's Disease Rep. 2019;3(1):71–6.
42. Tetz V, Tetz G. Effect of deoxyribonuclease I treatment for dementia in end- stage Alzheimer's disease: a case report. J Med Case Rep. 2016;10:1–3.

Publisher's note