Apply fine-mapping techniques to summary statistics from a GWAS of Alzheimer's disease in European Population

Jianing Wang, Hanqing Zhao

Abstract

Previous studies have shown that genetics may play a role in Alzheimer's disease (AD). The genome-wide association study (GWAS) has discovered many variants of AD, yet causal variants still remain unknown. Utilizing the fine-mapping techniques to refine the relocalization of causal variants, we want to see what insights the fine-mapping results could give us. In our final project, we used a ready-made pipeline[1] to compare the results of top causal variants found by three fine-mapping softwares as well as working on the interpretation by mainly doing literature research on the top variants.

Introduction

Alzheimer's disease is a fatal form of dementia and there is currently no cure for this disease. According to the Centers for Disease Control and Prevention (CDC), it is the sixth leading cause of death in the United States in 2014, and the number of people living with the disease doubles every 5 years beyond age 65. Research has shown that those who have a parent, brother or sister with AD are more likely to develop the disease and therefore researchers believe that genetics may play a role in developing AD. Therefore, we are interested in finding the potential causal variants in AD patients using fine-mapping on the genome-wide association study (GWAS) data.

Fine-mapping is a way to refine the localization of causal variants by statistical, bioinformatics or functional methods[2]. It is used to identify the particular genetic variants that are likely to influence the examined trait, which are known to be causal variants. The three fine-mapping tools we use are PAINTOR[3], CAVIARBF[4] and FINEMAP[5]. They all use similar models in the Bayesian framework, but have different algorithms to optimize the interferences. In particular, PAINTOR uses the MCMC algorithms with 1000 Genomes data to improve resolution of statistical fine-mapping[3]. CAVIARBF models the uncertainty of the effect size (or noncentrality parameters) and perform exhaustive model search[4], and FINEMAP uses Shotgun Stochastic Search (SSS) algorithms that explore the vast space of causal configurations by concentrating effects on the configurations with non-negligible probability[5].

Current studies have focused on improving the analysis methods on GWAS data, and interpreting the fine-mapping results, but have not identified the true causal variants of AD. It is challenging to identify the causal genes and variants from GWAS data because noncoding associations can affect genes that are far apart[6]. Restrained by the scope of the project, our motivation is to find what variants associated with AD might be the causal variants in European population, compare the top causal variants identified by the three tools, and interpret the results by literature research.

Methods

We obtained our GWAS summary statistics from GWAS Catalog (www.ebi.ac.uk/gwas/). The dataset was generated from a meta-analysis of 4 previously published GWAS summary statistics of AD[7]. It consists of 17,008 AD cases and 37,154 controls, and their imputed genotype data (7,055,881 SNPs). All the samples included are from European population.

The three well-known fine-mapping softwares we used are PAINTOR (v3.0), CAVIARBF, and FINEMAP (v1.3.1). Here, we ran the version of PAINTOR without feeding any functional annotations. These softwares also require a reference panel for linkage disequilibrium (LD) information, thus we used 1000 Genome Phase 1 data to build the reference panel[8]. To accelerate the computation, we partitioned all variants into relatively independent LD blocks detected by ldetect[9], and selected genome-wide significant variants (*P*-value < $5x10^{-8}$) in each block. Then, we used PAINTOR's framework to estimate Pearson's correlation coefficient between each variant in the LD block and convert the results into a genotype matrix for each block. After all data preparations, we ran three fine-mapping programs with the default parameters, and assumed one maximum causal variant is allowed in each region. The reported credible sets are organized into one single file for further analysis. The pipeline up to this step is automated using the script[1] from Wang *et al*.'s paper.

After gathering the credible sets for each causal block, we visualized and examined the results using LocusZoom[10] (locuszoom.org). Annotations of each SNP reported in the credible sets are retrieved from dbSNP[11] (www.ncbi.nlm.nih.gov/snp/). Along with the results from literature search, we evaluated each identified causal variant, and compared the credible sets from three softwares. The code and supplementary data are available in our GitHub repository (github.com/insanebruce/fine-mapping).

Results

We made the Manhattan plot (Fig. 1) for the dataset, and found mainly 13 peaks, or causal blocks. Additionally, we observe that there are many significant SNP variants especially in chromosome 2, 11, and 19. The distribution of *P*-values of each variant from the GWAS statistics further supports the existence of a genetic basis of AD, and sheds light on some potential genetic risk factors or causal variants.

To benchmark the performance of three fine-mapping programs, we first recorded the execution time spent by each software on Datahub. Despite the extremely long computation time for the preparation of the LD reference panel, all these softwares are very time-efficient to identify causal variants from the input dataset. FINEMAP runs the fastest, with an execution time of 335 seconds, and is then followed by CAVIARBF with a running time of 810 seconds. PAINTOR has the worst performance of execution time, and takes 1095 seconds to finish the computation. In conclusion, FINEMAP outperforms PAINTOR and CAVIARBF in terms of execution time as expected because the Shotgun Stochastic Search algorithm implemented by FINEMAP has the superiority in speed.

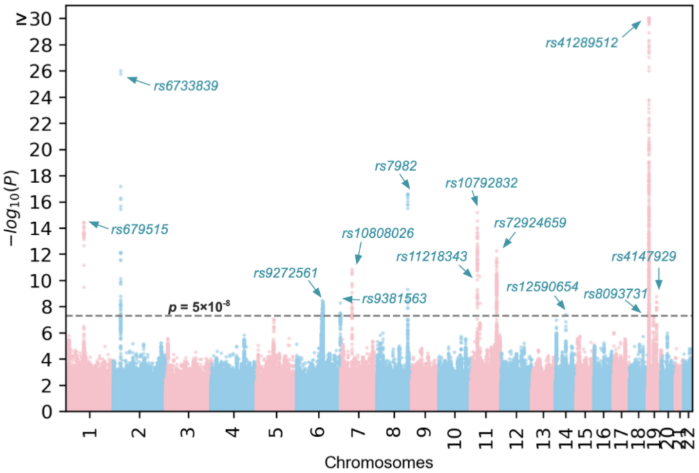


Figure 1 Manhattan plot of GWAS summary statistics of Alzheimer's disease. Our genome-wide significance level for *p*-value is $5×10^{-8}$, which is labelled by a black dashed line. The SNPs with highest posterior probabilities to be causal predicted by all three softwares in each causal block are labelled on the plot.

As all three softwares output posterior probabilities (PP) for each variant to be causal, and the credible sets of variants, we compare these results using Venn diagrams (Fig. 2). The credible set refers to a set of variants with a sum of PP of more than a threshold, which we used the conventional cutoff of 0.95. The number of variants in the credible set is generally inversely proportional to the respective PP predicted by the fine-mapping softwares. Overall, the results from three fine-mapping softwares show great agreement between each other, as the numbers of variants concentrate in the central intersection of the Venn diagrams.

Then, we explored the results generated by these fine-mapping softwares. The highest signal in the Manhattan plot (Fig. 1) lies in the causal block 13 on chromosome 19. A single SNP variant rs41289512 ($P = 2.24×10^{-167}$) is predicted to be the causal variant with PP equal to 1 by all three fine-mapping softwares. Besides, Fig. 3A shows that there are no highly correlated variants nearby, further supporting the prediction. dbSNP annotation indicates this SNP is mapped to the intron region of *PVRL2/NECTIN2*. This gene encodes a modulator of T-cell signaling, and can be either a costimulator of T-cell function, or a co-inhibitor depending on the receptor it binds to[12]. Further literature search indicates that *NECTIN2* is a strongly associated variant and also detected in many earlier studies. The potential involvement could be through its role in cell adhesion and brain's susceptibility to viral infections during aging, leading to neuronal loss, and may act together with nearby associated SNPs such as *APOE* and *T0MM* that are also commonly detected[13,14]. With all the information we obtained, we postulate that rs41289512 is likely to be the causal variant, but we admit there is still a possibility that some adjacent variants mapped to *APOE* or *T0MM* are the true causal variants instead. Even though those variants do not have a high correlation with rs41289512, it would not be surprising to find out that some of them are the main contributors considering that numerous variants with significantly low *P*-values are clustered in this region.

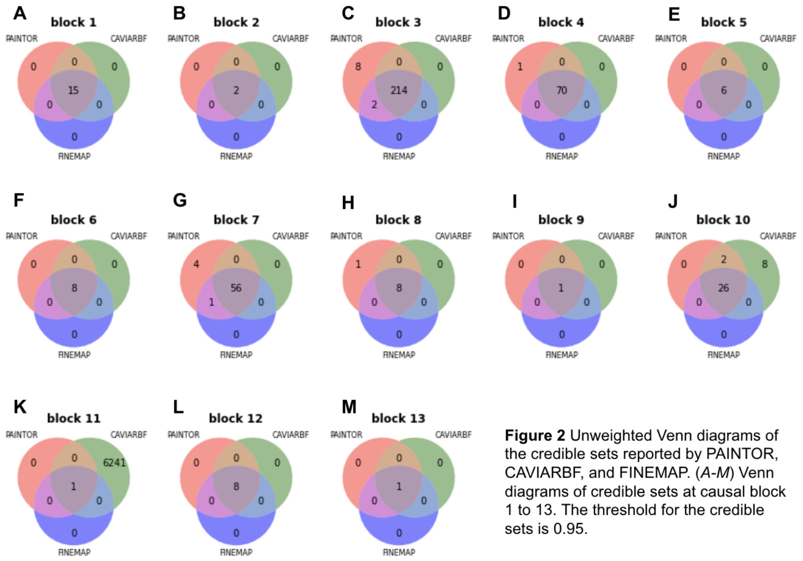


Figure 2 Unweighted Venn diagrams of the credible sets reported by PAINTOR, CAVIARBF, and FINEMAP. (*A-M*) Venn diagrams of credible sets at causal block 1 to 13. The threshold for the credible sets is 0.95.

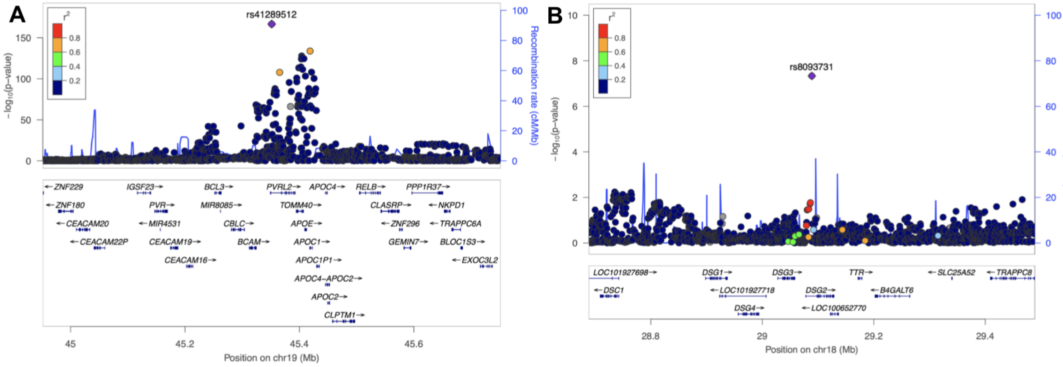
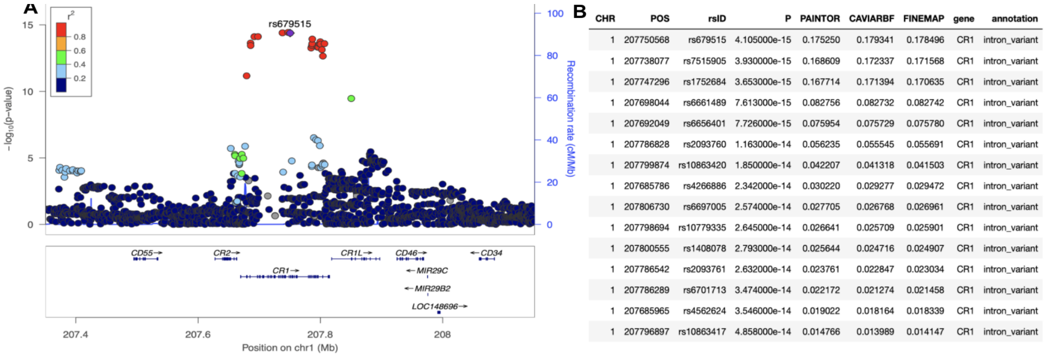


Figure 3 Regional plots made by LocusZoom. The diamond represents the causal variant identified by PAINTOR, CAVIARBF, and FINEMAP. (*A*) Region plot for the causal block 13 on chromosome 19. (*B*) Regional plot for the casual block 11 on chromosome 18.



CHR	POS	rsID	P	PAINTOR	CAVIARBF	FINEMAP	gene	annotation
1	207750568	rs679515	4.105000e-15	0.175250	0.179341	0.178496	CR1	intron_variant
1	207738077	rs7515905	3.930000e-15	0.168609	0.172337	0.171568	CR1	intron_variant
1	207747296	rs1752684	3.653000e-15	0.167714	0.171394	0.170635	CR1	intron_variant
1	207698044	rs6661489	7.613000e-15	0.082756	0.082732	0.082742	CR1	intron_variant
1	207692049	rs6656401	7.726000e-15	0.075954	0.075729	0.075780	CR1	intron_variant
1	207786828	rs2093760	1.163000e-14	0.056235	0.055545	0.055691	CR1	intron_variant
1	207799874	rs10863420	1.850000e-14	0.042207	0.041318	0.041503	CR1	intron_variant
1	207685786	rs4266886	2.342000e-14	0.030220	0.029277	0.029472	CR1	intron_variant
1	207806730	rs6697005	2.574000e-14	0.027705	0.026768	0.026961	CR1	intron_variant
1	207798694	rs10779335	2.645000e-14	0.026641	0.025709	0.025901	CR1	intron_variant
1	207800555	rs1408078	2.793000e-14	0.025644	0.024716	0.024907	CR1	intron_variant
1	207786542	rs2093761	2.632000e-14	0.023761	0.022847	0.023034	CR1	intron_variant
1	207786289	rs6701713	3.474000e-14	0.022172	0.021274	0.021458	CR1	intron_variant
1	207685965	rs4562624	3.546000e-14	0.019022	0.018164	0.018339	CR1	intron_variant
1	207796897	rs10863417	4.858000e-14	0.014766	0.013989	0.014147	CR1	intron_variant

Figure 4 Regional plot and the credible set for casual block 1. (*A*) Regional plot for the causal block 1 on chromosome 1. The diamond represents the SNP with the highest posterior probability in the credible sets. (*B*) The credible set for causal block 1. Columns *PAINTOR, CAVIARBF, FINEMAP* = posterior probabilities estimated by each software respectively; *gene, annotation* = the mapped gene and annotations from dbSNP.

Fig. 3B is another top causal variant captured by these fine-mapping tools. The variant rs8093731 lies in the block 11 on chromosome 18 (P = $4.63×10^{-8}$). It does not have highly correlated variants nearby either. The mapped gene is *DSG2,* a component of intercellular desmosome junctions. This gene is also involved in the interaction of plaque proteins and intermediate filaments mediating cell-cell adhesion. Studies have shown that, *DSG2* was previously found to be a risk factor for late-onset AD through GWAS studies, and their functional analysis also suggests *DSG2* may be functionally important in the development of *APOE* Ɛ3/4 allele, which is known to be the strongest genetic risk factor of sporadic AD[15]. With the information we found, we have some confidence to say this variant is likely to be a potential causal variant of AD.

Fig. 4 is an example of variants that have many highly correlated SNPs around. The red dots with high r^2 values indicate they are correlated with each other, with the top variant being rs679515 located in block 1 on chromosome 1 (P = $4.106×10^{-15}$). Those high correlated SNPs all mapped to *CR1,* which is a member of the receptors of complement activation (RCA) family. Decreases in expression of this protein and/or mutations in this gene have been associated with many diseases including AD[16]. *CR1* has been proved to affect the susceptibility of AD in many previous studies. Those studies found multiple SNPs in *CR1* were significantly linked to amyloid β (A**β**) metabolism of AD patients, which might be involved in developing AD via regulating A**β** accumulation. The amyloid β plaques have been regarded as the neuropathological hallmarks of AD[17,18]. Therefore, our results are consistent with the previous studies. The SNPs that are highly correlated might function together to affect the expression of *CR1* and contribute to the development of AD.

Discussion

We utilized three fine-mapping softwares to identify causal variants in 13 causal blocks for the GWAS summary statistics of AD in European population. The entire collection of credible sets can be found in the Supplementary data. With the assumption that a maximum of one causal SNP is allowed in each block, the results of PAINTOR, CAVIARBF, and FINEMAP reach some consensus in general. When only a few number of variants are reported in the credible sets and shared by all three softwares, the identified causal variant has a high chance of being the true causal variant based on the regional plot, dbSNP annotations, and previous studies. Nonetheless, these fine-mapping softwares fail to determine the causal variant when there are multiple candidate variants in certain regions, which suggests our assumption for the maximum number of causal variants needs refinement. For example, only two variants with PP close to 0.5 are reported at block 2 (Fig. 2B). Judging from the regional plot (see Supplementary data), we hypothesize that either only one of them is the true causal variant, and the other becomes the rival due to high correlation, or both are true causal variants. We actually ran fine-mapping that allows a maximum of two causal variants in each block, while the results demonstrate significant discrepancies from these softwares (see Supplementary data). For instance, FINEMAP reports rs406315 and rs12980613 with PP = 1 for block 13, while PAINTOR assigns 0 to both of them, but reports rs157580 and rs741780 with PP = 1. These two pairs of variants are mapped to two adjacent genes, but based on the regional plot, it is hard to make any promising conclusions. Also, PAINTOR and CAVIARBF tend to report similar results. More parameters should be tested, and a more thorough analysis of those results may offer some insight. Experimental validation is needed to help determine the true causal variants and build the foundation for a better comparison of these fine-mapping programs.

Author Contributions

H.Z. worked on the background information search, and results interpretation with literature search for the top variants identified by the three tools. J.W. ran the programs, created plots, and participated in literature search. H.Z. and J.W. wrote the report together.

Reflection

The project took about five days for us to generate the current results and come up with preliminary interpretations. The progress is relatively smooth without having too much trouble and team cooperation is good. The part that we might need more guidance is how to determine whether the variants we found are indeed causal variants. From the literature search, although we have some confidence in our results, it is still unsure whether they can be determined as causal variants without future functional analysis. One challenge we encountered was that preparing the LD reference panel is really time-consuming. We worked around by “divide-and-conquer” and ran tasks in parallel, but we also found a precomputated reference panel online afterwards.

References

1. Jianhua Wang, Dandan Huang, Yao Zhou, Hongcheng Yao, Huanhuan Liu, Sinan Zhai, Chengwei Wu, Zhanye Zheng, Ke Zhao, Zhao Wang, Xianfu Yi, Shijie Zhang, Xiaorong Liu, Zipeng Liu, Kexin Chen, Ying Yu, Pak Chung Sham, Mulin Jun Li, CAUSALdb: a database for disease/trait causal variants identified using summary statistics of genome-wide association studies, *Nucleic Acids Research*, 48 (1), 807–816 (2020), https://doi.org/10.1093/nar/gkz1026

2. Schaid, D. J., Chen, W., & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, *19*(8), 491–504, (2018). https://doi.org/10.1038/s41576-018-0016-z

3. Kichaev G, Yang W-Y, Lindstrom S, Hormozdiari F, Eskin E, et al. Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLoS Genet* 10(10): e1004722. (2014) doi:10.1371/journal.pgen.1004722

4. Chen W, Larrabee BR, Ovsyannikova IG, Kennedy RB, Haralambieva IH, Poland GA, Schaid DJ. Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics*, 200(3):719-36 (2015). doi: 10.1534/genetics.115.176107.

5. Benner C, Spencer CC, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493-501 (2016).doi: 10.1093/bioinformatics/btw018.

6. Schwartzentruber, J., Cooper, S., Liu, J.Z. *et al.* Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer’s disease risk genes. *Nat Genet* 53, 392–402 (2021). https://doi.org/10.1038/s41588-020-00776-w

7. Lambert, JC., Ibrahim-Verbaas, C., Harold, D. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* 45, 1452–1458 (2013). https://doi.org/10.1038/ng.2802

8. The 1000 Genomes Project Consortium., Corresponding authors., Auton, A. *et al.* A global reference for human genetic variation. *Nature* 526, 68–74 (2015). https://doi.org/10.1038/nature15393

9. Berisa, T., & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, *32*(2), 283–285 (2016). https://doi.org/10.1093/bioinformatics/btv546

10. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics,* 26(18): 2336.2337 (2010).

11. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Śmigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308-11 (2010).

12. Zhu Y, Paniccia A, Schulick AC, Chen W, Koenig MR, Byers JT, Yao S, Bevers S, Edil BH. Identification of CD112R as a novel checkpoint for human T cells. *J Exp Med*. 213(2):167-76 (2016).doi: 10.1084/jem.20150785.

13. Porcellini, Elisa et al. Alzheimer's disease gene signature says: beware of brain viral infections. *Immunity & ageing : I & A,* 7, 16.(2010), doi:10.1186/1742-4933-7-16.

14. Yashin AI, Fang F, Kovtun M, et al. Hidden heterogeneity in Alzheimer's disease: Insights from genetic association studies and other analyses. *Experimental Gerontology*. 107:148-160 (2018). doi: 10.1016/j.exger .2017.10.020.

15. Hongwon Kim, Junsang Yoo, Jaein Shin, Yujung Chang, Junghyun Jung, Dong-Gyu Jo, Janghwan Kim, Wonhee Jang, Christopher J Lengner, Byung-Soo Kim, Jongpil Kim, Modelling *APOE* ε3/4 allele-associated sporadic Alzheimer’s disease in an induced neuron, *Brain*, 140(8): 2193-2209 (2017), https://doi.org/10.1093 /brain/awx144

16. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*. **33**: 501-4. (2015) doi:10.1093/nar/gki025.

17. Hardy J, Selkoe DJ. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science*. 297(5580):353-6 (2002). doi: 10.1126/science.1072994.

18. Zhu, Xc., Dai, Wz. & Ma, T. Impacts of CR1 genetic variants on cerebrospinal fluid and neuroimaging biomarkers in Alzheimer's disease. *BMC Med Genet* 21, 181 (2020). https://doi.org/10.1186/s12881-020-01114-x.