GENOMICS AND COMPUTATIONAL BIOLOGY Vol. 4 No. 2 (2018): e100045

Research Article

Dynamics of Associations Between Single Nucleotide Polymorphisms in Relation to Alzheimer's Disease Captured with a New Measure of Linkage Disequilibrium

Stanislav J. Sys [1,4], David Fournier[1,4], Illia Horenko[2], Kristina Endres[3], Susanne Gerber[1,*]

[1]Faculty of Biology and Center for Computational Sciences, Johannes Gutenberg University Mainz Staudingerweg 9, 55128 Mainz, Germany
[2]Faculty of Informatics, Università della Svizzera italiana, Lugano, Via Buffi 13, 6900 Lugano, Switzerland
[3]Department of Psychiatry and Psychotherapy, University Medical Center of the Johannes Gutenberg-University Mainz, Untere Zahlbacher Str. 8, 55131 Mainz, Germany
[4]These Authors contributed equally to this work

*Correspondence: sugerber@uni-mainz.de

Received 2017-10-27; Accepted 2018-02-18

ABSTRACT

Genome-Wide-Association-Studies have become a powerful method to link point mutations (e.g. single nucleotide polymorphisms (SNPs)) to a certain phenotype or a disease. However, their power to detect SNPs associated to polygenic diseases such as Alzheimer's Disease (AD) is limited, since they can only infer the pairwise relation of single SNPs to the phenotype and ignore possible effects of various SNP combinations. The common method to probe these possible complex genetic patterns is to compute a measure called linkage disequilibrium (LD). Despite the fact that several predictive patterns found with LD could successfully be applied to medical diagnosis, this measure still holds several drawbacks as for example the difficulty to confirm and replicate experimental results as well as its sensitivity to statistical biases. Here, we present the application of an alternative method, Linkage Probability (LP) for genetic pattern identification that provides the posterior probability of a relation between two categorical data sets and simultaneously considers potential biases from latent variables, such as the recombination rate or the genetic structure of a population. By applying the LP framework to data from the ADSP-Project, we show that changes of linkage patterns between SNPs can be associated to Alzheimer's disease. Common genomic relation measures still fail to extract this link.

KEYWORDS

Alzheimer Disease; GWAS; Linkage Disequilibrium; Linkage Probability

INTRODUCTION

Aging populations are associated with a higher propensity for neurodegenerative diseases such as Alzheimer's disease (AD), the most common form of dementia. In only 1 to 5% of the cases, AD can be significantly traced back to dominantly inherited mutations in the APP and/or PS1/2 gene [1], but in most cases, except of the ApoE Gene [2, 3], AD only shows a weak genetic background while its causes still remain unknown. From all so-far known risk-factors age is usually considered to be the strongest predictor of AD. 30% of today's population develop AD beyond the age of 85 [4]. Furthermore, several risk factors related to lifestyle such as physical fitness, smoking status and body weight have been intensively studied and brought into the context of AD-risk [5, 6]. In terms of genetics, the APOE $\varepsilon 4$ allele has a significantly-higher relative frequency in populations with AD [7] compared to healthy people. Besides, various other single nucleotide polymorphisms located in the genes APOE CLU, PICALM, EPHA1 and FERMT2 have been linked to AD [8–10].

Within the recent decade, Genome-wide association studies (GWAS) [8, 9, 11] became increasingly important in unravelling the genetics of diseases and complex traits by identifying genetic variants and linking them to specific phenotypes via statistical tests. Regardless of scientific advances, this methodology is unable to detect complex patterns such as SNP clusters to explain multifactorial diseases such as AD.

Associations among SNPs are commonly detected using simple metrics such as a linkage disequilibrium (LD) [12, 13]. LD is defined as the non-random pairwise association between alleles at different gene loci [14]. Nevertheless, there are many latent factors that can influence LD: the recombination rate, the genetic structure of the population and the rate of mutations. Since these latent factors are not always known for the data to be analysed, like e.g. the genetic background or kinship of the subjects, LD values are usually biased, thus potentially leading to wrong interpretation of results. Despite these drawbacks, LD is still widely used in high profile studies, aside from other methodologies using frequencies or contingency tables to compute pairwise SNP-SNP relations [15, 16]. In order to take into account these potential latent factors, we applied a new method [17] that computes the linkage probability (LP) of a pairwise relation between two SNPs. The complete framework can be downloaded in form of a MATLAB toolbox

at https://github.com/SusanneGerber/PRP-Measure-MATLAB together with an extensive user manual. This method - based on [17–22] - allows to circumvent the typical biases that might become imposed by the common LD measure.

Due to the polygenic nature of AD, genetic factors underlying the onset of AD cannot be explained by one-dimensional models [23]. Applying instead a measure such as LP to uncover unbiased clusters of variants each associated with AD gives an opportunity to extract and analyse complex underlying networks.

METHODS

Whole-exome sequencing data were obtained from the Alzheimer's Disease Sequencing Project (ADSP) [Study report]. With a permission from NIH, data were retrieved directly from the dbGaP website (dbGaP Study Accession: phs000572.v7.p4). The dataset contains genotypes and phenotype information of 679 patients suffering from late onset AD and of 90 Polymorphisms control subjects.

First, we performed a standard GWAS to extract significant SNPs lying in (or in a close vicinity to) the genes already reported to be associated with AD in previous studies [4, 8, 9, 11]. Afterwards, we divided the available SNPs from the ADSP-dataset into two groups. The first group contains all those SNPs that are part of the aforementioned genes associated to AD (n=1070) whereas the second group contains all the remaining SNPs (n=1390369). This knowledge-driven preselection was performed due to the time-consuming nature of computing multivariate associations. The genetic data were downloaded as VCF files and converted into the PLINK binary format BED. We used the Fisher's exact test to calculate odds ratios and p-values (PLINK 1.9) [14]. The Manhattan plot generated to visualize the output of the GWAS was produced using the "qqman" - package in R 3.2.3.

A mathematical framework to analyse genetic data taking latent variables into account

The LP-measure differs from the standard measures for computing pairwise relations between categorical variables mainly in two ways:

1. LP provides the posterior probability of a relation between two categorical data sets and is able to consider an eventual bias coming from latent variables

2. Since LP provides a direct estimate of the posterior probability of a relation conditioned on the given data it does not introduce an additional model error that results from the mathematical effects such as scaling/transformation-dependence of variables, stationarity assumptions or others.

For further details we refer to the original publication [17].

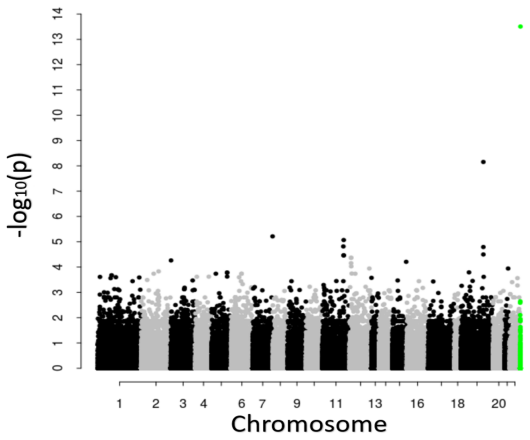

Figure 1: Manhattan Plot of 1.391.439 SNPs from 679 Individuals. p-values were obtained by performing the Fisher's exact test with PLINK 1.9. The vertical axis displays the p-values transformed with the formula -log10(p) and the horizontal axis contains the SNP positions on the genome. The green marked SNPs are SNPs that have been already reported to be associated with AD.

RESULTS

Analysing p-values and computing the Linkage Probability Measure

We calculated the p-values of the Fisher's exact test for all 1391439 SNPs (related or non-related to AD) from 769 individuals and visualized them by means of a Manhattan plot (Figure 1). Out of the whole dataset, only a well-known variant of the APOE gene (rs429358) showed a significant p-value below 5×10^{-12} [15, 16]. (T) is the common allele of rs429358. The APOE $\varepsilon 4$ allele is formed when both rs429358 C and the rs7412 (C) alleles are present. Another interesting candidate we found, called rs11556505, is also known to be associated with AD [16, 17].

We then performed analysis of linkage. One problem of LD is that it is difficult to interpret due to missing objective criteria regarding the relevance of the particular SNP association. Ranking SNP associations is usually carried out by examining the function of associated genes as a possible evidence for significance. As a possibly more objective criterion, we decided to separate the SNPs that are strongly related to AD from others, and study how the linkage patterns change between SNPs from genes associated to AD and genes which have not been reported to have an association to the disease. In our hypothesis, patterns associated to AD should emerge only from SNPs related to AD.

To elaborate on this, we created two samples: considering the runtime for computation, we chose one sample containing the 1000 SNPs which exhibit the strongest association ($p < 0.01$) to AD and a second sample containing 1000 SNPs with no relation to AD using the results from the Fisher's exact test. In order to examine the associations within these two samples, we computed LP and for comparison LD in both cases (Figure 2, Figure 3).

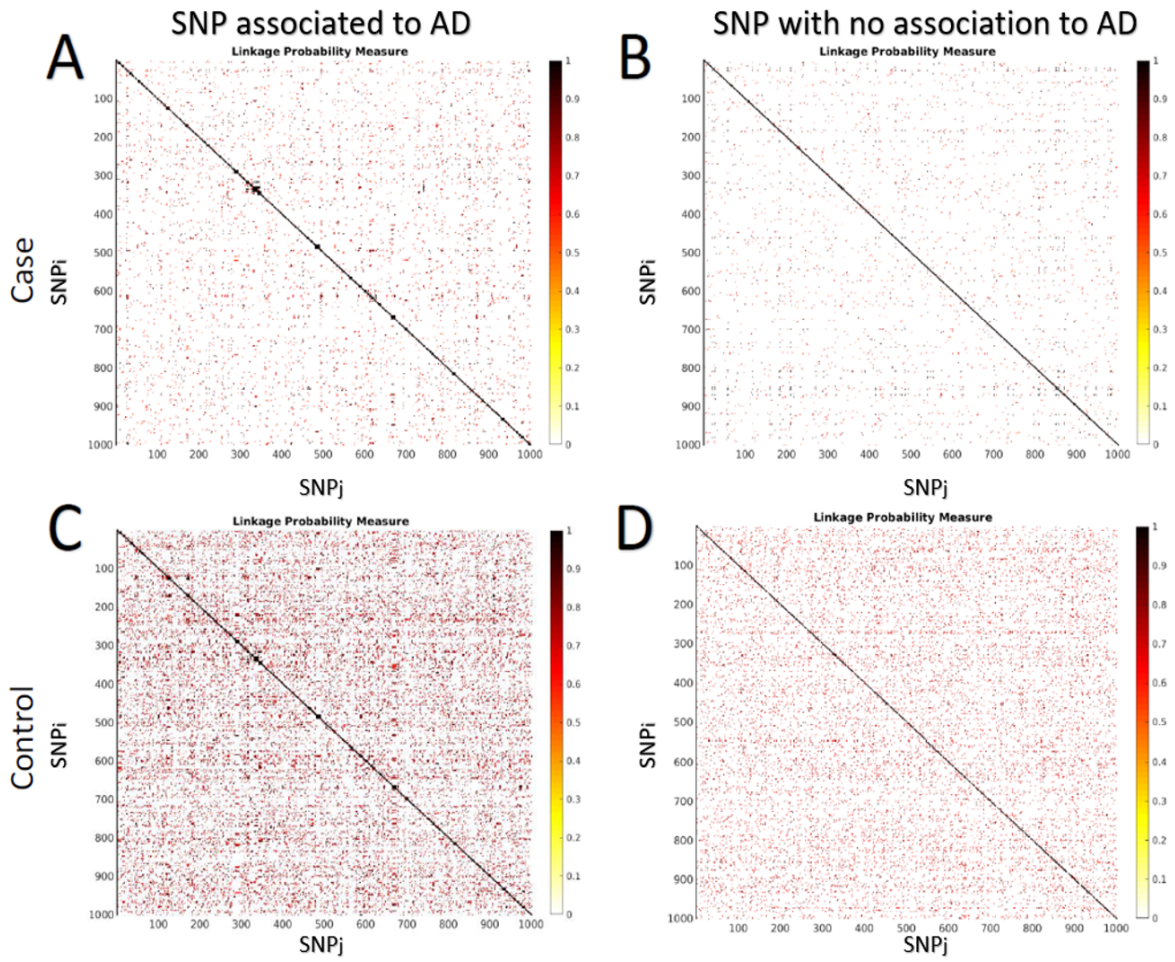

Figure 2: Linkage Probability for 1000 SNPs associated to AD (2A and 2C) and 1000 SNPs which are not associated to AD (2B and 2D). SNPs were selected based on their p-value from the Fisher's exact test. LP was computed for 679 cases (top left/right) and for 90 controls (bottom left/right). A pair is a (i,j) couple with i and j being the indices of the 1000 SNPs. On the map, the first SNP for i (vertical axis) is on the top and the first SNP for j (horizontal axis) is on the left. Values for a SNP with itself correspond to the values in the diagonal.

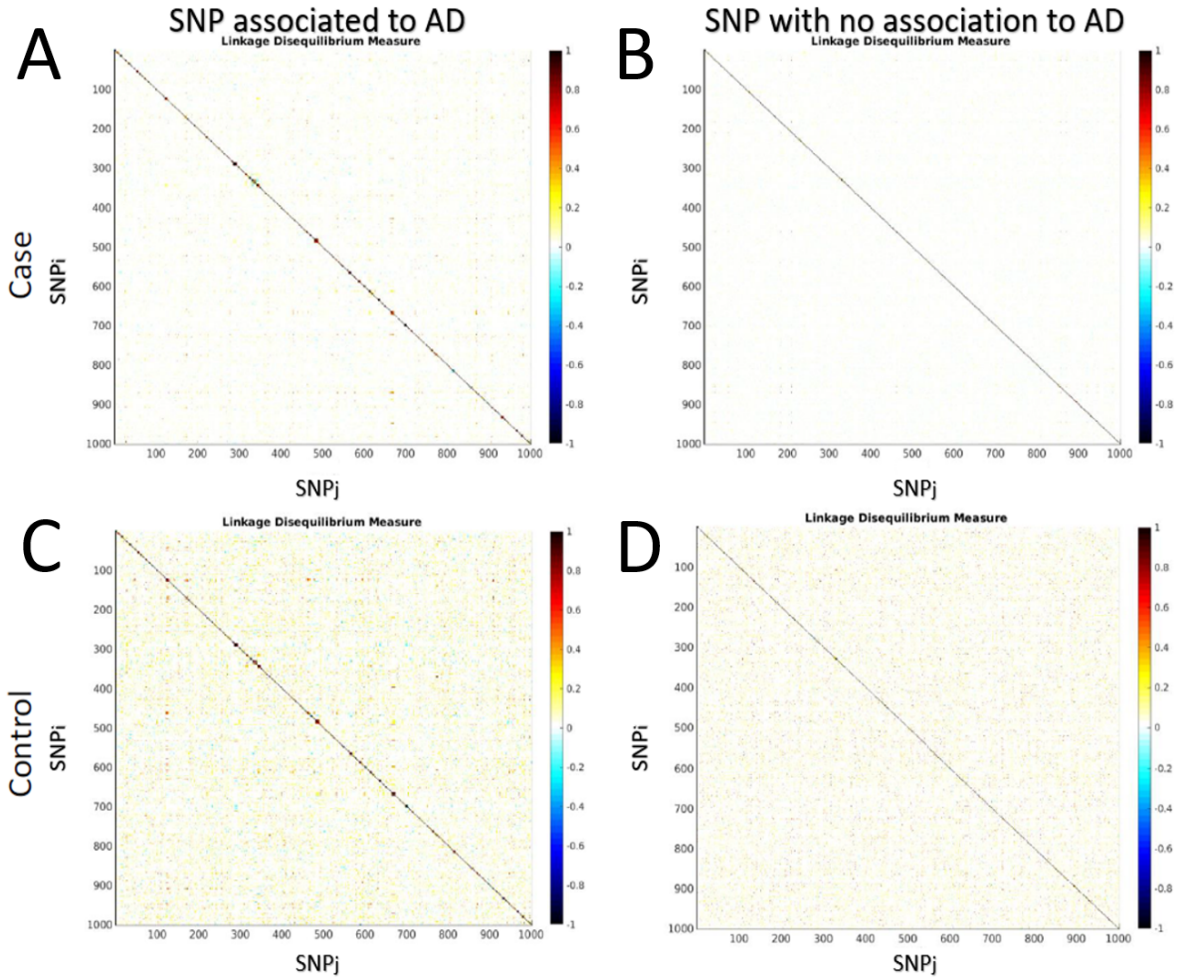

Figure 3: Disequilibrium for 1000 SNPs associated to AD (3A and 3C) and 1000 SNPs which are not associated to AD (3B and 3D). SNPs were chosen by p-value from the Fisher's exact test. LD was computed for 679 cases (top left/right) and for 90 controls (bottom left/right). A pair is a (i,j) couple with i and j being the indices of the 1000 SNPs. On the map, the first SNP for i (vertical axis) is on the top and the first SNP for j (horizontal axis) is on the left. Values for a SNP with itself correspond to the values in the diagonal.

Overall, the patterns of association appear to be similar in LP and LD for SNPs associated to the disease, as well as SNPs which have not been reported to have an association. For both LP and LD, more associations can be detected in the cohort of control subjects than in the cohort of AD-cases. This phenomenon could be partially explained by the differences in sample sizes (679 cases, 90 controls). Since we wanted to keep as much information as possible, we used the whole cohort for calculation. To what extent the sample size affects the outcome can be seen in Figure 5. The SNP patterns hold also true for a sample size of 90 cases. Furthermore, we found more associations between the top associated SNPs in the control group than in AD patients and less associations for SNPs which already had been interpreted as exhibiting "no association" in the Fisher's exact test. This result holds true for both, controls and AD patients. Despite the similar patterns in LD and LP, the Linkage Probability measure tends to find a larger number of SNPs with a high probability of linkage as compared to the LD method (Figure 4) and so is more discriminative.

Monitoring linkage disequilibrium dynamics using the Linkage Probability framework

We hypothesize that differences in linkage patterns between cases and controls could be derived from the fact that there is a certain linkage disequilibrium in healthy people, which might be disrupted by mutations in people with AD - leading to less linkage between specific SNP pairs. Therefore, people with AD will have less associations between any SNP pair typically occurring in the disease compared to healthy people.

Our results suggest such a conjecture. In Figure 2, we can see that the linkage patterns captured by the LP seem to vanish in patients suffering from AD (Figure 2 A&B), whereas healthy people have more complex LP relation patterns (Figure 2 C&D). This applies to SNPs associated with the disease, as well as for SNPs without association. LD is more difficult to capture (Figure 3), since we cannot see such clear trends as in the case of LP. Nevertheless, LD scores seem to show more erratic trends, hinting that high LP scores are better at capturing meaningful disease-related associations than LD (Figure 3).

To conclude, our results show that LP captures more significant pairwise SNP-SNP-relations than LD with a good agreement in terms of intersection for LD values found by LP (see Figure 6), tending to show that these SNPs are more representative. Moreover, we hypothesize that the occurrence of some SNPs associated with AD may disrupt linkage patterns of SNPs in people suffering from AD in comparison to the control group. In order to validate such an hypothesis, one would require more investigations and verifications to prove if our results still hold true with other datasets and larger cohorts. Furthermore, the detected discriminative patterns of pairwise SNP-SNP-relations in cases and controls should be related to biological functionality to confirm biological meaning. As a follow up study, systematically investigate the relation between the pattern of LP relations observed for AD (see Figure 1 and Figure 2) and the steric effects of mutations (e.g., captured by the Hi-C) to probe their possible involvement in disturbing linkage disequilibrium. We also intend to examine whether this finding can be generalized to other types of neurodegenerative diseases.

A subsequent multistage model integrating different layers of data such as transcription, protein structure and interactions between genes or metabolic pathways could provide further functional explanations for the involvement of these genetics patterns in AD [24, 25].

ACKNOWLEDGEMENTS

The authors thank the Center of Computational Sciences in Mainz (CSM) for partly funding the project. The work of SS was funded by the NMFZ (Naturwissenschaftlich-Medizinisches Forschungszentrum) of the University Medical Center of the Johannes Gutenberg-University Mainz. The work of SG and DF was funded by the CSM.

Acknowledgement for Alzheimer Disease Genetic Analysis Data: Biological samples and associated phenotypic data used in primary data analyses were stored at Principal Investigators' institutions, and at the National Cell Repository for Alzheimer's Disease (NCRAD) at Indiana University funded by NIA. Associated Phenotypic Data used in primary and secondary data analyses were provided by Principal Investigators, the NIA funded Alzheimer's Disease Centers (ADCs), and the National Alzheimer's Coordinating Center (NACC) and stored at Principal Investigators' institutions, NCRAD, and at the National Institute on Aging Alzheimer's Disease Data Storage Site (NIAGADS) at the University of Pennsylvania, funded by NIA. Contributors to the Genetic Analysis Data included Principal Investigators on projects that were individually funded by NIA, other NIH institutes, private U.S. organizations, or foreign governmental or nongovernmental organizations.

AUTHOR CONTRIBUTIONS

SS performed the data analysis and wrote the text. DF performed data analysis and interpreted the results. SG designed the study, supervised the project and edited the text. IH developed and provided analytical tools and edited the manuscript. KE edited the manuscript. All authors discussed the results and implications and commented on the manuscript at all stages.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ABBREVIATIONS

AD: Alzheimer's Disease

GWAS: Genome-wide-association-studies

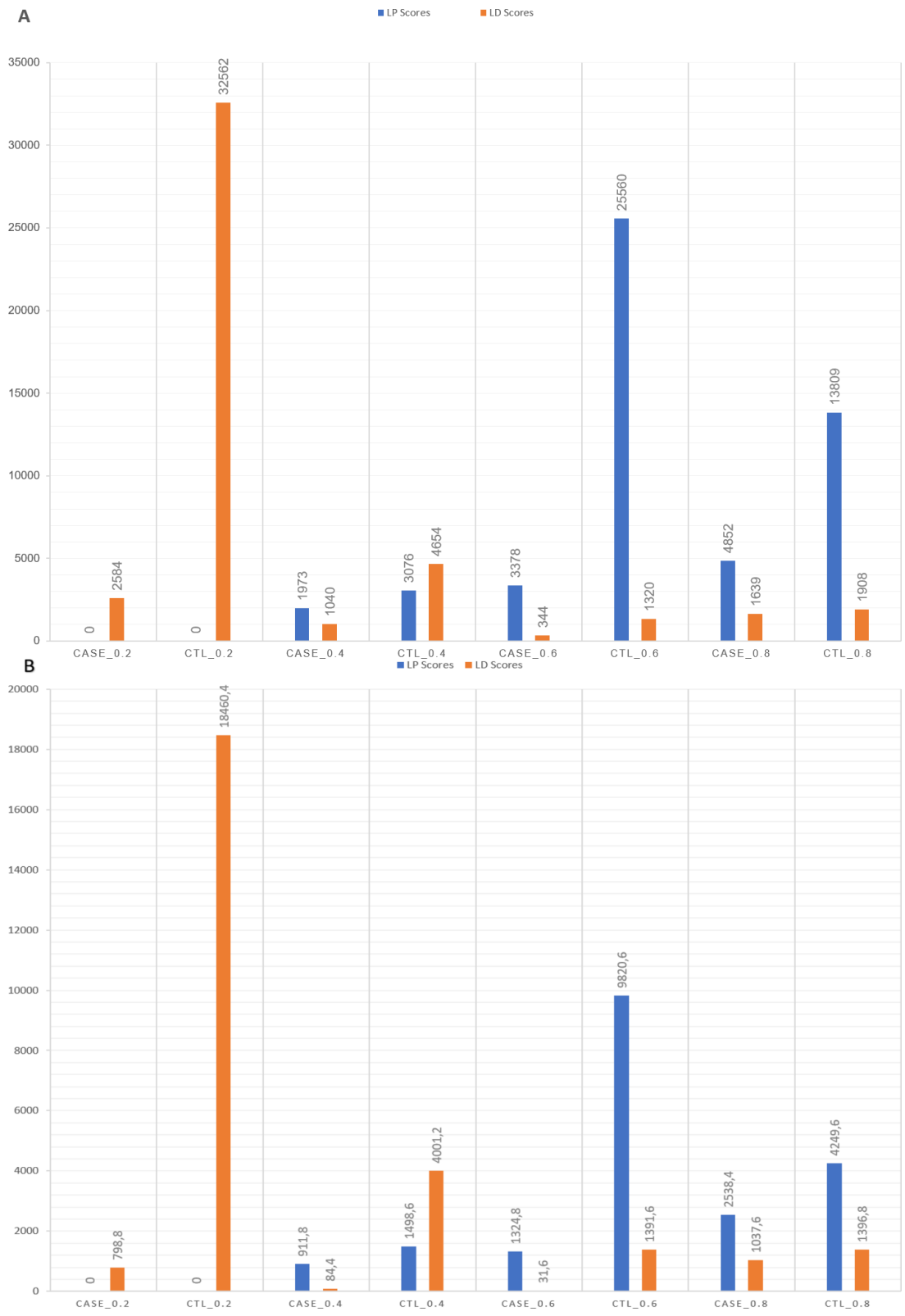

Figure 4: Distribution of LP and LD scores for genes associated to Alzheimer's disease and control patients.. 4A. Distribution of LP and LD for the 1000 SNPs with the strongest association to AD (smallest p-value in the Fisher's exact test). 4B. Distribution of LP and LD scores for the mean of five randomly-picked samples (n= 1000 SNP), which had a p-value of 1 in the Fisher-test, id est showed no association to AD.

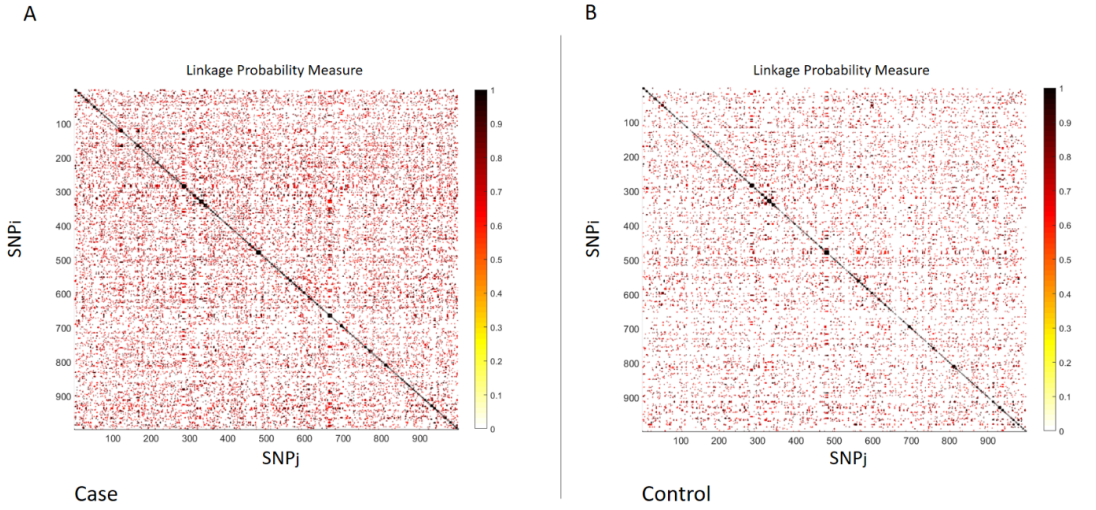

Figure 5: Linkage Probability for 1000 SNPs associated to AD (A) and 1000 SNPs which are not associated to AD (B). SNPs were selected based on their p-value from the exact Fisher test. LP was computed for 90 cases (A) and for 90 controls (B). A pair is a (i,j) couple with i and j being the indices of the 1000 SNPs. On the map, the first SNP for i (vertical axis) is on the top and the first SNP for j (horizontal axis) is on the left. Values for a SNP with itself correspond to the values in the diagonal.

LP: Linkage Probability

LD: Linkage Disequilibrium

ND: neurodegenerative disease

SNP: Single Nucleotide Polymorphism

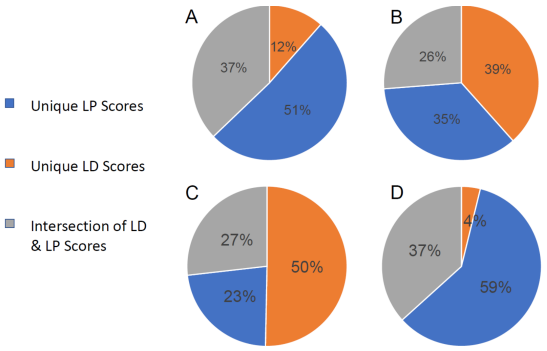

Figure 6: Pie-charts showing the percentage of captured LD & LP association scores and the intersection between them. 6A, 6B. Distribution of association scores for cases in SNPs associated to AD. 6B. Distribution of association scores for controls in SNPs associated to AD. 6C. Distribution of association scores for cases in SNPs not associated to AD. 6D. Distribution of association scores for controls in SNP not associated to AD.

REFERENCES

1. Campion D, Dumanchin C, Hannequin D, Dubois B, Belliard S, Puel M, et al. **Early-onset autosomal dominant Alzheimer disease: prevalence, genetic heterogeneity, and mutation spectrum.** American journal of human genetics. 1999;65(3):664–70. doi:10.1086/302553.
2. Tanzi RE. **The genetics of Alzheimer disease (7)**. Cold Spring Harbor perspectives in medicine. 2012;2(10):a006296–. doi:10.1101/cshperspect.a006296.
3. Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE. **Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database**. Nature Genetics. 2007;39(1):17–23. doi:10.1038/ng1934.
4. Chia-Chen L, Takahisa K, Huaxi X, Guojun B. **Apolipoprotein E and Alzheimer disease: risk, mechanisms, and therapy**. Nature reviews Neurology. 2013;9(2):106–118. doi:10.1038/nrneurol.2012.263.
5. Munoz DG, Feldman H. **Causes of Alzheimer's disease**. Cmaj. 2000;162(1):65–72. doi:10.1108/09526860010336984.
6. Barnes D, Yaffe K. **The Projected Impact of Risk Factor Reduction on Alzheimer's Disease Prevalence**. Lancet Neurology. 2013;10(9):819–828. doi:10.1016/S1474-4422(11)70072-2.
7. Bird TD. **Genetic Aspects of Alzheimer Disease**. Health Care. 2009;10(4):231–239. doi:10.1097/GIM.0b013e31816b64dc.
8. Harold D, Abraham R, Hollingworth P, Sims R, Hamshere M, Pahwa JS, et al. **Genome-Wide Association Study Identifies Variants at CLU and PICALM Associated with Alzheimer's Disease, and Shows Evidence for Additional Susceptibility Genes**. Nature Genetics. 2009;41(10):1088–1093. doi:10.1038/ng.440.
9. Lambert Jc, Heath S, Even G, Campion D, Sleegers K, Hiltunen M, et al. **Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer ' s disease**. Nature Genetics. 2009;41(10):1094–1099. doi:10.1038/ng.439.
10. Naj AC, Jun G, Beecham GW, Wang Ls, Narayan B, Buros J, et al. **Common variants in MS4A4/MS4A6E, CD2uAP, CD33, and**

EPHA1 are associated with late-onset Alzheimer's disease Adam. Nature genetics. 2011;43(5):436–441. doi:10.1038/ng.801.

11. Nettiksimmons J, Tranah G, Evans DS, Yokoyama JS, Yaffe K. **Gene-based aggregate SNP associations between candidate AD genes and cognitive decline**. Age. 2016;38(2). doi:10.1007/s11357-016-9885-2.
12. Squitti R, Polimanti R, Bucossi S, Ventriglia M, Mariani S, Manfellotto D, et al. **Linkage Disequilibrium and Haplotype Analysis of the ATP7B Gene in Alzheimer's Disease**. Rejuvenation Research. 2013;16(1):3–10. doi:10.1089/rej.2012.1357.
13. Hofmann-Apitius M, Ball G, Gebel S, Bagewadi S, De Bono B, Schneider R, et al. **Bioinformatics mining and modeling methods for the identification of disease mechanisms in neurodegenerative disorders**. International Journal of Molecular Sciences. 2015;16(12):29179–29206. doi:10.3390/ijms161226148.
14. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. **A comprehensive review of genetic association studies.** Genetics in medicine : official journal of the American College of Medical Genetics. 2002;4(2):45–61. doi:10.1097/00125817-200203000-00002.
15. Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NLS, et al. **BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies**. American Journal of Human Genetics. 2010;87(3):325–340. doi:10.1016/j.ajhg.2010.07.021.
16. Herold C, Steffens M, Brockschmidt FF, Baur MP, Becker T. **INTERSNP: Genome-wide interaction analysis guided by a priori information**. Bioinformatics. 2009;25(24):3275–3281. doi:10.1093/bioinformatics/btp596.
17. Gerber S, Fournier D, Hewel C, Horenko I. **Imputation of posterior linkage probability relations reveals a significant influence of structural 3D constraints on linkage disequilibrium**. bioRxiv. 2018 jan;doi:10.1101/255315.
18. Horenko I. **On Robust Estimation of Low-Frequency Variability Trends in Discrete Markovian Sequences of Atmospheric Circulation Patterns**. Journal of the Atmospheric Sciences. 2009;66(7):2059–2072. doi:10.1175/2008JAS2959.1.
19. Horenko I, Dolaptchiev S, Eliseev A, Mokhov I, Klein R. **Metastable decomposition of high-dimensional meteorological data with gaps**. J of Atmos Sci. 2008;65(11):3479–3496. doi:10.1175/2008JAS2754.1.
20. Meerbach E, Dittmer E, Horenko I, Schütte C. **Multiscale modelling in molecular dynamics: Biomolecular conformations as metastable States**. Lecture Notes in Physics. 2006;703(Fzt 86):495–517. doi:10.1007/3-540-35273-2_14.
21. Gerber S, Horenko I. **Toward a direct and scalable identification of reduced models for categorical processes**. Proceedings of the National Academy of Sciences. 2017;114(19):4863–4868. doi:10.1073/pnas.1612619114.
22. Gerber S, Horenko I. **On inference of causality for discrete state models in a multiscale context**. Proceedings of the National Academy of Sciences. 2014;111(41):14651–14656. doi:10.1073/pnas.1410404111.
23. Rosenthal SL, Kamboh MI. **Late-Onset Alzheimer ' s Disease Genes and the Potentially Implicated Pathways**. 2014;p. 85–101. doi:10.1007/s40142-014-0034-x.
24. Huang S, Chaudhary K, Garmire LX. **More is better: Recent progress in multi-omics data integration methods**. Frontiers in Genetics. 2017;8(JUN):1–12. doi:10.3389/fgene.2017.00084.
25. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. **Methods of integrating data to uncover genotype–phenotype interactions**. Nature Reviews Genetics. 2015;16(2):85–97. doi:10.1038/nrg3868.