ADAP Proposal
Seth D Temple
May 5, 2022

Introduction

I plan to conduct a genetic data analysis for the Alzheimer's Disease Sequencing Project (ADSP) group at University of Washington. ADSP is a broad initiative from the National Institute on Aging, an institute under NIH, with overarching goals to identify genomic variants involved in risk and/or prevention for Alzheimer's disease (AD) and other dementias. The statistical methods I intend to apply make use of long pairwise segments of DNA sharing (the result of recent shared common ancestry) to **associate any excess haplotype sharing with a phenotype case/control status**. This approach balances some of the benefits and limitations of targeted family-based studies and population genome wide association studies (GWAS). Family-based studies leverage close relatedness to detect associated low frequency variants that have moderate to large effect sizes (heritability), but are constrained by small sample sizes. GWAS uses large sample sizes to detect associated common variants of small effect sizes and requires a burdensome multiple testing correction. Identity-by-descent (IBD) mapping can gain more power to detect rare variants of large effects with a larger sample size and a less extreme genome-wide significance threshold (Browning and Thompson 2012). This study design may be especially promising for populations that experienced recent bottleneck and founding events (Acosta-Uribe et al. 2022), which is the case for populations involved in the migratory event from Europe and West Africa to the Americas.

Project Details

This enumerated list includes questions and answers concerning the details of the proposed applied data analysis project.

1) How did I come across this project?
    a) I learned about this project as an NIH Trainee in Statistical Genetics. I acted as a research assistant for Biostatistics Professor Timothy Thornton in Autumn 2020 and Winter 2021. I built out a bioinformatics pipeline to infer IBD segments from whole genome sequence data in the ADSP 5k dataset. Moreover, I developed data quality checks using pedigree analysis and relatedness inference to assess the performance of the IBD segment detection. In Spring 2021, I delayed the proposed analysis in waiting for the release of a follow-up study 17k dataset with more multiethnic genotyped individuals.

2) Who is the client for this project?
    a) The client is six analysis working groups (UW, UT - Houston, UT - San Antonio, Baylor, Boston, and Columbia) that are supported by an U01 grant from NIH. I will report directly to the UW group headed by Professor Ellen Wijsman. The UW group meets once monthly and consists of geneticists, epidemiologists, and engineers. The proposed research project is of interest to the group. (Ellen Wijsman is a Professor of Biostatistics, Medical Genetics, and Genome Sciences, but would only act in a supervisory role and

would not drive any analysis decisions. Timothy Thornton is not as involved with the working group since joining Regeneron in Atlanta.)

3) What scientific question will be addressed with this project?
    a) The scientific goal is to **identify new genes associated with AD** and other dementias. Previous research has implicated novel genetic loci showing excess IBD sharing in founder populations with Parkinson's neurodegenerative disease (Vacic et al. 2014; Gusev et al. 2011). Follow up analyses to the initial genome scan could fine map variants to motivate lab experiments into pharmacological and environmental interventions.

4) What data do you have access to?
    a) I have access to **nearly 17,000 whole genome sequences** from individuals of European, African, Hispanic, and indigenous ancestries. These individuals have been ascertained for **case and control status with AD** and other dementias. I can easily extend my existing bioinformatics pipeline for the 5k subset to this larger database. This is a unique opportunity to study neurodegenerative diseases in **historically underrepresented populations** using a heretofore unprecedented sample size. (I do not intend to conduct simulation studies as these could be challenging to design and implement; moreover, access to the ADSP data resource is for applied research, not methodological research. The paper by (Browning and Thompson 2012) describes a simulation study for IBD mapping.)

5) How do you plan to conduct the statistical analysis?
    a) Use principal component analysis (PCA) or ADMIXTURE (Alexander et al. 2009) to assign individuals to ancestry population groups for separate analyses.
    b) Detect IBD segments in the 17k WGS dataset
    c) IBD mapping in ancestry population groups
        i) Develop a genome-wide significant threshold for IBD mapping, similar to (Grinde et al. 2019)
        ii) Write scripts to conduct a case-control association test with pairwise IBD (Browning and Thompson 2012)
    d) Describe any results with data visualization
        i) Investigate IBD patterns between families and within families
        ii) Compare any associations to the current literature
        iii) Superimpose local ancestry over IBD segments: for admixed individuals, are the associations on the background of a given ancestry? (Local ancestry inference is done by an engineer in the group, not me.)
    e) (Optional) Burden tests
        i) Verify that these hits are due to a burden of rare variants (Wu et al. 2011)

6) To what extent do I have freedom to make decisions concerning statistical analysis?
    a) Ellen Wijsman has assured me that I have the freedom to independently drive the statistical analysis. I will report my findings to her in monthly working group meetings. I

will not be supported by a research assistantship; I am financially supported by an NDSEG fellowship.

7) **To what extent do I have freedom to make decisions concerning project scope?**
    a) Ellen Wijsman has assured me that I have the freedom to define the project scope as I see fit and to limit the work to Summer and Autumn 2022. Two ways in which I can limit the scope is (1) to study only a couple ancestry population groups (Europeans, Africans or Hispanics) and (2) to study only some chromosomal regions. Step 5b) is nearly complete from my previous traineeship. Steps 5a,c-d) are the focus of my ADAP. I can limit scope by not doing Step 5e).

8) To what extent does this project relate to my dissertation projects?
    a) My dissertation projects focus on developing statistical methods to infer natural selection from genetic data. This is a separate methodological framework from the proposed genome-wide association testing. My research advisor Sharon Browning will not be involved in this project. Ellen Wijsman and Timothy Thornton will not be on my dissertation committee.

9) **Follow-up** (May 11, 2022): can you identify an epidemiologist, geneticist, engineer, or MD that can be the pro-forma client for this project (someone from among the six working groups that is not a PhD Biostatistician to whom you would give the report)?
    a) Elizabeth Blue is a medical geneticist at UW who will see my reports and presentations.

Conclusion

I am excited about this project as it will (1) give me experience conducting genetic data analyses for multiethnic, admixed individuals, and it will (2) improve and extend my methodological skill set to genome-wide association testing and study design considerations. While this project does involve considerable background experience, I have gained this educational training as a statistical genetics trainee and the instructor of record for BIOST 550. I hope with my skill set and access to this unique genetic database that I can contribute to the expanding research in the neuropathology of underrepresented populations.

References

Acosta-Uribe, Juliana, David Aguillón, J. Nicholas Cochran, Margarita Giraldo, Lucía Madrigal, Bradley W. Killingsworth, Rijul Singhal, et al. 2022. “A Neurodegenerative Disease Landscape of Rare Mutations in Colombia due to Founder Effects.” *Genome Medicine* 14 (1): 27.

Alexander, David H., John Novembre, and Kenneth Lange. 2009. “Fast Model-Based Estimation of Ancestry in Unrelated Individuals.” *Genome Research* 19 (9): 1655–64.

Browning, Sharon R., and Elizabeth A. Thompson. 2012. “Detecting Rare Variant Associations by Identity-by-Descent Mapping in Case-Control Studies.” *Genetics* 190 (4): 1521–31.

Grinde, Kelsey E., Lisa A. Brown, Alexander P. Reiner, Timothy A. Thornton, and Sharon R. Browning. 2019. “Genome-Wide Significance Thresholds for Admixture Mapping Studies.” *American Journal of Human Genetics* 104 (3): 454–65.

Gusev, Alexander, Eimear E. Kenny, Jennifer K. Lowe, Jaqueline Salit, Richa Saxena, Sekar Kathiresan, David M. Altshuler, Jeffrey M. Friedman, Jan L. Breslow, and Itsik Pe’er. 2011. “DASH: A Method for Identical-by-Descent Haplotype Mapping Uncovers Association with Recent Variation.” *American Journal of Human Genetics* 88 (6): 706–17.

Vacic, Vladimir, Laurie J. Ozelius, Lorraine N. Clark, Anat Bar-Shira, Mali Gana-Weisz, Tanya Gurevich, Alexander Gusev, et al. 2014. “Genome-Wide Mapping of IBD Segments in an Ashkenazi PD Cohort Identifies Associated Haplotypes.” *Human Molecular Genetics* 23 (17): 4693–4702.

Wu, Michael C., Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. 2011. “Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test.” *American Journal of Human Genetics* 89 (1): 82–93.