Senior Thesis Proposal

Trang Dang

1 Student Information

Trang Dang - thdang@brynmawr.edu

2 Summary

Here, we hope to adjust Bonsai[2] Background IBD finding algorithm that is originally designed for general populations to accomodate endogamous populations with higher levels of in-group marrianges.

3 Problem Statement

Understanding inheritance in families is the first step towards understanding how rare genetics disorders can propagate within a family. A lot of these studies rely on Identical-By-Descent segments (IBD), which are segments of the DNA that descendants inherit from their ancestors.

However, while there are multiple IBD-finding approaches, these approaches are designed for general populations, and, therefore, cannot handle endogamous populations, which are populations with lots of in-group marriages and little external admixture. For general populations with lots of genetic admixtures, usually, long matching sequences can only be inherited, and can be confidently declared IBD. However, as member of endogamous populations tend to marry within their communities, their descendants have many matches in their genetic sequences. Hence, when working with these populations, algorithms designed for general population tend to overrestimate the amount of IBD by confusing Identical-By-Descent with Identical-By-State segments, which are segments that people share because they are popular in the population.

The Amish population, located in Lancaster, Pennsylvania, is an endogamous population with rates of Bipolar Disorder and Mood Disorder that are significantly higher than the US averages. We hope to explore how we can adapt Bonsai[2], an algorithm that removes false positive IBD for general populations, for this Amish population. Hopefully, this can further contribute to understanding how traits are inherited in this community.

4 Proposed Solution

We will use relationships from pedigrees to identify the true IBD among the mix of IBD and IBS that algorithms designed for general population output. We will be working with a pedigree of 1338 individuals from the Amish population in Lancaster, PA. In this pedigree, 394 out of the 1338 individuals are genotyped[1]. We have an approach, Bonsai[2], that takes in two ancestors, their descentdants, IBDs, computes the probabilities, and decides to accept or reject the IBDs. We hope to first adjust some of the graph traversal in Bonsai so that this algorithm can handle multiple paths. Then, we hope

5 Evaluation Plan

We have a complete pedigree of 1338 Amish individuals, and 394 genotypes. We plan to use `ped-sim` to simulate IBDs from the genotypes and the pedigree. Then, we will use these simulations as the ground truth to evaluate the IBDs that we've idenfied.

// what is the baseline for our evaluation?

6 Potential Challenges

The first challenge for this project is adopting our probability calculations for Bonsai. We believe that if it's too challenging, we can set this task aside. The second challenge is the implementation: it can be difficult working with graphs with cycles and many individuals.

7 Team Bios

References

[1] Kelly Finke et al. "Ancestral haplotype reconstruction in endogamous populations using identity-by-descent". In: *PLOS Computational Biology* 17.2 (Feb. 2021). Ed. by Degui Zhi, e1008638. DOI: 10.1371/journal.pcbi.1008638. URL: https://doi.org/10.1371/journal.pcbi.1008638.

[2] Ethan M Jewett et al. "Bonsai: An efficient method for inferring large human pedigrees from genotype data". In: *The American Journal of Human Genetics* 108.11 (2021), pp. 2052–2070.