Senior Thesis Proposal

Trang Dang

Advisors My advisors will be Professor Sara Mathieson and Professor Dianna Xu.

Background 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.

Problem Statement For general populations with lots of genetic admixtures, usually, long matching sequences can only be inherited, and can be confidently declared IBD. However, member of endogamous populations, populations that have lots of in-group marriages and little external admixture, have many matches in their genetic sequences. Hence, when working with these populations, algorithms designed for general population tend to confuse Identical-By-Descent with Identical-By-State segments, which are segments that people share because it's popular in the population, and overrestimate the amount of IBD.

Our Contribution 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[3], that takes in two ancestors, their descentdants, IBDs, computes the probabilities, and decides to accept or reject the IBDs. We hope to develop an algorithm that can select the ancestors and descentdants that will result in the most accurate Bonsai results.

Works We Built Upon First, we will preprocess the genotype with `phasedibd`, an algorithm that identifies IBD for general populations[2]. Then, we will use Bonsai, an algorithm that computes the likelihood that an observed IBD is an IBD and not an IBS. We will modify Bonsai's probabilities calculation using methods from our previous paper, `thread`, such that it accounts for how IBDs at different locations in the sequence are more or less likely inherited together[1]. Finally, we will design our approach that selects the ancestors to input into the pedigree.

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.

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] William A Freyman et al. "Fast and robust identity-by-descent inference with the templated positional burrows–wheeler transform". In: *Molecular Biology and Evolution* 38.5 (2021), pp. 2131–2151.

[3] 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.