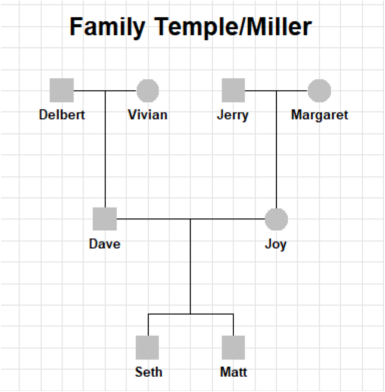
STAT 499
Winter 2022

Final Project

Given a pedigree (family tree), we can compute kinship, which is one half the expected proportion of the genome that relatives share identical by descent. (That is, kinship values range from 0 to 1/2.) These computations come from the path-counting formula of American population geneticist Sewall Wright. In this project, we will learn about the PED file format to encode family trees in data, and we will write some code to automate these path-counting calculations. The objectives are to (1) gain more practice in programming and to (2) learn through experience the challenges associated with computing statistics on trees and graphs. Below are small exercises to guide the mentee through the project. The mentor will be available to help over email and in-person for 2 meetings.

PED file format

1. Read about the PED file format on this page: PLINK.
2. What the six required columns? Note that columns are separated by tabs.
3. Open the stat550less.ped file in a text editor.
4. Open the temple.ped file in a text editor.
5. Download the HaploPainter software from Source Forge (link). Unzip (Windows) the folder. Double-click the HaploPainter application in the File Explorer (Windows) to run the program.
6. In HaploPainter application, File → Import Pedigree → Linkage → Select a PED file.
7. Open a text editor. Make your own family PED file. Include siblings, parents, maternal and paternal grandparents, and one set of uncle/aunt + first cousins. Visualize the family tree in the HaploPainter application.



Final Project

Programming

We want to write custom functions to implement the path-counting formula:

$$\sum_{A \in \mathcal{A}} \sum_{P \in \mathcal{P}(A)} (1 + f_A)(1/2)^{m(P)+1}$$

where A is an ancestor in the ancestors set $\mathcal{A}$, P is a path in the paths set through A, f_A is the inbreeding value for A, and $m(\cdot)$ is a function that counts the meioses in the path. For simplicity, we assume that $\mathcal{P}(A)$ is always a singleton set (there is only 1 path through each common ancestor). Implement the following functions. (Pseudocode reads like Python.)

1. find_ancestors(individual)
   - find_ancestors('Seth') = ['Joy','Dave','Margaret','Jerry','Vivian','Delbert','Seth']
   - find_ancestors('Joy') = ['Margaret','Jerry','Joy']
2. count_meioses(individual, ancestor)
   - count_meioses('Seth','Joy') = 1
   - count_meioses('Seth','Margaret') = 2
   - count_meioses('Seth','Adam') = print('Not related')
3. common_ancestors(individual1, individual2)
   - common_ancestors('Seth','Matt') = ['Joy','Dave']
4. path_formula(mcounts, fvalues)
   - path_formulas([2,2],[0,0]) = $(1+0)(1/2)^{2+1} + (1+0)(1/2)^{2+1}$

These are some of the main functions required. You can design more user-friendly functions that are derivatives of these core functions. Use the functions you write to compute kinship values for the given pedigrees. Calculate kinship values by hand and verify your program outputs. Some challenges you may face:

- Bringing the PED file in
- Systematically traversing the tree upwards so as to not miss common ancestors
- Removing common ancestors who are ancestors of another common ancestor (e.g. Seth and Matt have the maternal and paternal grandparents as common ancestors, but we only use the parents in the formula)
- One approach may be to use object-oriented programming (link)

Final Project

Publishing (Optional)

1. Make a GitHub profile (link). (Download GitHub Desktop?)

2. Review materials from Bryan D Martin regarding R and package development.

   - Introduction to R (link)
   - Fundamentals of R (link)
   - Software Development in R (link)

3. Publish your work as an R package on GitHub.

Alternative Project

Depending on the mentee's prior programming experience, the programming project may be involved. An alternative project is to (1) calculate inbreeding and kinship values by hand and (2) determine how these impact the prevalence of recessive conditions. These exercises are based on the following pedigree. (For the alternative project, still complete the PED file format exercises.)

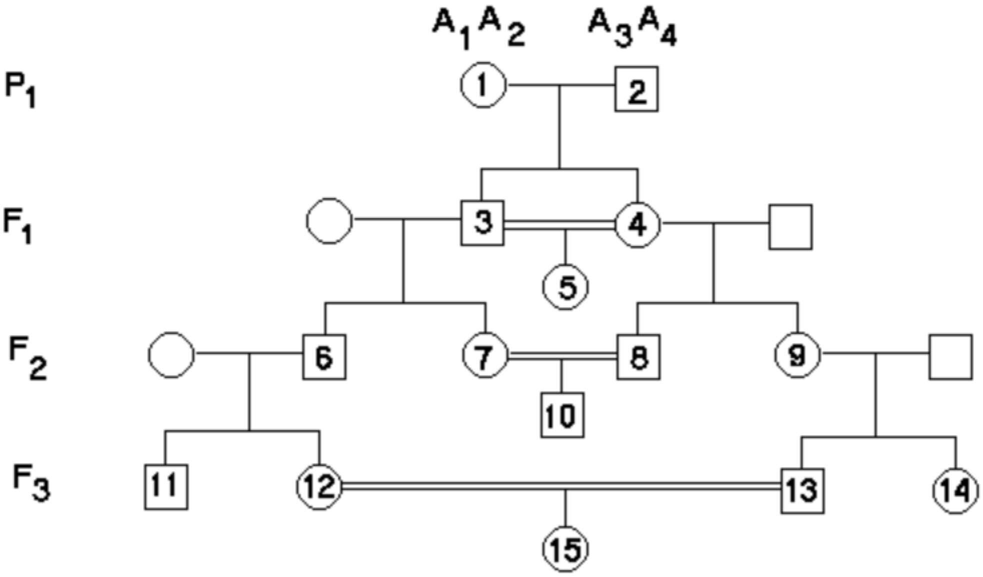


Figure 1: Family tree with three instances of inbreeding

Final Project

1. First level

   (a) Compute the kinship value between 3 and 4.

   (b) Compute the inbreeding value for 5.

   (c) Individual 3 has a recessive allele a. The recessive allele is present in the population at 1% frequency. Calculate the probability individual 5 has the homozygous recessive genotype.

   (d) Calculate the probability individual 6 has the homozygous recessive genotype.

2. Second level

   (a) Compute the kinship value between 7 and 8.

   (b) Compute the inbreeding value for 10.

   (c) Individual 7 has a recessive allele a. The recessive allele is present in the population at 1% frequency. Calculate the probability individual 10 has the homozygous recessive genotype.

   (d) Calculate the probability individual 6 has the homozygous recessive genotype.

3. Third level

   (a) Compute the kinship value between 12 and 13.

   (b) Compute the inbreeding value for 15.

   (c) Individual 12 has a recessive allele a. The recessive allele is present in the population at 1% frequency. Calculate the probability individual 15 has the homozygous recessive genotype.

   (d) Calculate the probability individual 11 has the homozygous recessive genotype.