Introduction

PONDEROSA is an algorithm designed to assist in pedigree construction. PONDEROSA is highly sensitive to phase quality and will see reduced performance in datasets with any of the following: admixture, sparse pedigrees (especially sparse parent-offspring data), samples from different populations, etc. While we have found PONDEROSA to work best in endogamous populations, PONDEROSA can work well in any population as long as high phase quality can be achieved. However, running PONDEROSA with a dataset with suboptimal phase quality should only affect avuncular/grandparent-grandchild vs. half-sibling distinction and may still be useful for datasets with few half-siblings or datasets with a narrow age range (which may be unlikely to have avuncular or grandparent-grandchildren pairs).

Please note that PONDEROSA is designed to *assist* pedigree construction and further steps by the user are required to construct the pedigree. We hope to change this in future versions of PONDEROSA. For now, PONDEROSA largely infers relationships in a vacuum (i.e. without considering the context of the pedigree). Relationship inference should be double-checked against the existing pedigree structure.

Requirements

python3.6
scikit-learn and its dependencies (install here: https://scikit-learn.org/stable/install.html)

Input

Required arguments

--king KING .seg file (or any .seg-formatted IBD file).

--map PLINK-formatted .map file. The chromosome number should be replaced with "%s". This .map file must be the same .map file used to generate IBD segments.

--fam PLINK-formatted .fam file. All PO present in the KING file must be present in the .fam file. If age data is unavailable/unreliable and the parent/offspring cannot be distinguished in the pair, PONDEROSA can be run and the haplotype scores of the individuals can be used to make the distinction (and then run again).

--match GERMLINE-formatted match file where the chromosome number is replaced with "%s". If GERMLINE file, must be generated with GERMLINE's **--haploid** flag (we suggest GERMLINE v1.5.3). iLASH .match files can also be used, but PONDEROSA's **--ilash** flag must be used.

Optional arguments

--out Output file prefix. *Default: "PONDEROSA"*

--ilash For use if .match file is in iLASH format.

--haps If PONDEROSA has already been run, supplying the haplotype score file will skip the haplotype score calculation step.

--age Age file where the first column corresponds to the individual ID and the second column corresponds to the age. Note that not all individuals need an age.

--gp_age Minimum age-gap for a grandparent-grandchild pair. Note that if you do not want PONDEROSA to consider age, use 0 for this flag. *Default: 30*

--mhs_age Maximum age-gap for maternal half-siblings. If you do not want PONDEROSA to consider age here, use an arbitrarily large age gap (e.g. 100) *Default: 30*

--ped PLINK-formatted .ped file used by PONDEROSA to stitch IBD segments together. If no .ped file is supplied, PONDEROSA stitches together two segments that are within 1 cM of each other. If .ped file is supplied, PONDEROSA only stitches two segments that are within 1 cM of each other *and* have, at most, one discordant homozygote.

Example
python3.6 PONDEROSA.py --map Himba_chr%s.map --ped Himba_chr%s.ped --king king.seg –fam Himba.fam --match Himba_chr%s.match --out Himba

Output

PONDEROSA outputs five files.

.pedigree file outputs every pairwise relationship found and the confidence. For inferred 2nd degree pairs, the confidence reported is P(2nd degree relationship | 2nd degree). Note that for each of these pairs, the P(2nd degree) > 0.80. The confidence reported for these 2nd degree relatives is a posterior probability, calculated after considering existing pedigree structure and age discrepancies; for the probabilities of each relationship class before this step (i.e. the probabilities calculated on the haplotype scores and number of IBD segments alone), refer to the .score file. Each line corresponds to a pair with the following fields:

1. Individual ID 1 (child in PO pair, grandchild in GP pair, niece/nephew in AV pair)
2. Individual ID 2 (parent in PO pair, grandparent in GP pair, aunt/uncle in AV pair)
3. Pedigree relationship
4. Confidence

.log file provides information about runtime, number of training relationships, and any warnings about inferred relationships. PONDEROSA assumes that all reported parent-offspring are true, but will output warnings when purported PO have IBD1 < 0.80. PONDEROSA will also report errors when FS or 2nd degree pairs have IBD1 < 0.30. Typically, these pairs have low IBD1 values because of an error with the reported PO. These should be double-checked, the .fam file should be updated, and PONDEROSA rerun.

.training file gives the haplotype scores, haplotype score ratios, and the number of IBD segments for each known second degree pair. This file is useful for assessing whether or not the phase quality of the dataset is high (e.g. by plotting). Each line corresponds to a pair with the following fields:

1. Individual ID 1 (child in PO pair, grandchild in GP pair, niece/nephew in AV pair)
2. IID1's haplotype score
3. Individual ID 2 (parent in PO pair, grandparent in GP pair, aunt/uncle in AV pair)
4. IID2's haplotype score
5. Haplotype score ratio
6. Number of IBD segments
7. Relationship

.score file provides information about each putative second degree relative. Each line corresponds to a pair with the following fields:

1. Individual ID 1 (putative grandchild/niece/nephew)
2. IID1's haplotype score
3. Individual ID 2 (putative grandparent/aunt/uncle)
4. IID2's haplotype score
5. Haplotype score ratio
6. Number of IBD segments
7. P(PHS)
8. P(MHS)
9. P(GP)
10. P(AV)

Pipeline

Scripts for the pipeline and sample data are available on Github.