Manual for 'find_pairs.f90'

A fortran program to find sample pairs belonging to the same individual, or to close relatives

Jisca Huisman

2024-01-04

Description

This program finds sample pairs that are likely to come from the same individual (duplicates), from a parent-offspring pair, or from other types of relatives. It does so by calculating for each pair the probabilities that they belong to Self (duplicate), Parent–offspring, Full Siblings, second degree relatives, third degree relatives, or unrelated. The probabilities sum to unity across these six categories, allowing for filtering using an easy to interpret threshold.

In addition, or alternatively, potential duplicates can be filtered based on the number of SNPs at which the genotypes differ, and potential parent-offspring pairs based on the number of SNPs at which they are opposing homozygotes.

The relationship probabilities are calculated for each sample pair independently, assuming all SNPs are independent, and not using any non-genetic information. Each sample pair is presumed to belong to exactly one category; consequently, for example a double grandparent–grand-offspring pair will have the highest probability to be full siblings. For more sophisticated identification of relative pairs and pedigree reconstruction, see `sequoia` (https://jiscah.github.io/).

Strengths & limitations

- [+] limited memory use; can be used in very large datasets with SNP genotypes of many tens of thousands of samples
- [+] very fast (e.g. 3.5 minutes to go through 60 thousand samples)
- [-] returns many false positives among the possible parent-offspring pairs, especially when not filtering on relationship probabilities
- [-] does not use birth year or sex information: no information on whether the second sample in a parent-offspring pair is the dam, sire, or an offspring of the first sample.

Input files

- `--geno` : genotype file, with 1 ID column followed by 1 column per SNP, no header row, 1 row per sample, and SNPs coded as 0/1/2 with a negative number for missing values.
- `--only` : individual subset, pairs where both individuals are not in the subset are ignored. Text file with a single column with IDs, no header row.
- `--af` : Optional text file with allele frequency at each SNP, e.g. when the genotype file contains a subset of a larger population. Only relevant in combination with `--min_prob`. Either 1 column and no header, or multiple columns with a column MAF, AF, or Frequency. E.g. output from PLINK –freq.

Data formatting

The genetic data can (for example) be formatted as follows:

This is very close to the output from PLINK's `--recodeA` option:

```
plink --file mydata --extract plink.prune.in --recodeA --out MyData
cat MyData.raw | tr -s ' ' | cut -d ' ' -f2,7- > MyData.txt
sed -i '1d' MyData.txt
sed -i 's/NA/-9/g' MyData.txt
```

Where PLINK performs the reformatting from a .ped/.map file pair (or use `--bfile` for a .bed/.bim/.fam trio) to a file with extension `.raw`. But in this file

- the first 6 column are family ID - Individual ID - father - mother - sex - phenotype
- there is a header row
- missing values are coded as NA.

The subsequent `cut` and `sed` commands will

- replace multiple adjacent spaces to a single space; then take column 2 (individual ID) and column 7 onwards
- delete the first (header) row
- replace all `NA`'s by -9

to create a file that can be read by `find_pairs.f90` (as well as by the stand-alone version of sequoia, https://github.com/JiscaH/sequoia_notR).

Program options

Relationships

The type of relatives to be found is set with the `--focal <REL>` option, where `<REL>` is one of the following six abbreviations:

- S : Self; duplicated samples from the same individual (identical to `--dup`)
- PO: Parent–Offspring (identical to `--po`)
- FS: Full Siblings
- GP: Grand-Parental; indistinguishable from other types of second degree relationships: half siblings and full-avuncular
- HA: Half-Avuncular; indistinguishable from other types of third degree relationships
- UU: Unrelated.

Multiple relationships may be chosen as e.g. `--focal S --focal FS` or `--po --focal GP`. Each relationship will be done separately and written to a separate output file; the same pair may occur in multiple output files.

Thresholds

There are three different thresholds that can be set:

- `--max_dif` : Maximum number of SNPs at which two samples may differ to be considered duplicates. SNPs at which either or both samples are not scored are not counted. Only used in combination with `--dup` / `--focal S`.
- `--max_oh` (`--max_OH`) : Maximum number of SNPs at which two samples may be Opposing Homozygotes to be considered parent-offspring. Only used in combination with `--po`/`--focal PO`.
- `--min_prob`: Minimum probability of the `--focal` relationship to write a pair to the output file. A value of 0 is allowed, to calculate all probabilities but not do any filtering.

Pairs passing the thresholds are not internally stored in any way, but directly written to a text file. There is therefore no upper limit to the number of pairs that may be identified, nor will the number of identified pairs affect memory usage.

The thresholds `--max_dif` and `--max_oh` are applied before `--min_prob` when `--focal` is `S` or `PO`, respectively, to reduce computation time. Both default to zero, but can be set equal to the number of SNPs to skip this

filtering step. To check the number of SNPs in a formatted genotype file, you can use "`awk -F' ' '{print NF; exit}' GenoFile.txt`".

Other

- `--err` : presumed genotyping error rate. Obligatory when using `--min_prob`. Must be between 0 and 0.5, and strictly positive (to avoid divisions by zero).
- `--af` : population allele frequencies. If not provided, they will be calculated from the genotype file.
- `--out` : The suffix `_DUP.txt`, `_PO.txt`, `_FS.txt`, etc. will be appended to the specified file name. Defaults to `Pairs_maybe`.
- `--quiet` turn off most messages

Output

The output is a text file with the following columns:

- row1, row2 : row numbers in the genotype file of the samples
- ID1, ID2 : IDs of the samples
- OH : number of SNPs at which the pair are opposing homozygotes (`--po` only)
- nDiff : number of SNPs at which the pair differ (both non-missing, `--dup` only)
- Snpd1, Snpd2 : number of SNPs at which sample 1/2 is successfully genotyped
- SnpdBoth : number of SNPs at which both samples are successfully genotyped
- prob_S, prob_PO, prob_FS, prob_GP, prob_HA, prob_UU : Probability that the pair has a certain relationship, conditional on their genotypes, the assumed genotyping error rate, the allele frequencies, and assuming HWE. The relationships are:
  - S : Self, i.e. the two samples come from the same individual
  - PO: Parent-Offspring. The second sample may be either parent or offspring!
  - FS: Full Siblings
  - GP: 2nd degree relatives, either half-siblings, grandparent-grandoffspring, or full avuncular
  - HA: 3rd degree relatives, e.g. half-avuncular or great-grandparent
  - UU: Unrelated

Background: OH count

The Opposing Homozygosity count is the number of SNPs at which one of a pair of individuals has 0 copies of the reference allele, and the other has 2. For parent–offspring pairs the count is always zero, except due to genotyping errors. For full sibling pairs it may occasionally be zero, simply due to chance.

It is extremely rare for other pairs of relatives or unrelated pairs to have an OH count of zero. However, since in any dataset the number of unrelated pairs is orders of magnitude larger than the number of parent-offspring pairs, a non-negligible proportion of the pairs with an OH count of zero may be unrelated.

When using a non-zero threshold on OH count to select parent-offspring pairs, to allow for occasional genotyping errors, the large majority of selected pairs may in fact be unrelated.

For further information, see https://jiscah.github.io/articles/mendelian_inheritance.html .

Background: Relationship probabilities

For each sample pair, the likelihood to be duplicates, parent-offspring, full siblings, 2nd degree relatives, 3rd degree relatives, or unrelated is calculated. This is the probability of observing their genotypes, conditional on each relationship, as well as conditional on the population allele frequencies, the provided genotyping error rates, and the Mendelian inheritance laws for autosomal, biallelic markers. When making the simplifying assumption that the sample pair belongs to exactly one of these six categories (and not two, e.g. due to inbreeding, or are fourth degree relatives or so), the likelihoods for each sample pair can be rescaled to probabilities summing to 1.

Typically, likelihood calculations are relatively computationally expensive. However, by considering each pair independently of the next, and due to the simple nature of SNPs, a look-up table can be used. The dimensions of this table are

- 4 possible genotypes of individual i (0, 1 or 2 copies of the reference allele, or missing (-1))
- 4 possible genotypes of individual j
- number of SNPs; likelihood contributions vary between SNPs due to their different allele frequencies
- the six relationship categories (S, PO, FS, GP, HA, UU)

The values in this table are calculated once at initiation, after which for each sample pair the likelihoods at each SNP can be very efficiently looked up, and then the sum of the logarithm calculated.