Manual for 'find_PO_dups.f90'

A fortran program to find parent-offspring or duplicated sample pairs

Jisca Huisman

2023-11-01

Description

This program finds sample pairs that are likely to come from the same individual (duplicates) or from a parent-offspring pair. It does so by using a simple threshold on the number of SNPs at which the genotypes of two samples differ (duplicates) or at which they are opposing homozygotes (parent-offspring pairs). Optionally, additional filtering based on relationship probabilities can be used.

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. 6 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.

Program options

Choose `--dup` and/or `--po`, to search for duplicated samples resp. parent-offspring pairs. Thresholds can be set with:

- `--max_dup` : Threshold, only used in combination with `--dup`. 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. Valid aliases: `--maxDUP`, `--max_DUP`.
- `--max_oh` : Threshold, only used in combination with `--po`. Maximum number of SNPs at which two samples may be opposing homozygotes to be considered parent-offspring. Valid aliases: `--maxOH`, `--max_OH`.

By default, all pairs not exceeding the threshold are written to a text file. They are not internally stored in any way, and so there is no upper limit to the number of pairs that may be identified, nor will the number of identified pairs affect memory usage.

Relationship probabilities - background

When using only a OH threshold to identify potential parent-offspring pairs, the resulting list will also contain many other close relatives. For example, with typical genotyping error rates an OH count of 2 will occur only rarely between true parent offspring-pairs (« 1%), but quite frequently among full siblings and even occasionally by chance among other pairs.

An alternative, additional approach is to calculate the likelihoods for the sample pair to be duplicates, parent-offspring, full siblings, 2nd degree relatives, 3rd degree relatives, or unrelated. When making the simplifying assumption that the sample pair belongs to one of these six categories (and not two, e.g. due to inbreeding), the likelihoods for each 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.

Relationship probabilities - program options

- `--min_prob`: Threshold; minimum value of `prob_S` (when `--dup`) or `prob_PO` (when `--po`) to write pair to output file. A value of 0 is allowed, which will add the `prob_xx` columns to the output but not do any filtering.
- `--err` : presumed genotyping error rate. Obligatory when using `--min_prob`.
- `--af` : file with allele frequencies (see section 'Input files')

Output

The output file name can be specified with `--out`. The suffix `_DUP.txt` or `_PO.txt` will be appended to the specified file name. Defaults to `Pairs_maybe`.

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)
- SnpdBoth : number of SNPs at which both of the pair are succesfully 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, and 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