Manual for `pedigree_checker.f90`

A fortran program to check consistency between pedigree and SNP data

Jisca Huisman

2024-01-11

Description

This program calculates for each offspring - dam - sire trio in a pedigree the probability that either or both parents are

- PO = parent - offspring,
- FS = full siblings
- GP = grandparent, half-sibling, or full avuncular,
- HA = half avuncular or other 3rd degree relationship,
- UU = unrelated;

returning all 5x5=25 probabilities for each trio.

In fact, 'pedigree' is here extremely loosely defined: the pedigree file may contain any arbitrary list of trios, such as offspring - sire - maternal grandsire. Each individual may occur multiple times in the offspring column, e.g. with different candidate parents.

Pairs can be included too, as trios where one parent is missing ('NA'); probabilities are only calculated for this parent being unrelated (UU). Non-genotyped parents are treated the same as missing parents.

Assumptions and simplifications

- The probabilities for each trio or pair are calculated independently;
- Individuals can only be related as PO, FS, GP/HS/FA, HA, or UU; the possibility of inbred and double relationships is ignored;
- One 'parent' is related via the maternal side, and the other via the paternal side; the two 'parents' are unrelated;
- All SNPs are independent, have the same (average) genotyping error rate, and genotype frequencies are in Hardy-Weinberg equilibrium.

For more sophisticated treatment of relative pairs, please see `sequoia` (https://jiscah.github.io/).

Requirements & preparation

- a Fortran compiler, e.g. gfortran (https://fortran-lang.org/en/learn/os_setup/install_gfortran/)

The source code first needs to be compiled with

```
gfortran -O3 pedigree_checker.f90 -o PedChecker
```

where `gfortran` is the name of a Fortran compiler and `-O3` is the optimisation level. On many platforms `-O4` may be available as well for a faster program. The choice of program name (after `-o`) is completely free.

Input

SNP data

The genotype filename and location is specified with `--geno`.

To convert genotype data from standard PLINK format to the format for `sequoia` and `pedigree_checker`, use

```
plink --bfile genoFile --recode A --out genoFile
cat genoFile.raw | tr -s ' ' | cut -d ' ' -f2,7- > genoFile.txt
sed -i '1d' genoFile.txt
sed -i 's/NA/-9/g' genoFile.txt
rm genoFile.raw
```

which

- recodes to 1 column per SNP, coded as 0/1/2 copies of minor allele, missing = NA
- drops columns 2–6 (family ID, sex, phenotype, etc)
- drops header row
- replaces missing value code NA by -9

This is implemented in `format_SNP_data_for_sequoia.sh`:

```
./format_SNP_data_for_sequoia.sh genoFile
```

which is also available at https://github.com/JiscaH/sequoiaExtra .

Pedigree / trios

The pedigree file is specified with `--pedigreeIN` or `--trios`. This can be a pedigree with columns id - dam - sire, but may for example also be id - sire - maternal grandsire.

Column names are recycled and used in the output, and should be maximum 7 characters long, longer names are shortened in the output. IDs may be maximum 40 characters long. (Both limits are easily adjustable in the source code, near the top of 'module FileIO')

Allele frequencies

Optionally a text file with allele frequency at each SNP can be provided. This can be useful when the genotype file contains a small subset of a larger population. If none is provided, allele frequencies are calculated from the genotype data.

Execute

```
./PedChecker --geno genoFile.txt --pedigreeIN MyPedigree.txt \
  --err 0.005 --out PedOUT_test.txt
```

Here `--err` is the presumed genotyping error rate. This must be strictly positive, to avoid divisions by zero.

An overview of the commands can be found with `./PedChecker --help`.

Output

The output consists of the the trios in the pedigree file, with 3+25 (or 16 with `--noFS`) added columns:

- OH_<parent1>, OH_<parent2>: the count of opposing homozygous SNPs between the focal (offspring) individual and , and between the focal individual and <parent2>. <parent1> and <parent2> are the column names of the 2nd and 3rd column in the input file. Missing value = -9
- ME_pair: number of Mendelian Errors between the focal individual and the parent pair. Missing value = -9
- prob_PO_PO : the probability that both the dam and the sire have a Parent-Offspring relationship with the focal individual. Missing value = 9999.0000
- prob_.._.. : as for prob_PO_PO, where the first abbreviation denotes <parent1> and the second <parent2>, and
  - PO = parent – offspring
  - FS = full sibling
  - GP = grand-parental, half sibling, or full avuncular
  - HA = half avuncular or other 3rd degree relationship
  - UU = unrelated.

With option `--LLR` log10-likelihoods (LL) are returned instead of probabilities, scaled for each trio by the LL that both parents are Unrelated, i.e. results are not scaled by $P(R)$.

Post-processing in R

A logical next step is to determine for each trio the relationship combination with the highest probability, and/or for each 'parent' the most likely relationship to the focal individual. For this the R function `pedChecker_topRel` is available, which can be accessed in R using

```
source('https://raw.githubusercontent.com/JiscaH/sequoiaExtra/main/pedigree_checker/pedChecker_toprel.R
```

(To get the URL: from https://github.com/JiscaH/sequoiaExtra browse to the R file, then click 'raw')

Some help is included at the top of this R file.

Note that `pair_TopRel` may occasionally appear inconsistent with `dam_TopRel`. For example, a trio may have the following probabilities:

```
#     sire
# dam     PO    FS GP HA UU
#   PO 0.000 0.000  0  0  0
#   FS 0.000 0.000  0  0  0
#   GP 0.275 0.304  0  0  0
#   HA 0.416 0.005  0  0  0
#   UU 0.000 0.000  0  0  0
```

Then `pair_TopRel` = HA_PO, but `dam_TopRel` = GP (0.275+0.304 > 0.416).

Calculations

For each trio, the probability of observing the offspring genotype is calculated for each of the 5x5 possible trio relationships, given the parents' observed genotypes, the presumed genotyping error rate, and Mendelian inheritance laws.

Parent genotype

The parents actual genotype, correcting for any genotyping errors and filling in any non-called SNPs, is estimated as the product of:

- the probability to draw 0,1, or 2 copies of the minor allele from the population at random, assuming HWE,
- the probability of observing the observed genotype, if the actual genotype were 0,1, or 2 copies.

The actual genotype probabilities are scaled to sum to one at each locus.

Note that this differs from the approach in sequoia, where a parent's parents' genotypes contribute to the probability. The benefits of this are increased speed and no confusing consequences of an incorrect grandparent. The downside is that if a parent has a low call rate or some genotyping errors, the parent's parents' genotypes are not there to improve the estimate of the actual genotype.

Scaling

Following Bayes' theorem, the probability of a relationship R given the genetic data G can be calculated from the probability of G given R as

$$P(R|G) = \frac{P(G|R)}{P(G)} P(R)$$

Where $P(G)$ is the probability of observing the data ignoring any relationships, which for simplicity can be approximated by the probability if both parents where unrelated, $P(G) \approx P(G|UU)$.

The non-genetic probability of the relationship $P(R)$ includes in `sequoia` the age-difference based probability. Here, it is merely a scaling factor to ensure that $\sum_{R \in \{PO,FS,GP,HA,UU\}} P(R|G) = 1$ (for a given combination of genotypes G, the probabilities of all considered relationships R sums to one).

Disclaimer

While every effort has been made to ensure that this program provides what it claims to do, there is absolutely no guarantee that the results provided are correct. Use is entirely at your own risk.
