sequoia

Jisca Huisman
jisca.huisman@gmail.com

April 29, 2016

`Sequoia` provides a method to reconstruct multi-generational pedigrees based on SNP data, as described in the manuscript "Pedigree reconstruction using SNP data: parentage assignment, sibship clustering, and beyond". The bulk of the algorithm is written in Fortran, to minimise computation times.

Example

An example pedigree and associated life history file are provided with the package, which can be used to try out the steps detailed in this vignette. The pedigree consists of 5 generations with interconnected half-sib clusters (Pedigree II in the manuscript).

```
# install the package (only required once)
install.packages("sequoia")

# set the working directory
setwd("E:/Sequoia/test")

# load the package
library(sequoia)

# copy the example pedigree and associated life history file to the working
# directory.
file.copy(system.file("Ped_HSg5.txt", package="sequoia"), getwd())
file.copy(system.file("LH_HSg5.txt", package="sequoia"), getwd())

# simulate genotype data for 200 SNPs, and otherwise default values
SimGeno(PedFile = "Ped_HSg5.txt", nSnp = 200)

# run the preparation step, duplicate checking and parentage assignment,
# but not yet the slower sibship clustering
sequoia(GenoFile = "SimGeno.txt", LifeHistFile = "LH_HSg5.txt", Sibships = FALSE)
```

```
# compare the assigned parents to those in the true pedigree
PedCompare(PedIN = "Ped_HSg5.txt", PedOUT = "Parents_assigned.txt")

# run sibship clustering
sequoia(Prep = FALSE, CheckDup = FALSE, Parentage = FALSE, Sibships = TRUE)

# compare the assigned real and dummy parents to the true pedigree
PedCompare(PedIN = "Ped_HSg5.txt", PedOUT = "PedSeq.txt")
```

The duplicate checking and parentage assignment can be iterated a number of times for an empirical dataset, weeding out duplicated or (potentially) erroneous samples (using PLINK's toolkit), and adding estimated birth years or sexes to the life history file.

1 Input

Only two files are required as input. The first one contains the SNP data, with one line per individual and one column per SNP, and each SNP coded as 0, 1, 2 or missing. This file format can for example be obtained using PLINK [Purcell et al., 2007], as described below.

The other file contains three columns: individual ID, sex (1 =females, 2 = males, other numbers = unknown), and birth year. Ideally all genotyped individuals are included in this life history file with sex and (estimated) birth year information, but this is not strictly necessary. The life history file may include many more individuals than the genotype file, or in a different order. In species with more than one generations per year, a finer time scale than year of birth ought to be used, ensuring that parents are never born within the same time unit as their offspring.

Selection of SNP markers Using tens of thousands of SNP markers for pedigree reconstruction is unnecessary, will slow down computation, and may even hamper inferences by their non-independence. Rather, a subset of SNPs in low linkage disequilibrium (LD) with each other , and with high minor allele frequencies (MAF), ought to be selected first. The calculations assume independence of markers, or absence of LD in the founder population. While low (background) levels of LD are unlikely to interfere with pedigree reconstruction, high levels may give spurious results. Markers with a high MAF provide the most information [Hill et al., 2008], as although rare allele provide strong evidence when they are inherited, this does not balance out the rarity of such events.

Creating a subset of SNPs can be done conveniently using PLINK (`http://pngu.mgh.harvard.edu/~purcell/plink/`), using for example the command

```
plink --file mydata --maf 0.4 --indep 50 5 2
```

which will create a list of SNPs with a minor allele frequency of at least 0.4, and which in a window of 50 SNPs, sliding by 5 SNPs per step, have a VIF of maximum 2. VIF,

or variance inflation factor, is $1/(1-r^2)$. It is advised to 'tweak' the parameter values until a set with a few hundred SNPs is created. For further details, see `http://pngu.mgh.harvard.edu/~purcell/plink/summary.shtml#prune`.

The resulting list ('plink.prune.in') can be used to create the genotype file used as input for Sequoia, with SNPs codes as 0, 1 or 2, with the command

```
plink --file mydata --extract plink.prune.in --recodeA --out inputfile_for_sequoia
```

which will create a file with the extension .RAW.

Simulating SNP data When SNP data is not (yet) available, but an approximate pedigree is, it is possible to test `sequoia` on a simulated dataset. This may be useful to for example explore the number of markers required to reliably infer a pedigree of that particular structure.

The function `SimGeno()` can be used for this, which lets the user specify the average proportion of missing genotypes per individual, the genotyping error rate, and the fraction of known parents (in the supposed 'true' pedigree) which have not been genotyped. Moreover, the data can be made to contain a fraction of low-quality samples, to assess whether inclusion of samples which did not pass stringent quality control would improve or hamper pedigree reconstruction.

2 Preparation

The program `sequoia` consists of various sub-programs, which are all called via the function

```
sequoia(RawFile = NULL,
GenoFile = NULL,
LifeHistFile = NULL,
Dir = NULL,
Prep = TRUE,
CheckDup = TRUE,
Parentage = TRUE,
AgePriors = "par",
Sibships = TRUE,
MinAFR = 1,
Err = 0.0001,
MaxMismatch = 3,
LLRThreshold = 3.0,
LLRThresholdLow = 1.0,
MaxSibshipSize = 100,
NumSibRounds = 3,
quiet = FALSE)
```

details of the various parameters can be found using `?sequoia`, and the various subprograms are described below. `RawFile` denotes the genotype file created by PLINK, and `LifeHistFile` the lifehistory file with invidual ID, sex and birth year.

The preparation step amongst others slightly reformats the genotype file, by removing the header row, removing columns 2–6, which in Plink are intended for family ID, sex and phenotypic data, and replacing 'NA' by '-9' for missing values. Genotype files which are already in this format, such as simulated genotype files, can be specified as `GenoFile`.

Setting parameter values When the program is called, with `Prep = TRUE`, it takes the parameter values supplied, or the default values, and writes those to the file 'SequoiaSpecs.txt' in the current working directory, or the specified directory (option `Dir`). This file is used by the Fortran part of the program, which does all the heavy lifting. Please do not alter the name of this file, or the number of rows in it, as it will cause the program to crash. Do however feel free to change parameter values within this file, provided the number of SNPs or individuals does not exceed the total of SNPs or individuals in the genotype file; these values are used to read in the data.

Data quality control The data may contain positive controls, as well as other intentional and unintentional duplicated samples, which ought to be removed prior to parentage assignment. Typically such samples seem to get assigned as full siblings, but this has not been extensively tested. Sequoia includes a function to quickly search the data for identical genotypes, called with the option `CheckDup = TRUE`. It allows one or two mismatches between the genotypes, with or without the same individual ID. Note that when the number of SNPs is limited, very inbred individuals may be nearly indistinguishable from their parent(s), e.g. when resulting from a parent-offspring mating; such individuals should not be excluded. The same function also searches the lifehistory file for duplicated entries, and will print a list of individuals included in the genotype file, but not in the life history file. This list is merely a service to the user; individuals without life history information can often be successfully included in the pedigree.

3 Age based prior

During preparation also the file 'AgePriors.txt' is created, which contains 8 columns, and as many rows as the birth year range detected in the life history data input file. It initially only indicates whether a given relationship is biologically possible (1) or not (0) for a given age difference between individuals, where the first row is for individuals born in the same year, the second row for individuals born one year apart, etc. The columns are labelled for various relationship categories, with M = mother, P = father, MS = maternal sibling, PS = paternal sibling, MGM = maternal grandmother, PGF = paternal grandfather, MGF = maternal grandfather and paternal grandmother, and AU = avuncular.

For example, the first value in the column 'MS' can be interpreted as 'if I were to pick two individuals born in the same year, and two individuals from my sample at random, how much more likely are the first pair to be maternal siblings, compared to the

second pair?' Values below 1 indicate less likely, and values above 1 more likely; values of 0 indicate a biological impossibility. MS, PS and AU are symmetrical around 0 (with overlapping generations, nephews may be older than their aunts), while M, P, MGM, PGF and MGF are not.

These age-difference based priors are by default automatically updated after parentage assignment and prior to sibship clustering, based on the empirical distribution of age differences between individual and their assigned fathers and mothers. This update can be prevented with the option `Agepriors="old"`, or enforced later with the option `AgePriors="par"`. As with the parameter specification file, feel free to alter the values in this file to match the biological characteristics of the species, but please do not alter the number of columns. The number of rows may be increased (but not decreased below the age range amongst the genotyped individuals), in which case the entry 'nAgeClasses' in the file 'SequoiaSpecs.txt' should be updated to match the new number of rows.

4 Parentage assignment

Parentage assignment is performed with the option `Parentage = TRUE`, which will use the parameter settings in the file 'SequoiaSpecs.txt', created by `Prep = TRUE`. Parentage assignment is quick, it takes about 10–20 seconds for a dataset with 2,500 genotyped individuals on a laptop with an intel i7 2.3 GHz CPU and 8GB RAM, and should not take more than a few minutes even on larger datasets and/or computers with lower processing speed.

The assigned parents are written to a text file, rather than returned within R. Besides the columns with the IDs of the offspring, its assigned mother and its assigned father, this file also contains columns with the log10 likelihood ratio (LLR) of the mother, father and the parent pair, the number of loci at which the offspring and the mother or father were opposite homozygotes, and the row number in the genotype file of the offspring, mother and father (the last three columns are used when reading in the file for subsequent sibship clustering). LLR is the ratio between the probabilities of the assigned parent being the parent, and the assigned parent being otherwise related (e.g. a full sibling), or unrelated, to the focal individual. This differs from for example Cervus [Marshall et al., 1998], which returns the natural log of the ratio between the probability that the assigned parent is the parent, and that the next most likely candidate is the parent. In the output file, LLR values of 999.0 and OH values of -9 represent missing values (NA).

Some parents may have a very small or even negative single-parent LLR, for example associated with close inbreeding. The LLR of the parent pair will however always be positive, and is relative to the most likely assignment of a single parent.

In addition, a file may be written with identified but non-assigned parent-offspring pairs. These are non-assigned either because it was not possible to tell which of the two was the parent and which was the offspring, due to either or both individuals having an unknown birth year, or because it was not possible to tell whether the candidate parent was the mother or the father. These situations can be remedied by providing estimated birth years, or guessed genders, for the individuals involved in the life history input file,

and re-running `sequoia`. The short computational time of parentage assignment means that this step can be repeated a number of times as a part of data quality control.

5 Sibship clustering

Sibship clustering amongst those individuals which have not been assigned a genotyped parent is performed with the option `Sibships = TRUE`, which will also use the parameter settings in the file 'SequoiaSpecs.txt'. Sibship clustering is considerably slower than parentage assignment, and may take from a few minutes to a few hours, depending on the number of individuals without a parent, and the number of sibships that is being constructed.

Number of iterations Convergence is typically reached within five iterations, even for complex pedigrees, but more iterations may be required when the proportion of genotyped parents is small and there are many interconnected sibships. When convergence is reached before the user-set maximum number of iterations, a final iteration with more lenient dependence on the age prior is ran, (dummy)parental likelihoods are calculated, and the algorithm is terminated.

Dummy names By default, dummy parents are denoted by increasing numbers, with prefix 'F' for females and 'M' for males. The prefixes can be altered in the 'SequoiaSpecs.txt' file, for example to avoid confusion with IDs of real individuals.

Output The output of the sibship clustering is written to a text file which by default is called 'PedSeq.txt', and which is highly similar to the output of the parentage assignment. For each individual it contains its assigned parents from the parentage step, and/or dummy-parents from the sibship clustering step, and the associated LLR. For dummy parents, the LLR reflects how much more likely the focal individual is to be a member of the sibship, versus otherwise related to its members (e.g. a full aunt).

Dummy individuals are appended at the bottom of this pedigree file, with their assigned parents, i.e. the sibship's assigned grandparents. Here, the LLR column denotes how much more likely the assigned parent is to be the grandparent of the dummy individuals' sibship, than otherwise related to it (e.g. a member or great-grandparent).

In addition, a file with non-assigned pairs of relatives may be created, containing for example half-siblings were it could not be determined whether they are maternal or paternal half-siblings, and grandparents of singletons.

6 Comparison with previous pedigree

Often times, a (part) pedigree is already available to which one wants to compare the results. This pedigree may consist only of maternal links, deduced from observations in

the field. Or, sequoia is run on simulated data and one wishes to compare the inferred pedigree to the true pedigree. The R function `PedCompare()` performs such comparisons, and takes as only arguments the file names of the 'true' pedigree and of the inferred pedigree.

Replacement of dummy IDs by IDs of non-genotyped individuals For many species, mothers can be assigned with fairly high accuracy based on field observations. In intensively monitored systems, a likely father or short-list of candidate fathers may be available as well, such as the mother's social mate, or based on proximity to a nest. While it is currently not possible to incorporate this type of information jointly with the genetic information to estimate parentage, as for example in MasterBayes [Hadfield et al., 2006], a function will be provided to integrate the inferred pedigree with these other sources of information.

The required input consists of a text file with four columns: the ID of the focal individual, the ID of the likely parent, the parent's sex (only known sex individuals are permitted), and the estimated probability that this is the parent of the focal individual. These probabilities do not need to sum to 1 per focal individual or per candidate parent.

References

JD Hadfield, DS Richardson, and T. Burke. Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a bayesian framework. *Molecular Ecology*, 15(12):3715–3730, 2006.

WG Hill, BA Salisbury, and AJ Webb. Parentage identification using snp genotypes: application to product tracing. *Journal of animal science*, 2008.

TC Marshall, JBKE Slate, LEB Kruuk, and JM Pemberton. Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular ecology*, 7(5): 639–655, 1998.

S Purcell, B Neale, K Todd-Brown, L Thomas, MAR Ferreira, D Bender, J Maller, P Sklar, PIW De Bakker, MJ Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3): 559–575, 2007.

Table 1: Overview of the algorithm

	Step	Calculate	Action	Comments
I	Parentage assignment	# opposite homozygous loci A–B	Continue if $< T_{OH}$	Very fast
		LLR(PO/U) A–B	Continue if $> -T$	
		Joined $\mathcal{L}$(H0) –$\mathcal{L}$(H6) :	Assign A if resulting $\mathcal{L}$ highest	Repeat 3x; assume A oldest
		pair A–B, given parents D_B, S_B		Current parents D_B & S_B optional
		pair A–B, given parent D_B		When sex A = sex S_B
		pair B–D_B, when A parent		
		pair B–S_B, when A parent		
II	Find sibling pairs			No age priors in 1st round
		LLR(HS/U) for $A + B$	Continue if $> -T$	
		$\mathcal{L}$(H0) –$\mathcal{L}$(H6) for $A + B$	List if $\mathcal{L}$(FS) or $\mathcal{L}$(HS) highest	
III	Cluster pairs			*
		If neither currently in sibship:		
		$\mathcal{L}$(H0) –$\mathcal{L}$(H6) for $A + B$ (again)	New sibship if $\mathcal{L}$(FS) or $\mathcal{L}$(HS) still highest	
		Else if B in sibship:		
		$\mathcal{L}$(H0) –$\mathcal{L}$(H6) for $A + \boldsymbol{B}$	Add A to $\boldsymbol{B}$ if $\mathcal{L}$(FS) or $\mathcal{L}$(HS) highest	$\boldsymbol{B}$ = half-sib cluster
		Else if $A + B$ in separate sibships:		
		$\mathcal{L}$(H0) –$\mathcal{L}$(H6) for $\boldsymbol{A} + S_B$	Merge if $\mathcal{L}$(PO) highest	S_B = dummy parent of $\boldsymbol{B}$
IV	Add to sibships			Consider all A, $\boldsymbol{B}$
		LLR(HS/U) for $A + \boldsymbol{B}$	Continue if $> -T \times n_B$	n_B = size of half-sib cluster $\boldsymbol{B}$
		$\mathcal{L}$(H0) –$\mathcal{L}$(H6) for $A + \boldsymbol{B}$	Add A to $\boldsymbol{B}$ if $\mathcal{L}$(FS) or $\mathcal{L}$(HS) highest	
V	Merge sibships			Consider all $\boldsymbol{A}$, $\boldsymbol{B}$
		LLR(HS/U) for $\boldsymbol{A} + \boldsymbol{B}$	Continue if $> -T \times n_A \times n_B$	
		$\mathcal{L}$(H0) –$\mathcal{L}$(H6) for $A + \boldsymbol{B}$	Add A to $\boldsymbol{B}$ if $\mathcal{L}$(FS) or $\mathcal{L}$(HS) highest	
VI	Sibship parents			E.g. when unknown birth year
		# opposite homozygous loci A - B_i	Continue if all $< T_{OH}$	For all B_i in $\boldsymbol{B}$
		$\mathcal{L}$(H0) –$\mathcal{L}$(H6) for $A + \boldsymbol{B}$	Assign A & dissolve $\boldsymbol{B}$ if $\mathcal{L}$(PO) highest	
VII	Sibship grandparents	LLR(GP/U) for A or $S_A + \boldsymbol{B}$	Continue if $> -T$	LRR(GP/U)=LLR(HS/U)
		Joined $\mathcal{L}$(H0) –$\mathcal{L}$(H6) :	Assign A (S_A) if resulting $\mathcal{L}$ highest	For both real A and dummy S_A
		$\boldsymbol{B}$–A, given GP $G_{B,1}$, $G_{B,1}$		Current grandparents optional
		$\boldsymbol{B}$–A, given parent GB1		
		$\boldsymbol{B}$–$G_{B,1}$, if A grandparent		
		$\boldsymbol{B}$–$G_{B,2}$, if A grandparent		

A and B denote focal individuals, $\boldsymbol{A}$ and $\boldsymbol{B}$ focal half-sib clusters.

Repeat steps II–VII 2–10×, until convergence of total likelihood. The first round consists of steps II-III-V only.

*: Only in the last round, allow pairs where distinction between HS, GP and FA can only be made based on age information.