Pedigree file and the inverse of the relationship matrix in honey bees

Version 20

Pim Brascamp, 26 June 2023
For comments and questions please contact pim.brascamp@wur.nl

This note describes two R-programs. The first produces a pedigree file for honey bees consisting of colonies, dams and sires. Colonies are groups of workers, dams are individual queens and sires are groups of drone-producing queens (DPQ). The programs allow for the situation that the sire is only one queen, which in particular is relevant for single-drone insemination. The program also allows natural mating. The second program produces the numerator relationship matrix **A** and its inverse.
For theory see Brascamp and Bijma (GSE, 2014, DOI:10.1186/s12711-014-0053-9), Brascamp *et al.* (Apidologie, 2016, DOI: 10.1007/s13592-016-0427-9) and a erratum to the Apidologie paper (Apidologie, 2018, DOI: 10.1007/s13592-018-0573-3). The erratum concerns the interpretation of the variance component for colony's worker effect. A paper going into this issue more extensively was published in GSE (Brascamp and Bijma GSE, 2019, DOI: 10.1186/s12711-019-0510-6.). The program computing the inverse of **A** does not include the inversion for very large datasets as developed by Bernstein et al. (2018, J Anim Breed Genet. 2018;135:323–332.)
The programs and files are the result of continuous development. The programs show traces of that and another consequence is that sequences of items in the files are sometimes no longer logical. In 2015 a FORTRAN-version of AMD-AINV was written by Jérémie Vandenplas which was extremely much faster that the R-version. Since then the R-version was extended with additional mating options. On the basis of Jérémie's work recently Tristan Kistler wrote a FORTRAN-version for his simulation work using versions 19 (T. Kistler, E.W. Brascamp, B. Basso, P. Bijma and F. Phocas, submitted).

To derive the relationship matrix **A** assumptions have to be made about the probability that two full-sibs descend from the same drone, and for the probability that they descend from the same drone-producing queen. Until recently the program inverting **A** assumed a Poisson distribution for both but as pointed out by Manuel Du this leads to probabilities larger than 1 if NS or ND equals 1, where NS is the number of DPQ constituting a sire and ND is the number of drones a queen on average fertilizes. See equation 23a in Brascamp & Bijma (2014). The inversion program now also accommodates the assumption that all drones have an equal contribution to the patrilines in the colony (equation 23b).
In Brascamp and Bijma, 2019, we show that the estimates of genetic parameters depend upon the assumptions about the base population. If the assumptions don't represent the real situation estimates of genetic parameters will be biased considerably. We distinguished the situation that drone-producing queens are unrelated in the base population or that their additive genetic relationship has equilibrium values because the breeding program already was

running some time preceding the data to be analysed. The default values for NS and ND that are input in the computer program setting up the pedigree relate to this base population. Sometimes, however, with more complicated (and probably more realistic) breeding programs it is not obvious what's taken to be the base population. As the levels of the estimated genetic parameters are influenced considerably by the assumptions about the base population it is important to explicitly pay attention to this issue when publishing results.

Both programs sometimes produce printed warnings and both programs produce log-files (.txt). Sometimes a program may run very short. This may relate to an early error, reported in the log-file. Apart from input-errors, reported problems generally concern sequence problems in the pedigree file or that the product of **A** and its inverse is not the identity matrix. If you can't solve these problems please let me know (pim.brascamp@wur.nl) and I can try to help you out.

The programs deal with the following situations or combinations of these:

1. A sire is identified by the dam of the full-sib group of drone-producing-queens that constitutes the sire. A particular dam may have different sires, as e.g. the case when she produced a sire for a mating station but also for instrumental insemination.
2. A sire is identified itself. There are a number of options for this case:
   a. The sire is a single sire producing one drone (single-drone insemination) or various drones for insemination. The program accommodates the situation that a particular single sire produces different numbers of drones.
   b. The sire is a group of open mating drones produced by unknown DPQ. In that case the additive genetic relationship between drone-producing queens is taken to be zero.

Alfons Willam is gratefully acknowledged because our collaboration led to the first version of the programs. Also, Ralph Büchler, Arista Bee Research (BartJan Fernhout, Bart Barten, Mari van Iersel and Tieme Wanders) and Florence Phocas and Tristan Kistler are gratefully acknowledged for let me work with their data such that I could extend and check the programs for various assumptions in real data. Tristan Kistler and Benjamin Basso used the programs extensively giving rise to some changes especially related to open mating.

1. pedigree.r (version 20).

Inputs are files steer.txt and input-pedigree.txt, the formats of which is given in Appendix 1. Steer.txt contains a number of steering parameters for both pedigree.r and AMD-AINV.r. For the former three parameters need to be defined.

- Default values for NS (the number of drone-producing queens) and ND (the number of drones fertilizing a queen) should be given. Usually this is something like 8-20 and e.g. 12. These values are taken for matings in the base population and also in later

generations if there is a record without mate specified. Special attention should be given to NS in case of open mating. See for this Appendix 2.

- There is a parameter codetest that equals 1 if a thorough test is carried out on the proper sequence of the pedigree. A second part of that test takes a lot of time and usually doesn't detect errors. For initial runs it may be useful to put codetest unequal 1 to limit the sequence test to the less time-consuming part, but ultimately codetest should be put to 1 to be pretty sure that inversion of the relationships matrix doesn't create problems.

The program produces several output files of which log-pedigree.txt reports whether the pedigree seems to be okay, especially with respect to the sequence of the individuals in the pedigree. There are two requirements. Firstly, parents should precede offspring, but specific for honey bees, aunts should precede cousins. Usually, the program takes care of this. If changes have to made in the input file, it may be recommendable to let the changes do by the program. Look for *file specific issues causing sequence problems* and add a script with the changes to be made. It often boils down to changing years of birth (column 5).

Among further output there are four files that are input for the inversion program: pedigree.txt, steer.txt, ident.txt and blocks.txt. In situation 2a, single sires, also singlesire.txt is input for the inversion program. The file steer.txt is input for both pedigree.r and the inversion program and some parameters in this file are output from pedigree.r.

The formats of these files are given in Appendix 1.

Various variations in input are accommodated in pedigree-input.txt and are described below.

1a, 2a and 4a are input

a. The input for the pedigree file usually consists of the identifications of the queen (1a) in the colony, its dam (2a) and the dam (4a) of the drone-producing queens (1b). Drones of those drone-producing queens are mated to the 1a. In addition the years of birth are to be added and, in case of colonies with observations, the year of observation.
b. In some applications a 4a may have more than one 1b, for example with mating on an island and instrumental insemination as well. In that case, an additional column with an additional single digit needs to be entered per 4a in column 18 to distinguish between the 1b. It may happen that beforehand it is not realized that more than one 1b exists per 4a. NS and ND are input for each record in pedigree-input.txt and the program tests whether for a 4a always the same NS and ND are present in the input file. If that

is not the case there is either a typo or different 1bs exist per 4a, and column 18 can be used if different 1bs exist per 4a.

Insemination from one drone-producing queen (1a, 2a, 1b and 4a are input)
In some applications only one drone-producing queen is used. In that case the identification of 1b has to be input as well. Otherwise it is zero. Note that in this case the identification for 4a should also be given, to distinguish from the case where queens are open mated.

Open mating (1a, 2a and 1b are input)
It may be that the 4a is not known and should be zero because queens are open mated. For each record with an open-mated queen, a 1b has to be added. There are different options for this, discussed in Appendix 2.

Deeper pedigree
It may be that the input file not only contains records that have observations, but also deeper pedigree information. For deeper pedigree information without observation the year of observation needs to be zero. Note that also for these records the numbers of DPQ and drones of the queen's mate should be given.

Sequence problems
In some cases the program doesn't arrive at a sequence suitable to derive **A**. In that case enter in the program an additional line (or lines) that makes specific changes to the modified year of birth of a specific queen, usually the 2a (See in the program the paragraph file specific issues causing sequence problems). The output file pedigree_ident.txt may facilitate to detect the reason of a sequence error. It may, however be more efficient to make these changes in the program that produces the input file for pedigree.r.

Pedigree.r is programmed such that all members in the pedigree have two or zero parents. If a parent is missing the program generates it. For dams it is a unique dam for each member and for sires it is a base sire with the specified assumptions, i.e. whether the drone-producing queens are unrelated or that their additive genetic relationship is in equilibrium, and NS and ND as specified in the beginning of the program. There may be reasons to deviate from that (e.g. in case of open mating) and then the sire should be added as with open mating.

2. AMD-AINV.r (version 20)

Input are the five or six files specified above.
The input file steer.txt specifies the following parameters.

- equi, being zero or one. In case of zero the drone-producing queens in base sires are taken to be unrelated. If one, the additive genetic relationship between drone-producing-queens is assumed to be in equilibrium. The latter is likely to be the case if a selection program is already running a couple of generations before the start of the pedigree of the data that are analysed.
- dist, being zero or one. In case of zero a Poisson distribution is assumed for the number of workers per drone and the number of workers per DPQ and one for equal proportions of workers per drone.
- pinv, being zero or one. In case of zero the inverse (AINV) of the numerator relationship matrix is computed utilizing mendelian sampling (co)variances. It is recommended to do this once because it is a fairly reliable test whether everything (i.e. the pedigree file but also the computer program related to special perhaps unforeseen cases) is okay.
  In case of one AINV is computed with the R-function 'solve'. It is likely that the first option (pinv=0) allows larger datasets than the latter, but the latter option (pinv=1) is (considerably) faster. So, particularly in the case of a simulation where several runs with different AINV need to be analysed, the latter option is to be preferred.
- priF, being zero or one. This concerns writing of inbreeding coefficients of possible future matings and if priF=0 these will not be computed nor written. If priF=1, an additional parameter should be entered, yip, that tells for how many years until the last year the coefficients should be written. The purpose of it is to give an indication of whether a particular planned mating doesn't result in too high inbreeding and that's only useful for queens that were born rather recently. Three coefficients are written: Mating of queens reared from queen A with drones of DPQ reared from queen B and vv, queens reared from queen A mated with drones of queen B, and queens reared from queen B mated to drones from queen A.

The program produces a log file: log-AMD-AINV.txt. Of special relevance is the result of the multiplication of **A** and its inverse that should be identity. Output file is AINV.giv which can be used as input for ASReml. Further output files are pedigree_complete.txt, ident_complete.txt and F-future.txt.

AMD-AINV.r assume that all individuals in the pedigree have two or zero known parents. In case an individual has one known parent, pedigree.r creates the second, that itself has no known parents. Animals with unknown parents form the base population.
The difference between pedigree.txt and pedigree_complete.txt is that for each individual the inbreeding coefficient and the mendelian sampling variance is added, and for sires also the additive genetic relationship between two drone-producing queens. Furthermore, the diagonal element of A-inverse is added, and the identification in column 1 (equals the sequence number in column 16) is replaced by the real identification in case of queens and

fabricated identification in the case of colonies (groups of workers), sires (groups of drone-producing queens) or base dams. Another difference is that in column 12 the diagonal element of AINV is given. This information is useful for the calculation of accuracies of EBVs.

3. **Y.txt**. This is input file for ASReml and contains the identification of the colony (if worker effect of workers is to be included in the statistical model) and also the dam of workers (1a, if also the queen effect of the dam of workers is to be included in the statistical model) and furthermore fixed effect(s) and observations. The identifications are sequence numbers 2 (column 17) as in pedigree_complete.txt. For flexibility the creation of Y.txt has been brought outside of AMD-AINV.r. Therefore, a specific program needs to be written to produce Y.txt. Apart from the input file (the file that was used to produce input-pedigree.txt) also pedigree_complete.txt is needed, to link identifications to sequence numbers 2 and also to the sequence number 2 of the dam in the colony. A small program Y.r is available to serve as a starting point.

4. **EBVs**. Often EBVs are a desired result of the analyses. In ASReml these are in the file with extension .sln. The identifications are the sequence numbers 2. Usually, the colonies are identified by the real identification of the 1a (the dam of workers) and therefore again pedigree_complete.txt is needed as input for a program to list EBVs.

Appendix 1. Formats of various files

steer.txt

This file specifies three parameters (1-3) for pedigree.r and five parameters (7-11) needed for AMD-AINV.r. Parameters 4-6 are needed for AMD-AINV.r too, but these are output of pedigree.r.

1 NS: number of DPQ in base sires
2 ND: number of drones in base sires
3 Codetest: zero if extensive testing of the sequence in the colony is not asked for, and 1 if it is.

4 NTOT: total number of colonies, queens and DPQ. This is output of pedigree.r.
5 ncol: number of colonies. This is output of pedigree.r.
6 single: Zero if there are no single sires, and 1 if there are. This is output of pedigree.r.
7 nblocks: Number of blocks in the relationship matrix. This is output of pedigree.r.

8 equi: zero if unrelated DPQ in the base population are assumed and 1 when additive genetic relationships between DPQ are assumed in equilibrium.
9 dist: zero when the numbers of workers per drone and per DPQ follow a Poisson distribution and 1 with equal proportion of workers for each drone.
10 pinv: zero when inversion is using mendelian sampling terms and 1 when direct inversion is carried out.
11 priF: zero if inbreeding coefficients of workers from possible matings are not computed nor written and 1 if these are for the yip most recent years of birth of colonies.
12 yip: only needed if priF=1.

input-pedigree.txt

1 year of observation (zero if it concerns deeper pedigree info without observations)
2 bee breeder
3 not used, zero
4 year of birth 1a
5 identification 1a
6 year of birth 2a
7 identification 2a
8 year of birth 4a
9 identification 4a
10 not used, zero
11 not used, zero
12 not used, zero
13 not used, zero

14 NS (number of drone-producing queens in the 1b mated to the 1a)

15 ND (average number of drones produced by the 1b mated to the 1a)

16 year of birth of 1b (if one single queen or open mating, otherwise zero)

17 identification of 1b (if known or fabricated in case of open mating, otherwise zero)

18 extra digit, with default zero. This is only needed if a particular 4a occurs with different sets of drone-producing queens, that may differ for NS and ND.

pedigree.txt, pedigree_complete.txt and pedigree_ident.txt

Table with in the first column pedigree.txt, in the second pedigree_complete.txt and in a third pedigree_ident.txt.

x item is present

xx sequence number 2

xxx identifications, for queens in colonies the original idents in input-pedigree.txt, for other queens, colonies, and sires, fabricated idents.

1	identification	x	xxx		xxx	colony, dam or sire
2	Sex (1=dam, 2=sire, 3=colony	x	x		x	x
3	mate (for dams only)	x	x		xxx	mate
4	type of mating (for sires only)	x	x		xxx	mate's dam
5	year of birth modified	x	x		x	x
6	sequence number of dam	xx	xx		xxx	dam
7	sequence number of sire	xx	xx		xxx	sire
8	a_{SS} (for sires only)		x		xxx	sire's dam
9	F		x			
10	var MS		x			
11	testing station	x	x		x	
12	diagonal element of A		x			
13						
14						
15						
16	sequence number 1	x	x		x	
17	sequence number 2	xx	xx		xx	
18	year of birth	x	x		x	
19	NS (number of DPQ)	x	x		x	
20	ND (number of drones)	x	x		x	

Note that NS and ND for dams relate to those of their mate. For sires they relate to the numbers of the sires themselves. For single sires the situation is divergent. First of all ND for a

single sire may vary among single sires and is therefore not informative. Furthermore, if the single sire appears as dam as well, the numbers relate to her mate.

Sequence number 2 is the sequence number of the individuals in the pedigree after sorting on year of birth modified in such a way that offspring of the same dam (and therefore sire) is in a block.

Column 12 gives the diagonal elements of A. These can be used to compute accuracies of EBVs for the case that the EBV only contains one random effect, for example worker effect. In that case the accuracy equals $r_{A_i\hat{A}_i} = \sqrt{1 - \frac{s_i^2}{\sigma_{A_i}^2}}$, where s_i^2 is the square of seEffect from ASReml (more generally, it equals the square of the standard error of prediction, SEP) and $\sigma_{A_i}^2$ equals the diagonal element in column 12 multiplied by the estimate of σ_A^2 from ASReml. If the EBV contains for example worker and queen effect, then the inverse of the block in AINV referring to the worker and the queen is needed. ASReml can produce SEP in that case as well. The SEP of the worker group can be obtained by inclusion of the command line (personal communication Arthur Gilmour)

PREDICT worker a: b queen a: b ! ONLYUSE worker queen ! PARALLEL worker queen

directly after the line defining the model. In this equation a and b refer to sequence numbers in the pedigree file. (Note that we experienced a limit to the size of $b - a$ that can be dealt with in a single run.)

ident.txt

This is an in-between file containing sorted on sequence number 1, after initially sorting the data on year of birth and identification of 1a.

1 true identification (for queens in the colonies) and fabricated for colonies, other queens (if present) and sires.

blocks.txt

The number of records is the 7th item in numbers.txt. For each record:

1 first sequence number 2 in the block
2 last sequence number 2 in the block

singlesires.txt

This file only is relevant in case of single sires with NS=1

1 sequence number 2 for the single sire
2 identification of the single sire
3 number of drones inseminated with

AINV.giv

1 sequence number 2 of the first member of the pedigree (i)
2 sequence number 2 of the second member of the pedigree (j)
3 element of AINV

The file is organized such that i ≥ j. It only contains non-zero elements.

F-future.txt

This file summarizes inbreeding coefficients (F) of possible matings using colonies of the last yip years. If there are two colonies x and y there are three possible matings.

a. A virgin queen from x with drones of drone-producing queens from y and the reciprocal; both lead to the same F.
b. A virgin queen from x with drones from y.
c. Drones from x with a virgin queen from y.

These latter became relevant because in case of single-drone insemination the drone usually is produced by the queen in a tested colony and not by drone-producing queens reared from tested colonies. These two F's differ because for b the pedigree of the mated queen x is relevant and that of the queen y itself, while for c it is the reverse.

1 sequence number 2 of the first colony (i)
2 sequence number 2 of the second colony (j)
3 real identification of the queen (dam) in colony i
4 real identification of the queen (dam) in colony j
5 F (%) for case a.
6 F (%) for case b.
7 F (%) for case c.

The file is organized such that i ≥ j. The file only contains colonies born in the last yip years.

Y.txt (for example)

1 sequence number 2 of the colony
2 sequence number 2 of the dam (if the statistical model needs it)
3 test location
4 Possible other fixed effects and observations

In earlier versions of pedigree-x.r the observations were part of input-pedigree.txt. Because the very variable numbers of observations that is no longer the case. There is a little example of a program Y.r that may serve as a starting point for Y.r.

Appendix 2

Open mating

Incorporation of open mating is still under development. This appendix describes the thoughts until now following from discussions with Tristan Kistler and Piter Bijma in particular as part of a study looking into datasets with open mating (Kistler et al. 2023, submitted).

Operationally, open mating is implemented by adding the identification of a sire (a set of DPQ) in column 17 of input-pedigree.txt plus its year of birth (column 16) and NS and ND (columns 14 and 15). Columns 8 and 9 (the dam and year of birth of the set of DPQ) should be zero.

Suppose the situation that all open mating takes place from one stable gene pool of drones. In that case one sire in column 17 suffices with a certain NS and ND, taking NS fairly large (hundreds) and set ND equal to 10-20. In the situation that there are different gene pools, the identification of several sets of DPQ seems appropriate. Kistler et al. show, however, that this has its problems. In particular if the genetic variance across these gene pools is (a lot) larger than that in the breeding population with its founding base population, the additive genetic variance estimated from this dataset will be (a lot) larger than that in the base population. The cause is that the variation among gene pools is co-estimated to arrive at the estimate of the additive genetic variance. If this is not intended, then the inclusion of a fixed (or random) effect for the gene pools in the statistical model to estimate the variance leads to a proper estimate in the base population.

For the estimation of variance components this solution seems okay, but if (some of) the open mating sires are genetically superior to the breeding population then a question is whether the inclusion of a fixed (or random) effect doesn't hinder the introduction of the open mating genes into the breeding population, while this introduction is desirable. Taking into consideration that it will slow down the introduction of these superior genes but likely not entirely prevent it in later generations, inclusion of a fixed (or random) effect still may be wise as breeding values then are estimated using proper genetic parameters.

The discussion above relates to the systematic use of open mating in a breeding program as in the case of Kistler et al., where DPQs were open mated and their colonies phenotyped. A question is what the consequences are of the discussion above for the incidental use of open mating in a breeding program, usually when a colony is included of which the queen's dam is open mated and this dam doesn't have a record. In the current implementation of pedigree.r it is assumed that this dam's mate is a set of DPQ unique for this dam and belongs to the base population. Again this may introduce unwanted additive genetic variance. Also, if the breeding population includes many generations with considerable genetic gain it is likely that the DPQ used for open mating show genetic gain, too. How to deal with this, especially if open mating is frequent, deserves research attention.