

Calculations involving genomic relationship matrices and friends available in WOMBAT

1	Introduction	1
2	Notation and abbreviations	2
3	Overview	2
4	Technical details	2
5	Input files	3
6	Specifications	4
6.1	Default settings	4
6.2	SPECIAL block options	5
6.2.1	Number of markers	5
6.2.2	Centering of allele counts	5
6.2.3	Method of calculating the GRM	5
6.2.4	Method of calculating the inverse of the GRM	5
6.2.5	Constant to ensure a positive definite GRM	6
6.2.6	Weighting of genomic and polygenic information	6
6.2.7	Determinant of \mathbf{H}	7
6.2.8	Diagonal elements of \mathbf{H}	7
6.2.9	Calculating the GRM only	7
6.2.10	Calculation of \mathbf{A}_{22}	7
6.2.11	Scaling the GRM	7
6.2.12	Calculating \mathbf{A}^{-1} with metafounders	8
6.2.13	Genetic groups for \mathbf{H}^{-1}	9
6.2.14	'Correction factors' τ and ω	9
6.2.15	Output options	10
7	Output files	10
8	Worked examples	12
8.1	A: Default settings	12
8.2	B: GRM only	12
8.3	H: \mathbf{H}^{-1} with APY approximation of \mathbf{G}^{-1}	13
8.4	I: $\mathbf{A}^{-\gamma}$ for a single meta-founder	13
8.5	J: $\mathbf{H}^{-\gamma}$ for two meta-founders	13
	References	14
	Appendix A Appendix: Valid lines in the parameter file (incomplete)	15

1 Introduction

Genetic evaluation and variance component estimation using genomic relationships have become common for mixed model analyses in quantitative genetics. A plethora of software is available to carry out calculations required to determine genomic relationship matrices and related quantities, differing in their capabilities, input and output formats available.

WOMBAT accommodates genomic relationships through externally calculated general covariance matrices (where the inverse needs to be supplied in a file with the extension **.gin**) or, for single-step genomic BLUP, through the inverse of the joint relationship matrix between genotyped and non-genotyped animals, \mathbf{H}^{-1} . A module has been added to WOMBAT to allow these quantities to be computed in a pre-analysis step (invoked by run option **--hinu**), with the input and output (mostly) in the format required by WOMBAT subsequently. In addition, miscellaneous related calculations are available, e.g. to set up an inverse numerator relationship matrix (\mathbf{A}^{-1}) incorporating meta-founders (Legarra et al., 2015). This note describes the variety of options currently available.



N.B.

This module is work in progress – testing so far has been rudimentary, specifications or formats may be subject to change and additional options are likely to be introduced.

- There are multiple options which can be combined – not all combinations have been tested and conflicting specifications may go unnoticed.
- Some options are currently available only for LINUX executables compiled using **ifort**.

2 Notation and abbreviations

NRM	Numerator relationship matrix
GRM	Genomic relationship matrix
G	symbol for GRM
A	symbol for NRM (all animals)
A₂₂	symbol for submatrix of A for genotyped animals
A^{ij}	symbol for <i>ij</i> -th submatrix of A ⁻¹
H	symbol for joint relationship matrix of genotyped and non-genotyped animals

3 Overview

Calculations available can be grouped as

1. Calculation of a genomic relationship matrix (GRM), its eigenvalues or its inverse.

This includes options for ‘centering’ marker counts, to scale the GRM (**G**) towards the corresponding submatrix of the numerator relationship matrix (NRM), **A₂₂** and to approximate the inverse using the so-called APY algorithm.

2. Calculation of **H**⁻¹.

This is the default. Output of **H**⁻¹ to file is available either as complete matrix (half-stored) of the ‘add-on’ part to **A**⁻¹ only. Calculation of log |**H**| – needed to compare models when estimating variance components (e.g. to estimate the weighing factor λ to combine genomic and pedigree based relationships) – or of selected diagonal elements of **H** – needed to compute accuracies of predicted breeding values – can be requested.

3. Calculation of **A**⁻¹ including meta-founders (MF).

Matrices of ‘self-relationships’ can be supplied or estimated as described by Garcia-Baccino et al. (2017).

4 Technical details

The GRM is calculated initially from the marker counts as

$$\mathbf{G}_M = (\mathbf{M} - 2\mathbf{P})(\mathbf{M} - 2\mathbf{P})' / s \quad (1)$$

or

$$\mathbf{G}_M = (\mathbf{M} - 2\mathbf{P})\mathbf{W}^{-1}(\mathbf{M} - 2\mathbf{P})' / m \quad (2)$$

following Van Raden (2008) or Yang et al. (2010), respectively, where **M** denotes the matrix of marker counts (of size number of genotyped animals \times number of markers), **P** the corresponding matrix of assumed frequencies p_i , **W** is a diagonal matrix with elements $2p_i(1 - p_i)$, m is the number of markers and $s = 2 \sum_i p_i(1 - p_i)$.

G_M is often modified to ensure that it can ‘safely’ be inverted, to improve alignment between GRM and NRM or to account for residual polygenic variation. Common types modifications can be summarised as

$$\mathbf{G} = \lambda [\beta(\mathbf{G}_M + \epsilon\mathbf{I}) + \alpha\mathbf{J}] + (1 - \lambda) \mathbf{A}_{22} \quad (3)$$

with $0 \leq \lambda \leq 1$ denoting the proportion of total genetic variance due to marker effects, α and β the ‘alignment’ factors proposed by Christensen (2012) or Vitezica et al. (2011), \mathbf{J} a matrix with all elements equal to unity, \mathbf{I} an identity matrix and ϵ a small constant.

The inverse of the joint relationship matrix is then calculated as

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \tau \mathbf{G}^{-1} - \omega \mathbf{A}_{22}^{-1} \end{pmatrix} \quad (4)$$

where τ and ω represent heuristic scale factors that have been suggested to reduce bias in estimated breeding values.

An alternative to scaling \mathbf{G} as above is to modify \mathbf{A} (Christensen, 2012) which (for $\alpha = 0$ and $\beta = \tau = \omega = 1$) gives

$$\mathbf{H}_\gamma^{-1} = \mathbf{A}_\gamma^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-\gamma} \end{pmatrix} \quad (5)$$

with

$$\mathbf{A}_\gamma = (1 - \gamma/2) \mathbf{A} + \gamma \mathbf{J} \quad (6)$$

where γ represents ‘self-relationships’ and $\mathbf{A}_{22}^{-\gamma}$ is the inverse of the submatrix of \mathbf{A}_γ for genotyped animals. Legarra et al. (2015) proposed a corresponding adjustment, capable of accounting for different base populations, by augmenting \mathbf{A} by one or more, so-called meta-founders.

5 Input files

Depending on the choice of calculations to be carried out, up to five input files are required:

1. A pedigree file, specified in the parameter file as usual in a line beginning with **PED** followed (space-separated) by the file name. It is expected to follow the same format as used by WOMBAT for other animal model analyses (i.e. three space separated columns with animal, sire and dam codes) but must be augmented with a fourth column, containing a “1” for non-genotyped and a “2” or “3” for genotyped animals.

Note that this should be a complete pedigree (i.e. include codes for all animals which might occur). Further, in contrast to other WOMBAT runs, there is no cross-check against animal codes in the data file.

A pedigree file is not required when calculation of the GRM only is specified.

2. A file with genotype information, expected to be read animal by animal. This has the default name **MarkerCounts.dat**. The default extension **.dat** implies a formatted file. If this is not found, WOMBAT will attempt to open and read from **MarkerCounts.BIN** or **MarkerCounts.BI1** as an unformatted file, which can be faster. An alternative filename can be used but requires adding the line

MRK filename

to the parameter file (where **filename** can have extension **.dat**, **.BIN** or **BI1**; see WOMBAT manual for details).

As for the single-step modules, each ‘row’ for an animal is expected to begin with the animal code (matching the codes in the pedigree file) followed by the marker allele counts. This is a list-directed FORTRAN read, reading the animal code as a standard length INTEGER and the allele counts as single precision REAL or INTEGER*1 variable. For large numbers of markers, the ‘row’ can be spread over several lines in the input file (but the next animal must always start with a new line). For the `--hinv` option, animal codes are NOT used – instead it is assumed that allele counts are given in the SAME sequence as the genotyped individuals occur in the list of animal codes in increasing order.

Marker counts must be complete. i.e. specified for all animals marked as genotyped in the pedigree file and missing counts for individual markers are not accommodated. Note also that WOMBAT does NOT perform any checks or quality control on the contents of this file.

A marker file is not required if only calculations involving A^{-1} or its submatrices are chosen.

3. A file named **MFGamma.dat** if calculation of A_{γ}^{-1} for more than one metafounder with pre-defined values for Γ is required. This formatted file should contain the upper triangle of the matrix of self-relationships (space or line separated).
4. A formatted file named **HdiagIDs.dat** if computation of a subset of diagonal elements of H for non-genotyped relatives of genotyped individuals is required. This file should give the numerical identities of those animals for which diagonals are needed. These should be given one per line.
5. A file named **Frequencies** if pre-computed values for marker allele frequencies are to be used to center allele counts. This is expected to be a formatted file, containing the frequencies to be used for centering as m space (or line) separated variables. Again, no checks for validity of the numbers given are carried out.

6 Specifications

The relevant module is invoked by specifying run option `--hinv`. All options are expected to be supplied in a **SPECIAL** block in the parameter file. For this run option, a reduced parameter file supplying only the names of input files and any special options suffices. If a ‘full’ parameter file (including model specification and covariance components) is given, information not relevant is skipped.

6.1 Default settings

If no other options (except for the mandatory number of markers) are given, the default for run option `--hinv` is to calculate H^{-1} (as given in (4)) and to write out its upper triangle element-wise to the file **Hinverse.gin**, in the format required by WOMBAT for subsequent REML or BLUP analyses. No intermediate results are written out and the determinant of H is not calculated – a dummy value of zero is written instead.

The default invokes calculation of the GRM (G) using Van Raden (2008)’s method I, centering allele counts by observed frequencies, and adding a value of $\epsilon = 0.01$ to the diagonal elements to ensure a positive definite matrix. The submatrix of the numerator relationship matrix for genotyped animals, A_{22} is calculated using Colleau (2002)’s method.

The default weight for \mathbf{G} in combining \mathbf{G} and \mathbf{A}_{22} is $\lambda = 1$, i.e. no weight is given to the latter. No alignment between \mathbf{A} and \mathbf{G} is attempted ($\alpha = 0$ and $\beta = 1$) and no heuristic scaling is carried out ($\tau = \omega = 1$).

6.2 SPECIAL block options

Relevant options are specified as ‘one per line’ (i.e. multiple options require multiple lines) with supplied keywords and values separated by spaces. All lines should begin with the keyword **HINVERSE** which can be abbreviated to **HINV**.

6.2.1 Number of markers

This option is required unless calculation of \mathbf{A}_{22} or \mathbf{A}_{22}^{-1} only is requested. It is specified by a line

```
HINV SNP m
```

with m an INTEGER value representing the number of markers.

6.2.2 Centering of allele counts

This is specified by the line

```
HINV CENTER keyword
```

where valid keywords are

NONE no centering,

HALF centering assuming all allele frequencies are equal to 0.5,

FREQ centering using observed frequencies (default), and

BASE centering using estimated founder frequencies (obtained using the method of McPeck et al. (2004)), and

FIXP centering using pre-calculated, fixed frequencies. If this option is given, WOMBAT will expect to read these from the file **Frequencies** (see Section 5).

6.2.3 Method of calculating the GRM

This can be selected by the line

```
HINV HOWGRM keyword
```

where valid keywords are

VRADEN1 for Van Raden (2008)’s method I (default), and

YANG for Yang et al. (2010)’s method.

6.2.4 Method of calculating the inverse of the GRM

The default calculation is a ‘full’ inverse, by carrying out a Cholesky factorisation and then inverting the factored matrix. If calculation of eigenvalues and eigenvectors of the GRM is specified (see Section 6.2.9), the inverse is obtained from the elements of the eigen-decomposition instead.

This requires the inverse of a matrix of size $n_2 \times n_2$, with n_2 denoting the number of genotyped animals. If n_2 is larger than the number of markers (m) it is computationally advantageous to utilise the Woodbury matrix identity, as suggested by Mäntysaari et al. (2017), as it reduces the size of the inverse required to $m \times m$. This calculation can be specified by the line

```
HINV GRMINV WOODBURY
```

if only G^{-1} is to be calculated, or, more generally,

HINV WOODBURY x

when only the eigenvectors explaining the first x percent of total variation are to be used in constructing G^{-1} to build H^{-1} ; see Mäntysaari et al. (2017) for a description of the “TBLUP” approach (to do: better description & worked example).

Note that use of this identity requires the GRM to be of the general form $B + FF'$: Currently the respective calculations are implemented for $G = G_M + \epsilon I$ or $G = \lambda (G_M + \alpha J) + (1 - \lambda)A_{22}$ (with G_M the GRM as computed from the marker information).

Approximate inverse A widely used approximation of the inverse for large values of n_2 is that of the so-called APY algorithm (Misztal et al., 2014), which requires dividing genotyped animals into so-called core and non-core animals. Calculation of the APY form of the inverse is specified with the line

HINV GRMINV APY keyword

where keyword specifies how to identify core animals. Four options are recognised:

PEDFILE (default) directs WOMBAT to identify core animals flagged in the pedigree file with a code of “3” in column 4 (replacing the standard “2” for genotyped animals).

The number of core animals is obtained by counting the number of such individuals.

FIRST x instructs WOMBAT to treat the first n genotyped animals as core animals.

RANDOM x selects n genotyped animals at random to represent the core.

PROGENY x counts the number of progeny (in the pedigree) for each animal and picks out those with the highest numbers.

If x is an integer number (>1) it presents the number of core animals, n . Alternatively, if $0 < x < 1$, x is interpreted as the proportion of genotyped animals to be treated as core and n is calculated as the nearest integer to xn .

If calculation of the inverse of the GRM only is specified, it is assumed that pedigree information is not given and animal ‘type’ codes are thus not available. Hence core selection defaults to the **FIRST** n animals found in the file with marker allele counts.

6.2.5 Constant to ensure a positive definite GRM

A constant is added to the diagonal elements of G to ensure a numerically stable inverse. This value can be set using

HINV EPSILON ϵ

with ϵ (default value 0.01) representing the constant (REAL) to be used.

Note that adding a constant may not be desired for centering options **HALF**, **FIXP** or **NONE** or when G_M is to be replaced by a weighted average of G_M and A_{22} . However, there is no automatic adjustment of the default value, i.e. a value of 0 needs to be specified explicitly to eliminate this step.

6.2.6 Weighting of genomic and polygenic information

Rather than using genomic information alone, it is common practice to use the weighted average $\lambda G + (1 - \lambda)A_{22}$ instead of G in constructing H^{-1} . The value of λ (default 1.0) can be set with the line

HINV LAMBDA λ

6.2.7 Determinant of H

By default this is not calculated and the value for the log determinant given in **Hinverse.gin** is zero. If required, calculation of this quantity (which can be computationally demanding if H^{-1} is large) can be requested via the line

```
HINV DET
```

6.2.8 Diagonal elements of H

By default these are not written out. If required, calculation can be specified via the line

```
HINV DIAGH
```

If this option is selected, diagonal elements of H are calculated as described by Legarra et al. (2020, Method3). The default is that this is done as an add-on to calculation of H^{-1} , i.e. that both G and A_{22} have been set up and held in core. Alternatively, specifying

```
HINV DIAGH ONLY
```

invokes a run which does not require these two matrices, providing only $\text{Diag}(H)$.

6.2.9 Calculating the GRM only

Calculation of G only is set with the line

```
HINV GRM
```

Calculation of G and its eigenvalues can be requested with the line

```
HINV GRMEIG
```

To obtain the eigen-vectors as well, add the keyword **VECTOR**, i.e. use

```
HINV GRMEIG VECTOR
```

instead. To write out the matrix of eigenvectors as a binary file, add the option **BIN** after the keyword **VECTOR** (space-separated) To write out the first two eigenvectors only (suitable for a biplot) use

```
HINV GRMEIG BIPLLOT
```

(formatted output). If G^{-1} is needed the line

```
HINV OUT GRMINV
```

should be given in addition.

6.2.10 Calculation of A_{22}

This is regulated by the line

```
HINV A22 keyword
```

where valid keywords are

COLLEAU selects calculation of A_{22} using Colleau (2002)'s method (default).

INDIRECT selects calculation of A_{22}^{-1} 'indirectly' from the submatrices of A^{-1} , i.e. as $A_{22}^{-1} = A^{22} - A^{21}(A^{11})^{-1}A^{12}$ which can be done without inverting A^{11} , requiring only the (sparse) factorisation of A^{11} . This option is available only for LINUX versions of WOMBAT compiled using **ifort**.

ONLY requests calculation of A_{22} (**COLLEAU**) or A_{22}^{-1} (**INDIRECT**) only, with the setting for the method of calculation determining which of the two is produced.

6.2.11 Scaling the GRM

There are two options for the modification of G to align better with A :

1. The line

HINV ALPHA α

causes the term $\alpha \mathbf{J}$ to be added to \mathbf{G} . If α is given as -9 , it is replaced by the difference in average elements of \mathbf{A}_{22} and \mathbf{G} , as suggested by Vitezica et al. (2011).

2. The line

HINV SCALEG

invokes calculation of α and β and scaling of \mathbf{G} as proposed by Christensen (2012); see (3) above.

Values of α and β used are reported in output file **SumPedigree.out**.

6.2.12 Calculating \mathbf{A}^{-1} with metafounders

For calculations fitting n_{MF} meta-founders, it is assumed that these are ‘first’ n_{MF} animals in the pedigree file, i.e. the animals with the *lowest* numerical codes (best coded 1 to n_{MF}) and the top n_{MF} records in the file. Pedigrees and animal codes in the marker counts file need to have been recoded accordingly.

Calculation \mathbf{A}_γ^{-1} alone can be selected by the line

HINV AGAMMA

in the parameter file. This requires that the degree(s) of ‘self-relationship’ have been determined.

As above, the default is to compute \mathbf{H}^{-1} , but replacing \mathbf{A}^{-1} with \mathbf{A}_γ^{-1} . This is specified using

HINV META γ SCALE

The keyword **SCALE** is optional. If given, it invokes scaling of \mathbf{H}^{-1} as described by Legarra et al. (2015) so that previously estimated variance components for the standard NRM are still applicable (default is no scaling).

- Values of $\gamma < 2$ imply that there is a single meta-founder. If $\gamma > 0$, it is taken to reflect the degree of self-relationship to be used. If γ is specified as -1 , it is attempted to estimate the it.
- Absolute values of $\gamma \geq 2$ are translated to the nearest INTEGER value, assumed to represent the number of meta-founders to be included. If this value is positive, the upper triangle of the matrix of self-relationships, $\mathbf{\Gamma}$, has to be supplied in a file named **MFGamma.dat** (see Section 5). If it is negative, it is attempted to estimate $\mathbf{\Gamma}$.

The resulting values for γ or $\mathbf{\Gamma}$ are written to screen and also given in **SumPedigree.out**. Estimation is carried out as described by Garcia-Baccino et al. (2017) and requires (uncentered) marker allele counts as well as pedigree information to be available. The default method is via generalised least squares. Use of the EM algorithm can be selected by adding the qualifier **EMALG** to the line above

HINV META γ EMALG

The maximum number of iterates performed by the EM algorithm can be regulated by adding **ITS n** (with n the number of iterates) to the line, e.g. for $n = 20$

HINV META γ EMALG ITS20

Calculations can be restricted to the estimation of $\mathbf{\Gamma}$ (or γ) only by adding the keyword **ONLY** to the line; e.g.

HINV META γ EMALG ONLY

This allows inspection of the estimate of $\mathbf{\Gamma}$ (and possibly manual fine tuning) prior to building \mathbf{A}_γ^{-1} or \mathbf{H}_γ^{-1} in a second step.

Insufficient marker information: Reliable estimation of Γ requires sufficient genotype information for all metafounders. For field data, this may not be the case. Estimation of Γ involves the matrix $\mathbf{Q}_2' \mathbf{A}_{22}^{-1} \mathbf{Q}_2$ (see Garcia-Baccino et al., 2017). This is the coefficient matrix in the generalised least squares estimator. Hence its diagonal elements can be interpreted as a (crude) measure of the information available. If there is no marker information for a relative of a particular metafounder, the corresponding rows and columns of this matrix are zero. By default, WOMBAT sets the respective diagonal elements to unity. This yields an estimate of Γ with corresponding rows and columns of zero.

Alternatively, WOMBAT provides the facility to restrict estimation to the submatrix of Γ for metafounders with those diagonal exceeding a specified value. This is invoked by adding a line

```
METAFOUND TRUNC x
```

where x is the minimum value allowed. On output, this submatrix is expanded by adding appropriate rows and columns with diagonal elements of -9 and off-diagonal elements of zero.

In either case, the resulting matrix is not positive definite and not suitable to build \mathbf{A}_y^{-1} without any manual modifications – hence WOMBAT will stop after the estimation step.

6.2.13 Genetic groups for \mathbf{H}^{-1}

Fitting genetic groups for single-step analyses in WOMBAT requires the proportions of memberships for individual groups to be specified in the pedigree file or the ***.codes** file corresponding to the inverse of a user specified relationship matrix (given as ***.gin** file).

For genetic groups to be fitted ‘explicitly’ (i.e. as additional random effect), WOMBAT offers the facility to transfer the group proportions from the pedigree file to the ***.codes** file while building \mathbf{H}^{-1} . This is specified by adding the line

```
GENGROUPS gname ng factor
```

which is in the same format as for other WOMBAT analyses. Here **gname** is a character variable and **factor** is an integer scale factor, both of which are not used in this module but are expected for consistency.

Alternatively, genetic groups can be fitted ‘implicitly’ by augmenting \mathbf{H}^{-1} with rows and columns corresponding to unknown parent groups *a.k.a* ‘phantom’ parents. This can be specified by adding the keyword **PHANTOM**, i.e. adding the line

```
GENGROUPS gname ng factor PHANTOM
```

to the **SPECIAL** block. This will add the rows and columns for genetic groups after the animals in the pedigree. If implicit genetic groups are to be fitted as fixed effects in subsequent analyses, it may be desirable to ‘zero out’ one level. This can be achieved by adding the line

```
HINVERSE GENGROUP0 n1
```

where **n1** is the running number of the genetic group to be omitted.

6.2.14 ‘Correction factors’ τ and ω

These values can be set using

```
HINV TAU  $\tau$ 
```

or

```
HINV OMEGA  $\omega$ 
```

with τ and ω the respective factors to be used (default values $\tau = \omega = 1$).

6.2.15 Output options

There are a number of options for the form and amount of output for individual matrices that is generated. Most (symmetric) matrices are written out element-wise to a formatted file, considering non-zero elements in the upper triangle only. Bear in mind that writing many large, formatted files to disk can be slow.

Output options are specified as “**HINV OUT** keyword” or “**HINV OUT** keyword *REname*”. The latter is optional and is only recognized for the first three keywords listed; if given it replaces the default term **Hinverse**.

Valid keywords for writing out H^{-1} are:

GIN specifies to write H^{-1} element-wise to a formatted file. This is the default and the filename is **Hinverse.gin** or *REname.gin*.

BIN is similar to option **GIN** but writes to an unformatted file named **Hinverse.BIN** or *REname.BIN*.

BIN22 causes the sparse blocks of H^{-1} , namely H^{11} and H^{12} , to be written to an unformatted file **Hinverse.BIN** or *REname.BIN*, as above. The upper triangle of the remaining dense block H^{22} is then written column-wise to an unformatted file **Hinverse.BIN22** or *REname.BIN22*.

DELTA writes out only the part of $H^{-1} - A^{-1}$ for genotyped animals to the unformatted file **Hinverse-Delta.BIN**.

For options **GIN**, **BIN** and **BIN22** output files are in the correct format to be used with WOMBAT for a subsequent REML or BLUP analysis to supply the relationship structure for a random effect *REname* with covariance option **GIN**. The accompanying ***.codes** file needed is also written out. The current version of WOMBAT will automatically read from file(s) *REname.BIN* and, if applicable, *REname.BIN22* (files **Hinverse.*** may need to be renamed).

Currently, no special form to exploit the sparsity of an APY inverse of the GRM is available and, if given, **BIN22** is automatically replaced by **BIN**.

Other options allow selected intermediate files to be written out:

GRM write out **G** to file **GRM.dat**,

GRMINV write G^{-1} to **GRMInv.dat**,

A22 write A_{22} to **A22.dat**,

A22INV write A_{22}^{-1} to **A22Inv.dat**, and

ALL write out all intermediate matrices ($G, G^{-1}, A_{22}, A_{22}^{-1}$).

7 Output files

A summary of options used and selected, additional information for values calculated are appended to the standard output file **SumPedigree.out**.

Output files with extension **.dat** are formatted, those with extension **.bin** or **.BIN** are binary.

Files with relationship matrices (or their inverses) are (mostly) written out element-wise, considering the non-zero elements of the upper triangle only. Each row is comprised of 3 elements: row number (INTEGER), column number (INTEGER) and coefficient (REAL8). Row and column numbers are ‘running numbers’, i.e. 1 to number of animals in total or 1 to number of genotyped animals, except for **Hinverse-Delta.BIN**.

Files **X.gin**, **Hinverse.gin** and **Hinverse-Delta.BIN** contain an additional first line with a `REAL8` variable (as placeholder for the log determinant of the matrix; set to zero unless calculation is explicitly requested) to yield the format expected for WOMBAT runs fitting random effects with the **GIN** option.

Hinverse.gin is the default for H^{-1} as given in (4) or (5), written as a formatted file. Alternative names or binary formats, provided for use in subsequent WOMBAT analyses, can be selected as described in Section 6.2.15.

Hinverse.codes is written alongside **Hinverse.gin** (or equivalent), listing the running numbers and original animal codes.

Hinverse.hdiags (or equivalent) contains the diagonal elements of **H** if requested. This is a formatted file with one line per animal containing

1. The running number
2. The original animal ID'
3. The diagonal element h_{ii}
4. The animal type (1: non-genotyped, 2/3: genotyped)
5. Related to genotyped animals(T: true, F: false)

Only the first three columns are used in a subsequent BLUP run to calculate accuracies.

Hinverse-Delta.BIN is the alternative output form for H^{-1} . It contains the non-zero coefficients of $\Delta = G^{-1} - A_{22}^{-1}$ (or equivalent) and row and column numbers are replaced by the original animal codes. This is the format as required by WOMBAT runs with option `--s2step`.

GRM.dat gives the genomic relationship matrix after adding a constant to the diagonal (if $\epsilon > 0$), i.e $G_M + \epsilon I$.

GRMEig.dat gives the eigenvalues of the GRM in descending order (column 2) together with the sum so far (column 3) and the proportion of total variance explained so far (column 4).

GRMEvecs.dat or **GRMEvecs.BIN** gives the eigenvectors of the GRM in descending order of eigenvalues. It is written out eigen-vector by -vector, using a list-directed FORTRAN write.

GRMEV1+2.dat gives the first and second eigenvector of the GRM only. It is written out individual by individual using a list-directed FORTRAN write.

GRMInv.dat gives the 'full' inverse of GRM. If only the GRM is calculated, this is the inverse of $G_M + \epsilon I$. Otherwise it is the inverse of the modified (if applicable) **G** as given in (3).

GPYInv.dat or **GPYInv.BIN** gives the approximate inverse of the GRM using the APY algorithm.

GPY.dat or **GPY.BIN** the approximate GRM corresponding to **GPYInv.***. Only calculated if output of the GRM is requested.

A22.dat gives the submatrix of the NRM for genotyped animals (A_{22}).

A22Inv.dat gives the inverse of A_{22} .

X.gin holds the inverse of NRM including meta-founders (A_y^{-1}).

X.codes is the 'codes' file to accompany **X.gin** needed for subsequent analyses in WOMBAT.

MFGammaNew.dat contains the upper triangle of the current estimate of matrix Γ .

8 Worked examples

Examples to illustrate the use of individual options are given in (directory) **Example20**. We show the parameter files for 4 selected tasks:

8.1 A: Default settings

Subdirectory **A** shows the minimal parameter file for the default settings, which produces the output file **Hinverse.gin**.

```
RUNOP --hinv -v
# optional comment line
COM Example20/A: Build H^{-1} with default settings

# need pedigree info
PED ../InputFiles/peds.dat

# need marker counts
MRK ../InputFiles/MarkerCounts.dat

SPECIAL
# there are 6000 markers
HINVERSE SNP 6000
END

# output file is Hinverse.gin
```

8.2 B: GRM only

Subdirectory **B** illustrates computation of G and G^{-1} only, generating output files for both.

```
RUNOP --hinv -v
COM Example20/B: calculate G and G^{-1} only

# no data or pedigree file or model specs required

# number of genotypes = no. of rows in "MarkerCounts.dat"
MRK ../InputFiles/MarkerCounts.dat

SPECIAL
# specify calculation of G-inverse only (not H-inverse)
HINVERSE grminv
# write out G as well as G-inverse
HINVERSE out grm
# there are 6000 marker counts
HINVERSE SNP 6000
# center marker counts assuming frequencies of 0.5
HINVERSE center half
# value to be added to diagonals of G
HINVERSE epsilon 0.01d0
END

# output files are GRM.dat and GRMInv.dat
```

8.3 H: H^{-1} with APY approximation of G^{-1}

```

RUNOP --hinv -v
COM Example20/H: Calculate  $H^{-1}$  with APY inverse of G

# need pedigree info
PED ../InputFiles/peds.dat

# need marker counts
MRK ../InputFiles/MarkerCounts.dat

SPECIAL
# there are 6000 markers
HINVERSE SNP 6000
# add constant to diagonals
HINVERSE epsilon 0.05d0
# APY inverse for G; select core: 150 animals with most progeny
HINVERSE APY PROGENY 150
END

# output files: Hinverse.gin and Hinverse.codes

```

8.4 I: $A^{-\gamma}$ for a single meta-founder

In subdirectory **I**, the example given by Legarra et al. (2015) is used to illustrate calculation of A^{-1} incorporating a single meta-founder. Only A_{γ}^{-1} is calculated. Output files **X.gin** and **X.codes** are generated to be used in a subsequent WOMBAT run with user-defined relationship matrix (rename as appropriate).

```

RUNOP --hinv -v
COM Example20/I:  $A^{\gamma}$  for single metafounder; toy example

# specify the pedigree file to be used - already renumbered
PED ../InputFiles/ped_1metafounder

SPECIAL
# gamma value to be used
HINVERSE METAFU 0.20
# specify construction of A-gamma inverse only
HINVERSE agammaonly
END

# outfiles files : X.gin and X.codes ->  $A^{\gamma}$  in GIN format

```

8.5 J: $H^{-\gamma}$ for two meta-founders

Subdirectory **J** shows the corresponding example for two meta-founders and **J** illustrates calculation of H^{-1} with A_{γ}^{-1} replacing A^{-1} .

```

RUNOP --hinv -v
COM Example20/J: Calculate  $H^{-1}$  for  $A^{-\gamma}$  with 2 metafounders

# specify pedigree - renumbered to include MF ; added column 4
PED ../InputFiles/pedigree2

# genotypes - same numbering for animals as pedigrees incl. MF
MRK ../InputFiles/MarkerCsMF.dat

SPECIAL
# no. of markers
HINVERSE SNP 500
# the no. gives number of MF; MFGamma.dat must exist; scaled = true
HINVERSE METAF 2 SCALE
# choose p = 0.5 as in baccino et al 2017
HINVERSE CENTER half
# write out intermediate results: X.gin contains A-gamma inverse

```

```

HINVERSE out AGAMMA
# no multiple of I added to G (not needed because p=0.5)
HINVERSE EPSILON 0.d0
END

# outfiles files : Hinverse.gin and Hinverse.codes
#                  X.gin and X.codes -> A^gamma in GIN format

```

References

- Christensen O.F. Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. *Genet. Sel. Evol.* 44 (2012) 37. doi: [10.1186/1297-9686-44-37](https://doi.org/10.1186/1297-9686-44-37).
- Colleau J.J. An indirect approach to the extensive calculation of relationship coefficients. *Genet. Sel. Evol.* 34 (2002) 409–422. doi: [10.1051/gse:2002015](https://doi.org/10.1051/gse:2002015).
- Garcia-Baccino C.A., Legarra A., Christensen O.F., Misztal I., Pocrnic I., Vitezica Z.G., Cantet R.J.C. Metafounders are related to F_{ST} fixation indices and reduce bias in single-step genomic evaluations. *Genet. Sel. Evol.* 49 (2017) 34. doi: [10.1186/s12711-017-0309-2](https://doi.org/10.1186/s12711-017-0309-2).
- Legarra A., Aguilar I., Colleau J. Short communication: Methods to compute genomic inbreeding for ungenotyped individuals. *J. Dairy Sci.* 103 (2020) 3363–3367. doi: [10.3168/jds.2019-17750](https://doi.org/10.3168/jds.2019-17750).
- Legarra A., Christensen O.F., Vitezica Z.G., Aguilar I., Misztal I. Ancestral relationships using metafounders: finite ancestral populations and across population relationships. *Genetics* 200 (2015) 455–468. doi: [10.1534/genet-ics.115.177014](https://doi.org/10.1534/genet-ics.115.177014).
- Mäntysaari E.A., Evans R.D., Strandén I. Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals. *J. Anim. Sci.* 95 (2017) 4728–4737. doi: [10.2527/jas2017.1912](https://doi.org/10.2527/jas2017.1912).
- McPeck M.S., Wu X., Ober C. Best linear unbiased allele-frequency estimation in complex pedigrees. *Biometrics* 60 (2004) 359–367. doi: [10.1111/j.0006-341X.2004.00180.x](https://doi.org/10.1111/j.0006-341X.2004.00180.x).
- Misztal I., Legarra A., Aguilar I. Using recursion to compute the inverse of the genomic relationship matrix. *J. Dairy Sci.* 97 (2014) 3943–3952. doi: [10.3168/jds.2013-7752](https://doi.org/10.3168/jds.2013-7752).
- Van Raden P.M. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91 (2008) 4414–4423. doi: [10.3168/jds.2007-0980](https://doi.org/10.3168/jds.2007-0980).
- Vitezica Z.G., Aguilar I., Misztal I., Legarra A. Bias in genomic predictions for populations under selection. *Genet. Res.* 93 (2011) 357–366. doi: [10.1017/S001667231100022X](https://doi.org/10.1017/S001667231100022X).
- Yang J., Benyamin B., McEvoy B.P., Gordon S., Henders A.K., Nyholt D.R., Madden P.A., Heath A.C., Martin N.G., Montgomery G.W., Goddard M.E., Visscher P.M. Common SNPs explain a large proportion of the heritability for human height. *Nature Genet.* 42 (2010) 565–569. doi: [10.1038/ng.608](https://doi.org/10.1038/ng.608).

Appendix A Appendix: Valid lines in the parameter file (incomplete)

HINV	SNP	m		Section 6.2.1
HINV	EPSILON	ϵ		Section 6.2.5
HINV	ALPHA	α		Section 6.2.11
HINV	LAMBDA	λ		Section 6.2.6
HINV	TAU	τ		Section 6.2.14
HINV	OMEGA	ω		Section 6.2.14
HINV	META	γ	SCALE, EMALG, ONLY, ITS	Section 6.2.12
HINV	SCALEG			Section 6.2.11
HINV	HOWGRM	VANRADEN1		Section 6.2.3
HINV	HOWGRM	YANG		Section 6.2.3
HINV	CENTER	FREQ		Section 6.2.2
HINV	CENTER	BASE		Section 6.2.2
HINV	CENTER	HALF		Section 6.2.2
HINV	CENTER	FIXP		Section 6.2.2
HINV	CENTER	NONE		Section 6.2.2
HINV	A22	COLLEAU		Section 6.2.10
HINV	A22	INDIRECT		Section 6.2.10
HINV	A22	ONLY		Section 6.2.10
HINV	GRM			Section 6.2.9
HINV	GRMEIG	VECTOR		Section 6.2.9
HINV	GRMEIG	BIPLOT		Section 6.2.9
HINV	GRMINV			Section 6.2.9
HINV	GRMINV	WOODBURY		Section 6.2.4
HINV	GRMINV	APY	FIRST,RANDOM,PROGENY	Section 6.2.4
HINV	WOODBURY	x		Section 6.2.4
HINV	AGAMMA			Section 6.2.12
HINV	OUT	GIN	REname	Section 6.2.15
HINV	OUT	BIN	REname	Section 6.2.15
HINV	OUT	BIN22	REname	Section 6.2.15
HINV	OUT	DELTA		Section 6.2.15
HINV	OUT	GRM		Section 6.2.15
HINV	OUT	GRMINV		Section 6.2.15
HINV	OUT	A22		Section 6.2.15
HINV	OUT	A22INV		Section 6.2.15
HINV	OUT	ALL		Section 6.2.15
HINV	DET			Section 6.2.7
HINV	DIAGH	ONLY		Section 6.2.8
HINV	GENGROUP0	n		Section 6.2.13
GENGROUPS	name	n	factor	Section 6.2.13
GENGROUPS	name	n	factor PHANTOM	Section 6.2.13