Bioinformatics and Biometrics for the Australian Grains Industry

CAIGE Pedigree Data Curation

GRDC Project No: UOS1901-002RTX

Ky Mathews[1], Robin Wilson[2]
email:kmathews@uow.edu.au

email:r.wilson@integratedbreeding.net

[1]Centre for Bioinformatics and Biometrics
National Institute for Applied Statistics and Research Australia
School of Mathematics and Applied Statistics
University of Wollongong

[2]Integrated Breeding Platform

August 2, 2021

Contents

Confidentiality and IP Statement

Acknowledgements: Thanks to the following people for assistance with the pedigree curation task: William Eusebio (IBP), Rowena Tulod (IBP), Clarissa Pimental (IBP), Dr. Filippo Bassi (ICARDA), Dr. Miguel Sanchez Garcia (ICARDA), Dr. Wuletaw Tadesse (ICARDA), Dr. Ravi Singh (CIMMYT), Dr. Karim Ammar (CIMMYT) and to Dr. Julie Nicol for co-ordinating.

1 Background

The CIMMYT Australian ICARDA Germplasm Evaluation project (CAIGE) is a Grains Research and Development Corporation (GRDC) funded project (UOS1901-002RTX) which commenced in 2005 and is managed by the University of Sydney in collaboration with the University of Queensland and Australian wheat, durum and barley breeders. From 2015 onwards statistical support has been provided by the University of Wollongong (UOW) employed on the GRDC projects: Statistics for the Australian Grains Industry (SAIG2) and Bioinformatics and Biometrics for the Australian Grains Industry (BBAGI) (UW00009). The key objective of CAIGE is to evaluate germplasm developed by CIMMYT (International Maize and Wheat Improvement Centre) and ICARDA (International Centre for Agricultural Research in the Dry Areas) for potential introgression into Australian bread and durum wheat and barley breeding programs.

Entries (varieties) are evaluated at different sites across the Australian wheat belt and selected by breeding companies to be included in their breeding programs and ultimately released to Australian wheat growers. Each year a new cohort of entries is evaluated in a multi-environment trial (MET) series. The majority of entries are from the CIMMYT and ICARDA wheat breeding programs and are related to other entries. In addition, entries from Resource Seeds International (RSI) led by Dr. Sanjaya Rajaram, who headed CIMMYT's global wheat breeding program and was also director of Director of Integrated Gene Management at ICARDA in the early 2000's, are also evaluated via the CAIGE project. Due to Dr. Rajaram's knowledge and experience the RSI material is somewhat related to both the CIMMYT and ICARDA germplasm.

The relationship between entries, pedigree information, is available in the CAIGE Breeding Management System (BMS) database, (http://www.caigeproject.org.au/breeding-management-system). The inclusion of pedigree information enhances the analysis of the CAIGE dataset in two ways. There is a significant lack of entry connectivity between years in the CAIGE datasets, less so for barley, and until now the CAIGE MET analyses have been within year only. However, it is well known that the year-to-year variability in Australia is large and selecting entries for use as parents or direct release based on a multi-year analysis is advised. Pedigree information enables an across year analysis through the ancestral relationships across years as the incoming germplasm from CIMMYT and ICARDA come from their breeding programs which have a high level of inbreeding from year to year. Further, the use of pedigree information in the analysis allows for the partitioning of the total genetic (variety) effects into additive and non-additive genetic effects. The additive genetic effects are equivalent to estimating the general combining ability of a line and are the most appropriate effects to use when selecting entries to use as parents in a breeding program. This meets a key aim of the CAIGE project - the identification of lines for use as parents and introgression into Australian breeding programs. Furthermore, the total genetic effects can be estimated from this analysis and can be used to identify entries that could potentially be directly released to Australian growers - subject to meeting disease and quality criteria.

A factor analytic linear mixed model (FA-LMM) framework is used for the analysis where the term Variety represents the entries/lines/varieties under evaluation and the term Environment represents a unique year-location combination. A one-stage approach, following Smith et al. (2001); Gogel et al. (2018), allows for the inherent imbalance in the dataset, individual trial (usually environment) spatial modelling and appropriate modelling of the variety by environment interaction (VEI). In addition, the genetic variance-covariance matrix is modelled using the numerator relationship matrix derived from the pedigree information (Oakey et al., 2006, 2007; Burgueno et al., 2007; Cullis et al., 2010).

In the CAIGE datasets pedigree data consists of passport information which describes and identifies each entry, selection history which indicates the cross type, level of selfing and bulking and the pedigree strings which describe the genealogy of an entry. Good quality pedigree data is a cornerstone of the FA-LMM. Despite pedigree information being fundamental to successful breeding program it is frequently not routinely curated to a standard sufficient for use in statistical models. A key reason for this is the data is not usually interrogated to the level required before generating a numerator relationship matrix and the data is rarely collated across years which enables inconsistencies in naming to be detected.

The purpose of this report is to document the process and reasoning for curating the bread, durum and barley pedigree data in 2021. The curated datasets were 2013-2021 for bread wheat, 2016-2021 for durum wheat and 2015-2021 for barley. This report is organised as follows: the passport data is described in Section 2, the selection history curation is described in Section 3, the process of curating the pedigree strings are described in Section 4 and recommendations presented in Section 5.

2 Passport Data

The genetic material consists of CIMMYT and ICARDA late stage breeding lines, Resource Seeds International breeding lines and Australian checks. The non-Australian lines are generally those ready to be deployed in international nurseries or have been selected by Australian breeders at CIMMYT and/or ICARDA breeding locations. Each year a new cohort of entries is released from quarantine and is evaluated in yield trials in Australian environments. There are 10-14 Australian released varieties included for each crop and these change over time with a core set retained over years.

Each entry has passport information which helps to uniquely identify it. This passport information includes the Quarantine Code (QCode) for the international nursery and Quarantine Number (QNo), the entry number within a nursery; EntryCD the concatenation of the QCode and QNo, such that QNo:QCode; Genotype Identifier (GID) from the CIMMYT/ICARDA databases or the CAIGE Breeding Management System (BMS); Cross Id (CID) and Sister Id (SID) from the CIMMYT IWIS2 database; accession identifier (acc_id) from the Australian Grain Genebank; the germplasm source (Source =

CIMMYT, CIMMYT-TURKEY, ICARDA, RSI, Australia), the pedigree string and selection history.

GID is assigned based on the selection history. For example, BORLAUG 100 had selection history CMSS06Y00605T-099TOPM-099Y-099ZTM-099Y-099M-11WGY-0B when introduced as 208:ZWB12 and was assigned a GID of 6176013. More recently, it was included as standard check in both the 6th Wheat Yield Consortium (ZWY19) and 6th Stress Adaptive Trait (ZSA19) CIMMYT nurseries with a new selection history CMSS06Y00605T-099TOPM-099Y-099ZTM-099Y-099M-11WGY-0B-0MEX to reflect a re-selection event. A new GID of 7806808 was assigned to these entries, 30:ZWY19 and 29:ZSA19. In this example, entries with GID 6176013 and 7806808 were considered duplicates and called BORLAUG 100. Each decision of this kind is made in collaboration with CAIGE staff.

Logic dictates that the GID could be used as the Variety field in both designs and analyses. This addresses concerns raised with Dr. Ky Mathews by collaborators that across years there may be duplicate entries (as described above for BORLAUG 100). However, GID as a 7 or 10 digit number does not provide collaborators with as much information as the EntryCD (QNo:QCode). For example, ZSA19 is known to be from CIMMYT's 6th Stress Adaptive Trait nursery which conveys extra information about the potential adaptation of entries in this nursery compared to entries in the 6th Wheat Yield Consortium (ZWY19) nursery. It is important to understand that QNo is nested within QCode, so an entry with QNo=1 with QCode=ZWB20 is different from an entry with QNo=1 and QCode=ZIF20. If the line is a check line from CIMMYT or ICARDA then the QNo:QCode designation is replaced with the check name, e.g. the ICARDA check, TERBOL, was introduced as both 1:ZIF19 and 198:ZIZ18. They both have the same GID and the name used in the Variety field was TERBOL.

The first curation step is to identify GID with multiple EntryCD and determine if they are true duplicates. This requires consultation with CAIGE staff, database managers and potentially CIMMYT or ICARDA breeders. In the first across year analysis of the CAIGE bread wheat dataset 2013-2019 documented in Mathews et al. (2020), GID was used as the key field for Variety in the variety by environment analysis. This resulted in curious results and a subsequent analysis of phenology and height data showed that some entries considered as duplicates were in fact not duplicates but probably entries re-selected for phenology and height.

There are three scenarios where duplicates may arise in these datasets:

1. Entries within the same nursery are not likely to be duplicate entries, e.g. in ZIG19, a nursery from Dr. Rajaram (RSI) entries 2, 3, 41 and 42 all had the same selection history, for some reason not clear from the dataset, entries 3 and 42 were assigned GID 200005302 and entries 2 and 43 GID 200005303. The phenology and height data showed that these entries should not be considered duplicates and the likely scenario is that they are re-selections for these traits but the selection histories have not been updated.

2. There is a flow of germplasm from CIMMYT Mexico to CIMMYT-Turkey where material are screened for soil borne pathogen tolerance/resistance. Thus, an entry can be sent to Australia from Mexico in one year and also sent to CIMMYT-TURKEY. Then, in a subsequent year it is sent to Australia from CIMMYT-TURKEY for its soil borne pathogen characteristics. An example of this is 92:ZWB13 which was evaluated in CAIGE yield trials in 2014 and then again as 29:SBP16 in 2018. If there is a re-selection event in CIMMYT-Turkey then ideally this should be added to the selection history and the GID should be changed. This scenario is possible in both the bread and durum wheat datasets.

3. CIMMYT Mexico distributes international nurseries for different purposes, as described above. Hence, there are sometimes duplicates across these nurseries which are ideally identified during quarantine and not included in yield trial entry lists. This process continues to be refined and thus for historical datasets a check for duplicates using GID and/or selection histories is recommended.

It is highly recommended that a biometrician determines duplicate lines in consultation with CAIGE staff. Phenology and height data are useful in identifying if entries are true duplicates.

The pedigree information exported out of the CAIGE BMS is converted to a data frame in the so-called “Me Mum Dad” format which is described in the R `pedicure` library (Butler, 2019) as a *data frame with (at least) three columns that correspond to the individual, female parent and male parent, respectively. The row giving the pedigree of an individual must appear before any row where that individual appears as a parent. Founders use 0 (zero) or NA in the parental columns.* The example presented in the 2021 CAIGE Bread Wheat Design report (Mathews et al., 2021) is repeated here for illustration (Table 1). The individual in the last row, 171:ZIZ20, has a ‘Mum’ (female parent) with pedigree SOONOT-10/HUBARA-15 and a ‘Dad’ (male parent) of JAWAHIR-14. These parents are in rows 10 and 8, respectively, in this table, *above* the individual that will be tested in the field.

The derivation of the values in columns four and five will be described in Section 3.

Table 1: Example of a *Me Mum Dad* file for the CAIGE Bread Wheat dataset. Me is the individual being tested or an ancestor of an entry being tested, Mum and Dad contain the name of the parents. A value of zero for Mum or Dad indicates unknown pedigree or base individuals. Fn is the filial generation number and is a single number for single plant selections (SPP) or of the form $n_b : n_f$ indicating the level of SPP and number of bulked generations. $\mathcal{F}$ is the inbreeding coefficient.

Me	Mum	Dad	Fn	$\mathcal{F}$
KAUZ'S'	0	0	6	0.96875
CHAM-4	FLICKER	HORK	7	0.98438
SHUHA'S'	0	0	6	0.96875
SHUHA-4	SHI4414	FCR/3/MCM/KT//Y50/4/ZA75/CM5287-J	6	0.96875
SAMAR-8	0	0	7	0.98438
HUBARA-15	FLORKWA-2	KAUZ	6	0.96875
JAWAHIR-14	SHUHA-4	NS-732/HERMOSILLO M77	7	0.98438
SOONOT-10	SAMAR-8/KAUZ'S'	CHAM-4/SHUHA'S'	7	0.98438
SOONOT-10/HUBARA-15	SOONOT-10	HUBARA-15	7	0.98438
171:ZIZ20	SOONOT-10/HUBARA-15	JAWAHIR-14	3:4	0.87500

3 Selection History

The selection history is a character string which documents the cross type, number of bulking and selfing events undertaken to develop an entry. CIMMYT and ICARDA have similar selection history naming protocols. A summary is provided here but readers are referred to the document *ICARDA-Standard pedigree and selection history codes- Dec 2020.docx* provided by ICARDA's durum breeder, Dr. Filippo Bassi (F.Bassi@cgiar.org) which provides a detailed explanation of the ICARDA durum breeding program selection histories. We anticipate that there are similar documents for the other breeding programs.

An example of an ICARDA durum wheat selection history is:
ICD11-129-0TR-7STR-0TR-1TR-0STR-5TR-0STR-0AUB-0AUB

An example of an CIMMYT bread wheat selection history is:
CMSS09Y00016S-099Y-099M-099Y-5WGY-0B

An example of an ICARDA barley selection history is:
ICM1213CJ24-3CJ-010CH-05CJ-1CH-0MR

The key concepts to understand about these selection histories are:

1. '–' separate each different selection event;

2. the first combination of letters represents the cross identifier and includes the institution (IC for ICARDA, C for CIMMYT or PT for CIMMYT physiology), the crop type D for durum, B for barley, M for Main wheat breeding program at CIMMYT, the year of crossing in YY format and sometimes information about the cross-type, e.g. T for top cross in CIMMYT durum or F2 in ICARDA durum;

3. for ICARDA entries the number after the first '–' is the cross number. For CIMMYT entries this cross number is incorporated into the first combination of letters;

4. each step of selection includes a number indicating the advancement strategy and two-three letters indicating the field station where the selection occurred;

5. values starting with zero, e.g. 0, 015, 020, 099 indicate a bulking event. Other integers, e.g. 5, indicate a single plant selection of the 5th plant.

These selection histories are used to determine the filial generation number, Fn, of each entry (fourth column of Table 1). This can take the form of an integer (1 = simple cross, 2 indicates the progeny of this cross selfed once, 3 is selfed twice, and so on). If the breeding program uses bulk selections before a single plant selection then Fn takes the form $n_b : n_f$, where n_b is the Fn of the single plant selection and n_f indicates the final generation of bulking. The difference between n_f and n_b is the number of generations of bulking. For example, an $F_{2:5}$ line, F_2 derived F_5 line, would have an Fn of 2:5, indicating a single plant selection at F2 followed by 3 generations of bulking. The fifth column self of Table 1 represents the level of selfing where

$$self = \begin{cases} Fn - 1, & \text{where } Fn \text{ is an integer} \\ n_f - 1, & \text{otherwise} \end{cases} \tag{1}$$

Jordan & Cullis (2020) have developed an improved algorithm for calculating the inbreeding coefficient of bulk selected lines. However, this was not successfully implemented for the CAIGE dataset. The CIMMYT and ICARDA breeders believe the inbreeding values to be higher than the values obtained from this algorithm and hence an approximation of $n_f - 1$ was used for these lines. The algorithm developed by Jordan & Cullis (2020) relies on the coefficient of parentage, a_{SD}, between any two entries to be available and accurate. However, due to the sparse nature of the CAIGE pedigree datasets (i.e. not all ancestral pathways are fully documented or connected) there is a high occurrence of low a_{SD} values and this leads to lower than expected inbreeding coefficients.

Australian checks are considered fixed and assigned a Fn=8, whilst all parents and grandparents are assigned a Fn=6.

A summary of the resulting selfing values by genetic sources for each crop (Table 2) shows that there is considerable variation within each source for each crop. For example, the selfing values ranging from 2 to 8 for ICARDA barley, 4 to 12 for CIMMYT bread wheat and 4 to 8 for CIMMYT-TURKEY. A simplified approach to estimating the inbreeding coefficients would be to nominate a single, or average selfing value, to each source for each crop. However, the results in Table 2 clearly show that this would lead to a loss of information and accuracy in the estimation of the inbreeding coefficients and subsequent coefficients of ancestry used in the numerator relationship matrix.

Table 2: Summary of the self values for the 4522 entries by genetic source (Australian, CIMMYT, CIMMYT-TURKEY, ICARDA and Resource Seeds International) and crop in the CAIGE bread wheat, durum wheat and barley datasets analysed in 2021. Dashed lines separate the crops.

Crop	Source	Min	Mean	Max
Barley	AUS-CHECK	7.0	7.0	7.0
Barley	ICARDA	2.0	4.7	8.0
Barley	ICARDA-CHECK	7.0	7.0	7.0
Bread	AUS-CHECK	7.0	7.0	7.0
Bread	CIMMYT	4.0	5.6	12.0
Bread	CIMMYT-CHECK	7.0	7.0	7.0
Bread	CIMMYT-TURKEY	4.0	5.1	9.0
Bread	ICARDA	3.0	5.0	11.0
Bread	ICARDA-CHECK	7.0	7.0	7.0
Bread	RSI	3.0	5.3	9.0
Durum	AUS	7.0	7.0	7.0
Durum	CIMMYT	4.0	5.2	6.0
Durum	CIMMYT-TURKEY	4.0	4.7	8.0
Durum	ICARDA	5.0	8.9	16.0

The inbreeding coefficient $\mathcal{F}$ (Henderson, 1976) for each record in the pedigree file is then calculated as

$$\mathcal{F} = 1 - \frac{1}{2}^{self} . \tag{2}$$

A summary of the resulting inbreeding coefficients by genetic sources for each crop (Table 3) shows that the Australian checks, which are fixed lines, have the highest level of inbreeding whilst material from ICARDA and RSI is lower than CIMMYT. This is expected as, in general, the CIMMYT material is selected at the F_4 or F_5 generations whereas the ICARDA and RSI lines are bulked from F_2 onward with single plant selections occurring in later generations.

The following data curation steps were necessary before calculating the self values and inbreeding coefficients. Essentially, these curation steps were required because the algorithm depend on splitting the selection history character string on '–'. If something impedes this process or causes incorrect processing then it needs to be addressed.

1. As noted above the ICARDA selection histories have a different protocol from the CIMMYT selection histories in that the ICARDA ones split the cross number from the breeding program-year designation by a '–' whereas CIMMYT do not. Thus, when processing a dataset which contains both ICARDA and CIMMYT selection histories it is necessary to convert this first '–' to a '.' to enable the algorithm which counts the number of self and bulking generations to be applied to both ICARDA

Table 3: Summary of the inbreeding coefficients for the 4522 entries by genetic source (Australian, CIMMYT, CIMMYT-TURKEY, ICARDA and Resource Seeds International) and crop in the CAIGE bread wheat, durum wheat and barley datasets analysed in 2021. Dashed lines separate the crops.

Crop	Source	Min	Mean	Max
Barley	AUS-CHECK	0.992	0.992	0.992
Barley	ICARDA	0.750	0.935	0.996
Barley	ICARDA-CHECK	0.992	0.992	0.992
Bread	AUS-CHECK	0.992	0.992	0.992
Bread	CIMMYT	0.938	0.975	1.000
Bread	CIMMYT-CHECK	0.992	0.992	0.992
Bread	CIMMYT-TURKEY	0.938	0.961	0.998
Bread	ICARDA	0.875	0.954	1.000
Bread	ICARDA-CHECK	0.992	0.992	0.992
Bread	RSI	0.875	0.944	0.998
Durum	AUS	0.992	0.992	0.992
Durum	CIMMYT	0.938	0.969	0.984
Durum	CIMMYT-TURKEY	0.938	0.955	0.996
Durum	ICARDA	0.969	0.994	1.000

and CIMMYT selection histories.

2. Not all ICARDA selection histories were in the format described above and they already had a '.' in place or no space at all, in line with the CIMMYT cross identifiers.

3. Identification of typographical errors in the place names - this is only necessary to fix when they contain '–' when they should not, or when the names makes the prefix number incomprehensible.

4 Pedigree Strings

A pedigree string for an entry describes its genealogy. The depth of the genealogy depends on what rules have been applied in the database which manages this information to help ease of reading. Long pedigree strings, more than 100 characters, say, are difficult to read and in the past have been accidentally truncated to conform to software constraints (e.g. 255 characters in Excel spreadsheet cells).

CIMMYT and ICARDA breeding programs follow the Purdy method (Purdy et al., 1968) and a concise description is available here https://cropforge.github.io/iciswiki/articles/w/h/e/Wheat_Pedigree_ecd4.html. For illustration some theoretical pedigrees are provided in Table 4. A forward slash, '/' is used to indicate a simple primary cross, two forward slashes '//' indicate a secondary simple cross, 'n' represents n crosses and an asterisk, '*', represents a back cross where "if the asterisk is on the left side of the simple cross symbol, then the back cross parent is the female. If the asterisk is on the right side of the simple cross symbol, then the back cross parent is male" (from cropforge ICIS WIKI page above).

Table 4: Theoretical pedigrees showing (Purdy et al., 1968) protocols used in CIMMYT and ICARDA pedigrees for simple and back cross pedigrees.

Simple	A/B	A	B
Simple	A/B//C	A/B	C
Simple	A/B//C/3/D	A/B//C	D
Backcross to Mum	A*2/B	A	A/B
Backcross to Dad	A/2*B	A/B	B

The only pedigree string required to generate a 'Me Mum Dad' file and subsequently the numerator relationship matrix is that of the individual being tested, called Pedigree.Me in the CAIGE Master Entry Lists. An R (R Core Team, 2020) function written by Dr. Ky Mathews and inspired by an Excel macro developed by William Eusebio and Robin Wilson (Integrated Breeding Platform), splits a pedigree string following the rules in Table 4 such that a "Me Mum Dad" file as presented in Table 1 can be generated for use in the analysis.

In practice, the pedigree strings for entries (Pedigree.Me), parents (Pedigree.Mum and Pedigree.Dad) and grandparents (Pedigree.MM = MumsMum, Pedigree.MD = MumsDad, Pedigree.DM = DadsMum and Pedigree.DD = DadsDad) were extracted by CAIGE staff from the CAIGE BMS database along with their respective GID. Both the GID and the pedigree strings require curation as it was often found that a GID was not correct, notably the swapping of mum and dad pedigrees at the grandparent level. Curating the parent and grandparent pedigree strings, and therefore GID, is not strictly necessary to generate a numerator relationship matrix but it does give great depth to the genealogy particularly when a mum is a released variety name.

In 2020 the pedigree analysis was based on the unique Genotype Identifier (GID) number stored in BMS to each breeding line. However, this approach only allowed for the pedigree depth to be back to grandparents. Whilst there is some uncertainty about how far back in the ancestral tree it is required for accurate estimation of the coefficient of parentage it is reasonable to assume that the greater the depth the more accurate the estimate. A combination of GID being unavailable for a considerable number of parents and grandparents for the durum and barley datasets and the realisation that by focusing on GID only for the parents and grandparents there was a loss of information led to large task of curating the pedigree strings for all crops being undertaken.

This data curation task involves:

1. Converting all text to uppercase so that discrepancies between lower and upper case are avoided.

2. Identifying and correcting typographical errors, e.g. removing spaces around '/'.

3. Identifying synonyms and ensuring consistency in naming, e.g. CHAM4, CHAM-4, CHAM_4 all converted to CHAM-4. Note this is true for single instances of CHAM-4 and within a pedigree string. Identifying synonyms is one of the most onerous steps in the data curation and requires care and collaboration with CAIGE staff and CIMMYT and ICARDA breeders. A biometrician should **never** make a decision that two names are synonyms without confirming with CAIGE staff as these pedigrees and their naming systems are complex. Dr. Ky Mathews has observed that the BMS database (or its precursors) use an abbreviated synonym for a released variety if it is within a pedigree string but the full name if it is on its own. See Table 5 for the example of ROELFS F2007.

4. Identifying when a pedigree string has multiple records in the 'Me Mum Dad' file due to different mums and dads. See the discussion on recurrent back crosses below.

5. Identifying when a record has no parents but in fact the parents are known because a synonym has the parents.

4.1 Recurrent back cross notation

The notation for recurrent back crosses requires special attention as it is not consistent within the data extracted from the database. It is important to note that the description given by Purdy et al. (1968) and detailed on the cropforge ICIS WIKI page (see above) for generating recurrent back crosses described in Table 4 is not always followed by the data in the database. The key discrepancy is that, based on GID, the A*2/B situation is A is the Dad (not the mum, as expected) and A/B is the mum (not the dad, as expected). An example is provided for EntryName 35:ZWB14 which is a recurrent back cross of ROELFS F2007 to ROELFS F2007/PAURAQ. Both Purdy et al. (1968) and crop forge ICIS WIKI define the recurrent parent to be the one with the numeral and asterisk on it. By this definition, ROELFS F2007 is the Mum. However, according to the GID defined in the database ROELFS F2007 is the Dad, Table 5.

Table 5: Genealogy back to grandparents for EntryName 35:ZWB14. The horizontal dashed line delineates between the maternal and paternal ancestors.

Me	6568360	ROLF07*2/PAURAQ
Mum	5848260	ROLF07/PAURAQ
MumsMum	4905617	ROELFS F2007
MumsDad	5398471	PAURAQ
Dad	4905617	ROELFS F2007
DadsMum	3551693	TACUPETO F2001/KUKUNA
DadsDad	2490771	TACUPETO F2001/BRAMBLING

A key anomaly noted in the bread wheat dataset in particular is that the rule is implemented differently when a recurrent back cross is a Mum compared to when it is a Dad. For example,

- A*2/B as Mum

- Mum's mum is A
- Mum's dad is A/B

But when

- A*2/B is a Dad, then
- Dad's mum is A/B and
- Dad's dad is A

Examples, of this are ATTILA*2/PBW65, ATTILA*2/CROW.

In addition, data curation problems occur when the recurrent back cross is further up the pedigree tree, say at grandparent level. In that situation, pedigree.cut() is used to split the grandparent pedigree string (Pedigree.MM, say) back to single varieties and hence in the 'Me Mum Dad' file two records for the recurrent back cross will occur and these requires fixing before the numerator relationship matrix can be generated. Table 6 illustrates this problem with the first record arising from the database (Table 5) and the second from pedigree.cut().

Table 6: Example of problems with inconsitent back cross nomenclature from IWIS2/ICIS and now in BMS.

Me	Mum	Dad
ROLF07*2/PAURAQ	ROLF07/PAURAQ	ROELFS F2007
ROLF07*2/PAURAQ	ROLF07	ROLF07/PAURAQ

The presentation, titled *Pedigrees wrongly uploaded from Fibos to PMS* by Jesper Nørgard Welen (International Wheat Information Systems, IWIS, developer) in 2012 (https://slideplayer.com/slide/2603351/) shows how a decision at CIMMYT to manage back cross notation when moving from IWIS2 to IWIS3 may have created this problem. In practice, an incorrect allocation of parents to mum or dad will not change the calculation of the coefficient of ancestry between them. Hence, the pedigree.cut() function follows the theory presented in Purdy et al. (1968) and has not been modified for CIMMYT pedigrees. Instead, Robin Wilson assisted in making the decision about which way the recurrent back cross should be and in this he was guided by the GID value in the database. He is confident that the decisions made in this curation process about which way round the parents are for this scenario is correct as parents in different crosses that have the same GID then have the same pedigree once corrections were made. Dr. Ky Mathews suggests that future curation activities ensure that the breeders are involved in checking the order of recurrent crosses to ensure that the database is managing this complex scenario appropriately.

One complication in the CAIGE pedigree datasets is that released lines are sometimes named after an institute with different selections having different numbers. One such example is in the "Me Mum Dad" file (Table 1) where SHUHA-4 and SHUHA'S' are present as *Me's* and have different *Mum* and *Dad* information. The parents of SHUHA-4 are known but of SHUHA'S' are unknown. If the pedigrees for these lines were included in the dataset then the coefficient of parentage between these two entries could be estimated. Currently this is not the case. The CHAM group is another example of this naming system. However, entries such as SERI'S' and SERI M82 do have the same parents. Again, clear, collaborative communication with CAIGE staff and CIMMYT and ICARDA breeders is required to clarify these scenarios.

5 Recommendations

The following recommendations are made regarding the curation of CAIGE pedigree data:

1. Data curation of pedigree files is a time consuming, iterative process. This task was only successful because Robin Wilson and Ky Mathews worked intensively for 4-6 months to curate these pedigree files and the relevant breeders were responsive and collaborative in assisting with queires. The combined skills of breeding knowledge and data management skills and joint tenacity resulted in pedigree files which are well curated - although there is always more to curate and check. This type of task requires investment into people with the appropriate skills - leaving it to untrained PhD students is a recipe for disaster.

2. CIMMYT and ICARDA breeders and their institutional hierarchy need to recognise the importance of maintaining pedigree records and invest in someone (or two) to assist with this task.

3. A BBAGI biometrician should only have to minimally curate the pedigree file. However, it is recognised that it is unlikely for pedigree files to be thoroughly inspected for synonyms, textual inconsistencies and parent swapping by personnel who do not have tools such as the `R pedicure` library available from the Mixed Models and Design of Experiments webpage (www.mmade.org).

4. BBAGI biometricians should only curate entries that are intended for evaluation in yield trials, not all all entries that enter quarantine.

5. Duplicate entries should be identified at quarantine so CAIGE staff can decide if an entry should be re-evaluated in Australian yield trials.

6. The recurrent back cross anomalies need addressing with the CIMMYT and ICARDA breeders. It is noted that these anomalies will not affect the calculation of the coefficient of ancestry.

7. Research is recommended to determine how many generations back is required for appropriate estimation of the inbreeding coefficient.

8. Research is recommended to determine how the Jordan & Cullis (2020) work to estimate inbreeding coefficients for bulked lines could be implemented in a dataset with varying levels of ancestry.

9. A comprehensive list of location codes used in CIMMYT and ICARDA selection histories could be collated. It is not important to the yield trials but could assist with data curation.

10. A comprehensive list of the full variety names and their synonyms could be extracted from the BMS database to assist with data curation.

11. The fixes and changes made to the bread, durum wheat and barley pedigree datasets and documented in the *Master Entry List* Excel files in 2021 be uploaded into BMS and checked by appropriate CIMMYT and ICARDA personnel.

6 Resources

This section is primarily as information for the author and team members of BBAGI.

The folder for this work is found on the BBAGI Dropbox account under `UOW-EIS-NIASRA-BBAGI/Projects/CAIGE/CAIGE20/`*`crop`*`/Analysis/1-Pedigree`, where *crop* is Bread wheat, Durum or Barley.

The scripts are located within the `1Scripts` folder and are broadly described below:

- SelectHistoryProcess.R - script curates the selection histories, determines the number of generations of bulks and selfing and calculates the inbreeding coefficient for each entry.

- PedigreeCurate.R - script which curates the pedigree strings. There can be multiple files (labelled alphabetically and by date) which addresses the iterative nature of data curation

- AmatrixCalculation.R - the numerator relationship matrix,$\boldsymbol{A}$ is generated.

The scripts for this report are found in
`UOW-EIS-NIASRA-BBAGI/Projects/CAIGE/CAIGE20/PedigreeCurationReport`.

References

Burgueno, J., Crossa, J., Cornelius, P. L., McLaren, G., Trethowan, R. & Krishnamachari, A. (2007). Modeling additve x environment and additive x additve x environment using genetic covariances of relatives of wheat genotypes. *Crop Science* **47**, 311–320.

Butler, D. (2019). *pedicure: pedigree tools*. www.mmade.org. R package version 2.0.0.

Cullis, B., Smith, A., Beeck, C. & Cowling, W. (2010). Analysis of yield and oil from a series of canola breeding trials. Part II. Exploring variety by environment interaction using factor analysisThis article is one of a selection of papers from the conference "Exploiting Genome-wide Association in Oilseed B. *Genome* **53**, 1002–1016.

Gogel, B., Smith, A. & Cullis, B. (2018). Comparison of a one- and two-stage mixed model analysis of Australia' s National Variety Trial Southern Region wheat data. *Euphytica* **214**, 1–21.

Henderson, C. R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* , 69–83.

Jordan, M. & Cullis, B. (2020). The Numerator Relationship Matrix.

Mathews, K., Nicol, J. & Trethowan, R. (2021). CAIGE Bread Wheat Designs 2021. Tech. rep., BBAGI, University of Wollongong.

Mathews, K. L., Baranawal, D. & Nicol, J. (2020). Caige bread wheat: Multi-year analysis report 2013-2019. Tech. rep., BBAGI, University of Wollongong.

Oakey, H., Verbyla, A., Pitchford, W., Cullis, B. R. & Kuchel, H. (2006). Joint Modelling of Additive and Non-Additive Genetic Line Effects in Single Field Trials. *Theoretical and Applied Genetics* **113**, 809–819.

Oakey, H., Verbyla, A. P., Cullis, B. R., Wei, X. & Pitchford, W. S. (2007). Joint modeling of additive and non-additive (genetic line) effects in multi-environment trials. *Theoretical and Applied Genetics* **114**, 1319–1332.

Purdy, L. H., Loegering, W., Konzak, C., Peterson, C. & Allan, R. (1968). A proposed standard method for illustrating pedigrees of small grain varieties. *Crop Science* **8**, 405–406.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Smith, A., Cullis, B. & Thompson, R. (2001). Analyzing Variety by Environment Data Using Multiplicative Mixed Models and Adjustments for Spatial Field Trend. *Biometrics* **57**, 1138–1147.