PIGBLUP Version 5.20: Flexible traits and data input

Tony Henzell and Ron Crump

Animal Genetics and Breeding Unit, a joint venture of NSW Department of Primary Industries and The University of New England.

Introduction

This paper briefly describes innovations in the forthcoming release of PIGBLUP, Version 5.20. This version of PIGBLUP will be circulated to all currently supported licensees in January of 2005.

The major areas of development for PIGBLUP Version 5.20 have been in the post-analysis use of molecular genetic information, increasingly generic data file reading and, in the background, more general handling of trait definitions.

Construction of PIGBLUP data files from generic data files that may have originated from multiple sources is of most immediate interest to new and potential PIGBLUP clients. The developments necessary to bring this about are an essential part of making PIGBLUP a more general, powerful and easier to upgrade program in the future. All clients in future versions will receive the benefits of the programming efforts that have gone into PIGBLUP Version 5.20.

When PIGBLUP was first developed in the late 1980s constraints on computer power and capacity meant that the system had to be particularly lean and efficient. Along with the desire to maintain ease of use for non-geneticists, this led to the program being restricted to a very specific set of traits and models. As time goes by interest in more demanding models, such as those incorporating molecular data in the BLUP evaluations and sow survival, and novel traits increases. The generalisation of the PIGBLUP package will allow a greater variety of analyses and more rapid future development of additional analyses.

Release 5.20

The current release, PIGBLUP Version 5.10, has had very few reported errors since its release. Most reported errors have derived from failure to identify erroneous data. Additional checks have been added to PIGCHECK and version 5.20 of PIGBLUP to detect and report these conditions.

The calculation of accuracies introduced in version 5.10 made PIGBLUP more sensitive to inconsistencies in littermates pedigree information. PIGCHECK's littermate pedigree correction logic has been enhanced. Display of littermate pedigree errors has been made easier to understand. The corrections that PIGCHECK infers are submitted for approval or modification by the user and appear to be very robust over a range of tested data sets.

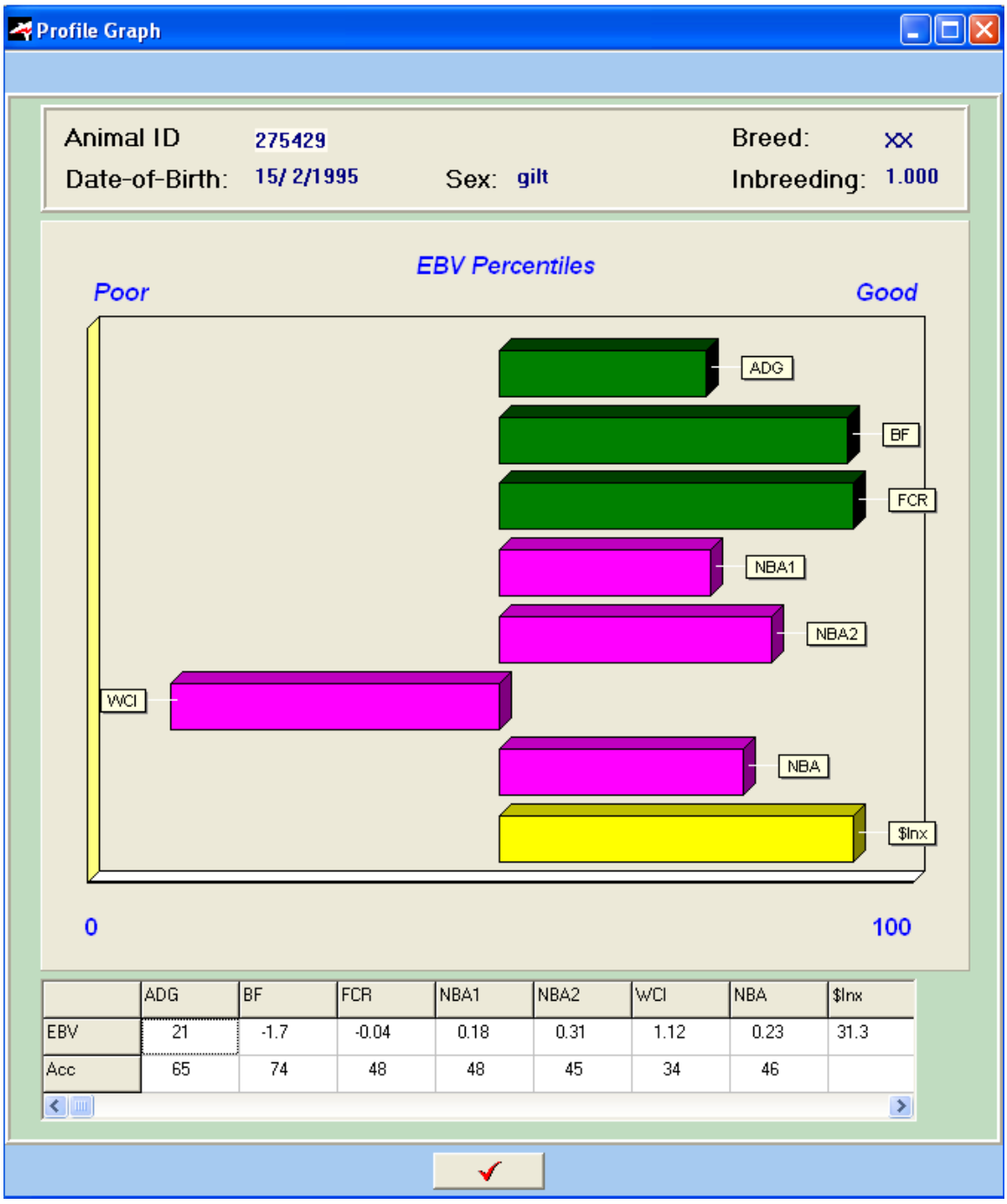
1. Markers

As the availability of molecular genetic information increases, it is essential that breeders have the tools to allow them to make appropriate use of it. To this end, a new module has been developed that enables data on marker genotypes to be combined with PIGBLUP EBVs to provide a good predictor of the overall genetic merit. Crump and Tier (these proceedings) cover this development in detail.

2. Pig Performance Graphic

A new graphic is available on the View EBVs screen. This graphic is enabled when 'All Sexes' is selected and no restrictions (such as the $INDEX being greater than a specified value) are applied. This is necessary so that statistics and percentiles can be calculated for the whole herd/breed.

On selecting a pig, a graphic appears that provides a visual comparison of the animal's EBVs and Indexes against all other animals in the herd/breed together with information on its inbreeding, date-of-birth, breed and sex.



3. Extra Carcase Trait

Another carcase trait slot was added using the current structures in PIGBLUP. This was done while determining ways to make trait definition more flexible and easier for clients to setup – and to meet the need of one of our clients.

4. Towards a Generic Genetic Analysis System

Considerable effort has been expended during the last year on the tandem problems of building a data input file creation system, defining the structures needed for client-defined traits and their analysis, and implementing the specification and editing interfaces.

Data file generation has been developed in tandem with trait definition because almost all the structures are common to the two systems.

Both data input file creation and client-defined traits have been designed to hide as much as possible of the complexity inherent in these tasks while minimising input of erroneous, inconsistent or incomplete information.

Current users will be able to continue with existing data input files and analyses; our intention is to minimise disturbance to current users. For example, all current data formats, trait definitions and analyses will have 'ready to go' specifications in place on installation.

The first version of the standalone data input file creation program will accompany PIGBLUP Version 5.20 while client-defined traits are scheduled for the following version.

5. Data Input File Creation

The purpose of this system is to minimise the time taken and effort required in massaging a new or prospective client's recorded herd data into one of the standard PIGBLUP data file formats.

Herd data will often exist in the form of industry standard databases such as Access or spreadsheets such as Excel. Rather than try interfacing with all possible database systems, it is assumed that the client will have exported his data base files into plain text files.

Commonly, the fields in such exported files are either enclosed in quotes and separated from other fields by commas or are in fixed position and width columns. Both forms are catered for. In addition, a mix of files of one format or the other may be processed together to generate a PIGBLUP compatible data file.

Common to the two systems is the ability to define files as being composed of records and records as being composed of fields. Files, records and fields may exist in a number of forms each imbued with particular properties. For example, there are *PIGBLUP-compatible* and *Non-PIGBLUP* files. PIGBLUP files can only contain *Animal*, *Production* (includes Carcase and IGF1) and *Reproduction* records. Date fields may be of one of the forms *ddmmyy*, *ddmmyyyy*, *mmddyy* or *mmddyyyy* and thus have an implicit width of 6, 8, 6 or 8 characters respectively. Dates may be specified as *Birth*, *Slaughter*, *Farrowing* dates etc. As a consequence, the system can enforce the constraint that there be exactly one Birth date, for example. ID fields may be *Animal*, *Dam*, *Sire*, *Service Sire* etc IDs – each imposing constraints on usage within the system. For example, sex is implicit in *Dam*, *Sire* and *Service Sire* ID fields. This in-built

information helps minimise the amount of input required by the user as well as providing lots of checking opportunities during input.

Two benefits flow from having in-built semantic information. Firstly, the user only needs to pick the type and subtype of a field from a list and is freed from having to provide or understand the semantic information. Secondly, the inherited information is used to hide all but the options that are still meaningful for the field. In effect, the user is provided a short and narrow trail to follow as he defines files, records and fields.

Some of the capabilities provided are:

- Predefined specifications for each standard PIGBLUP data format. These are non-editable specifications. However, a user may copy any of these specifications and add or replace fields or study them as exemplars.

- PIGBLUP record types contain generic and non-generic fields. *Generic* fields are those that must exist for PIGBLUP to be able to perform an analysis. For example, the PIGBLUP ANIMAL record must contain Ids for the animal and its sire and dam, a sex field and a date-of-birth date field. Also, it is possible to specify that a minimum or maximum number of instances of a field exist or that particular fields must have valid values for the animal to be analysed.

- Calculated fields may be defined. Expressions of almost arbitrary complexity using other fields in a file can be defined. For example, it can be used to calculate daily gain based on a date-of-birth, dates on and off test and weights on entering and exiting testing. Calculations are provided in both data file creation and trait definition. That is, a trait's value may be obtained from a calculated field.

- When a calculation capability is provided, it is necessary to ensure there are no circular dependencies between variables (e.g. field A is used to calculate B, B is used to calculate C, but C is used to calculate A). The system detects and advises of such occurrences and also sequences all calculations so that calculations are performed with valid fields.

- Input files do not need to be sorted prior to data merging. The system uses the provided information to read each file in the correct order so that all the information on a particular animal is gathered prior to selected fields being copied for output.

6. Trait Definition

The current plan is to provide flexible trait definition with the release of PIGBLUP following version 5.20. This version will initially replicate PIGBLUP's present traits and analysis capabilities – but in a flexible and editable form. There is a great deal of programming needed to replace present 'hard-wired' code with code that adapts to the user-defined traits.

The trait definition requirements were obtained initially by extracting PIGBLUP's present trait pre-processing code and abstracting its capabilities into a set of program structures. Data input file, data record, data field, trait and analysis specifications defined in terms of these structures are written to text files in a 'Specs' subdirectory.

The names and nature of each of these files are stored in directory text files. When the user selects a particular data file specification from the directory, the relevant structures are populated from settings saved in files the system knows to be associated with that specification. This minimises the amount a user needs to remember about the specifications. Furthermore, as traits are defined based on particular fields in a data file specification, only those trait sets associated with the selected data file specification become available for selection.

In order to meet the design requirement of minimising input of inconsistent information, the specification and editing interfaces display only those details that are consistent with previous inputs. Also, immediate feedback is provided when inputs fail error checks.

Users may need to interrupt file or trait definition for various reasons. In order that they not lose what work has already been done, it is possible to save incomplete specifications. Specification or editing can be resumed later. A consequence of this freedom is that incomplete specifications must be detectable and the incomplete details must be able to be identified. Incomplete specifications and details are flagged with special icons to assist the user.

Editing of specifications currently requires a password in case a client wishes to restrict who may make changes to the system.

Trait specification structures and editing interfaces appear stable. That is, no extra features have been found necessary for some months. The remaining stages are the addition of these structures to PIGBLUP, modifications of PIGBLUP's analysis routines to handle the new data input file specifications, and replacement of 'hard-coded' sections of the interface software – and lots of testing.

References

Crump, R., and B. Tier (2004). "Markers." *2004 Pig Genetics Workshop* **65**: pp. XX-YY.