Evaluating the suitability of subjectively defined base populations

Jan ten Napel[1*], Jérémie Vandenplas[1], Mario Calus[1], Ismo Strandén[2], Martin Lidauer[2] and Roel Veerkamp[1]

[1] Wageningen University & Research, P.O. Box 338, 6700 AH Wageningen, the Netherlands; [2] Natural Resources Institute Finland (Luke), Jokioinen, Finland; [*]jan.tennapel@wur.nl

Abstract

Defining multiple base populations for a genetic evaluation is subjective and there is generally no feedback on the suitability of the defined base populations prior to model validation. We present a number of statistics that can be used to evaluate defined base populations. Application of these statistics to one simulated and two practical datasets showed that in practical datasets the number of base animals per base population was in one case very low. In many cases the available genotype information for different base populations was coming from exactly the same genotyped animals. Both issues are likely to complicate genomic evaluation. A tool to estimate these statistics is available in the MiXBLUP software suite.

Introduction

Any pedigree contains founder animals with unknown parents. There may be other animals with unknown parents because of incorrectly or not recording parents at birth, or uncertainty about the sire in case of unobserved matings or use of pooled semen. We refer to these animals as base animals. In genetic evaluations, we generally assume that these base animals are sampled from an infinite population, so we consider them to be genetically unrelated. If base animals originate from various populations or different generations within the same population, we may define several base populations, based on an expected difference in genetic level, due to origin or selection history. Criteria often used for defining base populations are breed, line within breed, sex, years of birth and selection path (Quaas and Pollak, 1981; Mrode, 2014, p55), but the result is subjective. Legarra et al. (2015) suggested to use genomic information to take into account that base populations are finite in reality, and that genetic relationships may exist within and across base populations. Such related base populations are referred to as metafounders (MF). In contrast to MF, base populations that are assumed to be unrelated are referred to as unknown parent groups (UPG). Defining multiple base populations is aimed at minimizing bias in genetic trend and estimated breeding values of selection candidates and reducing the mean square error of the evaluation (Foulley et al., 1990). Whether the definition used is effective, depends on the true differences in genetic level and the precision with which genetic level of each base population is estimated. For MF, it also depends on the amount and quality of genomic data available to estimate genetic relationships within and between MF. In practice, the appropriateness of the defined base populations can only be assessed by extensive model validation. The aim of this paper is to present a number of base population statistics that can help to evaluate the suitability of defined base populations prior to any model validation.

Materials & Methods

Description of datasets. The base population statistics are illustrated with three example datasets (Table 1). Dataset A is a simulated set of data, which was used in the study of Van Grevenhof et al. (2019) that recommended the use of MF. Dataset B is a subset of a routine genetic evaluation and consists of purebred pigs of a single line, their crossbred progeny and ancestors of dams of crossbred pigs., In dataset B, base populations were defined by line and one or more years of birth. Dataset B had a history of poor convergence when using UPG. Dataset C is a set of a routine evaluation of trout using UPG, defined by breeding population. Each founder animal was assigned to a single base population in dataset A and B, but to two in dataset C.

Table 1. Descriptive statistics of example datasets.

	Dataset A	Dataset B	Dataset C
Lines + crosses, N	3 + 2	4 + 3	1 + 0
Animals in pedigree, N	51,500	421,511	385,200
Generations in pedigree, N	8	42	12
Animals genotyped, N	9,250	72,854	6,351

Base population statistics. We introduce three base population statistics, which all include the parameters c_{ij} and q_{ij}. We define a genomic base population which consists of all genotyped animals with a path of non-genotyped ancestors to a base population. Genotyped animals that are descendants of two genotyped parents are not included as they do not provide genotype information to a base population. Now c_{ij} quantifies the width of the path of non-genotyped ancestors to the base population and hence is the genomic contribution of the genomic base animal *j* to base population *i*. If c_{ij} is lower than 1, then animal *j* has genotyped ancestors. The parameter q_{ij} is the fraction relating the contribution of base population *i* to the total genetic value of the animal *j* and is an element of the Q matrix in QP transformation (Quaas and Pollak, 1981). If q_{ij} is lower than 1, then animal *j* is related to multiple base populations.

The first statistic quantifies the amount of genotype information available for each base population. Imputation of non-genotyped animals from genotyped descendants only using pedigree relationships causes loss of information, because the genotype information has to be distributed to two parents in every generation without genotypes. We therefore introduce equivalent number of base animals genotyped for a base population (*Neq_i*). It is calculated as:

$$Neq_i = \sum_{j=1}^{N} c_{ij} q_{ij} * (\tfrac{1}{2})^{gener_j} \quad (1)$$

where *N* is the number of animals in the genomic base population and $gener_j$ is the maximum number of generations between animal *j* and base population *i*.

The second statistic quantifies uniqueness of genotype information and is the extent to which the same genotype information was used for a pair of base populations, which we call auto-similarity of one base population to the other. To illustrate the concept, imagine 24 balls, 8 red, 8 blue and 8 orange. The balls are placed in two bowls, so the first bowl contains 5 red and 4 blue balls. The second contains the remaining 3 red, 4 blue and 8 orange. The similarity of bowl one to bowl two is 3 red + 4 blue over 9 balls is 0.78. The similarity of the second bowl to the first one is 3 red + 4 blue over 15 balls is 0.47. The colours in the illustration are genotyped animals, the bowls are base populations and the number of balls of the same colour in the same bowl is *q*. So auto-similarity of base population *i* to *j* ($AS_{i\ to\ j}$) is calculated as:

$$AS_{i\ to\ j} = \frac{\sum_{k=1}^{N} c_{ik} \min(q_{ik}, q_{jk})}{\sum_{k=1}^{N} c_{ik} q_{ik}} \quad (2)$$

An $AS_{i\ to\ j}$ of 0 means that genotype information available to two base populations is independent. A value of 1 means that genotype information available is identical.

The third statistic quantifies the proximity of genotype information and is the weighted average number of generations between base animals in a base population and the genomic base population. It is calculated as:

$$\delta Gener_i = \frac{\sum_{j=1}^{N} c_{ij} q_{ij} gener_j}{\sum_{j=1}^{N} c_{ij} q_{ij}} \quad (3)$$

δGener differentiates between many remote genotyped descendants and fewer proximate ones for a given *Neq*.

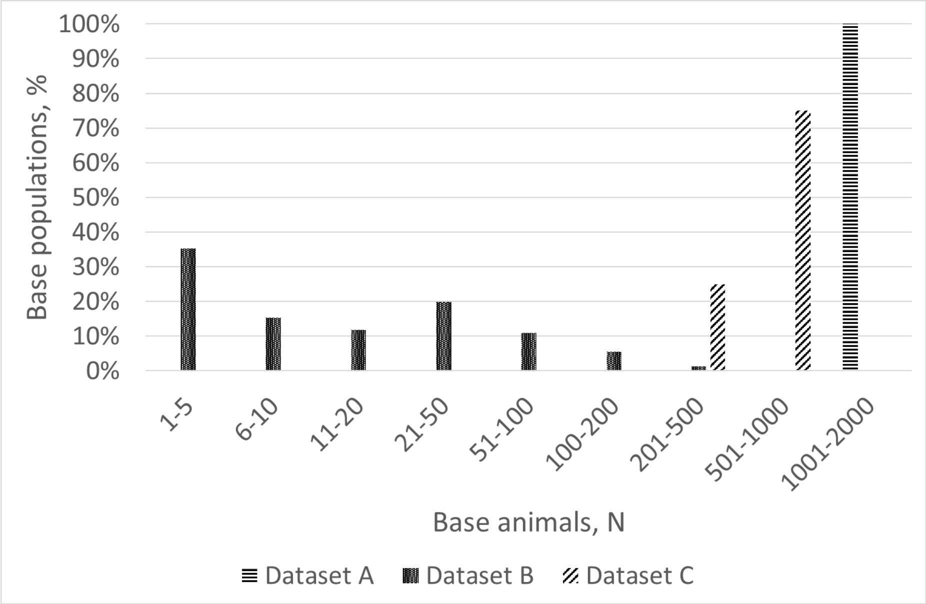

Figure 1. Histogram of number of base animals per base population.

Table 2. Statistics of base populations.

	Dataset A	Dataset B	Dataset C
Base populations, N	3	212	8
Base animals in total, N	1,030	5,702	2,028
Base populations without a link to genotypes, %	0%	47.5%	50.0%
Base populations Neq[1] < 0.05	0%	43.5%	0%
Average non-zero Neq[1]	9.5	8.0	5.3
Range non-zero Neq[1]	9.4 - 9.6	1.0 - 27.0	1.0 - 9.5
Average AS[2] across pairs of base populations	0.06	0.07	0.14
Full AS[2] with at least one other base population, %	0.0%	43.9%	50.0%
Average size of genomic base population[3] per base population	2,247	12.2	507
Average δGener[4]	8.4	25.7	8.3

[1] Neq: amount of genotype information expressed as equivalent number of base animals genotyped (Equation 1)
[2] AS: auto-similarity of genotype information of a base population to another and is a measure of the extent to which the same genotype information was used for both base populations (Equation 2)
[3] Genomic base population: genotyped animals with a path of non-genotyped ancestors to the base population
4 δGener: number of generations between genomic base population and base animals (Equation 3)

Results

The number of base animals as a proportion of the total number of animals in the pedigree was similar across datasets (0.5-2.0%). The number of defined base populations varied enormously across datasets and ranged from 3 to 212 (Table 2). Also, the number of base animals per base population varied considerably between datasets and within Dataset B (Figure 1).

In dataset A, all base populations were well-linked to genotyped animals (Table 2), whereas in datasets B and C, half the base populations had no genetic link to a genotyped animal. The average of non-zero Neq was similar across datasets, but variation was larger in datasets B and C. Of all base populations in datasets B and C, 43.9% and 50% had an AS of 1 with at least one other base population, meaning that their available genotype information was based on the same information (Table 2). In dataset A, there were no pairs of base populations with full AS. In datasets A and C, average δGener was 8.4 and 8.3, respectively. In dataset B, there were five clear clusters with a δGener of 0, 18, 24, 33 and 39 generations.

Discussion

The statistics presented quantify amount, uniqueness and proximity of genomic information for base populations and were useful for evaluating definitions of base populations to be fitted either as UPG or MF.

All studied datasets represented well-recorded populations with 98% or more of pedigree records being complete. Defining base populations based on an expected or *a priori* known difference in genetic level is only meaningful if these true differences can be estimated, so base populations should be defined to be sufficiently large.

The history of poor convergence when adding UPG to the model for dataset B is most likely caused by the large number of base populations with a small number of base animals (Figure 1). Poor connection of MF with genotyped animals, and pairs of MF with full auto-similarity (dataset B and C) will cause the gamma matrix of relationships within and between MF to be singular. If the gamma matrix (or Q'Q in QP transformation) is close to singularity, so is the coefficient matrix of the evaluation, which is known to slow down convergence when solving equations with a conjugate gradient method.

Another issue is the very deep pedigree of dataset B, in some cases, 30 generations of ancestors before the first animals with a genotype. The reasoning was not to lose any known pedigree relationships by removing generations of ancestors from the pedigree. For such data, the challenge is to reduce the number of generations in the pedigree without losing genetic relationships. A solution may be to define a set of base animals two or three generations before the first animals with data or genotypes and use pedigree records of ignored ancestors to construct relationships between these base animals, similar to MF.

Funding

This study was financially supported by the Dutch Ministry of Economic Affairs (TKI Agri and Food project 16022) and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin.

References

Quaas R.L. and Pollak E.J. (1981) Journal of Dairy Science 64(9): 1868-1872. https://doi.org/10.3168/jds.S0022-0302(81)82778-6

Mrode R.A. (2014) Linear Models for the Prediction of Animal Breeding Values, 3rd edition. CABI, Wallingford, UK

Legarra A., Christensen O.F., Vitezica Z.G., Aguilar I. and Misztal I. (2015) Genetics 200(2): 455-468. https://doi.org/10.1534/genetics.115.177014/-/DC1

Foulley J.L., Bouix J., Goffinet B. and Elsen J.M. (1990) Connectedness in Genetic Evaluation. In: Gianola D. and Hammond K. (eds) Advances in Statistical Methods for Genetic Improvement of Livestock. Springer-Verlag, Heidelberg

Van Grevenhof E.M., Vandenplas J. and Calus M.P.L. (2019) Journal of Animal Science 97(2): 548-558. https://doi.org/10.1093/jas/sky433