



**HAL**  
open science

# Quality of breeding value predictions from longitudinal analyses, with application to residual feed intake in pigs

Ingrid David, Anne Ricard, Van-Hung Huynh-Tran, Jack Dekkers, H el ene Gilbert

## ► To cite this version:

Ingrid David, Anne Ricard, Van-Hung Huynh-Tran, Jack Dekkers, H el ene Gilbert. Quality of breeding value predictions from longitudinal analyses, with application to residual feed intake in pigs. *Genetics Selection Evolution*, 2022, 54 (1), 8 p. 10.1186/s12711-022-00722-w . hal-03671277

**HAL Id: hal-03671277**

**<https://hal.inrae.fr/hal-03671277>**

Submitted on 14 Jun 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destin ee au d ep ot et  a la diffusion de documents scientifiques de niveau recherche, publi es ou non,  emanant des  tablissements d'enseignement et de recherche fran ais ou  trangers, des laboratoires publics ou priv es.




Distributed under a Creative Commons Attribution 4.0 International License

SHORT COMMUNICATION

Open Access



# Quality of breeding value predictions from longitudinal analyses, with application to residual feed intake in pigs

Ingrid David<sup>1\*</sup> , Anne Ricard<sup>2,3</sup>, Van-Hung Huynh-Tran<sup>1</sup>, Jack C. M. Dekkers<sup>4</sup> and H el ene Gilbert<sup>1</sup>

## Abstract

**Background:** An important goal in animal breeding is to improve longitudinal traits. The objective of this study was to explore for longitudinal residual feed intake (RFI) data, which estimated breeding value (EBV), or combination of EBV, to use in a breeding program. Linear combinations of EBV (summarized breeding values, SBV) or phenotypes (summarized phenotypes) derived from the eigenvectors of the genetic covariance matrix over time were considered, and the linear regression method (LR method) was used to facilitate the evaluation of their prediction accuracy.

**Results:** Weekly feed intake, average daily gain, metabolic body weight, and backfat thickness measured on 2435 growing French Large White pigs over a 10-week period were analysed using a random regression model. In this population, the 544 dams of the phenotyped animals were genotyped. These dams did not have own phenotypes. The quality of the predictions of SBV and breeding values from summarized phenotypes of these females was evaluated. On average, predictions of SBV at the time of selection were unbiased, slightly over-dispersed and less accurate than those obtained with additional phenotypic information. The use of genomic information did not improve the quality of predictions. The use of summarized instead of longitudinal phenotypes resulted in predictions of breeding values of similar quality.

**Conclusions:** For practical selection on longitudinal data, the results obtained with this specific design suggest that the use of summarized phenotypes could facilitate routine genetic evaluation of longitudinal traits.

## Background

Selecting animals for a better feed efficiency is one of the key levers to improve farm profitability while reducing the environmental impact of livestock farming [1, 2]. Thanks to the development of electronic devices on farms, the recording of repeated phenotypes over time is facilitated in different livestock species [3–5]. This recording of longitudinal data is beneficial in a genetic context because it allows the extraction of useful information for selection on more complex criteria than

the estimated breeding value (EBV) for the mean value of the trait over the studied period; thus, selection for an optimal shape of the curve [6–8] and/or for specific components of the curve, such as persistency of milk production [9] becomes possible. Longitudinal measurements are often correlated at both the genetic and environmental level, with, generally, a structured covariance pattern. To analyze such data with a limited number of parameters to be estimated, approaches that model the shape of the covariance functions, such as character process models (CP) [10], or that model the functions of the random effects, such as random regression (RR) [11, 12] or structured antedependence (SAD) [13] models, have been proposed. These models have proven to be efficient to model the covariance structure of the data and to

\*Correspondence: [ingrid.david@inrae.fr](mailto:ingrid.david@inrae.fr)

<sup>1</sup> GenPhySE, INRAE, Universit e de Toulouse, INPT, 31326 Castanet Tolosan, France

Full list of author information is available at the end of the article



  The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

accurately estimate breeding values for each time point in numerous studies [14–17]. However, these models are not widely used for routine genetic evaluation in livestock except for milk production traits in dairy cattle [6], for several reasons. First, combining time point EBV in a selection index remains an issue, and second, fitting these models is computationally demanding, especially when accounting for genomic information. To overcome these challenges, the use of eigenvectors of the genetic covariance matrix to linearly combine the time point breeding values into several summarized breeding values (SBV) has been proposed [18]. These eigenvectors have been shown to make sense biologically for several longitudinal traits [18–21] and allow predictions for independent SBV to be obtained for selection. In addition, the use of SBV results in fewer equations in the mixed model and better convergence properties [22].

Thanks to new genotyping technologies that render the genotyping of numerous single nucleotide polymorphisms (SNPs) cost-effective, genomic prediction (GP) has been implemented in multiple livestock species [23–28]. For example, genomic EBV (GEBV) can be obtained by single-step genomic best linear unbiased prediction (ssGBLUP), which combines in a single-step pedigree data, phenotypes, and genotypes for genetic evaluation [29]. The efficiency of genomic selection on these GEBV depends on their bias and accuracy [30]. Traditionally, the accuracy of GEBV is obtained by cross-validation approaches, based on the computation of correlations between adjusted phenotypes and GEBV of hypothetical selection candidates. This is, however, not straightforward for predictions of SBV. As an alternative, Legarra and Reverter [31] proposed the linear regression method (LR method), which provides a series of statistics to quantify the quality of predictions that can be used in complex scenarios where adjusted phenotypes or coefficients of determination are not straightforward to obtain, such as for SBV.

Thus, our goal was to propose and test operational solutions that could facilitate the implementation of selection for longitudinal traits, using the example of feed efficiency. Using the LR method, the objectives of the present study were to evaluate the quality of the predictions of SBV for longitudinal RFI in pigs, as well as the benefit of adding genomic information. In addition, for practical selection on longitudinal profiles, we evaluated the effect of using summarized phenotypes instead of longitudinal phenotypes on the quality of predictions of SBV.

## Methods

### Material

Phenotypes on feed intake (FI) and production (metabolic body weight (MBW), average daily gain (ADG), and backfat

thickness (BFT)) records measured weekly on 2397 growing French Large White pigs over a 10-week period from ~13 to ~22 weeks of age were used in this study (descriptive statistics are in Additional file 1 Table S1). These animals were from seven generations of a divergent selection experiment for RFI applied at the end of each test period (110 kg) [1]. Animal management and phenotype measurements are described in David et al. [32] and Huynh-Tran et al. [19]. The numbers of animals and records per low (LRFI), high RFI (HRFI) lines and generation are in Additional file 1: Table S2. All sires and dams (660 pigs) of the phenotyped animals of generations G1 to G7 were genotyped using the Illumina SNP60 Beadchip V2 (64,232 SNPs) or the Illumina Porcine HD Array GGP (68,516 SNPs). None of the parents had own phenotypes. After quality control, genotyped or imputed genotypes for 64,487 SNPs on 624 pigs were available for further analyses (see Additional file 1: Table S2). The inverse of the  $\mathbf{H}$  matrix that combines genomic and pedigree information was obtained using the method proposed by Legarra et al. [29] as  $\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{\Omega}_{\omega}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$ , where  $\mathbf{A}_{22}$  is the pedigree-based numerator relationship matrix for the genotyped animals and  $\mathbf{\Omega}_{\omega} = (1 - 0.05)\mathbf{\Omega}^* + 0.05\mathbf{A}_{22}$ . Matrix  $\mathbf{\Omega}^*$  was obtained by scaling the genomic relationship matrix  $\mathbf{\Omega}$  [33], to make the means of the diagonal and off-diagonal elements of  $\mathbf{\Omega}^*$  and  $\mathbf{A}_{22}$  matrices equal to each other. The  $\mathbf{A}$  matrix was obtained for all animals in the pedigree plus ancestors (grandparents), and comprised 3095 individuals over 10 generations. The  $\mathbf{H}$  matrix was obtained and formatted using the PreGSf90 software [34] and modified with R to be supplied to ASReml as a user defined relationship matrix.

### Methods

Longitudinal production and FI records of all phenotyped animals (whole data) were used to compute longitudinal RFI by a phenotypic regression of FI on production and maintenance traits, using the following RR model of degree 2 for the genetic and permanent environmental effects:

$$FI_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \beta_1 MBW_{ij} + \beta_2 ADG_{ij} + \beta_3 BFT_{ij} + \sum_{q=0}^2 a_{iq}\varphi_q(j) + \sum_{q=0}^2 b_{iq}\varphi_q(j) + e_{ij},$$

where  $FI_{ij}$ ,  $MBW_{ij}$ ,  $ADG_{ij}$ , and  $BFT_{ij}$  are the FI, MBW, ADG and BFT of animal  $i$  in week  $j$ ,  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  are the phenotypic regression coefficients linking production and maintenance traits to FI, and  $\varphi_q(j)$  is the  $(q + 1)^{th}$  Legendre polynomial at time  $j$ . Vectors  $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_{3095})$ ,  $\mathbf{a}_i = (a_{i0}, a_{i1}, a_{i2})$ , and  $\mathbf{b} = (\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_{2397})$ ,  $\mathbf{b}_i = (b_{i0}, b_{i1}, b_{i2})$  are the vectors of random coefficients for the genetic and permanent

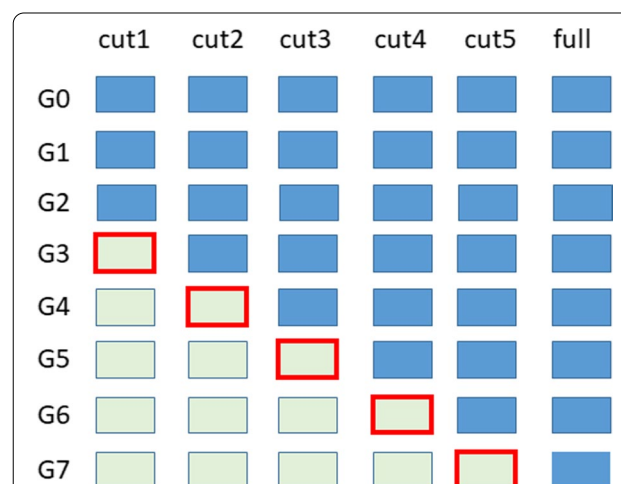
environmental effects, respectively, with  $\mathbf{a} \sim N(\mathbf{0}, \mathbf{A} \otimes \mathbf{K}_a)$  and  $\mathbf{b} \sim N(\mathbf{0}, \mathbf{I} \otimes \mathbf{K}_b)$ ,  $\mathbf{K}_a$  and  $\mathbf{K}_b$  being the  $3 \times 3$  covariance matrices between genetic and permanent random coefficient, respectively and  $\mathbf{e}$  is the residual vector, with heterogeneous variances over weeks ( $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I} \otimes \mathbf{D})$ , where  $\mathbf{D}$  is a  $10 \times 10$  diagonal matrix). Thus, the distribution of the additive genetic ( $u_{ij} = \sum_{q=0}^2 a_{iq} \varphi_q(j)$ ) and permanent environmental effects ( $p_{ij} = \sum_{q=0}^2 b_{iq} \varphi_q(j)$ ) are  $\mathbf{u} \sim N(\mathbf{0}, \mathbf{A} \otimes \mathbf{G})$  and  $\mathbf{p} \sim N(\mathbf{0}, \mathbf{I} \otimes \mathbf{P})$ , respectively, with  $\mathbf{G} = \boldsymbol{\Psi} \mathbf{K}_a \boldsymbol{\Psi}'$ , where  $\boldsymbol{\Psi}$  is the  $(10 \times 3)$  matrix of the Legendre polynomials for all time points and  $\mathbf{P} = \boldsymbol{\Psi} \mathbf{K}_b \boldsymbol{\Psi}'$ . For selection purposes, the 10 breeding values per animal (one for each time point) are summarized into a reduced number of values (SBV) that are interpretable and of interest for selection (i.e. that give information about the shape of the BV trajectory). These SBV were obtained by an Eigen decomposition of the genetic covariance matrix  $\mathbf{G}$ . The  $l^{\text{th}}$  SBV for animal  $i$  is then  $SBV_{l,i} = \mathbf{L}_{G,l} \mathbf{u}_i$ , where  $\mathbf{L}_{G,l}$  is the  $l^{\text{th}}$  eigenvector from the decomposition of the  $\mathbf{G}$  matrix and  $\mathbf{u}_i$  the vector of breeding values for animal  $i$  ( $\mathbf{u}_i = (u_{i1}, u_{i2}, \dots, u_{i10})$ ). Fixed effects included in the model ( $\mathbf{x}_{ij} \boldsymbol{\beta}$ ) were the combination week by generation ( $10 \times 8$  levels), the combination batch by sex ( $66 \times 3$  levels), birth herd (2 levels), age at start of the test (covariate), and pen (16 levels). Variance components for this full dataset were estimated using the REML approach in ASReml 4.0 [35] and were considered as known in all subsequent analyses. Estimates of heritability of each  $SBV_l$ , obtained by the Eigen decomposition of the genetic covariance matrix  $\mathbf{G}$ , were computed as

$$h_{SBV_l}^2 = \frac{\mathbf{L}'_{G,l} \widehat{\mathbf{G}} \mathbf{L}_{G,l}}{\mathbf{L}'_{G,l} (\widehat{\mathbf{G}} + \widehat{\mathbf{P}} + \widehat{\mathbf{D}}) \mathbf{L}_{G,l}}$$

The LR method applied to the SBV of a group of focal individuals (i.e. individuals of interest) consists in comparing the estimated SBV (ESBV) of these focal individuals based on less information to their ESBV based on more information, considering that more information provides a reasonable reference. ESBV of the focal individuals based on less information (partial) are referred to as  $ESBV_p$  and their ESBV based on more information (whole) as  $ESBV_w$ . For the sake of simplicity, the subscript of SBV is not added, thus ESBV refer to the first, second or third ESBV, indifferently. Three of the five statistics based on the comparison of these two sets of ESBV, as proposed by Legarra and Reverter [31], were used here, each referring to properties of  $ESBV_p$  for the focal individuals. The estimate of the standardized bias of  $ESBV_p$  was computed as (for the  $l^{\text{th}}$  SBV)  $\frac{\overline{ESBV_p} - \overline{ESBV_w}}{\mathbf{L}'_{G,l} \widehat{\mathbf{G}} \mathbf{L}_{G,l}}$ ,

where  $\overline{ESBV_p}$  and  $\overline{ESBV_w}$  are the mean of  $ESBV_p$  and  $ESBV_w$ , respectively. Its expected value is 0 if the evaluation with the partial dataset is unbiased. Dispersion of

$ESBV_p$  was evaluated by the regression of  $ESBV_w$  on  $ESBV_p$ :  $b_{w,p} = \frac{\text{cov}(ESBV_p, ESBV_w)}{\text{var}(ESBV_p)}$ . Its expected value is 1, if there is no over-/under-dispersion. Finally, the ratio of the prediction accuracy (i.e. the correlation between true and estimated breeding values) of  $ESBV_p$  to the prediction accuracy of  $ESBV_w$  was evaluated by the correlation between these ESBV:  $\rho_{w,p} = \frac{\text{cov}(ESBV_p, ESBV_w)}{\sqrt{\text{var}(ESBV_p) \text{var}(ESBV_w)}}$ . The lower this correlation is, the higher is the relative gain in accuracy due to additional information. We evaluated the quality of SBV predictions for genotyped dams, which did not have own phenotypes (but had phenotyped half-sibs and full-sibs, on average 24 and 4 per dam, respectively), as focal individuals by comparing their SBV estimated with less and more information. The additional information considered were the phenotypes of their descendants or their genomic information. To do so, five cut-offs were considered, corresponding to the number of generations with phenotypes in the dataset (from 3 generations (G0, G1, G2, cut-1) to 7 generations (cut-5)), as described in Fig. 1. Based on these cut-offs, five groups of focal individuals were defined (group\_FI*i*,  $i = 1, \dots, 5$ ), which corresponded to the dams of the first generation of animals without phenotypes. For instance, the first group of focal individuals (group\_FI1) consisted of the dams of the first generation without phenotypes in cut-1, i.e. the dams of the phenotyped animals in G3. For each group of focal individuals, the three LR statistics were computed to compare predictions obtained with less information to



**Fig. 1** Scenarios retained to test the effect of additional phenotypic or genomic information on predictions, depending on the number of phenotyped generations. Blue box: animals with phenotypes, green box: animals without phenotypes, red indicates the progeny of the focal individuals genotyped (sires and dams of the previous generation that do not have own phenotype). For instance in cut1, focal animals are the genotyped sire and dams of G3 animals

those obtained with more information. Thus, the following comparisons made were:

(1) SBV predicted for the group of focal individuals (group\_FI<sub>*i*</sub>) obtained with phenotypes in cut\_<sub>*i*</sub> compared to SBV predicted for the group of focal individuals obtained with phenotypes in cut\_<sub>*(i+1)*</sub>, cut\_<sub>*(i+1)*</sub> corresponding to the full dataset if (*i* = 5) using (a) pedigree information only (i.e. **A** matrix) or (b) genomic information (i.e. the **A** matrix is then replaced by the **H** matrix for the distribution of genetic effects in the RR model);

(2) SBV predicted for the group of focal individuals (group\_FI<sub>*i*</sub>) obtained with phenotypes in cut\_<sub>*i*</sub> compared to SBV predicted for the group of focal individuals obtained with phenotypes of the full dataset using (a) pedigree information only (i.e. **A** matrix) or (b) genomic information (i.e. **H** matrix);

(3) SBV predicted for the group of focal individuals using pedigree information (i.e. **A** matrix) compared to SBV obtained with phenotypes of the same dataset using genomic information (i.e. **H** matrix). SBV were obtained using phenotypes in (a) cut\_<sub>*i*</sub> or (b) phenotypes of the full dataset.

The first two comparisons allowed us to evaluate the quality of the model to predict SBV based on phenotypic information on ascendants and collateral relatives of the focal individuals. For these two comparisons, two “whole” datasets were considered as the gold standard (cut\_<sub>*(i+1)*</sub> or full data) to account for random errors in the estimates of the LR statistics [36]. The third comparison allowed us to evaluate the benefit of genomic information for predictions.

The comparisons were performed separately for each line and for the three first SBV. These comparisons led to 108 estimates of the three LR statistics when evaluating the quality of the model to predict SBV based on phenotypic information on ascendants and collateral relatives of the focal individuals (2 lines, 3 SBV, **A** or **H** matrix, 5 cut-offs), and 30 estimates of the three LR statistics when evaluating the quality of the model to predict SBV based on phenotypic and pedigree information (gold standard being GSBV, 2 lines, 3 SBV, 5 cut-offs). Tests of the impact of the different factors (line, SBV, type of genetic information) on the LR statistics were evaluated using linear models including these three factors, and likelihood ratio tests. The two-by-two interactions between all factors were tested and kept in the model only when they were significant for one LR statistic.

Mirroring the SBV, fixed effects and variance component estimates obtained for the full dataset and the pedigree relationship matrix were used to compute the following three summarized phenotypes ( $y_{SBV_l}, l = 1, 2, 3$ ):  $y_{SBV_l} = \mathbf{L}'_{\hat{\mathbf{G}},l} (\mathbf{FI} - \mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\beta}_1\mathbf{MBW} - \hat{\beta}_2\mathbf{ADG} - \hat{\beta}_3\mathbf{BFT})$ . In order to obtain summarized phenotypes for all animals

for which one of the longitudinal phenotype (FI, MBW, ADG or BFT) was missing (4.2% of the data corresponding to 20% of the animals with at least one missing weekly phenotype), missing values for  $\mathbf{FI} - \mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\beta}_1\mathbf{MBW} - \hat{\beta}_2\mathbf{ADG} - \hat{\beta}_3\mathbf{BFT}$  were replaced by the average value per week of the full population before multiplying by  $\mathbf{L}'_{\hat{\mathbf{G}},l}$ . Breeding values for the summarized phenotypes were obtained using the following model:  $y_{SBV_l,i} = \mu_l + u_{y_{SBV_l,i}} + e_{SBV_l,i}$ , with variance components derived from estimates obtained from the model for longitudinal RFI:  $\mathbf{u}_{y_{SBV_l}} \sim N(\mathbf{0}, \mathbf{A}\mathbf{L}'_{\hat{\mathbf{G}},l}\hat{\mathbf{G}}\mathbf{L}_{\hat{\mathbf{G}},l})$  and  $e_{SBV_l} \sim N(\mathbf{0}, \mathbf{I}\mathbf{L}'_{\hat{\mathbf{G}},l}(\hat{\mathbf{P}} + \hat{\mathbf{D}})\mathbf{L}_{\hat{\mathbf{G}},l})$ . To further evaluate the applicability of longitudinal models for routine genetic evaluation, we then compared the EBV of  $y_{SBV_l}(\widehat{\mathbf{u}}_{y_{SBV_l}})$  to those obtained for  $SBV_l$ , assuming that the eigenvectors do not have to be re-computed for each new evaluation.

## Results

The genetic parameters were estimated based on the full dataset and the pedigree relationship matrix. The three first eigenvalues from the eigen decomposition of  $\hat{\mathbf{G}}$  represented 59, 26 and 15% of the total genetic variance, respectively. Heritability estimates of the corresponding first three SBV were  $0.36 \pm 0.05$ ,  $0.20 \pm 0.04$ , and  $0.16 \pm 0.05$ , respectively. The LR statistics and p-values of the different factors evaluating the quality of (G)ESBV obtained at the time of selection are summarized in Table 1. On average, ESBV were unbiased, slightly overdispersed and less accurate than ESBV predicted with more phenotypic information ( $\Delta\mu_{wp} = 0.00$ , 95% Confidence Interval:  $[-0.01, 0.01]$ ,  $b_{w,p} = 0.84$  [0.79, 0.89], and  $\overline{\rho}_{w,p} = 0.61$  [0.58, 0.64]). Overdispersion was significantly larger for the LRFI line than for the HRFI line ( $b_{w,p} = 0.75$  [0.68, 0.82] versus 0.93 [0.86, 1.00]) and for SBV3 compared to SBV2 ( $b_{w,p} = 0.76$  [0.69, 0.85] versus 0.93 [0.85, 1.00]) (see Additional file 2: Fig. S1). The ratio of accuracy between SBV obtained with more or less phenotypic information was significantly lower for SBV3 than for SBV2 ( $\rho_{w,p} = 0.56$  [0.51, 0.61] versus 0.67 [0.61, 0.72]). Considering the SBV obtained using pedigree and genomic information as the gold standard, on average, SBV obtained with pedigree information only were biased downwards ( $\Delta\mu_{wp} = -0.08$   $[-0.09, -0.07]$ ). The bias differed between SBV: underestimated for SBV<sub>1</sub> and SBV<sub>2</sub> and overestimated for SBV<sub>3</sub> (see Additional file 2: Fig. S2). Finally, the SBV predicted using longitudinal phenotypes and EBV predicted using summarized phenotypes ( $\widehat{\mathbf{u}}_{y_{SBV_l,p}}$ ) for genotyped animals in the full dataset are summarized in Table 2: the EBV obtained from

**Table 1** Average LR statistics and p-values of the tested factors for the prediction of SBV

Partial data	Whole data		$\Delta\mu_{wp}$	$b_{w,p}$	$\rho_{w,p}$	
No phenotypic information from candidates' descendants	With phenotypic information from candidates' descendants	Mean <sup>a</sup>	0.00 [− 0.01,0.01]	0.84 [0.79,0.89]	0.61 [0.58,0.64]	
		p_value	SBV	0.78	0.03	0.02
			Line	0.22	<0.01	0.25
		<b>A</b> or <b>H</b> matrix	0.76	0.71	0.70	
Phenotypes, pedigree information	Phenotypes, pedigree and genomic information	Mean <sup>a</sup>	− 0.08 [− 0.09,− 0.07]	0.96 [0.92,1.00]	0.89 [0.87,0.92]	
		p_value	SBV	<0.01	0.68	0.37
			Line	0.41	0.97	0.61

<sup>a</sup> 95% confidence interval in bracket

Bias:  $\Delta\mu_{wp} = \frac{ESBV_p - ESBV_w}{L_{G1} L_{G1}}$ ; dispersion:  $b_{w,p} = \frac{cov(ESBV_p, ESBV_w)}{var(ESBV_p)}$ ; Ratio of accuracy:  $\rho_{w,p} = \frac{cov(ESBV_p, ESBV_w)}{\sqrt{var(ESBV_p)var(ESBV_w)}}$

Indices *w* = estimates obtained with more information, *p* = estimates obtained with less information

**Table 2** Comparison of SBV predicted using longitudinal phenotypes<sup>a</sup> and EBV predicted using summarized phenotypes<sup>b</sup>, using pedigree and pedigree plus genomic information for the low (LRFI) and high (HRFI) RFI lines

	Pedigree		Pedigree + genomics	
	LRFI	HRFI	LRFI	HRFI
Bias $\overline{SBV_w} - \overline{u_{ySBVw}}$				
<i>SBV</i> <sub>1</sub>	0.00	0.02	0.00	0.02
<i>SBV</i> <sub>2</sub>	0.03	0.00	0.02	0.00
<i>SBV</i> <sub>3</sub>	0.01	0.00	0.01	0.00
Dispersion $b_{SBV_w, u_{ySBVw}}$				
<i>SBV</i> <sub>1</sub>	1.05	1.03	1.04	1.03
<i>SBV</i> <sub>2</sub>	1.03	1.02	1.03	1.03
<i>SBV</i> <sub>3</sub>	1.00	1.01	0.97	1.00
Ratio of accuracies $\rho_{SBV_w, u_{ySBVw}}$				
<i>SBV</i> <sub>1</sub>	1.00	1.00	1.00	1.00
<i>SBV</i> <sub>2</sub>	1.00	0.99	0.99	0.99
<i>SBV</i> <sub>3</sub>	0.98	0.98	0.98	0.98

<sup>a</sup> Longitudinal phenotypes = weekly measurements of FI analyzed with an RR model:

$F_{ij} = \mathbf{x}_{ij}\beta + \beta_1 MBW_{ij} + \beta_2 ADG_{ij} + \beta_3 BFT_{ij} + \sum_{q=0}^2 a_{iq} \varphi_q(j) + \sum_{q=0}^2 b_{iq} \varphi_q(j) + e_{ij}$ ,  $SBV_{ij} = \mathbf{L}_{G1} \mathbf{u}_i$ , where  $\mathbf{L}_{G1}$  is the *i*<sup>th</sup> eigenvector from the decomposition of the **G** matrix

<sup>b</sup> Summarized phenotypes:  $\mathbf{y}_{SBV1} = \mathbf{L}'_{G1} (\mathbf{FI} - \mathbf{X}\hat{\beta} - \hat{\beta}_1 \mathbf{MBW} - \hat{\beta}_2 \mathbf{ADG} - \hat{\beta}_3 \mathbf{BFT})$

summarized phenotypes were unbiased, neither over- nor under-dispersed, and were as accurate as the SBV predicted using longitudinal phenotypes, when computed with the pedigree information only (**A** matrix) or combining pedigree and genotypes (**H** matrix).

### Discussion

The objectives of the present study were to evaluate, for longitudinal RFI in pigs, the quality of the model predictions for SBV that are useful for selection, and the benefit of adding genomic information. In addition, for practical selection on longitudinal profiles, we evaluated the

consequences of using summarized phenotypes instead of longitudinal phenotypes on estimates of (summarized) breeding values. To compare estimates of breeding values obtained in different situations, we applied the LR method proposed by Legarra and Reverter [31]. The use of the LR method in place of traditional cross-validation tests overcomes the limitations of the latter since it does not need “true” BV (i.e. highly accurate EBV) or pre-corrected phenotypes. Thus, the method is particularly suitable for pig populations where selection candidates do not necessarily have own performance or numerous recorded progeny, and it is all the more useful for longitudinal data for which missing data are frequent. In addition, the BV of interest for selection in the longitudinal case, i.e. SBV, are a linear combination of EBV that accumulates those aforementioned difficulties for each time point, making the computation of the accuracy of the resulting SBV with traditional approaches quite complex when information is heterogeneous across time points and candidates. To be relevant, the LR method should be applied to a set of focal individuals that is sufficiently large and diverse (i.e. animals from several families), that should represent the population of interest (selection candidates in our case), for which the quantity of “information” used to estimate SBV should be similar for the different focal individuals in the partial dataset as well as in the whole dataset and for which the reference EBV should be reasonably accurate. In the present study, progeny from all families were candidates to selection, while focal individuals were the genotyped dams used for breeding in the next generation. The dams were randomly selected within sire, one female replacing its dam, to maintain the genetic diversity, so they represent the population of interest and fulfilled the requirement in terms of diversity. Because experimental lines have a limited number of breeding animals, the groups of focal individuals were small and thus did not fulfil the ‘sufficiently large’ requirement. To partially counteract this,

we repeated, as Macedo et al. [36], the estimation of LR indicators in successive focal groups. In order to meet the requirement for the focal individuals to have the same quantity of information to predict their EBV, only genotyped dams were considered as focal individuals, and not genotyped sires. Indeed, the quantity of information to predict the SBV of sires and dams would have been similar in the partial dataset (phenotypes from ascendants), but not in the whole dataset, because sires had on average six times more phenotyped progeny than females. Finally, the theoretical accuracy of the reference EBV (whole population) for the focal individuals was around 0.6 for SBV1 and 0.5 for SBV2 and SBV3, and thus fulfilled the requested reasonable accuracy (Andres Legarra, personal communication). In this study, we were not interested in estimating the accuracy of SBV of the focal animals themselves, as the ratio of accuracies was sufficient to judge the quality of the SBV at the time of selection and to evaluate the gain from genomic information for selection. Nevertheless, it is possible to estimate the (selected) reliability of SBV at the time of selection with the LR method by computing  $\frac{cov(SBV_p, SBV_w)}{\sigma_{u^*}^2}$ , where  $\sigma_{u^*}^2$  is the genetic variance of the group of individuals of interest, which can be estimated by Gibbs sampling [37].

The bias ( $\Delta\mu_{wp}$ ) and dispersion ( $b_{w,p}$ ) obtained when evaluating the quality of predictions of (G)ESBV at the time of selection give indications on the errors that can be made on the expected genetic gain at the stage of selection [31]. On the one hand, the bias was null on average. On the other hand, the (G)ESBV were over-dispersed, especially in the LRFI line, which is in line with the difference in dispersion between the two lines reported for the same population by Aliakbari et al. [38]. This over-dispersion results in overestimation of the expected genetic gain (by  $\simeq 0.2\sigma_{SBV}$ , [31]). As expected, ESBV of the focal individuals predicted without phenotypes from descendants were significantly less accurate than ESBV predicted using all the phenotypic information ( $\rho_{w,p}$  significantly lower than 1). We did not detect differences in bias, dispersion, and relative gain in accuracy due to additional phenotypes, between ESBV obtained using genomic and pedigree information or pedigree information only. Yet, it is expected that “*additional phenotypic records would have lower impact on GEBV (compared to EBV) because they would contribute with less information than the direct genomic value*” [39]. The lack of effect of genomic information on SBV prediction in the present study is confirmed by the high correlation between ESBV and GSBV and the high regression coefficient of GSBV on ESBV that we obtained. In our study, the small benefit from genomic information is likely due to the small number of available genotyped animals with accurate phenotype information, much smaller than the expected

number (1690, [40]) that is necessary to represent the genomic structure of the population [41]. Consequently, too little of the genetic variance could be captured by the genomic information in this dataset to have a significant impact on ESBV. In addition, there were no animals with both genotype and phenotype, and the number of phenotyped progeny for each focal genotyped animal was small. It should be noted that a single set of scaling factors to construct the **H** matrix has been tested, although their values may impact bias and dispersion of GEBV [42].

Breeding values that summarized the trajectory of RFI over time using eigenvectors of the estimated variance–covariance structure of the longitudinal data were obtained using two approaches:  $SBV_l$ , which were extracted from the genetic analysis of the longitudinal phenotypes (considered as the reference), and  $u_{SBV_p}$  which corresponds to the breeding values in the analysis of summarized phenotypes. The latter requires much less computing time, and thus is more suitable for routine evaluation. When, in the model for summarized phenotypes  $y_{SBV}$ , the longitudinal phenotypes are pre-corrected for fixed effects and the variance components are assumed known and fixed, our results (Table 2) show that these two approaches give exactly the same EBV. In practice, pre-correction for all fixed effects is not always possible (i.e. contemporary group effects). To evaluate the additional noise generated in such a situation, we performed the same comparison between  $SBV_l$  and  $u_{SBV_l}$  using summarized phenotypes that were not corrected for fixed effects, except for regression on other phenotypes to obtain RFI (i.e.  $y_{SBV_l} = L'_{G,l} (FI - \hat{\beta}_1 MBW - \hat{\beta}_2 ADG - \hat{\beta}_3 BFT)$ ), but by including and estimating the effect of the fixed effects in the model used for their analysis (the same fixed effects as used in the longitudinal model except for time and interactions with time). This analysis resulted in similar bias and dispersion as those obtained when pre-correcting phenotypes for all fixed effects. The correlations between  $SBV_l$  and  $u_{y_{SBV_l}}$  were slightly lower but still large for SBV<sub>1</sub> (0.92 and 0.95 for LRFI and HRFI, respectively) and SBV<sub>2</sub> (0.94 for both lines), but lower for SBV<sub>3</sub> (0.78 for both lines). Since the first two SBV have been identified as sufficient to select for a desired trajectory pattern of RFI over time in pigs [32], using the model on summarized phenotypes seems a good alternative (similar accuracy with less computing time) to select for RFI trajectories in routine evaluation.

## Conclusions

Using the LR method, we evaluated, the quality of prediction of breeding values for candidates without own phenotypes in the study of longitudinal data, by using two different approaches (analysis of longitudinal or summarized phenotypes), and evaluated the benefit of adding genomic information for prediction. Predictions were of similar quality with the two approaches,

meaning that, in this population design, the use of summarized phenotypes would be of interest for routine evaluation of longitudinal traits. We did not highlight any benefit of genomic information for prediction in this study, which is certainly due to the number of genotyped animals with accurate estimation being too small in this dataset, contrary to the usual expectation.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12711-022-00722-w>.

**Additional file 1: Table S1.** Descriptive statistics of the data. **Table S2.** Numbers of records, animals with phenotype, sires, and dams per line and generation

**Additional file 2: Figure S1.** Ls means of the line effect on the LR statistics  $b_{w,p}$  (left panel) and of the type of SBV on  $b_{w,p}$  (middle panel), and  $\rho_{w,p}$  (right panel) evaluating the quality of the model to predict SBV based on phenotypic information on ascendants and collateral relatives of the focal individuals. **Figure S2.** Ls means of the type of SBV on the LR statistics  $\Delta\mu_{w,p}$ , evaluating the quality of the model to predict SBV based on phenotypic and pedigree information.

## Acknowledgements

The authors would like to thank the experimental farm staff for data collection and breeding of the animals.

## Author contributions

ID analysed the data and was the major contributor in writing the manuscript. VHT performed part of the data analyses. HG secured the funding, designed the experiment, HG, ID and JCMD provided scientific supervision. AR, ID and HG designed the data analysis and discussed the results. All the authors read and approved the final manuscript.

## Funding

This research was supported by the Feed-a-Gene Project funded by the European Union H2020 Programme under grant agreement EU 633531.

## Availability of data and materials

The datasets analysed during the current study are available from H el ene Gilbert (Helene.gilbert@inrae.fr) on reasonable request.

## Declarations

### Ethics approval and consent to participate

The study was conducted in accordance with the French legislation on animal experimentation and ethics. The experimental facilities are run under the certificate of Authorisation to Experiment on Living Animals number 86-213-01 issued by the Ministry of Higher Education, Research and Innovation.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>GenPhySE, INRAE, Universit e de Toulouse, INPT, 31326 Castanet Tolosan, France. <sup>2</sup>Universit e Paris Saclay, INRAE, AgroParisTech, GABI, 78352 Jouy-en-Josas, France. <sup>3</sup>D epartement Recherche et Innovation, Institut Franais du Cheval et de l'Equitation, 61310 Exmes, France. <sup>4</sup>Department of Animal Science, Iowa State University, Ames, IA 50011, USA.

Received: 26 November 2021 Accepted: 20 April 2022

Published online: 13 May 2022

## References

- Gilbert H, Bidanel JP, Gruand J, Caritez JC, Billon Y, Guillouet P, et al. Genetic parameters for residual feed intake in growing pigs, with emphasis on genetic relationships with carcass and meat quality traits. *J Anim Sci.* 2007;85:3182–8.
- Saintilan R, Merour I, Brossard L, Tributou T, Dourmad JY, Sellier P, et al. Genetics of residual feed intake in growing pigs: relationships with production traits, and nitrogen and phosphorus excretion traits. *J Anim Sci.* 2013;91:2542–54.
- Casey DS, Stern HS, Dekkers JCM. Identification of errors and factors associated with errors in data from electronic swine feeders. *J Anim Sci.* 2005;83:969–82.
- Bley TAG, Bessei W. Recording of individual feed intake and feeding behavior of Pekin ducks kept in groups. *Poult Sci.* 2008;87:215–21.
- Marie-Etancelin C, Francois D, Weisbecker J-L, Marcon D, Moreno-Romieux C, Bouvier F, et al. Detailed genetic analysis of feeding behaviour in Romane lambs and links with residual feed intake. *J Anim Breed Genet.* 2019;136:174–82.
- Oliveira HR, Brito LF, Lourenco DAL, Silva FF, Jamrozik J, Schaeffer LR, et al. Invited review: advances and applications of random regression models: from quantitative genetics to genomics. *J Dairy Sci.* 2019;102:7664–83.
- Schaeffer LR, Dekkers JCM. Random regressions in animal models for test-day production in dairy cattle. In *Proceedings of the 5th World Congress of Genetics Applied Livestock Production: 7–12 August 1994; Guelph. 1994.*
- Huynh-Tran VH, David I, Billon Y, Gilbert H. Changes of EBV trajectories for feed conversion ratio of growing pigs due to divergent selection for residual feed intake. In *Proceedings of the 11th World Congress on Genetics Applied to Livestock Production: 10–15 February 2018; Auckland. 2018.*
- Muir BL, Fatehi J, Schaeffer LR. Genetic relationships between persistency and reproductive performance in first-lactation Canadian Holsteins. *J Dairy Sci.* 2004;87:3029–37.
- Pletcher SD, Geyer CJ. The genetic analysis of age-dependent traits: modeling the character process. *Genetics.* 1999;153:825–35.
- Diggle PJ, Heagerty PJ, Liang KY, Zeger SL. *Analysis of longitudinal data.* Oxford: Oxford University Press; 2002.
- Jamrozik J, Schaeffer LR, Weigel KA. Genetic evaluation of bulls and cows with single- and multiple-country test-day models. *J Dairy Sci.* 2002;85:1617–22.
- Zimmerman DL, Nunez-Anton VA. *Antedependence models for longitudinal data.* Boca Raton: Chapman & Hall/CRC; 2010.
- Druet T, Jaffr ezic F, Ducrocq V. Estimation of genetic parameters for test day records of dairy traits in the first three lactations. *Genet Sel Evol.* 2005;37:257–71.
- Cai W, Wu H, Dekkers JCM. Longitudinal analysis of body weight and feed intake in selection lines for residual feed intake in pigs. *Asian-Australas J Anim Sci.* 2011;24:17–27.
- Speidel SE, Enns RM, Crews DH Jr. Genetic analysis of longitudinal data in beef cattle: a review. *Genet Mol Res.* 2010;9:19–33.
- David I, Ruesche J, Drouilhet L, Garreau H, Gilbert H. Genetic modeling of feed intake. *J Anim Sci.* 2015;93:965–77.
- van der Werf JHJ, Goddard ME, Meyer K. The use of covariance functions and random regressions for genetic evaluation of milk production based on test day records. *J Dairy Sci.* 1998;81:3300–8.
- Huynh-Tran VH, Gilbert H, David I. Genetic structured antedependence and random regression models applied to the longitudinal feed conversion ratio in growing Large White pigs. *J Anim Sci.* 2017;95:4752–63.
- Arnal M, Robert-Grani e C, Larroque H. Diversity of dairy goat lactation curves in France. *J Dairy Sci.* 2018;101:11040–51.
- Togashi K, Lin C. Selection for milk production and persistency using eigenvectors of the random regression coefficient matrix. *J Dairy Sci.* 2006;89:4866–73.



22. Druet T, Jaffrézic F, Boichard D, Ducrocq V. Modeling lactation curves and estimation of genetic parameters for first lactation test-day records of French Holstein cows. *J Dairy Sci.* 2003;86:2480–90.
23. Carillier C, Larroque H, Palhière I, Clément V, Rupp R, Robert-Granié C. A first step toward genomic selection in the multi-breed French dairy goat population. *J Dairy Sci.* 2013;96:7294–305.
24. Knol EF, Nielsen B, Knap PW. Genomic selection in commercial pig breeding. *Anim Front.* 2016;6:15–22.
25. Meuwissen T, Hayes B, Goddard M. Genomic selection: a paradigm shift in animal breeding. *Anim Front.* 2016;6:6–14.
26. Wolc A, Kranis A, Arango J, Settar P, Fulton JE, O'Sullivan NP, et al. Implementation of genomic selection in the poultry industry. *Anim Front.* 2016;6:23–31.
27. Raoul J, Swan AA, Elsen J-M. Using a very low-density SNP panel for genomic selection in a breeding program for sheep. *Genet Sel Evol.* 2017;49:76.
28. Wiggans GR, Cole JB, Hubbard SM, Sonstegard TS. Genomic selection in dairy cattle: the USDA experience. *Annu Rev Anim Biosci.* 2017;5:309–27.
29. Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. *J Dairy Sci.* 2009;92:4656–63.
30. Daetwyler HD, Calus MPL, Pong-Wong R, de los Campos G, Hickey JM. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics.* 2013;193:347–65.
31. Legarra A, Reverter A. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genet Sel Evol.* 2018;50:53.
32. David I, Huynh Tran VH, Gilbert H. New residual feed intake criterion for longitudinal data. *Genet Sel Evol.* 2021;53:53.
33. VanRaden PM. Efficient methods to compute genomic predictions. *J Dairy Sci.* 2008;91:4414–23.
34. Aguilar I, Misztal I, Tsuruta S, Legarra A, Wang H. PREGSF90—POSTGSF90: Computational tools for the implementation of single-step genomic selection and genome-wide association with ungenotyped individuals in BLUPF90 programs. In *Proceedings of the 10th Congress of Genetics Applied to Livestock Production: 17–22 August 2014; Vancouver.* 2014.
35. Gilmour AR, Gogel BJ, Cullis BR, Welham SJ, Thompson R. *ASReml User Guide Release 4.1 Functional Specification.* Hemel Hempstead: VSN International Ltd; 2015.
36. Macedo FL, Christensen OF, Astruc J-M, Aguilar I, Masuda Y, Legarra A. Bias and accuracy of dairy sheep evaluations using BLUP and SSGBLUP with metafounders and unknown parent groups. *Genet Sel Evol.* 2020;52:47.
37. Sorensen D, Fernando R, Gianola D. Inferring the trajectory of genetic variance in the course of artificial selection. *Genet Res.* 2001;77:83–94.
38. Aliakbari A, Delpuech E, Labrune Y, Riquet J, Gilbert H. The impact of training on data from genetically-related lines on the accuracy of genomic predictions for feed efficiency traits in pigs. *Genet Sel Evol.* 2020;52:57.
39. Hidalgo J, Lourenco D, Tsuruta S, Masuda Y, Miller S, Bermann M, et al. Changes in genomic predictions when new information is added. *J Anim Sci.* 2021;99:skab004.
40. Delpuech E, Aliakbari A, Labrune Y, Fève K, Billon Y, Gilbert H, et al. Identification of genomic regions affecting production traits in pigs divergently selected for feed efficiency. *Genet Sel Evol.* 2021;53:49.
41. Hollifield MK, Lourenco D, Bermann M, Howard JT, Misztal I. Determining the stability of accuracy of genomic estimated breeding values in future generations in commercial pig populations. *J Anim Sci.* 2021;99:ska085.
42. Martini JWR, Schrauf MF, Garcia-Baccino CA, Pimentel ECG, Munilla S, Rogberg-Muñoz A, et al. The effect of the H–1 scaling factors  $\tau$  and  $\omega$  on the structure of H in the single-step procedure. *Genet Sel Evol.* 2018;50:16.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

