AMERICAN SOCIETY OF ANIMAL SCIENCE

Journal of Animal Science, 2021, Vol. 99, No. 4, 1–8

doi:10.1093/jas/skab085
Advance Access publication March 17, 2021
Received: 30 November 2020 and **Accepted:** 15 March 2021
Animal Genetics and Genomics

ANIMAL GENETICS AND GENOMICS

Determining the stability of accuracy of genomic estimated breeding values in future generations in commercial pig populations

Mary Kate Hollifield,[†,1] Daniela Lourenco,[†] Matias Bermann,[†] Jeremy T. Howard,[‡] and Ignacy Misztal[†]

[†]Department of Animal and Dairy Science, University of Georgia, Athens, GA, USA, [‡]Smithfield Premium Genetics, Rose Hill, NC, USA

[1]Corresponding author: marykate.hollifield@uga.edu

ORCiD numbers: 0000-0003-3218-3174 (M. K. Hollifield); 0000-0002-5374-0710 (M. Bermann); 0000-0002-0382-1897 (I. Misztal).

Abstract

Genomic information has a limited dimensionality (number of independent chromosome segments [$\mathbf{M}_e$]) related to the effective population size. Under the additive model, the persistence of genomic accuracies over generations should be high when the nongenomic information (pedigree and phenotypes) is equivalent to M_e animals with high accuracy. The objective of this study was to evaluate the decay in accuracy over time and to compare the magnitude of decay with varying quantities of data and with traits of low and moderate heritability. The dataset included 161,897 phenotypic records for a growth trait (**GT**) and 27,669 phenotypic records for a fitness trait (**FT**) related to prolificacy in a population with dimensionality around 5,000. The pedigree included 404,979 animals from 2008 to 2020, of which 55,118 were genotyped. Two single-trait models were used with all ancestral data and sliding subsets of 3-, 2-, and 1-generation intervals. Single-step genomic best linear unbiased prediction (**ssGBLUP**) was used to compute genomic estimated breeding values (**GEBV**). Estimated accuracies were calculated by the linear regression (**LR**) method. The validation population consisted of single generations succeeding the training population and continued forward for all generations available. The average accuracy for the first generation after training with all ancestral data was 0.69 and 0.46 for GT and FT, respectively. The average decay in accuracy from the first generation after training to generation 9 was −0.13 and −0.19 for GT and FT, respectively. The persistence of accuracy improves with more data. Old data have a limited impact on the predictions for young animals for a trait with a large amount of information but a bigger impact for a trait with less information.

Key words: decay in accuracy, genomic selection, old data, predictive ability, young animals

Introduction

The addition of genomic information to routine genetic evaluations reduced generation interval and increased the accuracy of genomic estimated breeding value (**GEBV**), defined as the correlation between true and estimated breeding values (VanRaden et al., 2009). These factors are the main forces driving the increase in the rate of genetic gain over time (VanRaden, 2008; García-Ruiz et al., 2016). Genomic information helps to identify the best young animals accurately even before phenotypes are recorded; therefore, it is of interest to determine the accuracy of GEBV for generations without new

Abbreviations

BLUP	best linear unbiased prediction
EBV	estimated breeding value(s)
FT	fitness trait
GEBV	genomic estimated breeding value(s)
GRM	genomic relationship matrix
GT	growth trait
L	genome length
LR	linear regression, or Legarra–Reverter method
M_e	number of independent chromosome segments
N_e	effective population size
SNP	single nucleotide polymorphism
ssGBLUP	single step genomic best linear unbiased prediction

data recording and the magnitude of decay of accuracy over time. The selection of novel traits and traits difficult to measure is mainly dependent on the accuracies of GEBV. For example, milking speed and temperament have shown promising genetic progress due to genomics (Chen et al., 2020). Initial studies in genomic selection showed great persistence in the accuracy of genomic predictions over time. The results from the study of Meuwissen et al. (2001) showed marginal decay in accuracy with a decrease from 0.84 to 0.72 over five new generations without phenotypes. This created initial excitement for the potential of selection with genomic information; however, the parameters of the simulated population cannot be compared with present-day commercial livestock populations. In the simulation, there was no selection, and only a few major genes explained the additive genetic variance of the trait. Under strong selection, steep decay in accuracy occurs (Muir, 2007). In small, simulated populations, Muir (2007) found that the accuracy of GEBV decays more rapidly than expected when under strong selection compared with random selection.

We hypothesize that the decay will be minimized even under selection if enough phenotypes and genotypes are available to represent the population structure. The reason is that a limited number of independent chromosome segments (**M_e**) theoretically explain the additive genetic variance in a population (Pocrnic et al., 2016a). Therefore, if enough information exists to precisely estimate the effects of M_e, the additive genetic variance can be explained, and accuracies will be adequate and stable over time. The number of M_e is dependent on the effective population size (**N_e**) and genome length (**L**) (Stam, 1980). Pocrnic et al. (2016a) showed that the optimal amount of M_e can be estimated by computing the number of eigenvalues that explain a certain proportion of variation in the genomic relationship matrix (**GRM**), which is used in genomic best linear unbiased prediction (**GBLUP**; VanRaden, 2008) and single-step GBLUP (**ssGBLUP**; Aguilar et al., 2010). This creates a threshold for the amount of information that is nonredundant, that is, information that can increase accuracy, and the amount of which new data no longer increases the accuracy. Hence, the GRM has a limited dimension. Whereas N_eL eigenvalues explain most of the information, no new information is added after $4N_eL$ (Stam, 1980; Pocrnic et al., 2016a). Goddard (2009) showed that accuracy is inversely related to N_e. As N_e increases, accuracy decreases. It is estimated that genome lengths for pigs range from 18 to 23 Morgan (Rohrer et al., 1994; Archibald et al., 1995; Marklund et al., 1996; Tortereau et al., 2012), and N_e ranges from 55 to 113 (Welsh et al., 2009; Uimari and Tapio, 2011; Pocrnic et al., 2016b). Pocrnic et al. (2016b) found that 5,000 segments explain approximately 98% of the variation in commercial pig populations. With enough data relative to the independent chromosome segments, high accuracy could be achieved. Additionally, if the segments are well estimated, there should be less decay of predictivity under the additive model even under selection.

The inverse of the GRM can be obtained by recursion on a group of animals (Faux et al., 2012; Misztal et al., 2014), with the optimal group size equal to the dimensionality of the genomic information (Misztal, 2016). The recursion means that the breeding value of any animal can be estimated with near-perfect accuracy from the exact breeding values of $4N_eL$ other animals. Bradford et al. (2017) showed by simulation that the accuracy of GEBV was the same whether the recursion was based on animals from the last generation or a distant generation. Their results suggest that, under the additive model, the persistence of genomic evaluations is very high if the reference population includes $4N_eL$ animals with high accuracy or equivalent.

Although accuracy is dependent on the proportion of variance explained by the eigenvalues of the GRM, the distribution of eigenvalues is not consistent, and a small percentage of the largest eigenvalues explain the majority of the genetic variation (Pocrnic et al., 2019). Additionally, the animals necessary to explain the largest eigenvalues carry almost the same genomic information. Hence, selection by GBLUP-based models occurs on clusters of independent chromosome segments, not individual chromosome segments (Pocrnic et al., 2019). In pig populations, the segments can be well estimated if there are around 5,000 animals available with very high accuracy (e.g., theoretical EBV accuracy based on prediction error variance) or an equivalent number of animals with less accuracy. Despite a large amount of data available, the decay will be more dramatic if genomic selection induces faster epistatic changes (Huang and Mackay, 2016). Epistatic interactions between genes may reduce the value of old data, and epistatic effects may be unstable across populations because of the fluctuation in allele frequencies (Varona et al., 2018).

With the commercial pig production systems and population structure, the N_e and the M_e are small. The purpose of this study is to determine how accuracy and the decay in accuracy are affected by the quantity of data available, the heritability of the trait, and removing data from ancestral generations. With genotypes now available for many generations in pigs, reliable predictions for generations without new phenotype recordings may be possible.

Materials and Methods

Animal Care and Use Committee approval was not needed because information was obtained from preexisting databases.

Data

Data for animals born between 2008 and 2020 were provided by Smithfield Premium Genetics (Rose Hill, NC). The population consisted of 273,382 animals, of which 55,118 were genotyped or imputed to the 50k single-nucleotide polymorphism (**SNP**) panel for autosomal markers only. Quality control removed SNP with minor allele frequency lower than 0.05, SNP and animals with call rates lower than 0.9, SNP with the difference between expected and observed frequency of heterozygous greater than 0.15 (departure from the Hardy–Weinberg equilibrium), and animals with parent-progeny Mendelian conflicts. After quality control, 39,263 SNPs remained for 53,147 genotyped animals.

The dataset consisted of 27,669 records for a repeated fitness trait (**FT**) related to prolificacy from 13,883 animals and 161,495 records for a single growth trait (**GT**). The population consisted of 11 generations. Generations were constructed by tracing the population back to the oldest animals with no recorded parents. These animals were considered generation 1, and their progeny, grand-progeny, and great-grand-progeny were placed in generations 2, 3, and 4, respectively, and continued until generation 11. The birth year of the animals without parent records was considered when joining the successions to be more precise and to account for the age variation of animals without parent records. Table 1 presents the number of animals with genotypes, phenotypes, and pedigree per generation.

Model and analyses

Variance components were estimated using AIREMLF90 (Misztal et al., 2014) without genomic information. The heritabilities were 0.21 and 0.06 for GT and FT, respectively, with standard errors less than 0.01. GEBVs were computed using ssGBLUP (Aguilar et al., 2010). Two single-trait models were used in the analyses :

$$\mathbf{y}_{GT} = \mathbf{X}_{GT}\mathbf{b}_{GT} + \mathbf{Z}\boldsymbol{u}_{GT} + \mathbf{W}_1\mathbf{cl}_{GT} + \boldsymbol{e}_{GT} \quad (1)$$

$$\mathbf{y}_{FT} = \mathbf{X}_{FT}\mathbf{b}_{FT} + \mathbf{Z}\boldsymbol{u}_{FT} + \mathbf{W}_2\boldsymbol{pe}_{FT} + \boldsymbol{e}_{FT}, \quad (2)$$

where $\mathbf{y}_{GT}$ is a vector of GT observations; $\mathbf{b}_{GT}$ is a fixed vector of systematic effects, including contemporary group (farm, year, and week of birth), sex, and age in days at recording; $\boldsymbol{u}_{GT}$ and $\mathbf{cl}_{GT}$ are random vectors of direct additive genetic and common litter effects, respectively. Elements of $\mathbf{y}$ are related to elements of $\mathbf{cl}_{GT}$ by the incidence matrix $\mathbf{W}_1$. The $\mathbf{y}_{FT}$ is a vector of FT observations; $\mathbf{b}_{FT}$ is a fixed vector of systematic effects including contemporary group (farm, year, and month of birth) and parity; $\boldsymbol{u}_{FT}$ and $\boldsymbol{pe}_{FT}$ are random vectors of direct additive genetic and permanent environmental effects, respectively. Elements of $\mathbf{y}_{FT}$ are related to elements of $\boldsymbol{pe}_{FT}$ by the incidence matrix $\mathbf{W}_2$. In both models, $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices relating elements of $\mathbf{y}$ to $\mathbf{b}$ and $\boldsymbol{u}$, respectively, and $\boldsymbol{e}$ is a vector of random residuals. The covariance matrices were assumed to be:

$$Var\begin{bmatrix}\boldsymbol{u}_{GT}\\ \mathbf{cl}_{GT}\\ \boldsymbol{e}_{GT}\end{bmatrix} = \begin{bmatrix}\mathbf{H}\sigma^2_{uGT} & 0 & 0\\ 0 & \mathbf{I}\sigma^2_{cl} & 0\\ 0 & 0 & \mathbf{I}\sigma^2_{eGT}\end{bmatrix} \quad (3)$$

$$Var\begin{bmatrix}\boldsymbol{u}_{FT}\\ \boldsymbol{pe}_{FT}\\ \boldsymbol{e}_{FT}\end{bmatrix} = \begin{bmatrix}\mathbf{H}\sigma^2_{uFT} & 0 & 0\\ 0 & \mathbf{I}\sigma^2_{pe} & 0\\ 0 & 0 & \mathbf{I}\sigma^2_{eFT}\end{bmatrix}, \quad (4)$$

where σ^2_{uGT} and σ^2_{uFT} are variances for additive genetic effects for GT and FT, respectively; σ^2_{cl} is the variance for the common litter effect; σ^2_{pe} is the variance for the permanent environmental effect; σ^2_{eGT} and σ^2_{eFT} are the variances for the residual effects for GT and FT, respectively; $\mathbf{I}$ is the identity matrix; $\mathbf{H}$ is a matrix combining pedigree and genomic relationships among animals as applied in ssGBLUP (Aguilar et al., 2010). The inverse of the pedigree-based relationship matrix $(\mathbf{A}^{-1})$ is replaced by the inverse of $\mathbf{H}$ $(\mathbf{H}^{-1})$ in the ssGBLUP mixed model equations, which is written as follows:

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

where $\mathbf{G}$ was constructed using the first method of VanRanden (2008), then 95% of $\mathbf{G}$ was blended with 5% of the pedigree relationship matrix for genotyped animals $(\mathbf{A}_{22})$, and finally tuned so that the means of the diagonal and off-diagonal elements were similar to those of $\mathbf{A}_{22}$ (Chen et al., 2011). The allele frequencies used to compute $\mathbf{G}$ were calculated based on all genotyped animals in the dataset.

In this study, the accuracy and dispersion of GEBV were estimated with the linear regression (**LR**) method (Legarra and Reverter, 2018). This method uses two datasets, namely the *whole* dataset and the *partial* dataset, hereinafter denoted with the subscripts *w* and *p*, respectively. The former contains all the available phenotypes up to a certain time *t*, whereas the latter contains phenotypes up to a time period before *t*. The focal individuals, that is, the individuals for whom the accuracy of GEBV will be estimated, are defined as the genotyped animals with phenotypes in the whole dataset but without in the partial dataset.

To investigate the impact of the amount of data on the accuracy of GEBV for focal individuals, GEBV were sequentially estimated by changing the definition of focal individuals and partial datasets using a sliding approach based on generation. Figure 1 shows four definitions of focal groups that included generations 5 to 9, 6 to 9, 7 to 9, and 8 and 9. Accuracy and dispersion were then calculated separately for each generation of focal individuals. Additionally, to investigate the impact of ancestral data, four partial datasets were created for each focal group: 1) the *ancestral group*: contained all the ancestors of the focal individuals, 2) the *3-generation group*: consisted of the ancestors up to the great-grandparents of the focal individuals, 3) the *2-generation group*: included the grandparents and parents of the focal individuals, and 4) the *1-generation group*: contained only the parents of the focal individuals. A total of 16 different combinations of groups of focal individuals and partial datasets were created (Figure 1).

The benchmark for each validation, that is, GEBV_w, remained unchanged, whereas GEBV_p was updated as the partial datasets were modified. Due to the lack of phenotypes and genotypes in generations 10 and 11, these animals were removed from all analyses as they were incomparable with the other validation generations. Accuracies were estimated for each generation in each set of focal individuals using: $\hat{\rho}_{\text{cov(w,p)}} = \sqrt{\frac{\text{cov}(\hat{\boldsymbol{u}}_w, \hat{\boldsymbol{u}}_p)}{(1-\bar{F})\hat{\sigma}^2_u}}$ (Legarra and Reverter, 2018; Macedo et al., 2020b), where $\bar{F}$ is the average inbreeding coefficient

Table 1. Number of animals in the pedigree, genotyped animals, and records for GT and FT per generation

Generation	Pedigree	Genotypes	GT	FT
1	758	214	658	1,991
2	12,513	384	4,767	2,098
3	15,190	831	7,697	3,447
4	29,017	1,929	16,491	3,753
5	38,316	2,775	23,211	4,302
6	42,476	6,158	26,474	4,278
7	44,363	10,769	28,260	3,348
8	39,082	11,345	25,002	2,290
9	27,445	8,636	16,989	1,435
10	17,084	6,149	8,762	570
11	7,138	3,957	3,184	157

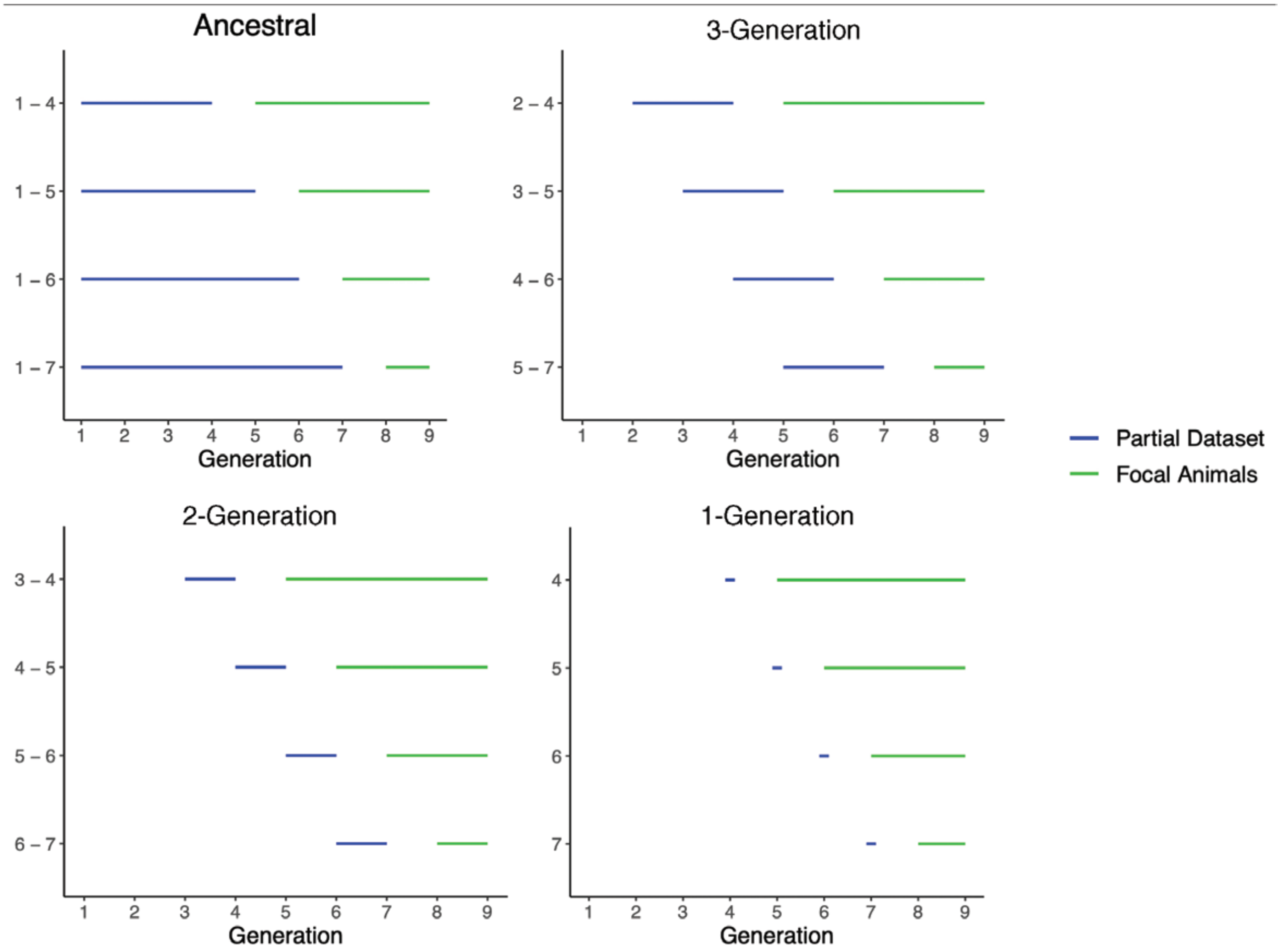

Figure 1. Scheme for partial datasets and focal animals. The four partial dataset groups include ancestral, 3-, 2-, and 1-generation subsets. In each scenario, the genomic and pedigree information is included for all animals and remains unchanged, but only phenotypes exist for animals in the partial dataset. Generations are not grouped for the focal animals, and accuracies are calculated for each generation separately.

among focal individuals in a specific generation and $\hat{\sigma}_u^2$ is the estimated additive genetic variance of the population. Inbreeding coefficients for each animal were calculated with a recursive method based on pedigree using INBUPGF90 (Aguilar and Misztal, 2008). The slope of the regression of $\hat{\mathbf{u}}_w$ on $\hat{\mathbf{u}}_p$ is used to assess the dispersion of partial GEBV and is equal to $b_{w,p} = \frac{cov(\hat{\mathbf{u}}_w,\hat{\mathbf{u}}_p)}{var(\hat{\mathbf{u}}_p)}$. The primary purpose of this research was to compare accuracies over time with varying amounts of ancestral data for two traits of differing heritabilities; therefore, other statistical parameters were not used. Accuracy and dispersion are well researched and logical to use as a function over time (Macedo et al., 2020a). Additional statistics proposed by the LR method have not been widely tested as a function of time. Including those values would output uninterpretable comparisons and should be further researched.

Results and Discussion

Figures 2 and 3 show the accuracy for GT and FT over time using the partial datasets belonging to each group. When comparing traits, GT had higher accuracy and less decay in accuracy over time compared with FT. For example, when considering the partial dataset composed of generations 1 to 4 from the ancestral group, the accuracy decreased from 0.55 in generation 5 to 0.42 in generation 9 for GT (Figure 2A), and from 0.46 to 0.22 for FT (Figure 3A), respectively. These results are expected and agree with those from Muir (2007) since GT has higher heritability than FT, and low heritability traits require a large number of records to achieve high accuracy; FT had about one-sixth of the records compared with GT.

Persistence for both traits can be inferred by observing the initial and final accuracy for each line in Figures 2 and 3. The slopes for FT are greater in magnitude than the slopes for GT, meaning that the latter showed more persistence. The differences in persistence between the two traits may be explained by the heritability and the amount of phenotypic information. Roughly, the amount of information in this study can be approximated as accuracies of hypothetical 5,000 ($4N_eL$) sires with as many progeny as the number of animals with records and with progeny equally distributed per sire. For a trait with 32 progeny per sire and heritability of 0.21, the accuracy per sire would be approximately 0.80. For a trait with five progeny per sire and heritability of 0.06, the equivalent accuracy would be only 0.25.

The distance between different lines in Figures 2 and 3 shows the impact that the different sources of information, namely parents, grandparents, etc., have on the estimation of the accuracy of GEBV. This fact can be observed for the focal

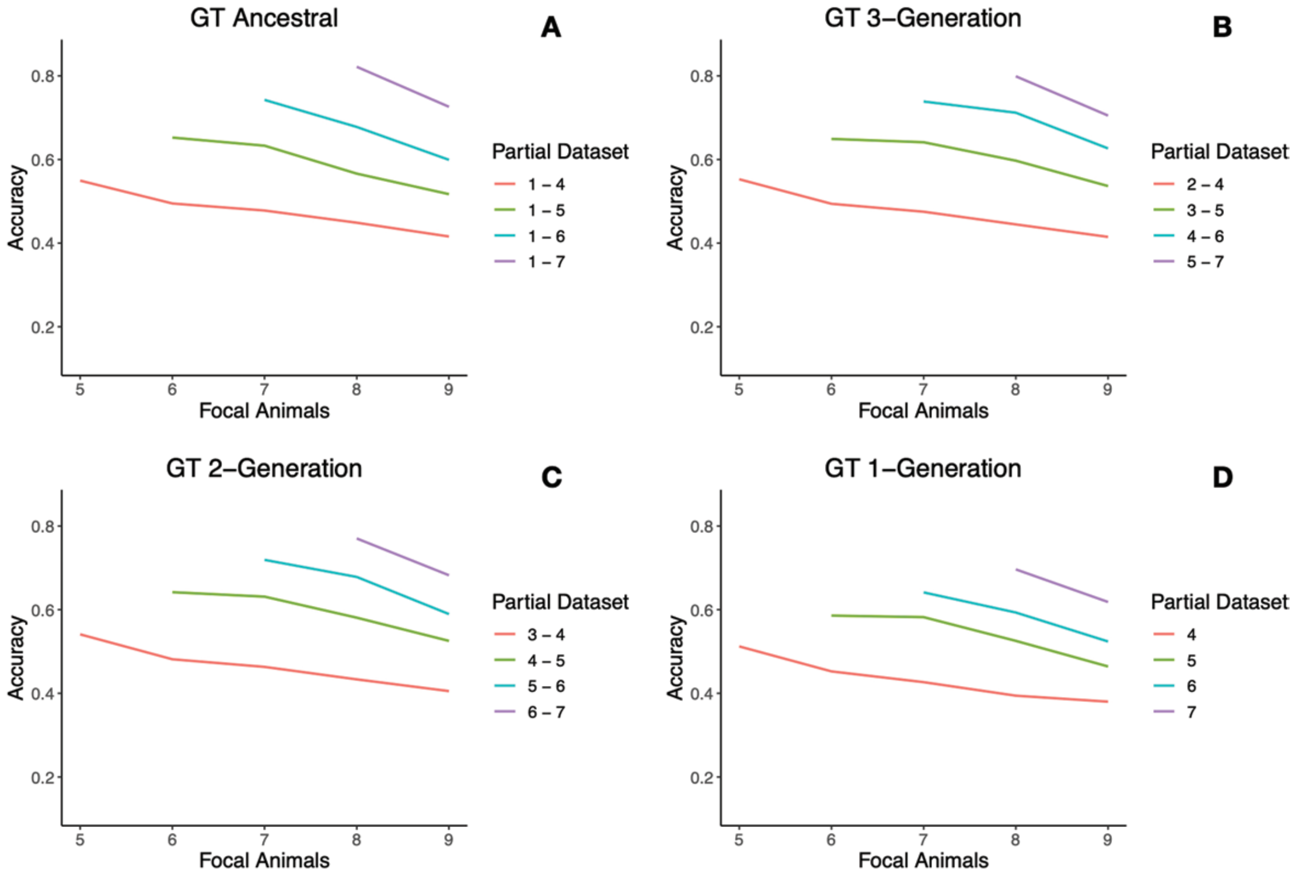

Figure 2. Accuracy over time with four partial dataset groups for GT. The partial datasets are updated over time, increasing a generation of data for the ancestral groups (A) and adding a recent generation of data while removing the oldest generation of data for 3-, 2-, and 1- generation subsets (B, C, and D, respectively). Accuracy is calculated for each generation separately, beginning with the first generation following the partial dataset and ending at generation 9.

individuals in generation 8 (Figures 2 and 3). In this case, the purple line includes the parents of the named focal individuals, whereas, for the blue line, the closest generation used to estimate their accuracies was that of their grandparents. When comparing the difference between both lines, it can be deduced that removing the parents drops the accuracy for about 0.11, on average for GT, whereas the average drop for FT was about 0.04. To compare the two traits across time, the average decreases in accuracy for GT (FT) were 16.0% (10.1%) after removing parents and 79.3% (34.4%) after removing three generations (parents, grandparents, and great-grandparents).

The magnitude and slope of the regression of $\hat{\boldsymbol{u}}_w$ on $\hat{\boldsymbol{u}}_p$ over time for both traits explain the effect of heritability and quantity of data on GEBV prediction. Regression coefficient less than one indicates that the GEBV of the focal animals are over-dispersed (overestimated) compared with GEBV from the whole dataset. In Figure 4, the partial datasets include generations 1 through 4 for both traits. The partial datasets are not updated over time; therefore, the focal animals become less related to the partial datasets as generations proceed. In relation to animals in generation 4, the GEBV for focal animals were overestimated for progeny, grand-progeny, great-grand-progeny, great-great-grand-progeny, and great-great-great-grand-progeny, which are generations 5, 6, 7, 8, and 9, respectively. Analogously to accuracy, $b_{w,p}$ remained greater and more persistent over time for GT than FT. The $b_{w,p}$ decreased from 0.84 to 0.66 for GT from generations 5 and 9, respectively. Similarly, it decreased from 0.63 to 0.21 for FT. A steep negative trend for $b_{w,p}$ over time indicates that there was not enough information available to predict the amount of dispersion in further generations. The differences in the persistence of accuracy and dispersion confirm that for traits with low heritability, the impact of information from closely related individuals is less than traits with high heritability.

Apparently, this is subject to the fact that all chromosome segments are represented in the population (Pocrnic et al., 2016a). Thus, with sufficient genotyped animals, it is expected that chromosome segments would be well represented in the population. Consequently, the gain in accuracy when adding information from individuals more closely related will be minimal if the corresponding trait has low heritability. It is important to highlight that, in this study, the accumulation of ancestors was considered a new source of information, not the addition of progeny of the focal individuals. Logically, the accuracy of GEBV for focal individuals will largely depend on the incorporation of their progeny in the genetic evaluation, regardless of the heritability of the trait and the representation of the chromosome segments in the population.

To maximize the accuracy of genomic predictions, an optimal size of the training population is necessary to capture most of the variation in the population. This optimal subset is theoretically related to a limited dimension of the genomic information. This limited dimension is a function of N_e and L. If ~4N_eL largest eigenvalues are contained in the GRM, the M_e is likely obtained, and ample information is provided to achieve high accuracies (Pocrnic et al., 2016a). According to Misztal (2016), each independent chromosome segment has an additive

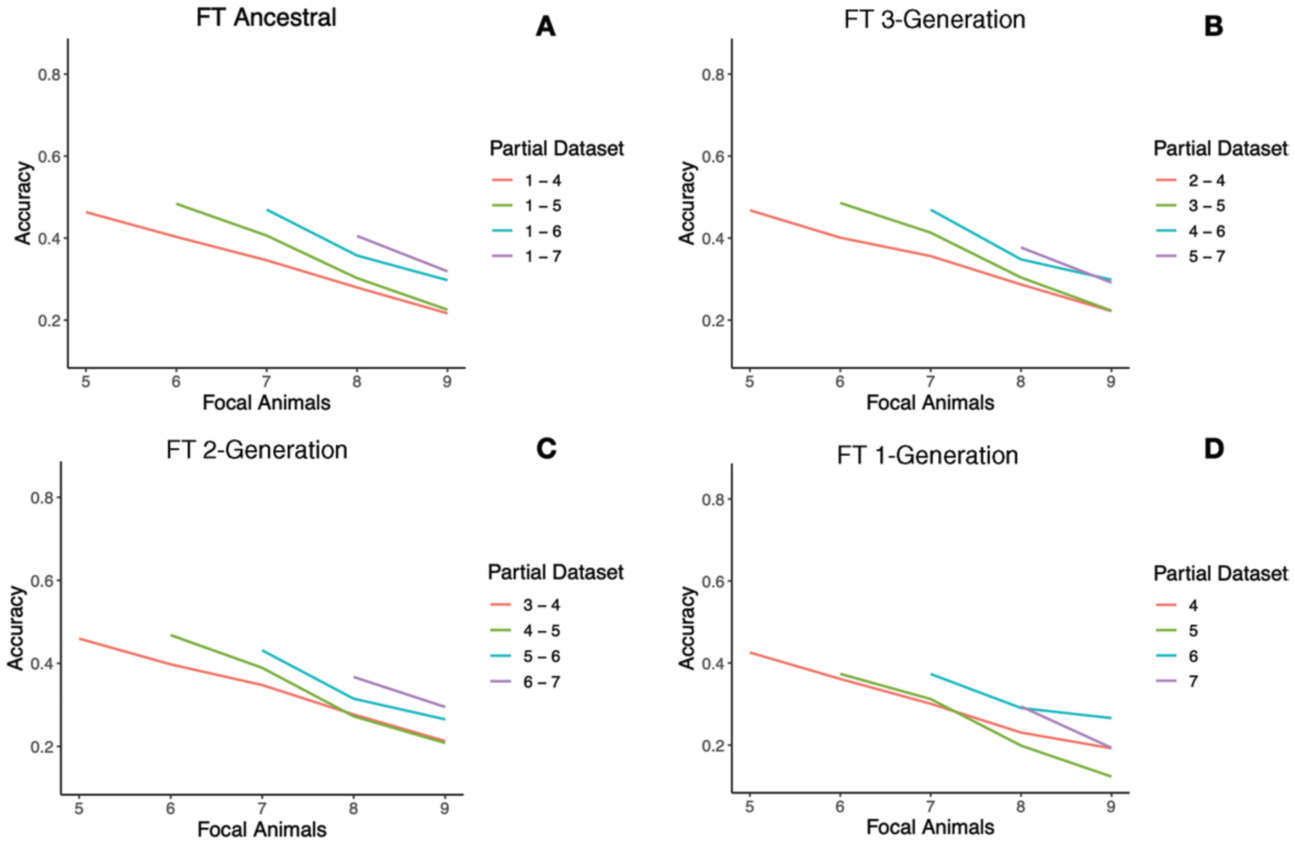

Figure 3. Accuracy over time with four partial dataset groups for FT. The methods are the same as in Figure 2.

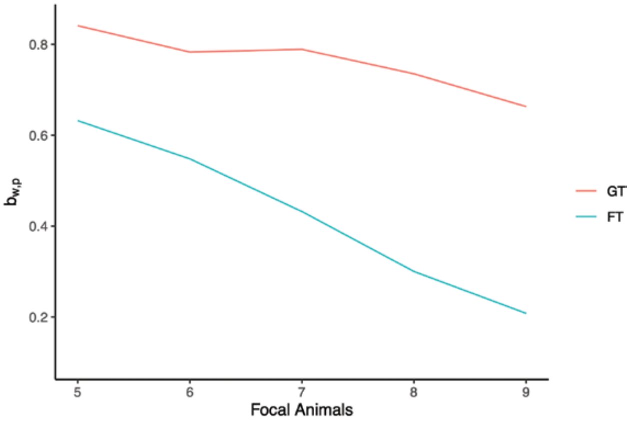

Figure 4. Dispersion trends over time for GT and FT. The partial datasets include ancestral data from generations 1 to 4 and are not updated over time. Each generation beyond generation 4 is a generation of focal animals becoming less related to the partial dataset animals. The slope of the regression of GEBV whole on GEBV partial ($b_{w,p}$) was used to estimate dispersion. Dispersion was calculated for each generation separately, beginning with generation 5 and ending at generation 9.

effect, and the sum of the effects of the existing chromosome segments in individual animals composes the breeding values. If enough chromosome segment effects are captured in the population, more variation is explained in the population, and thus, it is expected that accuracies will also show more persistence over time.

As explained in a study conducted by Hayes et al. (2009), the accuracy of genomic selection is crucially dependent on the number of phenotypic records available and the heritability of a trait. In their study, approximately 5,000 phenotypes were required to achieve an accuracy of GEBV equal to 0.6 for a trait with a heritability of 0.2 in a population with an N_e of 1,000. In our study, for FT, generations 6 and 7 contained 4,278 and 3,348 records, respectively. Compared with GT that had 26,474 records for generation 6 and 28,260 for generation 7, it can be concluded that FT does not have enough information to achieve an accuracy as high as GT. This can explain the lack of persistency and low accuracy over time when analyzing FT with 1-generation partial datasets. In every analysis for FT and GT, 2 or 3 generations of data seem sufficient enough to reach a comparable maximum accuracy to all ancestral data. As heritability decreases, the number of required phenotypic records to achieve the desired accuracy of GEBV increases (Hayes et al., 2009).

The selection pressure and complexity of a trait significantly affect the accuracy of GEBV over time (Muir, 2007; Gorjanc et al., 2015). In this study, different intensities and types of selection pressure were placed on the two separate traits. GT was heavily selected upon over time, and this trait was directly selected across all generations. FT, however, was only indirectly selected, meaning that the selection pressure on FT depended on the selection pressure of a different trait with a more favorable relationship with preweaning mortality. These differences in selection for both traits can be observed in Figure 5, where the genetic trends of GEBV across generations for both GT and FT are shown. To make both traits comparable, GEBV were standardized. As seen in the trends over time, GT increased at a steadier rate, whereas FT increased less directly, implying less selection. Also, FT is more challenging to select upon and predict its performance since it is a categorical trait, compared with the continuity of GT.

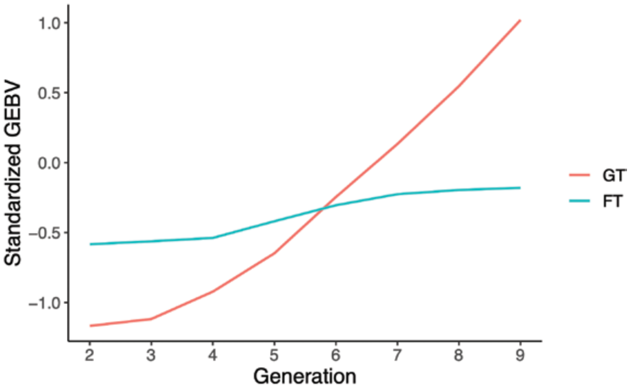

Figure 5. Genetic trends for GT and FT with average standardized GEBV. Generation 1 was excluded from the trend due to the lack of animals with phenotypic records.

One important limitation of this is that the accuracy for generations that were distant from the reference populations was computed for preselected animals, and preselection decreases realized accuracies (Bijma, 2012; Lourenco et al., 2015). Therefore, the future accuracies may be underestimated, although the LR method may partially account for the preselection.

The issue of persistence of GEBV is also important in the dairy industry where young bulls are selected from other young bulls only based on the genomic information. For Holsteins with a large amount of information and the genomic dimensionality of around 15,000 (Pocrnic et al., 2016b), the reliability for production traits two generations ahead of the reference population was 90% of that of one generation ahead (VanRaden et al., 2010). If the persistence of the evaluations is high, the importance of phenotyping may be reduced. However, the persistence is likely to be lower for lower heritability traits, especially with fewer records, keeping phenotyping relevant. Additionally, in the long run, very strong selection and epistatic interactions may possibly reduce the persistence, keeping the need for phenotype recording.

Conclusions

When the reference population is large enough to accurately estimate the effects of the independent chromosome segments, GEBV can be persistent, with minimal decay of accuracy over generations. In such a case, the impact of old data is minimal. The decay is larger with less information, particularly for lower heritability traits, and with necessarily lower selection pressure, the impact of old data is likely larger. It would be desirable to estimate the decay as a function of many parameters analytically; however, the complexity of selection and side effects of faster selection (e.g., Bulmer effect and epistasis) are likely to make such a theory complex.

Acknowledgment

This study was supported by Smithfield Premium Genetics, Rose Hill, NC.

Conflict of interest statement

The authors declare no real or perceived conflicts of interest.

Literature Cited

Aguilar, I., and I. Misztal. 2008. Technical Note: Recursive algorithm for inbreeding coefficients assuming nonzero inbreeding of unknown parents. *J. Dairy Sci.* **91**:1669–1672. doi:10.3168/jds.2007-0575

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot Topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.* **93**(2):743–752. doi:10.3168/jds.2009-2730

Archibald, A. L., C. S. Haley, J. F. Brown, S. Couperwhite, H. A. McQueen, D. Nicholson, W. Coppieters, A. Van de Weghe, A. Stratil, A. K. Winterø, *et al*. 1995. The PiGMaP consortium linkage map of the pig (*Sus scrofa*). *Mamm. Genome* **6**(3):157–175. doi:10.1007/BF00293008

Bijma, P. 2012. Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations. *J. Anim. Breed. Genet.* **129**:345–358. doi:10.1111/j.1439-0388.2012.00991.x

Bradford, H. L., I. Pocrnić, B. O. Fragomeni, D. A. L. Lourenco, and I. Misztal. 2017. Selection of core animals in the algorithm for proven and young using a simulation model. *J. Anim. Breed. Genet.* **134**:545–552. doi:10.1111/jbg.12276

Chen, C. Y., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2011. Effect of different genomic relationship matrices on accuracy and scale. *J. Anim. Sci.* **89**:2673–2679. doi:10.2527/jas.2010-3555

Chen, S. Y., H. R. Oliveira, F. S. Schenkel, V. B. Pedrosa, M. G. Melka, and L. F. Brito. 2020. Using imputed whole-genome sequence variants to uncover candidate mutations and genes affecting milking speed and temperament in Holstein cattle. *J. Dairy Sci.* **103**:10383–10398. doi:10.3168/jds.2020-18897

Faux, P., N. Gengler, and I. Misztal. 2012. A recursive algorithm for decomposition and creation of the inverse of the genomic relationship matrix. *J. Dairy Sci.* **95**:6093–6102. doi:10.3168/jds.2011-5249

García-Ruiz, A., J. B. Cole, P. M. VanRaden, G. R. Wiggans, F. J. Ruiz-López, and C. P. Van Tassell. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proc. Natl. Acad. Sci. U. S. A.* **113**(28):E3995–E4004. doi:10.1073/pnas.1519061113

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**:245–257. doi:10.1007/s10709-008-9308-0

Gorjanc, G., P. Bijma, and J. M. Hickey. 2015. Reliability of pedigree-based and genomic evaluations in selected populations. *Genet. Sel. Evol.* **47**:65. doi:10.1186/s12711-015-0145-1

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited Review: Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* **92**:433–443. doi:10.3168/jds.2008-1646

Huang, W., and T. F. Mackay. 2016. The genetic architecture of quantitative traits cannot be inferred from variance component analysis. *PLoS Genet.* **12**:e1006421. doi:10.1371/journal.pgen.1006421

Legarra, A., and A. Reverter. 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genet. Sel. Evol.* **50**:53. doi:10.1186/s12711-018-0426-6

Lourenco, D. A., B. O. Fragomeni, S. Tsuruta, I. Aguilar, B. Zumbach, R. J. Hawken, A. Legarra, and I. Misztal. 2015. Accuracy of estimated breeding values with genomic information on males, females, or both: an example on broiler chicken. *Genet. Sel. Evol.* **47**:56. doi:10.1186/s12711-015-0137-1

Macedo, F. L., O. F. Christensen, J. M. Astruc, I. Aguilar, Y. Masuda, and A. Legarra. 2020a. Bias and accuracy of dairy sheep evaluations using BLUP and SSGBLUP with metafounders and unknown parent groups. *Genet. Sel. Evol.* **52**:47. doi:10.1186/s12711-020-00567-1

Macedo, F. L., A. Reverter, and A. Legarra. 2020b. Behavior of the linear regression method to estimate bias and accuracies with correct and incorrect genetic evaluation models. *J. Dairy Sci.* **103**:529–544. doi:10.3168/jds.2019-16603

Marklund, L., M. Johansson Moller, B. Høyheim, W. Davies, M. Fredholm, R. K. Juneja, P. Mariani, W. Coppieters, H. Ellegren, and L. Andersson. 1996. A comprehensive linkage map of the pig based on a wild pig-Large White intercross. *Anim. Genet.* **27**:255–269. doi:10.1111/j.1365-2052.1996.tb00487.x

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**:1819–1829.

Misztal, I. 2016. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics* **202**:401–409. doi:10.1534/genetics.115.182089

Misztal, I., S. Tsuruta, D. A. L. Lourenco, I. Aguilar, A. Legarra, and Z. Vitezica 2014. Manual for BLUPF90 family of programs. Available from http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all7.pdf. Accessed August 16, 2020.

Muir, W. M. 2007. *Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters*. Germany: Blackwell Publishing Ltd.; p. 342.

Pocrnic, I., D. A. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2016a. The dimensionality of genomic information and its effect on genomic prediction. *Genetics* **203**:573–581. doi:10.1534/genetics.116.187013

Pocrnic, I., D. A. Lourenco, Y. Masuda, and I. Misztal. 2016b. Dimensionality of genomic information and performance of the algorithm for proven and young for different livestock species. *Genet. Sel. Evol.* **48**:82. doi:10.1186/s12711-016-0261-6

Pocrnic, I., D. A. L. Lourenco, Y. Masuda, and I. Misztal. 2019. Accuracy of genomic BLUP when considering a genomic relationship matrix based on the number of the largest eigenvalues: a simulation study. *Genet. Sel. Evol.* **51**:75. doi:10.1186/s12711-019-0516-0

Rohrer, G. A., L. J. Alexander, J. W. Keele, T. P. Smith, and C. W. Beattie. 1994. A microsatellite linkage map of the porcine genome. *Genetics* **136**:231–245.

Stam, P. 1980. The distribution of the fraction of the genome identical by descent in finite random mating populations. *Genet. Res.* **35**(2):131–155. doi:10.1017/S0016672300014002

Tortereau, F., B. Servin, L. Frantz, H. J. Megens, D. Milan, G. Rohrer, R. Wiedmann, J. Beever, A. L. Archibald, L. B. Schook, *et al*. 2012. A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC Genomics* **13**:586. doi:10.1186/1471-2164-13-586

Uimari, P., and M. Tapio. 2011. Extent of linkage disequilibrium and effective population size in Finnish Landrace and Finnish Yorkshire pig breeds. *J. Anim. Sci.* **89**:609–614. doi:10.2527/jas.2010-3249

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **91**:4414–4423. doi:10.3168/jds.2007-0980

VanRaden, P., J. O′Connell, G. R. Wiggans, and K. Weigel. 2010. Combining different marker densities in genomic evaluation. *Interbull Bull.* **42**.

VanRaden, P. M., C. P. Van Tassel, G. R. Wiggians, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* **92**(1):16–24. doi:10.3168/jds.2008-1514

Varona, L., A. Legarra, M. A. Toro, and Z. G. Vitezica. 2018. Non-additive effects in genomic selection. *Front. Genet.* **9**:78. doi:10.3389/fgene.2018.00078

Welsh, C. S., H. D. Blackburn, and C. Schwab. 2009. Population status of major U. S. swine breeds. Proceedings of American Society of Animal Science Western Section; June 16–18, 2009; Fort Collins, CO. 60. p. 42–45.