animals

MDPI

Article

A study of Genomic Prediction across Generations of Two Korean Pig Populations

Beatriz Castro Dias Cuyabano [1,*], Hanna Wackel [1], Donghyun Shin [2] and Cedric Gondro [1]

[1] Department of Animal Science, Michigan State University, East Lansing, MI 48824, USA
[2] The Animal Molecular Genetics and Breeding Center, Chonbuk National University, Jeonju 54932, Korea
[*] Correspondence: castrod7@msu.edu

Received: 29 June 2019; Accepted: 28 August 2019; Published: 11 September 2019

Simple Summary: Commercial genotyping has become accessible at a relatively low cost and nowadays it is widely used by breeders to predict production and economic traits. Many studies explored the benefits of using DNA information in breeding programs, and many methods have been established to optimize the use of such information. To date, however, very few studies have explored how prediction accuracies change across generations. Here we present a short evaluation across five generations in two pig breeds and evaluate the accuracy of the prediction of relevant production traits using different generational groups.

Abstract: Genomic models that incorporate dense marker information have been widely used for predicting genomic breeding values since they were first introduced, and it is known that the relationship between individuals in the reference population and selection candidates affects the prediction accuracy. When genomic evaluation is performed over generations of the same population, prediction accuracy is expected to decay if the reference population is not updated. Therefore, the reference population must be updated in each generation, but little is known about the optimal way to do it. This study presents an empirical assessment of the prediction accuracy of genomic breeding values of production traits, across five generations in two Korean pig breeds. We verified the decay in prediction accuracy over time when the reference population was not updated. Additionally we compared the prediction accuracy using only the previous generation as the reference population, as opposed to using all previous generations as the reference population. Overall, the results suggested that, although there is a clear need to continuously update the reference population, it may not be necessary to keep all ancestral genotypes. Finally, comprehending how the accuracy of genomic prediction evolves over generations within a population adds relevant information to improve the performance of genomic selection.

Keywords: animal breeding; SNP-chips; genomic evaluation; generations

1. Introduction

Genomic models that incorporate dense single nucleotide polymorphism (SNP) marker information are widely used for the prediction of genomic values [1] to select animals and plants in breeding programs [2–5] or to predict susceptibility to diseases in humans [6,7].

Within a genotyped population, the extent of the genomic relationships [8,9], linkage disequilibrium (LD), and co-segregation of QTL with markers [10] contribute to the accuracy of genomic predictions. In livestock populations such as beef and dairy cattle, sheep, and pigs, the level of relatedness between individuals is higher compared to the relatedness observed in other populations, such as humans.

Genomic relationships are higher within families, rather than across families, and genomic prediction in breeding populations generally relies on the genotypes and phenotypes from ancestral individuals to predict the breeding values of subsequent generations. In fact, this practice of using information from previous generations to predict the next one has been in use for many years and predates the genomic technologies. Before genotypes were available, only the known family relationships from the current and previous generations (in the form of a pedigree relationship matrix) were combined with recorded phenotypes, to predict breeding values.

When genomic prediction is performed over successive generations of the same population, the prediction accuracy is expected to decay if the reference population is not updated, even if the base population initially genotyped was large, due to the decline in the extent of the relationships between reference and test individuals, and the breakdown of the linkage disequilibrium (LD) between SNPs and the quantitative trait loci (QTL) across generations [11].

The practical challenges of updating the reference population across generations to find a balance between genotyping costs and the economic gain from a more accurate selection was discussed by Pszczola and Calus [11], where it was empirically evaluated how many individuals should be updated in the reference population in each generation, in order to maintain a stable prediction accuracy. That evaluation was performed on a simulated population that resembled the genetic structure of dairy cattle. We are not aware of studies reporting results with real data across a large number of genotyped individuals for all generations. Weng et al. [12] evaluated a large number of generations in a layer chicken population, however most of the predictions were pedigree-based, rather than genotype-based.

In this study we performed an empirical evaluation of the prediction accuracy of genomic breeding values for two small populations of Korean pigs across five generations, in which all individuals were genotyped and phenotyped. One population consisted of Landrace pigs (3149 individuals) and the other population consisted of Yorkshire pigs (5053 individuals). We evaluated the decay in the prediction accuracy when the reference population was not updated and the prediction accuracy trends when individuals from only the immediate previous generation were used as the reference population. These were then compared to the prediction accuracy trends when individuals from all previous generations were used as the reference population. Additionally, prediction accuracy trends across generations were compared to the genomic relationships observed between the reference and test populations across generations.

2. Materials and Methods

A genetic model assuming additive SNP-effects was used for the genomic prediction of five traits (daily weight gain, weight at 90 days of age, average back fat, back fat depth, and meat percentage) in two populations of Korean pigs (Landrace (LL) and Yorkshire (YY), evaluated separately). Animals were genotyped on the Illumina PorcineSNP60 array (http://www.illumina.com). Quality control of the genotypes was performed with the program snpQC [13]. Individuals with call rates <0.9 were excluded from the data set, and only autosomal SNPs were considered for the analyses. SNPs with minor allele frequencies <0.01, with GC (GenCall) scores <0.6, or that failed in more than 5% of the samples were excluded from the genotype data. After quality control 47,225 SNPs remained for a total of 8202 individuals (3149 LL and 5053 YY). Missing SNP (<0.01%) were imputed with Eagle [14] using the software's default parameters. Seven discrete generations (generation zero to generation six) were available in the populations, but generations four to six were combined into a single generation (4–6), due to the small number of individuals in generations five and six. Table 1 summarizes the number of individuals by breed per generation.

The phenotypes were pre-corrected for fixed effects (overall mean, sex, company, and parity) using a linear regression model, $\boldsymbol{y}_{original} = \boldsymbol{Xb} + \boldsymbol{e}$, in which $\boldsymbol{b}$ were the coefficients for fixed effects and $\boldsymbol{X}$ their design matrix, and the pre-corrected phenotypes were defined as $\boldsymbol{y} = \boldsymbol{y}_{original} - \boldsymbol{X}\hat{\boldsymbol{b}}$. Then, as previously stated, we assumed an additive genomic model $\boldsymbol{y} = \boldsymbol{Ma} + \boldsymbol{\varepsilon}$ for the prediction of breeding values, in which $\boldsymbol{y}$ were the pre-corrected phenotypes, $\boldsymbol{M}$ the SNP genotypes, $\boldsymbol{a} \sim N\left(\mathbf{0}, \mathbf{I}\sigma_a^2\right)$ the additive random SNP effects, and $\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \mathbf{I}\sigma_\varepsilon^2\right)$ the random residuals. The breeding values were defined as $\boldsymbol{g} = \boldsymbol{Ma}$

and the genomic relationship matrix (GRM) as $\boldsymbol{G} = \boldsymbol{MM'} / \sum_{j=1}^{m} 2p_jq_j$ [15], with $\boldsymbol{g} \sim N\left(\boldsymbol{0}, \boldsymbol{G}\sigma_g^2\right)$, $p_j^{\prime}s$ were the minor allele frequencies of the SNP genotypes ($q_j = 1 - p_j$), and $\sigma_g^2 = \sum\limits_{j=1}^{m} 2p_jq_j\sigma_a^2$. Heritabilities ($h^2 = \sigma_g^2 / \left(\sigma_g^2 + \sigma_e^2\right)$) for the entire dataset and for each generation group (Table 2) were estimated using the restricted maximum likelihood (REML) [16] and breeding values were estimated using GBLUP [15]. All analyses were coded and conducted with the R statistical programming language.

Table 1. Number of individuals by breed per generation (LL—Landrace, YY—Yorkshire).

Generation	Breed		
	LL	YY	Total
0	431	1475	1906
1	520	1172	1692
2	418	827	1245
3	916	987	1903
4	832	508	1340
5	32	74	106
6	0	10	10
all	3149	5053	8202

Table 2. Heritabilities of the five traits for different generations in reference populations, by breed (LL and YY).

Generation	Traits									
	Daily Weight		Weight at Day 90		Back Fat Average		Back Fat Depth		Meat Percentage	
	LL	YY	LL	YY	LL	YY	LL	YY	LL	YY
0	0.456	0.502	0.354	0.493	0.826	0.365	0.146	0.246	0.502	0.389
1	0.344	0.487	0.277	0.414	0.570	0.334	0.296	0.250	0.613	0.309
2	0.239	0.441	0.210	0.389	0.311	0.364	0.305	0.246	0.515	0.374
3	0.361	0.435	0.344	0.412	0.316	0.486	0.254	0.164	0.310	0.456
4–6	0.346	0.431	0.298	0.403	0.305	0.424	0.295	0.237	0.289	0.397
0–1	0.373	0.485	0.291	0.456	0.634	0.334	0.193	0.196	0.476	0.336
0–2	0.323	0.490	0.251	0.450	0.579	0.347	0.199	0.206	0.426	0.331
0–3	0.326	0.494	0.271	0.458	0.481	0.392	0.192	0.200	0.311	0.345
0–6	0.332	0.492	0.282	0.453	0.466	0.385	0.193	0.201	0.308	0.337

Along with the genomic evaluations across generations, we evaluated the genomic relationships between reference and test populations across these generations. For this evaluation, we calculated the mean values of the GRM between the reference and the test individuals ($mean(\boldsymbol{G}_{ref,test})$), and we also calculated a genomic correlation between reference and test populations with a singular value decomposition (SVD) of the genotype matrix [17]. For this genomic correlation, we assumed $SVD(\boldsymbol{M}) = \boldsymbol{UDV'}$ and $\boldsymbol{R} = \sqrt{n_{test}/n_{ref}}\boldsymbol{V}'_{test}\boldsymbol{V}_{ref}\boldsymbol{D}_{ref}$, with $SVD(\boldsymbol{R}) = \boldsymbol{U}_R\boldsymbol{T}\boldsymbol{V}_R{}'$. $\boldsymbol{R}$ is a matrix that correlates the reference and the test individuals through their genotypes and accounts for their variance components and differences in population sizes. The prediction accuracy reaches its maximum ($\sqrt{h^2}$) when $\boldsymbol{R}$ and $\boldsymbol{D}_{test}$ are similar matrices, and this similarity can be summarized as $r = a + b$, such that a and b are the coefficients of the regression $\boldsymbol{T} \sim \boldsymbol{D}_{test}$. Because $\boldsymbol{R}$ is balanced for different population sizes, it is expected that $-1 < a < 0$ and $0 < b < 1$. Thus, $r = a + b$ is the correlation between reference and test populations using the SVDs of $\boldsymbol{M}_{ref}$ and $\boldsymbol{M}_{test}$ [17].

Prediction accuracy and the relationships between reference and test populations were obtained when the reference population was not updated (maintaining population zero as the reference for all subsequent generations), as well as when individuals from only the immediately previous generation were used as the reference population. A cumulative reference population with all individuals from previous generations was used to serve as the baseline (best case scenario). For each scenario, all animals available within these groupings where included in the analyses, there was no data sub-setting or resampling. Prediction accuracy trends across generations were then compared to the relationships between reference and test populations across generations.

3. Results

We performed a brief analysis of the two pig populations (breeds LL and YY) and their generations, regarding the principal components of the GRM, and the phenotypic values. Figure 1 presents the first two principal components, that represent 27% and 3.2% of the total variance, respectively, by breed and generation. With the principal components analysis (PCA), we can observe a clear separation between the two breeds. Within breeds, the distinction between generations is less discrete, although a slight trend is discernable. The across generation variance for LL was 5.6% and for YY it was 9.2%, for the sum of the first two principal components of the GRM, when the PCA was performed on each breed independently.

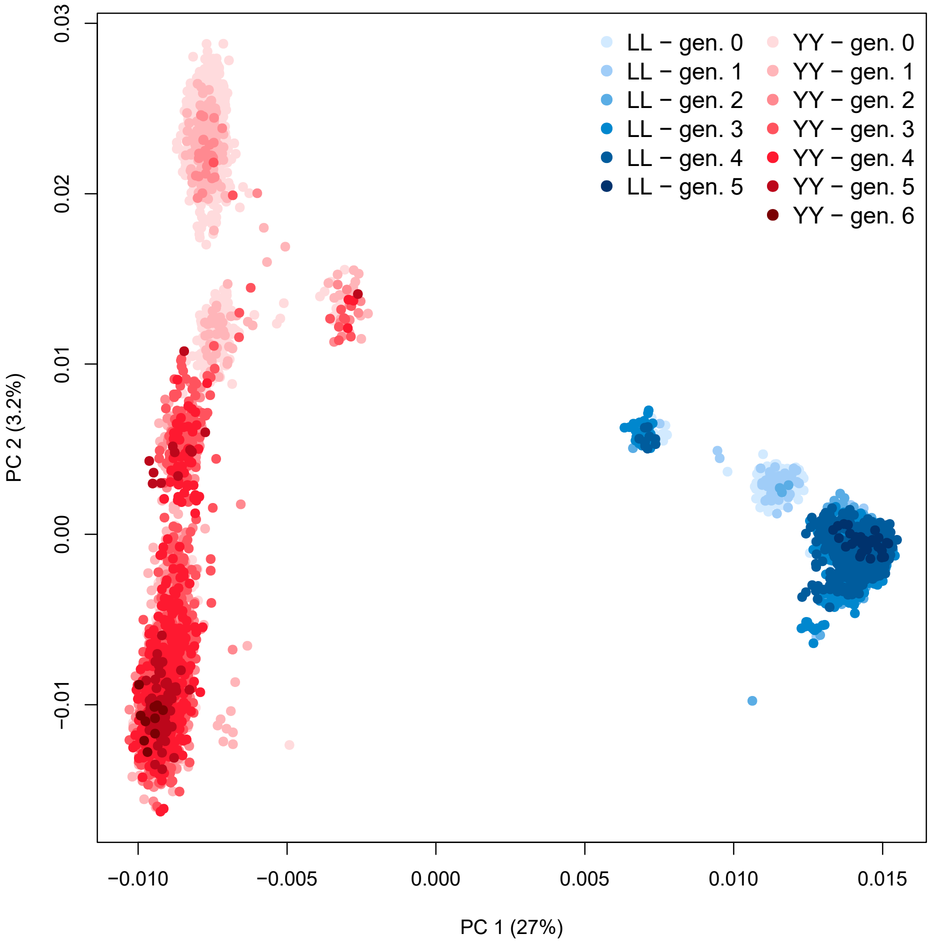

Figure 1. Principal components analysis of the genomic relationship matrix (GRM) by breed (Landrace (LL) and Yorkshire (YY)) and generations (zero to six).

Figure 2 presents the phenotypic values by breed and generations, for the five production traits (daily weight, weight at 90 days, average back fat, back fat depth and meat percentage). With the exception of meat percentage, the trends in the phenotypic values across the generations are very similar for both breeds.

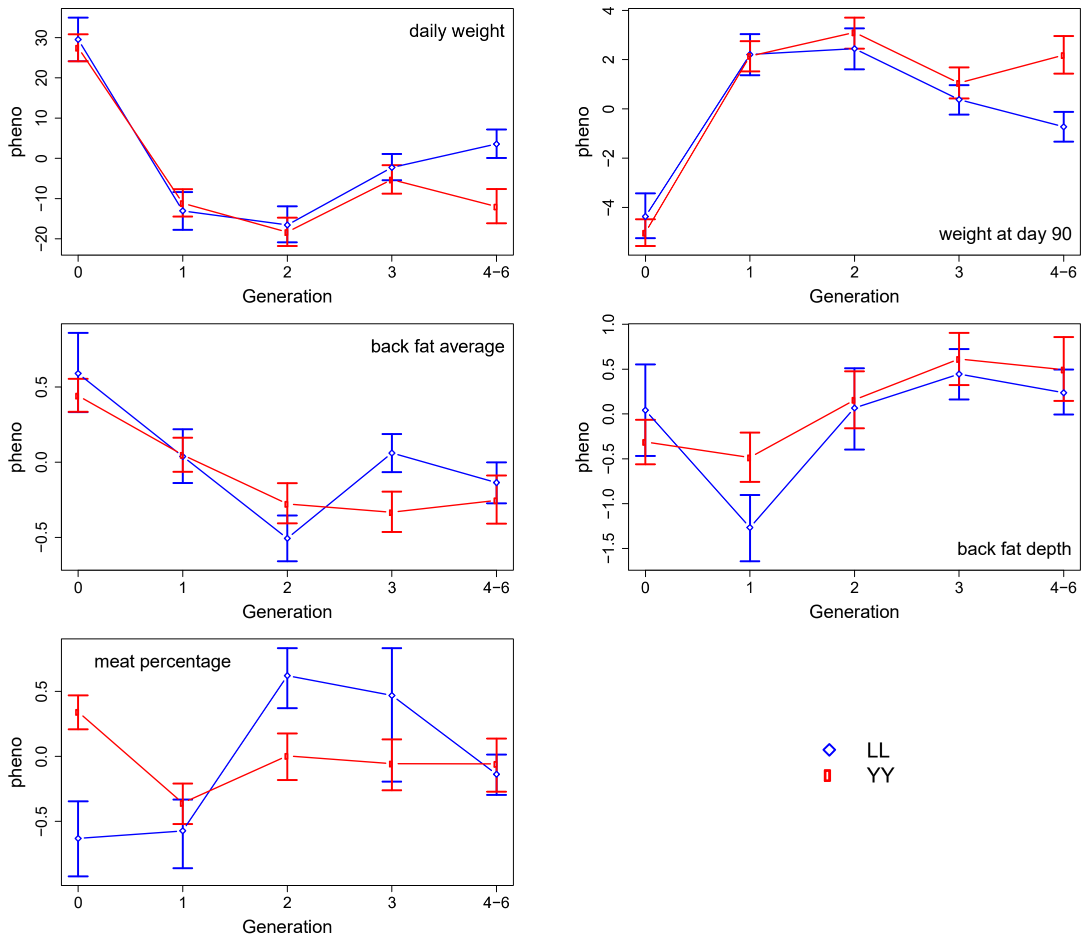

Figure 2. 95% bootstrap confidence intervals from 1,000,000 resamples of the phenotypic values (corrected for fixed effects) by breed (Landrace (LL) and Yorkshire (YY)) and generations (zero to six).

Table 2 presents the heritabilities obtained with REML by breed and generations, for the five traits (daily weight, weight at 90 days, average back fat, back fat depth, and meat percentage). The standard deviation of heritability estimates ranged from 0.025 to 0.048.

Figure 3 presents the prediction accuracies obtained by breed and generations, for the five production traits (daily weight, weight at 90 days, average back fat, back fat depth, and meat percentage). We observed distinct levels of prediction accuracy when different generations were considered as the reference population. When generation zero was used as the reference to all subsequent generations, we observed a loss in prediction accuracy of the next generations for all traits, on both breeds. When the immediate previous generation or all the previous generations were used as reference populations, we observed an increase in prediction accuracy for all traits, in both breeds.

Figure 4 presents the relationships between individuals in reference and test populations by breed and generations. We observed a distinction of the different generations considered as the reference population. When generation zero was fixed as the reference to all subsequent generations, we observed a decay in the relationships to further generations for all the traits, on both breeds. When the immediate previous generation was used as reference population, we observed that $mean(\boldsymbol{G}_{ref,test})$ remained stable across generation for both breeds, while $r = a + b$ increased for YY, and remained stable for LL. When all previous generation were used as reference population we observed that $mean(\boldsymbol{G}_{ref,test})$ remained stable across generations for LL and increased for YY, while $r = a + b$ increased for both LL and YY (with the exception of generation 4-6, on which there was a decrease of $r = a + b$ for YY).

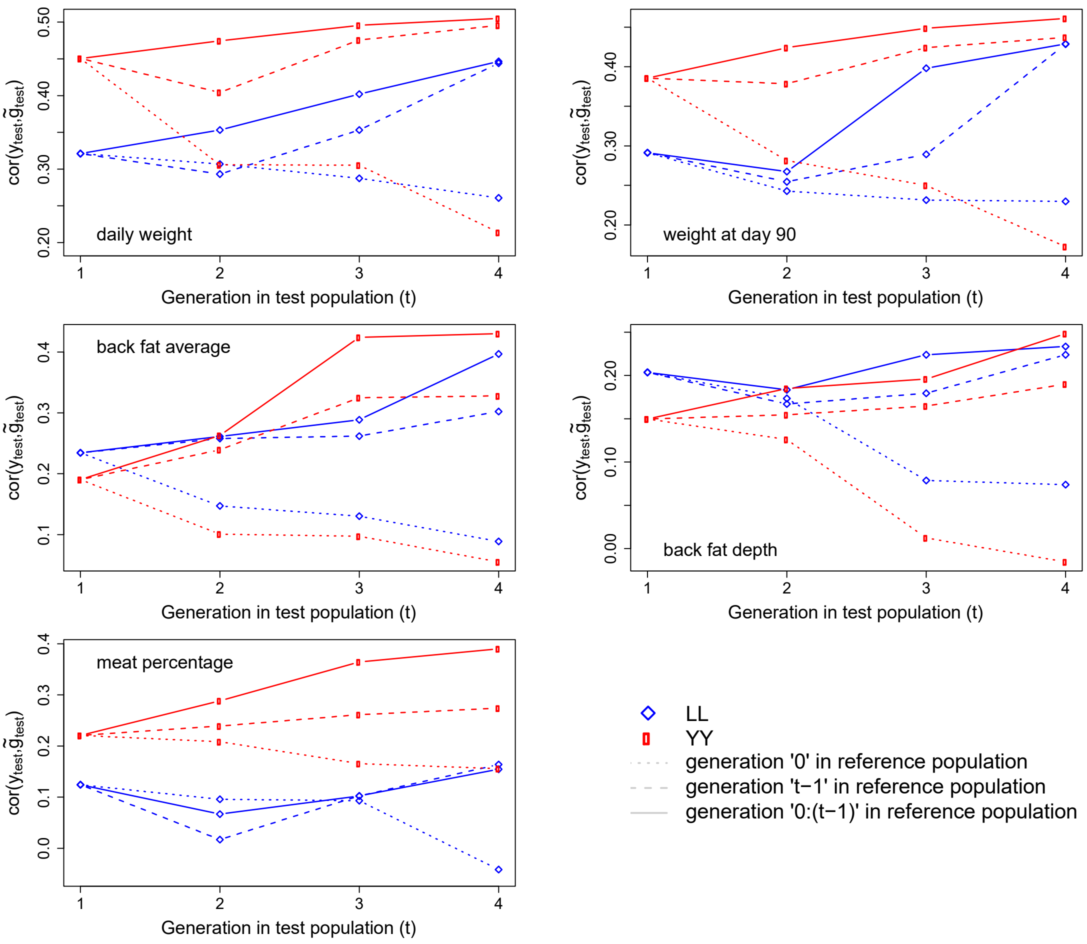

Figure 3. Prediction accuracies $\left(\hat{cor}\left(y_{test}, \tilde{g}_{test}\right)\right)$, the correlation between real phenotypes in the test population, y_{test}, and predicted breeding values, $\tilde{g}_{test}$) by breed (Landrace (LL) and Yorkshire (YY)) and generations (one to four, the last one combining generations four to six). Results are presented for the different generations considered as the reference population to perform genomic prediction, to predict a test generation (t).

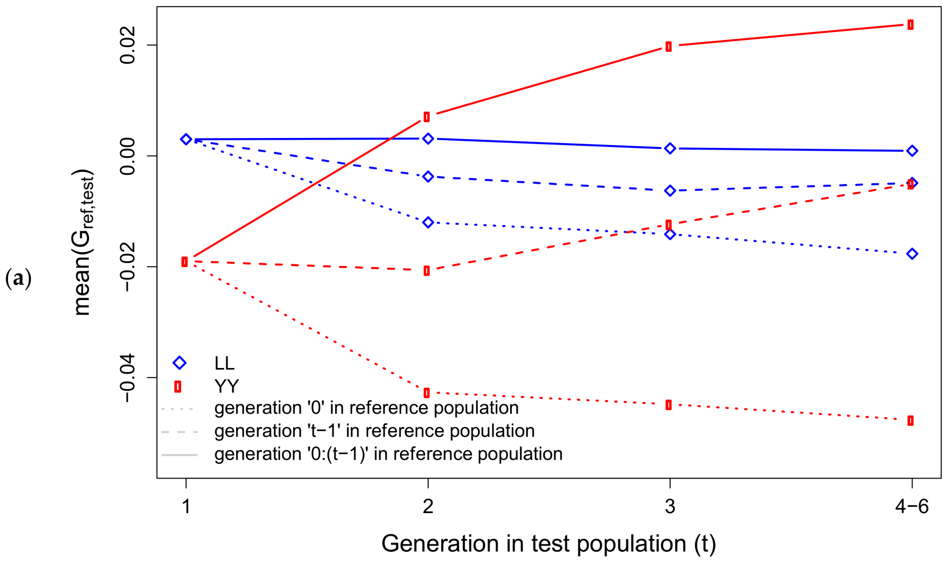

Figure 4. *Cont.*

(**b**)

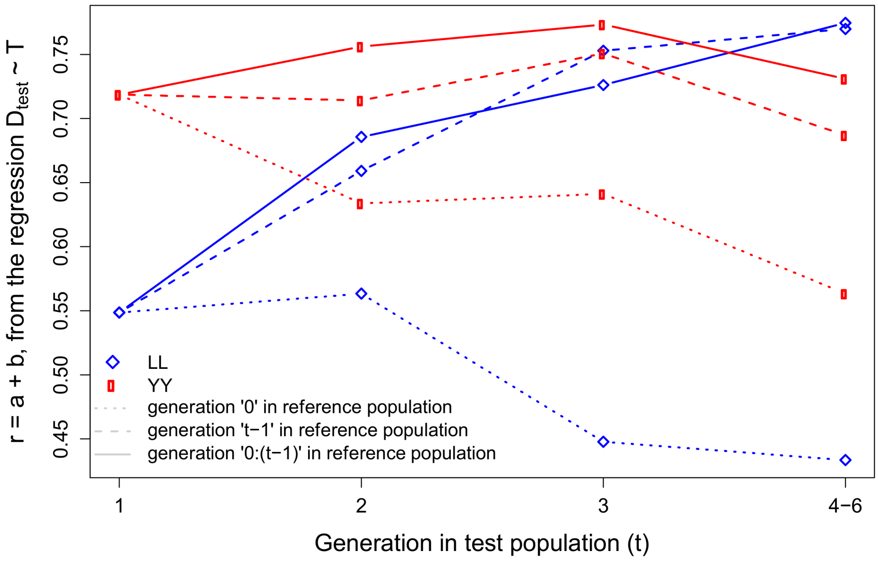

Figure 4. Genomic relationships between individuals in reference and test populations by breed (Landrace (LL) and Yorkshire (YY)) and generations (1 to 4–6). Panel (**a**) presents the average values in the off-diagonals of the GRM, and panel (**b**) presents the genomic correlation $r = a + b$, based on the singularvalue decomposition (SVD) of the genotype matrices.

4. Discussion

Our study aimed to evaluate the accuracy of prediction across generations, and possibly relate trends in prediction accuracy to patterns in the relationships between individuals from reference and test populations. We did observe that the closer the individuals from reference and test populations are in terms of generations, both relationships and prediction accuracy increase.

While the close relationship between test and reference populations is highly relevant to prediction accuracy, it is also well known that sample size, and genetic and phenotypic variation are important contributors to genomic prediction. As expected, we did observe that when aggregating all past generations to the reference population, prediction accuracy indeed is higher than when only the immediately previous generation features in the reference population. This was intentional in this study as we primarily wanted to contrast the decay in accuracy across generations, and by using all previous aggregated generation data we could set the baseline of "best possible prediction" which could then be used to compare close versus further apart generations.

This same behavior in prediction accuracy was also observed in the relationships between individuals, and even clearer in our measure based on the SVD of the genotype matrices. The gain in prediction accuracy, from aggregating all past generations is, however, greater than the gain in the relationship between individuals. This is most likely a result of an additional gain in the prediction accuracy due to the greater size of the reference.

We did observe some fluctuations in the prediction accuracies, and this is most probably related to sample sizes. The number of individuals available for each generation in our data set was not homogeneous, and this may be a cause for the observed fluctuations. In addition, the scenario that incorporated all previous generations in the reference population represents the best-case scenario, serving as a comparative baseline for the prediction accuracies. An interesting result observed was that for LL, with the exception of back fat average, the prediction accuracy of generation 4–6 was very similar when either all previous generations were incorporated in the reference population, or only the immediately previous generation was used as reference. The same was observed for YY, for daily weight and weight at day 90. This indicates that potentially, if the population evolves mostly within itself, ancestral generations can be discarded from the reference over time, without compromising prediction accuracy. The sample sizes of the LL breed were quite small when compared to the YY breed, and the phenotypic values of LL did fluctuate more because of the small sample size, potentially resulting in less reliable prediction accuracies.

Cross-generational persistency of prediction accuracy should be further explored in animal breeding. One topic that should be addressed in future studies is how many generations back should a reference population incorporate, to predict the subsequent generation. We did observe that when all previous generations are aggregated to train the model, both the relationships between individuals and prediction accuracies start to reach a maximum limit. We did not however, have enough generations in our data set to explore this in detail. This is particularly interesting, in the sense that it can give a perspective of how much information is necessary to aggregate to reference populations, without adding too many individuals from ancient generations that do not incorporate any relevant information and increase the computer burden of performing the prediction models. Another topic that should be explored in these studies is the behavior of prediction accuracy in comparison to the intensity of selection, and the incorporation of new individuals into a herd.

While trends were observed in the prediction accuracy of the traits evaluated over generations (increasing accuracy of prediction when the reference population was updated, and decreasing accuracy of prediction when the reference populations was not updated), no clear or conclusive trends were observed on the estimated heritabilities. This is potentially a topic that would generate more discussion when combined with evaluations that explore intensity of selection, and the incorporation of new individuals into a herd.

Finally, comprehending how the accuracy of genomic prediction evolves over generations within a population may add extra and relevant information to improve the performance of genomic selection. Overall these results confirm the need to continuously update the reference with new genotypes and phenotypes, but they also show that it may not be necessary to keep all ancestral genotypes indefinitely in the reference population.

5. Conclusions

This study presented an empirical assessment of the prediction accuracy of genomic breeding values of production traits, across five generations in two Korean pig breeds. Overall, the results suggested that, although there is a clear need to continuously update the reference population, it may not be necessary to keep all ancestral genotypes. Finally, comprehending how the accuracy of genomic prediction evolves over generations within a population may add extra and relevant information to improve the performance of genomic selection.

Author Contributions: B.C.D.C. and C.G. conceptualized the study; D.S. generated the data; C.G., D.S., and H.W. prepared the data for the analyses; B.C.D.C. performed the analyses and wrote the manuscript.

Funding: This work was supported by a grant from the Next-Generation BioGreen 21 Program (Project No. PJ01322204 and PJ01337702), Rural Development Administration, Korea.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Meuwissen, T.H.; Hayes, B.J.; Goddard, M.E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **2001**, *157*, 1819–1829. [PubMed]
2. Dekkers, J.C.M. Prediction of response to marker-assisted and genomic selection using selection index theory. *J. Anim. Breed. Genet.* **2007**, *124*, 331–341. [CrossRef] [PubMed]
3. De Roos, A.; Schrooten, C.; Veerkamp, R.; Van Arendonk, J. Effects of genomic selection on genetic improvement, inbreeding, and merit of young versus proven bulls. *J. Dairy Sci.* **2011**, *94*, 1559–1567. [CrossRef] [PubMed]
4. Heffner, E.L.; Sorrells, M.E.; Jannink, J.L. Genomic Selection for Crop Improvement. *Crop Sci.* **2009**, *49*, 1–2. [CrossRef]
5. Jannink, J.-L.; Lorenz, A.J.; Iwata, H. Genomic selection in plant breeding: From theory to practice. *Brief. Funct. Genom.* **2010**, *9*, 166–177. [CrossRef] [PubMed]

6. Campos, G.D.L.; Gianola, D.; Allison, D.B. Predicting genetic predisposition in humans: The promise of whole-genome markers. *Nat. Rev. Genet.* **2010**, *11*, 880–886. [CrossRef] [PubMed]
7. Wray, N.R.; Goddard, M.E.; Visscher, P.M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **2007**, *17*, 1520–1528. [CrossRef] [PubMed]
8. Hayes, B.J.; Visscher, P.M.; Goddard, M.E.; Goddard, M. Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.* **2009**, *91*, 47–60. [CrossRef] [PubMed]
9. Hayes, B.J.; Bowman, P.J.; Chamberlain, A.C.; Verbyla, K.; Goddard, M.E. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet. Sel. Evol.* **2009**, *41*, 51. [CrossRef] [PubMed]
10. Habier, D.; Fernando, R.L.; Garrick, D.J. Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction. *Genetics* **2013**, *194*, 597–607. [CrossRef] [PubMed]
11. Pszczola, M.; Calus, M.P.L. Updating the reference population to achieve constant genomic prediction reliability across generations. *Animal* **2016**, *10*, 1018–1024. [CrossRef] [PubMed]
12. Weng, Z.; Wolc, A.; Shen, X.; Fernando, R.L.; Dekkers, J.C.M.; Arango, J.; Settar, P.; Fulton, J.E.; O'Sullivan, N.P.; Garrick, D.J. Effects of number of training generations on genomic prediction for various traits in a layer chicken population. *Genet. Sel. Evol.* **2016**, *48*, 22. [CrossRef] [PubMed]
13. Gondro, C.; Porto-Neto, L.R.; Lee, S.H. Snpqc—An R pipeline for quality control of Illumina SNP genotyping array data. *Anim. Genet.* **2014**, *45*, 758–761. [CrossRef] [PubMed]
14. Loh, P.R.; Danecek, P.; Palamara, P.F.; Fuchsberger, C.; Reshef, Y.A.; Finucane, H.K.; Schoenherr, S.; Forer, L.; McCarthy, S.; Abecasis, G.R.; et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **2016**, *48*, 1443–1448. [CrossRef] [PubMed]
15. VanRaden, P. Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.* **2008**, *91*, 4414–4423. [CrossRef] [PubMed]
16. Patterson, H.D.; Thompson, R. Recovery of Inter-Block Information when Block Sizes are Unequal. *Biometeika* **1971**, *58*, 545–554. [CrossRef]
17. Cuyabano, B.C.; Gondro, C. Limits for the Accuracy of Genomic Prediction Using a Genotype Based Correlation between Populations. Unpublished Work. 2019.