Impact of genotyping strategy on the accuracy of genomic prediction for survival in pigs

T. Liu[1,2], B. Nielsen[2,3], O.F. Christensen[2], M. S. Lund[2], G. Su[2]

[1] Institute of Animal Science, Guangdong Academy of Agricultural Sciences, Dafeng 1st Street 1, Wushan, Tianhe District, 510640 Guangzhou, China; [2] Center for Quantitative Genetics and Genomics, Aarhus University, Blichers Alle 20, 8830 Tjele, Denmark; [3] SEGES Pig Research Centre, Axelborg, Axeltorv 3, 1609 Copenhagen, Denmark; liutfei@gmail.com

Abstract
This study compared four genotyping scenarios of genomic prediction for survival in a simulated pig population with survival rate of 80%. Breeding values were predicted using a linear mixed model and a logit model, based on individual records. The scenarios in order of prediction accuracy were: genotyping all individuals, randomly genotyping 80% individuals, genotyping alive individuals, and no genotyping. Using variance components estimated from pedigree-based model, genomic predictions with genotypes only from alive animals were unbiased. Given the same number of genotyped individuals, random genotyping 80% individuals led to higher prediction accuracy than only genotyping alive individuals, indicating the importance of genotype information of dead pigs. Moreover, the linear and the logit models achieved similar accuracy. We conclude that genotyping dead individuals should be considered, and a linear model can be used for predicting breeding values of survival in pigs.

Introduction
Survival rate can affect not only economic efficiency, but also welfare in pigs. Genetic improvement will be a good solution to improve pig survival. Compared with traditional pedigree-based selection, genomic selection can increase of accuracy of estimated breeding value (EBV) of mortality rate up to 50% (Guo et al., 2015;Knol et al., 2016). Usually, survival rate is recorded as phenotypes of sow. However, survival rate is a complex trait that is also affected by the pig's own genotype. It may therefore be more appropriate to perform genetic evaluation of survival at individual level. Genomic prediction for survival at individual level may encounter problems in the case where dead individuals do not have genotypes. Using only the genotyping data of living individuals may cause bias of genomic prediction. Furthermore, at individual level, survival is a binary trait, and therefore a logit model could be more appropriate, and thus may lead to higher accuracy than a linear mixed model. The objective of this study was to: 1) investigate genetic parameters estimated using a linear model and a logit model; 2) explore the impact of genomic information of dead individuals on genomic prediction; 3) compare the accuracy of genomic prediction of a linear model and the logit model for survival in pigs.

Materials & Methods
Data simulation. The data was simulated using QMSim software (Sargolzaei and Schenkel, 2009) mimicking a pig population. We simulated 18 chromosomes, each 100cM with 3,100 markers and 50 QTLs. The simulation started with a founder population of 200 males and 200 females, and went through 300 historical generations to generate linkage disequilibrium between markers and QTLs. After historical generations, a base population was created and went through eight non-overlapping generations. In each generation, 30 sire and 300 dams were randomly selected from alive animals,

a sire mated 10 dams randomly, and each dam produced one litter. The litter size was not generated as a genetic trait but randomly assigned to 10, 12, 14, 16, or 18 with the probabilities 0.02, 0.14, 0.68, 0.14, 0.02, and sex ratio was 1:1.

The liability of an individual to be alive was generated as the sum of direct additive genetic effect (a) of the individual, maternal additive genetic effect (m) of the dam, litter effect and random residual, and independent of litter size. The heritabilities in terms of *a* and *m* as well as proportion of litter variance were all set as 0.04, and genetic correlation between *a* and *m* was 0.3. The phenotype on observed scale was scored as 1 if the liability was the top 80%, and otherwise 0. Four genotyping scenarios were studied: (1) all pigs were genotyped (G_all); (2) only alive pigs (80%) (G_alive); (3) 80% of pigs randomly selected from the whole population (G80_ran); (4) no pig was genotyped (G_none). In total, 30 replicates were simulated for the four scenarios. The reported results were the mean and standard deviation of the 30 replicates.

Statistical analysis. The data were analyzed using a linear mixed model and a logit model.

The linear mixed model is,

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{W_l l} + \mathbf{Z_a a} + \mathbf{Z_m m} + \mathbf{e} \quad (1)$$

where, $\boldsymbol{y}$ is the vector of the observed phenotypes (0 or 1), **1** is a vector of 1s, μ is overall mean, **l** is the vector of litter effects, **a** is the vector of direct additive genetic effects, **m** is the vector of maternal additive genetic effects, and **e** is the vector of random residuals. $\mathbf{W_l}$, $\mathbf{Z_a}$, $\mathbf{Z_m}$ are incidence matrixes associating **l, a, m** with $\boldsymbol{y}$.

The logit model is,

$$\boldsymbol{\eta} = \mathbf{1}\mu + \mathbf{W_l l} + \mathbf{Z_a a} + \mathbf{Z_m m} \quad (2)$$

where, $\boldsymbol{\eta}$ is the vector of logit function of probabilities of surviving, $\eta_i = \log_e \frac{\mu_i}{1-\mu_i}$ and μ_i is the expected value of y_i. The other notations are the same as in the linear model, but in logit scale.

For both modes, it was assumed that **l, a** and **m** have the following distributions: $\boldsymbol{l} \sim N(\mathbf{0}, \boldsymbol{I}\sigma_l^2)$ and $\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{m}\end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix}\sigma_a^2 & \sigma_{am}\\ \sigma_{am} & \sigma_m^2\end{bmatrix} \otimes \mathbf{K}\right)$, where $\sigma_{\mathbf{l}}^2$, $\sigma_{\mathbf{a}}^2$, $\sigma_{\mathbf{m}}^2$ and $\sigma_{\mathbf{am}}$ are the variance of *l, a, m* and covariance between *a* and *m*, respectively, and **K** is an additive genetic relationship matrix. When using traditional BLUP method, **K** was constructed from pedigree information. When using single-step GBLUP model (ssGBLUP), **K** was constructed from pedigree and genotypes (Christensen and Lund, 2010). It was assumed that in the linear model $\boldsymbol{e} \sim N(\mathbf{0}, \boldsymbol{I}\sigma_e^2)$, and in the logit model residuals followed a standard logistic distribution with variance of $\pi^2/3$. In this study, total phenotypic variance was defined as $\sigma_p^2 = \sigma_l^2 + \sigma_a^2 + \sigma_m^2 + \sigma_{am} + \sigma_e^2$.

The variance components estimated from the model with pedigree-based relationship matrix based on data of generations 5-8 were used for prediction of breeding values in all models. The estimation of variances and breeding values were performed using the DMU software (Madsen et al., 2010).

Evaluation of genomic prediction. A validation procedure was performed to assess genomic prediction using different genotyping strategies and statistical models, in which the 5~7th generations were used as reference population, and the 8th generation as validation population. Genomic predictions were evaluated using the following criteria: 1) The correlation between EBV and the true breeding value to assess prediction accuracy; 2) Average true breeding value of the top 5% of individuals in EBVs to assess the realized selection differential; 3) Regression of EBV from whole data on the EBV from reference data (Legarra and Reverter, 2018) to evaluate dispersion bias. Here, EBV from whole data was defined as the EBV obtained from the whole data with genotypes of all individuals.

Results

As shown in Table 1, estimates of parameters were different between the two models due to different scales. By a transformation, the estimates in observed scale from the linear model were consistent with those used in the simulation. For the logit model, the estimated h_a^2 was much lower than the estimated h_m^2, which was unexpected since direct and maternal heritabilities were the same in the simulation. For both models, the estimates of genetic correlations between *a* and *m* (around 0.4) were higher than correlation used in simulation (0.3), and also had large standard error.

Table 1. Estimates of variance components in proportion to phenotypic variance using models with pedigree-based relationship matrix.

Parameters	True parameters of liability	Linear Model	Logit Model
lit^2	0.04	0.019(0.008)	0.026(0.012)
h_a^2	0.04	0.019(0.007)	0.023(0.007)
h_m^2	0.04	0.020(0.010)	0.037(0.016)
r_{am}	0.30	0.401(0.465)	0.400(0.407)

Accuracies measured as correlation between total EBV (EBV_a+EBV_m) and total true BV are shown in Table 2. The scenario of G_all had the highest accuracy, and the scenario of no genotyping had the lowest. With the same proportion of genotyped pigs (80%), the accuracy of G80_ran was 4.73 to 11.2 percentage points higher than G_alive, indicating that the genotypes of dead pigs greatly increased the accuracy of EBV. The regression coefficients were around 1 for all genotyping scenarios including G_alive, indicating that genotyping only for alive animal did not lead to biased prediction when using variances estimated from the model with pedigree-based relationship matrix. The results from the logit model were almost the same as those from the linear model.

Table 2. Accuracy and bias of EBV.

Model	Genotyping scenario	Accuracy			Unbiasedness		
		All	Genotyped	Nongenotyped	All	Genotyped	Nongenotyped
Linear	G_all	0.511	0.511	/	0.986	0.986	/
	G_alive	0.464	0.477	0.374	1.029	1.006	0.991
	G80_ran	0.487	0.504	0.415	0.983	0.983	0.986
	G_none	0.334	/	0.334	0.985	/	0.985
Logit	G_all	0.511	0.511	/	0.996	0.996	/
	G_alive	0.465	0.479	0.375	1.016	1.015	1.018
	G80_ran	0.487	0.504	0.417	0.997	0.996	0.998
	G_none	0.334	/	0.334	0.993	/	0.993

The 5% top animals selected from the scenario of G_all has the highest true breeding value, followed by G80_ran, then G_alive, and then G_none (Table 3). The results of the linear model and the logit model were very similar, and the difference in true total breeding value of top 5% EBV individuals between the two models was less than 1%.

Table 3. The true breeding value of the top 5% individuals in EBV of survival.

Genotyping scenario	Linear model			Logit model		
	a	*m*	*a*+*m*	*a*	*m*	*a*+*m*
G_all	1.847	1.596	3.443	1.781	1.654	3.435
G_alive	1.602	1.604	3.206	1.555	1.658	3.213
G80_ran	1.745	1.529	3.274	1.699	1.609	3.308
G_none	1.104	1.022	2.126	1.122	1.047	2.169

Discussion

In this study, we compared four genotyping scenarios for the accuracy of predictions of survival in pigs. In a pig breeding program, dead animal is usually not genotyped, which may lead to biased estimation of variance components and breeding values when using a genomic model. We carried out a pilot study, and found that a ssGBLUP model with genotypes only from alive animals severely overestimated additive genetic variance and led to a residual variance near to zero (results not shown). Similarly, Wang et al. (2020) reported that selective genotyping severely overestimated additive genetic variance. Therefore, in the case of no dead animals being genotyped, genomic prediction should use the variances estimated from the model with pedigree-based relationship matrix to avoid or reduce bias of genomic prediction due to selective genotyping.

With the same size of genotyped individuals, genotyping both alive and dead pigs (G80_ran) has a higher accuracy than genotyping only for alive pigs (G_alive), indicating that the genotypes of dead pigs has an important influence on prediction accuracy. Therefore, it could be a good strategy to genotype dead animals. The predictive powers of the linear model and the logit model are similar, suggesting that linear mixed model is robust to predict breeding value of survival traits. We conclude that the genomic information of dead individuals is very useful, and linear model is a good choice for genomic prediction of survival in pigs.

References

Christensen O.F., and Lund M.S. (2010). Genetics Selection Evolution 42(1): 2. https://doi.org/10.1186/1297-9686-42-2

Guo X., Christensen O.F., Ostersen T., Wang Y., Lund M.S., et al. (2015). Journal of Animal Science 93(2): 503-512. https://doi.org/10.2527/jas.2014-8331

Knol E.F., Nielsen B., and Knap P.W. (2016). Animal Frontiers 6(1): 15-22. https://doi.org/10.2527/af.2016-0003

Legarra A., and Reverter A. (2018). Genetics Selection Evolution 50(1): 53. https://doi.org/10.1186/s12711-018-0426-6

Madsen P., Su G., Labouriau R., and Christensen O.F. (Year). Proc. of 9th World Congr. Genet. Appl. Livest. Prod, Leipzig, Germany.

Sargolzaei M., and Schenkel F.S. (2009). Bioinformatics 25(5): 680-681. https://doi.org/10.1093/bioinformatics/btp045

Wang L., Janss L.L., Madsen P., Henshall J., Huang C.-H., et al. (2020). Genetics Selection Evolution 52(1): 31. https://doi.org/10.1186/s12711-020-00550-w