VINICIUS SILVA DOS SANTOS

GENOMIC PREDICTION MODELS WITH ADDITIVE AND DOMINANCE EFFECTS FOR CENSORED TRAITS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae.*

VIÇOSA
MINAS GERAIS - BRASIL
2017

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
S237g
2017

Santos, Vinicius Silva dos, 1987-
Genomic prediction models with additive and dominance effects for censored traits / Vinicius Silva dos Santos. – Viçosa, MG, 2017.
xi, 85 f. : il. ; 29 cm.

Texto em inglês.
Orientador: Sebastião Martins Filho.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Genômica. 2. Análise de sobrevivência (Biometria). 3. Suínos - Genética - Métodos estatísticos. 4. Análise de regressão. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 572.86

VINICIUS SILVA DOS SANTOS

GENOMIC PREDICTION MODELS WITH ADDITIVE AND DOMINANCE EFFECTS FOR CENSORED TRAITS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae.*

APROVADA: 31 de março de 2017.

Adésio Ferreira

Renata Veroneze

Camila Ferreira Azevedo

Fabyano Fonseca e Silva
(Coorientador)

Sebastião Martins Filho
(Orientador)

Aos meus pais, Dilma (in memoriam) e Nazareno,

À minha esposa Karolayne e à minha irmã Camila,

Aos meus amigos

AGRADECIMENTOS

Primeiramente agradeço a Deus, pelo dom da vida e seu amor incondicional, sempre conduzindo meus passos e ajudando-me todos os dias a superar as dificuldades e a realizar mais este sonho. Sem Ele, nada posso fazer.

À minha mãe Dilma (*in memoriam*), por todo amor sem igual, dedicação, cuidado e carinho dados a mim e a minha irmã. Pelas orações, ensinamentos e horas de estudos, na infância, com a matemática e a redação, que tanto me ajudaram para que eu chegasse até aqui. Como queria ter dado essa notícia à senhora!

Ao meu pai Nazareno, por todo amor, dedicação e cuidado dados a mim e minha irmã. Por todo o esforço para que tivéssemos sempre o que precisássemos. Muito obrigado pai e mãe por dedicar suas vidas a nós!

À minha irmã Camila, pelo amor, amizade e ajuda em todos os momentos que preciso. Vocês, pai, mãe e irmã foram meu incentivo a continuar em meio às dificuldades e à saudade.

À minha esposa Karolayne, por todo amor, companheirismo, paciência, apoio e incentivo dado em todos os momentos.

Aos meus amigos, em especial Caio e Leonardo, parentes e irmãos na fé, pelo apoio e incentivo.

Aos amigos de república, Leonardo, Wesly e Jean, e aos amigos do grupo de convivência, pelos momentos de conversa e descontração.

À Universidade Federal de Viçosa (UFV) e à Coordenação do Programa de Pós-Graduação em Estatística Aplicada e Biometria (PPESTBIO), pela criação do curso de doutorado.

À CAPES pela bolsa de estudos concedida no Brasil e no exterior.

Ao orientador, professor Dr. Sebastião Martins Filho, pela instrução, amizade, confiança, incentivo e paciência durante a realização deste trabalho.

Ao professor e coorientador Dr. Fabyano Fonseca e Silva, pela troca de saberes, pela disponibilidade em me atender com as dúvidas e incentivo na conclusão deste trabalho.

Ao professor e coorientador Dr. Marcos Deon Vilela de Resende, pelos ensinamentos, apoio, confiança, disposição em ajudar-me e dedicação ao ensino e à pesquisa. Aproveito para relatar a grande contribuição desses dois pesquisadores (Fabyano e Marcos Deon) à pesquisa brasileira na área de seleção genômica, dentre outras.

Aos membros da banca de qualificação, professor Dr. Fabyano Fonseca e Silva, professor Dr. Paulo César Emiliano, Dr. Rodrigo Reis Mota e professor Dr. Sebastião Martins Filho, pela disponibilidade e pelas sugestões para o aprimoramento deste trabalho.

Aos membros da banca de defesa, professores Dr. Adésio Ferreira, Dra. Camila Ferreira Azevedo, Dr. Fabyano Fonseca e Silva, Dra. Renata Veroneze e Dr. Sebastião Martins Filho, pela disponibilidade e pelas valiosas sugestões para a conclusão deste trabalho.

Ao Departamento de Zootecnia da Universidade Federal de Viçosa, pelos dados concedidos para realização deste trabalho.

A todos os professores do Programa de Pós-graduação em Estatística Aplicada e Biometria (PPESTBIO) pelos ensinamentos repassados;

Aos secretários do PPESTBIO, Carla e Júnior por todo apoio para sanar questões pendentes da vida acadêmica.

Ao Wagner pela ajuda na entrega da tese e emissão de documentos.

À Camila, Elisabeth, Leandro e Lidiane pela troca de saberes e pela companhia nos momentos de estudos.

A todos que de alguma forma contribuíram para a conclusão deste trabalho e realização deste sonho. Meu muito obrigado.

SUMÁRIO

ABSTRACT

SANTOS, Vinicius Silva dos, D.Sc., Universidade Federal de Viçosa, March, 2017. **Genomic prediction models with additive and dominance effects for censored traits**. Adviser: Sebastião Martins Filho. Co-advisers: Fabyano Fonseca e Silva and Marcos Deon Vilela de Resende.

Recently, dominance effects have been included in the genomic selection of several species, with the GBLUP-D method being the most used. This method consists in replacing, in the REML / BLUP procedure, the pedigree-based relationship matrices by marker-based relationship matrices. This method can be performed using the GVCBLUP software or through BGLR R-package, which is based on Bayesian regression via the Reproduction Kernel Hilbert Space. The objective of this work was to evaluate the possibility and effectiveness of GBLUP-D implementation via the *lmekin* function implemented in the coxme package of R through the inclusion of additive and dominance genomic matrices. Thus, through simulated data analyzes, the results obtained by the *lmekin* function were compared with those obtained by the GVCBLUP software and the BGLR package. Subsequently, the analysis was extended considering phenotypes with censored observations in a F_2 population of pigs, where the time (in days) of the birth to the slaughter of the animal was evaluated through the Cox model and the truncated normal model, in that the censoring was considered or not in the analysis. Finally, the inclusion of the polygenic effect in the additive-dominant models was evaluated in three traits with complete and normally distributed observations of a mice population, and in censored data from a F_2 population of pigs. The results showed that the *lmekin* function is an efficient alternative for the fit of genomic linear models with additive and dominance effects, since it results were identical to those obtained through GVBLUP software. For the censored data, it was observed a high agreement between the Cox model and the truncated normal model in selecting the best individuals and the highest marker effects. Thus, it was possible to show the possibility of predicting genomic genetic values for censored data, considering the Cox survival model with additive and dominance effects. The inclusion of the polygenic effect in

the evaluated models allowed a significative increase in the additive heritabilities of the evaluated traits.

RESUMO

SANTOS, Vinicius Silva dos, D.Sc., Universidade Federal de Viçosa, março de 2017. **Modelos de predição genômica com efeitos aditivos e de dominância para características censuradas**. Orientador: Sebastião Martins Filho. Coorientadores: Fabyano Fonseca e Silva e Marcos Deon Vilela de Resende.

Recentemente, efeitos de dominância, têm sido incluídos na seleção genômica de várias espécies, sendo o método GBLUP-D o mais utilizado. Esse método consiste em substituir, no procedimento REML/BLUP, as matrizes de parentesco baseadas no *pedigree* pelas matrizes com base nos marcadores moleculares. Este método pode ser realizado por meio do software GVCBLUP ou por meio do pacote BGLR do software R, o qual se baseia em regressão bayesiana via Kernel de Reprodução do Espaço de Hilbert. Este trabalho teve como objetivo inicial avaliar a possibilidade e efetividade de implementação do GBLUP-D via a função *lmekin* implementada no pacote *coxme* do software R por meio da inclusão das matrizes de parentesco genômico aditivo e de dominância. Assim, comparou-se, via análises de dados simulados, os resultados obtidos pela função *lmekin* com os obtidos pelo software GVCBLUP e pacote BGLR. Posteriormente, estendeu-se a análise considerando fenótipos com observações censuradas numa população F_2 de suínos, onde avaliou-se o tempo (em dias) do nascimento ao abate do animal por meio do modelo de Cox e do modelo normal truncado, nas situações em que a censura foi considerada ou não na análise. E por fim, avaliou-se a inclusão do efeito poligênico nos modelos aditivos-dominante em três características com observações completas e normalmente distribuídas de uma população de camundongos e em dados censurados de uma população F_2 de suínos. Os resultados comprovaram inicialmente que a função *lmekin* é uma alternativa eficiente para o ajuste, no software R, de modelos lineares genômicos com efeitos aditivos e de dominância, uma vez que apresentou resultados idênticos aos obtidos por meio do software GVCBLUP. Para os dados censurados, observou-se alta concordância entre o modelo de Cox e o modelo normal truncado em

selecionar os melhores indivíduos e os maiores efeitos de marcas. Com isso, mostrou-se a possibilidade de predizer valores genéticos genômicos para dados censurados, considerando o modelo de sobrevivência de Cox com efeitos aditivos e de dominância. A inclusão do efeito poligênico nos modelos avaliados permitiu um aumento significativo nas herdabilidades aditivas das características avaliadas.

GENERAL INTRODUCTION

To accelerate the selection process, Meuwissen, Hayes, and Goddard (2001) developed the genome-wide selection (GWS), which predicts the genomic genetic values using a large number of molecular markers, such as single nucleotide polymorphisms (SNPs).

However, the following 2 problems arise in the estimation of marker effects and in the prediction of genomic genetic values: the large number of parameters to be estimated (markers), in which multicollinearity is present and is usually much more than the number of observed individuals—a problem called as *'large p, small n'* (JANNINK; LORENZ; IWATA, 2010).

Several methods have been proposed to resolve this problem, which can be divided into 3 classes (RESENDE; SILVA AZEVEDO, 2014): explicit, implicit, and dimensional reduction regression. In the first class, the methods RR-BLUP, *Least Absolute Shrinkage and Selection Operator* (LASSO), Bayes A and Bayes B, among others, are well known. In the implicit regression class, Gianola and Campos (2009) suggested the semiparametric method *Reproducing Kernel Hilbert Spaces* (RKHS). Among the regression methods with dimensional reduction, independent components, partial minimum squares, and principal components (RESENDE; SILVA AZEVEDO, 2014) can be highlighted.

Another approach that can be used in genomic selection is the mixed model methodology (HENDERSON, 1984), in which the *pedigree*-based relationship matrix is replaced with a relationship matrix estimated by the markers; this method denominated G-BLUP (Genomic BLUP) (MEUWISSEN; HAYES GODDARD, 2001; VANRADEN, 2008). Among the additive models, Bayesian regressions and the GBLUP method are the most commonly used. The G-BLUP method is attractive because its computational implementation is simple as it utilizes programs that are based on the REML+BLUP procedure (Restricted Maximum Likelihood - REML and Best Linear Unbiased Predictor - BLUP) (DE LOS CAMPOS et al., 2013).

The superiority of G-BLUP versus the traditional BLUP (based on the *pedigree*) in predicting the genetic effects with greater accuracy has been proven in several animal species (DAETWYLER et al.; 2010; GARRICK, 2011; GONZALEZ-RECIO et al., 2008; WOLC et al., 2011) as well as in plant species (CROSSA et al., 2010; HESLOT et al., 2012; RESENDE et al., 2012; SPINDEL et al., 2015).

Genomic selection was initially proposed after considering the methods with only additive effects. Recently, the inclusion of non-additive effects, such as the dominance effect has also been evaluated in animal (ERTL et al., 2014; SU et al., 2012) and plant species (MUÑOZ et al., 2014; TECHNOW et al., 2012).

The dominance effect can be defined as the interaction between the alleles of the same locus (FALCONER; MACKAY, 1996). In the traditional genetic evaluation, this effect is rarely included in the models due to the need for a high proportion of full-sibs in the study population. In GWS, with the recent availability of genetic information through molecular markers and the development of genomic prediction methods, the estimation of dominance effects based on the markers has become more viable (XIANG et al., 2016).

One of the methods that were initially proposed for additive and dominance genomic models was GBLUP-D (TORO; VARONA, 2010; SU et al., 2012), which consists of replacing the additive and dominance relationship matrices, based on the pedigree, with the relationship matrices, based on markers. In addition to GBLUP-D, other methods such as the LASSO, Ridge and Bayesians regressions (Bayes A, Bayes B, and T-BLASSO) were proposed for the genomic models with additive and dominance effects (AZEVEDO et al., 2015; TORO VARONA, 2010; WELLMANN BENNEWITZ, 2012).

All these above-mentioned methods were proposed for situations in which the analyzed traits follow normal distribution and complete observations. However, according to de Los Campos, Gianola, and Allison (2010), the

genomic selection can be applied while also considering censored data, through the fit of survival models, such as the Cox model or different modifications of the linear model.

According to Schaeffer (2013), there are 3 possible situations for the analysis of censored data: (a) when censoring is removed from the analysis; (b) when censoring is included in the analysis, but are not properly considered, or (c) when censoring is included in the analysis and is properly considered. The latter approach constitutes survival analysis.

In this context, Santos et al. (2015) performed genomic prediction of censored phenotypes of a pig data set considering the Cox survival model with gaussian fragility, while considering only the additive genetic random effect. Thus, the main objective of this study was to perform genomic prediction of the censored phenotypes considering the dominance and polygenic effects, in addition to the additive effects, on the Cox model.

This manuscript discusses the present work in different chapters, as described: Chapter 1 presents a comparison between the proposed GBLUP-D method when using the *lmekin* function of R software and the GBLUP-D methods available in the GVCBLUP software and BGLR R-package. Chapter 2 discusses the genomic prediction considering different situations for censored data based on the additive-dominant models. Finally, Chapter 3 states the estimation of genomic variance components while considering combined marker and pedigree information in 2 datasets (without and with censoring).

REFERENCES

AZEVEDO, C. F. et al. Ridge, Lasso and Bayesian additive-dominance genomic models. **BMC genetics**, v. 16, n. 1, p. 105, 2015.

CROSSA, J. et al. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. **Genetics,** Austin, v. 186, p. 713-724, 2010.

DAETWYLER, H. D. et al. Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. **Animal Production Science**, v. 50, n. 12, p. 1004-1010, 2010.

DE LOS CAMPOS, G. et al. Whole-genome regression and prediction methods applied to plant and animal breeding. **Genetics**, v. 193, n. 2, p. 327-345, 2013.

DE LOS CAMPOS, G.; GIANOLA, D.; ALLISON, D. B. Predicting genetic predisposition in humans: the promise of whole-genome markers. **Nature Reviews Genetics**, London, v. 11, p. 880 – 886, 2010.

ERTL, J. et al. Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. **Genetics Selection Evolution**, v. 46, n. 1, p. 40, 2014.

FALCONER, D. F.; MACKAY, T. F. C. Variance. In Introduction and Quantitative Genetic. Longman 4 th edition, **Malaysia**. p.122-143, 1996.

GARRICK, D. J. The nature, scope and impact of genomic prediction in beef cattle in the United States. **Genetics Selection Evolution**, v. 43, n. 1, p. 17, 2011.

GIANOLA, D.; CAMPOS, G. de los. Inferring genetic values for quantitative traits non-parametrically. **Genetic Research**, v. 90, p. 525-540, 2009.

GONZÁLEZ-RECIO, O. et al. Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. **Genetics Selection Evolution**, v. 41, n. 1, p. 3, 2009.

HENDERSON, C. R. **Applications of linear models in animal breeding**. Guelph: University of Guelph, 1984.
HESLOT, N. et al. Genomic selection in plant breeding: a comparison of models. **Crop science**, v. 52, n. 1, p. 146-160, 2012.

JANNINK, J. L.; LORENZ, A. J.; IWATA, H. Genomic selection in plant breeding: from theory to practice. **Briefings in Functional Genomics**. v. 9, p. 166-177, 2010.

LEE, E. T.; WANG, W. J. **Statistical methods for survival data analysis**. Wiley series in probability and statistics, 2003. 513 p.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome wide dense marker maps. **Genetics**, v. 157, p. 1819-1829, 2001.

MUÑOZ, P. R. et al. Unraveling additive from nonadditive effects using genomic relationship matrices. **Genetics**, v. 198, n. 4, p. 1759-1768, 2014.

RESENDE, M. D. V.; SILVA, F. F.; AZEVEDO, C. F. **Estatística matemática, biométrica e computacional: modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência**. Viçosa: Editora Suprema, 2014. 881 p.

RESENDE, M. F. R. et al. Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). **Genetics**, v. 190, n. 4, p. 1503-1510, 2012.

SANTOS, V. S. et al. Genomic selection for slaughter age in pigs using the Cox frailty model. **Genetics and molecular research: GMR**, v. 14, n. 4, p. 12616-12627, 2015.

SCHAEFFER, L. Survival. In: History of genetic evaluation methods in dairy cattle (Grosu H, Schaeffer L, Oltenacu PA, et al., eds.) 279-298. 2013. https://xa.yimg.com/kq/groups/18395782/1926111600/name/FINAL_BOOK_29.04.2013.pdf

SPINDEL, J. et al. Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. **PLoS genetics**, v. 11, n. 2, p. e1004982, 2015.

SU, G. et al. Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. **PloS one**, v. 7, n. 9, p. e45293, 2012.

TECHNOW, F. et al. Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. **Theoretical and Applied Genetics**, v. 125, n. 6, p. 1181-1194, 2012.

TORO M. A.; VARONA L. A note on mate allocation for dominance handling in genomic selection. **Genetics Selection Evolution**, v. 42, n. 33, 2010.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, v. 91, n. 11, p. 4414 – 4423, 2008.

WELLMANN, R; BENNEWITZ, J. Bayesian models with dominance effects for genomic evaluation of quantitative traits. **Genetics research**, v. 94, n. 1, p. 21-37, 2012.

WOLC, A. et al. Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. **Genetics Selection Evolution**, v. 43, n. 1, p. 5, 2011.

XIANG, T. et al. Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. **Genetics Selection Evolution**, v. 48, n. 1, p. 92, 2016.

CHAPTER I

PROPOSAL OF GENOMIC BLUP WITH ADDITIVE AND DOMINANCE EFFECTS IN ENVIRONMENT *R*

Submitted to Biometric Brazilian Journal
Received April 15, 2016
Accepted September 29, 2016

ABSTRACT

Recently, dominance effects have been included in genomic selection of various species, being the GBLUP-D method the most frequently employed. This method consists in replacing, in the REML/BLUP procedure, the pedigree-based relationship matrices by marker-based relationship matrices. To implement this method the GVCBLUP software or the R-package BGLR, based on the Bayesian regression via the Hilbert Space Reproduction Kernel are utilized. This study aimed at assessing the possibility and effectiveness of applying the GBLUP-D through the *lmekin* function implemented in the R software *coxme* package via the inclusion of the additive and dominance genomic relationship matrices. Thus, by utilizing simulated data analyses, the findings resulting from the *lmekin* function were compared with those achieved through the GVCBLUP software and BGLR package. The results revealed that both the GBLUP and GBLUP-D methods fitted by REML in the GVCBLUP software and through the *lmekin* function were equivalent. The *lmekin* function in the R software has thus been shown to be an effective option for the fit of the genomic models which consider the additive and dominance effects.

KEY WORDS: Gibbs sampler, Variance components, Mixed linear model, SNPs.

1. INTRODUCTION

Animal and plant breeding traditionally employs the selection method based on the expected genetic values, normally estimated by the REML / BLUP procedure (estimation of the variance components through Residual Maximum Likelihood - REML, and the prediction of random effects by BLUP - Best Linear Unbiased Predictor), based on the phenotypic data and pedigree relationship matrix (GODDARD e HAYES, 2007).

However, as low cost and effective molecular markers became available, Meuwissen et al., (2001) proposed the Genomic Wide Selection (GWS), which implies the simultaneous selection of thousands of markers extending across the entire genome to facilitate the prediction of an individual's genetic value (MEUWISSEN et al., 2001; Et al., 2008).

Initially, the research on GWS had considered the additive genetic effect alone. However, more lately, the non-additive effects, like the dominance effect in several species, were also included in the genomic selection (SU et al., 2012; ERTL et al., 2014; MUÑOZ et al., 2014). The GBLUP method (referred to in the literature as GBLUP-D) was initially proposed for GWS, which involved both the additive and dominance effects; while it showed similarities to the REML / BLUP procedure, it differed by replacing the additive and dominance genetic relationship matrices based on the pedigree, with those based on the markers. This procedure is more advanced as it records the real and real relationship, and not a mean relationship based on pedigree (MEUWISEEN et al., 2001; RESENDE et al., 2008).

The GBLUP-D method is implemented in the GVCBLUP software (WANG et al., 2014) as well as in the R software (R DEVELOPMENT CORE TEAM, 2016) through the BGLR package - Bayesian Generalized Linear Regression (PÉREZ and DE LOS CAMPOS, 2014) based on the Reproducing Kernel Hilbert Space Regression (RKHS), applying the Bayesian technique.

Santos et al. (2015), applied the GBLUP method using the *lmekin* function implemented in the R-package *coxme*. This study aimed to extend this methodology including the dominance effect (GBLUP-D) and to compare, using simulated data, the results obtained using the *lmekin* function with those obtained in the GVCBLUP program and in the BGLR package.

2. MATERIAL AND METHODS

The data for analysis were drawn from a simulation study performed by Vitezica et al., (2013). These authors simulated a quantitative trait with additive and dominance effects, reflecting 20% and 10% of the phenotypic variance, respectively. The data included 2100 genotyped and phenotyped individuals, and 10,000 single nucleotide polymorphism markers (SNPs). The variance components were estimated, and the genomic genetic effects were predicted using the GBLUP and GBLUP-D methods, involving the GVCBLUP, *lmekin* and BGLR package.

2.1 GBLUP AND GBLUP-D METHODS IMPLEMENTED IN GVCBLUP SOFTWARE

The GBLUP method was implemented in the GVCBLUP software per the mixed linear model given below:

$$y = 1\mu + Za + e, \qquad (1)$$

in which *y* represents the vector of the phenotypes, μ is the general mean, a is the vector of the *i*-th additive genetic effects (*i* = 1, ..., n), Z is the incidence matrix for random effects *a*, that display normal distribution with mean 0 and covariance matrix equal to $G\sigma_a^2$; where G represents the additive genomic relationship matrix and σ_a^2 is the additive genetic variance. The error vector (e) also shows normal distribution with the variance-covariance matrix, equal to $I\sigma_e^2$.

The additive genomic relationship matrix (G) is expressed as:

$$G = WW' \Big/ 2\sum_{j=1}^{m} p_j \left(1 - p_j\right),$$

in which the w_{ij} values of the matrix W are equal to 0-2p, 1-2p and 2-2p for the marker genotypes of the mm, Mm and MM types, respectively; pj is the allelic frequency of M at locus j (VITEZICA et al.; RESENDE et al., 2014).

By including the dominance effect, we have the GBLUP-D method based on the mixed linear model as given below:

$$y = 1\mu + Za + Td + e, \quad (2)$$

where y, μ, a, Z and e are defined similarly as in model (1) and d is the vector of the i-th dominant deviations (i = 1, ..., n), in which T is the incidence matrix of the respective random effects d, which show normal distribution with mean 0 and covariance matrix equal to $D\sigma_d^2$, where $D\sigma_d^2$ is the dominance genetic variance and D is the dominance genomic relationship matrix, as expressed by the expression given below (VITEZICA et al. , 2007).

$$D = SS' \Big/ \sum_{j=1}^{m} \left\{2p_j \left(1 - p_j\right)\right\}^2,$$

in which the s_{ij} values of the matrix S are equal to $-2p^2$, $2p(1-p)$ e $-2(1-p)^2$ for the marker genotypes of the mm, Mm and MM types, respectively; and p_j is the same as in matrix G.

Both matrices were obtained in the GVCBLUP software, which shows five other different methods of identifying the genomic relationship matrices. In this work, definition 1 was employed for the G and D matrices (VANRADEN, 2008; WANG et al., 2014).

The variance components in both methods were estimated by applying the REML procedure using the maximization method AI (Average Information), based on first and second orders of partial derivatives (RESENDE et al., 2014). Besides the AI-REML, the EM-REML (Expectation-Maximization) method was also implemented in the GVCBLUP software (WANG et al., 2014). The additive

and of dominance genetic values of the individuals were predicted through the mixed model equations, which in the case of the complete model involving the additive and dominance effects, are expressed as:

$$\begin{bmatrix} X'X & X'Z & X'Z \\ Z'X & Z'Z+G^{-1}\sigma_e^2/\sigma_a^2 & Z'Z \\ Z'X & Z'Z & Z'Z+D^{-1}\sigma_e^2/\sigma_d^2 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{a} \\ \hat{d} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ Z'y \end{bmatrix}$$

in which, σ_a^2 and σ_d^2 are the additive and dominance genetic variances, respectively and σ_e^2 is the residual variance. For the model with additive effects only, the mixed model equations are expressed as follows:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z+G^{-1}\sigma_e^2/\sigma_a^2 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

2.2 THE GBLUP AND GBLUP-D METHODS USING THE BGLR PACKAGE

The GBLUP and GBLUP-D methods were implemented in the BGLR package as specific cases of the Bayesian reproducing kernel Hilbert spaces regressions (RKHS), as the kernel matrix K can be calculated from the pedigree and / or markers (DE LOS CAMPOS et al., 2009).

The form of fit of the method is presented considering the model with additive and dominance effects, with the GBLUP method being a specific case (when the dominance effect is not estimated). For the GBLUP-D method utilizing the RKHS, the following model was expressed (PÉREZ e DE LOS CAMPOS, 2014):

$$y = 1\mu + Za + Td + e, \quad (3)$$

in which y, μ, Z, T and *e* are defined as in model (2), *a* is the vector of the individual additive genetic values and *d* is the vector of the dominance deviations. Multivariate normal distribution is noted for the conditional data, whereas for μ, *a* and *d*, the *a priori* distributions given below were assumed:

$$p(\mu) \propto \text{constant};\ a|K_1,\sigma_a^2 \sim N(0,K_1\sigma_a^2) \text{ and } d|K_2,\sigma_d^2 \sim N(0,K_2\sigma_d^2),$$

in which the kernel matrices K_1 and K_2 show dimensions equal to the number of individuals; in the GBLUP-D method, K_1 and K_2 were replaced by additive (G) and dominance (D) genomic relationship matrices. The BGLR package implements the GBLUP-D method through the RKHS regression by directly providing the G and D matrices or by decomposing their eigenvalues. When relationship matrices are used, the eigenvalues are internally decomposed (PÉREZ e DE LOS CAMPOS, 2014). For the variance components σ_a^2, σ_d^2 and σ_e^2, *a priori* distributions were assumed $\sigma_i^2 \sim \chi^{-2}\left(\sigma_i^2 \left| df_i, S_i \right.\right)$, with the degrees of freedom df_i and parameters $S_i > 0$ specified, based on the standard configurations of the package (PÉREZ e DE LOS CAMPOS, 2014). In the GBLUP method involving the additive effects alone, the process is similar considering the effects of dominance (D) equal to 0 (DE LOS CAMPOS, 2009).

The GBLUP and GBLUP-D methods involving RKHS were implemented by applying the Bayesian approach with the Gibbs sampler algorithm. Of the 60,000 samples generated, the initial 10,000 were rejected to ensure stationarity of the stochastic process and to guarantee independence between the samples. An interval of 5 was considered between the saved samples. The chain convergence was confirmed employing the Geweke criterion (GEWEKE, 1992) implemented in the BOA R-package (SMITH, 2007).

2.3 THE GBLUP AND GBLUP-D METHODS EMPLOYING THE *LMEKIN* FUNCTION OF R

The GBLUP method involving the *lmekin* function implemented in the *coxme* package (THERNEAU, 2012) was defined similarly as in (1). By including the dominance effect, the function disallows the simultaneous prediction of two random effects. A good alternative for the GBLUP-D method is the linear mixed model given below:

$$y = 1\mu + Zg + e, \quad (4)$$

where y, μ, and e are as defined earlier in (1) and (2), and $g = a + d$ is the vector of the i-th predicted total genotypic values, which are equal to the additive genetic effects and dominance, assumed to have normal distribution, with mean 0 and matrix of variances and covariances $\Sigma = G\sigma_a^2 + D\sigma_d^2$, with *a, d,* σ_a^2, σ_d^2, G and D defined above. The equivalence between models 2 and 4 is because the covariance between the additive and dominance effects is considered null (VITEZICA et al., 2013). Mrode (2005, p.195), proposed that the individual additive genetic values (*a*) and the dominance deviations (*d*) can be calculated separately from the total genetic values (*g*) via the respective expressions:

$$\hat{a} = \sigma_a^2 G\Sigma^{-1}\hat{g} \text{ and } \hat{d} = \sigma_d^2 D\Sigma^{-1}\hat{g}.$$

Similar to the GVCBLUP software, the variance components were estimated in both GBLUP and GBLUP-D, via the REML procedure, by maximizing the marginal likelihood of residues from the generalized least squares fit (VENABLES e RIPLEY, 2002). To maximize this likelihood function, the *lmekin* function initially employed the EM algorithm, and following a suitable number of iterations, the Newton-Raphson algorithm (PINHEIRO and BATES, 2000) was applied.

2.4 HERITABILITY ESTIMATES AND MODEL COMPARISON

The estimates of heritability additive or restricted sense, dominance and the broad sense (H2), considering the three methods explained above, were obtained by the respective expressions:

$$h_a^2 = \left[\sigma_a^2 / \left(\sigma_a^2 + \sigma_d^2 + \sigma_e^2\right)\right];\ h_d^2 = \left[\sigma_d^2 / \left(\sigma_a^2 + \sigma_d^2 + \sigma_e^2\right)\right];\ H^2 = h_a^2 + h_d^2.$$

The dominance effect in the GBLUP method was included by employing the *lmekin* function which was tested with the Likelihood Ratio Test (LRT), and expressed as,

$$\Lambda = 2\ln L(V) - 2\ln L(U),$$

where L (V) is the likelihood of the model having additive and dominance effects and L (U) is the likelihood of the model having additive effects alone (likelihood under H0). The test statistic has a chi-square distribution with degrees of freedom given by the difference between the number of variance components between the two models evaluated. The p-value for the test is given by 0.5 [1-P ($\chi_1^2 \leq \lambda$)], where λ is the observed value of Λ (VISSCHER, 2006; ERTL et al., 2014).

For the fitted models employing the BGLR package, the inclusion of the dominance effect was assessed by utilizing the DIC (Deviance Information Criterion) criterion, as expressed below:

$$DIC = \bar{D} + P_D$$

where $\bar{D}$ represents the posterior mean of the deviance and P_D implies the "effective number of parameters", the best model being the one showing the lowest DIC value (SPIEGELHALTER et al., 2002).

The predictive ability of the methods employed in each approach was evaluated depending upon the accuracy achieved by the correlation between the true genetic values (TGV) and the predicted ones (PGV). The regression coefficients of TGV on PGV were also estimated, which revealed whether the prediction was biased or not (RESENDE et al., 2014; VITEZICA et al., 2013). The results were based on the average of ten replications of the data set available at http://genoweb.toulouse.inra.fr/~zvitezic/simu_for_genetics.tar.gz (VITEZICA et al., 2013).

3. RESULTS AND DISCUSSION

The chain convergence to the Bayesian techniques of GBLUP (additive effects) and GBLUP-D (additive and dominance effects) adjusted in the BGLR package was pleased, in keeping with the Geweke criterion, where the p-value obtained was always higher than the 5%, not excluding the

hypothesis of equality between the means of the first 10% iterations post burn-in and the last 50% iterations of the chain (NOGUEIRA et al., 2004).

It is evident in the GBLUP method (additive genetic variance only - Fig. 1a) and GBLUP-D (both additive and dominance genetic variances - Figs. 1b, 1c) that the variances exhibited similarity when estimations were done using the *lmekin* and GVCBLUP software, but were not the same when the BGLR package was used. When this same package was employed to assess the additive variances via the GBLUP method, values marginally higher than those obtained by the *lmekin* function and GVCBLUP software were observed (Fig.1a). However, this was not the case when the GBLUP-D method was utilized (Fig. 1b).

The estimates of the genetic variances resulting from dominance were noted to be more than the simulated value of 10 in all the replications of the simulated data set, with estimates drawn from utilizing the BGLR package with the GBLUP-D method (Fig. 1c) in the range of 11.34 to 18.12. However, when the GVCBLUP program and the *lmekin* function were used, these estimates were in the range of 5.81 to 14.81, and between 4.81 and 16.07, respectively.

From Table 1, the heritability estimates recorded when the GVCBLUP software and *lmekin* function were used, revealed similar values, around 0.202, 0.125 and 0.327 for the additive, dominance, and wide-sense heritabilities, respectively, with 0.03 standard deviation values; the results agreed with the simulated values and indicated the accurate implementation of the GBLUP and GBLUP-D methods using the *lmekin* function.

When the dominance effect was included, a drop of around 8.0% in the additive heritability was observed, only when the BGLR package was used. De Los Campos et al., (2009) highlighted that the GBLUP and GLUP-D methods fitted by the BGLR package were regarded as cases of the RKHS method, in which the genomic relationship matrices G and D are treated as the parametric kernels.

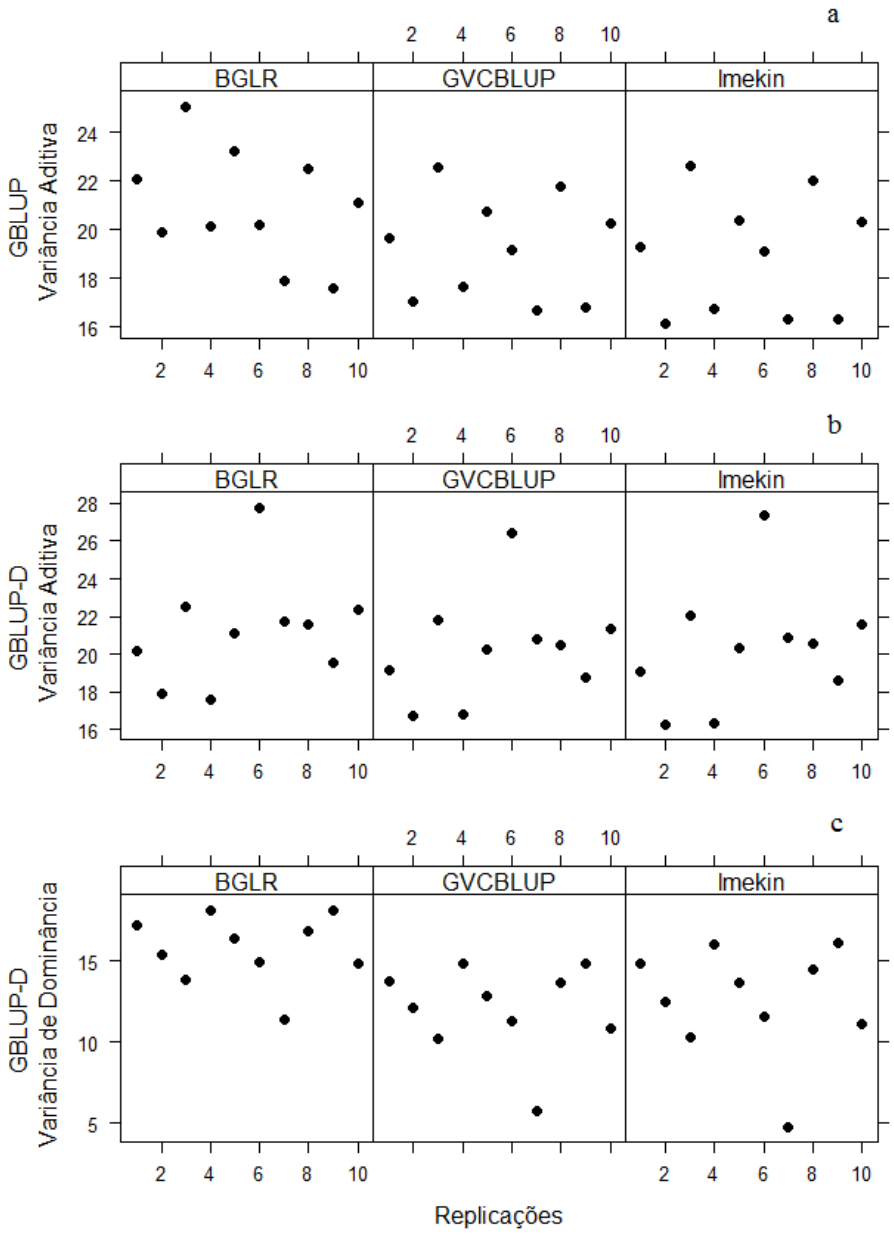

Figure 1: Genetic additive and dominance variances were assessed using the GBLUP model (with additive effects only) and GBLUP-D model (with the additive and dominance effects) using the BGLR, GVCBLUP and *lmekin* function.

When the BGLR method was used the estimates of dominance heritability were about 23.2% higher than those obtained via the GVCBLUP and *lmekin*, thus overestimating the true simulated value of 0.10. This occurred because the dominance variance was overestimated, a value equal to 15.68, with a standard deviation of 2.09, based on the ten replicates of the dataset

assessed. Thus, the broad sense heritability was also overestimated, showing a value of 0.362 ± 0.02. Morota et al., (2014) reported similar results assigning them to the absence of orthogonality between the genomic relationship matrices G and D, being one of the likely reasons for this overestimation. A suitable alternative suggested by the authors was to achieve the matrix $GD=X_{AD}X'_{AD}$, where $GD=X_{AD}$ is the additive x dominance matrix (Xu et al., 2013).

Table 1 - Additive (h_a^2), dominance (h_d^2), total (H_2) heritabilities and time of execution, considering the GVCBLUP software, *lmekin* function and the BGLR package in simulated data.

Models	h_a^2	h_d^2	H^2	Time*
True value	0.20	0.10	0.30	-
GBLUP				
BGLR	0.226 ± 0.03	-	-	32.52
GVCBLUP	0.202 ± 0.03	-	-	1.10
Lmekin	0.202 ± 0.03	-	-	21.04
GBLUP-D				
BGLR	0.208 ± 0.03	0.154 ± 0.02	0.362 ± 0.02	49.25
GVCBLUP	0.202 ± 0.03	0.125 ± 0.03	0.327 ± 0.03	2.41
Lmekin	0.202 ± 0.03	0.125 ± 0.03	0.327 ± 0.03	44.12

* Time in minutes employing an Intel Core i5 1.70 GHz computer and 4GB RAM.

The analysis of Vitezica et al., (2013) using this very same data set with Bayesian inference, reported the dominance genetic variance estimates of 13.2 ± 3.5, bearing close similarity to the findings in this work, using the *lmekin* function and the GVCBLUP software.

Although identical results were obtained, the GVCBLUP software shows faster prediction of the genetic values than the *lmekin* function, as is evident from Table 1, where the software fitted the models in 1 minute and 10 seconds on average for the GBLUP method and in 2 minutes and 41 seconds on average for the GBLUP-D method. The GVCBLUP software, like the R, is free and available online at http://animalgene.umn.edu/. It is programmed in C ++ and employing parallel computing with shared memory in the inversion of dense matrices such as (G) and dominance (D), thus optimizing the

computational time in the GBLUP and GBLUP-D methods (WANG et al., 2014).

The *lmekin* function fitted the same models in roughly 21 minutes (GBLUP) and 44 minutes (GBLUP-D). From these three approaches, the one that involved the longest time to perform was the model fitted using the BGLR package, showing 33 and 49 minutes on average for the GBLUP and GBLUP-D methods, respectively, utilizing 60,000 iterations in each method. The *lmekin* function showed the benefit of extending the proposed methodology, like the Cox fragility model for censored data, as done by Santos et al. (2015) for additive effects. Another benefit revealed by the *lmekin* function was the availability of the log-likelihood values, which enabled testing the significance of the dominance effect by the LRT, as shown in Table 2.

Table 2 - -2 log-likelihood values, χ2 and their corresponding p-values of the likelihood ratio test for the fitted models employing the *lmekin* function, besides the adjusted DIC values in the BGLR package.

Replications	Models	*-2logL*	valor-χ^2	valor-p	DIC
1	GBLUP	15491.08			15411.40
	GBLUP-D	15472.41	18.671	$7.76*10^{-6}$	15282.81
2	GBLUP	15530.00			15465.98
	GBLUP-D	15515.4	14.599	$6.64*10^{-5}$	15368.96
3	GBLUP	15471.45			15372.69
	GBLUP-D	15459.89	11.554	0.00034	15283.43
4	GBLUP	15539.04			15473.90
	GBLUP-D	15519.36	19.677	$4.58*10^{-6}$	15346.65
5	GBLUP	15499.99			15413.64
	GBLUP-D	15485.44	14.545	$6.84*10^{-5}$	15295.45
6	GBLUP	15472.31			15347.86
	GBLUP-D	15458.72	13.591	0.00011	15227.57
7	GBLUP	15467.66			15378.73
	GBLUP-D	15465.57	2.093	0.07399	15325.05
8	GBLUP	15509.24			15427.04
	GBLUP-D	15490.48	18.761	$7.40*10^{-6}$	15293.45
9	GBLUP	15501.25			15427.79
	GBLUP-D	15481.31	19.938	$3.99*10^{-6}$	15285.03
10	GBLUP	15473.51			15379.00
	GBLUP-D	15462.24	11.271	0.00039	15279.20

Table 2 also shows the DIC values for the GBLUP and GBLUP-D methods adjusted via the BGLR package, indicating that the best model is the one that includes both the additive and dominance effects.

Table 3 shows the accuracies and regression coefficients for each model. By including the dominance effect, the accuracy between the predicted and true genetic values remain unaltered, at about 0.72 ± 0.02, despite the significant dominance effect (Table 2). Similar findings were reported by Ertl et al., (2014) where a significant dominance effect was found in five of the nine traits evaluated; however, when this effect was included, the accuracy remained constant.

Table 3 - The accuracy (r) and regression coefficients (*B*, measure of bias) between the predicted and true genetic values for ten simulated data replications, utilizing the GVCBLUP software, *lmekin* function and BGLR package.

Models	*r*	*B*
GBLUP		
GVCBLUP	0.72 (0.02)	0.97 (0.06)
Lmekin	0.72 (0.02)	0.97 (0.06)
BGLR	0.72 (0.02)	0.93 (0.05)
GBLUP-D		
GVCBLUP	0.72 (0.02)	0.97 (0.06)
Lmekin	0.72 (0.02)	0.97 (0.06)
BGLR	0.72 (0.02)	0.96 (0.06)

High accuracy values enable higher genetic gain over the short term. Several works on animal and plant breeding have emphasized that the GWS is superior in terms of the usual selection, in which, for instance, a 33% increase was reported in the selective accuracies for beef cattle (VANRADEN et al., 2009), and 25% in sheep (DUCHEMIN et al., 2012) when marker information was considered instead of only pedigree.

Genomic accuracy above 90% was reported in plant species like corn (FRITSCHE NETO et al., 2012) and wheat (HESLOT et al., 2012). One of the factors involved in this superior genomic selection, results from utilizing the

relationship matrix performed instead of the average relationship matrix as in the usual selection (RESENDE, 2008).

Resende et al. (2014) showed that there were five factors which determined the accuracy of GWS, namely: (i) the spacing between the markers, based on their number and genome size; (ii) the number of individuals phenotyped and genotyped in the training population; (iii) the heritability of the trait; (iv) the number of loci that control the trait and distribution of its effects; and (v) the effective population size.

In all the three approaches employed as the regression coefficients were around 1, it implied that the predictions were non-biased (RESENDE et al., 2014). Like the accuracy, the regression coefficients obtained agreed with the reports of Vitezica et al., (2013) in their estimation of the same data set, but with smaller standard deviations (0.96 ± 0.08).

The only way to implement the GBLUP-D method via REML, using software R, was via the *lmekin* function. Besides the programs listed here, software like GS3 (LEGARRA et al., 2011; VITEZICA et al., 2013; AZEVEDO et al., 2015); ASReml (GILMOUR et al., 2015, MUÑOZ et al., 2014) and DMU (SU et al., 2012; MADSEN, 2013) also enable genomic prediction based on mixed models, which include the additive and non-additive effects.

Estimating the dominance effects using the genetic correlation matrix, involving the markers, is easier in terms of implementation than the fit of the pedigree-based models (SU et al., 2012; VITEZICA et al., 2013). Besides, several studies have utilized additive effects alone (HESLOT et al., 2012; RESENDE et al., 2012; VANRADEN, 2008) and only more recently, both additive and non-additive effects have been explored (ERTL et al.2014; VITEZICA et al., 2013). Also, the genotypic BLUP method (GBLUP-D) is proven to be superior to that of the conventional pedigree-based BLUP (AZEVEDO et al., 2015; MUÑOZ et al., 2014).

4. CONCLUSIONS

The possibility and effectiveness of accomplishing the genomic prediction of the individual genetic values that consider both additive and dominance effects, have been confirmed via the *lmekin* function of R software. The results obtained exactly matched those of the GVCBLUP software already available for such analysis. The GBLUP-D method implemented in the BGLR package through the Reproduction Kernel Hilbert Space (RKHS) requires greater study, as the variance estimates of dominance were overestimated.

Acknowledgments

The authors express their gratitude to CAPES for the grant awarded to the first author for assistance in the doctoral program in Brazil and Sandwich Doctorate scholarship abroad under the process of nº BEX 9415 / 14-9.

REFERENCES

AZEVEDO, C. F.; RESENDE, M. D. V.; SILVA, F. F.; VIANA, J. M. S.; VALENTE, M. S. F.; RESENDE, M. F. R.; MUÑOZ, P. Ridge, Lasso and Bayesian additive-dominance genomic models. *BMC Genetics*, London, v.16, n.105, p.1-13, 2015.

DE LOS CAMPOS, G.; GIANOLA, D.; ROSA, G. J. M. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *Journal of Animal Science*, Champaign, v.87, n.6, p.1883-1887, 2009.

DUCHEMIN, S. I.; COLOMBANI, C.; LEGARRA, A.; BALOCHE, G.; LARROQUE, H.; ASTRUC, J. M; BARILLET, F.; ROBERT-GRANIÉ, C.; MANFREDI, E. Genomic selection in the French Lacaune dairy sheep breed. *Journal of Science Dairy*, Champaign, v.95, n.5, p.2723–2733, 2012.

ERTL, J.; LEGARRA, A.; VITEZICA, Z.; VARONA, L.; EDEL, C.; EMMERLING, R.; GÖTZ, K. U. Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. *Genetics Selection Evolution*, London, v.46, n.1, p.40-49, 2014.

FRITSCHE-NETO, R., RESENDE, M. D. V., MIRANDA, G. V., VALE, J. C. Seleção genômica ampla e novos métodos de melhoramento do milho. *Ceres*, Viçosa, v.59, n.6, p.794-802, 2015.

GEWEKE, J. Evaluating the accuracy of sampling based approaches to the calculation of posterior moments. In: BERNARDO, J. O.; BERGER, J. M.; DAWID, A. P.; SMITH, A. F. M. (Ed.). *Bayesian statistics 4*. Oxford: Oxford University Press, 1992. p. 169–194.

GILMOUR, A. R.; GOGEL, B. J.; CULLIS, B. R.; WELHAM, S. J.; THOMPSON, R. (2015). ASReml User Guide Release 4.1 Structural Specification, VSN International Ltd, Hemel Hempstead, HP11ES, UK <www.vsni.co.uk>

GODDARD, M. E.; HAYES, B. J. Genomic selection. *Journal of Animal Breeding and Genetics*, Austin, v.124, n.6, p.323-330, 2007.

HESLOT, N.; YANG, H. P.; SORRELLS, M. E.; JANNINK, J. L. Genomic selection in plant breeding: a comparison of models. *Crop Science*, Madison, v.52, n.1, p.146-160, 2012.

LEGARRA, A.; RICARD, A.; FILANGI, O. (2011). GS3: Genomic Selection, Gibbs Sampling, Gauss-Seidel (and BayesCpi). <http://genoweb.toulouse.inra.fr/~alegarra/> Acesso em: 27 mar. 2015.

MADSEN, P.; JENSEN, J. (2013). A user's guide to DMU. A Package for Analysing Multivariate Mixed Models. Version 6, release 5.2.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome wide dense marker maps. *Genetics*, Austin, v.157, n.4, p.1819-1829, 2001.

MOROTA, G.; BODDHIREDDY, P.; VUKASINOVIC, N.; GIANOLA, D.; DENISE, S. Kernel-based variance component estimation and whole-genome prediction of pre-corrected phenotypes and progeny tests for dairy cow health traits. *Frontiers in Genetics*, Lausanne, v.5, n.56, doi: 10.3389/fgene.2014.00056, 2014.

MRODE, R. A. *Linear models for the prediction of animal breeding values*. 2th Edition. Wallingford: CAB International, 2005. 358p.

MUÑOZ, P. R.; RESENDE, M. F.; GEZAN, S. A.; RESENDE, M. D. V.; DE LOS CAMPOS, G.; KIRST, M.; HUBER, D.; PETER, G. F. Unraveling Additive

from Nonadditive Effects Using Genomic Relationship Matrices. *Genetics*, Austin, v.198, n.4, p.1759-1768, 2014.

NOGUEIRA, D. A.; SAFÁDI, T.; FERREIRA, D. F. Avaliação de critérios de convergência para o método de Monte Carlo via Cadeias de Markov. *Revista Brasileira de Estatística*, Rio de Janeiro, v.65, n.224, p.59-88, 2004.

PINHEIRO, J. C.; BATES, D.M. *Mixed-Effects Models in S and S-PLUS*. New York: Springer-Verlag, 2000. 537 p.

PÉREZ, P.; DE LOS CAMPOS, G. Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, Austin, v.198, n.2, p.483-495, 2014.

R DEVELOPMENT CORE TEAM. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, Version 3.2.4 (16-03-2016). Disponível em: <http://www.R-project.org>

RESENDE, M. D. V.; LOPES, P. S.; SILVA, R. L.; PIRES, I. E. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. *Pesquisa Florestal Brasileira*, Colombo, v.56, n.1, p.63-78, 2008.

RESENDE JR., M. F. R.; MUÑOZ, P.; RESENDE, M. D. V.; GARRICK, D. J.; FERNANDO, R. L.; DAVIS, J. M.; JOKELA, E. J.; MARTIN, T. A.; PETER, G. F.; KIRST, M. Accuracy of Genomic Selection Methods in a Standard Dataset of Loblolly Pine (Pinus taeda L.). *Genetics*, Austin, v.190, n.1, p.1503-1510, 2012.

RESENDE, M. D. V.; SILVA, F. F.; AZEVEDO, C. F. *Estatística matemática, biométrica e computacional: modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência*. Viçosa: Editora Suprema, 2014. 881 p.

SANTOS, V. S.; MARTINS, F. S.; RESENDE, M. D.; AZEVEDO, C. F.; LOPES, P. S.; GUIMARÃES, S. E.; GLÓRIA, L.; SILVA, F. F. Genomic selection for slaughter age in pigs using the Cox frailty model. *Genetics and Molecular Research*, Ribeirão Preto, v.14, n.4, p.12616-12627, 2015.

SMITH, B. J. boa: An R Package for MCMC Output Convergence Assessment and Posterior Inference. *Journal of Statistical Software*, Washington, v.21, n.11, p.1-37, 2007.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; VAN DER LINDE, A. *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*), v.64, n.4, p.583-639, 2002.

SU, G.; CHRISTENSEN, O.F.; HENRYON, M.; LUND, M.S. Estimating additive and non-additive genetic variances and prediction genetic merits using genome-wide dense single nucleotide polymorphism markers. *PLoS One*, San Francisco, v.7, n.9, p.e45293, 2012.

THERNEAU, T. coxme: Mixed Effects Cox Models (2012). R package version 2.2-3. <http://CRAN.R-project.org/package=coxme>. Acesso em: 04 mar. 2016.

VANRADEN, P. M. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, Champaign, v.91, n.11, p.4414 – 4423, 2008.

VANRADEN, P. M.; VAN TASSELL, C. P.; WIGGANS, G. R.; SONSTEGARD, T. S.; SCHNABEL, R. D.; TAYLOR, J. F.; SCHENKEL, F. S. Invited review: reliability of genomic predictions for North American Holstein bulls, *Journal of Dairy Science*, Champaign, v.92, n.1, p.16–24, 2009.

VENABLES, W.N.; RIPLEY, B.D. *Modern Applied Statistics with S-PLUS*. 4th Edition. New York: Springer Verlag, 2002. 504p.

VISSCHER, P. M. A note on the asymptotic distribution of likelihood ratio tests to test variance components. *Twin research and human genetics*, Cambridge, v.9, n.4, p.490-495, 2006.

VITEZICA, Z. G.; VARONA, L.; LEGARRA, A. On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics*, Austin, v.195, n.4, p.1223-1230, 2013.

WANG, C.; DA, Y. Quantitative genetics model as the unifying model for defining genomic relationship and inbreeding coefficient. *PloS One*, San Francisco, v.9, n.12, p.e114484, 2014.

XU, S. Mapping quantitative trait loci by controlling polygenic background effects. *Genetics*, Austin, v.195, n.4, p.1209-1222, 2013.

CHAPTER II

GENOMIC PREDICTION FOR ADDITIVE AND DOMINANCE EFFECTS OF CENSORED TRAITS IN PIGS

Running title: Genomic additive and dominance variance of censored traits

V.S. Santos[1], S. Martins Filho[1], M.D.V. Resende[2], C.F. Azevedo[1], P.S. Lopes[3], S.E.F. Guimarães[3] and F. F. Silva[3]

[1]Departamento de Estatística, Universidade Federal de Viçosa, Viçosa, MG, Brasil

[2]Empresa Brasileira de Pesquisa Agropecuária, Centro Nacional de Pesquisa de Florestas, Colombo, PR, Brasil

[3]Departamento de Zootecnia, Universidade Federal de Viçosa, Viçosa, MG, Brasil

Corresponding author: V.S. Santos

E-mail: 2santosvinicius@gmail.com

Departamento de Estatística, Universidade Federal de Viçosa, Avenida Peter Henry Rolfs, s/n, Campus universitário 36570-900, Viçosa, MG, Brasil

Published in Genet Mol Res. 2016 Oct 17;15(4). doi: 10.4238/gmr15048764.

ABSTRACT

Age at the time of slaughter is a commonly used trait in animal breeding programs. Since studying this trait involves incomplete observations (censoring), analysis can be performed using survival models or modified linear models, for example, by sampling censored data from truncated normal distributions. For genomic selection, the greatest genetic gains can be achieved by including non-additive genetic effects like dominance. Thus,

censored traits with effects on both survival models have not yet been studied under a genomic selection approach. We aimed to predict genomic values using the Cox model with dominance effects and compare these results with the linear model with and without censoring. Linear models were fitted via the maximum likelihood method. For censored data, sampling through the truncated normal distribution was used, and the model was called the truncated normal via Gibbs sampling (TNL). We used an F_2 pig population; the response variable was time (days) from birth to slaughter. Data were previously adjusted for fixed effects of sex and contemporary group. The model predictive ability was calculated based on correlation of predicted genomic values with adjusted phenotypic values. The results showed that both with and without censoring, there was high agreement between Cox and linear models in selection of individuals and markers. Despite including the dominance effect, there was no increase in predictive ability. This study showed, for the first time, the possibility of performing genomic prediction of traits with censored records while using the Cox survival model with additive and dominance effects.

Key words: GBLUP; Censored data; Mixed model; Survival models

1. INTRODUCTION

Genomic selection (GS) in pigs has been performed mainly for traits of economic importance, for those related to growth and reproduction, and for carcass traits. However, GS can also be applied toward traits such as age at slaughter, which is defined as the length of productive life (for instance, the time in days from birth to slaughter). Since this involves the presence of incomplete observations (i.e. censored data), not all individuals have achieved the desired slaughter weight, and the analysis of this trait can be performed by means of survival analysis models or different modifications of the linear model (Serenius et al., 2006; Hou et al., 2009).

In animal breeding, Weibul and Cox survival models have been fitted using Survival Kit software (Ducrocq et al. 2010), which only estimates the additive effects based on pedigree using the Bayesian approach.

In the context of GS, Santos et al. (2015) were the pioneers when it came to be performing the genomic prediction of censored phenotypes while using the Cox survival model. As such, the model with only additive effects was considered, and was fitted by using the *coxme* function of R software, since the relationship matrix based on pedigree was replaced by the relationship matrix based on markers. Furthermore, in GS, Kärkkäinen and Sillanpaa (2013) proposed the use of Bayesian threshold models for data in binary and ordinal scale in order to evaluate censored traits.

Recently, non-additive effects such as dominance have been included in the GS of pigs (Su et al., 2012; Costa et al., 2015). This effect is defined as the interaction between alleles at the same locus, and it is expected that the inclusion of this effect in genomic models will increase prediction accuracy (Su et al., 2012).

To study additive and non-additive effects in GS, the method that was originally proposed was the GBLUP-D, which consists of replacing additive and dominance genomic relationship matrices based on pedigree with marker-based relationship matrices. Besides the GBLUP-D, other methods such as LASSO, Ridge, and Bayesian (Bayes, Bayes B, t-BLASSO) were proposed for additive-dominance genomic models (Azevedo et al., 2015).

However, no study was performed based on survival models considering both effects (additive and dominance) so much by usual analysis based on pedigree by genomic selection. Thus, we aimed to build upon the method proposed by Santos et al. (2015) to include dominance effects and evaluate their performance. In addition, we also aimed to compare the proposed model with alternative linear models in the presence and absence of censored observations.

2. MATERIAL AND METHODS

Population establishment and the phenotypic measurements were performed at the Pig Breeding Farm of the Department of Animal Science, Universidade Federal de Viçosa (UFV), MG, Brazil. The phenotypic data consisted of an F_2 population that was generated by crossing 11 boars and 54 dams that were randomly selected from the F_1 generation, which was initially created by crossing two natives Brazilian Piau boars with 18 commercial sows (Landrace × Large White × Pietrain). The use of these animals was reviewed and approved by the Bioethics committee of the Department of Veterinary Medicine (DVT-UFV) in agreement with the Guide to the Care and Use of Experimental Animals of the Canadian Council on Animal Care.

DNA was extracted from the white blood cells of parental, F_1, and F_2 animals at the Animal Biotechnology Laboratory of the Animal Science Department of the Universidade Federal de Viçosa. More details can be found in Band et al. (2005). From the F_2 population, 345 animals were genotyped for 384 single nucleotide polymorphisms (SNPs). The low-density customized SNPChip used was based on the Illumina Porcine SNP60 BeadChip.

These SNPs were selected according to QTL positions that were previously identified in this population using meta-analyses and fine mapping (Verardo et al., 2015). Therefore, although a small number of markers were used, the customized SNPchip based on previously identified QTL positions ensures appropriate coverage of the relevant genome regions in this population. From these, 66 SNPs were discarded because of a low-genotyping call rate (< 0.95), and from the remaining 318 SNPs, 81 were discarded due to a minor allele frequency (MAF) < 0.05. Thus, 237 SNP markers were distributed on the *Sus scrofa* chromosomes (SSC) as follows: SSC1 (n = 56), SSC4 (n = 54), SSC7 (n = 59), SSC8 (n=31), SSC17 (n = 25), and SSCX (n = 12).

In order to identify individuals with rapid weight gain for slaughter, the time in days from birth until the slaughter of the animal was considered a

response variable. The desired weight of the animals at the time of slaughter in this population was around 65 kg (Band et al., 2005). The exact time it took an animal to gain the desired weight was not known, as daily weighing was impractical. Only the ages and weights of the animals at the time of slaughter were known.

As a result, censoring was created based on slaughter weight, i.e., animals that did not reach 65 kg were referred to as censored (event = 0), whereas animals that reached that weight or more were referred to as a failure (event = 1). In this dataset, the proportion of censoring was around 0.561, i.e., about 44% of the animals weighed at least 65 kg. The data used in the analysis was adjusted for the fixed effects of sex and contemporary group (Silva et al., 2013), and the halothane gene was included as an additional marker.

2.1. STATISTICAL MODELS

The models shown in Table 1 were fitted to compare estimates of variance components and predicted genetic values using linear and survival models, with and without the effects of dominance and censoring.

According to Schaeffer (2013), there are three possible situations for the analysis of censored data: (1) when the censoring is removed in the analysis; (2) when the censoring is included in the analysis, but is not properly taken into account, or (3) when the censoring is included in the analysis and is properly taken into account. AL, ADL, AUC, and ADUC models belong to situation (2) and ATN, ADTN, AC, and ADC models to situation (3). In the ATN and ADTN models, the censored records were simulated from truncated normal distributions (Pérez and de los Campos, 2014).

Table 1. Evaluated models for time in days from birth to slaughter of the animal, considering censoring and additive and dominance effects.

Models	Additive Effect	Additive and Dominance Effect
Uncensored	Additive Linear (AL)	Additive-dominance Linear via ML (ADL)
	Additive Uncensored Cox (AUC)	Additive-dominance Uncensored Cox (ADUC)
Censored	Additive Truncated Normal (ATN)	Additive-dominance Truncated Normal (ADTN)
	Additive Cox (AC)	Additive-dominance Cox (ADC)

2.1.1 ADDITIVE (AL) AND ADDITIVE-DOMINANCE (ADL) LINEAR MODELS

The AL and ADL models were fitted using the *lmekin* function of the *coxme* package (Therneau, 2012) of R software (R Development Core Team, 2016). The linear mixed model with only additive effects (AL) was defined as follows:

$$y = 1\mu + Za + e, \qquad \text{(Equation 1)}$$

where y is a vector of length n of adjusted phenotypes for fixed effects of sex and contemporary group, ignoring censoring; μ is an intercept; a is the vector of random additive effects of individuals; Z is the incidence matrix for the random effects; and e is the vector of errors assumed as $N(0,\sigma_e^2 I)$. The additive random effect was assumed to be $N(0,G\sigma_a^2)$, where G is the additive genomic relationship matrix given by VanRaden, 2008:

$$G = \frac{WW'}{\sum_{j=1}^{m}\left(2p_j q_j\right)}, \qquad \text{(Equation 2)}$$

where W has the dimension of the number of animals (n) by the number of loci (m), with elements that are equal to $-2p_j$, $1-2p_j$, and $2-2p_j$ for the genotypes mm, Mm, and MM, respectively; p_j is the allelic frequency of M at locus j and $q_j = 1 - p_j$.

The *lmekin* function does not allow for the prediction of two separate random effects. Thus, the linear mixed model with additive and dominance effects (ADL) was defined as follows:

$$y = 1\mu + Zg + e \quad \text{(Equation 3)}$$

where *y, μ, Z,* and *e* were defined in the same way as in model AL presented in Equation 1, and $g = a + d$ is a vector of the total predicted genetic values, which were assumed as $N(0, G\sigma_a^2 + D\sigma_d^2)$, where *a* is the vector of additive genetic values and *d* is the vector of dominance deviations, such that *G* given in Equation 2 and *D* is the dominance genomic relationship matrix:

$$D = \frac{SS'}{\sum_{j=1}^{m}\left(2p_j q_j\right)^2}, \quad \text{(Equation 4)}$$

where *S* has the dimension of the number of animals (*n*) by the number of loci (*m*), with elements that are equal to $-2p_j^2$, $2pq$, and $-2q_j^2$ for the genotypes mm, Mm, and MM, respectively. The parameterization $g = a + d$ implies that, in a population with Hardy-Weinberg equilibrium, the covariance between the additive genetic effects and dominance deviation is zero (Resende et al., 2014). The *G* and *D* matrices were obtained using the GVCBLUP software (Wang and Da, 2014).

From the total genetic values (*g*), the additive genetic values (*a*) and dominance deviations (*d*) can be obtained as follows (Mrode, 2005):

$$\hat{a} = \sigma_a^2 G\Sigma^{-1}\hat{g}, \quad \text{(Equation 5)}$$

$$\hat{d} = \sigma_d^2 D\Sigma^{-1}\hat{g}. \quad \text{(Equation 6)}$$

In models with additive effects, $\hat{g}$ is equal to $\hat{a}$, since the dominance effects are not considered in the model. The AL and ADL models were fitted by the method of maximum likelihood (ML) while initially using the EM algorithm and then switching to the Newton–Raphson algorithm to complete the convergence (Pinheiro and Bates, 2000).

2.1.2 ADDITIVE (ATN) AND ADDITIVE-DOMINANCE (ADTN) TRUNCATED NORMAL MODELS

The ATN and ADTN models were fitted via Gibbs sampling (GS) in the BGLR package of R by using Bayesian Reproducing Kernel Hilbert Spaces (RKHS) regression. This package allows for the fitting of models with censored records, which are considered missing and are sampled from a truncated normal distribution (Pérez and de los Campos, 2014). Fitting of the ATN and ADTN models will be presented by considering the more general model with additive and dominance effects (ADTN), particularly the ATN model (without the dominance effect). Thus, the following linear mixed model was considered:

$$y = 1\mu + Za + Zd + e \qquad \text{(Equation 7)}$$

where μ, Z, and e are defined as in the AL and ADL models, a is the vector of additive genetic values, and d is the vector of dominance deviations. The above model induces the conditional distribution:

$$y|\theta \sim N\left(1\mu + Za + Zd, I\sigma_e^2\right) \qquad \text{(Equation 8)}$$

where θ represents the set of unknown parameters μ, a, d, σ_a^2, σ_d^2, and σ_e^2. For μ, a, and d, the following prior distributions were assumed:

$$p(\mu) \propto \text{constant};\ a|K_1, \sigma_a^2 \sim N(0, K_1\sigma_a^2) \text{ and } d|K_2, \sigma_d^2 \sim N(0, K_2\sigma_d^2) \qquad \text{(Equation 9)}$$

where K_1 and K_2 are the kernel square matrices with a size equal to the number of individuals, and in the ADTN model, K_1 and K_2 were replaced by the additive (G) and dominance (D) genomic relationship matrices (de los Campos et al., 2009). For variance- components σ_a^2, σ_d^2, and σ_e^2 were assumed weak priors. The vector of phenotypes includes censored and uncensored records. Thus, the density function of the conditional distribution in (8) is given by

$$p\left(y|\theta\right) \propto \left(2\pi\sigma_e^2\right)^{-n_u/2} \times \exp\left\{-\frac{1}{2\sigma_e^2}\left[\sum_{i=1}^{n_u}\left(y_{obs_i} - z_i'a - z_i'd\right)^2 \times \prod_{i=n_u+1}^{n}\left[1 - \Phi\left(\frac{t_i - z_i'a - z_i'd}{\sigma_e}\right)\right]\right.\right.,$$

(Equation 10)

where y_{obs} is the dataset with uncensored records of dimension n_u, n is the total number of observations, Φ is the cumulative distribution function of the standard normal, and t_i is the truncation point, which was 150 days in this study.

By using a data augmentation technique for censored observations, the conditional distribution of the observed data in Equation 8 can be rewritten as:

$$p\left(y|\theta\right) \propto \left(2\pi\sigma_e^2\right)^{-n/2} \times exp\left\{-\frac{1}{2\sigma_e^2}\left[\sum_{i=1}^{n_u}\left(y_{obs_i} - x_i'\beta - z_i'a - z_i'd\right)^2 + \sum_{i=n_u+1}^{n}\left(y_{cen_i} - x_i'\beta - z_i'a - z_i'd\right)^2\right]\right\}$$

Equation (11)

where $y_{cen_i} \geq t_i$, i.e., the unobserved value, must be greater or equal to the truncation point. For more details, see Sorensen et al. (1998) and Guo et al. (2001). For parameter estimation, the Gibbs sampler algorithm was used while assuming 120,000 iterations. The first 20,000 iterations were discarded and, to ensure independence, a spacing of 10 between samples was considered. The convergence was verified by the Geweke criterion in the BOA package of R software (Smith, 2007).

2.1.3 ADDITIVE (AC) AND ADDITIVE-DOMINANCE (ADC) COX MODELS

In the context of GS, the additive Cox model (AC) is defined as follows:

$$h(t) = h_0(t)\exp\{Za\}, \qquad \text{(Equation 12)}$$

where $h_0(t)$ is an unspecified nonnegative function of time called the *baseline hazard*, Z is the incidence matrix for random effects, and a is the vector of additive genetic effects, which were assumed with a zero mean and variance-covariance matrix $\Sigma_1 = G\sigma_a^2$, where σ_a^2 is the additive genetic variance and G is the additive genomic relationship matrix given in Equation 2.

When extending the AC model presented in Equation 12, the Cox model with additive and dominance effects (ADC) was defined as follows:

$$h(t) = h_0(t)\exp\{Zg\}, \qquad \text{(Equation 13)}$$

where $h_0(t)$ is defined as in Equation 12 and Z is the incidence matrix of the total genetic effects (g), where $g = a+d$ is a vector of total predicted genetic values, as in Equation 3, with zero for the mean and a variance-covariance matrix of $\Sigma_2 = G\sigma_a^2+D\sigma_d^2$. G and D matrices are given in Equation 2 and 4, respectively. Like the ADL model, the a and d random effects were independent. The individual additive genetic values (a) and dominance deviations (d) were obtained from equations (5) and (6), respectively.

Unlike the linear model, in the Cox model, the intercept is absorbed in the baseline hazard function. In the Cox mixed model fitted with the coxme package, the estimation is based on a penalized partial likelihood that was developed by applying the Laplace approximation to the marginal likelihood function, since the solution of the multidimensional integral is intractable (Pankratz et al., 2005).

2.1.4 ADDITIVE (AUC) AND ADDITIVE-DOMINANCE (ADUC) UNCENSORED COX MODELS

In order to compare the linear and Cox models in equivalent situations (normal data and uncensored), the uncensored Cox models with additive (AUC) and additive-dominance (ADUC) effects were also fitted. The difference between these models when compared to the AC and ADC models is that the observed time for all individuals was regarded as failure time (indicator variable of censoring equal to 1). The AC, AUC, ADC, and ADUC models were fitted by the method of maximum likelihood (ML), which was implemented in the *coxme* package of R (Therneau, 2012). For the Cox model, the *coxme* function provides only ML estimates. Thus, in order to compare the estimates that were obtained in the Cox model without censoring with estimates obtained from the linear model, we used the method ML and not REML for linear models.

2.2 ADDITIVE ($\hat{\mathbf{m}}_{\mathbf{A}}$) AND DOMINANCE ($\hat{\mathbf{m}}_{\mathbf{D}}$) MARKERS EFFECT ESTIMATES

The $\hat{m}_a$ and $\hat{m}_d$ effects were obtained from their respective expressions (Resende et al., 2014):

$$\hat{m}_a = (W'W)^{-1}W'\hat{a}, \qquad \text{(Equation 14)}$$

$$\hat{m}_d = (S'S)^{-1}S'\hat{d}, \qquad \text{(Equation 15)}$$

where $\hat{a}$ and $\hat{d}$ are the vectors of additive genetic values and dominance deviations in each model, and *W* and *S* are the matrices that were previously defined.

2.3 HERITABILITY ESTIMATES

Heritability estimates were obtained from the estimated variance components. The phenotypic variance is given by the following $\sigma_y^2 = \sigma_a^2 + \sigma_d^2 + \sigma_e^2$, where σ_a^2 and σ_d^2 are the additive and dominance genetic variances, respectively, and σ_e^2 is the residual variance. The following estimates of heritability were calculated: additive or narrow sense heritability ($h_a^2 = \frac{\sigma_a^2}{\sigma_y^2}$), heritability of dominance ($h_d^2 = \frac{\sigma_d^2}{\sigma_y^2}$) and total or broad sense heritability ($H^2 = h_a^2 + h_d^2$).

For the additive (AC) and additive-dominance (ADC) Cox models, the term that was relative to the random error was not included in the model because it was incorporated in the *baseline hazard* function (Pankratz et al., 2005). Thus, residual variance cannot be directly obtained. An alternative was proposed by Yazdi et al. (2002) and was used by Schneider et al. (2005) and Santos et al. (2015). Under this approach, the residual variance is replaced by *1*/(*1-c*) where *c* is the proportion of censored data. Therefore, the phenotypic

variance is given by $\sigma_y^2 = \sigma_a^2 + \sigma_d^2 + \frac{1}{1-c}$, and the additive, dominance and total heritabilities are given, respectively, by:

$$h_{ac}^2 = \sigma_a^2 \Big/ (\sigma_a^2 + \sigma_d^2 + 1/(1-c)) \qquad \text{(Equation 16)}$$

$$h_{ad}^2 = \sigma_d^2 \Big/ (\sigma_a^2 + \sigma_d^2 + 1/(1-c)) \qquad \text{(Equation 17)}$$

$$H_c^2 = h_{ac}^2 + h_{dc}^2. \qquad \text{(Equation 18)}$$

For the additive (AUC) and additive-dominance (ADUC) uncensored Cox models, the proportion of censored observations was zero; thus, the residual variance was assumed to be equal to 1 and the phenotypic variance is given by $\sigma_y^2 = \sigma_a^2 + \sigma_d^2 + 1$. This expression is equivalent to random error under a latent scale with a standard normal distribution and probit link function (Resende et al., 2014). For models with only additive effects, the term σ_d^2 was excluded from the phenotypic variation.

2.4 ADDITIVE MODELS *VS* ADDITIVE-DOMINANCE MODELS

In order to verify the superiority of the models with additive and dominance effects over models with only additive effects, the Likelihood Ratio Test (LRT) was used while considering the maximum likelihood estimation method. This test is based on the following: D = (-2*log-likelihood for the restricted model – (-2*log-likelihood for the more general model), which considers the model with only additive effects (H0) versus the model with additive and dominance effects (H1). The statistics of the test follow a Chi-square distribution with degrees of freedom based on the difference between the number of parameters (variance components). The p-value for the test is given by 0,5[1-P($\chi_1^2 \leq D$)] (Ertl et al., 2014). Thus, the *LRT* was used to compare the following models: AL *vs* ADL, AC *vs* ADC, and AUC *vs* ADUC.

For the ATN and ADTN models that were fitted by Gibbs sampling, the goodness-of-fit was verified by the Deviance Information Criterion (*DIC*

$= \overline{D} + P_D$), where $\overline{D}$ is the posterior mean deviance and P_D is the "effective number of parameters". The best models were those with the lowest DIC.

2.5 MODEL COMPARISON

The predicted genomic values in the additive (AC) and additive-dominance (ADC) Cox models are inversely proportional to the observed values and to the predicted genomic values in the additive and additive-dominance linear models. This is because in the Cox model, the predicted values are determined while using the hazard scale, whereas the observed data and the predicted values in the linear models are determined on the observed scale of days (Hou et al. 2009). Thus, the correlation was obtained between the ranks of sorted genomic breeding values (GBVs) in decreasing order for the Cox model, and in ascending order for the observed values and predicted GBVs in the linear models. The same idea was used for the predicted total genomic values (TGVs) with additive-dominance models. In addition, Spearman correlations between the predicted GBVs and TGVs in the eight models were also calculated. To verify whether the estimates differ from zero, Spearman's *rho* was used to estimate a rank-based measure of association, with p-values computed via the asymptotic *t* approximation.

In the Cox models (AUC and ADUC) and in the linear models fitted via ML (AL and ADL), censoring was not considered, i.e., the time of all individuals was failure time. This procedure was used to evaluate the prediction consistency. Similar comparisons were performed for the additive (ATN) and additive-dominance (ADTN) truncated normal model with the additive (AC) and additive-dominance (ADC) Cox model with censored data.

The equivalence between the Cox model and linear model when selecting the best individuals and the largest marker effects was evaluated by means of the agreement rates between the 10% largest predicted GBVs in similar additive models (AL and AUC, ATN and AC) and between the 10% largest predicted TGVs in the additive-dominance models (ADL and ADUC,

ADTN and ADC). The same agreement rates were calculated for the 10% highest additive and dominance marker effects. In linear models, the selected animals were those whose genomic genetic values minimized the time until slaughter. In the Cox model, the selected animals were those whose genomic genetic values maximized the hazard (Hou et al., 2009; Sobczyńska and Blicharski, 2015).

In addition to calculating the agreement rates between models, the *Cohen's kappa* coefficient (*k*) was obtained, which assessed the degree of agreement between two models in the above-described equivalent situations. The statistical significance of *k* was assessed.

3. RESULTS

For ATN and ADTN models fitted via Gibbs sampling, the *p*-values that were obtained for the Geweke criterion were higher than the predetermined significance level of 5%, thereby indicating that convergence was attained for the MCMC chain. For linear (ADL) and Cox (ADC and ADUC) models, the dominance variance was null, and on occasion the dominance heritabilities were also null (Table 2). For the ADTN model fitted via RKHS, the dominance variance was equal to 12.36. When including the dominance effect in this model, the additive genetic variance decreased, and 11.2% of the phenotypic variation was explained by the dominance variation. Su et al. (2012), when evaluating the daily gain in Danish Duroc pigs by means of the linear model, reported dominance values explaining 5.6% of the phenotypic variation.

The additive genetic variances that were estimated in linear models with truncated normal distribution (ATN and ADTN) were higher than the estimated variances in the classical linear models, with an increase of about 192.94% and 124.38% for additive and dominance-additive models, respectively. The same occurred for the dominance variance, which was equal to 0.007 in the ADL model and 12.356 in the ADTN model.

In the linear analysis, broad-sense heritability estimates ranged between 0.08 and 0.25. These values correspond with those obtained by de Hollander et al. (2015) and Sobczyńska and Blicharski (2015). In these studies, the pedigree-based relationship matrix was considered. In addition, the dominance effect was not considered.

The narrow-sense heritability values using the Cox survival model were around 0.02 in the presence and absence of censoring and the dominance effect. Low heritability values (0.05 and 0.08) were also found by Mészáros et al. (2010) in Large White and Landrace pigs when using the Weibull survival model.

Table 2. Estimates of additive genetic variance (σ_a^2), dominance variance (σ_d^2), residual variance (σ_e^2), phenotypic variance (σ_y^2), narrow-sense (or additive) heritability (h_a^2), dominance heritability (h_d^2) and broad-sense (or total) heritability (H^2) for age at slaughter in pigs, considering the Cox and linear model.

		Variance components						
Models	Y	σ_a^2	σ_d^2	σ_e^2	σ_y^2	h_a^2	h_d^2	H^2
AL	NC	9.103	-	108.985	118.088	0.077	-	-
ADL	NC	9.101	0.007	108.981	118.089	0.077	0.000	0.077
ATN	NTC	26.666	-	79.756	106.422	0.247	-	-
ADTN	NTC	20.421	12.356	77.782	110.559	0.182	0.112	0.294
AC	C+B	0.057	-	-	-	0.024[(a)]	-	-
ADC	C+B	0.057	0.000	-	-	0.024[(a)]	0.000[(d)]	0.024
AUC	C+B*	0.018	-	-	-	0.018[(a)]	-	-
ADUC	C+B*	0.018	0.000	-	-	0.018[(a)]	0.000[(d)]	0.018

y = response variable; NC = continuous response variable and without censoring; NTC = continuous response variable with truncated normal distribution for censored data; C+B = continuous response variable + indicator variable of censoring; C+B* = continuous response variable + indicator variable equal to 1; AL = additive linear model; ADL = additive-dominance linear model; ATN = additive truncated normal model; ADTN = additive-dominance truncated normal model; AC = additive Cox model; ADC = additive-dominance Cox model; AUC = additive uncensored Cox model; ADUC = additive-dominance uncensored Cox model; [(a)]narrow-sense (or additive) heritability calculated as in the Equation (16); [(d)]dominance heritability calculated as in the Equation (17).

For all the models evaluated, the correlations were significant at 1% according to Spearman's test (Table 3). In the models with only additive effects, TGVs were equal to GBVs (g = a). The linear models (LA and LA) and uncensored Cox (AUC and ADUC) presented with similar predictive ability

values (0.40). The same occurred for models where censoring was considered (ATN, ADTN, AC, and ADC), with predictive capacity values around 0.20.

In the models that were fitted by maximum likelihood, the correlations between GBVs and between TGVs in the same models, with or without dominance effect, were equal to 1, since the dominance effect in these models was null. For the truncated normal model where the dominance effect was not zero, the correlation between GBVs in the additive and additive-dominant models was equal to 0.97, and between TGVs was equal to 0.88. As previously reported, in the ATN model, TGV is equal to GBV since the dominance effect is not estimated.

Table 3. Spearman's rank correlations (above the diagonal) between total genetic values (TGV, defined as the sum of the genetic effects in the model) and Spearman's rank correlations (below the diagonal) between genomic breeding values (GBV) from different models.

Models	AL	ADL	ATN	ADTN	AC	ADC	AUC	ADUC
AL	0.42	1	0.40	0.29	-0.62	-0.62	-0.93	-0.93
ADL	1	0.42	0.40	0.29	-0.62	-0.62	-0.93	-0.93
ATN	0.40	0.40	0.18	0.88	-0.78	-0.78	-0.35	-0.35
ADTN	0.47	0.47	0.97	0.19	-0.63	-0.64	-0.26	-0.26
AC	-0.62	-0.62	-0.78	-0.84	0.21	1	0.53	0.53
ADC	-0.62	-0.62	-0.78	-0.84	1	0.21	0.53	0.53
AUC	-0.93	-0.93	-0.35	-0.38	0.53	0.53	0.38	1
ADUC	-0.93	-0.93	-0.35	-0.38	0.53	0.53	1	0.38

AL = additive linear model; ADL = additive-dominance linear model; ATN = additive truncated normal model; ADTN = additive-dominance truncated normal model; AC = additive Cox model; ADC = additive-dominance Cox model; AUC = additive uncensored Cox model; ADUC = additive-dominance uncensored Cox model.

Among the different models, the highest Spearman correlation values between TGVs (above the diagonal) and GBVs (below the diagonal) were obtained between the linear model and uncensored Cox model (-0.93). In both models, censoring was not considered. The truncated normal and Cox models, which considered censoring, presented with Spearman correlations equal to -0.78 between GBVs of additive models, as well as -0.84 and -0.64 between GBVs and TGVs of additive-dominant models, respectively.

The lowest correlations were obtained between TGVs and GBVs of the truncated normal model and the uncensored Cox model, with values equal to -0.26 between TGVs, -0.35 between GBVs for ATN, and -0.38 between GBVs for ADTN. The correlation between the linear model and truncated normal model was 0.40 between GBVs in the additive models and increased to 0.47 in the additive-dominant models. Between TGVs, the correlation was 0.29 in the additive-dominant models.

The Spearman correlation values below the diagonal in Table 3 were higher than the values above the diagonal only for the truncated normal model, where the dominance effect was not zero. This was expected, since the values above the diagonal are the correlations between TGVs (sum of additive and dominance effects) and GBVs (only additive effects), while the values below the diagonal are the correlations between predicted GBVs by the additive and additive-dominance models.

The likelihood ratio test (*LRT*) for models fitted via maximum likelihood and *DIC* values for the models fitted via the Bayesian approach are shown in Table 4. According to the *LRT*, for the three evaluated cases, the hypothesis that the additive model is adequate was not rejected (*p*-value greater than 5%), so we can conclude that the model with additive effects only was the best. The same occurred for the truncated normal model, where the model with additive effects presented with a lower DIC and was thus more adequate than the model with additive and dominance effects. Ertl et al. (2014), when evaluating the inclusion of dominance effects in dairy cattle, found that *LRT* was not significant for four of the nine evaluated traits.

The proportion of agreement among the 10% smallest predicted GBVs in the linear model and the 10% largest predicted GBVs in the additive uncensored Cox model was equal to 76% (Table 5), with a kappa equal to 0.74; i.e., of the 34 individuals, 26 were in agreement for both models. For the marker effects, the agreement rate was 58% (14 markers of 24), with a kappa of 0.54. The same agreement rates were observed among TGVs and additive

marker effects in the additive-dominant models, since the dominance effect was null in these models.

Table 4. -2 log-likelihood, χ^2-value and its corresponding p-value of the likelihood ratio test and DIC values for Bayesian models.

Models	*-2logL*	χ-value^2	p-value	*DIC*
AL[a]	2541.214			-
ADL[a]	2541.218	-0.004	0.475	-
ATN[b]	-	-	-	1025.127
ADTN[b]	-	-	-	1032.066
AC[b]	1378.220			-
ADC[b]	1378.221	-0.0006	0.490	-
AUC[a]	3232.948			-
ADUC[a]	3232.950	-0.002	0.482	-

[a]models without taking into account the censoring; [b]models considering censoring; AL = additive linear model; ADL = additive-dominance linear model; ATN = additive truncated normal model; ADTN = additive-dominance truncated normal model; AC = additive Cox model; ADC = additive-dominance Cox model; AUC = additive uncensored Cox model; ADUC = additive-dominance uncensored Cox model. χ^2-value = $-2ln\left(\frac{likelihood\ for\ additive\ models}{likelihood\ for\ additive-dominance\ models}\right)$.

Table 5. Proportion of agreement and *Cohen's kappa* coefficient (in parentheses) between the smallest 10% and largest 10% predicted genetic values using linear and Cox models, respectively, and between the 10% largest additive (AME_{GBV} and AME_{TGV}) and dominance (DME_{TGV}) marker effects for the linear and Cox models

Models	TGV	GBV	AME_{GBV}	DME_{TGV}
Additive				
AL *vs* AUC	-	0.76 (0. 74**)	0.58 (0.54**)	-
ATN *vs* AC	-	0.53 (0.48**)	0.46 (0.40**)	-
Additive-dominance			AME_{TGV}	
ADL *vs* ADUC	0.76 (0. 74**)	0.76 (0.74**)	0.58 (0.54**)	0.71 (0.68**)
ADTN *vs* ADC	0.44 (0.38**)	0.59 (0.54**)	0.54 (0.49**)	0.50 (0.44**)

AL = additive linear model; ADL = additive-dominance linear model; ATN = additive truncated normal model; ADTN = additive-dominance truncated normal model; AC = additive Cox model; ADC = additive-dominance Cox model; AUC = additive uncensored Cox model; ADUC = additive-dominance uncensored Cox model. *Significant at 5%, **significant at 1% level, [ns]not significant.

When considering the censoring, the percentage of agreement between the ATN model and the AC model was 53% between the GBVs and 46% between the largest marker effects, with kappa coefficients of 0.48 and

0.40, respectively. When including the dominance effect, this agreement percentage increased to 59% between GBVs and 54% for the additive marker effects.

When evaluating the hypothesis that the kappa is null, the same was rejected at a significance level of 1% for all cases evaluated, thus showing that there is a direct and significant correlation between the GBVs, TGVs, and marker effects in the comparative models.

4. DISCUSSION

This study aimed to estimate additive and dominance genetic variance components using genomic prediction models for censored data, such as the Cox survival model and the linear model with truncated normal distribution.

For the linear and Cox models that were fitted by using the lmekin and coxme functions of the coxme package of R, the additive and dominance effects were obtained through equations 5 and 6, respectively, since in these models only the total genetic effects (sum of additive and non-additive effects) were predicted. This approach is conceptually allowed because it considers the independence between both predicted effects (Mrode, 2005).

As reported, the truncated normal model fitted via RKHS presented with additive and dominance variance estimates that were much higher than the estimated variances of the linear model fitted via maximum likelihood. A similar result was observed by Morota et al. (2014) in a study on beef cattle, where they also found that additive and dominance variance estimates were overestimated when using parametric kernels. The authors attributed this to the possible lack of orthogonality between additive (G) and dominance (D) genomic relationship kernels, in that a single kernel captures multiple sources of genetic information.

When including the dominance effect in the truncated normal model, the broad sense heritability decreased by approximately 26%. According to

Muñoz et al. (2014), this result is expected due to the proportion of dominance variance that can be explained by the additive genetic variance.

In the additive and additive-dominance uncensored Cox models, the residual variance was equal to 1, in that the random error in latent scale had a standard normal distribution with the probit link function. By accounting for censoring, Yazdi et al. (2002) showed that the error variance can be estimated by the expression 1/(1-c) in the Cox model, where c represents the proportion of censored data. In this study, c was equal to 0.56. A summary of the estimations used for heritability in the survival analysis was presented by Resende et al. (2014), who showed that the expression that was used here and was proposed by Yazdi et al. (2002) is the best estimator, as it is valid for the Cox and Weibull models.

There are few studies that include dominance effects in the usual analysis as well as in genomic selection. Moreover, no reports were found in the literature that considered the estimation of the two variance components (additive and dominance) when using survival models. The Survival Kit software (Ducrocq et al., 2010) permits the estimation of only one component at a time and only allows for the use of the pedigree-based relationship matrix, which makes it impossible to predict genetic values using the genomic relationship matrix. Therefore, comparing both approaches (classical and Bayesian) while using the genomic relationship matrix in the Cox model is not yet possible.

In genomic selection, the correlation between the predicted genomic values and the corrected phenotypes is called of the predictive ability, which is used to assess the quality of fit of the models in the estimation population (Resende et al., 2014). When including the dominance effect, the predictive abilities of the GBVs in the evaluated models did not increase. Similar results were obtained by Ertl et al. (2014) in dairy cattle. According to Nishio and Satoh (2014), this was due to the low and null dominance effects that were obtained in the evaluated models. De Almeida Filho et al. (2016) obtained the highest

accuracies in additive-dominant models in relation to additive models only with regards to simulated traits with great dominance effects.

According to Muñoz et al. (2014), additive effects can capture a large proportion of the genetic variance of non-additive effects, such as dominance, due to a possible lack of independence between these effects in improved populations.

The predictive ability values were lowest (around 0.20) in the models where censoring was considered (ATN, ADTN, AC, and ADC) versus in the models where it was not considered (LA, LAD, AUC, and ADUC). This occurred because of the greater distance between the predicted genetic values and the adjusted phenotypes of the censored times. Although they presented with the smallest predictive abilities, these models are conceptually more appropriate, since they express the correlations with the phenotypes in the latent scale (Santos et al., 2015).

The uncensored Cox models (AUC and ADUC) presented with the highest Spearman correlations in the AL and ADL models in comparison to the ATN, ADTN, AC, and ADC models. This reveals the influence of censoring, since in the AUC and ADUC models, all data were uncensored.

Negative correlations between the linear and Cox models are to be expected, since the random effects in both models are inversely proportional. In the linear model, the fit is performed directly on the time variable, while in the Cox model, the fit is based on the hazard of the animal with regards to obtaining the desired weight at slaughter; i.e., the best animals in the linear models were those with the lowest predicted TGVs and GBVs, since they reached the desired slaughter weight in a shorter period of time. In the Cox models, the best animals were those with the greatest predicted TGVs and GBVs; thus, the risk function grows rapidly, thereby indicating that the animal's weight also grows quickly (Hou et al., 2009; Giolo and Demétrio, 2011; Santos et al., 2015).

The agreement percentages between predicted genomic values and marker effects were lowest in the models where censoring was considered. In addition to censoring, another factor that played a role in the occurrence was the difference in model fits. The ATN and ADTN models were fitted via the Bayesian approach using RKHS, while the AC and ADC models were fitted by the maximum likelihood method.

In addition, as has been reported, the estimates that were obtained in the ADTN model were biased because of a lack of orthogonality among additive and dominance genomic relationship kernels (Morota et al., 2014).

In addition to the predictive ability, Cohen's kappa coefficient has also been employed for evaluating genomic models (Ornella et al., 2014).

Contrary to what has been reported by Onteru et al. (2010), where the implementation of survival models in GS is not yet possible, this study showed that for the method where the pedigree-based relationship matrix is replaced by a marker-based relationship matrix (GBLUP and GBLUP-D in linear models), fitting the Cox model can be conducted relatively simply, without the need for a large computational demand. Furthermore, the fit has allowed for the estimation of non-additive effects, such as the dominance effect.

In addition to the Cox model, other models such as the linear model with truncated normal distribution have been proposed for censored data in GS (Pérez and De Los Campos, 2014), thereby expanding the analytical possibilities in additive and non-additive models. However, as mentioned above, the approach that was used for the truncated normal model while using the RKHS method overestimated the additive and dominance genetic variance estimates, thus indicating the need for further investigation of this method.

5. CONCLUSION

This paper showed for the first time and through real data analyses the possibility of performing genomic prediction of censored phenotypes using the Cox survival model with additive and dominance effects. In general, for the assessed trait in the study population, the dominance variance was null, but on occasion the augmentation of predictive abilities did not include the dominance effect in the assessed models.

ACKNOWLEDGMENTS

The first author would like to thank the CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) for a Sandwich Doctorate scholarship (grant no. BEX 9415/14-9). This study was supported by the CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and the FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais).

REFERENCES

Azevedo CF, Resende MDV, Silva FF, Viana JMS, et al. (2015). Ridge, Lasso and Bayesian additive-dominance genomic models. *BMC Genet.* 16: 1.

Band GO, Guimarães SEF, Lopes PS, Peixoto JDO, et al. (2005). Relationship between the Porcine Stress Syndrome gene and carcass and performance traits in F2 pigs resulting from divergent crosses. *Genet. Mol. Biol*. 28: 92-96.

Costa EV, Diniz DB, Veroneze R, Resende MD, et al. (2015). Estimating additive and dominance variances for complex traits in pigs combining genomic and pedigree information. *Genet. Mol. Res.* 14: 6303-6311.

Cox DR (1972). Regression models and life tables (with ssion). *J. R. Stat. Soc. Series B Stat. Methodol.* 34: 187-220.

de Almeida Filho JEA, Guimarães JFR, Silva FF, Resende MDV, et al. (2016). The contribution of dominance to phenotype prediction in a pine breeding and simulated population. *Heredity*. 117: 33-41.

de Hollander CA, Knol EF, Heuven HCM and van Grevenhof EM (2015). Interval from last insemination to culling: II. Culling reasons from practise and the correlation with longevity. *Livest. Sci*. 181: 25-30.

de Los Campos G, Gianola D and Rosa GJM (2009). Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *J. Anim. Sci*. 87: 1883-1887.

Ducrocq V, Sölkner J and Mészáros G (2010). Survival Kit v6-a Software Package for Survival Analysis (ID232). Proceedings of the 9th World Congress of Genetic and Applied Livestock Production, Leipzig, 232.

Ertl J, Legarra A, Vitezica ZG, Varona L, et al. (2014). Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. *Genet. Sel. Evol.* 46: 40.

Giolo SR and Demétrio CGB (2011). A frailty modeling approach for parental effects in animal breeding. *J. Appl. Stat*. 38: 619-629.

Guo SF, Gianola D, Rekaya R and Short T (2001). Bayesian analysis of lifetime performance and prolificacy in Landrace sows using a linear mixed model with censoring. *Livest. Prod. Sci.* 72: 243-252.

Hou Y, Madsen P, Labouriau R, Zhang Y, et al. (2009). Genetic analysis of days from calving to first insemination and days open in Danish Holsteins using different models and censoring scenarios. *J. Dairy Sci*. 92: 1229-39.

Kärkkäinen HP and Sillanpaa MJ (2013). Fast genomic predictions via Bayesian G-BLUP and multilocus models of threshold traits including censored Gaussian data. *G3 (Bethesda*) 3: 1511-1523.

Mészáros G, Palos J, Ducrocq V and Solkner J (2010). Heritability of longevity in Large White and Landrace sows using continuous time and grouped data models. *Genet. Sel. Evol*. 42: 13.

Morota G, Boddhireddy P, Vukasinovic N, Gianola D, et al. (2014). Kernel-based variance component estimation and whole-genome prediction of pre-corrected phenotypes and progeny tests for dairy cow health traits. *Front. Genet*. 5: 56.

Mrode RA (2005). Linear models for the prediction of animal breeding values. CAB International, Wallingford.

Muñoz PR, Resende MF, Gezan SA, Resende MDV, et al. (2014). Unraveling additive from non-additive effects using genomic relationship matrices. *Genetics*. 198: 1759-1768.

Nishio M and Satoh M. (2014). Including dominance effects in the genomic BLUP method for genomic evaluation. *PLoS One*. 9: e85792.

Onteru SK, Fan B, Nikkilä MT, Garrick DJ, et al. (2011). Whole-genome association analyses for lifetime reproductive traits in the pig. *J. Anim. Sci*. 89: 988-995.

Ornella L, Pérez P, Tapia E, González-Camacho JM, et al. (2014). Genomic-enabled prediction with classification algorithms. *Heredity*. 112: 616-626.

Pankratz VS, Andrade M and Therneau TM (2005). Random-effects Cox proportional hazards model: general variance components methods for time-to-event data. *Genet. Epidemiol.* 28: 97-109.

Pérez P and De Los Campos G (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics*. 198: 483-495.

Pinheiro JC and Bates DM (2000). Mixed-Effects Models in S and S-PLUS. Springer-Verlag, New York.

R Development Core Team (2016). R: A Language and Environment for Statistical Computing. Available at http://www.R-project.org. Accessed March 16, 2016.

Resende MDV, Silva FF and Azevedo CF (2014). Estatística matemática, biométrica e computacional: modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência. Editora Suprema, Viçosa.

Santos VS, Martins Filho S, Resende MDV, Azevedo CF, et al. (2015). Genomic selection for slaughter age in pigs using the Cox frailty model. *Genet. Mol. Res* 14: 12616-12627.

Schaeffer L (2013). Survival. In: History of genetic evaluation methods in dairy cattle (Grosu H, Schaeffer L, Oltenacu PA, et al., eds.) 279-298. https://xa.yimg.com/kq/groups/18395782/1926111600/name/FINAL_BOOK_29.04.2013.pdf, Accessed 12 April, 2016

Schneider MDP, Strandberg E, Ducrocq V and Roth A (2005). Survival analysis applied to genetic evaluation for female fertility in dairy cattle. *J. Dairy Sci*. 88: 2253-2259.

Serenius T, Stalder KJ, and Puonti M (2006). Impact of dominance effects on sow longevity. *J. Anim. Breed. Genet.* 123: 355-361.

Silva FF, Resende MDV, Rocha GS, Duarte DA, et al. (2013). Genomic growth curves of an outbred pig population. *Genet. Mol. Biol*. 36: 520-527.

Smith BJ (2007). boa: An R Package for MCMC Output Convergence Assessment and Posterior Inference. *J. Stat. Softw.* 21: 1-37.

Sobczyńska M and Blicharski T (2015). Phenotypic and genetic variation in longevity of Polish Landrace sows. *J. Anim. Breed. Genet.* 132: 318-327.

Sorensen DA, Gianola D and Korsgaard IR (1998). Bayesian mixed-effects model analysis of a censored normal distribution with animal breeding applications. *Acta Agric. Scand. A Anim. Sci.* 48: 222-229.

Su G, Christensen OF, Ostersen T, Henryon M, et al. (2012). Estimating additive and non-additive genetic variances and prediction genetic merits using genome-wide dense single nucleotide polymorphism markers. *PLoS One*. 7: e45293.

Therneau T (2012). Mixed effects Cox models. R package version 2.2-3. http://cran.r-project.org/web/packages/coxme/vignettes/coxme.pdf. Accessed April 12, 2016.

VanRaden PM (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci*. 91: 4414-4423.

Verardo LL, Silva FF, Varona L, Resende MDV, et al. (2015). Bayesian GWAS and network analysis revealed new candidate genes for number of teats in pigs. *J. Appl. Genet.* 56: 123-132.

Wang C and Da Y (2014). Quantitative genetics model as the unifying model for defining genomic relationship and inbreeding coefficient. *PLoS One.* 9: e114484.

Yazdi MH, Visscher PM, Ducrocq V and Thompson R (2002). Heritability, reliability of genetic evaluations and response to selection in proportional hazard models. *J. Dairy Sci.* 85: 1563-1577.

CHAPTER III

GENETIC PARAMETERS FOR CENSORED TRAITS COMBINING MARKERS AND PEDIGREE IN ADDITIVE AND DOMINANCE MODELS

ABSTRACT

This study aimed to estimate the additive and dominance genetic variation based on a linear model of a mouse data set and on the Cox survival model of a censored pig data set, while considering the pedigree and markers information in both the models. For normal data (mouse data), the variance components were estimated for three growth traits, including 1884 individuals and 10,946 markers. Phenotypic data were initially precorrected by fixed effects (i.e., gender, season, and month). These data were analyzed by the GBLUP-D method, fitted via restricted maximum likelihood using the *lmekin*) function of the R software and the Bayesian GBLUP-D method, fitted via Gibbs Sampler in the BGLR-R Package. For the censored data of pigs, the response variable was time (in days) from birth to slaughter. The animals who did not reach the desired slaughter weight (of approximately 65 kg) were considered to be censored. All animals were genotyped to 384 SNP markers, of which a total of 237 SNPs remained for 345 animals after quality control. To these data, the truncated normal distribution model and the Cox model were fitted with additive, dominance, and polygenic effects. For the frequentist models fitted by the functions *lmekin* and *coxme*, the inclusion of the dominance and polygenic effects in the additive models was tested by the Likelihood Ratio Test (LRT). For the Bayesian models, the DIC criterion was considered and the model with the lowest value was selected. For the mouse data, in the 3 evaluated traits, both the dominance and polygenic effects were found to be significant by the LRT method. For the censored data of pigs, both the models indicated statistical significance of the polygenic effect, with high percent of additive genetic variance explained by the pedigree. This study is a pioneer in

predicting additive and dominance genomic genetic values by using the Cox survival model and by combining pedigree and marker information; hereby promising to further explore the range of statistical models that can be applied to genomic selection.

1. INTRODUCTION

Since long, the method commonly used for the selection of the best individuals, especially in animal breeding and perennial plant species, was based on the REML/BLUP procedure (estimation of variance components by restricted maximum likelihood - REML and prediction of random effects by BLUP - best linear unbiased predictor), which were based on phenotypic information and the kinship matrix (GODDARD; HAYES, 2007).

However, with the development of molecular markers and their reduced cost, Meuwissen, Hayes, and Goddard (2001) developed the genomic wide selection (GWS), a method that uses, beyond the phenotype, the information obtained directly from the DNA of individuals, thereby improving the accuracy of the selection.

Several methods have been employed in GWS, of which Genomic BLUP (GBLUP) stands out owing to its ease of implementation. This method consists of replacing, in the mixed linear model, the pedigree-based relationship matrix by a relationship matrix estimated by markers (VANRADEN, 2008). The superiority of GBLUP in relation to the traditional BLUP has been confirmed in several animal (DAETWYLER et al.; 2010; GARRICK, 2011; GONZALEZ-RECIO et al., 2008; WOLC et al., 2011) as well as plant species (CROSSA et al., 2010; HESLOT et al., 2012; RESENDE et al., 2012; SPINDEL et al., 2015).

Some studies have also shown the superiority of models that simultaneously consider pedigree and markers information on those based only on the pedigree or marker (BURGUEÑO et al., 2012; CROSSA et al., 2013; CROSSA et al., 2014). According to de los Campos et al. (2013), the

benefits of jointly modeling pedigree and marker data relative to a markers-only model tend to decrease as marker density increases. According to Solberg et al. (2009), when a sparse marker map is used, the inclusion of a polygenic effect can capture genetic variation that are not intimately linked to markers.

In the usual genetic evaluation, a few studies have evaluated the inclusion of non-additive effects in the models, such as dominance. This is mainly due to the greater computational complexity and the low evidence of non-additive genetic variance in the literature. Nevertheless, past studies have shown that, in traits with low heritability, high proportion of dominance variance and high percent of full sibs, the prediction of dominance effects increased the selection gain by approximately 10% (TORO; VARONA, 2010; VARONA MIZTAL, 1999).

Recently, the dominance effects were included in genomic selection of pigs (COSTA et al., 2015; LOPES et al., 2016; NISHIO; SATOH, 2014; SU et al., 2012), dairy cattle (ERTL et al., 2014; SUN et al., 2014; WINTTENBURG; MELZER REINSCH, 2014), maize (Technow et al., 2012), and pinus (DE ALMEIDA FILHO et al., 2016). The method initially proposed consideration of the additive and dominance effects in genomic selection (denominated GBLUP-D), as an extension of GBLUP, which uses the additive and dominance relationship matrices based on markers instead of the relationship matrices based on the pedigree. In this context, Santos et al. (2017) reported an alternative to the performance of the GBLUP-D method when using the lmekin function of the package coxme (THERNEAU, 2012) of R software (R DEVELOPMENT CORE TEAM, 2016), which obtained identical results to the GLUP-D method available in the GVCBLUP software (WANG et al., 2014).

Genomic Selection is widely used for phenotypic traits, with full observations and normal distribution. However, statistical software's such as the R-package BGLR (PÉREZ; DE LOS CAMPOS, 2014) allows genomic prediction for binary and censored traits. Santos et al. (2015, 2016) were

pioneers in performing genomic predictions for censored traits using the Cox model with the additive and additive–dominant effects, respectively.

This study aimed to estimate the additive and dominance genetic variation based on the linear model to a mouse data set and based on the Cox survival model to a censored pig data set, considering the pedigree and markers information in both the models.

2. MATERIAL AND METHODS

Two data sets were analyzed: one referring to 3 growth traits in mice and the other referring to the age at the slaughter of pigs. For the mouse data, two approaches of the GBLUP-D method: frequentist model fitted by the *lmekin* function and the Bayesian model fitted by using the BGLR package, were compared. In both the approaches, the inclusion of the polygenic effect was evaluated in the additive and additive–dominant genomic models.

The same procedure was performed for the age data at the slaughter of pigs, albeit considering the presence of censored observations, since not all individuals achieved the desirable slaughter weight. In this case, the Cox survival model fitted with the *coxme* function of the R (THERNEAU, 2012) and the truncated normal model fitted in the BGLR package were used. In both the models, the additive and dominance effects were estimated based on the markers and pedigree information. The data sets and models used are detailed in the following sections.

2.1 MICE DATA SET

The dataset was composed of 1884 mice with 10,946 single nucleotide polymorphisms (SNP) markers. This data set was selected owing to the ease of access to phenotypic and genotypic information (available at http://gscan.well.ox.ac.uk/) and also for its use in other studies (LEGARRA et al., 2008; DE LOS CAMPOS et al., 2009; OKUT et al., 2011; VITEZICA et al.,

2013; PÉREZ ; DE LOS CAMPOS, 2014). Three traits related to the growth of the animals were evaluated: GS, BL, and BMI, as these were the phenotypes precorrected for their fixed effects of sex, season, and month. Genealogical information was composed only by parents, totaling 2,272 individuals organized in 168 families of full-sibs. Further details of the description of these data are available in Valdar et al. (2006).

The details of models with the additive, dominance, and polygenic effects, implemented via the *lmekin* function, proposed in this work, and in the BGLR package, are given below.

2.1.1 LINEAR MODEL WITH ADDITIVE, DOMINANCE AND POLYGENIC EFFECTS (L_ADP) FITTED VIA *LMEKIN* FUNCTION

The L_ADP model was implemented using the *lmekin* function of the *coxme* R-package (Therneau, 2012) by means of the following expression:

$$y = 1\mu + Zg + e, \quad (1)$$

where, y is the vector of phenotypes, μ is the general mean, $g = a + d + p$ is a vector of total predicted genetic values, which are given by the sum of additive (a), dominance (d) and polygenic (p) genetic effects, assumed with the normal distribution with the mean 0 and of variance-covariance matrix $\Sigma = G\sigma_a^2 + D\sigma_d^2 + A\sigma_p^2$, in which σ_a^2 and σ_p^2 are the additive genetic variances associated to marker and the pedigree, respectively, and σ_d^2 is the variance due dominance deviation. G and D are the additive and dominance genomic relationship matrices, of dimension $n \times n$, respectively, and A is the pedigree-based relationship matrix of dimension $n \times n$. Assuming independence between the effects, the individual additive genetic values (a), of dominance deviation (d) and polygenic (p) can be obtained separately as from the total genetic values (g) by the respective expressions (MRODE, 2005):

$$\hat{a} = \sigma_a^2 G\Sigma^{-1}\hat{g}\,, \quad \hat{d} = \sigma_d^2 D\Sigma^{-1}\hat{g} \quad \text{and} \quad \hat{p} = \sigma_p^2 A\Sigma^{-1}\hat{g}\;; \quad (2)$$

where, $a \sim N(0, G\sigma_a^2)$, $d \sim N(0, D\sigma_d^2)$ and $p \sim N(0, A\sigma_p^2)$. The vector of errors (*e*) also showed normal distribution with a zero average and variance–covariance matrix equal to $I\sigma_e^2$, where *I* is an identity matrix and σ_e^2 is the residual variance.

The additive (G) and dominance (D) genomic relationship matrices are given below, respectively, by:

$$G = \frac{MM'}{2\sum_{j=1}^{m} p_j \left(1 - p_j\right)} \tag{3}$$

and

$$D = \frac{SS'}{\sum_{j=1}^{m} \left\{2p_j \left(1 - p_j\right)\right\}^2}, \tag{4}$$

where, the m_{ij} values of the *M* matrix are equal to 0-2p, 1-2p, and 2-2p for the genotypes mm, Mm, and MM, respectively; and s_{ij} values of matrix *S* are equal to $-2p^2$, $2p(1-p)$, and $-2(1-p)^2$ for the genotypes mm, Mm, and MM, respectively; and p_j is the allelic frequency of M at locus *j* (RESENDE et al., 2014; VITEZICA; VARONA; LEGARRA, 2013). These matrices were obtained by using the GVCBLUP software (WANG et al., 2014), which were created for the genomic prediction of phenotypes normally distributed after considering the additive and dominance effects.

The additive relationship matrix based on the pedigree was obtained by using the *kinship2* package (SINNWELL; THERNEAU; SCHAID, 2014) of R, where the elements a_{ij} of matrix A represent the expected proportion of the genome that are shared by each pair of individuals. The diagonal values of the matrix are equal to 1, and off the diagonal, are equal to 0 for the pairs of unrelated individuals and $\left(\frac{1}{2}\right)^r$ for the related pairs, where *r* denotes the degree of kinship between the 2 individuals *i* and *j*.

The linear model with additive and dominance effects (L_AD) was adjusted in the same way as the L_ADP, that is, by excluding the pedigree-

based relationship matrix. Therefore, the total genetic value (*g*) can be given by *g* = *a* + *d*, where $g \sim N(0, \Sigma)$ and $\Sigma = G\sigma_a^2 + D\sigma_d^2$. In the linear model with additive effects only, the total genetic values (*g*) were equal to the additive genetic effects (*a*).

2.1.2 LINEAR MODEL WITH ADDITIVE, DOMINANCE AND POLYGENIC EFFECTS (BL_ADP) FITTED VIA GIBBS SAMPLER

The BL_ADP model was adjusted via the Gibbs sampler in the BGLR package of the R software using the Bayesian reproducing kernel Hilbert spaces regressions, RKHS. In this context, the BLUP and GBLUP methods can be considered as particularities of the RKHS method, since the parametric kernel matrix *K* can be computed from the pedigree and/or markers (DE LOS CAMPOS; GIANOLA; ROSA, 2009).

Thus, the Bayesian linear model with additive, dominance and polygenic effects (BL_ADP) is defined as follows (PÉREZ; DE LOS CAMPOS, 2014):

$$y = 1\mu + Za + Wd + Tp + e, \quad (5)$$

where *y*, *μ*, and *e* are defined as for the L_ADP and L_AD models, *a* is the vector of individual additive genetic values, *d* is the vector of dominance deviations and *p* is the vector of polygenic effects. Z, W and T are incidence matrices. The conditional distribution of the data is normal multivariate, as given below:

$$y|\theta \sim N\left(1\mu + Za + Wd + Tp, I\sigma_e^2\right) \quad (6)$$

where, θ represents the set of unknown parameters *μ*, *a*, *d*, *p*, σ_a^2, σ_d^2, and σ_e^2. For *μ*, *a*, *d*, and *p* were assumed the following *a priori* distributions:

$p(\mu) \propto$ constant; $a|K_1, \sigma_a^2 \sim N\left(0, K_1\sigma_a^2\right)$, $d|K_2, \sigma_d^2 \sim N\left(0, K_2\sigma_d^2\right)$ and

$$p|K_3, \sigma_p^2 \sim N\left(0, K_3\sigma_p^2\right) \quad (7)$$

where K_1, K_2, and K_3 are the Kernel matrices with dimension equal to the number of individuals. In the BL_ADP model, K_1 and K_2 were replaced by additive (G) and dominance (D) genomic relationship matrices and K_3 by the additive relationship matrix based on the pedigree. For the variance components σ_a^2, σ_d^2, and σ_e^2 assumed the following *a priori* distributions, as given below:

$$\sigma_i^2 \sim \chi^{-2}\left(\sigma_i^2 \left| df_i, S_i \right.\right); \tag{8}$$

with degrees of freedom df_i and scale parameters $S_i > 0$. Thus, the joint *a priori* distribution of θ is given by the following:

$$p\left(\theta \left| df_e, S_e, df_a, S_a, df_d, S_d, df_p, S_p \right.\right) \propto N\left(a \left| 0, K_1\sigma_a^2 \right.\right) \times N\left(d \left| 0, K_2\sigma_d^2 \right.\right) \times N\left(p \left| 0, K_3\sigma_p^2 \right.\right)$$

$$\times \chi^{-2}\left(\sigma_a^2 \left| df_a, S_a \right.\right) \times \chi^{-2}\left(\sigma_d^2 \left| df_d, S_d \right.\right) \times \chi^{-2}\left(\sigma_p^2 \left| df_p, S_p \right.\right) \times \chi^{-2}\left(\sigma_e^2 \left| df_e, S_e \right.\right). \tag{9}$$

The Bayesian estimation maximizes *a posteriori* distribution and is proportional to the product of the likelihood stated in (6) by the *a priori* distribution stated in (9). For the model with additive and dominance effects (BL_AD), the process is identical, not considering the polygenic effects, p = 0.

Heritability estimates were obtained from the estimated variance components. Considering the full model, with additive, dominance and polygenic effects, the following heritabilities estimates were calculated: additive or narrow sense heritability (h_a^2), dominance heritability (h_d^2) and total or broad sense heritability (H^2), given, respectively, by the following:

$$h_a^2 = \left[(\sigma_a^2 + \sigma_p^2) / \sigma_y^2 \right];$$

$$h_d^2 = \sigma_d^2 / \sigma_y^2$$

and

$$H^2 = h_a^2 + h_d^2;$$

where σ_y^2 is the phenotypic variance given by $\sigma_y^2 = \sigma_a^2 + \sigma_p^2 + \sigma_d^2 + \sigma_e^2$, in which σ_a^2, σ_p^2, σ_d^2, and σ_e^2 have been already explained.

In summary, for each one of 3 variables, the following models were considered, as shown in the table below.

Table 2.1 – Linear models fitted for three variables of mice growth, considering additive, dominance and polygenic effects.

Models	Estimated effects	Fit type
L_ADP	Additive + Dominance + Polygenic	REML – *lmekin* function of R software
L_AD	Additive + Dominance	
L_AP	Additive + Polygenic	
L_A	Additive	
BL_ADP	Additive + Dominance + Polygenic	Gibbs Sampler – BGLR R-package
BL_AD	Additive + Dominance	
BL_AP	Additive + Polygenic	
BL_A	Additive	

2.2 AGE AT THE TIME OF SLAUGHTER OF PIGS

In order to identify individuals with rapid weight gain for slaughter, the time (in days) from birth until slaughter of the animals was considered as a variable response. As not all animals reached the desired slaughter weight [fixed at 65 kg for this population (BAND et al., 2005)], a binary censoring variable was created, where animals weighing ≥65 kg received value 1 and animals weighing <65 kg were considered censored and received the value 0. The proportion of censoring was around 0.561, that is, approximately 44% of the animals weighed at least 65 kg.

The phenotypic data arising from the F2 population generated by crossing 11 boars and 54 dams were randomly selected from the F1 generation, which was initially created by crossing 2 native Brazilian Piau boars with 18 commercial sows (Landrace x Large White x Pietrain). The

experiment was conducted at the Universidade Federal de Viçosa and the use of these animals was reviewed and approved by the Bioethics committee of the Department of Veterinary Medicine (DVT-UFV) in agreement with the Guide to the Care and Use of Experimental Animals of the Canadian Council on Animal Care.

The 345 animals were genotyped for 384 SNPs, and, after the quality control of these markers, 66 SNPs were discarded because of a low-genotyping call rate (<0.95) and 81 were discarded due to a minor allele frequency (MAF) <0.05, such that 237 markers remained in the analysis. Despite the small number and sparse distribution of the SNPs markers, they were distributed in previously identified QTL regions, thus ensuring appropriate coverage of the relevant genome regions in this population. For this trait, the truncated normal distribution model and the COX model were fitted, with additive, dominance, and polygenic effects, as presented below.

2.2.1 TRUNCATED NORMAL MODEL WITH ADDITIVE, DOMINANCE AND POLYGENIC EFFECTS (TN_ADP)

The TN_ADP model was also fitted in the BGLR package using the technique of data augmentation (TANNER; WONG, 1987). Censored data in BGLR is described using 3 vectors $\{t_i, y_i, u_i\}$, satisfying the condition: $t_i < y_i < u_i$, where y_i is the vector of observed phenotypes (e.g., a time-to event variable, observable only in un-censored data points, otherwise missing, NA) and t_i and u_i are defined as lower and upper-bounds for the response, respectively. For right censoring, the configuration of the triplet is $\{t_i, NA, Inf\}$, where, the censored values de y_i are sampled from truncated normal densities (PÉREZ; DE LOS CAMPOS, 2014). The TN_ADP model is defined as that for uncensored normal data [given in (5)]. However, the conditional density of the data is given by the following equation:

$$p(y|\theta) \propto \left(2\pi\sigma_e^2\right)^{-m/2} \times \exp\left\{-\frac{1}{2\sigma_e^2}\left[\sum_{i=1}^{m}\left(y_{obs_i} - x_i'\beta - z_i'a - z_i'd - z_i'p\right)^2\right.\right.$$

$$\left.\left. \times \prod_{i=m+1}^{n}\left[1-\Phi\left(\frac{t_i - x_i'\beta - z_i'a - z_i'd - z_i'p}{\sigma_e}\right)\right]\right\}, \qquad (10)$$

where, y_{obs} is the dataset with uncensored records of dimension n_u, n is the total number of observations, Φ is the cumulative distribution function of the standard normal, and t_i is the truncation point, which was 150 days in this study.

For the censored data, Taner and Wong (1987) proposed the technique of data augmentation, in which the conditional distribution of the observed data given in (10) can be rewritten as follows:

$$p(y|\theta) \propto \left(2\pi\sigma_e^2\right)^{-n/2} \times \exp\left\{-\frac{1}{2\sigma_e^2}\left[\sum_{i=1}^{m}\left(y_{obs_i} - x_i'\beta - z_i'a - z_i'd - z_i'p\right)^2\right.\right.$$

$$\left.\left. + \sum_{i=m+1}^{n}\left(y_{cen_i} - x_i'\beta - z_i'a - z_i'p\right)^2\right]\right\}, \qquad (11)$$

where, $y_{cen_i} \geq t_i$, i.e., the unobserved value, must be greater or equal to the truncation point. For more details, see Sorensen, Gianola and Korsgaard (1998) and Guo et al. (2001). Models without dominance and/or polygenic effects were considered as particular cases of the full model.

For Bayesian models, the Gibbs sampler algorithm was used while assuming 120,000 iterations. The first 20,000 iterations were discarded and, to ensure independence, a spacing of 10 between samples was considered. The convergence was verified by the Geweke criterion in the BOA package of R software (SMITH, 2007).

2.2.2 COX MODEL WITH ADDITIVE, DOMINANCE AND POLYGENIC EFFECTS (C_ADP).

The Cox fragility model considering the three genetic effects (additive, dominance and polygenic) was defined as follows:

$$h(t) = h_0(t)\exp\{Zg\}, \quad (12)$$

where $h_0(t)$ is the unspecified baseline hazard function, g is the vector of total genetic effects, given by the sum of additive, dominance and polygenic effects, and Z is the incidence matrix. The values of g are assumed to be normal with a mean value of 0 and covariances matrix $\Sigma = G\sigma_a^2 + D\sigma_d^2 + A\sigma_p^2$, in which G, D, A, σ_a^2, σ_d^2 and σ_p^2 have been already explained.

Similar to the linear model adjusted by the *lmekin* function, the additive, dominance, and polygenic genetic effects can be obtained separately according to the expression given in equation (2). The Cox models were fitted using the *coxme* function of the R software, which is based on partial penalized likelihood to estimate the random effects.

In the Cox model, heritability estimates were not obtained directly, since the error is not included in the model, which made it impossible to estimate the residual variance. In this study, an approximation was used for the calculation of heritability, as proposed by Yazdi et al. (2002) for survival models, which consisted of replacing the residual variance by $1/(1-c)$, where c is the proportion of censored data. Thus, the approximate additive and dominance heritabilities were given by the following:

$$h_{ac}^2 = \left[(\sigma_a^2 + \sigma_p^2)/\sigma_y^2\right]; \quad (13)$$

$$h_{dc}^2 = \sigma_d^2/\sigma_y^2 \quad \text{and} \quad (14)$$

$$H_c^2 = h_{ac}^2 + h_{dc}^2; \quad (15)$$

where σ_y^2 is the phenotypic variance given by $\sigma_y^2 = \sigma_a^2 + \sigma_p^2 + \sigma_d^2 + \left[1/(1-c)\right]$, in which σ_a^2, σ_p^2 and σ_d^2 have been already explained.

For the truncated normal distribution model, the heritabilities were calculated in the same way as for the linear models. In Table 2.2, the fitted models for the variable age at the time of slaughter are presented, considering the additive, dominance, and polygenic effects.

Table 2.2 – Truncated normal distribution model and Cox model fitted for slaughter age on pigs considering additive, dominance and polygenic effects.

Models	Estimated effects	Fit type
TN_ADP	Additive + Dominance + Polygenic	Truncated normal distribution model fitted via GS in the BGLR R-package
TN_AD	Additive + Dominance	
TN_AP	Additive + Polygenic	
TN_A	Additive	
C_ADP	Additive + Dominance + Polygenic	Cox model fitted via *coxme* function of R
C_AD	Additive + Dominance	
C_AP	Additive + Polygenic	
C_A	Additive	

GS: Gibbs Sampler.

2.3 COMPARISON OF MODELS

For the linear and Cox frailty models fitted by using the *lmekin* and *coxme* functions of R, respectively, the statistical significance of the polygenic and dominance effects was verified by the Likelihood Ratio Test (LRT), considering the null hypothesis H_0: model with only additive effects and the alternative hypotheses $H_{1\,(DP)}$: model with additive, dominance, and polygenic effects; $H_{1\,(D)}$: model with additive and dominance effects and $H_{1\,(P)}$: model with additive and polygenic effects. The test statistics are given, respectively, by the following equations:

$$LRT_{DP} = -2ln\left(\frac{likelihood\ of\ additive\ model}{likelihood\ of\ additive\ +dominance+polygenic\ \ model}\right) \quad (16)$$

$$LRT_{D} = -2ln\left(\frac{likelihood\ of\ additive\ model}{likelihood\ of\ additive\ +dominance\ \ model}\right) \quad (17)$$

and

$$LRT_{P} = -2ln\left(\frac{likelihood\ of\ additive\ model}{likelihood\ of\ additive\ +polygenic\ \ model}\right). \quad (18)$$

The distribution of the LRT values for the test that *k* variance components are zero involves a mixture of χ^2 variates from 0 to *k* degrees of freedom (ZHANG; LIN, 2008; ERTL et al., 2014; GILMOUR et al., 2015). The 5% significance level was used as a threshold.

For the models adjusted via the Gibbs sampler in the BGLR package, the inclusion of the dominance and polygenic effects was assessed by using the DIC (Deviance Information Criterion) criterion, given by the following equation

$$DIC = \bar{D} + P_D, \tag{18}$$

where, $\bar{D}$ is the posterior mean deviance and P_D is the effective number of parameters. The best models were those with the lowest DIC values (SPIEGELHALTER et al., 2002).

For the mouse data, the equivalence between the linear models adjusted via *lmekin* and BLGR was calculated using Pearson correlations between the total genetic values estimated in the different models. For the censored data of slaughter age on pigs, the Spearman correlation was calculated between the total genetic values estimated in the truncated normal distribution model and the Cox fragility model.

3. RESULTS AND DISCUSSION

3.1 DATA WITH NORMAL DISTRIBUTION (MOUSE DATA)

The chain convergence was verified for all models adjusted via the Gibbs sampler (i.e., BL_A, BL_AD, BL_AP, BL_ADP), for the 3 evaluated traits (i.e., GS, BMI, and BL), where the p-values obtained by the Geweke criterion were higher than the predetermined significance level of 5%, indicating that the Gibbs chains did converge for all models.

The statistics and the p-values of the LRT for the variance components estimated by the *lmekin* function and the DIC values for the Bayesian models are shown in Table 3.1. The inclusion of both the dominance and polygenic

effects in the additive models was initially tested and the 3 evaluated characteristics showed the significance of these effects, with p-values of the LRT being <5% for the models adjusted via *lmekin* and lower DIC values in the model with the additive, dominance, and polygenic effects (L_ADP) than in the model with only the additive effects (L_A).

The separate inclusion of the polygenic and dominance effects it was also verified. In both the approaches (i.e., *lmekin* and BGLR), the inclusion only of the dominance effect in the additive model resulted in statistical significance (p-value of the LRT <0.05 by using *lmekin*) and DIC value of the L_AD model lower than the DIC value of the L_A model), for the 3 evaluated traits. The inclusion only of the polygenic effect in the additive genomic model was also statistically significant (p-value of LRT being <5% by using *lmekin* function, and DIC value of the L_AP model being lower than the DIC value of the L_A model).

Table 3.1 - Values of -2*log likelihood, χ^2 and their corresponding p-values of the likelihood ratio test (LRT) for the models adjusted using the *lmekin* function and the DIC values for the Bayesian models in 3 traits of mice.

Traits	Models	*lmekin* function			BGLR
		-2logL	χ^2-value	p-value	DIC
GS	L_A	-3076.27			-3100.191
	L_AD	-3081.96	5.69	0.009	-3110.422
	L_AP	-3104.30	28.03	0.000	-3154.293
	L_ADP	-3104.29	28.02	0.000	-3151.778
BMI	L_A	-1026.687			-1057.921
	L_AD	-1029.874	3.187	0.037	-1067.381
	L_AP	-1040.18	13.493	0.000	-1093.799
	L_ADP	-1041.062	14.375	0.000	-1100.283
BL	L_A	2681.678			2613.077
	L_AD	2678.876	2.80	0.047	2612.618
	L_AP	2649.374	32.304	0.000	2535.566
	L_ADP	2649.384	32.294	0.000	2539.828

L_A: linear model with additive effects; L_AD: Linear model with additive and dominance effects; L_AP: Linear model with additive and polygenic effects; L_ADP: Linear model with additive, dominance, and polygenic effects; GS: Growth speed; BMI: Body mass index; BL: Body length.

In the Bayesian estimation, among all evaluated models, the L_AP (additive and polygenic effects) model presented with the lowest DIC values for the traits GS and BL. For the trait BMI, the selected model was the L_ADP, with additive, dominance and polygenic effects, with lower DIC value.

The correlations between total genetic values (TGV) predicted by the different models for the three traits evaluated are presented in Table 3.2. The correlations between the predicted values and the corrected phenotypes increased around 18%-23% by including the polygenic effect in both approaches (*lmekin* and BGLR) for the three evaluated traits. For example, for trait BMI, the correlation of the L_AD model that was 0.57 increased to 0.69 (increase of 21%) by including the polygenic effect on the model. The same occurred in the Bayesian estimation, in which the correlation that was 0.62 in the BL_AD model increased to 0.73 in the BL_ADP model, with an increase of 18%.

Table 3.2 - Correlations between the total genetic values (TGV, defined as the sum of the genetic effects in the model) of different models and correlations (in the diagonals) between TGV and corrected phenotypes for the traits: growth speed (GS), body mass index (BMI) and body length (BL) in mice.

Traits	Models	L_AD	L_ADP	BL_AD	BL_ADP
GS	L_AD	0.56	0.82	0.99	0.92
	L_ADP		0.67	0.83	0.96
	BL_AD			0.60	0.94
	BL_ADP				0.72
BMI	L_AD	0.57	0.90	0.99	0.93
	L_ADP		0.69	0.91	0.99
	BL_AD			0.62	0.95
	BL_ADP				0.73
BL	L_AD	0.61	0.91	0.99	0.92
	L_ADP		0.75	0.92	0.99
	BL_AD			0.65	0.94
	BL_ADP				0.78

L_AD: Linear model with additive and dominance effects adjusted via *lmekin*; L_ADP: Linear model with additive, dominance and polygenic effects adjusted via *lmekin*; BL_AD: Bayesian linear model with additive and dominance effects adjusted in the BGLR package; BL_ADP: Bayesian linear model with additive, dominance and polygenic effects adjusted via BGLR.

In genomic selection, the correlation between the observed and predicted values is called as the predictive ability, which can be considered as the measure of the goodness-of-fit between the data and the model, when validation is not performed (RESENDE et al., 2014).

High correlations between the total genetic values predicted in the frequentist (*lmekin*) and Bayesian models were observed, with correlations of 0.99 between the L_AD and BL_AD models in the 3 evaluated traits. The same occurred among the L_ADP and BL_ADP models for the BMI and BL traits. This almost perfect correlation between the total genetic values predicted by the Bayesian method (BGLR) and that proposed by using the *lmekin* function indicates that the models lead to a similar ranking of the individuals. This result demonstrates that the genomic model considering 3 random effects (i.e., additive, dominance, and polygenic effects) was implemented correctly by using the *lmekin* function of the R software, thus being an efficient alternative in the genomic prediction considering these effects.

When considering the polygenic effect in the additive–dominant models, adjusted via *lmekin* and BGLR package, highest correlations occurred between the total genetic values predicted in the Bayesian models (BGLR) in comparison to the models adjusted via function *lmekin* in the 3 evaluated traits. For example, in the trait GS, the correlation between BL_AD and BL_ADP was 0.94, while the correlation between L_AD and L_ADP was 0.82. For BL, this difference was lower, with the correlation being equal to 0.91 between the L_AD and L_ADP models and being 0.94 between the BL_AD and BL_ADP models.

The estimates of variance components based on the selected models, with dominance and polygenic effects, for the 3 studied traits are presented in Table 3.3. In order to compare the inclusion of the polygenic effect in the models, the estimates were presented without the inclusion of this effect. For the model with additive and dominance effects, in the 3 evaluated traits, the dominance heritabilities presented values different of zero, thus indicating the presence of dominance effect on the growth of mice. This result corroborates

those obtained by the LRT (Table 3.1), in which the null hypothesis of a non-significant dominance effect was rejected for the 3 evaluated traits.

The genetic effect of dominance is defined as the interaction between alleles of the same locus (SU et al., 2012; FALCONER MACKAY, 1996). In previous studies, a considerable increase in the predictive accuracy was observed by including the dominance effect (DA et al., 2014; MUNÕZ et al., 2014; NISHIO SATOH, 2014). However, some studies have also shown that the inclusion of this effect did not increase the predictive accuracy for some traits (ERTL et al., 2014; SUN et al., 2014; SANTOS et al., 2016).

Table 3.3 - Estimates of additive genetic variance (σ_a^2), dominance variance (σ_d^2), residual variance (σ_e^2), phenotypic variance (σ_y^2), narrow-sense (or additive) heritability (h_a^2), dominance heritability (h_d^2) and broad-sense (or total) heritability (H^2) for three mice traits.

Parameters	GS		BMI		BL	
Ad+Dom	*Lmekin*	BGLR	*Lmekin*	BGLR	*Lmekin*	BGLR
σ_a^2	0.001	0.001	0.003	0.003	0.038	0.038
σ_d^2	0.001	0.001	0.001	0.002	0.009	0.018
σ_e^2	0.010	0.010	0.031	0.03	0.209	0.205
σ_y^2	0.01	0.012	0.035	0.035	0.256	0.261
h_a^2	0.045	0.071	0.075	0.091	0.150	0.146
h_d^2	0.045	0.067	0.034	0.066	0.035	0.068
H^2	0.091	0.138	0.110	0.157	0.184	0.214
Ad+Dom+Pol	*Lmekin*	BGLR	*Lmekin*	BGLR	*Lmekin*	BGLR
σ_{a*}^2	0.002	0.002	0.005	0.005	0.07	0.07
σ_d^2	0.000	0.001	0.001	0.002	0.000	0.011
σ_e^2	0.01	0.009	0.029	0.028	0.186	0.181
σ_y^2	0.012	0.010	0.034	0.032	0.256	0.262
h_a^2	0.151	0.164	0.142	0.167	0.272	0.266
h_d^2	0.002	0.051	0.017	0.052	0.000	0.042
H^2	0.153	0.209	0.16	0.219	0.272	0.308

Ad + Dom: Model with additive and dominance effects; Ad + Dom + Pol: Model with additive, dominance and polygenic effects; GS: Growth speed; BMI: Body mass index; BL: Body length. $\sigma_{a*}^2 = \sigma_a^2 + \sigma_P^2$; σ_a^2 and σ_p^2 are the genetic variances associated to the markers and pedigree, respectively.

Considering several levels of dominance in a simulated data set, De Almeida Filho et al. (2016) observed that the additive–dominant models presented with the highest accuracies only for simulated traits with a wide effect of dominance. The authors therefore concluded that the inclusion of the dominance effect, in genomic selection, depends on the genetic architecture of the trait in a specific population.

For the variable GS, the dominance heritability was found to be equal to additive heritability when considering the linear model adjusted via the *lmekin* function. In the Bayesian model, the dominance heritability estimate was lower, but extremely close to additive heritability. Greater difference between the additive and dominance heritabilities was observed for the trait BL, in which additive heritability was 4-times higher than dominance heritability.

By including the polygenic effect in the additive–dominant models, the additive heritabilities increased considerably in the 3 evaluated traits, which highlights the variable GS, in which the heritability was 3-times higher by including the polygenic effect in the model fitted by the *lmekin* function. In the Bayesian model fitted by the BGLR package, this increase was 2-times higher (0.071 for BL_AD and 0.164 for BL_ADP).

For the variable BMI, the increase in additive heritability by including the pedigree information was 89% in the model fitted by *lmekin* and 84% in the model fitted by BGLR. For the variable BL, the increase was 81% and 82%, respectively. Costa et al. (2015), by including the polygenic effect in the additive–dominant model, found an increase of 56% and 28% in additive heritabilities of the traits birth weight and weight at 21 days in pigs, respectively.

The additive heritabilities estimated by the frequentist (*lmekin*) and Bayesian (BGLR) models were close for the 3 evaluated traits in both the models, with the additive and dominance effects and with the additive, dominance, and polygenic effects. However, in both the evaluated situations,

dominance heritabilities were approximately higher in the Bayesian models than in the models fitted via *lmekin* function, and this increase was more pronounced in models with additive, dominance, and polygenic effects (Table 3.3).

Santos et al. (2017) analyzed the simulated data set provided by Vitezica et al. (2013) and found overestimated dominance heritabilities in models with additive and dominance effects fitted by BGLR. Morota et al. (2014) also found the same results compared to those obtained by Su et al. (2012). The authors cite that one of the possible causes of this overestimation is the lack of orthogonality among the additive and dominance genomic relationship matrices.

Santos et al. (2017) compared the obtained results by the analysis proposed using the *lmekin* function with those obtained by the GVCBLUP usual software (WANG et al., 2014) and verified the possibility and effectiveness of predicting additive and dominance effects through the *lmekin* function. Expanding the work of Santos et al. (2017), this study depicts the possibility of predicting additive and dominance effects while considering the pedigree information in genomic prediction. In addition, the *coxme* R package, in which the *lmekin* function is implemented, also allows predicting of the genomic genetic values for censored traits by using the *coxme* function, as performed by Santos et al. (2015, 2016), based on the Cox fragility model.

3.2 SWINE CENSORED DATA

The values of -2log-likelihood, χ^2 and their corresponding p-values of the LRT for the Cox model, and the DIC values of the truncated normal models are presented in Table 3.4. For the truncated normal models with additive and dominance effects (TN_AD) and with the additive, dominance, and polygenic effects (TN_ADP), the Gibbs chains converged for both models by using the Geweke's criterion.

Table 3.4 - Values of -2log-likelihood, χ^2 and their corresponding p-values of the likelihood ratio test considering the Cox model fitted by the *coxme* function, and the DIC values of the truncated normal models fitted in the BGLR package.

Models	*lmekin* function			BGLR
	-2logL	χ^2-value	P-value	DIC
TN_A	-	-	-	1025.127
TN_AD	-	-	-	1032.066
TN_AP	-	-	-	1022.438
TN_ADP	-	-	-	1025.517
C_A	1378.220			
C_AD	1378.221	-0.001	0.487	-
C_AP	1373.050	5.170	0.011	-
C_ADP	1373.054	5.166	0.049	-

TN_A = truncated normal model with additive effects; TN_AD = truncated normal model with additive and dominance effects; TN_AP = truncated normal model with additive and polygenic effects; TN_ADP = truncated normal model truncated with additive, dominance and polygenic effects; C_A = Cox model with additive effects; C_AD = Cox model with additive and dominance effects; C_AP = Cox model with additive and polygenic effects; C_ADP = Cox model with additive, dominance and polygenic effects.

Similar to normal mouse data, it was initially tested the joint inclusion of dominance and polygenic effects in both the additive models (i.e., Cox and truncated normal). For the Cox model, both dominance and polygenic effects were significant, with p-value of LRT equal to 4,9%. For the truncated normal distribution model, both the effects were not significant, with a DIC value of TN_ADP (1025.517) being higher than the DIC value of TN_A (1025.127).

It was also evaluated that the inclusion of the dominance and polygenic effects separately in both Cox and normal truncated models. The inclusion of the dominance effect in both the additive models was not significant, with p-value of LRT equal to 48.7% for the Cox model and DIC value of the TN_AD model (1032.066) being higher than the DIC value of the TN_A model (1025.127). The inclusion of the polygenic effect in both the additive models showed statistical significance, with p-value of LRT being equal to 1.1% for the Cox model and DIC value of the TN_AP model (1022.438) being lower than the DIC value of the TN_A model (1025.127).

Among the 4 models evaluated in the Bayesian fit, the truncated normal model with the additive and polygenic effects (TN_AP) presented with

lower DIC value. Although the LRT indicates statistical significance when including both the dominance and polygenic effects in the additive Cox model, it was decided to consider the Cox model with additive and polygenic effects (C_AP), since the genetic variance of dominance was null in both the models C_ADP and C_AD (data not shown).

The dominance effects on the genomic models were evaluated for various carcass and growth traits in pigs, with different results. The first study to consider the dominance effect in the genomic selection of pigs was by Su et al. (2012), who analyzed the daily gain trait and noted a substantial contribution of the dominance variance to the total genetic variation. Guo et al. (2016), when analyzing backfat thickness and average daily gain in the animals, observed that, the inclusion of dominance effects in the models resulted in more accurate genomic genetic values and less bias of prediction. Xiang et al. (2016) showed that the inclusion of dominance did not increase the predictive ability for the total number of piglets born. Costa et al. (2015), who analyzed 15 growth and carcass traits in the same pig population used in the present study, noted significant dominance effects only for the carcass traits.

As already mentioned, de Almeida Filho et al. (2016) reported, using simulated data, that the inclusion of the dominance effect in genomic models depends on the genetic architecture of each studied trait. Other factors such as large environmental variance, low relationship among individuals, predominance of additive effects, and confounding with other effects, such as common environment or maternal effect, may result in the lack of power to estimate the dominance variance (BOLORMAA et al., 2015).

Spearman correlations between the predicted genomic genetic values in the Cox and normal truncated models, with additive and polygenic effects, are presented in Table 3.5. The correlation between the predicted genomic genetic values and the corrected phenotypes was higher in the Cox model with the additive and polygenic effects.

Table 3.5 - Spearman correlations between genomic genetic values (VGG, defined as the sum of genetic effects in the model) of truncated normal and Cox models and Spearman correlations (in the diagonals) between VGG and corrected phenotypes for trait age at the time of slaughter of pigs.

Models	TN_A	TN_AP	C_A	C_AP
TN_A	0.17	0.79	-0.78	-0.57
TN_AP		0.23	-0.58	-0.82
C_A			0.21	0.58
C_AP				0.37

TN_A = additive truncated normal model; TN_AP = additive-polygenic truncated normal model; C_A = additive Cox model; C_AP = additive-polygenic Cox model.

The TN_A, TN_AP, and C_A models presented predictive abilities equal to 0.17, 0.23, and 0.21, respectively. These low predictive ability values occur due to the presence of censoring, thus reflecting a greater distance between the predicted genomic genetic values and the corrected phenotypes for the censored times. By not considering censoring, Santos et al. 2016 noted predictive ability values of approximately 0.53 for the additive–dominant Cox model and 0.42 for the additive–dominant linear model.

By including the polygenic effect in the Cox model and in the truncated normal distribution model, the predictive abilities of the models increased; this increase was more pronounced in the Cox model, where the correlation in the C_A model was 0.21 increased to 0.37 in the C_AP model. The correlation between the additive model and the additive–polygenic model was higher in the truncated normal model (0.79) than in the Cox survival model (0.58), thereby indicating that the predicted genomic values in these models were closer to each other than to those in the Cox models.

The correlation between the TN_A and C_A models was equal to -0.78. By including the polygenic effect on the model, this correlation increased to -0.82. In dairy cattle, Hou et al. (2009) noted correlations of -0.83 and-0.90 between the linear model similar to those used in the present study and the Cox model, for 2 fertility traits in more than 450,000 cows. Negative correlation values between the linear and Cox survival models were expected because

the random effects predicted in the Cox model were inversely proportional to those obtained in the linear models. The Cox model fit the risk of the animal obtaining the desired weight at the time of the slaughter, that is, the best animals were considered to be those with the greatest predicted genomic values, occasioning a rapid growth in the hazard function, which indicates that the animal's weight also grows rapidly. In linear models, the animals with lowest genomic values were selected because they reached the desirable weight of slaughter in less time (HOU et al., 2009; GIOLO DEMÉTRIO, 2011; SANTOS et al., 2015; SANTOS et al., 2016).

The estimates of variance components considering the truncated normal distribution model and the Cox model for the trait age at the time of slaughter in pigs are presented in Table 3.6. In order to compare the inclusion of the polygenic effect in the models, the estimates without and with the inclusion of this effect were presented, considering the models based on the markers. An increase of approximately 93% in additive heritability was noted by including the polygenic effect in the truncated normal distribution model. In the Cox model, this increase was even more significative, passing from an additive heritability of 0.02 using only markers information to 0.19 when considering information of markers and pedigree.

Table 3.6 - Estimates of genetic variance associated to markers (σ_a^2) and pedigree (σ_p^2), residual variance (σ_e^2) and additive heritability (h_a^2) for slaughter age in pigs, considering the Cox model and the truncated normal distribution model, respectively.

Models	Variance components			
	σ_a^2	σ_p^2	σ_e^2	h_a^2
TN_A	26.895	--	80.090	0.248
TN_AP	19.413	36.010	57.982	0.479
C_A	0.057	--	-	0.024[(a)]
C_AP	0.027	0.502	-	0.188[(a)]

TN_A = additive truncated normal model; TN_AP = additive-polygenic truncated normal model; C_A = additive Cox model; C_AP = additive-polygenic Cox model. [(a)]additive heritability calculated as in equation (13).

In the truncated normal model, the percent of total genetic variance explained by the markers was approximately 35% and the remainder (65%)

was explained by the pedigree. In the Cox model, this proportion of the total genetic variance explained by the pedigree was even higher, that is, approximately 95%. One of the possible causes for a high proportion of the genetic variance explained by the pedigree is the small number of markers used.

In this study, the genomic models were fitted considering only 237 markers. Presently, the literature presents studies in which more than 10000 markers were used. For instance, Crossa et al. (2010) commented that it is reasonable, as the number of markers increases, that the contribution due to the pedigree information decreases. Some other studies (DAETWYLER et al., 2010; SAATCHI et al., 2011) have shown that factors such as the size of the training data set and problems with the phenotypes as a mixture of breeds and the relatively small family size may limit the potential accuracy that can be obtained with genomic selection in comparison with the models based on pedigree (DE LOS CAMPOS et al., 2013).

Previous studies have proved that the combined use of markers and pedigree information has improved the prediction of the models only in situations with low density of markers. In the presence of a large number and high density of markers, these same studies did not find that the inclusion of the pedigree improve the predictions (CALUS; VEERKAMP, 2007; CROSSA et al., 2010; VAZQUEZ et al.; 2010).

According to Almeida Filho (2016), the inclusion of pedigree in genomic models is an alternative to decreasing the cost in genomic selection, since the cost of genotyping depends on the number of markers. Thus, the authors claim that, instead of genotyping only a few individuals with high-density markers, it is better to genotype several individuals with low-density markers and include the information of the pedigree in the genomic model, because the accuracy tends to increase with the numbers of sampled cases.

In the truncated normal distribution model, the residual variance decreased by approximately 27.6% when the polygenic effect was included in

the model. According to Crossa et al. (2010), the residual variance (σ_e^2) can be used as a criterion to assess model goodness-of-fit, since it gives an indication of how much of the phenotypic variance is not explained by the model.

In survival models, such as Cox model, residual variance is not estimated, since the model does not include the error component. As a result, several heritability estimators considering the Cox and Weibull survival models were proposed (DUCROCQ, 1999; KORSGAARD et al., 1999; YAZDI et al., 2002). According to Resende et al. (2014), the expression used (equations 13 and 14) and proposed by Yazdi et al. (2002) is the most suitable one for obtaining an approximate heritability for the Cox model, and it is also valid for the Weibull model.

In the context of genomic selection, other softwares such as GS3 (LEGARRA; RICARD; FILANGI, 2011), ASReml (GILMOUR et al., 2015), and DMU (Madsen et al., 2014) also allow prediction of the genomic genetic values considering additive and non-additive effects, based on the markers and pedigree information. In the R software, one of the first implementations of the mixed models in the genomic era was using the *rrblup* package (ENDELMAN, 2011). However, this package estimates only the additive effects. Based on the REML+BLUP procedure, this work acts as a pioneer in the use of a function available in the R software (*lmekin*) to predict the genomic genetic values while considering simultaneously the additive, non-additive, and polygenic effects.

This work also demonstrated, for the first time, the possibility of predicting additive and dominance genomic genetic values by using the Cox survival model and by combining pedigree and marker information. Until date, it is known that this analysis is only allowed through the *coxme* function of R, as shown in the present work. The *Survival Kit* software, widely used in animal breeding, which adjusts the Cox and Weibull models, performs genetic evaluation of censored phenotypes based only on the pedigree relationship matrix and does not simultaneously estimate the additive and non-additive effects, such as dominance. Thus, this work demonstrates an efficient

alternative in the genomic selection era that allows the prediction of censored phenotypes based on survival models which best describes data of this nature.

REFERENCES

ALMEIDA FILHO, J. E. de. **Genomic prediction of additive and non-additive effects in a pine breeding and simulated population**. 2016. 107 f. Tese (Doutorado em Genética e Melhoramento), Universidade Federal de Viçosa - UFV, Viçosa, 2016.

AZEVEDO, C. F. et al. Ridge, Lasso and Bayesian additive-dominance genomic models. **BMC genetics**, v. 16, n. 1, p. 105-118, 2015.

BAND, G. O. et al. Relationship between the Porcine Stress Syndrome gene and pork quality traits of F2 pigs resulting from divergent crosses. **Genetics and Molecular Biology**, v. 28, n. 1, p. 88-91, 2005.

BOLORMAA, S. et al. Non-additive genetic variation in growth, carcass and fertility traits of beef cattle. **Genetics Selection Evolution**, v. 47, n. 1, p. 26, 2015.

CALUS, M. P. L.; VEERKAMP, R. F. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. **Journal of Animal Breeding and Genetics**, Austin, v. 124, n. 6, p. 362-368, 2007.

COSTA, E. V. et al. Estimating additive and dominance variances for complex traits in pigs combining genomic and pedigree information. **Genetics and Molecular Research**, v. 14, n. 2, p. 6303-6311, 2015.

CROSSA, J. et al. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. **Genetics,** Austin, v. 186, p. 713-724, 2010.

CROSSA, J. et al. Genomic prediction in maize breeding populations with genotyping-by-sequencing. **G3: Genes, Genomes, Genetics**, p. g3. 113.008227, 2013.

CROSSA, J. et al. Genomic prediction in CIMMYT maize and wheat breeding programs. **Heredity**, v. 112, n. 1, p. 48, 2014.

DA, Y. et al. Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. **PloS one**, v. 9, n. 1, p. e87666, 2014.

DAETWYLER, H. D. et al. Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. **Animal Production Science**, v. 50, n. 12, p. 1004-1010, 2010.

DE ALMEIDA FILHO, J. E. et al. The contribution of dominance to phenotype prediction in a pine breeding and simulated population. **Heredity**, v. 117, n. 1, p. 33, 2016.

DE LOS CAMPOS, G.; GIANOLA, D.; ROSA, G. J. M. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. **Journal of Animal Science**, v. 87, n. 6, p. 1883-1887, 2009.

DE LOS CAMPOS, G. et al. Whole-genome regression and prediction methods applied to plant and animal breeding. **Genetics**, v. 193, n. 2, p. 327-345, 2013.

DUCROCQ V., SÖLKNER J., MÉSZÁROS G. *Survival Kit v6* – a software package for survival analysis. In: Proc. **9th World Congress on Genetics Applied to Livestock Production**. Leipzig, Germany, 2010.

ERTL, J. et al. Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. **Genetics Selection Evolution**, v. 46, n. 1, p. 40-49, 2014.

FALCONER, D. F.; MACKAY, T. F. C. **Variance. In Introduction and Quantitative Genetic**. Malaysia: Longman, p.122-143, 1996.

GARRICK, D. J. The nature, scope and impact of genomic prediction in beef cattle in the United States. **Genetics Selection Evolution**, v. 43, n. 1, p. 17, 2011.

GEWEKE, J. Evaluating the accuracy of sampling based approaches to the calculation of posterior moments. In: BERNARDO, J. O.; BERGER, J. M.; DAWID, A. P.; SMITH, A. F. M. (Ed.). **Bayesian statistics 4**. Oxford: Oxford University Press, 1992. p. 169–194.

GILMOUR, A. R. et al. **ASReml user guide release 4.1 structural specification**. 2015. Disponível em <https://asreml.kb.vsni.co.uk/wp-content/uploads/sites/3/2018/02/ASReml-4.1-Structural-Specification.pdf>. Acesso em 15 fev. 2017.

GIOLO, S. R.; DEMÉTRIO, C. G. B. A frailty modeling approach for parental effects in animal breeding. **Journal of Applied Statistics**, v. 38, n. 3, p. 619 – 629, 2011.

GODDARD, M. E.; HAYES, B. J. Genomic selection. **Journal of Animal Breeding and Genetics**, Austin, v. 124, p. 323-330, 2007.

GONZÁLEZ-RECIO, O. et al. Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. **Genetics Selection Evolution**, v. 41, n. 1, p. 3, 2009.

GUO, S. F. et al. Bayesian analysis of lifetime performance and prolificacy in Landrace sows using a linear mixed model with censoring. **Livestock Production Science**, v. 72, n. 3, p. 243-252, 2001.

HESLOT, N. et al. Genomic selection in plant breeding: a comparison of models. **Crop Science**, Madison, v.52, n.1, p.146-160, 2012.

HOU, Y. et al. Genetic analysis of days from calving to first insemination and days open in Danish Holsteins using different models and censoring scenarios. **Journal Dairy Science**, Champaign, v. 92, n. 3, p. 1229 – 1239, 2009.

LEGARRA, A. et al. Performance of genomic selection in mice. **Genetics**, v. 180, n. 1, p. 611-618, 2008.

LEGARRA, A.; RICARD, A.; FILANGI, O. **GS3: Genomic Selection, Gibbs Sampling, Gauss-Seidel (and BayesCpi)**. 2011. Disponível em <http://genoweb.toulouse.inra.fr/~alegarra/> Acesso em: 27 mar. 2015.

LOPES, M. S. et al. Genomic prediction of growth in pigs based on a model including additive and dominance effects. **Journal of Animal Breeding and Genetics**, v. 133, n. 3, p. 180-186, 2016.

MADSEN, P. et al. DMU-a package for analyzing multivariate mixed models in quantitative genetics and genomics. In: **Proceedings of the 10th world congress of genetics applied to livestock production**. 2014. p. 18-22.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome wide dense marker maps. **Genetics**, Austin, v. 157, p. 1819-1829, 2001.

MOROTA, G. et al. Kernel-based variance component estimation and whole-genome prediction of pre-corrected phenotypes and progeny tests for dairy cow health traits. **Front. Genet.**, v. 5, n. 56, doi: 10.3389/fgene.2014.00056, 2014.

MRODE, R. A. **Linear models for the prediction of animal breeding values**. Wallingfors: CAB international, 2005.

MUÑOZ, P. R. et al. Unraveling Additive from Nonadditive Effects Using Genomic Relationship Matrices. **Genetics**, Austin, v. 198, n. 4, p. 1759-1768, 2014.

NISHIO, M.; SATOH, M. Including dominance effects in the genomic BLUP method for genomic evaluation. **PloS one**, v. 9, n. 1, e85792, 2014.

OKUT, H. et al. Prediction of body mass index in mice using dense molecular markers and a regularized neural network. **Genetics research**, v. 93, n. 3, p. 189-201, 2011.

PÉREZ, P.; DE LOS CAMPOS, G. Genome-wide regression and prediction with the BGLR statistical package. **Genetics**, Austin, v. 198, n. 2, p. 483-495, 2014.

R DEVELOPMENT CORE TEAM (2016). **R: a language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, Version 3.2.4. Available at http://www.R-project.org (Accessed 16 March 2016).

RESENDE, M. F. R. et al. Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). **Genetics**, v. 190, n. 4, p. 1503-1510, 2012.

RESENDE, M. D. V.; SILVA, F. F.; AZEVEDO, C. F. **Estatística matemática, biométrica e computacional: modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência**. Viçosa: Editora Suprema, 2014. 881 p.

SANTOS, V. S. et al. Proposal of genomic blup with additive and dominance effects in the environment R. **Biometric Brazilian Journal**, 2017. No prelo.

SANTOS, V. S. et al. Genomic prediction for additive and dominance effects of censored traits in pigs. **Genetics and molecular research**, v. 15, n. 4, 2016.

SANTOS, V. S. et al. Genomic selection for slaughter age in pigs using the Cox frailty model. **Genetics and Molecular Research**, v. 14, n. 4, p. 12616-12627, 2015.

SAATCHI, M. et al. Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. **Genetics Selection Evolution**, v. 43, n. 1, p. 40, 2011.

SCHAEFFER, L. **Survival**. In: GROSU, H.; SCHAEFFER, L.; OLTENACU, P. A.; et al. History of genetic evaluation methods in dairy cattle, 2013, p. 279-298. Available at

https://xa.yimg.com/kq/groups/18395782/1926111600/name/FINAL_BOOK_29.04.2013.pdf (Accessed 12 April 2016).

SINNWELL, J.P.; THERNEAU, T.M.; SCHAID, D.J. The kinship2 R package for pedigree data. **Human heredity**, v. 78, n. 2, p. 91-93, 2014.

SMITH, B. J. boa: An R Package for MCMC Output Convergence Assessment and Posterior Inference. **Journal of Statistical Software**, v. 21, n. 11, p. 1-37, 2007.

SORENSEN, D. A.; GIANOLA, D.; KORSGAARD, I. R. Bayesian mixed-effects model analysis of a censored normal distribution with animal breeding applications. **Acta Agriculturae Scandinavica A-Animal Sciences**, v. 48, n. 4, p. 222-229, 1998.

SPIEGELHALTER, D. J. et al. Bayesian measures of model complexity and fit (with discussion). **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v.64, n.4, p.583-639, 2002.

SPINDEL, J. et al. Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. **PLoS genetics**, v. 11, n. 2, p. e1004982, 2015.

SU, G. et al. Estimating additive and non-additive genetic variances and prediction genetic merits using genome-wide dense single nucleotide polymorphism markers. **PLoS One**, v. 7, n. 9, p. e45293, 2012.

SUN, C. T. et al. Improvement of prediction ability for genomic selection of dairy cattle by including dominance effects. **PloS one**, v. 9, n. 8, p. e103934, 2014.

TANNER, M. A.; WONG, W. H. The calculation of posterior distributions by data augmentation. **Journal of the American statistical Association**, v. 82, n. 398, p. 528-540, 1987.

TECHNOW, F. et al. Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. **Theoretical and Applied Genetics**, v. 125, p.1181-1194, 2012.

THERNEAU, T. coxme: **Mixed Effects Cox Models** (2012). R package version 2.2-3. Available at http://CRAN.R-project.org/package=coxme (Accessed 12 April 2016).

TORO M. A.; VARONA L. A note on mate allocation for dominance handling in genomic selection. **Genetics Selection Evolution**, v. 42, n. 33, p. 1-9, 2010.

VALDAR, W. et al. Genome-wide genetic association of complex traits in heterogeneous stock mice. **Nature genetics**, v. 38, n. 8, p. 879-887, 2006.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, Champaign, v. 91, n. 11, p. 4414 – 4423, 2008.

VARONA, L.; MISZTAL, I. Prediction of parental dominance combinations for planned matings, methodology, and simulation results. **Journal of dairy science**, Champaign, v. 82, n. 10, p. 2186-2191, 1999.

VAZQUEZ, A. I. et al. Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. **Journal of dairy science**, Champaign, v. 93, n. 12, p. 5942-5949, 2010.

VITEZICA, Z. G.; VARONA, L.; LEGARRA, A. On the additive and dominant variance and covariance of individuals within the genomic selection scope. **Genetics**, Austin, v. 195, n. 4, p. 1223-1230, 2013.

WANG, C et al. GVCBLUP: a computer package for genomic prediction and variance component estimation of additive and dominance effects. **BMC Bioinformatics**, v. 15, n. 1, p. 270, 2014.

WELLMANN, R.; BENNEWITZ J. Bayesian models with dominance effects for genomic evaluation of quantitative traits. **Genetics research**, v. 94, n. 1, p. 21-37, 2012.

WITTENBURG, D.; MELZER, N.; REINSCH, N. Genomic additive and dominance variance of milk performance traits. **Journal of Animal Breeding and Genetics**, v. 132, n. 1, p. 3-8, 2014.

WOLC, A. et al. Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. **Genetics Selection Evolution**, v. 43, n. 1, p. 5, 2011.

YAZDI, M. H. et al. Heritability, reability of genetic evaluations and response to selection in proportional hazard models. **Journal Dairy Science**, Champaign, v. 85, n. 6, p. 1563 – 1577, 2002.

ZHANG, D.; LIN, X. Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. In: **Random effect and latent variable model selection**. Springer, New York, NY, 2008. p. 19-36.

CHAPTER IV

GENERAL CONCLUSIONS

Until date, the present work can be considered as the pioneer in the use of survival models, such as the Cox model, in predicting additive and nonadditive random effects in the context of genomic selection. For this purpose, the Cox model with Gaussian distribution for random effects was considered, by using the *coxme* function (*Mixed Effects Cox Models*) of the *coxme* R-package, being the variance–covariance matrix of random effects provided by the additive and dominance genomic relationship matrices. For the linear models, this method is called the GBLUP-D (genomic BLUP with additive and dominance effects).

The *coxme* package of R also implements the *lmekin* function, which predicts random effects while considering the mixed linear model. Thus, in Chapter I of this work, we compared the GBLUP-D method implemented in the available software's GVCBLUP and BGLR by using the *lmekin* function of R. The results obtained by using the *lmekin* function were found to be identical to those of the GVCBLUP software, thus indicating the possibility and effectiveness of using the proposed methodology for predicting genomic genetic values while considering the additive and dominance effects.

On comparison of the equality of results between the proposed methodologies by using the *lmekin* function and the GVCBLUP software, the analysis was extended to the Cox model, when the control code for *lmekin* was found to be identical to *coxme* with respect to specifying the random effects.

Thus, Chapter II of this manuscript discussed the genomic prediction in pigs using models for censored data with additive and dominance genetic effects. Considering the impossibility of comparing the proposed methodology with the other available ones considering the Cox model in genomic selection, if no program supports this type of analysis, the Cox model is used to compared to the truncated normal model, which is implemented in the BGLR

R-package, wherein censored data are sampled from a truncated normal distribution. The inclusion of the dominance effect was tested by the LRT for the Cox model and by the DIC criterion for the truncated normal models, as they are based on Bayesian regression. Both the approaches indicated the best model that includes only the additive effects, and, for the Cox model, the dominance effect was found to be null. In addition, a high degree of agreement was noted between the Cox model and the truncated normal linear model with respect to the selection of the best individuals and the greatest marker effects.

In addition to the dominance effect, the inclusion of one more random genetic effect in the Cox model was evaluated—the polygenic effect—which was compared again with the truncated normal model, as presented in chapter III.

In this chapter, we have also presented the comparison between the GLUP-D method fitted by using the *lmekin* function (frequentist fit) and that by using the BGLR package (Bayesian fit) while combining marker and pedigree information in mouse traits. In the 3 evaluated traits, both the dominance and polygenic effects were found to be significative for both the approaches (i.e., *lmekin* and BGLR). For the censored data of pigs, both the models indicated statistical significance of the polygenic effect, with a high percent of additive genetic variance explained by the pedigree.