ALESSANDRA ALVES DA SILVA

AUTOREGRESSIVE SINGLE-STEP TEST-DAY MODELS FOR GENOMIC PREDICTION AND GWAS IN PORTUGUESE HOLSTEIN CATTLE

Thesis presented to the Animal Science Graduate Program of the Universidade Federal de Viçosa, in partial fulfillment of the requirements for degree of *Doctor Scientiae*.

Adviser: Fabyano Fonseca e Silva

Co-advisers: Júlio Gil Vale Carvalheira
Paulo Sávio Lopes
Cláudio Napolis Costa

VIÇOSA - MINAS GERAIS
2019

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

T
S586
2019

Silva, Alessandra Alves da, 1991-
Autoregressive single-step test-day models for genomic prediction and GWAS in Portuguese Holstein cattle / Alessandra Alves da Silva. – Viçosa, MG, 2019.
89 f. : il. (algumas color.) ; 29 cm.

Texto em inglês.
Orientador: Fabyano Fonseca e Silva.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Bovinos de leite - Melhoramento genético. 2. Autocorrelação. 3. Genômica. 4. Lactação. I. Universidade Federal de Viçosa. Departamento de Zootecnia. Programa de Pós-Graduação em Zootecnia. II. Título.

CDD 22 ed. 636.20821

ALESSANDRA ALVES DA SILVA

AUTOREGRESSIVE SINGLE-STEP TEST-DAY MODELS FOR GENOMIC PREDICTION AND GWAS IN PORTUGUESE HOLSTEIN CATTLE

Thesis presented to the Animal Science Graduate Program of the Universidade Federal de Viçosa, in partial fulfillment of the requirements for degree of *Doctor Scientiae*.

APPROVED: July 15th, 2019.

Alessandra Alves da Silva
Author

Fabyano Fonseca e Silva
Adviser

ACKNOWLEDGMENTS

The Universidade Federal de Viçosa (UFV), especially to the Animal Science Graduate Program and the Department of Animal Science, for the opportunity of carrying out the course;

The University of Porto, especially to the Research Center in Biodiversity and Genetic Resources (CIBIO-InBio), and the Institute of Biomedical Sciences Abel Salazar (ICBAS) for theoretical and technical support;

The National Council of Technological and Scientific Development (CNPq, Brazil), Coordination for the Improvement of Higher Education Personnel (CAPES, Brazil), and Portuguese National Funding Agency for Science, Research and Technology (FCT, Portugal) for financial support;

The Portuguese Dairy Cattle Breeders Association (ANABLE) and Embrapa Dairy Cattle for providing the data and to have become this work possible;

I would like to express my sincere gratitude to my advisor Prof. Fabyano Fonseca e Silva for the continuous support of my Ph.D study, for all his supervision, motivation, friendship, advice, and opportunities given to me. You are an amazing person and I have learned a lot from you. Thank you very much for everything that you have done for me.

My sincere thanks also to my co-advisor Prof. Júlio Carvalheira, for making me feel welcome in your team and opportunities given to me. Without their valuable support, it would not be possible to conduct this research. Thank you very much for all his suggestions and comments in our studies, for their supervision, friendship, patience, and advice. I have learned a lot from you.

My sincere thanks also to my co-advisor Prof. Paulo Sávio Lopes, for his valuable teachings, friendship, advice, for all your great suggestions and comments in our studies. Thank you for the great professional experiences, which would not have been possible without your support.

My sincere thanks also to my co-advisor Prof. Cláudio Napolis Costa, for all teachings, support, friendship, for all your great suggestions and comments in our studies. Without their precious support, it would not be possible to conduct this research;

I also thank Prof. Renata Veroneze for their contributions, for the great ideas, support, discussions, and friendship;

I also would like to thank Dr. Alexandre Rodrigues Caetano and Dr. Moysés Nascimento, for accepting being part of this defense committee, and for all their contributions to this thesis;

I would also like to thank all friends, professors and staff of the DZO (Brazil), CIBIO and ICBAS (Portugal);

A special thank you to all friends from “salinha”, GDMA, and “Labtec” for the good conversations, discussions, barbecues and for the volleyball games, especially Teo, Darlene, Hugo, Hinayah, Ivan, Sirlene, Giovani, Eula, Amanda, Matheus, Natalia, Talita, Ingrid, Layla, Daniele, Margareth, Susana, Karine, Haniel, and Pamela;

A big thank you to my old friends, Denise, Nayanne, and Patricia, which even far away the friendship is always the same;

I would also like to thank my friends from Portugal, especially Fernanda and Luciano for friendship and great weekends;

Finally, I must express deep gratitude to my family for providing me with unfailing support and continuous encouragement throughout of years. This accomplishment would not have been possible without them: you always will be with me, mom (Joana) and my brother (Alexandre); sister and brother-in-law (Rosimary and Mário); brother and sister-in-law (Júnior and Sabrina); nephew and niece (Victor and Janete); father-in-law and mother-in-law (David and Geralda); and finally my husband and the best research partner (Delvan). Thank you very much for everything. I love you all.

BIOGRAPHY

Alessandra Alves da Silva was born in Tupã, São Paulo, Brazil on February 1991. In March 2009, she started her undergraduate studies in Animal Science at Universidade Federal de Mato Grosso (UFMT), located in Sinop, Mato Grosso, Brazil. During her course period, she was involved in several tutoring and scientific initiation activities in the animal breeding area, under the supervision of Professor Dr. Cláudio Vieira de Araújo.

In August 2013, she started her graduate program at Animal Science Graduate Program by UFV to obtain her degree of Magister Scientiae in Animal Science, under the supervision of Professor Dr. Simone Eliza F. Guimarães. She presented her dissertation in July 2015 in nutrigenomics area.

In August 2015, started her Ph.D at Animal Science Graduate Program by UFV, with the objective to expand her training in Animal Breeding and Genomics areas, under the supervision of Dr. Fabyano Fonseca e Silva. During her PhD, she had the opportunity to work in a research project of international cooperation between Brazil (Universidade Federal de Viçosa and Embrapa Gado de Leite) and Portugal (Universidade do Porto), under the supervision of Professor Dr. Julio Carvalheira. This project provided the results which are summarized in this thesis. Her thesis will be defended in July 2019.

ABSTRACT

SILVA, Alessandra Alves, D.Sc., Universidade Federal de Viçosa, July, 2019. **Autoregressive single-step test-day models for genomic prediction and GWAS in Portuguese Holstein cattle.** Adviser: Fabyano Fonseca e Silva. Co-advisers: Júlio Gil Vale Carvalheira, Paulo Sávio Lopes, and Cláudio Napolis Costa.

The milk traits are termed longitudinal traits because they are measured over time. Therefore, these traits need to be evaluated using appropriate statistical models to take into account the covariance structure that exists among the repeated records. Autoregressive test-day (AR) model for multiple lactations has been routinely used for the national genetic evaluation of dairy cattle in Portugal. Under this model, the animals' permanent environment are assumed to follow a first order autoregressive process as a long-term (auto-correlations between parities) and a short-term (auto-correlations between test-day within lactations) effects, taking into account the non-genetic correlations due to the cows' repeated performance. Currently, given the relevance of genomic prediction in dairy cattle, it is essential to include dense marker information in national genetic evaluations. Therefore, the general objective with this thesis was to evaluate the inclusion of genomic information in the AR test-day model for multiple lactations for better understand the genetic and genomic aspects of milk related traits in Portuguese Holstein cattle. Firstly, to perform the genomic evaluation under AR model we evaluated the imputation accuracy for Portuguese Holstein cattle using several commercially available SNP panels in different densities with a relatively small number of genotyped animals. Genotype imputation was feasible and may be advantageous to the National genomic evaluations. Thus, we analyzed the feasibility of applying the single-step GBLUP (ssGBLUP) to analyze milk yield using the AR (H-AR) model in Portuguese Holstein cattle. The use of H-AR increased the reliability and reduced the bias of GEBVs compared to traditional evaluation. Therefore, these results suggest that the ssGBLUP methodology applied to AR models is feasible and may be advantageous to the National genetic evaluations. With the anticipated increase in the number of genotyped animals (for example by including females), it is expected that the H-AR will provide even higher reliabilities especially for the young stock, thus contributing to the improvement in the genetic progress. In a

second step, we evaluated the feasibility of using a weighted ssGWAS methodology under a multiple lactations AR test-day model to find genomic regions associated with milk, fat, and protein yields, and score somatic cells (SCS). Genomic regions associated with the analyzed traits were also identified simultaneously throughout the lactations, to provide a better understanding of the genetic architecture for these traits. The findings described in this thesis will contribute to advance the knowledge about genomic prediction and GWAS for milk related traits. In addition, this thesis provides the first results about the inclusion of genomic information in AR models, which will be important for future national genetic evaluations.

Keywords: Autocorrelation. Dairy cattle. Gene function. Genomic evaluation. Imputation. Multiple lactations.

RESUMO

SILVA, Alessandra Alves, D.Sc., Universidade Federal de Viçosa, julho de 2019. **Modelos *test-day* autorregressivos para predição genômica em passo único e GWAS para bovinos Holstein Portugueses.** Orientador: Fabyano Fonseca e Silva. Coorientadores: Júlio Gil Vale Carvalheira, Paulo Sávio Lopes e Cláudio Napolis Costa.

As características de leite são denominadas características longitudinais porque são mensuradas ao longo do tempo. Portanto, essas características precisam ser avaliadas por meio de modelos estatísticos apropriados, que levem em consideração a estrutura de covariância existente entre as observações repetidas. Nesse contexto, o modelo autorregressivo *test-day* (AR) para múltiplas lactações tem sido rotineiramente usado na avaliação genética nacional de bovinos de leite em Portugal. Sob este modelo, assume-se que o ambiente permanente dos animais segue um processo autorregressivo de primeira ordem denominado *long-term* (auto correlações entre lactações) e *short-term* (auto correlações entre *test-day* dentro das lactações), levando em conta as correlações não-genéticas devido ao desempenho repetido das vacas. Atualmente, dada a relevância da predição genômica em bovinos leiteiros, é essencial incluir informações densas de marcadores em avaliações genéticas nacionais. Portanto, o objetivo principal desta tese foi avaliar a inclusão de informações genômicas no modelo *test-day* AR para múltiplas lactações para melhor compreensão dos aspectos genéticos e genômicos de características relacionadas ao leite em bovinos Holstein. Primeiramente, para realizar a avaliação genômica sob o modelo AR, foi necessário avaliar a acurácia da imputação para bovinos Holstein Portugueses, utilizando vários painéis de SNP comercialmente disponíveis em diferentes densidades, com um número relativamente pequeno de animais genotipados. A imputação de genótipos foi viável e pode ser vantajosa para as avaliações genômicas nacionais. Assim, foi possível investigar a viabilidade do *single-step* GBLUP (ssGBLUP) para produção de leite usando o modelo AR (H-AR). O modelo H-AR proporcionou um aumento na acurácia e reduziu o viés dos GEBVs em comparação com a avaliação tradicional. Portanto, estes resultados sugerem que a metodologia ssGBLUP aplicada aos modelos AR é viável e pode ser vantajosa para as avaliações genéticas nacionais. Com o aumento do

número de animais genotipados (incluindo fêmeas), espera-se que o H-AR proporcione uma acurácia ainda maior, especialmente para animais jovens, contribuindo assim para melhorias no progresso genético. Em uma segunda etapa, foi avaliado a viabilidade de usar a metodologia ssGWAS ponderado sob o modelo AR, para encontrar regiões genômicas associadas a produção de leite, gordura e proteína, e escore de células somáticas (SCS). As regiões genômicas associadas às características analisadas foram identificadas simultaneamente ao longo das lactações, para melhor compreensão da arquitetura genética dessas características. Os achados descritos nesta tese poderão contribuir para o avanço do conhecimento sobre predição genômica e GWAS para as características relacionadas a produção de leite. Além disso, esta tese fornece os primeiros resultados sobre a inclusão de informações genômicas em modelos AR, o que será importante para futuras avaliações genéticas nacionais.

Palavras-chave: Autocorrelação. Avaliação genômica. Bovinos de leite. Função gênica. Imputação. Múltiplas lactações.

SUMMARY

CHAPTER 1

1.1. General introduction

Milk and dairy products provide essential nutrients (e.g. minerals and vitamins) and are an important source of dietary energy, high-quality proteins and fats. More than 6 billion people worldwide consume milk and dairy products. In general, dairy consumption has increased worldwide, and is projected to increase almost 58% by 2050 (FAO, 2017). Given the increasing demand for food production, it is essential a continuous improvement in milk production.

In animal breeding context, the Holstein is the most widespread cattle breed in the world for milk production. The milk traits (such as milk yield, and milk composition) are termed longitudinal traits because they are measured over time. In this sense, longitudinal traits need to be evaluated using appropriate statistical models to take into account the covariance structure that exists among the repeated records. Thus, the genetic evaluation for these traits has been routinely performed using test-day (TD) models (Schaeffer et al., 2000; Carvalheira et al., 2002). These models consider the direct use of TD records and provide more precise adjustment for permanent and temporary environmental effects on TD.

Among the TD models, the autoregressive test-day (AR) model for multiple lactations proposed by Carvalheira et al. (2002) has been routinely used for the national genetic evaluation of dairy cattle in Portugal. Under this model, the animals' permanent environment are assumed to follow a first order autoregressive process as a long-term (auto-correlations between parities) and a short-term (auto-correlations between test-day within lactations) effects, taking into account the non-genetic correlations due to the cows' repeated performance.

Currently, given the relevance of genomic prediction in dairy cattle, it is essential to include dense marker information in national genetic evaluations. In this sense, the imputation strategies have become an important approach to make efficient use of all available genotype information. Currently, there are several commercial SNP panels with different densities in dairy cattle providing many genotypic datasets that are routinely shared between countries to minimize costs with genomic information (Nicolazzi et al., 2015). Thus, imputation strategies have contributed to increasing the genotyped animals number included in further analyses, such as genomic prediction or genome-wide association studies (GWAS). Most importantly, the accuracy of imputation has to be sufficiently high to allow for reliable conclusions in these analyses.

Several factors are known to influence the accuracy of imputation (Larmer et al., 2014; Boison et al., 2015; García-Ruiz et al., 2015). However, most of these studies involved a large number of genotyped animals using low-density panels derived from only an HD panel. Thus, results from these studies may not be directly applied to populations consisting of a limited number of individuals, since the population structure affect the imputation accuracy, reinforcing the need and importance for reliable imputation strategies.

Genomic selection has become a standard procedure in dairy cattle breeding given its potential to increase selection accuracy and to reduce generation interval (VanRaden, 2008). Despite the advancements and practical applications of AR model in genetic evaluations, yet there are no reports in the literature which investigating the appropriateness of using AR in genomic evaluations. One of the most used methods in genomic selection is based on the multistep procedure (VanRaden, 2008). In brief, the multistep procedure requires the estimation of traditional EBV, and the direct genomic

values (DGV) are predicted for genotyped animals using daughter yield deviations (DYD) or deregressed phenotypes from the previous step (VanRaden, 2008; Hayes et al., 2009; VanRaden et al., 2009). Finally, the DGVs are combined with the parent averages to predict the genomic EBV (GEBV) (VanRaden et al., 2009). In general, this procedure has shown some disadvantages, such as the loss of information attributed to pseudo-phenotypes instead true phenotypes. Additionally, this approach is very complex and requires several approximations, which may reduce the accuracy and induce bias on the GEBV (Koivula et al., 2015).

The single-step genomic BLUP (ssGBLUP), where phenotypes, pedigree, and genotypes are jointly used to generate GEBV as a direct output, is a suitable alternative to multistep approach (Misztal et al., 2009; Aguilar et al., 2010; Christensen and Lund, 2010). The ssGBLUP has been successfully applied to Holstein cattle (Lourenco et al., 2014; Koivula et al., 2015; Jattawa et al., 2016) and outperformed the multistep procedure. Therefore, using the ssGBLUP method under AR model could increase the reliability of the genomic prediction.

In GWAS context, most GWAS studies have carried out using the total accumulated yield from first lactation (Nayeri et al., 2016; Yue et al., 2017). In addition, these studies have usually used de-regressed EBVs as pseudo-phenotypes, and fitted the SNPs (one at a time) as a fixed effect. The weighted single-step GWAS (ssGWAS) proposed by Wang et al. (2012), is a suitable alternative to traditional GWAS. In this methodology, SNP effects are estimated using GEBVs, which are then used in an iterative approach to update solutions and weight for SNP are used when forming the genomic relationship matrix. Thus, this method allows unequal variances for SNPs, which results in improved precision of the estimation of SNP effects (Wang et al., 2012). The ssGWAS has been successfully applied to milk related traits (Iung et

al., 2019; Zhou et al., 2019), which revealed important candidate genes and QTLs regions associated with these traits, but still using the total accumulated yield. Thus, identifying QTL regions and candidate genes associated with milk related traits, using the weighted ssGWAS methodology under AR test-day model for multiple lactations could provide a better understanding of the genetic architecture for these traits.

1.2. Objectives

The general objective with this thesis was to evaluate the inclusion of genomic information in the AR test-day model for multiple lactations for better understand the genetic and genomic aspects of milk related traits in Holstein cattle. The specific objectives were to: i) evaluate the imputation accuracy for Portuguese Holstein population considering several commercially available SNP panels in different densities with a small number of genotyped animals; ii) evaluate the feasibility of using ssGBLUP methodology to predict GEBVs for milk yield under a multiple lactations AR model; iii) identify QTL regions and important genes associated with milk related traits, using the weighted ssGWAS methodology under a AR test-day model; and investigate the differences in QTL regions and important genes found using the first 3 lactations or only the first lactation.

1.3. References

Aguilar, I., I. Misztal, D.L. Johnson, A. Legarra, S. Tsuruta, and T.J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score1. J. Dairy Sci. 93:743–752. doi:10.3168/jds.2009-2730.

Boison, S.A., D.J.A. Santos, A.H.T. Utsunomiya, R. Carvalheiro, H.H.R. Neves, A.M.P. O'Brien, J.F. Garcia, J. Sölkner, and M.V.G.B. da Silva. 2015. Strategies for single nucleotide polymorphism (SNP) genotyping to enhance genotype

imputation in Gyr (Bos indicus) dairy cattle: Comparison of commercially available SNP chips. J. Dairy Sci. 98:4969–4989. doi:10.3168/jds.2014-9213.

Carvalheira, J., E.J. Pollak, R.L. Quaas, and R.W. Blake. 2002. An autoregressive repeatability animal model for test-day records in multiple lactations. J. Dairy Sci. 85:2040–2045. doi:10.3168/jds.S0022-0302(02)74281-1.

Christensen, O.F., and M.S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42:1–8.

FAO. 2017. The state of food and agriculture: leveraging food systems for inclusive rural transformation. Rome, Food and Agriculture Organization. ISBN 978-92-5-109873-8.

García-Ruiz, A., F.J. Ruiz-Lopez, G.R. Wiggans, C.P. Van Tassell, and H.H. Montaldo. 2015. Effect of reference population size and available ancestor genotypes on imputation of Mexican Holstein genotypes. J. Dairy Sci. 98:3478–3484. doi:10.3168/jds.2014-9132.

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

Iung, L.H.S., J. Petrini, J. Ramírez-Díaz, M. Salvian, G.A. Rovadoscki, F. Pilonetto, B.D. Dauria, P.F. Machado, L.L. Coutinho, G.R. Wiggans, and G.B. Mourão. 2019. Genome-wide association study for milk production traits in a Brazilian Holstein population. J. Dairy Sci. 1:1–10. doi:10.3168/jds.2018-14811.

Jattawa, D., M.A. Elzo, S. Koonawootrittriron, and T. Suwanasopee. 2016. Imputation accuracy from low to moderate density single nucleotide polymorphism chips in a Thai multibreed dairy cattle population. Asian-Australasian J. Anim. Sci. 29:464–470. doi:10.5713/ajas.15.0291.

Koivula, M., I. Strandén, J. Pösö, G.P. Aamand, and E.A. Mäntysaari. 2015. Single-step genomic evaluation using multitrait random regression model and test-day data. J. Dairy Sci. 98:2775–2784. doi:10.3168/jds.2014-8975.

Larmer, S.G., M. Sargolzaei, and F.S. Schenkel. 2014. Extent of linkage disequilibrium, consistency of gametic phase, and imputation accuracy within and across Canadian dairy breeds. J. Dairy Sci. 97:3128–3141. doi:10.3168/jds.2013-6826.

Lourenco, D.A.L., I. Misztal, S. Tsuruta, I. Aguilar, E. Ezra, M. Ron, A. Shirak, and J.I. Weller. 2014. Methods for genomic evaluation of a relatively small genotyped

dairy population and effect of genotyped cow information in multiparity analyses. J. Dairy Sci. 97:1742–1752. doi:10.3168/jds.2013-6916.

Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92:4648–4655. doi:10.3168/jds.2009-2064.

Nayeri, S., M. Sargolzaei, M.K. Abo-ismail, N. May, S.P. Miller, F. Schenkel, S.S. Moore, and P. Stothard. 2016. Genome-wide association for milk production and female fertility traits in Canadian dairy Holstein cattle. BMC Genet. 1–11. doi:10.1186/s12863-016-0386-1.

Nicolazzi, E.L., S. Biffani, F. Biscarini, P. Orozco Ter Wengel, A. Caprera, N. Nazzicari, and A. Stella. 2015. Software solutions for the livestock genomics SNP array revolution. Anim. Genet. 46:343–353. doi:10.1111/age.12295.

Schaeffer, L.R., J. Jamrozik, G.J. Kistemaker, and B.J. Van Doormaal. 2000. Experience with a Test-Day Model. J. Dairy Sci. 83:1135–1144. doi:10.3168/jds.S0022-0302(00)74979-4.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions.. J. Dairy Sci. 91:4414–23. doi:10.3168/jds.2007-0980.

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

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W.M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. (Camb). 94:73–83. doi:10.1017/S0016672312000274.

Yue, S.J., Y.Q. Zhao, X.R. Gu, B. Yin, Y.L. Jiang, Z.H. Wang, and K.R. Shi. 2017. A genome-wide association study suggests new candidate genes for milk production traits in Chinese Holstein cattle. Anim. Genet. 48:677–681. doi:10.1111/age.12593.

Zhou, C., C. Li, W. Cai, S. Liu, H. Yin, S. Shi, Q. Zhang, and S. Zhang. 2019. Genome-Wide Association Study for Milk Protein Composition Traits in a Chinese Holstein Population Using a Single-Step Approach. Front. Genet. 10:1–17. doi:10.3389/fgene.2019.00072.

CHAPTER 2

Genotype imputation strategies for Portuguese Holstein cattle using different SNP panels

Alessandra Alves Silva[1], Fabyano Fonseca Silva[1], Delvan Alves Silva[1], Hugo Teixeira Silva[1], Cláudio Napolis Costa[2], Paulo Sávio Lopes[1], Renata Veroneze[1], Gertrude Thompson[3,4], Julio Carvalheira[3,4]

[1]*Department of Animal Science, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil*
[2]*Embrapa Dairy Cattle, Juiz de Fora, Minas Gerais, Brazil*
[3]*Research Center in Biodiversity and Genetic Resources (CIBIO-InBio), University of Porto, Vairão, Porto, Portugal*
[4]*Institute of Biomedical Sciences Abel Salazar (ICBAS), University of Porto, Porto, Portugal*

Paper submitted in Czech Journal of Animal Science (May 2019).

2.1. Abstract

We aimed to compare different imputation strategies for the Portuguese Holstein cattle population considering several commercially available SNP panels with a relatively small number of genotyped animals. Data from 1,358 genotyped animals were used to evaluate imputation in 7 different scenarios. In the scenarios S1 to S6, imputations were performed from LDv1, 50Kv1, 58K, HD, HDv3 and Ax58K panels to 50Kv2 panel. In these scenarios, the bulls in 50Kv2 were divided into reference (352) and validation (101) populations based on birth year. In the scenario S7, the validation population consisted of 566 cows genotyped with the Ax58K panel with their genotypes masked to LDv1. In general, all sample imputation accuracies were high with correlations ranging from 0.94 to 0.99 and concordance rate ranging from 92.59 to 98.18%. SNP-specific accuracy was consistent with that of sample imputation. S4 (40.32% of SNPs imputed) had higher accuracy than S2 and S3, both with less than 7.59% of SNPs imputed. Most probably, this was due to the high number of imputed SNPs with minor allele frequency (MAF) < 0.05 in S2 and S3 (18.43% and 16.06% higher than in S4, respectively). Therefore, for these two scenarios, the MAF was more relevant than panel density. These results suggest that genotype imputation using several commercially available SNP panels is feasible for the Portuguese National genomic evaluation.

Keywords: dairy cattle, genomic evaluation, imputation accuracy

2.2. Introduction

In practice, several studies in animal breeding with genomic data such as genomic prediction or genome-wide association studies (GWAS) use imputed data

(Jattawa et al., 2016; Wang et al., 2016). Thus, imputation strategies have become an important approach to make efficient use of all available information. More, there are several commercial SNP panels with different densities in dairy cattle (Nicolazzi et al., 2015) providing many genotypic datasets that are routinely shared between countries to minimize costs with genomic information, reinforcing the need and importance for reliable imputation strategies.

Briefly, imputation refers to statistical and computational tools applied to infer SNP genotypes, which are not obtained from a low-density panel using information from a reference population genotyped with a higher density panel (Ventura et al., 2014). Several studies have investigated the factors affecting imputation accuracy in dairy cattle population, such as, the number of reference individuals, the relationship between reference population and target population, minor allele frequency (MAF), linkage disequilibrium (LD), and the difference between marker densities of the reference and imputed sets (Larmer et al. 2014; Boison et al. 2015; García-Ruiz et al. 2015). However, most of these studies involved large number of genotyped animals using low-density panels derived from only an HD panel. Thus, results from these studies may not be direct applied to small populations, since the population structure affect the imputation accuracy. In addition, other factors affecting imputation accuracy may also be intensified in small populations. Therefore, we aimed to test different imputation strategies for Portuguese Holstein population considering several commercially available SNP panels in different densities with a small number of genotyped animals.

2.3. Material and methods

Data from 1,358 genotyped animals were used in this study. Of these, 50.85% were foreign bulls, mainly American and Canadian (82.63%), while 49.15% of the animals were Portuguese, mainly cows (83.83%) and only 16.17% bulls. The animals were genotyped using different panels: LDv1 (GeneSeek Genomic Profiler, Neogen Corp., Lincoln, NE, USA), 50Kv1 and 50Kv2 (Bovine SNP50v.1 and Bovine SNP50v.2 BeadChips, Illumina, San Diego, CA, USA), 57K (USDA Illumina, San Diego, CA, USA), Ax58K (Affymetrix, Santa Clara, CA, USA), 77K and HDv3 (GeneSeek Genomic Profiler, Neogen Corp., Lincoln, NE, USA), including 8,610, 54,001, 54,609, 56,947, 57,497, 76,883 e 139,376 markers, respectively. All the cows were genotyped with Ax58K. The numbers of genotyped animals by panel and birth year are shown in Table 1.

Table 1. Number of genotyped animals by panel according to birth year

Birth year	50Kv2	LDv1	50Kv1	57K	Ax58K[a]	77K	HDv3
1966-2003	352	-	23	76	-	31	18
2004	9	-	6	-	-	-	6
2005	3	-	22	-	-	-	7
2006	7	-	25	-	-	-	8
2007	2	-	20	-	-	-	7
2008	5	-	25	1	-	-	13
2009	6	1	15	2	-	-	6
2010	23	3	8	1	71	-	11
2011	35	-	-	-	96	-	1
2012	10	-	-	-	166	3	1
2013	-	-	-	-	193	-	-
2014	1	-	-	-	39	-	-
2015	-	-	-	-	1	-	-
Total	453	4	144	80	566	34	78

[a]Females.

Marker positions and chromosomes in the map for each panel were standardized according to the UMD v3.1 assembly (Zimin et al., 2009). Quality control analysis

(QC) for SNP and sample were done separately for each panel using PLINK v1.07 (Purcell et al., 2007). In QC for samples, errors of sex disagreement were checked by the heterozygosity on the X chromosome; animals with call rate < 0.90 were discarded; and deviations from heterozygosity were controlled by removing animals with ± 3 standard deviations. In QC for SNPs, markers with call rate < 0.95; minor allele frequency (MAF) < 0.02; and Hardy-Weinberg equilibrium with χ^2 lower than 10^{-6} were excluded. SNPs with positions unknown or located on sex chromosome were not considered in the analysis. The number of SNPs after QC for each panel are described in Table 2.

When using a diversity of genotype panels, mainly from different technologies (e.g., Illumina and Affymetrix), it is important to assess their consistency by performing a population structure analysis. A genotype data set contained the 3,239 SNPs that were shared between all panels were selected and used to evaluate the population structure through principal components analysis. For this analysis, principal components were calculated from the genomic relationship matrix (**G**) obtained according to VanRaden, (2008) as follows:

$$\mathbf{G} = \frac{(\mathbf{M} - 2\mathbf{P})(\mathbf{M} - 2\mathbf{P})'}{2\sum p_i (1 - p_i)}$$

in which, **M** is a matrix of minor allele (with dimension equal to number of animals by number of SNP markers), p_i is the frequency of the allele A of the *i-th* SNP and **P** is a matrix (with dimensions equal to the number of animals by number of SNPs) with each row containing the p_i values. The **G** matrix and principal components were obtained using the PreGSF90 software (Misztal et al., 2014). The first and second principal components calculated based on the **G** matrix are shown in Figure 1. The animals from different panels showed high connectivity, indicating high consistency between the studied panels.

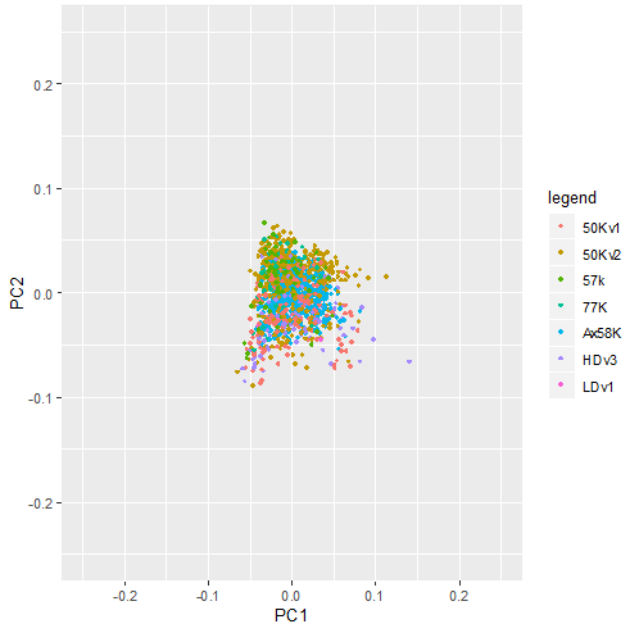

Figure 1. Plot of the first 2 principal components (PC) of the genomic relationship matrix between animals of each Panel.

The genotypes from different panels were imputed to 50Kv2 because this panel included the highest number of males. In addition, several studies have shown that increasing the SNP density above 50,000 markers (50K) added small gains in the reliability of genomic prediction for Holstein cattle (Vanraden et al., 2011; VanRaden et al., 2013). Imputations were performed using the FImpute 2.2 software (Sargolzaei et al., 2014). According to Boison et al. (2015) and Jattawa et al. (2016), this software presents high imputation accuracies and high computational performance in cattle populations. The FImpute option combining family and population-based algorithms was considered here.

Seven imputation scenarios based on panel densities were investigated. In the scenarios S1 to S6, imputations were performed from LDv1, 50Kv1, 58K, HD, HDv3 and Ax58K panels, to 50Kv2 panel. In these scenarios, the bulls in 50Kv2 panel were divided into reference (N =352) and validation (N = 101) populations based on their

birth year. The validation bulls had their 50K genotypes masked to each imputed panel (S1 to S6). The last scenario (S7) was defined to assess the quality of the imputation from Ax58K to 50Kv2 (Affymetrix and Illumina technologies, respectively), where the validation population consisted of 566 cows from Ax58K panel. Their 58K genotypes were masked to the LDv1 panel and the reference population was given by 352 bulls from 50Kv2. The seven scenarios are shown in Table 2.

Table 2. Description of imputation scenarios evaluated S1 to S7

Scenarios	Description[a]	No of animals Validation/Target	No of SNPs after QC	No of SNPs common to 50Kv2
S1	LDv1	101/4	7,306	6,304
S2	50Kv1	101/144	39,910	35,866
S3	57K	101/80	40,602	35,692
S4	77K	101/34	69,403	23,051
S5	HDv3	101/78	118,298	33,383
S6	Ax58K	101/566	42,123	32,476
S7	Ax58K-a[b]	566/566	42,123	5,437

[a]The reference panel for all scenarios studied was the 50Kv2 including 38,624 SNPs after QC analysis.
[b]The validation population consisted of 566 cows, which had their genotypes (32,476 SNPs in common to 50Kv2) masked to the LDv1 panel. QC: quality control analysis.

Imputation accuracies (per sample and SNP-specific) were assessed using the Spearman correlation coefficient (r) between the imputed and true SNPs markers and the concordance rate (CR) as a proportion of correctly imputed SNPs in relation to all imputed SNPs.

The effect of having relatives in the reference population for imputation was also investigated. Based on the **G** matrix (3,239 SNPs shared between all panels), we calculated the average of top 10 relationships and the average of all relationships higher than zero between each imputed animal and the those in the reference population. The relationship between relatedness and imputation accuracy was assessed by regressing r on the average of top 10 relationship.

To evaluate the effect of the MAF on imputation accuracy, the SNPs to be imputed were classified according two levels of MAF obtained from reference population: MAF < 0.1 and MAF < 0.05. Linkage disequilibrium (LD) between markers was measured using r^2 (Hill and Robertson, 1968) expressed as follow:

$$r^2 = \frac{D^2}{f(A)f(a)f(B)f(b)}$$

in which, $D = f(AB) - f(A)\,f(B)$ and $f(AB)$, $f(A)$, $f(a)$, $f(B)$, and $f(b)$, are observed frequencies of haplotype AB and alleles A, a, B, and b, respectively. We calculated r^2 among SNPs in the reference group between pairs of loci within chromosomes that were less than or equal to 1 Mb apart using PLINK v1.07 (Purcell et al., 2007). The average LD by SNP pair was then calculated as the average of all LD values between a given SNP and all others within a window of 1 Mb.

2.4. Results

Sample imputation accuracy and effect of relatedness on accuracy

Table 3 shows the sample imputation accuracy (r and CR) for each evaluated scenario. In general, the accuracy means were high for all scenarios, ranging from 0.94 to 0.99. The CR was also high ranging from 92.59 to 98.18%. The lowest values of accuracy were observed in the S1 and S7 scenarios, which had a higher number of imputed SNPs (83.68 and 83.26%, respectively). Although the mean of imputation accuracy was similar over the scenarios, the standard deviation was higher for the S2 and S3, which had the lower number of imputed SNPs (7.14 and 7.59%, respectively). On the other hand, the scenarios that had a higher number of imputed SNPs (> 13.57%), presented the lowest standard deviations.

The S7 was performed to evaluate the imputation between panels from different companies (Illumina and Affymetrix), in which 27,039 SNPs were imputed.

Accuracies were high (mean of 0.94 for r and of 92.59% for CR) and similar to those found in the other scenarios that consisted of panels from Illumina only (Table 3). In addition, the S7 also allow us to access accuracy for females whose genotype imputation was done using a reference population composed only by males.

Table 3. Sample imputation accuracy for all scenarios evaluated. Spearman correlation coefficient (r) between the imputed and true SNPs markers and Concordance rate (CR), as a proportion of correctly imputed SNPs versus the true SNPs

Scenarios	Number of SNPs imputed	r				CR			
		Mean	Min.[a]	Max.	SD	Mean	Min.	Max.	SD
S1	32,320 (83.68%)	0.96	0.93	0.99	0.012	95.39	91.46	98.27	1.45
S2	2,758 (7.14%)	0.98	0.80	0.99	0.032	97.25	73.46	99.27	4.26
S3	2,932 (7.59%)	0.98	0.80	0.99	0.032	97.24	75.03	99.28	4.02
S4	15,573 (40.32%)	0.99	0.95	0.99	0.008	98.11	93.37	99.22	1.07
S5	5,241 (13.57%)	0.99	0.93	0.99	0.011	98.18	91.18	99.43	1.48
S6	6,148 (15.92%)	0.98	0.93	0.99	0.010	97.87	91.64	99.15	1.31
S7	27,039 (83.26%)	0.94	0.79	0.98	0.02	92.59	76.54	97.40	2.41

[a]Min: Minimum; Max: Maximum; SD: standard deviation.

The average relatedness between validation and reference individuals as well as between target and reference individuals are shown in Table 4. The higher genomic relatedness with reference individuals was observed in scenarios S3 e S4, in which 58.75 and 73.53% of individuals had at least one genotyped half-sib in the reference population, respectively. The remaining scenarios including validation had less than 15.38% of individuals with at least one genotyped half-sib in the reference population.

The impact of relatedness between validation and reference animals on imputation accuracy is shown in Figure 2. The greatest influence of relatedness with the reference population on the r was observed for S1 ($P < 0.01$), in which a higher number of SNPs were imputed. On the other hand, this behavior was not observed (P

> 0.05) in the remaining scenarios (S2, S3, S4, S5, and S6), in which lower number of SNPs were imputed.

Table 4. Average and standard deviation (SD) genomic relationships between animals in the imputed and reference data set

Panel (Imput/Ref[a])	Genomic relationships[b]	
	Top10	Relationships > 0
Val/Ref	0.11 (0.05)	0.031 (0.008)
S1/Ref	0.09 (0.01)	0.028 (0.003)
S2/Ref	0.13 (0.04)	0.034(0.006)
S3/Ref	0.18 (0.06)	0.046(0.010)
S4/Ref	0.19 (0.05)	0.049 (0.012)
S5/Ref	0.14 (0.06)	0.035 (0.009)
S6/Ref	0.1 (0.03)	0.031 (0.005)

[a]Imput: Scenarios imputed; Val: Validation population; Ref: Reference population.

[b]Top10: Average genomic relationships of top10 and average relationships > 0, between imputed individuals and reference individuals.

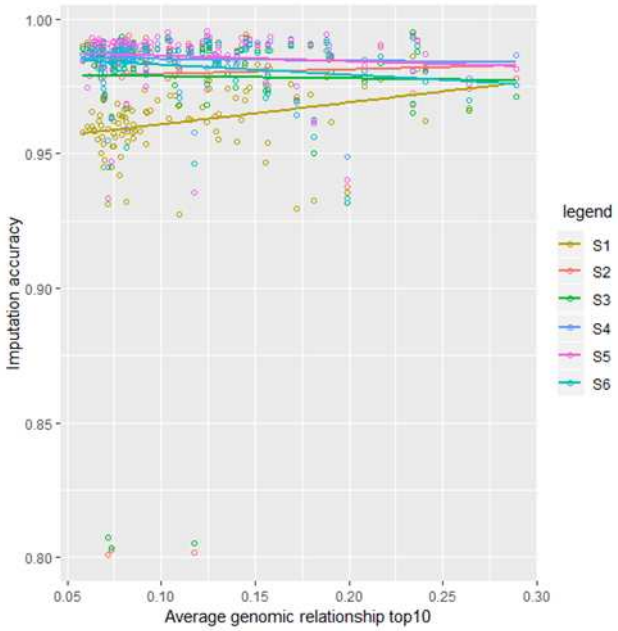

Figure 2. Imputation Accuracy (r) as a function of genomic relatedness (top10).

SNP-specific imputation accuracy and effect of MAF

In order to graphically display the results of imputation accuracy by chromosome, we built a circular plot using the *circlize* package (Gu et al., 2014) of R software (R Core Team, 2018). The average accuracy (r and CR) for SNPs by chromosomes are shown in Figure 3. In general, the r and (CR) accuracies (among chromosomes) were 0.91 (95.25), 0.92 (97.22), 0.93 (97.20), 0, 96 (98.09), 0.95 (98.14), 0.96 (97.87) and 0.87 (92.43%), for S1, S2, S3, S4, S5, S6 and S7 scenarios, respectively. As expected, the S1 and S7 presented the lowest accuracy due to the higher number of imputed SNPs (83.68% and 83.26%, respectively). On the other hand, although with minimal differences, the S4 containing 40.32% of imputed SNPs presented higher accuracy when compared to S2 or S3 scenarios, in which only 7.14% and 7.59% of SNPs were imputed, respectively.

The average MAF (< 0.1 and < 0.05) in the reference population was calculated for each set of SNPs imputed in each scenario. The number of SNPs in each level of MAF is described in Table 5. In the S1, in which a higher number of SNPs were imputed (83.68%), it was observed lower proportion of SNPs with MAF lower than 0.1 or 0.05. Similar result was observed for S4, that contained 40.32% of imputed SNPs. Differently, in the S2 and S3, in which a lower number of SNPs were imputed (7.14 and 7.59%, respectively), we observed higher proportion of SNPs with MAF lower than 0.1 or 0.05. For simplicity and clarity, only the MAF lower than 0.05 will be discussed here.

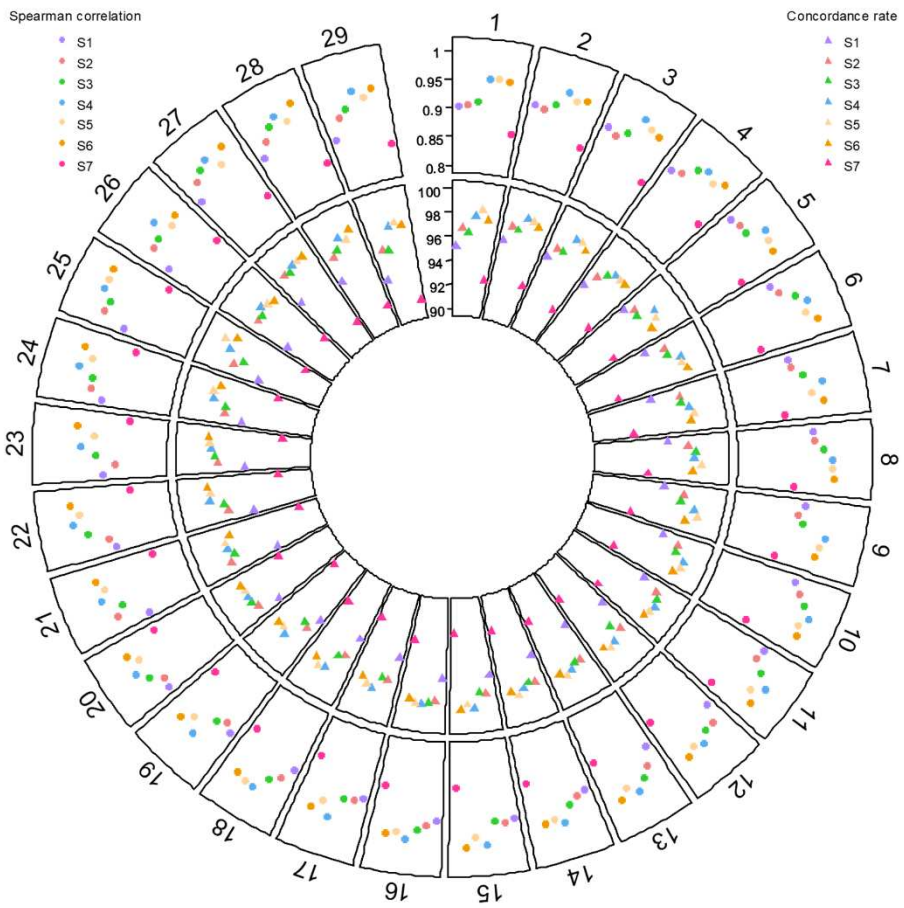

Figure 3. SNP-specific imputation accuracy for all scenarios evaluated.

Table 5. Number of imputed SNPs and number of SNPs in different levels of minor allele frequency (MAF) evaluated

Scenarios	Number of imputed SNPs	MAF	
		< 0.1	< 0.05
S1	32,320	5,743 (17.77%)	2,094 (6.48%)
S2	2,758	1,176 (42.64%)	769 (27.88%)
S3	2,932	1,065 (36.32%)	748 (25.51%)
S4	15,573	3,453 (22.17%)	1,471 (9.45%)
S5	5,241	1,790 (34.15%)	982 (18.74%)
S6	6,148	1,158 (18.84%)	638 (10.38%)

Linkage disequilibrium (LD)

The mean of LD estimated for the 29 chromosomes considering the animals of the reference population is shown in Figure 4. It was observed that 52.82% of the SNPs had low levels of LD (< 0.10). The mean LD for the population was 0.107. The highest mean LD was observed on chromosome 14 (r^2 = 0.135), while chromosome 27 had the lowest mean LD (r^2 = 0.083). These results are in agreement with those reported by Salem et al. (2018), in which the authors studied the level of LD in Portuguese Holstein cattle.

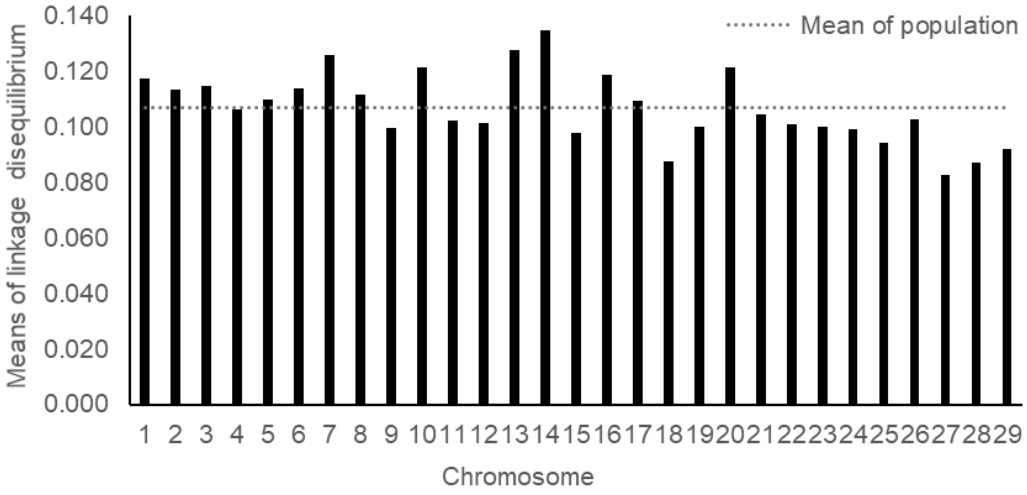

Figure 4. Mean of linkage disequilibrium (r^2) between adjacent SNPs makers separated by at most 1 Mb within each chromosome.

2.5. Discussion

Our study evaluated the accuracy of imputation for Portuguese Holstein cattle using several commercially available SNP panels in different densities and a relatively low number of genotyped animals. Imputation was performed using FImpute software (Sargolzaei et al., 2014), and we evaluated sample imputation accuracy, the effect of

relatedness on sample imputation accuracy, SNP-specific imputation accuracy, effect of MAF on SNP-specific imputation, and linkage disequilibrium.

Concordance rate and squared Pearson correlation coefficient have been reported in several studies (Boison et al. 2015; Jattawa et al. 2016; Ventura et al. 2016) as measure of imputation accuracy. Nevertheless, Pearson's correlation coefficient assumes that the two samples are normally distributed. If the assumption of normality is violated, Pearson's correlation coefficient may produce unreliable results. Differently, the Spearman rank correlation coefficient is a non-parametric measure of correlation, calculated on ranks and it depicts a monotonic relationship (Göktaş and İşçi 2011). Therefore, Spearman correlation seems to be more appropriate to measure accuracy for data obtained from genotypes.

In general, sample imputation accuracies were within the range of those reported for dairy cattle (Boison et al. 2015; Jattawa et al. 2016). Several studies have shown that imputation accuracy increases according to reduction in SNPs number to be imputed (Khatkar et al., 2012; Chud et al., 2015); however, this pattern was not clearly observed in the present study. Although with minimal differences, better results were observed for the S4 scenario, in which higher number of SNPs were imputed (40.32%) in relation to scenarios with lower number of SNPs imputed (< 7.59% - Table 3). Several factors may have influenced these results, such as the size and structure of the reference population, level of relationship between the animals to be imputed and the reference population size, the position of the SNP on the chromosome and its frequency (MAF) in the population (Ventura et al., 2014).

In the S7 scenario, which the panels are from different technologies (Affymetrix and Illumina), the imputation for the 50Kv2 using bulls in the reference population was successful, presenting imputation accuracy similar to those found in the other

panels (Table 3). Similar results were observed by Berry et al. (2016) and Zhou et al. (2014). In addition, with the expected increase in the number of genotyped females in the genetic evaluations, it is essential to evaluate the imputation accuracy for this group. The validation in females (S7 scenario) also indicates that females may be imputed using males in the reference population. Similar results were observed by Chud et al. (2015), García-Ruiz et al. (2015) and Jattawa et al. (2016) using larger populations.

Studies have shown that the imputation accuracy is strongly associated with the level of relationship between the animals from imputed and reference data set (Carvalheiro et al., 2014; Boison et al., 2015; Ventura et al., 2016). In general, imputation accuracy is measured only as a function of the density of SNPs to be imputed using a validation population composed by animals from the reference panel, in which their SNPs were masked to the target SNP panel. Nevertheless, the validation population may not represent the real population to be imputed. Therefore, to assist the comparison of the imputation accuracies in the different scenarios, a summary of the genomic relationship between the reference and validation population has been exploited (Daetwyler et al., 2013; Boison et al., 2015). In our study, the mean of the top10 genomic relationship between animals of imputed scenarios and reference was similar (ranging from 0.09 to 0.19) to the mean of genomic relationship between animals of validation and reference populations (Table 4). This result indicated that the animals selected to compose the validation population are representative of all evaluated scenarios, thus the imputation accuracy can be compared. We observed a strong relationship between genomic relatedness and imputation accuracy only for S1, wherein higher number of SNPs were imputed (Figure 2). These results indicate that relatedness has a greater influence on imputation accuracy when higher SNP densities

are imputed. Similar results were also found by Carvalheiro et al. (2014) and Chud et al. (2015).

SNP-specific imputation accuracy was consistent with the results obtained for the sample imputation accuracy, where S4 (40.32% of SNPs imputed) presented higher accuracy when compared to scenarios wherein less than 7.59% of SNPs (e.g., S2 and S3) were imputed (Table 3, Figure 3). Most probably this occurred due to the imputation of SNPs with low frequency in the population, since the number of imputed SNPs with MAF lower than 0.05 were 18.43% and 16.06% higher in S2 and S3 compared to the S4, respectively (Table 5). This is consistent with a previous study of Ventura et al. (2016), in which authors investigated the accuracy imputation for rare alleles according to MAF level (ranging from 0 to 0.05) and observed low accuracy for rare alleles (up to 57.8%). In addition, Boison et al. (2015) showed that the Illumina 50Kv2 panel presented higher proportion of markers with low MAF compared to the panels from GeneSeek. Therefore, for these scenarios (S1 and S2), the MAF was more relevant in the imputation accuracy than panel density. Probably, a strategy to reduce the effect of MAF on imputation accuracy would be to increase the size of the reference population (Heidaritabar et al., 2015).

In agreement with Carvalheiro et al. (2014), the r values were higher than the corresponding CR values for sample imputation accuracy because the penalty for one incorrectly imputed allele is relatively higher for CR than for r (Table 3). On the other hand, the opposite behavior was observed in SNP-specific imputation accuracy, in which the r values were lower than the corresponding values in CR (Figure 3). There are no reports comparing both criteria in SNP-specific imputation accuracy, but most probably this occurred due to the MAF effect. As reported by Hickey et al. (2012), a marker with very low MAF increases the probability of being homozygous for the

common allele, which, in general produces an increase in CR. The opposite behavior is expected for r, where the imputation accuracy is lower for markers with low MAF.

In summary, we evaluated the imputation accuracy for Portuguese Holstein cattle using several commercially available SNP panels in different densities with a relatively small number of genotyped animals. Sample imputation accuracy was higher than 0.93 for Spearman's correlation and higher than 92% for concordance rate. In addition, relatedness has a large influence on accuracy when higher SNPs densities are imputed. Our results also indicated that females may be imputed using males in the reference population. SNP-specific imputation accuracy was higher than 0.86 for Spearman's correlation and higher than 92% for concordance rate. Moreover, MAF was more relevant for accuracy than panel density, probably due to the small number of animals in the reference population used in this study.

2.6. Conclusion

Our results suggest that genotype imputation for Portuguese Holstein cattle using several commercially available SNP panels in different densities with a relatively small number of genotyped animals is feasible and may be advantageous to the National genomic evaluations of dairy cattle.

2.7. Acknowledgments

The authors acknowledge Portuguese Dairy Cattle Breeders Association (ANABLE) and Embrapa Dairy Cattle for providing the used data. This study was partially financed by CAPES/FCT (99999.008462/2014-03) and CNPq/INCT-CA.

2.8. References

Berry, D.P., A. O'Brien, E. Wall, K. McDermott, S. Randles, P. Flynn, S. Park, J. Grose, R. Weld, and N. McHugh. 2016. Inter- and intra-reproducibility of genotypes from sheep technical replicates on Illumina and Affymetrix platforms. Genet. Sel. Evol. 48:1–5. doi:10.1186/s12711-016-0267-0.

Boison, S.A., D.J.A. Santos, A.H.T. Utsunomiya, R. Carvalheiro, H.H.R. Neves, A.M.P. O'Brien, J.F. Garcia, J. Sölkner, and M.V.G.B. da Silva. 2015. Strategies for single nucleotide polymorphism (SNP) genotyping to enhance genotype imputation in Gyr (Bos indicus) dairy cattle: Comparison of commercially available SNP chips. J. Dairy Sci. 98:4969–4989. doi:10.3168/jds.2014-9213.

Carvalheiro, R., S.A. Boison, H.H.R. Neves, M. Sargolzaei, F.S. Schenkel, Y.T. Utsunomiya, A.M.P. O'Brien, J. Sölkner, J.C. McEwan, C.P. Van Tassell, T.S. Sonstegard, and J.F. Garcia. 2014. Accuracy of genotype imputation in Nelore cattle. Genet. Sel. Evol. 46:1–11. doi:10.1186/s12711-014-0069-1.

Chud, T.C.S., R. V. Ventura, F.S. Schenkel, R. Carvalheiro, M.E. Buzanskas, J.O. Rosa, M. de Alvarenga Mudadu, M.V.G.B. da Silva, F.B. Mokry, C.R. Marcondes, L.C.A. Regitano, and D.P. Munari. 2015. Strategies for genotype imputation in composite beef cattle. BMC Genet. 16:1–10. doi:10.1186/s12863-015-0251-7.

Daetwyler, H.D., M.P.L. Calus, R. Pong-Wong, G. de los Campos, and J.M. Hickey. 2013. Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. Genetics 193:347–365. doi:10.1534/genetics.112.147983.

García-Ruiz, A., F.J. Ruiz-Lopez, G.R. Wiggans, C.P. Van Tassell, and H.H. Montaldo. 2015. Effect of reference population size and available ancestor genotypes on imputation of Mexican Holstein genotypes. J. Dairy Sci. 98:3478–3484. doi:10.3168/jds.2014-9132.

Göktaş, A., and Ö. İşçi. 2011. A comparison and normality test of some measures of association via simulation for rectangular doubly ordered cross tables. Metod. Zv. 8:17–37. doi:10.5897/SRE11.1283.

Gu, Z., L. Gu, R. Eils, M. Schlesner, and B. Brors. 2014. circlize implements and enhances circular visualization in R. Bioinformatics 30:2811–2812. doi:10.1093/bioinformatics/btu393.

Heidaritabar, M., M.P.L. Calus, A. Vereijken, M.A.M. Groenen, and J.W.M. Bastiaansen. 2015. Accuracy of imputation using the most common sires as reference population in layer chickens. BMC Genet. 16:1–14. doi:10.1186/s12863-015-0253-5.

Hickey, J.M., J. Crossa, R. Babu, and G. de los Campos. 2012. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52:654–663. doi:10.2135/cropsci2011.07.0358.

Hill, W.G., and A. Robertson. 1968. Linkage disequilibrium in finite populations. TAG Theor. Appl. Genet. 38:226–231. doi:10.1007/bf01245622.

Jattawa, D., M.A. Elzo, S. Koonawootrittriron, and T. Suwanasopee. 2016. Imputation accuracy from low to moderate density single nucleotide polymorphism chips in a Thai multibreed dairy cattle population. Asian-Australasian J. Anim. Sci. 29:464–470. doi:10.5713/ajas.15.0291.

Khatkar, M.S., G. Moser, B.J. Hayes, and H.W. Raadsma. 2012. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. BMC Genomics 13. doi:10.1186/1471-2164-13-538.

Larmer, S.G., M. Sargolzaei, and F.S. Schenkel. 2014. Extent of linkage disequilibrium, consistency of gametic phase, and imputation accuracy within and across Canadian dairy breeds. J. Dairy Sci. 97:3128–3141. doi:10.3168/jds.2013-6826.

Misztal, I., S. Tsuruta, D.A.L. Lourenco, I. Aguilar, A. Legarra, and Z.G. Vitezica. 2014. Manual for BLUPF90 family of programs.

Nicolazzi, E.L., S. Biffani, F. Biscarini, P. Orozco Ter Wengel, A. Caprera, N. Nazzicari, and A. Stella. 2015. Software solutions for the livestock genomics SNP array revolution. Anim. Genet. 46:343–353. doi:10.1111/age.12295.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly, and P.C. Sham. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. 81:559–575. doi:10.1086/519795.

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Salem, M.M.I., G. Thompson, S. Chen, A. Beja-Pereira, and J. Carvalheira. 2018. Linkage disequilibrium and haplotype block structure in Portuguese holstein cattle. Czech J. Anim. Sci. 63:61–69. doi:10.17221/56/2017-CJAS.

Sargolzaei, M., J.P. Chesnais, and F.S. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15. doi:10.1186/1471-2164-15-478.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions.. J. Dairy Sci. 91:4414–23. doi:10.3168/jds.2007-0980.

VanRaden, P.M., D.J. Null, M. Sargolzaei, G.R. Wiggans, M.E. Tooker, J.B. Cole, T.S. Sonstegard, E.E. Connor, M. Winters, J.B.C.H.M. van Kaam, A. Valentini, B.J. Van Doormaal, M.A. Faust, and G.A. Doak. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. J. Dairy Sci. 96:668–678. doi:10.3168/jds.2012-5702.

Vanraden, P.M., J.R. O'Connell, G.R. Wiggans, and K.A. Weigel. 2011. Genomic evaluations with many more genotypes. Genet. Sel. Evol. 43:1–11. doi:10.1186/1297-9686-43-10.

Ventura, R. V., D. Lu, F.S. Schenkel, Z. Wang, C. Li, and S.P. Miller. 2014. Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle. J. Anim. Sci. 92:1433–1444. doi:10.2527/jas.2013-6638.

Ventura, R. V., S.P. Miller, K.G. Dodds, B. Auvray, M. Lee, M. Bixley, S.M. Clarke, and J.C. McEwan. 2016. Assessing accuracy of imputation using different SNP panel densities in a multi-breed sheep population. Genet. Sel. Evol. 48:1–20. doi:10.1186/s12711-016-0244-7.

Wang, Y., G. Lin, C. Li, and P. Stothard. 2016. Genotype Imputation Methods and Their Effects on Genomic Predictions in Cattle. Springer Sci. Rev. 4:79–98. doi:10.1007/s40362-017-0041-x.

Zhou, L., B. Heringstad, G. Su, B. Guldbrandtsen, T.H.E. Meuwissen, M. Svendsen, H. Grove, U.S. Nielsen, and M.S. Lund. 2014. Genomic predictions based on a joint reference population for the Nordic Red cattle breeds. J. Dairy Sci. 97:4485–4496. doi:10.3168/jds.2013-7580.

Zimin, A. V., A.L. Delcher, L. Florea, D.R. Kelley, M.C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C.P. Van Tassell, T.S. Sonstegard, G. Marçais, M. Roberts, P.

Subramanian, J.A. Yorke, and S.L. Salzberg. 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol. 10. doi:10.1186/gb-2009-10-4-r42.

CHAPTER 3

Autoregressive single-step test-day model for genomic evaluations of Portuguese Holstein cattle

Alessandra Alves Silva[1], Delvan Alves Silva[1], Fabyano Fonseca Silva[1], Cláudio Napolis Costa[2], Paulo Sávio Lopes[1], Alexandre Rodrigues Caetano[3], Gertrude Thompson[4,5], Julio Carvalheira[4,5]

[1]*Department of Animal Science, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil*
[2]*Embrapa Dairy Cattle, Juiz de Fora, Minas Gerais, Brazil*
[3]*Embrapa Genetic Resources & Biotechnology, Brasília, Distrito Federal, Brazil*
[4]*Research Center in Biodiversity and Genetic Resources (CIBIO-InBio), University of Porto, Vairão, Porto, Portugal*
[5]*Institute of Biomedical Sciences Abel Salazar (ICBAS), University of Porto, Porto, Portugal*

Paper published in Journal of Dairy Science, 2019, doi: 10.3168/jds.2018-15191.

3.1. Abstract

The multiple lactations autoregressive test-day (AR) is the adopted model for the national genetic evaluation of dairy cattle in Portugal. Under this model, the animals' permanent environment effects are assumed to follow a first order autoregressive process as a long (auto-correlations between parities) and a short (auto-correlations between test-day within lactation) terms. Given the relevance of genomic prediction in dairy cattle, it is essential to include marker information in national genetic evaluations. In this context, we aimed to evaluate the feasibility of applying the single-step GBLUP to analyze milk yield using the AR model in Portuguese Holstein cattle. A total of 11,434,294 test-day records from the first 3 lactations collected between 1994 and 2017 and 1,071 genotyped bulls were used in this study. Rank correlations and differences in the reliability among bulls were used to compare the performance of the traditional (A-AR) vs single-step (H-AR) models. Additionally, these two modelling approaches were also applied to reduced data sets with records truncated after 2012 (deleting daughters of tested bulls) to evaluate the predictive ability of the H-AR. Validation scenarios were proposed taking into account young and proven bulls. Average EBV reliabilities (R^2_{ebv}), empirical reliabilities (R_{cor}) and genetic trends predicted from the complete and reduced data sets were used to validate the genomic evaluation. Average R^2_{ebv} for H-AR (A-AR) using the complete data set were 0.52 (0.16) and 0.72 (0.62) for genotyped bulls with no daughters and bulls with 1 to 9 daughters, respectively. These results showed a R^2_{ebv} increase of 0.10 to 0.36 when the genomic information was included, corresponding to a reduction of up to 43% in the prediction error variance. Considering the three validation scenarios, the inclusion of genomic information improved the average R^2_{ebv} in the reduced data set, which ranged on average, from 0.16 to 0.26 indicating an increase in the predictive ability. Similarly,

R_{cor} increased up to 0.08 between validation tests. The H-AR outperformed A-AR in terms of genetic trends when unproven genotyped bulls were included. The results suggest that the single-step GBLUP AR model is feasible and may be applied to National Portuguese genetic evaluations for milk yield.

Keywords: autoregressive test-day model, Holstein cattle, single-step GBLUP

3.2. Introduction

Genomic selection has become a standard procedure in dairy cattle breeding given its potential to increase selection accuracy and to reduce generation interval. One of the most used methods in genomic selection is based on the multistep procedure (VanRaden, 2008). In this procedure, the first step requires the estimation of traditional EBV. In the second step, daughter yield deviations (**DYD**) or deregressed phenotypes (**DD**) are calculated. In the third step, direct genomic values (**DGV**) are predicted for genotyped animals using DYD or DD as pseudo-phenotypes (VanRaden, 2008; Hayes et al., 2009; VanRaden et al., 2009). Finally, the blending of DGV with parent averages (**PA**) is performed to yield genomic breeding values (**GEBV**) (VanRaden et al., 2009). In general, this procedure has shown some disadvantages, such as the loss of information attributed to pseudo-phenotypes instead true phenotypes. Additionally, this approach is very complex and requires several approximations, which may reduce the accuracy and induce bias on the GEBV (Koivula et al., 2015).

The single-step genomic BLUP (**ssGBLUP**), where phenotypes, pedigree, and genotypes are jointly used to generate GEBV as a direct output, is a suitable alternative to multistep method (Misztal et al., 2009; Aguilar et al., 2010; Christensen and Lund, 2010). In ssGBLUP, the traditional pedigree based relationship matrix (**A**) is replaced

by the **H** matrix that integrates **A** and a genomic relationship matrix (**G**). The ssGBLUP has been successfully applied to Holstein cattle (Lourenco et al., 2014; Koivula et al., 2015; Jattawa et al., 2016) and outperformed the multistep procedure.

In Portugal, the autoregressive test-day (**AR**) model for multiple lactations proposed by Carvalheira et al. (2002) has been routinely used for the national genetic evaluation in dairy cattle. Under this model, the animals' permanent environment are assumed to follow a first order autoregressive process as a long-term (auto-correlations between parities) and a short-term (auto-correlations between test-day within lactations) effects, taking into account the non-genetic correlations due to the cows' repeated performance. Currently, given the relevance of genomic prediction in dairy cattle, it is essential to include dense marker information in national genetic evaluations. Therefore, we aimed to evaluate the feasibility of using ssGBLUP methodology to predict GEBVs for milk yield of Holstein cattle in Portugal under a multiple lactations AR model.

3.3. Materials and methods

Data

A total of 11,434,294 test-day records from the first 3 lactations of Portuguese Holstein cows, calving between 1994 and 2017, were provided by the Portuguese Dairy Cattle Breeders Association (Aveiro, Portugal). The data set was edited according to predefined criteria for genetic analysis with AR models (Carvalheira *et al.*, 2002) and used to calculate genomic predictions (GEBV) using ssGBLUP methodology and the traditional EBV for comparisons. The data consisted of 4,725,673 test-day records from 578,552 cows in the first lactation, 3,910,679 test-day records from 486,177 cows in the second lactation, and 2,797,942 test-day records

from 353,753 cows in the third lactation. To further test the feasibility of the ssGBLUP methodology, a reduced data set with records truncated at 2012 (excluding records of daughters of young bulls) was also used in the statistical comparisons to evaluate the prediction ability and bias of the new methodology. The reduced data consisted of 3,961,378 test-day records from 488,811 cows in the first lactation, 3,286,254 test-day records from 412,141 cows in the second lactation, and 2,369,824 test-day records from 302,196 cows in the third lactation.

A total of 1,081 bulls genotypes were provided by the Portuguese Dairy Cattle Breeders Association. Five hundred and twenty five were genotyped using the Illumina BovineSNP50v2 BeadChip (Illumina Inc., San Diego, CA, USA) with 54,609 markers. The remaining 556 bulls were genotyped with low-density (GeneSeek Genomic Profiler GGP8K, GGP20K, and GGP25K, Neogen Corp., Lincoln, NE, USA), medium-density (Illumina BovineSNP50v1, San Diego, CA, USA), and high-density (GeneSeek Genomic Profiler GGP80K, and GGP150K) SNP chips. The genotypes of these 556 bulls were imputed to the 50Kv2 using FIMPUTE 2.2 software (Sargolzaei et al., 2014). Imputation accuracy was estimated based on the Spearman correlation between imputed and known 50Kv2 genotypes. In general, sample accuracies were higher than 0.96 for all platforms. Quality control analysis retained SNPs and animals with call rates > 0.9, SNPs with minor allele frequency > 0.02 and lower deviations from Hardy-Weinberg equilibrium ($P > 1 \times 10^{-6}$). Errors of sex disagreement were checked by the heterozygosity on the X chromosome and parent–progeny pairs were tested for Mendelian conflict. SNPs with unknown position or located on sex chromosomes were not considered in the analyses. Finally, a total of 38,286 autosomal SNPs and 1,071 bulls were retained for further analyses. The numbers of genotyped bulls by year of birth are shown in Table 1.

Table 1. Frequency distribution of genotyped bulls according to the year of birth

	Bulls					
Birth year	All	With more than 9 daughters	Young	No daughters	Validation	With daughters in 2012 or earlier
1966-1975	13	4	-	2	-	11
1976-1980	34	15	-	4	-	30
1981-1985	43	37	-	-	-	43
1986-1990	68	64	-	2	-	66
1991-1995	94	90	-	-	-	94
1996-2000	99	85	-	7	-	91
2001-2005	164	86	-	29	33	93
2006-2010	279	46	-	91	90	18
2011-2012	158	12	107	-	11	-
2013-2015	119	1	118	-	-	-
Total	1071	440	225	135	134	446

Statistical Modeling

The AR model for multiple lactations proposed by Carvalheira et al., (2002) was used to implement the traditional genetic (**A-AR**) and the single-step genomic (**H-AR**) evaluations for milk yield. In matrix notation, this model may be described as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Za} + \boldsymbol{Wc} + \boldsymbol{Mp} + \boldsymbol{Qt} + \boldsymbol{r},$$

where $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V})$ is the vector of test-day **(TD)** records with covariance matrix **V**, β is the vector of fixed effects (age class at calving nested within herd, and days in milk class nested within herd and lactation); $\boldsymbol{a} \sim N(\boldsymbol{0}, \boldsymbol{A}\sigma_a^2)$, is the vector of random animal additive genetic effects with **A** representing the pedigree-based additive genetic relationship matrix; or $\boldsymbol{a} \sim N(\boldsymbol{0}, \boldsymbol{H}\sigma_g^2)$, in which the combined relationship matrix **H** was constructed including both marker and pedigree information (Aguilar et al., 2010). The variance components σ_a^2 and σ_g^2 represent the additive genetic variance obtained from **A** and **H** matrices, respectively; $\boldsymbol{c} \sim N(\boldsymbol{0}, \boldsymbol{N})$ is the vector of random contemporary group (**HTD**) effects; $\boldsymbol{p} \sim N(\boldsymbol{0}, \boldsymbol{J})$ is the random long-term

environmental effects (**LTE**); $\boldsymbol{t} \sim N(\mathbf{0}, \boldsymbol{S_L})$ is the random short-term environmental effects (**STE**); so that L = 1, 2 or 3 corresponding to first, second and third lactations, respectively; $\boldsymbol{r} \sim N(\mathbf{0}, \boldsymbol{R})$ is the vector of random residual effects. $\boldsymbol{X, Z, W, M}$ and $\boldsymbol{Q}$ are incidence matrices relating observations to fixed and random effects. A first order autoregressive covariance structure was assumed for HTD (within herds), LTE (between parities) and STE (between TD, within lactations) effects. In this context, we have:

$$\boldsymbol{V} = \boldsymbol{ZGZ'} + \boldsymbol{WNW'} + \boldsymbol{MJM'} + \sum_{L=1}^{3} (\boldsymbol{Q}_L \, \boldsymbol{S}_L \, \boldsymbol{Q}'_L) + \boldsymbol{R},$$

Where $\boldsymbol{G} = \boldsymbol{A}\sigma_a^2$ for (A-AR) or $\boldsymbol{G} = \boldsymbol{H}\sigma_g^2$ for (H-AR), and $\boldsymbol{N}$, $\boldsymbol{J}$ and $\boldsymbol{S_L}$ are first order autoregressive covariance structures of the appropriate dimension as follows (for simplicity, let's assume that each herd have 3 HTD levels and cows have 3 TD records in each lactation):

$$\boldsymbol{N} = \sigma_c^2 \begin{bmatrix} 1 & \rho_c & \rho_c^2 \\ & 1 & \rho_c \\ sym & & 1 \end{bmatrix} \otimes \boldsymbol{I}_q,$$

$$\boldsymbol{J} = \sigma_p^2 \begin{bmatrix} 1 & \rho_p & \rho_p^2 \\ & 1 & \rho_p \\ sym & & 1 \end{bmatrix} \otimes \boldsymbol{I}_m,$$

$$\boldsymbol{S_L} = \sigma_{tL}^2 \begin{bmatrix} 1 & \rho_{tL} & \rho_{tL}^2 \\ & 1 & \rho_{tL} \\ sym & & 1 \end{bmatrix} \otimes \boldsymbol{I}_{mL}, \text{ and}$$

$$\boldsymbol{R} = I_{nL}\sigma_{eL}^2,$$

where $\boldsymbol{I}$ is the identity matrix; $\boldsymbol{q}$ is the number of herds; $\boldsymbol{m}$ is the number of cows with records; $\boldsymbol{mL}$ is the number of cows within the $\mathbf{L^{th}}$ lactation; and $\boldsymbol{nL}$ is the number of records in each lactation; σ_c^2 is the HTD variance component, ρ_c is the HTD autocorrelation coefficient, σ_p^2 is the LTE variance component, ρ_p is the LTE autocorrelation coefficient, σ_{tL}^2 is the STE variance component, ρ_{tL} is the STE

autocorrelation coefficient and $\sigma^2_{e_L}$ is the residual variance components for the records of each lactation. The index **L= 1, 2 or 3** correspond, respectively, to first, second and third lactations. More details about the autoregressive covariance structure fitted in the AR model may be found in Carvalheira et al. (1998, 2002).

In the H-AR model, the inverse of the **H** matrix in the mixed model equations is obtained as follows:

$$\boldsymbol{H}^{-1} = \boldsymbol{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \tau(0.95G + 0.05\boldsymbol{A}_{22})^{-1} - \omega\boldsymbol{A}_{22}^{-1} \end{bmatrix},$$

where: **G** is the genomic relationship matrix and $\mathbf{A}_{22}$ is the pedigree-based numerator relationship matrix for genotyped animals. The **G** matrix was obtained by the first method proposed by VanRaden (2008) where the allele frequencies were calculated from the current genotyped animals. Weights for **G** (0.95) and $\mathbf{A}_{22}$ (0.05) can avoid singularity problems (VanRaden, 2008). The τ is the scaling factor for $\mathbf{G}^{-1}$ and ω is the scaling factor for $\mathbf{A}_{22}^{-1}$. The τ value of 1.0 and ω of 0.7 were chosen according to Tsuruta et al. (2011). The $\mathbf{H}^{-1}$ matrix was calculated using the PreGSF90 software (Misztal et al., 2014).

Preliminary analysis indicated that using the complete data set for variance components estimations using MATLAB was not practical because of the excessive memory required to store the Cholasky factor of the Coefficient matrix of the mixed model equations, necessary to get solutions and the log determinant of this matrix in each iteration. For this reason, we opted to build six data sub-sets consisting of 20 herds randomly sampled from the data set with a minimum of 5,000 cows with records spanning across the all period of study and representing all regions of Portugal in order to estimate the variance components and autocorrelations coefficients applying DFREML methodology (Smith and Graser, 1986). Likelihood functions were maximized by the multivariate simplex algorithm (Nelder and Mead, 1965)

programmed in MATLAB. The convergence criterion was defined as 10^{-8}. The occurrence of local maxima was checked by running 6 consecutive cold starts without significant changes in the log-likelihood (up to four decimal places). This process was repeated for each data sub-set.

The reliabilities of the breeding values ($\mathbf{R^2_{ebv}}$) were estimated for each bull as:

$$\boldsymbol{R^2_{ebv} = 1 - \frac{PEV_i}{\sigma^2_a(1 + F_i)}},$$

where **PEV** is the estimated prediction error variance extracted from the diagonal of the inverse of the coefficient matrix for animal ***i***, $\mathbf{1+F_i}$ corresponds to the diagonal of the **A** (**H)** matrix for animal ***i***, and σ^2_a (σ^2_g) is the additive genetic variance.

Statistical Analyses

Rank correlation and differences in the reliability among bulls were used to compare the performance of the A-AR vs H-AR models using the complete data sets. In order to better differentiate the impact of the ssGBLUP on the evaluations, these comparisons were done within specific bull categories: Genotyped (**G0**) or non-genotyped (**NG0**) bulls without daughters; Genotyped (**G1-9**) or non-genotyped (**NG1-9**) bulls with one to nine daughters; Genotyped (**GM9**) or non-genotyped (**NGM9**) bulls with more than nine daughters.

The predictive ability of the H-AR model was based on the methodology proposed by Mäntysaari et al. (2010). The comparisons were based in the results obtained by each model using different data sets ($\text{A-AR}_{\text{complete}}$ vs $\text{A-AR}_{\text{reduced}}$ and $\text{H-AR}_{\text{complete}}$ vs $\text{H-AR}_{\text{reduced}}$), where young bulls had no daughter information in the reduced data sets, but more than 19 daughters in the complete data sets. Three scenarios were defined for validation of the predictive ability of the H-AR model. In the first scenario (Validation 1), 134 genotyped young bulls composed the validation

population, where their daughter (15,256) was excluded from the reduced data set. Taking into account that most of these bulls' dams have no daughters or own performance records, the bulls estimated PA may be biased. To evaluate the impact of a complete pedigree on this estimate, the second scenario (Validation 2) assumed the validation population including only 96 genotyped bulls with Portuguese origin, born between 1981 and 2009, in which all their descendent were removed in the reduced data set. For both these scenarios (Validations 1 and 2), the data sets included all 1,071 genotyped bulls in the pedigree file. In the third scenario (Validation 3), the aim was to evaluate the effect of including non-contributing genotyped bulls (340 genotyped bulls without daughters or relationships on the complete data set). In this scenario, these bulls were excluded from the pedigree file and the 134 young bulls (as in the first scenario) composed the validation population.

To validate the genomic prediction, the DYD computed from the H-$AR_{complete}$ or A-$AR_{complete}$ were regressed on the GEBV (H-$AR_{reduced}$) or on PA (A-$AR_{reduced}$), respectively. This regression model was as follows (Mäntysaari et al., 2010):

$$\boldsymbol{y_c = 1b_0 + b_1 \hat{a}_r + e},$$

where y_c is the DYD of the validation bulls calculated in the complete data set (H-AR or A-AR), b_0 and $b_1$are the intercept and regression coefficients, $\hat{a}_r$ is the bull GEBV or PA calculated in the reduced data set, e is the residual error. The linear regression was weighted by the number of records from their daughters. The $b_1$was used as an indicator of bias in the genomic prediction. The empirical reliability ($\mathbf{R_{cor}}$) was measured as (Meuwissen et al., 2015):

$$\boldsymbol{R_{cor} = Corr(GEBV\ or\ PA, DYD) / \sqrt{R^2_{DYD}}}.$$

The DYD reliability ($\mathbf{R^2_{DYD}}$) was calculated for each bull as:

$$\boldsymbol{R^2_{DYD} = EDC/(EDC + K)},$$

where, EDC is the effective daughter contribution, in which $\mathbf{EDC = K*R^2_{ebv}/ (1 - R^2_{ebv})}$, and $\mathbf{K = (4 - h^2)/h^2}$ (Mäntysaari et al., 2010). Differences of $\mathbf{R^2_{ebv}}$ between models were also analyzed for each validation test.

Genetic trends for genotyped bulls were estimated by regressing the animals' GEBV and EBV on birth year. Bias of predicted genetic trends was assessed by comparing GEBV or PA averages per year from the reduced data set with GEBV and EBV averages per year from the complete data set. These evaluations were standardized by deviating all genetic values from the mean EBV of cows born in 2007.

3.4. Results

Genetic parameter, EBV Reliability and Rank Correlation

The variance components, autocorrelations and heritabilities estimates for milk yield obtained from the A-AR model are shown in Table 2. These estimates were used in all subsequent analyses. The relative magnitude these estimates and autocorrelations were consistent with the literature (Carvalheira et al., 1998, 2002).

Figure 1A shows the individual reliability (R^2_{ebv}) obtained from the two models (A-AR and H-AR) using the complete data set. In general, the R^2_{ebv} obtained from the H-AR was always higher. The greatest increase in R^2_{ebv} was observed in genotyped bulls without daughters (G0) and bulls with less than 10 daughters (G1-9). For G0 bulls, the mean (SD) and minimum R^2_{ebv} were 0.52 (0.07) and 0.21 for H-AR, whereas the traditional A-AR estimates were only 0.16 (0.13) and 0.00, respectively. For G1-9 bulls, the mean (SD) and minimum for H-AR were 0.72 (0.09) and 0.44, compared with 0.62 (0.16) and 0.24 for A-AR, respectively. The difference in R^2_{ebv} for these two categories between models was highly significant ($P < 0.001$). For bulls with more than 9 daughters (GM9), the mean R^2_{ebv} was similar among models but its minimum

was higher for the H-AR than for the A-AR (0.82 vs 0.78). On the other hand, no changes were observed in R^2_{ebv} for non-genotyped bulls (NG1-9 and NGM9). Figure 1B shows the PEV per bull category estimated using the A-AR or H-AR models also using the complete data set. For G0, G1-9, and GM9, the PEV was 43%, 25%, and 8% lower with H-AR than the one obtained with the A-AR model. Again, there no changes was observed in the PEV for non-genotyped bulls across models (NG1-9 and NGM9).

Table 2. Variance components, autocorrelation coefficients and genetic parameters estimated from traditional evaluation (**A-AR** model) for milk yield of Portuguese Holstein cattle

Parameter	Estimates ± SE
Additive genetic variance	10.80 ± 0.2087
Error variance for lactation 1	7.66 ± 0.9331
Error variance for lactation 2	10.65 ± 1.4126
Error variance for lactation 3	11.80 ± 1.4432
LTE[1] variance	<0.001
LTE autocorrelation	<0.001
STE[2] variance for lactation 1	11.25 ± 0.3076
STE variance for lactation 2	23.24 ± 0.5681
STE variance for lactation 3	28.36 ± 0.6120
STE autocorrelation for lactation 1	0.83 ± 0.0036
STE autocorrelation for lactation 2	0.84 ± 0.0031
STE autocorrelation for lactation 3	0.84 ± 0.0035
HTD[3] variance	4.21 ± 0.1471
HTD autocorrelation	0.51 ± 0.0230
heritability for lactation 1	0.32 ± 0.0102
heritability for lactation 2	0.22 ± 0.0076
heritability for lactation 3	0.20 ± 0.0001

[1]LTE: random long-term environmental effects (between parities); [2]STE: random short-term environmental effects (between TD within lactations); [3]HTD: herd test-date.

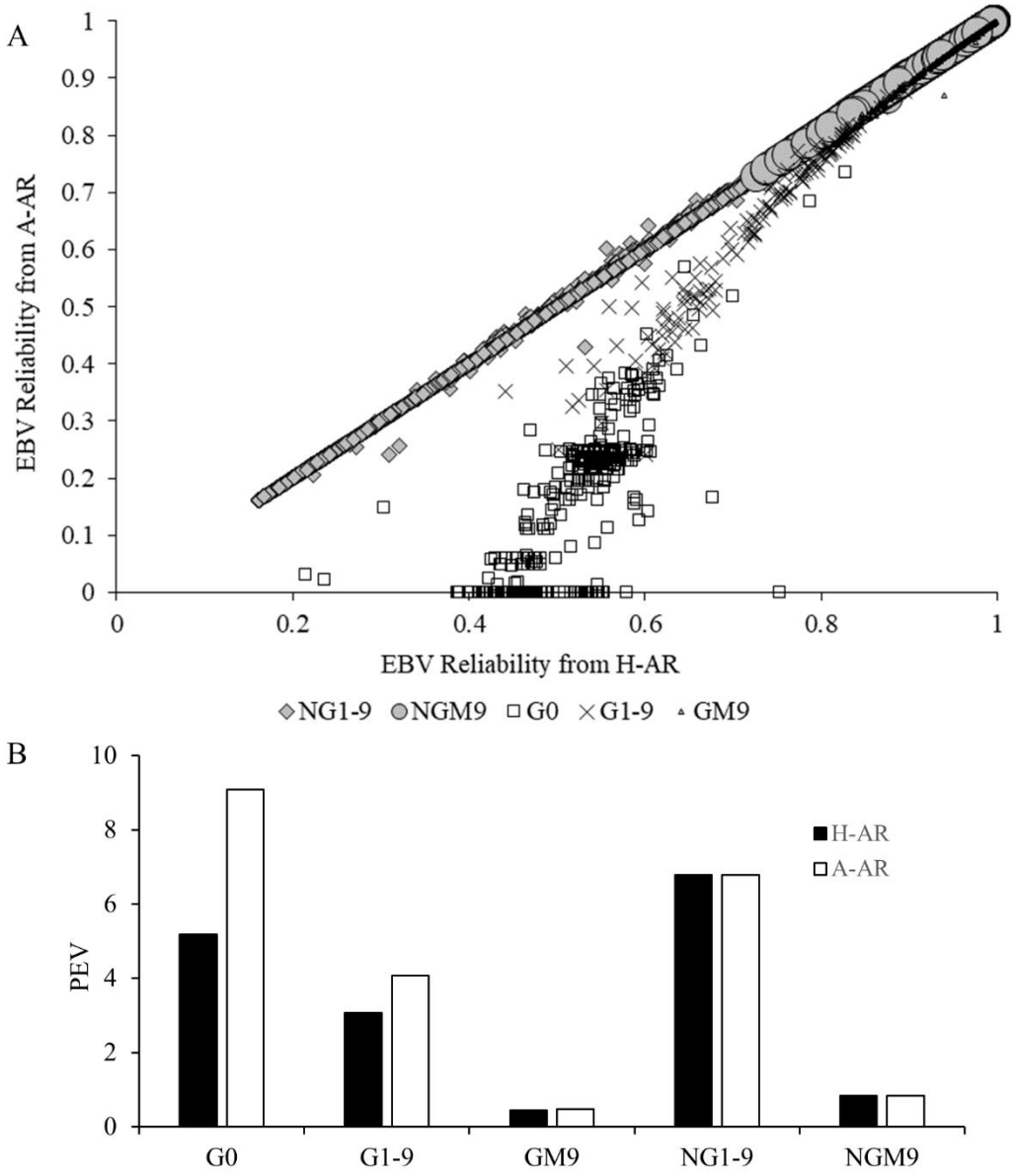

Figure 1. EBV reliabilities (R^2_{ebv}; **A**) and prediction error variance (PEV; **B**) estimated using genomic (H-AR) and the traditional (A-AR) autoregressive models. G0: genotyped bulls with no daughters; G1-9: genotyped bulls with 1 to 9 daughters; NG1-9: non-genotyped bulls with 1 to 9 daughters; GM9: genotyped bulls with more than 9 daughters; NGM9: non-genotyped bulls with more than 9 daughters.

Spearman (rank) correlation coefficients between the animals' genetic merit obtained from H-AR or A-AR models (GEBV vs EBV) across bull categories, are given in Table 3. No relevant changes were observed for both models in the rank of

non-genotyped bulls or genotyped bulls with more than 1 daughter ($r > 0.90$). In contrast, there was a significant re-ranking among the G0 bulls ($r = 0.61$) between the two models.

GEBV Validation

Table 4 shows the average R^2_{ebv} in reduced data sets obtained using the H-AR or A-AR models for the three validation tests. When young bulls (Validation 1 or 3) were the validation population, the R^2_{ebv} obtained from the H-AR increased on average 0.25 compared to the A-AR ($P < 0.001$). Similarly, for the Validation 2 (only Portuguese bulls), the R^2_{ebv} obtained from the H-AR increased on average 0.16.

The R_{cor} obtained for GEBV (H-AR) or EBV (A-AR) for the three validation tests are also shown in Table 4. For Validation 1 and 2, there were no expressive changes in R_{cor} between models. On the other hand, when non-contributing genotyped bulls were excluded from data sets (Validation 3), the R_{cor} for H-AR increased on average 0.08 compared with the A-AR model. Removing the non-contributing genotyped bulls (Validation 1 vs Validation 3) had no effect on the R_{cor} for the A-AR. On the other hand, the H-AR model showed an average improvement of 0.07. The regression coefficients ($\mathbf{b_1}$), used as an indicator of bias in the genomic prediction are also in Table 4. The b_1 were lower than expected for all validation tests, except when the GEBV or PA from Portuguese bulls was tested (Validation 2). The H-AR model was 0.11 and 0.18 less biased than the A-AR model for Validation 1 and 3, respectively. Nevertheless, for validation 2, the $\mathbf{b_1}$ values were within ±15% from optimal value (Tsuruta et al. 2011) for both models (H-AR or A-AR).

Table 3. Number of bulls by category and rank correlation between (G)EBVs obtained from the genomic and the traditional autoregressive models

Category[1]	Number of bulls	Rank correlation
	Genotyped bulls	
G0	360	0.61
G1-9	137	0.94
GM9	574	0.99
	Non-genotyped bulls	
NG0	5000	0.98
NG1-9	21317	0.99
NGM9	4420	0.99

[1]Genotyped (G0) or non-genotyped (NG0) bulls with no daughters; Genotyped (G1-9) or non-genotyped (NG1-9) bulls with 1 to 9 daughters; Genotyped (GM9) or non-genotyped (NGM9) bulls with more than 9 daughters.

Table 4. Mean (SD) of reliabilities (R^2_{ebv}), empirical reliabilities (R_{cor}) and regression coefficients (b_1) for three validation tests, obtained from the genomic (H-AR) and the traditional (A-AR) autoregressive models using the reduced data set. R_{cor} (b_1) were correlated (regressed) with the DYDs obtained from the respective models using the complete data set

Test[1]	H-AR			A-AR		
	R^2_{ebv}	R_{cor}	b_1	R^2_{ebv}	R_{cor}	b_1
Validation 1	0.52 (0.06)	0.17	0.36	0.26 (0.10)	0.16	0.25
Validation 2	0.52 (0.06)	0.82	1.15	0.36 (0.11)	0.79	0.99
Validation 3	0.51 (0.06)	0.24	0.43	0.26 (0.10)	0.16	0.25

[1]Validation 1: 1071 genotyped bulls included in pedigree and 134 younger bulls in validation test; Validation 2: 1071 genotyped bulls included in pedigree and 96 Portuguese bulls in validation test; Validation 3: 731 genotyped bulls included in pedigree (non-contributing genotyped bulls excluded) and 134 younger bulls in validation test.

Figure 2 depicts the genetic progress for bulls born between 1985 and 2011 where the GEBV and EBV were predicted using the H-AR and A-AR models, either with the complete, or the reduced data sets. The graphs are arranged to show the progress of all genotyped bulls (Figure 2A, C and E) versus only genotyped bulls with more than 9 daughters (Figure 2B, D and F) in each of the validation scenarios:

Validation 1 (Panels A and B), Validation 2 (Panels C and D) and Validation 3 (Panels E and F). Except when Portuguese bulls composed the validation test (Panels C and D), bias in predicted genetic trend was observed in all validation tests by visual inspection.

3.5. Discussion

Our study evaluated the feasibility of using the ssGBLUP methodology under an AR model for the Portuguese Holstein cattle taking into account the relatively low number of genotyped bulls available. Comparisons were based on differences between individual EBV reliabilities, rank correlation, predictive ability and genetic trends.

The additive genetic relationship between the animals affected the individual reliability of genomic predictions, particularly in genotyped bulls with few or no daughters. The improvement of the R^2_{ebv} ranged from 0.10 to 0.36 when including the genomic information in the analysis (Figure 1A), a direct consequence of the reduction observed (up to 43%) in PEV (Figure 1B). This finding is supported by previous studies in which the genomic evaluation was more beneficial for young animals with low accuracy on a traditional EBV evaluation (Hayes et al., 2009; VanRaden et al., 2009). Similar results were observed by Forni et al. (2011) and Oh et al. (2017). These authors also used a small number of genotyped animals in the population (e.g. 1,038 and 1,989 pigs, respectively) and, as in the present study, found higher estimates of reliabilities only for genotyped animals. On the other hand, Guo et al. (2015) using 3,445 genotyped animals, reported an effective improvement in the reliabilities also for non-genotyped animals. Moreover, in our study there was also a significant re-ranking among bulls with no daughters ($r = 0.61$) when genomic information was included (Table 3).

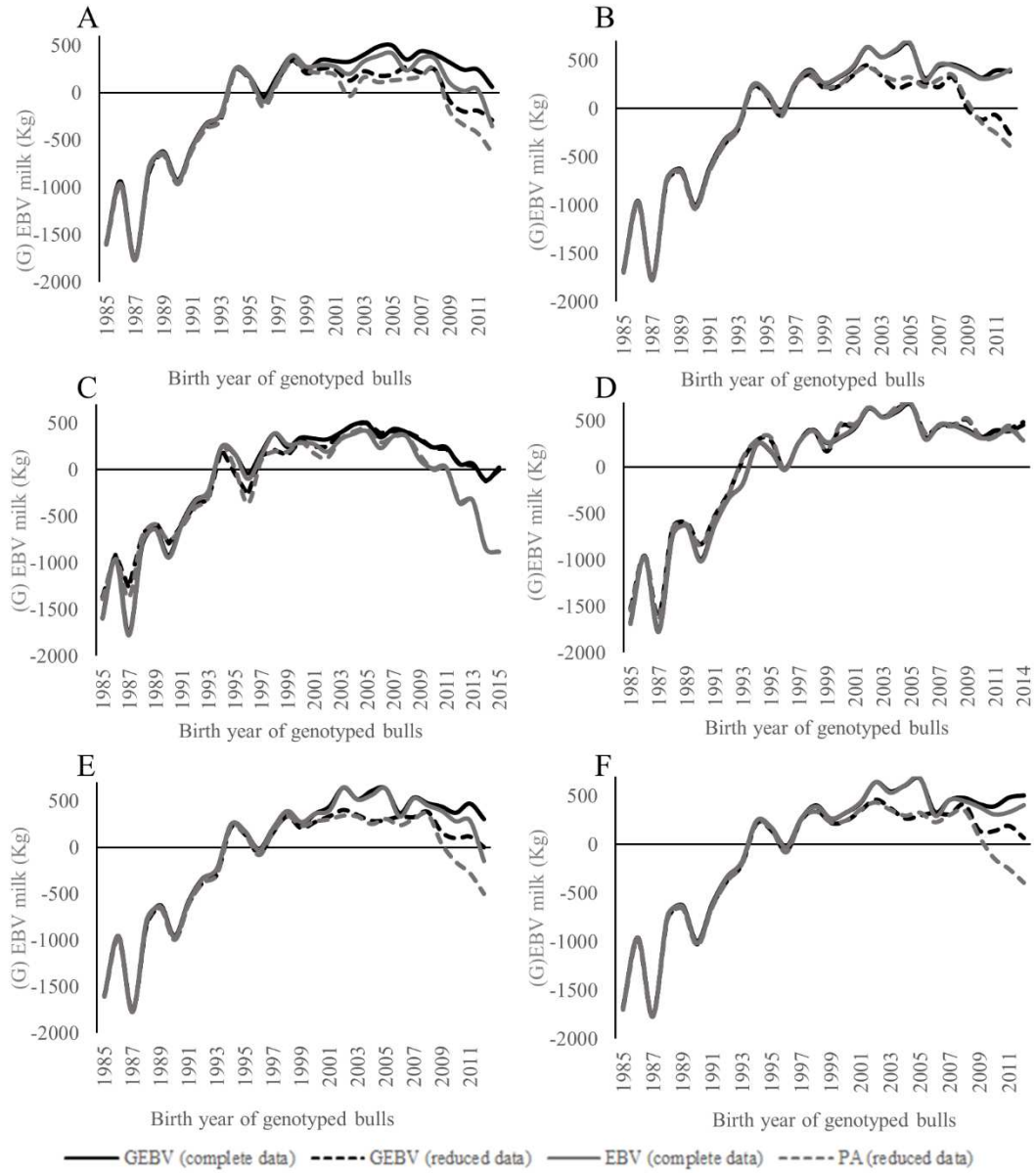

Figure 2. Bulls genetic trends for the 3 scenarions of validation tests. On the left (Panels A, C and E), shows the genetic trends for all bulls (genotyped or not) for each validation test. On the right (Panels B, D and F), shows the genetic trends for bulls with more than 9 daughters (genotyped or not) for each validation test. Validation 1 (Panels A and B) - 134 younger bulls in validation test; Validation 2 (Panels C and D) - 96 Portuguese bulls in validation test; Validation 3 (Panels E and F) - 134 younger bulls in validation test but excluding 340 non-contributing genotyped bulls from the pedigree.

In terms of predictive ability, the inclusion of genomic information improved the R^2_{ebv} from 0.16 to 0.26 for all validation scenarios (Table 4). Similarly, the empirical reliability (R_{cor}) increased up to 0.08 (Table 4). As a measure of accuracy, our results of R_{cor}, were in general lower than in other reports (e.g., Koivula et al., 2015; Baba et al., 2016). Using ssGBLUP, these authors included more than 5,000 genotyped animals in their study and found reliabilities ranging from 0.12 to 0.24. VanRaden et al. (2009) showed that reliability values increased linearly with the number of genotyped bulls under GBLUP. Although ssGBLUP is more versatile due to the incorporation of phenotypes of non-genotyped relatives of the selection candidates, the number of genotyped bulls seems to affect the reliabilities values. Also, the work of Lourenco et al. (2014) using ssGBLUP in a population with a relatively small number of genotyped animals (1,648), found an increase in reliability of 0.06, similar to results of the present study. In previous study of Clark et al. (2012), the authors showed that close relationship between the animals in the validation and reference resulted in higher reliability for GEBV. In this context, considerable gains in reliability may also be expected with ssGBLUP with a relatively small number of genotyped animals, since there is a closer relationship between reference and predicted animals. The average relationship within the reference population may also affect the average reliability, as was reported by Pszczola et al., (2012). These authors showed that when the animals in the reference population were related to each other, the reliabilities were also higher, implying that the design of the reference population is also important when genotyped and non-genotyped animals are jointly evaluated using the H matrix.

In agreement with the study of Li et al. (2014), the estimated average R^2_{ebv} were much higher than R_{cor} when young bulls were the validation test population (Table 4). This result occurs because the bulls selected for the validation tests and selection itself

reduce the additive genetic variance and therefore also the empirical reliability of selection. This is consistent with previous studies (Bijma, 2012; Gorjanc et al., 2015), which showed that prediction accuracies or predictive ability are biased downward by selection.

The regression coefficients measuring bias from validation tests in young bulls were smaller than 1.0 (Table 4). A possible reason could be that 85.5% of the bulls were foreigners and only 9.8% of the genotyped bulls were from the Portuguese population. This implies that most of bull dams had no daughters or own phenotypic performance records which may have induced bias on the PA estimation for young bulls. A similar situation was also reported by Pribyl et al. (2014) where, for a small Holstein population the majority of the artificial insemination were made with imported semen with bulls having low and/or only indirect genetic relationships with the population. This interpretation is in line with our results on the validation tests for the Portuguese bulls where the prediction bias was within an acceptable range.

Bias on predicted genetic trends was consistent with the interpretation of bias from the regression analysis (Figure 2). The reason for the bias on the genetic trend could be explained in part by the fact that the reference animals did not have all the information required to trace selection, especially on bull dams, since most of them are from foreign origin with no other source of information. Similar to reports by Ma et al. (2015) and Baba et al. (2016), the bias found on GEBV prediction was lower than in PA. Ma et al., (2015) evaluated several strategies to reduce bias on genetic trend and concluded that the most efficient way is to implement a single-step approach. In addition, the superiority of H-AR over A-AR is well depicted on the genetic progress when unproven genotyped bulls (bulls with 1 to 9 daughters) were included (Figure 2).

We hypothesized that the inclusion of non-contributing genotyped bulls (with no reliability or no daughters in the traditional evaluation) in the pedigree would contribute by identical-by-state (IBS) to improve the relationship matrix and, therefore, to the reliability of the genomic prediction (13% of them were from an older generation with the remaining ones being young bulls). In the traditional evaluation, the inclusion of those bulls has no effect on the predictive ability or bias because of the almost total lack of relationship with the rest of the population. On the other hand, in the genomic prediction, the inclusion of these bulls was more beneficial for themselves if compared to the traditional evaluation (Figure 1), improving their individual R^2_{ebv}, and having only a minor effect on the predictive ability. Previous studies showed that when relatives are in the reference population, the importance of information on distantly related animals may be reduced (Clark et al., 2012; Pocrnic et al., 2017).

In the present study, we analyzed the feasibility of using the ssGBLUP approach with an AR model with a relatively small number of genotyped bulls. The H-AR model provided higher individual EBV reliabilities for young bulls. In terms of predictive ability, bias occurred mostly due to the lack of phenotypic information on bull dams (foreign origin). In general, H-AR promoted higher average EBV, better empirical reliabilities and reduced the bias when compared with the traditional evaluations.

3.6. Conclusions

These results suggest that the ssGBLUP methodology applied to AR models is feasible and may be advantageous to the Portuguese National genetic evaluations. With the anticipated increase in the number of genotyped animals (for example by including females), it is expected that the H-AR will provide even higher reliabilities

especially for the young stock, thus contributing to the improvement of the genetic progress of the Portuguese dairy cattle population.

3.7. Acknowledgments

The authors acknowledge Portuguese Dairy Cattle Breeders Association (Aveiro, Portugal) and Embrapa Dairy Cattle (Juiz de Fora, Brazil) for providing data for this study. This study was partially financed by Coordination for the Improvement of Higher Education Personnel and Portuguese National Funding Agency for Science, Research and Technology (CAPES/FCT, nº 99999.008462/2014-03 and 88887.136171/2017-00), and National Council of Technological and Scientific Development (CNPq 465377/2014-9 - PROGRAMA INCT).

3.8. References

Aguilar, I., I. Misztal, D.L. Johnson, A. Legarra, S. Tsuruta, and T.J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score1. J. Dairy Sci. 93:743–752. http://dx.doi:10.3168/jds.2009-2730.

Baba, T., Y. Gotoh, S. Yamaguchi, S. Nakagawa, H. Abe, Y. Masuda, and T. Kawahara. 2016. Application of single-step genomic best linear unbiased prediction with a multiple-lactation random regression test-day model for Japanese Holsteins. Anim. Sci. J. 88:1226–1231. http://dx.doi:10.1111/asj.12760.

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

Carvalheira, J., E.J. Pollak, R.L. Quaas, and R.W. Blake. 2002. An autoregressive repeatability animal model for test-day records in multiple lactations. J. Dairy Sci. 85:2040–2045. http://dx.doi:10.3168/jds.S0022-0302(02)74281-1.

Carvalheira, J.G., R.W. Blake, E.J. Pollak, R.L. Quaas, and C. V. Duran-Castro. 1998. Application of an autoregressive process to estimate genetic parameters and breeding values for daily milk yield in a tropical herd of Lucerna cattle and in United States Holstein herds. J. Dairy Sci. 81:2738–2751. http://dx.doi:10.3168/jds.S0022-0302(98)75831-X.

Christensen, O.F., and M.S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42:1–8. http://dx.doi.org/10.1186/1297-9686-42-2.

Clark, S.A., J.M. Hickey, H.D. Daetwyler, and J.H.J. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock beeding schemes. Genet. Sel. Evol. 44:1–9. http://dx.doi.org/10.1186/1297-9686-44-4.

Forni, S., I. Aguilar, and I. Misztal. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet. Sel. Evol. 43:1–7. http://dx.doi:10.1186/1297-9686-43-1.

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

Guo, X., O.F. Christensen, T. Ostersen, Y. Wang, M.S. Lund, and G. Su. 2015. Improving genetic evaluation of litter size and piglet mortality for both genotyped and nongenotyped individuals using a single-step method1. J. Anim. Sci. 93:503–512. http://dx.doi:10.2527/jas.2014-8331.

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

Jattawa, D., M.A. Elzo, S. Koonawootrittriron, and T. Suwanasopee. 2016. Genomic-polygenic and polygenic evaluations for milk yield and fat percentage using random regression models with Legendre polynomials in a Thai multibreed dairy population. Livest. Sci. 188:133–141. http://dx.doi:10.1016/j.livsci.2016.04.019.

Koivula, M., I. Strandén, J. Pösö, G.P. Aamand, and E.A. Mäntysaari. 2015. Single-step genomic evaluation using multitrait random regression model and test-day data. J. Dairy Sci. 98:2775–2784. http://dx.doi:10.3168/jds.2014-8975.

Li, X., S. Wang, J. Huang, L. Li, Q. Zhang, and X. Ding. 2014. Improving the accuracy of genomic prediction in Chinese Holstein cattle by using one-step blending. Genet. Sel. Evol. 46:1–5. http://dx.doi:10.1186/s12711-014-0066-4.

Lourenco, D.A.L., I. Misztal, S. Tsuruta, I. Aguilar, E. Ezra, M. Ron, A. Shirak, and J.I. Weller. 2014. Methods for genomic evaluation of a relatively small genotyped dairy population and effect of genotyped cow information in multiparity analyses. J. Dairy Sci. 97:1742–1752. http://dx.doi:10.3168/jds.2013-6916.

Ma, P., M.S. Lund, U.S. Nielsen, G.P. Aamand, and G. Su. 2015. Single-step genomic model improved reliability and reduced the bias of genomic predictions in Danish Jersey. J. Dairy Sci. 98:9026–9034. http://dx.doi:10.3168/jds.2015-9703.

Mäntysaari, E., Z. Liu, and P. VanRaden. 2010. Interbull Validation Test for Genomic Evaluations. Interbull Bull. No. 41 17–22.

Meuwissen, T.H.E., M. Svendsen, T. Solberg, and J. Ødegård. 2015. Genomic predictions based on animal models using genotype imputation on a national scale in Norwegian Red cattle. Genet. Sel. Evol. 47:79. http://dx.doi:10.1186/s12711-015-0159-8.

Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92:4648–4655. http://dx.doi:10.3168/jds.2009-2064.

Misztal, I., S. Tsuruta, D.A.L. Lourenco, I. Aguilar, A. Legarra, and Z.G. Vitezica. 2014. Manual for BLUPF90 family of programs. http://nce.ads.uga.edu/wiki/doku.php?id=application_programs.

Nelder, J.A., and R. Mead. 1965. A simplex method for function minimization. Comput. J. 7:308–313. https://dx.doi.org/10.1093/comjnl/7.4.308.

Oh, J.D., C.S. Na, and K. Do Park. 2017. Validation of selection accuracy for the total number of piglets born in Landrace pigs using genomic selection. Asian-Australasian J. Anim. Sci. 30:149–153. https://dx.doi:10.5713/ajas.16.0394.

Pocrnic, I., D.A.L. Lourenco, H.L. Bradford, C.Y. Chen, and I. Misztal. 2017. Technical note: Impact of pedigree depth on convergence of single-step genomic BLUP in a purebred swine population. J. Anim. Sci. 95:3391–3395. https://dx.doi:10.2527/jas.2017.1581.

Pribyl, J., J. Bauer, P. Pesek, J. Pribylová, L. Vostrý, and L. Zavadilová. 2014. Domestic and Interbull information in the single step genomic evaluation of Holstein milk production. Czech. J. Anim. Sci. 9:409–415.

Pszczola, M., T. Strabel, J.A.M. van Arendonk, and M.P.L. Calus. 2012. The impact of genotyping different groups of animals on accuracy when moving from traditional to genomic selection. J. Dairy Sci. 95:5412–5421. http://dx.doi.org/10.3168/jds.2012-5550.

Sargolzaei, M., J.P. Chesnais, and F.S. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15. https://dx.doi:10.1186/1471-2164-15-478.

Smith, S.P., and H.-U. Graser. 1986. Estimating Variance Components in a Class of Mixed Models by Restricted Maximum Likelihood. J. Dairy Sci. 69:1156–1165. https://dx.doi:10.3168/jds.S0022-0302(86)80516-1.

Tsuruta, S., I. Misztal, I. Aguilar, and T.J. Lawlor. 2011. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. J. Dairy Sci. 94:4198–4204. https://dx.doi:10.3168/jds.2011-4256.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions.. J. Dairy Sci. 91:4414–23. https://dx.doi:10.3168/jds.2007-0980.

VanRaden, P.M., C.P. Van Tassell, G.R. Wiggans, T.S. Sonstegard, R.D. Schnabel, J.F. Taylor, and F.S. Schenkel. 2009. Invited review : Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16–24. https://dx.doi:10.3168/jds.2008-1514.

CHAPTER 4

Genome wide association studies using autoregressive test-day model for milk related traits in Portuguese Holstein cattle

Alessandra Alves Silva[1], Fabyano Fonseca Silva[1], Delvan Alves Silva[1], Hugo Teixeira Silva[1], Cláudio Napolis Costa[2], Paulo Sávio Lopes[1], Renata Veroneze[1], Gertrude Thompson[3,4], Julio Carvalheira[3,4]

[1]*Department of Animal Science, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil*

[2]*Embrapa Dairy Cattle, Juiz de Fora, Minas Gerais, Brazil*

[3]*Research Center in Biodiversity and Genetic Resources (CIBIO-InBio), University of Porto, Vairão, Porto, Portugal*

[4]*Institute of Biomedical Sciences Abel Salazar (ICBAS), University of Porto, Porto, Portugal*

4.1. Abstract

This study implemented the weighted ssGWAS using an autoregressive test-day model for evaluating milk, fat, and protein yields, and somatic cell score (SCS) in Holstein cattle. In addition, this study was also focused in the analysis of the differences in QTL regions and important genes found using either data from the first 3 lactations (LAC3), or from the first lactation (LAC1) in the ssGWAS analyses. Comparisons were based on the genetic variance explained by genomic windows and sets of important genes. A total of 30, 12, 22 and 3 windows that explained 1% or more of the total genetic variance were identified for milk, fat, and protein yields, and SCS, respectively, using LAC3. For milk, fat, and protein yields, and SCS, the mean (maximum) proportion of explained genetic variance were 2.42 (4.16), 2.10 (3.27), 1.78 (2.76), and 1.15% (1.28%), respectively. A total of 51, 5, 24, and 4 genes were associated in relevant QTL regions for milk, fat, and protein yields, and SCS, respectively. The gene that explained the highest proportion of genetic variance for milk, fat, and protein yields was the *CACNA2D1* gene on BTA4. This gene has been associated to the calcium transport in the mammary gland. For SCS, the *LOC104973995*, *TRNAG-CCC* genes on BTA14 explained the highest proportion of genetic variance. In addition, we propose that more phenotypic information, such as in the LAC3 data set, would contribute to access genes with major effects in all lactations simultaneously, providing a better understanding of the genetic architecture of these traits. When using only LAC1, the candidate gene is representing genomic regions with substantial effect at only a certain stage of the animal's productive life. Although most of the genes found in our study were present in LAC3 or LAC1 for milk, fat and protein yields, the genes that accounted for most of the variance in LAC3 or LAC1 were not always the same. In SCS for example, no common important genes were found between LAC3 and

LAC1. The ssGWAS methodology using the AR model was feasible and the results presented in this study helped to provide a better understanding of the genetic architecture of milk related traits in Portuguese Holstein cattle.

Keywords: gene function; multiple lactations, QTL region; SNP windows; Wss-GWAS

4.2. Introduction

Test-day (TD) models from multiple lactations have been routinely used in genetic evaluation in dairy cattle breeding. These models take into account the genetic and environmental effects for each TD, which allows predicting more accurate EBVs compared those based on the cumulative 305-d models. In Portugal, the autoregressive test-day (AR) model for multiple lactations proposed by Carvalheira et al. (2002) has been routinely used for the national genetic evaluation in dairy cattle. Under this model, the animals' permanent environment are assumed to follow a first order autoregressive process as a long-term (auto-correlations between parities) and a short-term (auto-correlations between test-day within lactations) effects, taking into account the non-genetic correlations due to the cows' repeated performance. Currently, given the relevance of genomic prediction in dairy cattle, it is essential to include genomic information in the national genetic evaluations. A recent study has reported applications of AR models to predict GEBVs (Silva et al., 2019), but no genome-wide association studies (GWAS) have been found using GEBVs predicted by this model.

Most GWAS studies have been done using the total accumulated yield (Nayeri et al., 2016; Yue et al., 2017). These studies have usually used de-regressed EBVs as pseudo-phenotypes, and fitted the SNPs (one at a time) as a fixed effect. However, the

use of de-regressed EBVs may lead to biases and losses of accuracy (Vitezica et al., 2011).

The weighted single-step GWAS (ssGWAS), proposed by Wang et al. (2012), is a suitable alternative to traditional GWAS. In this method, all genotypes, phenotypes and pedigree information are jointly considered to predict GEBVS, which are then used in an iterative approach to update GEBVs and SNP solutions. The weighted ssGWAS allows unequal variances for SNPs, which may give more weight to SNPs that are in high LD with a causal mutation or associated with QTL with a relatively large effect (Wang et al., 2012).

The weighted ssGWAS has been successfully applied to milk related traits (Iung et al., 2019; Zhou et al., 2019), revealing important candidate genes and QTLs regions associated with the traits. Nevertheless, there are no reports in the literature about weighted ssGWAS for milk related traits using AR models. Thus, our aim with this study was to identify QTL regions and important genes associated with milk, fat, protein yields, and somatic cell score (SCS) in Holstein cattle, using the weighted ssGWAS methodology under a AR test-day model. In addition, differences in QTL regions and important genes found using either, the first 3 lactations (LAC3), or just the first lactation (LAC1) in the weighted ssGWAS analyses were also studied.

4.3. Materials and Methods

Phenotypic and genotypic data

A total of 11,434,294 TD records from the first 3 (LAC3) lactations (with 4,725,673 TD records corresponding to the first lactation - LAC1) of milk, fat and protein yields and somatic cell counts (SCC) of Portuguese Holstein cows were provided by the Portuguese Dairy Cattle Breeders Association. The data were collected

between 1994 and 2017. The SCC was log transformed in somatic cell scores (SCS) as: SCS = log2 (SCC / 100) + 3. The data set were edited according to predefined criteria for genetic analysis with AR models (Carvalheira et al., 2002).

Data from 1,338 genotyped animals (785 bulls and 553 cows) were used in this study. The bulls were genotyped using different panels: LDv1 (GeneSeek Genomic Profiler, Neogen Corp., Lincoln, NE, USA), 50Kv1 and 50Kv2 (Bovine SNP50v.1 e Bovine SNP50v.2 BeadChips, Illumina, San Diego, CA, USA), 57K (USDA Illumina, San Diego, CA, USA), 77K and HDv3 (GeneSeek Genomic Profiler, Neogen Corp., Lincoln, NE, USA), including 8,610, 54,001, 54,609, 56,947, 76,883 and 139,376 markers, respectively. The cows were genotyped using Ax58K panel (Affymetrix, Santa Clara, CA, USA), which included 57,497 SNP markers. Genotype quality control and imputation procedures to 50Kv2 were implemented and detailed in Chapter 2.

Statistical modeling

The AR model for multiple lactations was used to predict GEBVs used in the weighted ssGWAS for milk, fat, and protein yield, and SCS. In matrix notation, this model may be described as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Za} + \boldsymbol{Wc} + \boldsymbol{Mp} + \boldsymbol{Qt} + \boldsymbol{e},$$

where $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V})$ is the vector of TD records with covariance matrix **V**, β is the vector of fixed effects (i.e., age class at calving nested within herd, and days in milk class nested within herd and lactation); $\boldsymbol{a} \sim N(\mathbf{0}, \boldsymbol{H}\sigma_g^2)$, in which the combined relationship matrix **H** was constructed including both marker and pedigree information (Aguilar et al., 2010). The variance component σ_g^2 represent the additive genetic variance obtained from the **H** matrix; $\boldsymbol{c} \sim N(\mathbf{0}, \boldsymbol{N})$ is the vector of random

contemporary groups (**HTD**) effects; $\boldsymbol{p} \sim N(\mathbf{0}, \boldsymbol{J})$ is the random long-term environmental effects (**LTE**); $\boldsymbol{t} \sim N(\mathbf{0}, \boldsymbol{S_L})$ is the random short-term environmental effects (**STE**); and $e \sim N(\mathbf{0}, \boldsymbol{R_L})$ is the vector of random residual effects with L = 1, 2 or 3 corresponding to first, second and third lactations, respectively. The $\boldsymbol{X, Z, W, M}$ and $\boldsymbol{Q}$ are incidence matrices relating observations to fixed and random effects. A first order autoregressive covariance structure was assumed for HTD (within herds), LTE (between parities) and STE (between TD, within lactations) effects. The models for LAC1 has the same definitions, except the terms that imply more than one lactation (LTE information). In this context, **V** is defined as:

$$\boldsymbol{V = ZKZ' + WNW' + MJM' + \sum_{L=1}^{3}(Q_L\ S_L\ Q_L') + R,}$$

where $\boldsymbol{G = H\sigma_g^2}$ and $\boldsymbol{N}$, $\boldsymbol{J}$ and $\boldsymbol{S_L}$ are the first order autoregressive covariance structures of the appropriate dimension. More details about the autoregressive covariance structure fitted in the AR model may be found in Carvalheira et al. (2002) and Silva et al. (2019).

Usually **H** is computationally demanding to calculate, but its inverse has a simple structure (Aguilar et al., 2010) and was obtained as follows:

$$\boldsymbol{H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \tau(0.95G + 0.05A_{22})^{-1} - \omega A_{22}^{-1} \end{bmatrix}},$$

Where $\mathbf{A^{-1}}$ is the inverse of pedigree-based numerator relationship matrix, **G** is the genomic relationship matrix and $\mathbf{A_{22}}$ is the pedigree-based numerator relationship matrix for genotyped animals. Weights for **G** (0.95) and $\mathbf{A_{22}}$ (0.05) were used to avoid singularity problems (VanRaden, 2008). In addition, the scaling factors τ and ω used to make $\mathbf{G^{-1}}$ (τ) and $\mathbf{A_{22}^{-1}}$ (ω) compatible, were assumed based on pilot validation analysis as 1.0 and 0.7, respectively. The **G** matrix was obtained by the first method proposed by VanRaden (2008) as follow:

$$\mathbf{G} = \frac{\mathbf{ZDZ'}}{2\sum \mathrm{p_i}\,(1-\mathrm{p_i})},$$

where $\mathbf{Z}$ is the matrix containing marker information; $\mathbf{D}$ is a diagonal matrix of weights for SNP variances (initially $\mathbf{D} = \mathbf{I}$); and $\mathrm{p_i}$ is the minor allele frequency of the i-th SNP. The $\mathbf{H^{-1}}$ matrix was calculated using the preGSF90 software (Misztal et al., 2015). Variance components were estimated according to Carvalheira et al. (2002) and Silva et al. (2019).

Estimates of SNP effects and weights for ssGWAS were obtained according to Wang et al. (2012) by following these steps programed in R (R Core Team, 2018):

i. In the first iteration ($\mathrm{t} = 1$): $\mathbf{D} = \mathbf{I}$; $\mathbf{G}_{(\mathrm{t})} = \mathbf{D}_{(\mathbf{t})}\mathbf{Z'}\lambda$, where $\lambda = \frac{1}{\sum_{i=1}^{M} 2p_i(1-p_i)}$, in which M is the number of SNPs;

ii. GEBVs were calculated for the entire dataset using ssGBLUP;

iii. GEBVs were converted to SNP effect estimates ($\hat{u}$): $\hat{u}_{(\mathrm{t})} = \lambda\mathbf{D}_{(\mathrm{t})}\mathbf{Z'}\mathbf{G}^{-\mathbf{1}}_{(\mathrm{t})}\hat{\boldsymbol{a}}_g$, where $\hat{\boldsymbol{a}}_g$ is the GEBV of animals that were also genotyped;

iv. The weight for each SNP to be used in the next iteration was calculated as: $d_{\mathrm{i(t+1)}} = \hat{u}^2_{\mathrm{i(t)}} 2p_{\mathrm{i}}(1-p_{\mathrm{i}})$, where i is the i-th SNP;

v. The SNP weights were normalized to keep the total genetic variance constant: $\mathbf{D}_{(\mathrm{t+1})} = \frac{\mathrm{tr}(\mathbf{D}_{(1)})}{\mathrm{tr}(\mathbf{D}_{(\mathrm{t+1})})}\mathbf{D}_{(\mathrm{t+1})}$; *vi.* $\mathbf{G}_{(\mathrm{t+1})} = \mathbf{ZD}_{(\mathrm{t+1})}\mathbf{Z'}\lambda$ was calculated;

vi. $\mathrm{t} = \mathrm{t} + 1$ and loop to step *ii*. This script is available at the end of this manuscript (Supplementary material).

This procedure was run for three iterations (t = 3) based on accuracy prediction from pilot validation analysis (results not shown) as recommended by Wang et al. (2014). At each iteration, the weights for SNPs were updated (steps *iv* and *v*), and were used to construct the G matrices (step *vi*), update the GEBVs (step *ii*) and, consequently, update the SNP effects (step *iii*). The proportion of genetic variance

explained by the i-th set of consecutive SNPs (i-th SNP window) was calculated as described by Wang et al. (2014):

$$\frac{Var\,(a_i)}{\sigma_a^2} \times 100\% = \frac{Var\,(\sum_{j=1}^{x} \mathbf{Z}_j \widehat{u_j})}{\sigma_a^2} \times 100\%,$$

where $Var(a_i)$ is the variance for the i-th SNP window. SNP windows consisted of a region of x consecutive SNPs located within 100kb, which was defined by an appropriate disequilibrium linkage (LD) level. The LD (r^2) concept was relevant to inform an appropriate window length for ssGWAS analyses. To examine the decay of LD with physical distance and the appropriate LD to create the SNP windows, SNP pairs on the autosomes were sorted based on pair-wise marker distance and the average of each interval was calculated. The σ_a^2 is the total additive genetic variance; $\mathbf{Z}_j$ is the matrix of gene content of the j-th SNP for all individuals and $\widehat{u_j}$ is the effect of the j-th SNP within the i-th window.

SNP windows, important genes, and gene network analyses

The SNP windows that explained 1% or more of the total genetic variance for milk, fat, and protein yields, and SCS were considered as important windows to be further investigated. The threshold of 1% was chosen based on the literature (e.g., Iung et al., 2019).

Based on the start and end positions of each window selected as relevant, positional important genes were identified using the Gene database for Bos taurus available at National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/genome/gdv/?org=bos-taurus). For all important genes identified, a complete research in the literature was performed to identify the relationship between genes and traits under study. In addition, gene network analyses were carried out to describe the biological processes related to the sets of genes

identified, using the ClueGO and CluePedia Cytoscape plug-ins (Bindea et al., 2009, 2013).

4.4. Results

The average heritabilities for milk, fat, protein yields and SCS were 0.25, 0.20, 0,24 and 0.18, respectively. The estimated autocorrelation associated with the non-genetic component were especially important between TD within lactations (STE), in which a strong correlation (> 0.76) between TD were observed for all traits studied. The relative magnitude of these estimates and autocorrelations were consistent with the literature (Carvalheira et al., 2002; Silva et al., 2019).

Figure 1 displays the average LD at given distances. High LD values were observed only at small distances between pairs of SNPs. The LD decays rapidly as distance between SNPs increases. These results are in agreement with those reported by Salem et al. (2018), in which the authors studied the level of LD in Portuguese Holstein cattle. The LD was relevant to inform an appropriate window length to be used in the ssGWAS analyses. When omitting the LD analysis, window's length are empirically assumed, which may decrease the power to detect true QTL regions (Resende et al., 2018). According to Qanbari et al., (2010), a convenient value of LD for association studies was 0.25.

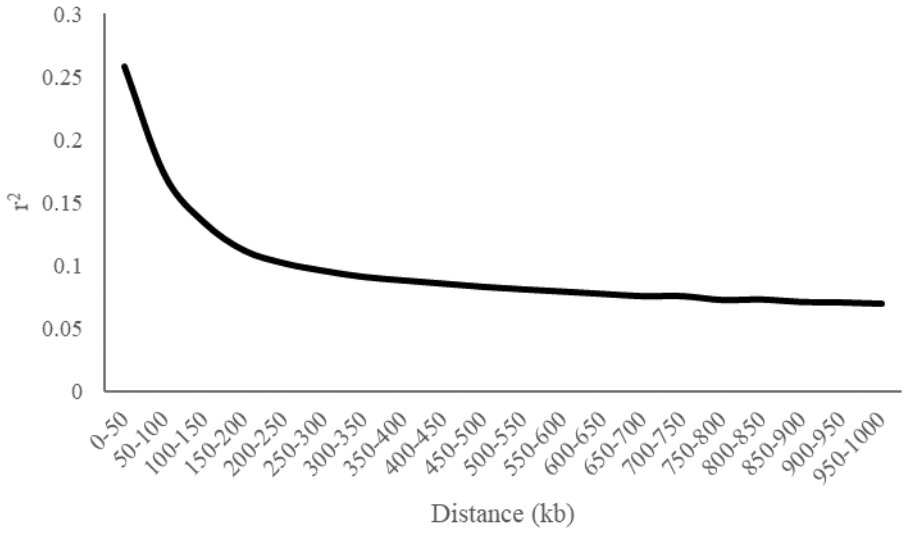

Figure 1. Average linkage disequilibrium (r^2) at given distances (Kb) for Portuguese Holstein cattle.

LAC3

A total of 30, 12, 22, and 3 windows that explained 1% or more of the total genetic variance were identified for milk, fat, and protein yields, and SCS, respectively. In average, a density of 4 SNPs per window was observed. The variance explained by each window along the chromosomes for milk, fat, and protein yields, and SCS are shown in Figure 2. The greatest proportion of genetic variance explained was obtained in milk yield, whereas the lowest was in SCS. For milk, fat, and protein yields, and SCS, the mean, and (maximum) proportion of genetic variance explained were 2.42 (4.16), 2.1 (3.27), 1.78 (2.76), and 1.15% (1.28%), respectively.

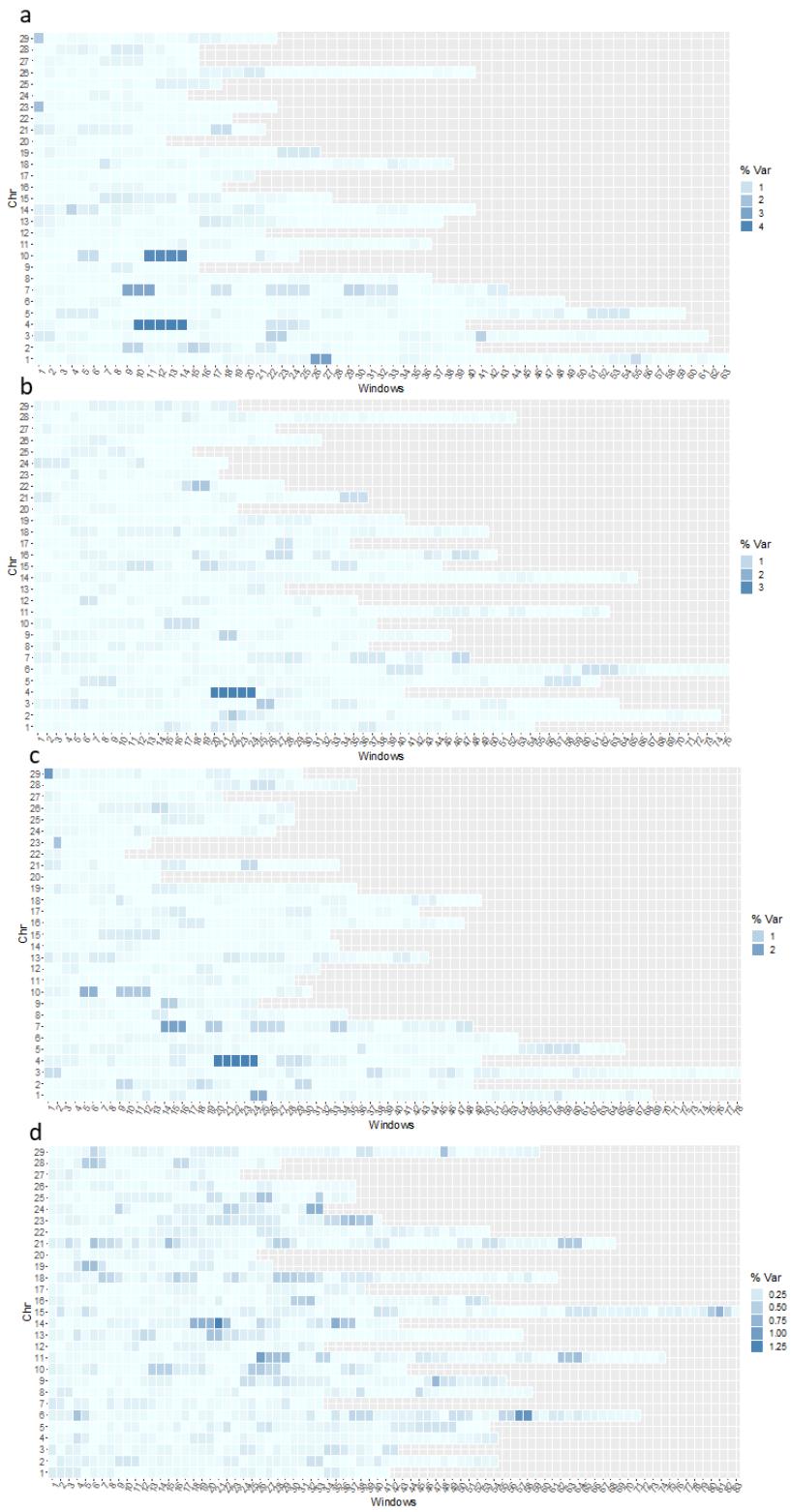

Figure 2. GWAS results of milk trait for the first 3 lactations (LAC3). **a** milk yield, **b** fat yield, **c** protein yield and **d** somatic cell score (SCS). Each dot represents one SNP window of 100 kb. On the y-axis is the chromosomes (Chr).

The QTL regions (where important genes were found) and the variance explained by them for milk, fat, and protein yields, and SCS are described in Table 1. Important genes found in the relevant QTL regions are presented in Table 2. For milk, fat, and protein yields, and SCS a total of 51, 5, 24, and 4 genes were associated in relevant QTL regions (Table 2), which explained 29.99, 8.17, 14.96, and 2.38% of the total genetic variance, respectively (Table 1). Several genes were associated for milk yield due to pattern polygenic observed for this trait. The important gene that explained the highest proportion of genetic variance for milk, fat, and protein yields was the *CACNA2D1* gene on BTA4. For SCS, the *LOC104973995*, *TRNAG-CCC* genes on BTA14 explained the highest proportion of genetic variance (Tables 1 and 2). The gene network of biological processes for these important genes are depicted in Figure 3.

Table 1. QTL regions in which important genes were found to be associated with milk, fat, and protein yields, and somatic cell score (SCS) for first 3 lactations (LAC3) or just the first lactation (LAC1). QTL regions without genes are not shown

Trait	LAC3				LAC1			
	BTA[1]	QTL (region) bp	N SNPs[2]	Var (%)[3]	BTA	QTL (region) bp	N SNPs	Var (%)
Milk	1	87675213-87824971	2	3.1	1	87724971-87879634	2	3.57
	1	143565536-143761195	3	1.08	2	54469795-54638834	3	1.74
	2	54469795-54638834	3	1.32	2	64954490-65119112	6	1.18
	2	64954490-65119112	6	1.4	3	1173441-1360626	4	1.22
	3	98662726-98824663	3	1.45	4	38163706-38347560	4	2.18
	4	38163706-38347560	4	4.16	7	48858617-49049205	3	3.93
	7	48858617-49049205	3	2.89	10	46436647-46621563	4	2.86
	7	88490604-88678942	3	1.31	10	46521563-46715928	4	2.86
	10	20412779-20600545	3	1.1	14	1873292-2052873	3	2.75
	10	46436647-46621563	4	3.85	14	1952873-2104457	2	1.59
	10	46521563-46715928	4	3.82	23	5434630-5597603	3	1.64
	14	1873292-2052873	3	1.37				
	21	27968589-28144869	2	1.06				
	23	5434630-5597603	3	2.08				
Fat	2	64954490-65119112	6	1.28	2	64954490-65119112	6	1.02
	4	38163706-38347560	4	3.26	4	38163706-38347560	4	1.34
	9	76726811-76918290	2	1.09	7	48858617-49049205	3	1.44
	9	76818290-76978003	3	1.08	22	59117318-59287587	3	1.45
	22	59117318-59287587	3	1.46				
Protein	1	87675213-87824971	2	1.72	1	87675213-87824971	2	1.3
	4	38163706-38347560	4	2.76	3	1173441-1360626	4	1.19
	7	48858617-49049205	3	2.03	4	38163706-38347560	4	1.56
	7	66901622-67083026	4	1.09	7	48858617-49049205	3	2.89
	9	76726811-76918290	2	1.05	9	76726811-76918290	2	1.21
	9	76818290-76978003	3	1.01	9	76818290-76978003	3	1.17
	10	20412779-20600545	3	1.67	10	20412779-20600545	3	1.53
	10	46436647-46621563	4	1.17	10	46436647-46621563	4	1.11
	10	46521563-46715928	4	1.17	10	46521563-46715928	4	1.12
	23	5434630-5597603	3	1.29	23	5434630-5597603	3	1.1
SCS	6	103305581-103501359	3	1.1	10	1429316-1608353	3	1.31
	14	13530277-13730187	3	1.28	15	84999720-85199373	4	1.16

[1]BTA Bos Taurus autosome. [2]Number of SNPs within the QTL region. [3]Proportion.

Table 2. Important genes for milk, fat, and protein yields, and somatic cell score (SCS) for first 3 lactations (LAC3) or only the first lactation (LAC1). Genes in bold were reported only in LAC3 or LAC1

Trait	BTA	LAC3	LAC1
Milk	1	*PEX5L, LOC104970945,* ***PRDM15, LO104971061, C2CD2, ZBTB21***	*PEX5L, LOC104970945*
	2	*LOC101908745, NCKAP5*	*LOC101908745, NCKAP5*
	3	***SKINT1, TRNAR-ACG***	***LOC104971407, CREG1, CD247, POU2F1***
	4	*CACNA2D1*	*CACNA2D1*
	7	*SLC25A48, IL9, FBXL21, LECT2, TGFBI,* ***COX7C***	*SLC25A48, IL9, FBXL21, LECT2, TGFBI*
	10	***TBC1D21, LOC104973073, SDR39U1, KHNYN, CBLN3, NYNRIN****, FBXL22, USP3, LOC104973148, CA12, APH1B*	*FBXL22, USP3, LOC104973148, CA12, APH1B*
	14	*MROH1, MIR1839, LOC523023, HGH1, LOC104973958, MAF1, SHARPIN, CYC1, GPAA1, LOC107133095, EXOSC4, OPLAH, SMPD5, SPATC1, TRNAG-UCC, LOC101908059, GRINA, PARP10*	*MROH1, MIR1839, LOC523023, HGH1, LOC104973958, MAF1, SHARPIN, CYC1, GPAA1, LOC107133095, EXOSC4, OPLAH, SMPD5, SPATC1, TRNAG-UCC, LOC101908059, GRINA, PARP10,* ***LOC786966, MIR2309***
	21	***MTMR10, LOC101902644, TRPM1, MIR211***	-
	23	*FAM83B*	*FAM83B*
Fat	2	*NCKAP5*	*NCKAP5*
	4	*CACNA2D1*	*CACNA2D1*
	7	-	***TGFBI, LECT2, SLC25A48, IL9, FBXL21***
	9	***TNFAIP3, PERP***	-
	22	*NUP210*	*NUP210*
Protein	1	*PEX5L, LOC104970945*	*PEX5L, LOC104970945*
	3	-	***LOC104971407, CREG1, CD247, POU2F1***
	4	*CACNA2D1*	*CACNA2D1*
	7	*SLC25A48, IL9, FBXL21, LECT2, TGFBI,* ***GRIA1, LOC100140426***	*SLC25A48, IL9, FBXL21, LECT2, TGFBI*
	9	*TNFAIP3, PERP*	*TNFAIP3, PERP*
	10	*TBC1D21, LOC104973073, SDR39U1, KHNYN, CBLN3, NYNRIN, FBXL22, USP3, LOC104973148, CA12, APH1B*	*TBC1D21, LOC104973073, SDR39U1, KHNYN, CBLN3, NYNRIN, FBXL22, USP3, LOC104973148, CA12, APH1B*
	23	*FAM83B*	*FAM83B*
SCS	6	***PTPN3, MAPK10***	-
	10	-	***EPB41L4A***

14	***LOC104973995, TRNAG-CCC***	-
15	-	***LOC519145, B3GAT1, LOC107133208, LOC100335879, LOC104969826***

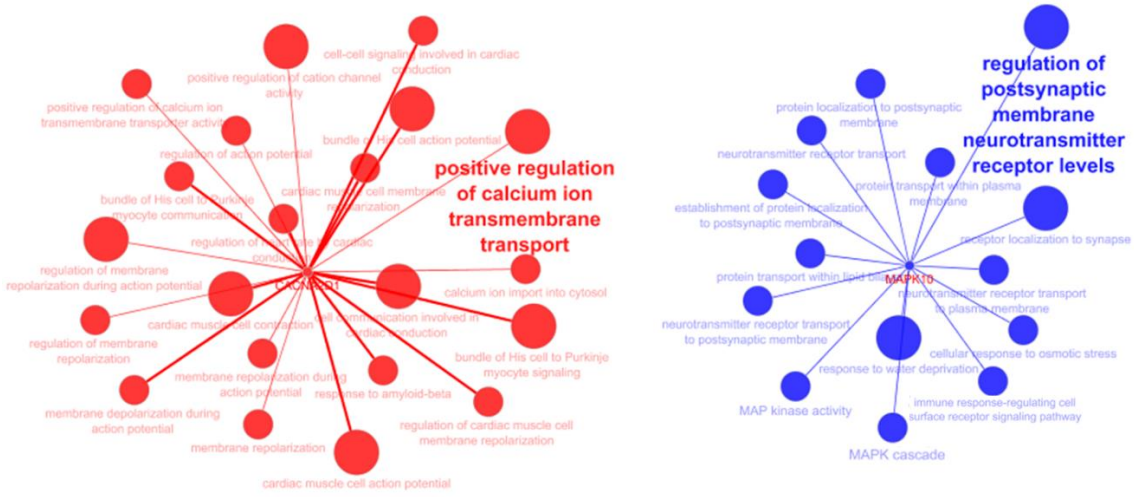

Figure 3. Gene network of biological processes for milk, fat, and protein yields, and somatic cell score (SCS) for the important genes obtained from the first 3 lactations (LAC3). Red color indicates pathways for *CACNA2D1* gene and blue color indicates pathways for *MAPK10* gene. Nodes named in red are the observed genes.

LAC3 vs LAC1

In general, when comparing LAC3 to LAC1, the greatest proportions of genetic variance explained for milk, fat and protein yields were obtained from LAC3. For SCS, although with minimal differences, the greatest proportion of genetic variance explained was obtained from LAC1. The variance explained by each window along the chromosomes for milk, fat and protein yields, and SCS are shown in Figure S1 for LAC1. For milk, fat, and protein yields, the mean and (max) proportion of genetic variance explained were 0.23 (0.23), 0.73 (1.82), 0.34% (0.10%) higher for LAC3 compared to LAC1. On the other hand, for SCS, the mean and (max) was 0.09% (0.03%) lower for LAC3 than in LAC1.

In addition, a total of 34, 3, and 22 genes, respectively for milk, fat and protein yields were shared by LAC3 and LAC1. No important genes in common were found for SCS between LAC3 and LAC1 (Tables 1 and 2). For LAC1 the important genes that explained the highest proportion of genetic variance were not the same as in LAC3. The most important genes based on the proportion of genetic variance explained were the *SLC25A48*, *IL9*, *FBXL21*, *LECT2*, *TGFBI* genes on BTA7, for milk, fat and protein yields. For fat yield, the *NUP210* gene on BTA22 also explained a large proportion of genetic variance. For SCS, the *EPB41L4A* gene on BTA10 was considered as a potential important gene (Tables 1 and 2).

4.5. Discussion

This study implemented the weighted ssGWAS using an AR test-day model for milk, fat, and protein yields, and SCS traits in Holstein cattle. In addition, differences in QTL regions and important genes found using the first 3 lactations (LAC3) or only the first lactation (LAC1) were also studied. Comparisons were based on the magnitude of genetic variance explained by the genomic windows and the set of important genes that were found.

LAC3

Several of the important genes found in our study were previously reported in GWAS based on total accumulated yield for the same or correlated traits (Nayeri et al., 2016; Yue et al., 2017). For instance, in LAC3, the calcium channel voltage-dependent alpha-2/delta subunit 1 (*CACNA2D1*) gene on BTA4 which explained the highest proportion of genetic variance for milk, fat, and protein yields (Tables 1 and 2), encodes a member of the alpha-2/delta subunit family, which regulates calcium

current density and activation/inactivation kinetics of the calcium channel (Figure 3). It plays an important role in muscle physiology i.e., in muscles contraction (Magotra et al., 2017, 2018), thus helping in opening and closing the teat canal during milk let down (Gabashvili et al., 2007). In addition, it seems to have a small effect on calcium transport in the mammary gland (VanHouten et al., 2007). The *CACNA2D1* gene has been considered to be one of the potential candidate gene influencing SCS and mastitis (Magotra et al., 2017, 2018). Although this gene was not associated with SCS, it was associated with milk, fat and protein yields in our study.

For SCS, the transfer RNA glycine anticodon CCC (*TRNAG-CCC*) gene on BTA14 explained the highest proportion of genetic variance (Tables 1 and 2) in this study. It was also been associated with udder health and mastitis in a previous study on Holstein cattle (Wu et al., 2015). Nevertheless, there are no reports in the literature regarding the *TRNAG-CCC* gene action in responses to mastitis. Moreover, other genes associated with SCS, such as the Mitogen-Activated Protein Kinase 10 (*MAPK10*) gene, seem to be directly involved in inflammatory processes (Figure 3).

The *MAPK10* gene is involved in cell migration, inflammation, stress response, pathogenicity, and apoptosis (Ryman et al., 2015). Exposure to pathogens initially triggers a response from mammary epithelial cells and resident immune cells, which produce and secrete a variety of inflammatory mediators, such as cytokines. These inflammatory mediators also activate the endothelial cells, increasing vascular permeability which is necessary for the influx of neutrophils to ingest pathogens and limit extravascular tissue damage (Ryman et al., 2015).

GWAS studies frequently report strong associations between milk related traits and the diacylglycerol acyltransferase 1 (*DGAT1*) (e.g., Iung et al., 2019; Zhou et al., 2019). In this study, the *DGAT1* gene (located between 1795425 to 1804838 bp in the

BTA14) has not been found inside any important genomic window. Nevertheless, the *DGAT1* gene is 68.45 kb away from an important window found on BTA14 (1873292 to 2052873 bp, Table 1), for milk yield. Similar result was reported by Carvalheira et al. (2014), in which the authors performed a GWAS analysis in Portuguese Holstein cattle. In addition, Olsen et al. (2017) performed GWAS for milk fat composition in cows and also found no significant associations with the *DGAT1*. The authors showed that several animals were all homozygous for the A variant of the *DGAT1 K232A* polymorphism. In contrast to the A variant, the K variant is associated with increased fat yield, fat and protein percentage, and decreased milk and protein yields. They concluded that selection may have favored the A variant in the population, because most selection pressure was applied on milk and protein yield in the breeding goal.

Moreover, several genes close to the *DGAT1* that were previously associated with milk traits (Nayeri et al., 2016), such as *CYHR1, VPS28, MROH1, OPLAH, GRINA, SMPPD5, MAF1* and *GPR20*, were found as important genes in this study (Table 2).

LAC3 vs LAC1

Most GWAS studies have been done using the total accumulated yield from first lactation (Nayeri et al., 2016; Yue et al., 2017). We hypothesized that admitting more phenotypic information (as it happens in LAC3), could contribute to access genes that simultaneously have major effects throughout the lactations and may provide a better understanding of the genetic architecture of these traits. When using only the first lactation, the candidate gene is representing genomic regions with substantial effect only at a certain stage of the animal's productive life. Although most of the genes identified in our study are present in both LAC3 and LAC1 for milk, fat and protein

yields, the genes that accounted for most of the variance in LAC3 or LAC1 were not always the same (Tables 1 and 2).

On the other hand, for SCS no common genes between LAC3 and LAC1 were identified (Tables 1 and 2). It is important to mention that, for LAC1, the data set for SCS has fewer records compared to milk yield. In the AR model approach, TD records for multiple lactations (i.e. in LAC3) are considered simultaneously, providing a higher number of daughter-records per bull and, consequently, higher accuracy in predictions of GEBVs (Carvalheira et al., 2002). Therefore, GEBVs that are more accurate may have a positive influence on the estimation of the SNP effects and thus in the proportion of the variance explained per window. Similar results were observed by Lu and Bovenhuis, (2019), in which the authors compared different GWAS approaches using just the first lactation and observed that separate GWAS for specific stages within lactation were less powerful than GWAS for complete lactation (all TD) based on the repeatability model. They referred to the enormous difference in the number of records (on average,10 times as much), concluding for the importance of the number of records as they are expected to affect the GWAS results.

4.6. Conclusion

The inclusion of more phenotypic information (as it happens in LAC3), contributed to access genes that simultaneously have major effects throughout the lactations and provided a better understanding of the genetic architecture of milk traits on Portuguese Holstein cattle. Most of the genes identified in our study contributing with more than 1% of the total genetic variance were present in both (LAC3 and LAC1) for milk, fat and protein yields, although not always the same. In SCS, no common important genes were found between LAC3 and LAC1. These results suggest

that the weighted ssGWAs methodology using the AR model is feasible and may be applied in the National genetic evaluations.

4.7. Acknowledgments

The authors acknowledge Portuguese Dairy Cattle Breeders Association (ANABLE) and Embrapa Dairy Cattle for providing the used data. This study was partially financed by CAPES/FCT (99999.008462/2014-03) and CNPq/INCT-CA.

4.8. References

Aguilar, I., I. Misztal, D.L. Johnson, A. Legarra, S. Tsuruta, and T.J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score1. J. Dairy Sci. 93:743–752. doi:10.3168/jds.2009-2730.

Bindea, G., J. Galon, and B. Mlecnik. 2013. Systems biology CluePedia Cytoscape plugin : pathway insights using integrated experimental and in silico data. Bioinformatics 29:661–663. doi:10.1093/bioinformatics/btt019.

Bindea, G., B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W. Fridman, F. Pagès, Z. Trajanoski, and J. Galon. 2009. ClueGO : a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics 25:1091–1093. doi:10.1093/bioinformatics/btp101.

Carvalheira, J., E.J. Pollak, R.L. Quaas, and R.W. Blake. 2002. An autoregressive repeatability animal model for test-day records in multiple lactations. J. Dairy Sci. 85:2040–2045. doi:10.3168/jds.S0022-0302(02)74281-1.

Carvalheira, J., M.M.I. Salem, G. Thompson, S.Y. Chen, and A. Beja-Pereira. 2014. Genome-Wide Association Study for Milk and Protein Yields in Portuguese Holstein Cattle. Pages 1–3 in Proceedings, 10th World Congress of Genetics Applied to Livestock Production.

Gabashvili, I.S., B.H.A. Sokolowski, C.C. Morton, and A.B.S. Giersch. 2007. Ion channel gene expression in the inner ear. JARO - J. Assoc. Res. Otolaryngol. 8:305–328. doi:10.1007/s10162-007-0082-y.

Iung, L.H.S., J. Petrini, J. Ramírez-Díaz, M. Salvian, G.A. Rovadoscki, F. Pilonetto,

B.D. Dauria, P.F. Machado, L.L. Coutinho, G.R. Wiggans, and G.B. Mourão. 2019. Genome-wide association study for milk production traits in a Brazilian Holstein population. J. Dairy Sci. 1:1–10. doi:10.3168/jds.2018-14811.

Lu, H. and H. Bovenhuis. 2019. Genome-wide association studies for genetic effects that change during lactation in dairy cattle. J. Dairy Sci. 102:1-14. doi:10.3168/jds.2018-15994.

Magotra, A., I.D. Gupta, A. Verma, R. Alex, M. Vineeth, and T. Ahmad. 2018. Candidate SNP of CACNA2D1 Gene Associated with Clinical Mastitis and Production Traits in Sahiwal (Bos taurus indicus) and Karan Fries (Bos taurus taurus × Bos taurus indicus). Anim. Biotechnol. 0:1–7. doi:10.1080/10495398.2018.1437046.

Magotra, A., I.D. Gupta, A. Verma, M. V Chaudhari, A. Arya, M.R. Vineeth, R. Kumar, and A.S. Selvan. 2017. Characterization and validation of point mutation in exon 19 of CACNA2D1 gene in Karan Fries (Bos taurus × Bos indicus) cattle. Indian J. Anim. Res. 51:227–230. doi:10.18805/ijar.5668.

Misztal, I., S. Tsuruta, D. Lourenco, I. Aguilar, A. Legarra, and Z. Vitezica. 2015. Manual for BLUPF90 Family of Programs. Athens: University of Georgia. Retrieved from http://nce.ads.uga.edu/wiki/doku.php?id=application_programs.

Nayeri, S., M. Sargolzaei, M.K. Abo-ismail, N. May, S.P. Miller, F. Schenkel, S.S. Moore, and P. Stothard. 2016. Genome-wide association for milk production and female fertility traits in Canadian dairy Holstein cattle. BMC Genet. 1–11. doi:10.1186/s12863-016-0386-1.

Olsen, H.G., T.M. Knutsen, A. Kohler, M. Svendsen, L. Gidskehaug, H. Grove, T. Nome, M. Sodeland, K.K. Sundsaasen, M.P. Kent, H. Martens, and S. Lien. 2017. Genome - wide association mapping for milk fat composition and fine mapping of a QTL for de novo synthesis of milk fatty acids on bovine chromosome 13. Genet. Sel. Evol. 1–13. doi:10.1186/s12711-017-0294-5.

Qanbari, S., E.C.G. Pimentel, J. Tetens, G. Thaller, P. Lichtner, A.R. Sharifi, and H. Simianer. 2010. The pattern of linkage disequilibrium in German Holstein cattle. Anim. Genet. 346–356. doi:10.1111/j.1365-2052.2009.02011.x.

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Resende, R.T., M.D. V. de Resende, C.F. Azevedo, F. Fonseca e Silva, L.C. Melo,

H.S. Pereira, T.L.P.O. Souza, P.A.M.R. Valdisser, C. Brondani, and R.P. Vianello. 2018. Genome-Wide Association and Regional Heritability Mapping of Plant Architecture, Lodging and Productivity in *Phaseolus vulgaris*. G3: Genes, Genome and Genetics 8:2841–2854. doi:10.1534/g3.118.200493.

Ryman, V.E., N. Packiriswamy, and L.M. Sordillo. 2015. Role of endothelial cells in bovine mammary gland health and disease. Anim. Heal. Res. Rev. 16:135–149. doi:10.1017/S1466252315000158.

Salem, M.M.I., G. Thompson, S. Chen, A. Beja-Pereira, and J. Carvalheira. 2018. Linkage disequilibrium and haplotype block structure in Portuguese holstein cattle. Czech J. Anim. Sci. 63:61–69. doi:10.17221/56/2017-CJAS.

Silva, A.A., D.A. Silva, F.F. Silva, C.N. Costa, P.S. Lopes, A.R. Caetano, G. Thompson, and J. Carvalheira. 2019. Autoregressive single-step test-day model for genomic evaluations of Portuguese Holstein cattle. J. Dairy Sci. 1–10. doi.org/10.3168/jds.2018-15191.

VanHouten, J.N., M.C. Neville, and J.J. Wysolmerski. 2007. The calcium-sensing receptor regulates plasma membrane calcium adenosine triphosphatase isoform 2 activity in mammary epithelial cells: A mechanism for calcium-regulated calcium transport into milk. Endocrinology 148:5943–5954. doi:10.1210/en.2007-0850.

VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91:4414–4423. doi:10.3168/jds.2007-0980.

Vitezica, Z.G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. Genet. Res. (Camb). 93:357–366. doi:10.1017/S001667231100022X.

Wang, H., I. Misztal, I. Aguilar, A. Legarra, R.L. Fernando, Z. Vitezica, R. Okimoto, T. Wing, R. Hawken, and W.M. Muir. 2014. Genome-wide association mapping including phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body weight in broiler chickens. Front. Genet. 5.

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W.M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. (Camb). 94:73–83. doi:10.1017/S0016672312000274.

Wu, X., M.S. Lund, G. Sahana, B. Guldbrandtsen, D. Sun, Q. Zhang, and G. Su. 2015. Association analysis for udder health based on SNP-panel and sequence data in Danish Holsteins. Genet. Sel. Evol. 47:1–14. doi:10.1186/s12711-015-0129-1.

Yue, S.J., Y.Q. Zhao, X.R. Gu, B. Yin, Y.L. Jiang, Z.H. Wang, and K.R. Shi. 2017. A

genome-wide association study suggests new candidate genes for milk production traits in Chinese Holstein cattle. Anim. Genet. 48:677–681. doi:10.1111/age.12593.

Zhou, C., C. Li, W. Cai, S. Liu, H. Yin, S. Shi, Q. Zhang, and S. Zhang. 2019. Genome-Wide Association Study for Milk Protein Composition Traits in a Chinese Holstein Population Using a Single-Step Approach. Front. Genet. 10:1–17. doi:10.3389/fgene.2019.00072.

4.9. Supplementary material

Figure S1 GWAS results of milk trait for the first lactation only (LAC1). **a** milk yield, **b** fat yield, **c** protein yield and **d** somatic cell score (SCS). Each dot represents one SNP window of 100 kb. On the y-axis is the chromosomes (Chr).

Programming codes in R to access SNP weights according to Wang et al. 2012

```
#These analyses were set up by Alessandra A. Silva; Fabyano F.
Silva; Delvan A. Silva; Hugo T. Silva; Cláudio N. Costa; Paulo S.
Lopes; Renata Veroneze; Gertrude Thompson and Julio Carvalheira

# Weight for SNPs according to Wang et al. 2012
#Genomic Breending value (BV) file for genotyped animals
#     column 1 - ID
#     column 2 - BV
#Genotype file
#     column 1 - ID
#     column 2:m - markers - one column per marker
#G matrix
#     column 1 – ID1
#     column 2 - ID2
#     column 3 - Value
#wt file
# if iteration 1, D=I, with dimension according to SNPs number.
# Denote t as an iteration number.

# NO HEADERS on input tables
# Returns u and wt vectors.
# Install required packages from internet if is not already
installed.
install.packages('reshape2')
install.packages('psych')
# path and setwd
path='add here the setwd'
setwd(path)    # sets path as the working directory
getwd()   #set and check directory
# To read BV file
bv = read.table('BV data here')
# To read Genotype file
gen = read.table('genotype data here')
# To read wt
wt = read.table('./wt here',sep=' ', h=F)
# To read G matrix
g = read.table('G_Orig file here',sep=' ',h=F)
################################################################
#Genomic Breending value (BV) file for genotyped animals
colnames(bv)=c('code','bv')
bv = bv[with(bv,order(code)),]
#GEBV
a = bv$bv
#Genotype file
gen = gen[with(gen, order(V1)),]
################################################################
#Markers information
W=as.matrix(gen[,-c(1)] )
################################################################
# allele frequency
p1=matrix(0,ncol(W),1)
q1=matrix(0,ncol(W),1)
for(i in 1:ncol(W))
{
p1[i,]=(2*length(which(W[,i]==0))+length(which(W[,i]==1)))/(2*
length(na.omit(W[,i])))
q1[i,]=(2*length(which(W[,i]==2))+length(which(W[,i]==1)))/(2*
length(na.omit(W[,i])))
}
```

```
p=p1
q=1-p1
#####################################################################
#making pi
pi = matrix(0,ncol(W),1)
for(i in 1:ncol(W))
{
pi[i,] = (sum(W[,i])/(2*nrow(W)))
}
#####################################################################
#Z matrix
P = 2*pi
P = as.matrix(P)
Z=matrix(0,nrow(W),ncol(W))
for (i in 1:nrow(W))
{
Z[i,]= W[i,]-(t(P))
}
#lambda
lamb = 1/(sum(2*pi*(1-pi)))
#SNP weighted = 1
D1 = diag(1,ncol(W),ncol(W))
#SNP weighted wt
wt=as.vector(wt$V1)
Dt = diag(wt,ncol(W),ncol(W))
#####################################################################
#G matrix
g = g[with(g, order(V1,V2)),]
#Full matrix
library(reshape2)
g=acast(g, V1~V2, value.var='V3', fill=0)
#G inverse
ginv = solve(as.matrix(g))
#SNP effects (Wang et al. 2012)
u = lamb*Dt%*%t(Z)%*%ginv%*%a
write.table(u, file ='./u.txt',sep=' ',col.names=F,row.names=F,
quote=F)
#new SNP weighted
d = (u^2)*(2*pi*(1-pi))
#Normalize new D
library(psych)
diag(Dt) = d
Dt = (tr(D1)/tr(Dt))*Dt
#Weight
wt = diag(Dt)
write.table(wt, file ='./wt.txt',sep=' ',col.names=F,row.names=F,
quote=F)
```

CHAPTER 5

5.1. General conclusions

The main objective with this thesis was to evaluate the inclusion of genomic information in the AR test-day model for multiple lactations for better understand the genetic and genomic aspects of milk related traits in Portuguese Holstein cattle.

To perform the genomic evaluation under AR model we firstly evaluated the imputation accuracy for Portuguese Holstein cattle using several commercially available SNP panels in different densities with a relatively small number of genotyped animals. Genotype imputation was feasible and may be advantageous to the National genomic evaluations of dairy cattle.

In this sense, the ssGBLUP methodology applied to AR models was feasible and may be advantageous to genetic evaluation. With the anticipated increase in the number of genotyped animals, it is expected provide even higher reliabilities especially for the young stock, thus contributing to the improvement of the genetic progress of the Portuguese dairy cattle population.

The weighted ssGWAs methodology using the AR model also was feasible. The GWAS results helped to provide a better understanding of the genetic architecture of milk related traits in Portuguese Holstein cattle. In addition, the inclusion of more phenotypic information, such as TD records from the first 3 lactations may contribute to access genes with major effects simultaneously throughout the lactations.

The results described in this thesis will contribute to advance the knowledge about genomic prediction and GWAS for milk related traits. In addition, this thesis provided the first results about the inclusion of genomic information in AR models, which will be important for future national genetic evaluations.