UNIVERSIDADE DE SÃO PAULO
FACULDADE DE ZOOTECNIA E ENGENHARIA DE ALIMENTOS

BRUNO DA COSTA PEREZ

**Strategies to improve results from genomic analyzes in
small dairy cattle populations**

BRUNO DA COSTA PEREZ

**Strategies to improve results from genomic analyzes in
small dairy cattle populations**

Thesis submitted to the College of
Animal Science and Food Engineering
University of São Paulo (FZEA/USP)
in partial fulfillment of the
requirements for the degree of Doctor
in Science.

Concentration area: Animal quality and
productivity

Supervisor: PhD. Júlio Cesar de
Carvalho Balieiro

Ficha catalográfica elaborada pelo
Serviço de Biblioteca e Informação, FZEA/USP,
com os dados fornecidos pelo(a) autor(a)

dP438s da Costa Perez, Bruno
Strategies for improving results from genomic
analyzes in small dairy cattle populations. / Bruno
da Costa Perez ; orientador Júlio Cesar de Carvalho
Balieiro. -- Pirassununga, 2019.
149 f.

Tese (Doutorado - Programa de Pós-Graduação em
Zootecnia) -- Faculdade de Zootecnia e Engenharia
de Alimentos, Universidade de São Paulo.

1. genomic selection . 2. graph-theory. 3. graph
community. 4. population partition. 5. genotyping.
I. de Carvalho Balieiro, Júlio Cesar, orient. II.
Título.

BRUNO DA COSTA PEREZ

**Strategies to improve results from genomic analyzes in
small dairy cattle populations**

Thesis submitted to the College of
Animal Science and Food Engineering
University of São Paulo (FZEA/USP)
in partial fulfillment of the
requirements for the degree of Doctor
in Science.

Concentration area: Animal quality and
productivity

Date of approval: ____/____/____

1st Referee _____

Institution _____

2nd Referee _____

Institution _____

3rd Referee _____

Institution _____

4th Referee _____

Institution _____

5th Referee _____

Institution _____

ABSTRACT

PEREZ, B.C. **Strategies to improve results from genomic analyzes in small dairy cattle populations.** 2019,149p. Thesis – College of Animal Science and Food Engeneering, University of São Paulo (FZEA/USP).

The main objective of the present thesis was to propose a procedure to optimize genotypic information value in small dairy cattle populations and investigate the impacts of including genotypes and phenotypes of cows chosen by different strategies over the performance of genome-wide association studies and genomic selection. The first study was designed to propose innovative methods that could support alternative inference over population structure in livestock populations using graph theory. It reviews general aspects of graphs and how each element relates to theoretical and practical concepts of traditional pedigree structure studies. This chapter also presents a computational application (PedWorks) built in Python 2.7 programming language. It demonstrates that graph theory is a suitable framework for modeling pedigree data. The second study was aimed asses how graph community detection algorithms could help unraveling population partition. This new concept was considered to develop a method for stablishing new cow genotyping strategies (community-based). Results obtained showed that accounting for population structure using community detection for choosing cows to get included in the reference population may improve results from genomic selection. Methods presented are easily applied to animal breeding programs. The third study aimed to observe the impacts of different genotyping strategies (including the proposed community-based) over the ability to detect quantitative trait loci in genome-wide association studies. Distinct models for genomic analysis were also tested. Results obtained showed that including

cows with extreme phenotypic observations proportionally sampled from communities can improve the ability to detect quantitative trait loci in genomic evaluations. The last chapter was designed study possible deleterious impacts of the presence of preferential treatment (in different levels) in a small dairy cattle population environment over accuracy and bias of genomic selection. Different proportions of cows with artificially increased phenotypic observations were included in the reference population. Observed results suggest that both accuracy and bias are affected by the presence of preferential treatment of cows in the evaluated population. Preferential treatment is expected to have much more effect on the performance of genomic selection in small than in large dairy cattle populations for the higher (proportional) value of the information from cows in such reduced-size breeds.

RESUMO

PEREZ, B.C. **Estratégias para aprimorar os resultados de análises genômicas em pequenas populações de gado de leite.** 2019,149p. Tese de Doutorado– Faculdade de Zootecnia e Engenharia de Alimentos, Universidade de São Paulo (FZEA/USP).

O principal objetivo da presente tese foi propor um procedimento capaz de otimizar o valor da informação genotípica em pequenas populações de gado de leite e investigar os impactos da inclusão de genótipos e fenótipos de vacas escolhidas por diferentes estratégias sobre o desempenho de estudos de associação genômica ampla e seleção genômica. O primeiro estudo foi delineado para elaborar um método que permita uma inferência alternativa sobre a estrutura populacional de populações de animais de produção usando como base a teoria de grafos. Este revê os aspectos gerais de grafos e como cada elemento se relaciona com conceitos teóricos e práticos de estudos de estrutura de pedigree tradicionais. Este capítulo também apresenta um aplicativo computacional (PedWorks) construído em linguagem de programação Python 2.7. Resultados observados demonstraram que a teoria de grafos é uma estrutura adequada para modelar dados de pedigree. O segundo estudo teve como objetivo avaliar como os algoritmos de detecção de comunidades de grafos poderiam ajudar revelar o particionamento de uma população. Este novo conceito foi considerado para desenvolver um método para o estabelecimento de novas estratégias de genotipagem de vacas (baseadas em comunidades). Os resultados obtidos mostraram que a contabilização da estrutura populacional usando a detecção de comunidades para a escolha de vacas a serem incluídas na população de referência pode melhorar os resultados da seleção genômica. Os métodos apresentados sugerem ser facilmente introduzidos em

programas de melhoramento animal. O terceiro estudo teve como objetivo observar os impactos de diferentes estratégias de genotipagem (incluindo a anteriormente proposta baseada em comunidades) sobre a capacidade de detectar locos relacionados características quantitativas por meio de estudos de associação genômica ampla. Modelos distintos para análise genômica também foram testados. Os resultados obtidos mostraram que incluir vacas com observações fenotípicas extremas amostradas proporcionalmente das comunidades pode melhorar a capacidade de detectar locos de características quantitativas em avaliações genômicas. O último capítulo foi desenhado para estudar possíveis impactos deletérios da presença de tratamento preferencial no ambiente de pequenas populações de gado leiteiro sobre resultados da seleção genômica. Diferentes proporções de vacas com observações fenotípicas aumentadas artificialmente foram incluídas na população de referência. Os resultados observados sugerem que tanto a acurácia quanto o viés são afetados pela presença de tratamento preferencial de vacas na população avaliada. Espera-se que o tratamento preferencial tenha muito mais efeito sobre o desempenho da seleção genômica em populações pequenas de gado de leite que em grandes populações devido a maior relevância das informações de vacas em raças de tamanho reduzido.

TABLE OF CONTENTS

CHAPTER 1. GENERAL INTRODUCTION.....	1
1.1 THESIS BACKGROUND AND MOTIVATION.....	1
1.2 THESIS OUTLINE	2
1.3 THESIS OBJECTIVES	4
1.3.1 Main objective.....	4
1.3.2 Specific objectives.....	4
1.4 REFERENCES.....	5
 CHAPTER 2. PEDWORKS: MODELING GENEALOGICAL STRUCTURED DATA BY GRAPH-THEORY APPROACH FOR POPULATION STUDIES IN ANIMAL GENETICS.....	 8
2.1 INTRODUCTION.....	9
2.2 MATERIAL AND METHODS.....	11
2.2.1 Graph theory and properties of graphs.....	11
2.2.2 Graph centrality.....	13
2.2.2.1 Degree centrality.....	13
2.2.2.2 Closeness centrality.....	15
2.2.2.3 Betweenness centrality.....	15
2.2.2.4 Katz centrality.....	16
2.2.3 Degree Rank Function.....	17
2.2.4 Community detection.....	18
2.2.5 Graph visualization.....	22
2.2.5.1 Force directed algorithm and large pedigrees.....	26
2.2.6 Application of graph framework to genealogical structured data.....	27
2.2.6.1 Simulated data.....	28
2.2.6.2 Real data.....	29
2.2.7 Correlation between relationship coefficients and the distance between vertices.....	30
2.3 RESULTS AND DISCUSSION.....	31
2.3.1 Graph general structure.....	31
2.3.2 Graph analysis.....	35
2.3.3 Linear association between elements of A and D matrices.....	38
2.3.4 Further studies, current and future applications.....	41
2.4 CONCLUSION.....	43
2.5 REFERENCES.....	43
 CHAPTER 3. ACCOUNTING FOR POPULATION STRUCTURE IN COW GENOTYPING STRATEGIES.....	 49
3.1 INTRODUCTION.....	50
3.2 MATERIAL AND METHODS.....	53
3.2.1 Simulated population.....	53
3.2.2 Genome simulation.....	54

3.2.3 Phenotype simulation.....	55
3.2.4 Pedigree graphs, community detection and population partitioning.....	56
3.2.5 Selective genotyping designs.....	58
3.2.6 Quality control and genomic prediction model.....	61
3.2.7 Accuracy, prediction error and bias.....	62
3.2.8 Mimicking selection for an expensive- of difficult-to-measure trait in dairy cattle.....	64
3.3 RESULTS.....	66
3.3.1 Population structure description.....	66
3.3.2 Community detection and population partitioning	67
3.3.3 Accuracy of genomic evaluations.....	70
3.3.4 Prediction error and bias.....	74
3.3.4 Community-based genotyping design for a reference population of cows.....	77
3.4 DISCUSSION.....	79
3.5 CONCLUSION.....	85
3.6 REFERENCES.....	86

CHAPTER 4. ASSESSING GENOTYPE AND PHENOTYPE INFORMATION VALUE WHEN CONSIDERING DISTINCT WELL-CONNECTED GROUPS OF COWS OVER GENOME-WIDE ASSOCIATION STUDIES RESULTS FOR LOWLY HERITABLE TRAITS.....94

4.1 INTRODUCTION.....	95
4.2 MATERIAL AND METHODS.....	98
4.2.1 Simulated population and genome.....	98
4.2.2 Phenotypes.....	99
4.2.3 Graph community detection and population partitioning.....	100
4.2.4 Selective cow genotyping designs.....	100
4.2.5 Quality control and genome-wide association study.....	102
4.2.5.1 GBLUP.....	102
4.2.5.2 Weighted single-step GBLUP.....	103
4.2.5.3 Bayes R.....	105
4.2.6 Comparison of models and scenarios.....	106
4.3 RESULTS.....	107
4.3.1 Genomic data structure description and population partition.....	107
4.3.2 Parameters of comparison.....	108
4.4. DISCUSSION.....	114
4.5 CONCLUSION.....	117
4.6 REFERENCES.....	117

CHAPTER 5. IMPACTS OF SELECTIVE GENOTYPING OVER RESULTS OF GENOMIC EVALUATIONS UNDER THE ASSUMPTION OF PREFERENTIAL TREATMENT OF COWS IN SMALL DAIRY CATTLE POPULATIONS.....123

5.1 INTRODUCTION.....	124
-----------------------	-----

5.2 MATERIAL AND METHODS.....	127
5.2.1 Simulated population.....	127
5.2.2 Genome simulation.....	128
5.2.3 Phenotypic observations and preferential treatment simulation.....	129
5.2.4 Selective genotyping designs.....	130
5.2.5 Genotype quality control and genomic analysis.....	132
5.2.6 Assessing accuracy and bias.....	133
5.3 RESULTS.....	133
5.3.1 Accuracy for de-regressed estimated breeding values of genotyped cows.....	133
5.3.2 Performance of genomic evaluations in the absence of preferential treatment.....	134
5.3.3 Accuracy and bias of genomic evaluation in the presence of preferential treatment.....	137
5.4 DISCUSSION.....	139
5.5 CONCLUSION.....	143
5.6 REFERENCES.....	143

CHAPTER 1. GENERAL INTRODUCTION

1.1 THESIS BACKGROUND AND MOTIVATION

Studies proposed and outlined in the present thesis were elaborated after preliminary studies using phenotypic, genotypic and pedigree data from the Brazilian Guzerá population. The Guzerá is a *Bos indicus* breed, one of the first zebu breeds to arrive in the first importation waves in the early 90's, is largely considered as dual-purposed in national territory, having both economic and social relevance in pasture-based tropical production systems in Brazil (PEIXOTO et al., 2010). In the last decades, historical occurrences have decreased the numbers of Guzerá individuals in Brazil, mainly for the large use of Guzerá females as dams in a diversity of crossbreeding schemes, either for milk, beef and dual-purpose objectives (PEIXOTO et al., 2010).

Today, there are national breeding programs that consider selection of Guzerá populations for dairy and beef purposes. In the national dairy Guzerá animal breeding program, conducted by EMBRAPA (Empresa Brasileira de Pesquisa Agropecuária) in partnership with ABCZ (Associação Brasileira de Criadores de Zebu), animals have been genotyped in the last few years aiming to achieve a better comprehension of the population structure, genetic architecture of economic traits and for further implementation of a genomic selection program for this breed. In the year 2018, the complete data had a total of 50 bulls and 1,053 Guzerá cows with genotypic information available. There is an understanding, from researchers and breeders, that the numbers of genotyped individuals should be increased in the next years in order to accomplish the designed objectives. However, as a dairy breed of reduced size under selection by a

relatively recent animal breeding program, there is just a limited number of Guzerá bulls with reliable estimated breeding values, which increases the importance of genotyping Guzerá cows as possible candidates to be included in a reference population in the future (JIMÉNEZ-MONTERO et al., 2012; PRYCE AND DAETWYLER, 2012; THOMASEN et al., 2014; DING et al., 2015; GAO et al., 2015; PLIESCHKE et al., 2016, JENKO et al., 2016; UEMOTO et al., 2017). To accomplish that objective in an efficient manner, strategies for selective genotyping of cows could help to optimize genotypic and phenotypic information value, as budget for genotyping is often limited.

From this background, we aimed to propose innovative strategies that could help to establish new genotyping designs based on an optimized allocation of genotyping resources. Results presented in this thesis will help to comprehend limitations and opportunities generated by specific aspects of dairy breeds of reduced size for the implementation of genomic selection.

1.2 THESIS OUTLINE

Chapter 2 aimed to present and explain methods for alternative inference over population structure in livestock populations using graph theory. It reviews general aspects of graphs and how each element relates to theoretical and practical concepts of traditional pedigree structure studies. This chapter also presents a computational application (PedWorks) built in Python 2.7 programming language, which was elaborated to implement all features presented in this thesis.

Chapter 3 was designed to investigate how a graph community detection algorithm (BLONDEL et al., 2008) could support assessing population partition. We used this concept and method as a tool for developing new cow genotyping strategies. Those strategies were called “community-based” and aimed to take into account the complete pedigreed population structure to obtain the best set of cows to be genotyped for maximizing performance of genomic selection.

In Chapter 4, we aimed to assess the impacts of different genotyping strategies (including the community-based) over the ability to detect quantitative trait loci in genome-wide association studies for a low heritability sex-limited trait. We also tested different models for genomic analysis to understand how each model is affected by the criteria adopted to obtain the set of genotyped individuals.

Chapter 5 was elaborated from the perception that in small dairy cattle breeds, the proportion of elite herds is often much higher than what is found in larger breeds, such as the Holstein cattle. As previously described in literature (KUHN, 1994; STRANDEN AND GIANOLA, 1998; TSURUTA et al., 2001; DASSONEVILLE et al., 2012; DEHNAVI et al., 2017), cows from elite herds are more susceptible to preferential treatment, which is often associated to the inclusion of bias when estimating breeding values. From this, we decided to investigate possible deleterious impacts of the presence of preferential treatment (in different levels) over genomic evaluation accuracy and bias.

1.3 THESIS OBJECTIVES

1.3.1 Main objective

Propose a procedure to optimize genotypic information value in small dairy cattle populations and investigate the impacts of including genotypes and phenotypes of cows chosen from different strategies over the performance of genomic evaluations.

1.3.2 Specific objectives

- a) Investigate the graph theory framework as a tool to support alternative inference over population structure in livestock populations.
- b) Investigate the use of a graph-community algorithm as a tool to obtain population partition and develop new cow genotyping strategies that can indirectly take into account the complete population partition.
- c) Assess the impacts of choosing cows to get genotyped and included in the reference population (previously containing a reduced number of progeny-tested bulls) by traditional and community-based methods over accuracy, bias and prediction error in genomic selection; and to
- d) Comprehend the possible impacts of different levels of preferential treatment of cows over the performance of genomic selection in small dairy cattle populations. Investigate the influence of the structure and size of the set of cows included in the reference population over accuracy and bias of multi-step genomic evaluations.

1.4 REFERENCES

- BLONDEL, V. D., GUILLAUME, J. L., LAMBIOTTE, R., & LEFEBVRE, E. Fast unfolding of communities in large networks, 1–12. **Journal of Statistical Mechanics: Theory and Experiment**, 10, 10008, 2008.
- DASSONNEVILLE, R., BAUR, A., FRITZ, S., BOICHARD, D., & DUCROCQ, V. Inclusion of cow records on genomic evaluations and impact on bias due to preferential treatment. **Genetics Selection Evolution**, 44:40, 2014.
- DEHNAVI, E., MAHYARI, S. A., SCHENKEL, F. S., & SARGOLZAEI, M. The effect of using cow genomic information on accuracy and bias of genomic breeding values in a simulated Holstein dairy cattle population. **Journal of Dairy Science**, 101(6), 5166–5176, 2018.
- GAO, H., LUND, M. S., ZHANG, Y. & SU, G. Accuracy of genomic prediction using different models and response variables in the Nordic Red cattle population. **Journal of Animal Breeding and Genetics**, 130:333-340, 2013.
- GIRVAN, M., & NEWMAN, M. E. J. Community structure in social and biological networks, 99(12), 2002.
- JENKO, J., WIGGANS, G. R., COOPAR, T. A., EAGLEN, S. A. E., LUFF, W. G. L., BICHARD, M., PONG-WONG, R., & WOOLIAMs J. A. **Journal of Dairy Science**, 100(1):439-452, 2017.
- JIMÉNEZ-MONTERO, J. A., GONZÁLEZ-RECIO O., & ALENDA R. Genotyping strategies for genomic selection in small dairy cattle populations. **Animal**, 6, 1216–

1224, 2012.

KUHN, M. T., BOETTCHER, P. J., & FREEMAN, A. E. Potential Biases in Predicted Transmitting Abilities of Females from Preferential Treatment. **Journal of Dairy Science**, 77(8), 2428–2437, 1994.

PEIXOTO, M. G. C. D., POGGIAN, C. F., VERNEQUE, R. S., EGITO, A. A., CARVALHO, M. R. S., PENNA, V. M., MACHADO, M. A. Genetic basis and inbreeding in the Brazilian Guzarat (*Bos indicus*) subpopulation selected for milk production. **Livestock Science**, 131(2-3), 168–174, 2010.

PLIESCHKE, L., EDEL, C., PIMENTEL, E. C. G., EMMERLING, R., BENNEWITZ, J., & GOTZ K. U. Systematic genotyping of groups of cows to improve genomic estimated breeding values of selection candidates. **Genetics Selection Evolution**, 48:73, 2016.

PRYCE, J.E., & DAETWYLER, H. D. Designing dairy cattle breeding schemes under genomic selection: a review of international research. **Animal Production Science**, 52, 107-114, 2012.

STRANDÉN I. & GIANOLA, D. Attenuating effects of preferential treatment with Student-t mixed linear models: a simulation study. **Genetics Selection Evolution**, 30 (6), pp.565- 583, 1998.

TSURUTA, S.; KEOWN, JEFFREY F.; VAN VLECK, L. DALE; AND MISZTAL, I. Bias in Genetic Evaluations by Records of Cows Treated with Bovine Somatotropin. **Faculty Papers and Publications in Animal Science**, 150, 2000.

UEMOTO, Y., OSAWA, T., & SABURI, J. Effect of genotyped cows in the reference population on the genomic evaluation of Holstein cattle. **Animal**, 11:3, 382-393, 2017.

CHAPTER 2. PEDWORKS: MODELING GENEALOGICAL STRUCTURED DATA BY GRAPH-THEORY APPROACH FOR POPULATION STUDIES IN ANIMAL GENETICS.

Bruno da Costa Perez^a, Stephen P. Millerⁱ, Fabio Tirelo^b, Júlio Cesar de Carvalho Balieiro^d, Bruno Valente^h, Guilherme J.M Rosa^h, Juliana Demenshuk Machado^f, André Rabelo Fernandes^e, Aníbal Vercesi Filho^g, Maria Gabriela Campolina Diniz Peixoto^c, Ricardo Vieira Ventura^{a,i,j}

^a *Department of Animal Sciences, College of Animals Science and Food Engineering (FZEA), University of São Paulo (USP), Pirassununga, São Paulo, Brazil*

^b *Google, Program development, Montreal, Quebec, Canada*

^c *National Center of Research on Dairy Cattle (CNPGL), Brazilian Agricultural Research Corporation (EMBRAPA), Juiz de Fora, Minas Gerais, Brazil*

^d *College of Veterinary Medicine and Animal Science (FMVZ), University of São Paulo (USP), Pirassununga, São Paulo, Brazil*

^e *Dairy Gir Breeder's Association (ABCGIL), Uberaba, MG, Brazil*

^f *Department of Animal Science, Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brazil*

^g *Agência Paulista de Tecnologia dos Agronegócios (APTA), Mococa, SP, Brazil*

^h *Department of Animal Sciences, University of Wisconsin – Madison, Madison, WI 53706, USA*

ⁱ *Beef Improvement Opportunities (BIO), Elora, Ontario, Canada*

^j *Department of Biosciences, University of Guelph, Guelph, Ontario, Canada*

ABSTRACT Graph theory has been vastly used to better understand complex structured systems and how its functional units relate to each other. Pedigree data, which represent the expectation of genetic connectedness between individuals in a population, perfectly adapt into the graph framework. The detailed comprehension of the topology of pedigree graphs can help to detect individuals (and groups of individuals) that represent influential roles in the population and reveal population partitioning, among other features. This innovative conceptual approach can support further animal breeding studies, both for quantitative genetic and genomic areas. In this study, we present and discuss methods

based exclusively on graph theory to represent, analyze and visualize pedigree data. Both simulated and real pedigree data were used as practical examples. Also, we present PedWorks, a computer program, written in Python 2.7 programming language, developed to apply all methods here discussed to animal breeding problems (Available at <https://github.com/BrnCPPrz/PedWorks>).

Keywords: data visualization, pedigree, population structure, graph centrality, community detection

2.1 INTRODUCTION

Nowadays, data has been increasingly represented by networks, often called graphs. Graphs are mathematical structures used to represent relationship between variables in many fields, notably in computer, social and data sciences (NEWMAN, 2003). Formed by a collection of vertices (V) and edges (E), a graph (G) is represented by ordered pairs defined as $G = (V, E)$, which implies the perception that every edge connects two vertices. Edges in a graph represent the relationship between the vertices which it establishes a connection.

Pedigrees are sets of structured data that exhibit complex and discernible relationship patterns. It expresses (a belief of) the genetic relationships between individuals and have been used for both inference and decision making in livestock populations. In the scope of graph theory, a pedigree may be represented as a directed acyclic graph (DAG) (YANG et al., 2014). A DAG, also called “acyclic digraph”, is a graph in which each edge connects one vertex to another in only one direction, and in which it

is not impossible to start at some vertex v_i and follow a sequence of edges that eventually loops back to v_i . That property agrees with the obvious practical concept of pedigrees that no individual may be an ancestor of itself. When graph theory is applied to model pedigrees, each vertex represents an individual in the population and each edge indicates a parent-child relationship. In practice, the number of (outgoing) connections a vertex (understood here as an “individual”) has, determines the number of progeny of that same individual. Differently from more general DAGs, in pedigree structures each vertex cannot have more than two incoming edges, as any individual must have only up to two parents.

Much effort has been put into quantifying the genetic relationship among individuals in animal populations. The concept of inbreeding, often measured as the inbreeding coefficient (F) (WRIGHT, 1931; MEUWISSEN and LUO, 1992) has contributed remarkably to the development of the animal and plant breeding as a science, and has been successfully applied by quantitative geneticists to characterize animal populations according to its level of genetic connectivity. Many other approaches have been introduced (BOICHARD et al. 1997; GUTIÉRREZ et al., 2003) for population studies in the course of history. However, not much focus has been given to pedigree data visualization, as well as to new frameworks able to assess the underlying configuration of complex structured data.

The constant growth on the influx of information in any livestock production enterprise requires novel processes, both to analyze and visualize massive data. This investigation aims to introduce innovative methods using exclusively graph theory to handle, represent, analyze and visualize pedigree data. Also, we present PedWorks (Available at <https://github.com/BrnCPrz/PedWorks>), a computer program written in

Python (version 2.7) programming language (also making use of third party modules), primarily for handling pedigree problems converted to the DAG concept. We implement the methods presented in this paper for describing genealogical structures using simulated and real data recorded by dual purpose and dairy cattle associations in Brazil.

2.2 MATERIAL AND METHODS

2.2.1 Graph theory and properties of graphs

A graph is a flexible structured system containing a finite number of points, called vertices (or nodes), and edges (or arcs), which essentially connects pairs vertices (REINHARD, 2017). Given a graph G , $V(G)$ comprehends the vector of vertices and $E(G)$ the vector of edges contained in G . An edge, defined as $E = \{(x, y) \mid x, y \in V(G)\}$, in practice, represents the connection between vertices x and y . Vertices connected by an edge are said to be neighbors. Likewise, edges may be weighted according to the strength of the relationship, the distance between the vertices they connect, or the magnitude of information flux they represent (NEWMAN, 2004). Graphs are classified as directed if connections between vertices have a unique direction, and undirected when connections are symmetrical between neighbor vertices. In directed graphs, edges are graphically represented as an “arrow”, given the concept of directionality assumed. Also, graphs may be cyclic or acyclic, considering the presence/absence of paths (sequence of edges) from any vertex that ultimately leads back to the same vertex (ALBERT AND BARABÁSI, 2002).

A graph is typically represented by an adjacency matrix, meaning that it can easily

adapt to linear algebra operations. The adjacency matrix is a square matrix with $n \times n$ elements, where n refers to the number of vertices in the graph (BEINEKE AND WILSON, 2004). Rows and columns of nonzero elements refer to source and target vertices of the edges, respectively. The adjacency matrix of an undirected graph is symmetric to the main diagonal, while for directed graphs it is frequently non-symmetric (CHARTRAND AND LENIAC, 1996). Edge weights are an important feature for describing large-scale complex graphs as they denote the magnitude of the connections between vertices. When drawing graphs with huge numbers of vertices (>5,000 for example), edge weights may markedly support inference, especially when analyzing the topological structure configuration of a graph.

Since genetic information flows from parents to offspring, pedigrees are optimally modeled as directed graphs where each between-vertex connection reflects a parent-child relationship. The in-degree (edges that arrive at the vertex) of each vertex is maximally two, while the out-degree (edges that depart from the vertex) will depend on how many offspring an individual has. Since individuals cannot be their own ancestors, the graph is acyclic as well as being directed.

Different graph properties can provide valuable inference support on both the complexity and organization of genealogical structures. In the following section we briefly describe some properties of standard graph theory, and how each feature may be associated with current applications in animal population studies. Code (in Python 2.7 programming language) for the implementation of all methods discussed are included as supplementary material, and can be easily applied or extended to other pedigree files from other sources. It is important, though, to emphasize that as not much optimizations

were applied to the code, some presented functions may take a long time for pedigrees with more than 100,000 individuals.

2.2.2 Graph centrality

Centralities are quantitative measures that refer to properties of the vertices and indicate their importance based on their position in the graph. Every vertex has some degree of influence over the graph structure. Thus, vertices with high values of centrality assume a critical role for the organization of a graph. For each type of centrality measure, a different concept of “importance” is assumed (BONACICH, 1987; BORGATTI AND EVERETT, 2005). Distinct degrees of complexity presented by different graphs may hamper using their comparison by the absolute centrality values. Thus, calculated centrality measures may be (and often are) normalized in order to allow the comparison of different graphs. In the following sub-section, some different centrality measures are presented. Here we present four centrality measures included in the PedWorks computer program and discuss their application and inference when applied over genealogical data.

2.2.2.1 Degree Centrality

The degree centrality is considered a local measure and refers to the number of edges connected to a vertex. It may be also understood as the fraction of vertices a vertex v_i is connected to (FREEMAN, 1979). In the case of directed graphs, one may split the concept of degree centrality into “in-degree” and “out-degree” centralities. In the case of directed graphs, in- and out-degrees of a vertex can be calculated as $k_i^{in} = \sum_j A_{ij}$ and $k_i^{out} = \sum_j A_{ji}$, respectively, where A is the adjacency matrix. In practice, the in-degree for

a given vertex i is the sum of all columns in its respective row in the adjacency matrix. Similarly, the out-degree of a given vertex i is the sum of all rows in its given column in the adjacency matrix. In the scope of genealogical structured data, these can be understood as parent-child (out-degree) and child-parent (in-degree) relationships. For obvious reasons, maximum in-degree of a vertex located inside a pedigree graph is 2, its sire and dam. This is an easy concept that should be used to check for errors in the pedigree; that is, when modeling a pedigree by graph theory, the occurrence of vertices (animals) with in-degree higher than 2 must always be checked and corrected. For simplicity, from here on the term “degree centrality” will be used to refer only to out-degree centrality in this paper.

Animals with high degree centrality in a pedigree-graph are important for many reasons. First, these individuals are connected to a larger number of other individuals in the next generation (i.e., their progeny), making them key elements for the topological organization of the studied population. However, degree centrality of a certain vertex does not consider connections beyond direct/immediate (parent-child) connections. The number of out-connections (progenies) from the progeny of an individual is not considered for its degree centrality calculation. In practical terms, this means that individuals with large progeny will be considered “relevant” and have a high (out)degree centrality, even if their progeny do not have descendants. When deeper connections are of interest to determine vertex importance the Katz centrality is more suitable and will be discussed further in this paper.

2.2.2.2 Closeness centrality

Closeness centrality is based on the average length of the shortest paths between one vertex and every other vertex in the network (FREEMAN, 1979). In practical terms, it is a measure of how close a vertex is to every other vertex in the network at the same time. In computer and social sciences, closeness centrality is also referred to as the grade that defines how fast the information flowing in the network gets to a specific vertex (OKAMOTO et al., 2008).

Closeness centrality is calculated as:

$$C_c(v) = \left[\frac{1}{n-1} \sum_{u=1}^{n-1} d(v, u) \right]^{-1}$$

where $d(v, u)$ is the shortest-path distance between v and u ; and n is the number of vertices in the graph.

Information definition in genealogical data is an important subject as it may be seen as the gene flow in the population. In that aspect, individuals with higher closeness centrality are the ones “located” near to all sources of different genotypes in the population.

2.2.2.3 Betweenness Centrality

The formal definition of betweenness centrality refers to the likelihood that information originated anywhere else in the network reaches a specific vertex (ANTHONISSE, 1971; FREEMAN, 1977). Thus, betweenness centrality shows important

vertices that lie on a high proportion of paths between other vertices in the network.

Betweenness centrality, as defined by BRANDES (2008), is calculated as:

$$C_B(v) = \sum_{s,t \in V} \frac{n(s, t | v)}{n(s, t)}$$

where V is the set of vertices, $n(s, t)$ is the number of shortest (s, t) -paths, and $n(s, t | v)$ is the number of those paths passing through some vertex v other than s, t . If $s = t$, $n(s, t | v) = 1$, and if $v \in s, t$, $n(s, t | v) = 0$.

When applied over genealogical data, individuals with high betweenness centrality in a pedigree-graph may have, for example, a relevant role of connecting isolated groups/families in the pedigree. This property is particularly useful when analyzing populations with sparse genetic relationship, defined by animals from distant families (or lineages).

2.2.2.4 Katz centrality

The Katz centrality (KATZ, 1953) is a generalization of the Eigenvector centrality. This algorithm computes the centrality for a node based on the centrality of its neighbors. The Katz centrality for node i is

$$x_i = \alpha \sum_j A_{ij} x_j + \beta$$

where A_{ij} is the adjacency matrix of the graph with eigenvalues λ .

The parameter β controls the initial centrality and $\alpha < \frac{1}{\lambda_{max}}$.

Katz centrality computes the relative influence of a node within a network by measuring the number of the immediate neighbors (first degree nodes) and also all other nodes in the network that connect to the node under consideration through these immediate neighbors. The parameter β controls the weight given for the number of directly connected nodes (here represented by progeny). Distant connections are, however, penalized by the parameter α which acts as an attenuation factor, giving more or less importance on deep connections. It must be warned that α should be strictly less than the inverse largest eigenvalue of the adjacency matrix in order for the Katz centrality to be computed correctly.

When observed under genealogical structured data (pedigrees) files modelled as networks, individuals with high Katz centrality values are the ones whose progeny also has a large number of offspring. This value may help detection of important ancestors in the population, especially when referring to important bulls, which are often responsible for a large number of descendants.

2.2.3 Degree rank function

It has been argued that the degree rank function, which describes the relationship between the degree d (often expressed in a logarithmic scale) and the rank r of a degree sequence, is an important statistical property to characterize a graph (WU et al., 2008). In the scope of genealogical structured data, these measures are naturally split into in- and out-degree rank functions, for the previously discussed inherent properties. In-degree rank function is highly associated with pedigree completeness, as it will explicit the proportions of vertices with 0, 1 and 2 in degrees, which is easily understood as individuals with 0, 1 or 2 known parents. The out-degree rank function, on the other hand,

provides information on the frequency of vertices with a given number of out-degree, which can be interpreted as the number of individuals with a given progeny size. Measuring the frequency of individuals with specific numbers of direct descendants may help to unravel differential contribution of progenitors in the population. The PedWorks software is capable of generating degree rank function plots for pedigrees.

2.2.4 Community detection

A key property of many graphs is their community structure, defined as a set of vertices that are highly connected to each other, but weakly connected to other vertices in the graph (BALL, et al., 2011). These may be calculated considering weights attached to the edges, if not, all edge weights are considered to be identical. In practice, community identification helps to uncover existent functional structures in the network (Fortunato, 2010).

Network community detection, in essence, requires the partition of a network into modules. That is not always an easy task, especially for larger graphs as is the case of genealogical structured data in livestock populations. Solutions for this optimization problem are known to be complicated. Several algorithms have therefore been proposed to find reasonably good partitions in a reasonably fast way, but no method is perfect and different results may arise depending on the algorithm used and the population structure.

The quality of the partitions resulting from these methods is often measured by the modularity (NEWMAN and GIRVAN, 2002). Modularity is a metric that quantifies the quality of the assignment of vertices to communities by evaluating how much more

connected the vertices inside a community are compared to how connected they would be, on average, in a random graph. The modularity of a partition is a scalar value between -1 and 1 that measures the density of connections inside communities as compared to connections between communities. In general, modularity is defined by the Louvain method (BLONDEL et al., 2008):

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i c_j),$$

where A_{ij} represents the magnitude of the connection between vertices i and j , which is often assumed to be 1 (connected) or 0 (not connected) in standard graphs, $k_i = \sum_j A_{ij}$ is the sum of the weights of the edges attached to vertex i , c_i is the community to which vertex i is assigned, the δ function $\delta(u, v)$ is 1 if u is connected to v and 0 otherwise and $m = \frac{1}{2} \sum_{i,j} A_{ij}$. In the context of weighted graphs (graphs that have weights “attached” to their edges), A_{ij} values are not identical for all edges, as each connection is attached to a certain value (weight) that express the magnitude of that connection.

Blondel et al., 2008 implemented a community detection method that can deal with genealogical structured data (modeled as a graph), even with larger pedigrees, as it minimizes graph size limits imposed by other methods in the literature. The algorithm consists of two main components, repeated iteratively. At a first moment, it is assigned a different community to each vertex of the graph. Thus, the number of communities is the same as the number of vertices. From that, for each vertex i , its neighbors (vertices directly connected to vertex i) are considered in order to evaluate the change in modularity (which is a global parameter) that would take place by adding vertex i to the community

of a vertex j . After iterating over all $n - 1$ vertices in the graph, vertex i is then placed in the community for which the gain in modularity is maximum (and positive). If no positive gain is possible, i is maintained on its original community. This process is applied repeatedly and sequentially for all vertices until no further improvement can be achieved. This first phase stops when a local maxima of the modularity is attained (when no individual move can improve the modularity). Gain in modularity (ΔQ) obtained by moving an isolated vertex i into a community C is computed by:

$$\Delta Q = \left[\frac{\sum i n + k_{i,in}}{2m} - \left(\frac{\sum tot + k_i}{2m} \right)^2 \right] - \left[\frac{\sum i n}{2m} - \left(\frac{\sum tot}{2m} \right)^2 - \left(\frac{\sum k_i}{2m} \right)^2 \right]$$

where in is the sum of the weights of the edges inside C , $\sum tot$ is the sum of the weights of the edges incident to vertices in C , k_i is the sum of the weights of the edges incident to vertex i , $k_{i,in}$ is the sum of the weights of the edges from i to the other vertices in C and m is the sum of the weights of all the links in the graph. A similar expression is used in order to evaluate the change of modularity when i is removed from its community. In practice, the change of modularity is checked twice per iteration, first by removing i from its community and then by moving it into a neighboring community.

The second part of the algorithm consists on building a new network whose vertices are now the communities found during the first phase. To do so, the weights of the edges between the new vertices are given by the sum of the weight of the edges between vertices in the corresponding pair of the connected communities (ARENAS et al., 2008). Once this second phase is completed, it is then reapplied the first phase of the algorithm to the resulting weighted network. By construction, the number of meta-communities decreases at each pass. The passes are iterated until there are no more

changes and a maximum of modularity is attained. The algorithm incorporates a notion of hierarchy, as communities of communities are built during the process. The height of the hierarchy that is constructed is determined by the number of passes and is generally a small number.

The described algorithm has relevant properties. First, it is extremely fast, handling larger networks (hundreds of thousands of vertices). Second, it is intuitive and easy to implement, and the results does not need much effort to interpret. Resolution limit problem of modularity also seems to be overcome by the intrinsic multi-level nature of this algorithm. Fortunato & Barthélemy (2007) pointed that modularity optimization could fail for communities smaller than $\sqrt{2L}$, where L is the total number of connections in the graph. however, small communities are of no interest for genealogical structured data studies. This issue is often referred at as “resolution limit” and involves the wrongful merging of smaller communities due to higher grades of complexity in larger graphs. However, the algorithm applied by Blondel et al., (2008), for displacing vertices one by one iteratively through the communities, diminishes the probability of merging communities.

Additional methods to deal with the resolution parameter come from the definition of a “time” parameter (LAMBIOTTE et al., 2009), which intend to accommodate different community sizes for the same graph. In practical terms, when decreasing this “time” parameter, the number of communities tend to V (the total number of vertices), where its maximum value yields a partition of V one vertex communities. In the other hand, when increasing this parameter, the partition tends to stabilize in a two-way configuration. This approach helps when studying genealogical structured data under a graph framework, in which different populations have distinct grades of complexity and therefore, one may be

interested in assessing the partition configuration that yield the best modularity.

The PedWorks software makes use of the method proposed by Blondel et al. (2008) to assess community partition over pedigree-graphs. Results of the community detection routine included in the program indicate the number of detected communities, the final modularity values for the current analysis and the community for which each animal was determined.

2.2.5 Graph visualization

As stated before, not much effort has been given to genealogical structure data visualization in the literature. Based on that, we discuss the application of a specific type of graph drawing algorithm that can help to represent on a 2-dimensional space the relationship between individuals (and groups of individuals) in a pedigree.

Force-directed network drawing algorithms (and the Fruchterman & Reingold method)

Force-directed layout algorithms are well-known in graph drawing literature, and its currently the one of the most used class of algorithms for calculating layouts to solve common network problems. Such algorithms calculate the layout of a graph based exclusively on information contained within the structure of the graph itself, rather than relying on domain-specific knowledge (KOBOUROV, 2012). Graphs drawn by these algorithms tend to exhibit symmetry and to produce crossing-free layouts for planar graphs (reduced number of edges crossing to facilitate visualization). Traditionally, force-directed methods use the graph's structure to mimic a physical system of attractive spring

forces along edges and global repulsive forces emanating from the vertices (Bannister et al., 2012).

The Fruchterman & Reingold algorithm (FRUCHTERMAN and REINGOLD, 1991) focuses on dealing with the most general class of graphs: undirected graphs drawn with straight edges. It attempts to produce two-dimensional drawings of graphs by performing simplified simulations of physical systems. Also, Fruchterman & Reingold's force-directed method presents the concept of "temperature", which is a control parameter for vertex displacement between iterations. At the beginning of the algorithm, the "temperature" parameter starts at an arbitrary value, and then it approaches linearly towards zero as the algorithm iterates. As the temperature decreases, the vertex adjustments between iterations become smaller. On each iteration, the basic algorithm computes $O(|E|)$ attractive forces and $O(|V|^2)$ repulsive forces. Considering that repulsive forces will act from all vertices to all other vertices, distant vertices have their repulsive forces dispersed in order to control the quadratic complexity imposed by the algorithm when it is applied to large graphs.

As proposed by Fruchterman & Reingold (1991), those forces are defined by:

$$f_a(d) = \frac{d^2}{s} \quad \text{and} \quad f_r(d) = \frac{-s^2}{d},$$

where $f_a(d)$ and $f_r(d)$ are the attractive and repulsive forces, respectively, d is the distance between two vertices, and s is the optimal distance between vertices, which is often defined as

$$s = C \sqrt{\frac{\text{area}}{V}}$$

where C is a constant (often obtained by experimentation), $area$ is the total area where the graph is to be drawn and V is the total number of vertices in the graph. As the number of iterations increases, theoretical attractive forces attached to the edges make strongly connected vertices to get closer, defining distinct interconnected groups of vertices.

An illustration of how the Fruchterman and Reingold's algorithm designates vertices positions as it iterates is presented in Figure 2.1.

As mentioned before, edge weights are commonly considered graph attributes, which allow for modeling more complex and inconstant relationship measures between different nodes. In the scope of pedigree graphs, the relationship coefficient (f_i) between all $n \times n$ individual connections, when attached to the edge weights connecting those respective nodes, can help to characterize different magnitudes of connectedness. In practice, and for the analyzes presented in this study, these weights were added to the off-diagonals of the adjacency matrix, by adding f_{ij} to each element of A_{ij} that share a connection. This approach has the objective of achieving a final node positioning that best reflects the relationship of all individuals in the pedigree.

It needs to be noted that it has been mentioned before that force-directed algorithms were based on undirected graphs to work properly (EADES, 1984, FRUCHTERMAN and REINGOLD, 1991 KAMADA and KAWAI, 1989). However, in this paper we have also stated that genealogical structured data are optimally modelled as directed networks, which leads to a theoretical conflict.

Figure 2.1 Illustration of the progress made by Fruchterman and Reingold's force directed algorithm at iterations 1 (A - vertices randomly placed), 50 (B) and 500 (C) when applied over a pedigree.

Authors have described different strategies to deal with directed graphs drawing. One common way to work around that obstacle is to consider the directed graph as undirected in order to apply forced-directed algorithms (by assuming symmetry of the edge's directionality). There is some criticism on this approach, as it, in many cases, may assume

a nonexistent symmetry among connections over vertices in a graph. When considering a pair of connected vertices A and B as undirected, it means to assume that the connection coming from A to B is identical to the connection from B to A. In many fields, this assumption will interfere over domain-specific properties of the studied object. But when considering genealogical structured data modeled as a graph, as stated before, connections between vertices express parent-progeny relationship, which in practice may be easily understood as an indication of direct kinship. Considering that the (expectation of the) genetic relationship between animal A and animal B is symmetrical to the genetic relationship from B to A, it gets clear that, for drawing purposes, we don't really lose or misinterpret information when considering a "pedigree graph" as undirected.

2.2.5.1 Force directed algorithm and large pedigrees

Graphs larger than 20 000 nodes or strongly connected (too many edges) graphs are known obstacles for when drawing graphs by Fruchterman and Reingold's algorithm. The latter is not a real problem when modeling genealogical structured data, since connections between individuals (represented in the adjacency matrix) are absent (zero) in the same intersections as in the relationship matrix, which is predominantly sparse. However, pedigrees with more than 20 000 individuals are much more common, especially when considering bigger cattle breeds. In this case, Fruchterman and Reingold's algorithm struggle to maintain nodes well segregated in the resulting final draw, generating groups of overlapping nodes in the center of the graph representation. Those overlapping nodes markedly impair the ability to infer over any topological structure of pedigree graphs. Although such optimizations are not the main focus of this paper,

pointing and discussing this issue may help future applications of force directed algorithms applied to drawing pedigree graphs. Thus, we present a method to prevent nodes from overlapping (not implemented in the code presented in supplementary material).

The concept of non-overlapping nodes in graph representations require the knowledge that despite being theoretically represented by points, nodes are often illustrated as circles, thus, each node occupies a specific area in the plane. A simple implementation for preventing node overlapping consists of defining a minimum distance between nodes $d'(n_1, n_2)$ that is bounded between $s(n_1) + s(n_2)$ and ∞ , where s is the size of the respective node. This minimum distance may be achieved by increasing the repulsive force for each pair of nodes that have $d'(n_1, n_2) < s(n_1) + s(n_2)$ after each iteration of the Fruchterman and Reingold's algorithm. Although this method may increase computational cost, it is simple to implement in any programming language and in many cases may be a way to turn visualization of larger graphs a feasible task.

2.2.6 Application of the graph framework to genealogical structured data (pedigrees)

In this section, we implement the above discussed methods on both simulated and field genealogical data using the routines included in the PedWorks software. Comparison of the results obtained in different proposed examples are discussed to support inference for future applications.

2.2.6.1 Simulated data

To illustrate how pedigrees may be optimally modeled by graph theory and how differences in genealogical structure data are reflected in the graph architecture, two populations were simulated using QMSim (version 1.10, Linux) software package (SARGOLZAEI and SCHENKEL, 2009). Both simulations were idealized to represent cattle populations, one as a strongly connected population and the other as a less (genetically) connected population, in order to show and discuss how both pedigrees can be analyzed as graphs. Simulations were firstly based on Brito et al. (2001), with slight modifications to obtain the specific characteristics desired for this study, which were, in practice, a low mean inbreeding coefficient (SprPop) and a high mean inbreeding coefficient (NrwPop) populations. A historical population was created starting with 1000 animals, which randomly mated for 1000 generations maintaining the same size. From generation 1000 this HP was reduced to 200 individuals in 20 generations. An expansion population (EP) was then created, considering 100 males and 100 females (SprPop) and 5 males / 100 females for (NrwPop), which lasted for 8 generations of random mating and had an exponential growth rate. From this, a recent population (RP) was created, using 500 males / 10 000 females (SprPop) or 50 males / 10 000 females (NrwPop) from the last generation of EP, during 10 generations. This last population had also differences on sire and dam replacement rates, respectively, 0.60/0.20 for SprPop and 0.20/0.20 for Nrwpop. All other details were set exactly as described in Brito et al. (2001). Inference was taken considering the last 5 generations from RP.

2.2.6.2 Real data

Two existing cattle populations were also analyzed in order to demonstrate how real pedigrees can be studied when modelled by graph theory and to discuss practical applications of this methodology. Genealogical structured data from two *B. indicus* breed populations in Brazil, Guzerá and dairy Gir, were analyzed as graphs. Pedigree files used in the present study were obtained from CBMG (Centro Brasileiro de Melhoramento Genético do Guzerá) for the Guzerá breed, and from ABCGIL (Associação Brasileira de Criadores de Gir Leiteiro) for the dairy Gir breed population.

The Guzerá pedigree file considered in this research contained 11,776 individuals, 1,280 unique sires and 5,568 unique dams. The dairy Gir population consisted of 42,420 individuals, of which 2,796 were unique sires and 17,228 were unique dams. Table 2.1 shows population structure data for both populations obtained using CFC software (SARGOLZAEI, 2006).

Table 2.1 Population structure results for simulated populations A and B; and for real Guzerá and dairy Gir populations considered in the present study.

Item	NrwPop	SprPop	Gir	Guzerá
Total number of individuals	60,050	60,500	42,248	11,776
Unique sires	90	1,700	2,796	1,280
Unique dams	18,000	18,000	17,228	5,568
Indiv. with no progeny	41,960	40,800	22,472	4,930
Number of founders	10,050	10,500	8,160	2,710
% of inbred individuals	09.22	00.90	37.47	31.19
Average F	00.70	00.03	0.90	0.80
Average F inbred	7.63	3.84	2.41	2.57
Range	0.00 – 37.50	0.00 – 0.25	0.00 – 37.50	0.00 – 31.64

The community detection algorithm proposed by Blondel et al. (2008) was used to generate the best partitioning for both Guzerá and Gir pedigrees (as graphs), using the modularity as a measure of quality of the partition. For each population, five individuals with the highest values of out-degree centrality (from the five communities with the highest out-degree centrality means) were considered in order to illustrate how community detection algorithms behave when applied to real pedigree data. The average relationship, obtained by $AR_i = \frac{1}{N-1} \sum A_{i-}$, was calculated for each considered individual, accounting for its relationship with the overall population and within its own community.

Since no application of Fruchterman and Reingold's algorithm over genealogical data was found in the literature, there was no information about the number of iterations needed to obtain a minimum energy state, and consequently, the node positioning that best reflect relationship between the individuals in the pedigree. For that, the stability of the final node positioning and its variation between different iterations was investigated by evaluating the distribution of node distances between iterations intervals. Tested iteration number varied from 1,000 to 20,000 iterations. Stability was considered to be achieved when the distribution of node distances was minimal between iteration intervals.

2.2.7 Correlation between relationship coefficients and the distance between vertices

Once modeled under graph theory framework and plotted by Fruchterman & Reingold's force directed algorithm, pedigrees may be represented as a vector of coordinates (x and y-axis), which locate each vertex (individual) in a two-dimension plane.

From that, the distance (on a straight line) from each vertex and every other vertex in the graph can be calculated as $d(i, j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$, where x and y are the coordinates in the x and y-axis, respectively, for the pair of vertices i and j . This allows the construction of a matrix of distances between vertices, here referred as D matrix (of order $v \times v$), where v is the number of vertices in the graph. At the same time, the conventional relationship matrix, often called the A matrix (HENDERSON, 1976) for animal breeding purposes, which contains the genetic relationship expectancy between all individuals in the pedigree, was created and all its off diagonal elements (containing the expected relationship coefficients between the n individuals) was estimated. After that, each distance between vertex x_i and vertex x_j (where $i \neq j$) was correlated to the relationship coefficients between animals i and j from the A matrix (considering only non-zero relationship coefficients) in order to investigate the presence of linear associations between both calculated measures.

2.3 RESULTS AND DISCUSSION

2.3.1 Graph general structure

As stated before, visual analysis of graphs after the application of drawing algorithms may help to unravel underlying structures, not recognizable by other methods. Supplementary material 1 shows drawings of populations A and B as graphs using an implementation of the Fruchterman and Reingold's force directed algorithm, implemented by the code in supplementary material making use of NetworkX python package (HAGBERG et al., 2008) for its capabilities towards dealing with graphs. The circular

topology of both graphs are a result of the symmetry imposed by this drawing method, which is more evident in smaller graphs. Vertices (representing the individuals) were colored according to the community which they were designated by the community detection algorithm (BLONDEL et al., 2008). This algorithm, when applied for pedigree graphs, may yield a large number of one-vertex communities. Those are individuals with no connections within the population (no parents or progeny). Communities with less than three individuals were colored white/transparent to simplify visualization and to identify one- and two-vertex communities, as those can't really be considered as consistent communities.

Table 2.2 shows a general graph structure description for both simulated and real populations studied.

Table 2.2. General graph structure for the Guzerá and Gir population pedigrees.

Item	Vertices	Edges	Avr. Outdegree ¹	Density	Number of Communities ²
SprPop	60,500	100,000	1.653	2.73e-5	110
NrwPop	60,050	100,000	1.665	2.77e-5	85
Guzerá	11,776	18,018	1.530	1.30e-4	61
Gir	42,428	68,292	1.610	3.79e-5	34

¹ Average node's outdegree centrality, ² Considering a resolution parameter of 1.

When comparing both simulated populations (SprPop and NrPop), which had almost the same number of individuals, it gets clear that the sparser the population (more poorly connected), the higher the number of independent communities (Table 2.2). A similar behavior was observed for the real (field) data. The Guzerá population showed a higher

number of communities when compared to the Gir population. Differences in population size should be considered cautiously, however, it is plausible to consider that a less connected population will exhibit a higher number of communities, as families will be smaller than in very well connected pedigrees.

A general stability test for both Guzerá and Gir pedigrees drawn by Fruchterman and Reingold's algorithm is illustrated in Table 2.3. The mean Euclidean distance of the nodes between iteration intervals clearly diminished as iteration number increased. After 4,000 iterations, node distances between iteration intervals start to oscillate around a minimum value. This was considered as the (possible) minimum energy state for each given pedigree. When applying the Fruchterman and Reingold's algorithm over such large graphs (over 5,000 node), it is really difficult to obtain an absolute minimum-energy state (null node displacement on subsequent iterations), as there are too many attractive and repulsive forces acting over each node. Node displacement, in practice, may never reach an absolute zero-energy state because of the natural axis rotation promoted by Fruchterman and Reingold's algorithm. We have considered that after 4,000 iterations, most of the node displacement (between iteration intervals) observed (for both Guzerá and Gir pedigrees) occurred due to axis rotation. Thus, after the (possible) minimum-energy state, graph's visual aspect remains almost intact, and consequently, so is inference. Bigger populations may require more iterations to stabilize when under a graph approach.

Table 2.3. Mean and median of the difference of the Euclidean distance of nodes position between iterations y_1 and iteration y_2 . Total graph area (scaling parameter) was considered to be 5,000.

Iteration intervals	Guzerá		Gir	
	Mean	Median	Mean	Median
100 x 1,000	484.77	419.12	372.77	311.38
1,000 x 2,000	380.82	298.57	297.82	225.65
2,000 x 3,000	272.94	195.54	245.36	189.42
3,000 x 4,000	190.34	125.84	202.55	132.87
4,000 x 5,000	238.72	187.85	188.35	154.81
5,000 x 6,000	159.62	141.22	169.17	143.48
6,000 x 7,000	186.51	110.58	158.88	105.94
7,000 x 8,000	161.28	94.90	145.23	99.48
8,000 x 9,000	166.72	87.46	165.44	102.35
9,000 x 10,000	170.27	92.70	182.78	96.29

Visual assessment of pedigree data by graph theory may be optimized according to the population structure and size. Different values for the resolution parameter (Lambiotte, 2009), often called “time” parameters, previously (although briefly) described in this study, may help to define the best partition for a given pedigree graph. To establish the best community partition structure using real genealogical data, three different values for the resolution parameter (RP) were tested: 0.50, 1.00 and 1.50. The RP is often set to one in most cases, and may often yield the best modularity values, but it can be shifted in order to establish better partition configurations for each specific situation. Modularity values were 0.76, 0.77 and 0.76 for the Guzerá population and 0.72, 0.73 and 0.71 for the Gir population, respectively for RP=0.50, RP=1.00 and RP=1.50. The number of communities detected was 65, 34, 25 for the Guzerá population and 100, 61 and 46 for the Gir population, respectively, using the above RP values. This results reveal that there is no linear association between the RP values and the changes in modularity. This must

be taken into account when applying community detection to other populations.

2.3.2 Graph analysis

Rank plots were built for all populations (simulated and real) studied. Vertices are ordered by their (out)degrees, and then plotted against the frequency of each values in a log scale. This type of plot helps to elucidate the distribution of the number outgoing edges from the vertices in a network and check for the occurrence of gaps or inequalities. The (out)degree rank plots for the simulated (and from real) populations are shown in Figure 2.2. For population NrwPop, the difference in number the of progenies from the individual with most progenies and the bottom 50% is much larger than observed for population SprPop. The higher number of sires/dams per generation on population SprPop caused the minor difference between the number of progenies per individuals. The visual analysis of the (out)degree rank plot helps to infer over differences in the number of progenies from individuals throughout the entire population, giving valuable information about the population structure.

Both Guzerá and Gir populations exhibited similar rank plots (Figure 2.3), both indicating fewer individuals with higher number of progenies and higher number of individuals with fewer progeny. However, the decay in the number of progenies behave differently between both populations. Also, the magnitude of node's out-degree in Guzerá and Gir pedigrees remarkable. The individual with highest out-degree had around 10^3 (Gir), while in Guzerá population, the sire occupying the same position had around 10^2 out-degrees (progenies). In the rank plot from the Guzerá population there is a clear

smooth decay directed to the fewer progeny individuals, while in Gir population, there is a slight persistency when decaying from the individuals with very high numbers of progenies to individuals with fewer progeny.

Figure 2.3 (Out)Degree rank plots from simulated SprPop (upper left) and NrWPop (upper right) populations; and for real Guzerá (lower left) and Gir (lower right) populations considered in this study.

This detour may indicate pronounced variation on the number of progenies of certain

Gir sires when compared to sires from the Guzerá population. When comparing the percentage of unique sires (over the total population) for both populations of 6.60% (Gir) and 10.87% (Guzerá), and the percentage of individuals with no progeny (Table 2.1) of 37.47% (Gir) and 31.49% (Guzerá) corroborate to the idea that there is a concentration of sires with a high number of progenies in the Gir population. In Figures 2.4 and 2.5, it is shown the Guzerá and Gir pedigrees drawn by the Fruchterman and Reingold's algorithm (4,000 iterations, considering a scaling parameter of 5,000 and $k=0.003$) using the previously mentioned code built for this purpose. In addition, nodes were colored by the community in which they were allocated by the Louvain community detection method, considering a time parameter of 1.00. Five bulls (of each population) pointed as having the highest outdegree centralities in their own communities were highlighted. In both cases, these bulls were coincident with breeders' opinions on the most important sires for the Guzerá and dairy Gir populations.

A brief analysis of community structure applied to livestock populations pedigree is shown in Table 2.5. For both the Guzerá and Gir populations, the relevant individuals (that had the higher out-degree centrality value) inside each considered community had much higher average relatedness between individuals within its community ($\bar{f}_C$) than when considering the entire population ($\bar{f}_P$). This emphasizes the importance of community partitioning algorithms for detecting groups of strongly connected individuals, but sparsely connected with the rest of the population.

Figure 2.4. Guzerá pedigree drawn by Fruchterman & Reingold's force directed algorithm, considering 4,000 iterations ($k=0.003$). Node's colors determine its community.

2.3.3 Linear association between values of A and D matrices

Two different sets were used for each population: the first one contained all elements from A and D matrices (excluding diagonals) and the second contained only non-zero elements of the A matrix and its correspondent i and j values of the D matrix. Results for the study on linear associations between elements in the A and D matrices are presented in Table 2.5.

Figure 2.5. Gir pedigrees by Fruchterman & Reingold's force directed algorithm, considering 4,000 iterations ($k=0.003$). Node's colors determine its community

As expected, the number of non-zero elements in the A matrix was relatively high in Gir population when compared to the Guzerá population, where the latter exhibited higher percentage of zero elements in the A matrix. Similarly, the average numerator relationships (excluding self-relationship) were higher for Gir, in detriment to Guzerá population. These results support the observation that the Gir population is slightly more connected than the Guzerá population.

Table 2.4. Number of individuals within each considered community (N), the individual with the highest out-degree centrality value within its community (topID), its inbreeding coefficient (F), average relationship ($\bar{f}_P$), average relationship with individuals from out of its community ($\bar{f}_C$) and number of progeny (NoP) for the Guzerá and Gir pedigrees.

Item	C (1)	C (2)	C (3)	C (4)	C (5)
N	417	509	720	348	343
topID	"7866"	"TABO636"	"9940"	"A1437"	"CNS4995"
F	0.062	0.000	0.000	0.000	0.072
Guzerá $\bar{f}_P$	0.032	0.026	0.041	0.034	0.025
$\bar{f}_C$	0.144	0.152	0.159	0.197	0.135
NoP	169	142	152	207	134
N	2,111	1,766	1,632	1,362	1,144
topID	"KCA427"	"A7481"	"A7368"	"GAV291"	"EFC441"
F	0.012	0.091	0.000	0.000	0.020
Gir $\bar{f}_P$	0.066	0.067	0.040	0.042	0.047
$\bar{f}_C$	0.214	0.135	0.147	0.150	0.161
NoP	2,461	1,657	1,147	978	771

C = Community, (*) = community identification number.

The Pearson correlation between elements of the A and D matrices varied from -0.42 to -0.44 when considering all elements; and from -0.60 to -0.68 when considering only non-zero elements in the A matrix. The negative nature and magnitude of the coefficients indicate that the distance between vertices after Fruchterman & Reingold's drawing algorithm is linearly associated to the relationship coefficients described by Wright (1931). Therefore, higher distances between vertices are associated to lower relationship coefficients. The distance between each vertex of a pedigree graph drawn by Fruchterman & Reingold's force-directed algorithm may, thus, be considered as an indicator of the expectancy of genetic relationship between those individuals. The slight lower P_{AxD} values for Gir populations explicit that as general relative connectedness increases between individuals in a pedigree graph, the relationship between repulsive

(f_r) and attractive (f_a) forces calculated by the Fruchterman and Reingold's algorithm tend to mutually interact, deviating the distance between vertices i and j from their relationship coefficient calculated in the A matrix.

Table 2.5. Total number of elements in the A matrix, total number of non-zero elements in the A matrix, average numerator relationship (excluding self-relationship) on the A matrix, Pearson correlation coefficient between upper diagonal elements of A and D matrices ($P_{AxD(z)}$) and between only non-zero elements of A matrix and its correspondent in D matrix (P_{AxD}) for all populations studied.

Population	Total elements in A	Non-zero elements in A	Average NR (%) ¹	$P_{AxD(z)}$ (with zeros)	P_{AxD} (non-zeros)
Guzerá	138,674,176	40,224,725	0.89	-0.44	-0.68
Gir	1,799,626,084	719,850,434	1.76	-0.42	-0.60

¹ Average of the numerator relationship in the A matrix, excluding self-relationship (diagonal).

2.3.4 Further studies, current and future applications

Graph theory framework offers a vast diversity of applications when modeling genealogical structured data. We believe that one of the most promising opportunities involve the usage of centrality measures to identify key individuals in the population in order to support decision making in a variety studies. This information, in conjunction with results yielded from community detection algorithms, are the basis for consistent innovative methods for detecting important, but mutually sparsely-connected individuals. Those would fit as candidates for genotyping strategies in livestock populations with few or no molecular information derived from Single Nucleotide Polymorphism (SNP) panels.

Improvements on pedigree visual evaluation methods, considering the distance between vertices as approximations of the “genetic relationship” between individuals, also have a lot of potential. This spatial-like approach for assessing relationship between individuals and/or groups of individuals in the population could support further population

studies in animal science. A “visual mating tool”, making use of this later approach may easily be implemented by extending the code presented in supplementary material. A possible resulting plot is presented in Figure 2.5, showing males colored by their EBV (varying from red to blue, respectively, for low and high EBV values), and females colored white, kept in the graph for structure maintenance purposes. With this information, one can easily locate high genetic merit and sparsely connected (implying a lower relationship coefficient) candidates for planned matings with previously selected females (in this example, pointed in the graph by the arrow and colored yellow).

At last, graph theory offers an innovative approach for analyzing multi-breed livestock populations. As an example, we included a function for completing breed composition values for individuals with no breed information, by assessing compositions from in degrees of each vertex in the graph, in the Python (2.7) code presented in the supplementary material.

Figure 2.6. Illustration of a possible application of graph theory for the development of a

visual mating tool.

2.4 CONCLUSION

Graph theory framework was successfully applied using PedWorks software using both simulated and real pedigree data. Graph centrality seems to be a reliable parameter for finding influential individuals in animal populations. Community detection algorithms, when applied over pedigrees, help to identify its underlying structure and may have many usages such as developing new genotyping strategies for populations with few or no genotype records. Drawing pedigrees (as graphs) using Fruchterman and Reingold's force directed algorithm yield plots that may indirectly represent expected genetic connection between different groups and may be used in further population studies. The presented software is already computationally efficient but further improvements in terms of run time could be considered by implementing the same functions using another programming language such as C++ or FORTRAN to run even faster. More studies will help to develop more accurate methods and new applications for graph theory when applied to pedigrees.

2.5 REFERENCES

- ALBERT R. & BARABÁSI A.L Statistical mechanics of complex networks. **Reviews of Modern Physics**, 74, 47-97, 2002
- ANTHONISSE J.M. The rush in a directed graph. **Technical Report**, BN 9/71, Stichting Mathematisch Centrum, 2e Boerhaavestraat 49 Amsterdam, 1971.

- ARENAS A.; FERNÁNDEZ A.; & GÓMEZ S. Analysis of the structure of complex networks at different resolution levels. **New Journal of Physics**, 10, 053039, 2008.
- BALL B.; KARRER B.; NEWMAN M.E.J. Efficient and principled method for detecting communities in networks. **Physical Review**, 84 (3): 036103, 2011.
- BANNISTER M.J.; EPPSTEIN D.; GOODRICH M.T.; TROTT L. Force-directed graph drawing using social gravity and scaling. **In Proceedings of the 20th International Symposium of Graph Drawing**, arXiv:1209.0748, 2012.
- BEAUCHAMP, M.A. An improved index of centrality. **Behavioral Science**, 10, 161–163, 1965.
- BEINEKE L.W.; WILSON R.J. 'Topics in algebraic graph theory'. **Cambridge University Press**, Cambridge, UK, 2004.
- BLONDEL V.D.; GUILLAUME J.L.; LAMBIOTTE R.; & LEFEBVRE E. Fast unfolding of community hierarchies in large networks. **Journal of Statistical Mechanics**, 10, P10008, 2008.
- BOICHARD D.; Maignel L.; & VERRIER E. The value of using probabilities of gene origin to measure genetic variability in a population. **Genetic Selection Evolution**, 29, 5–23, 1997.
- BONACICH P. "Power and Centrality: A Family of Measures". **American Journal of Sociology**. University of Chicago Press, 92: 1170–1182, 1987.
- BONACICH P. Factoring and weighting approaches to status scores and clique identification. **Journal of Mathematical Sociology** 2, 113–120, 1972.

- BORGATTI P. & EVERETT M.G. A Graph-Theoretic Perspective on Centrality. **Social Networks**, 28, 466–484, 2006.
- BRANDES U. On Variants of Shortest-Path Betweenness Centrality and their Generic Computation. **Social Networks**, 30, (2):136-145, 2008.
- CHARTRAND G. & LESNIAK L. 'Graphs and Digraphs', **Chapman and Hall/CRC**, 1966.
- EADES P. A Heuristic for Graph Drawing. **Congressus Numerantium**, 42, (11): 149–160, 1984.
- FORTUNATO S. & BARTHÉLEMY M. Resolution limit in community detection. **Proceedings of the National Academy of Sciences**, USA, 104, 36, 2007
- FREEMAN L.C.; BORGATTI S.P.; & WHITE DR. Centrality in valued graphs: A measure of betweenness based on network flow. **Social Networks**, 13 (2):141–154, 1991.
- FREEMAN, L.C. Centrality in networks: I. conceptual clarification. **Social Networks**, 1:215–239, 1979.
- FRUCHTERMAN, T.M.J.; & REINGOLD, E.M. Graph Drawing by Force-Directed Placement, **Software – Practice & Experience**, 21 (11): 1129–1164. doi:10.1002/spe.4380211102, 1991.
- FORTUNATO S. "Community detection in graphs". **Physics Report**, 486, (3-5): 75–174, 2010.
- GIRVAN M. & NEWMAN M.E.J. Community structure in social and biological networks. In **Proc. Natl. Acad. Sci. USA**, (12): 7821–7826, 2002.

- GUTIÉRREZ J.P.; ALTARRIBA J.; DÍAZ C.; QUINTANILLA R.; CAÑÓN J.; & PIEDRAFITA J. Pedigree analysis of eight Spanish beef cattle breeds. **Genetics, Selection and Evolution**, 35, 43-63, 2003.
- HAGBERG A.A.; SCHULT D.A.; & SWART P.J. Exploring network structure, dynamics, and function using NetworkX. In **Proceedings of the 7th Python in Science Conference** (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), Pasadena, (CA USA), pp. 11–15, 2008.
- HENDERSON C.R. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. **Biometrics**, 32, 69-83, 1976.
- KAMADA T. & KAWAI S. An algorithm for drawing general undirected graphs. **Information Processing Letters**, 31, (1): 7–15, 1989.
- KATZ L. A New Status Index Derived from Sociometric Index. **Psychometrika**, 18, 39–43, 1953.
- KAHN A.B. Topological sorting of large networks. **Communications of the ACM**, Volume 5, Issue 11, 1962.
- KOBOUROV, STEPHEN G. Spring Embedders and Force-Directed Graph Drawing **Algorithms**, [arXiv:1201.3011](https://arxiv.org/abs/1201.3011), 2012.
- LAMBIOTTE R.; DELVENNE J.C.; & BARAHONA M. Laplacian Dynamics and Multiscale Modular Structure in Networks. **IEEE Transactions on Network Science and Engineering**, 1, 2:76-90, 2008.

- LIU R. A Low Complexity Topological Sorting Algorithm for Directed Acyclic Graph. **International Journal of Machine Learning and Computing**, 4, 2:194-197, 2014.
- MEUWISSEN T.H.E; & LUO Z. Computing inbreeding coefficients in large populations. **Genetics Selection Evolution**, 24, 305-313, 1992.
- NEWMAN M.E.J. The structure and function of complex networks. **SIAM Rev**, 45, (2) 167–256, 2003.
- NEWMAN M.E.J. Analysis of weighted networks. **Physical Review**, 70, Issue, 2004.
- OKAMOTO K., CHEN W., & LI X. Ranking of closeness centrality for large-scale social networks. Conference: **Frontiers in Algorithmics, Second Annual International Workshop, FAW 2008**, Changsha, China, June 19-21, 2008.
- REINHARD D. Graph Theory. **Graduate Texts in Mathematics**, Springer-Verlag, v.173, p.428, 2017.
- SABIDUSSI, G. The centrality index of a graph. **Psychometrika** 31, 581–603
- SARGOLZAEI M; IWAISAKI H.; & COLLEAU JJ. CFC: A tool for monitoring genetic diversity. In **Proceedings of the 8th World Congress of Genetics Applied Livestock Production, CD-ROM Communication**, 13-18, 2006.
- SARGOLZAEI, M. and F. S. SCHENKEL. QMSim: a large-scale genome simulator for livestock. **Bioinformatics**, 25, 680-681, 2009.
- YANG L.; CHENG E.; & ÖZSOYOGLU. Efficient path-based computations on pedigree graphs with compact encodings. **BMC Bioinformatics**, 13, Supplement 3, S14, 2012.
- WRIGHT S. Evolution of Mendelian populations. **Genetics**, 16, 97-159, 1931.

WU J.; TAN Y.; DHENG H.; & ZHU D. Relationship between degree-rank function and degree distribution of protein-protein interaction networks. **Computational Biology and Chemistry**, 32, (1):1-4, 2008.

CHAPTER 3. ACCOUNTING FOR POPULATION STRUCTURE IN SELECTIVE COW GENOTYPING STRATEGIES.

Bruno da Costa Perez¹, Julio Cesar de Carvalho Balieiro², Roberto Carneiro⁴, Fabio Tirelo⁵, Gerson Antonio Oliveira Junior⁷, Juliana Machado Dementshuk³, Joanir Pereira Eler⁶, José Bento Sterman Ferraz⁶, Ricardo Vieira Ventura²

¹ *Universidade de São Paulo, Faculdade de Zootecnia e Engenharia de Alimentos, Pirassununga, SP, Brasil.*

² *Universidade de São Paulo, Faculdade de Medicina Veterinária e Zootecnia, Pirassununga, SP, Brasil.*

³ *Universidade Federal do Rio Grande do Sul, Departamento de Zootecnia, Porto Alegre, RS, Brasil.*

⁴ *Universidade Estadual Paulista Julio de Mesquita Filho, Campus de Jaboticabal, Departamento de Zootecnia.*

⁵ *Google Inc., Toronto, Canada.*

⁶ *Grupo de Melhoramento Animal e Biotecnologia, Departamento de Ciências Veterinárias, Faculdade de Zootecnia e Engenharia de Alimentos, Universidade de São Paulo (GMAB-FZEA/USP), Pirassununga, São Paulo, Brasil*

⁷ *Centre for Genetic Improvement of Livestock, Department of Animal Biosciences, University of Guelph, Guelph, ON, Canada.*

ABSTRACT The objective of the present study was to investigate the impact of considering population structure in cow genotyping strategies over the accuracy and bias of genomic predictions. A small dairy cattle population was simulated to address these objectives. Based on four main traditional designs (random, top-yield, extreme-yield and top-accuracy cows), different numbers (1,000; 2,000 and 5,000) of cows were sampled and included in the reference population. Traditional designs were replicated considering or not population structure and compared among and with a reference population containing only bulls. The inclusion of cows increased accuracy in all scenarios compared with using only bulls. Scenarios accounting for population structure when choosing cows to the reference population slightly outperformed their traditional versions by yielding

higher accuracy and lower bias in genomic predictions. Building a cow-based reference population from groups of related individuals considering the frequency of individuals from those same groups in the validation population yielded promising results with applications on selection for expensive- or difficult-to-measure traits. Methods here presented may be easily implemented in both new or already established breeding programs, as they improved prediction and reduced bias in genomic evaluations while demanding no additional costs.

Keywords: graph theory, community detection, genotyping designs, genomic selection, proportional sampling

3.1 INTRODUCTION

Genomic selection (GS) has been implemented in several countries and is now a recognized tool to reduce generation intervals and promote improvements in genetic gains in dairy cattle (PRYCE and DAETWYLER, 2012; GARCIA-RUIZ, 2016). It uses information from genome-wide marker (SNP) effects, estimated from a reference population of previously phenotyped and genotyped individuals, to enhance the prediction of breeding values and to further establish new selection candidates (HAYES et al., 2009), often referred at as the validation population. The accuracy of estimated SNP marker effects is proportional to the size (LUND et al., 2011; JIMÉNEZ-MONTERO et al., 2012), the structure of the reference population (PSZCZOLA et al., 2012; RINCENT et al., 2017), the connectiveness between the reference population and the selection candidates (CLARK et al., 2012; WU et al., 2015, VENTURA et al., 2016), the density of markers recorded (MOGHADDAR et al., 2015), QTL allele frequencies (WIENTJES et al.,

2014) and the effective size of the population from which the reference population was sampled (DAETWYLER et al., 2008).

Mainly in dairy cattle, sires have played a more relevant role than cows in GS for having high numbers of progeny and, therefore, more reliable estimated breeding values (EBV), being main candidates to be genotyped (BOUQUET and JUGA, 2013; CALUS, 2016). However, in specific situations, such as in small populations or when selecting for difficult- or expensive-to-measure traits (SCHÖPKE and SWALVE, 2016), the number of sires with accurate EBV is usually low, impairing the ability to build a robust reference population. With decreasing costs of genotyping in general, a promising way to overcome this obstacle is to include females in the reference population (PRYCE and DAETWYLER, 2012; THOMASEN et al., 2014; PLIESCHKE et al., 2016; JENKO et al., 2016).

As the number of cows is much higher than the number of bulls in dairy cattle populations, there is a demand on the elaboration of strategies to assess an optimal genotyping design (SCHÖPKE and SWALVE, 2016), in which cows included in the reference population maximize SNP information value. Some cow genotyping strategies have been investigated using both simulated and real data (JIMÉNEZ-MONTERO et al., 2012; PSZCZOLA et al., 2012; CALUS et al., 2013; KOIVULA et al., 2016; UEMOTO et al., 2017), and results point to a general increase in genomic prediction accuracy when the reference population is supplemented with genotyped and phenotyped cows. Jiménez-Montero et al. (2012) have suggested, through simulation, that in case of small dairy populations, a reference population formed exclusively by cows could overcome both pedigree index and reference population containing only progeny-tested bulls. Dassonneville et al. (2012) and Gao et al. (2015) have addressed the impact of including

pre-selected cows into a reference population of bulls and suggested that this strategy would introduce bias into genomic predictions. Plieschke et al. (2016) reported that including large (random) samples of first-crop daughters in the reference set composed by bulls had a positive effect not only on prediction accuracy, but on preventing deterioration of reliabilities caused by intense pre-selection of young males.

Clark et al. (2012) and Wientjes et al. (2013) have demonstrated that the relationship between animals in the reference population and validation population has major effect on the accuracy of genomic prediction. Results suggest that an optimum balance of relationships within the reference population and between reference population and validation population is crucial to avoid under/over estimation of accuracies (CALUS et al., 2013). Despite numerous studies investigating the effectiveness of different genotyping designs to include cows into the reference population, not many studies considered population structure and partition to define selective genotyping strategies.

Innovative methods have been crucial for the evolution of animal breeding, helping to improve decision making ability, and consequently, efficiency and productivity in livestock populations. In this manner, graph theory has previously been proposed as a tool for better understanding genealogical data (COLE, 2007; YANG et al, 2012). However, no further investigation on how pedigree graphs can support comprehension on underlying configurations of livestock population pedigrees is found in literature.

Graphs are flexible mathematical systems containing a finite amount of points (nodes) and edges, which are essentially used to represent and analyze relationship between structured data (GIRVAN and NEWMAN, 2002). When modeling pedigrees as graphs, it is possible to subset a population into partitions by using community detection

algorithms (BLONDEL et al., 2008). Graph communities are defined as groups of nodes (here representing individuals) strongly connected among themselves but sparsely connect with the rest of the system (GIRVAN and NEWMAN, 2002). In practical terms, a graph community could be considered one of the distinct families or groups of related individuals found in the population. In literature, some attempts to account for population structure in genotyping designs have been investigated in plant breeding (ISIDORO et al., 2014; RINCENT et al., 2017), but applications for cattle populations are rare or non-existent.

In the present study, we investigate the use of graph communities as tools for developing selective cow genotyping strategies that consider population structure. Different approaches were considered and compared to previously reported cow genotyping designs (random, EBV- and accuracy-dependent) to determine the efficiency of the proposed methods. The impact of each genotyping design studied over the genomic prediction accuracy, error and bias.

3.2 MATERIAL AND METHODS

3.2.1 Simulated population

The open access software QMSim (SARGOLZAEI and SCHENKEL, 2009) was considered to run the simulations for the present study, considering five replicates. We simulated a relatively small dairy cattle population, similarly to previous simulation studies (JIMÉNEZ-MONTERO et al., 2012; PLIESCHKE et al., 2016; ANDONOV et al., 2016). To accomplish that, a historical population consisting of 1,500 unrelated individuals with

balanced sex ratio was created. Those animals were randomly mated for 1,000 generations. A bottleneck was introduced from generation 1,001, reducing the historical population size from 1500 to 400 animals in 20 generations. From the historical population, a second population, here referred as expansion population, was created in which founders were 100 males and 300 females randomly sampled from the last generation of historical population (1020). This expansion population was run for 60 generations in which animals were also randomly mated, considering a growth rate of 1.2 for both sexes by generation. From the expansion population, a recent population was created by sampling 400 males and 20,000 females from the last generation of the expansion population. The recent population was run for 20 overlapping generations. In each generation, 10,000 male and 10,000 female offspring were simulated. At the same time, 30% of the dams and 50% of the sires were replaced. Genealogical information was recorded for all individuals from the recent population. Breeding animals were selected based on their BLUP-based breeding values (here we used EBV) and culled by age. Cows from generations 16 to 20 and sires from generations 15 to 18 had genotypic data recorded. Bulls with more than 40 daughters were considered as progeny tested individuals ($n = 764$) and assigned to the reference set. Animals from generation 20 were considered as validation population.

3.2.2 Genome simulation

The simulated genome had 29 pairs of autosome chromosomes with total length of 2333 cM, in which each chromosome's length mimicked the bovine genome without sex chromosomes. The intent was to generate a more plausible scenario when

considering real distances between markers and QTL loci. A total of 50,000 bi-allelic SNP markers were evenly distributed throughout the genome, in which the number of SNP per chromosome was proportional to its size. The additive genetic component was determined by 750 randomly distributed QTL along the genome, which effects were sampled from a gamma distribution with a shape parameter of 0.4 (HAYES AND GODDARD, 2001). The number of QTL (750) was defined aiming to assure a polygenic nature for the simulated trait. The mutation rate was identical (2.5×10^{-5}) for both markers and QTL (JIMÉNEZ-MONTERO et al., 2012). The rate of missing marker genotypes was 0.01 and the rate of marker genotyping error was 0.005. A schematic overview of the simulation process is presented in Figure 3.1.

3.2.3 Phenotype simulation

A single sex-limited trait with heritability of 0.15 and phenotypic variance of 1.0 was simulated. The true breeding value (TBV) for each animal was calculated as the sum of the QTL additive effects, as follows:

$$TBV_k = \sum_{j=1}^{qtl} \beta_j \cdot Q_{kj}, \quad [1]$$

where qtl is the total number of QTL, β_j is the additive effect of QTL genotype (j) and Q_{kj} is the QTL genotype at locus j , coded as 0,1 or 2, representing the number of copies of the QTL allele that an animal (k) carries.

Pedigree-based EBV were estimated for all individuals from generations 12 to 20 of the recent population. The analysis was carried out under a restricted maximum likelihood (REML) approach using an animal model. Implementation was performed in the BLUPF90 family software (MIZSTAL et al., 2002). The model was as follows:

$$\mathbf{y} = \mu \mathbf{1} + \mathbf{Z}\mathbf{a} + \mathbf{e}, \quad [2]$$

where $\mathbf{y}$ is the vector of phenotypes for cows; μ is the general mean (no fixed effects were simulated); $\mathbf{1}$ is a vector of ones; $\mathbf{Z}$ is an incidence matrix allocating records to breeding values; $\mathbf{a}$ is a vector of EBV with $var(\mathbf{a}) = \mathbf{A}\sigma_a^2$, where $\mathbf{a} \sim N(0, \mathbf{A}\sigma_a^2)$, in which σ_a^2 is the additive genetic variance and $\mathbf{A}$ (Henderson, 1975) is the pedigree-based relationship matrix containing the expectancy of genetic connection between all pedigreed individuals, and $\mathbf{e}$ is the random residual effect, where $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$ in which σ_e^2 is the residual variance.

The EBV accuracies were calculated as $r = \sqrt{1 - PEV/\sigma_a^2}$, where PEV is the prediction error variance obtained from the estimate's standard error. The EBV and respective reliabilities were then used to obtain de-regressed proofs (dEBV) for all individuals following (GARRICK et al., 2009), which were then used as pseudo-phenotypes in the genomic evaluation. All phenotypic information included as dependent variable in the model was weighted by its reliability.

3.2.4 Pedigree graphs, community detection and population partitioning

A “pedigree graph” was generated by transforming genealogical information into a matrix (adjacency matrix) of zeros and ones, with dimensions $n \times n$, where n is the total number of individuals in the pedigree, that represent direct connections between individuals (parent-progeny). In practical terms, this basic approach consists of transforming the numerator relationship matrix ($\mathbf{A}$; HENDERSON, 1975) into a graph adjacency matrix where the following criteria must be met: elements of the pedigree

graph input adjacency matrix receive the value 1 when the respective A_{ij} element is 0.50, and 0 otherwise.

Graph community detection results in the partition of a graph into modules, in which vertices within their module are more strongly connected than to the rest of the system. The quality of the partitions resulting from these methods is often measured by the modularity (NEWMAN and GIRVAN, 2002) parameter, which quantifies the quality of the assignment of vertices to communities by comparing how much more connected vertices are inside a community to how connected they would be, on average, in a random graph. The modularity of a partition is a scalar value between -1 and 1 that measures the density of connections inside communities as compared to connections between communities. In general, modularity is defined as (BLONDEL et al., 2008):

$$Mod = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i c_j), \quad [4]$$

where A_{ij} represents the magnitude of the connection between vertices i and j , which may assume values of 1 (connected) or 0 (not connected) in standard graphs, $k_i = \sum_j A_{ij}$ is the sum of the weights of the edges attached to vertex i (here 1 or 0), c_i is the community to which vertex i is assigned, the δ function $\delta(u, v)$ is 1 if u is connected to v and 0 otherwise and $m = \frac{1}{2} \sum_{i,j} A_{ij}$.

Community detection algorithms may include a resolution parameter (RsPr), which can be understood as a “time” parameter (BLONDEL et al., 2008; Lambiotte et al., 2008), intended to accommodate different community sizes for the same graph. In practical terms, when decreasing RsPr, the number of detected communities tend to the total number of vertices (individuals). On the other hand, when increasing RsPr, the

partition tends to stabilize in a two-community's configuration. In the present study we considered different values for RsPr (low = 0.5, medium = 1.0 and high = 1.5) when running the community detection algorithm in order to investigate its flexibility and assess the impact on the accuracy of genomic evaluations.

3.2.5 Selective genotyping designs

Cows from generations 16 through 19 of the recent population represented a contemporary overlapping active population of 40,000 individuals. From them, 1000, 2000 and 5000 were sampled as female candidates to be genotyped and included in the reference population, based on the methods described further in this section.

For all reference population sizes proposed (1000, 2000 and 5000 cows), a total of 13 different genotyping designs were considered to supplement the bull-based reference population with cows. Of these, four traditional designs were included as a base for comparison, being: 1) ϕ females randomly selected (RND), 2) ϕ females with the highest dEBV (T_{dEBV}), 3) $\phi/2$ females with the highest and $\phi/2$ females with the lowest dEBV (EX_{dEBV}) and 4) ϕ females with the highest EBV accuracies (T_{EBVAC}), where ϕ was 1000, 2000 or 5000, depending on the reference population scenario.

Nine other designs were proposed as a way to investigate both efficiency and adaptability of the community partitioning algorithm (BLONDEL et al., 2008). To accomplish that, a genealogical dataset containing all individuals from generations 12 to 20 ($n = 180,000$) of the recent contemporary population was built and analyzed as a graph. Animals from generation 12 had their parents omitted (set as "unknown") to represent founder individuals of the pedigreed population of a relatively small cattle

population. The algorithm designated a community for every animal in the pedigree. Community-based designs were primarily aimed to integrate communities and the traditional “high or extreme” dEBV and top accuracy concept. The representativeness of each detected community in the validation population was considered when sampling cows to be included into the reference population in each situation. These cow genotyping designs were elaborated as follows: 5) α females with the highest dEBV inside each detected community ($T_{dEBV}C$); 6) $\alpha/2$ females with the highest and $\alpha/2$ females with the lowest dEBV inside each detected community ($EX_{dEBV}C$); 7) α females with the highest EBV accuracies inside each detected community ($T_{EBV}ACC$); where α is the representativeness of the respective community in the total number of individuals defined as the validation population, calculated as $\alpha = (N_{ref_i} * N_{smpl_j}) / n_{total}$, where N_{ref_i} is the size of reference population for the i^{th} scenario (1000, 2000 or 5000), N_{smpl_j} is the size of the j^{th} community and n_{total} is the total number of candidate cows for the reference population. Therefore, the number of cows sampled from each community was directly linked to the proportion of individuals from this same community in the validation population. This implies that the reference population is built after knowing the selection candidates for the current generation. For further discussion on this document, this procedure will be referred at as “proportional sampling”. In Figure 3.2 it is shown a brief description of the community-based cow sampling method proposed. In order to investigate the flexibility of the community detection algorithm, $T_{dEBV}C$, $EX_{dEBV}C$ and $T_{EBV}ACC$ concepts were replicated considering low (0.5; $_L$), medium (1.0; $_M$) and high (1.5; $_H$) values for the $RsPr$, resulting in a total of 9 community-based cow genotyping designs: $T_{dEBV}C_L$, $T_{dEBV}C_M$, $T_{dEBV}C_H$, $EX_{dEBV}C_L$, $EX_{dEBV}C_M$, $EX_{dEBV}C_H$ and

T_{EBVACC_L} , T_{EBVACC_M} , T_{EBVACC_H} . All proposed scenarios are described in Table 3.1.

Table 3.1. Summary of genotyping scenarios and number of genotypes for bulls and cows in the reference population for each method proposed.

Design	Summary of reference population
RND	764 progeny tested bulls + ϕ cows randomly genotyped.
T_{dEBV}	764 progeny tested bulls + ϕ cows with highest dEBV genotyped.
EX_{dEBV}	764 progeny tested bulls + $\phi/2$ cows with highest and $\phi/2$ cows with lowest dEBV genotyped.
$T_{dEBVACC}$	764 progeny tested bulls + ϕ cows with highest accuracy of dEBV genotyped.
T_{dEBV_L}	
T_{dEBV_M}	764 progeny tested bulls + α females with the highest dEBV from within each detected community.
T_{dEBV_H}	
EX_{dEBVC_L}	
EX_{dEBVC_M}	764 progeny tested bulls + $\alpha/2$ females with the highest and $\alpha/2$ females with the lowest dEBV from within each detected community.
EX_{dEBVC_H}	
$T_{dEBVACC_L}$	
$T_{dEBVACC_M}$	764 progeny tested bulls + α females with the highest dEBV accuracies from within each detected community.
$T_{dEBVACC_H}$	

ϕ = depended on the scenario considered (1000, 2000 or 5000); $\alpha = (Nref_i * Nsmpl_j) / n_{total}$, where $Nref_i$ is the size of reference population for the i^{th} scenario (1000, 2000 or 5000), $Nsmpl_j$ is the size of the j^{th} community and n_{total} is the total number of candidate cows for the reference population.

RND = at random; T_{dEBV} = top de-regressed proof values; EX_{dEBV} = extreme de-regressed proof values;

$T_{dEBVACC}$ = top dEBV accuracy;

T_{dEBVC_L} = top de-regressed proof values inside communities (low resolution parameter);

EX_{dEBVC_L} = extreme de-regressed proof values inside communities (low resolution parameter);

$T_{dEBVACC_L}$ = top EBV accuracy inside communities (low resolution parameter);

T_{dEBVC_M} = top de-regressed proof values inside communities (medium resolution parameter);

EX_{dEBVC_M} = extreme de-regressed proof values inside communities (medium resolution parameter);

$T_{dEBVACC_M}$ = top EBV accuracy values inside communities (medium resolution parameter);

T_{dEBVC_H} = top de-regressed proof values inside communities (high resolution parameter);

EX_{dEBVC_H} = extreme de-regressed proof values inside communities (high resolution parameter);

$T_{dEBVACC_H}$ = top EBV accuracy inside communities (low resolution parameter).

3.2.6 Quality control and genomic prediction model

Genotype data quality control was performed prior to all genomic analyzes in the present study. In all proposed scenarios, criteria were: SNPs with minor allele frequency lower than 0.02 and call rate lower than 0.95, as well as samples with call rate lower than 0.90 were excluded from analysis. After quality control, around 43,500 SNPs remained, with slight variations depending on the reference population considered, the simulation replicate and the scenario proposed.

A genomic best linear unbiased (GBLUP) model (VANRADEN, 2008) was chosen for genomic prediction analyzes. The general equation is as follows:

$$\mathbf{y} = \mu \mathbf{1} + \mathbf{Z}\mathbf{g} + \mathbf{e}, \quad [6]$$

where $\mathbf{y}$ is the vector of the dEBV for the genotyped individuals in the reference population; μ is the general mean (no fixed effects were simulated); $\mathbf{1}$ is a vector of ones; $\mathbf{Z}$ is an incidence matrix allocating records to breeding values; $\mathbf{g}$ is a vector of GEBV with $var(\mathbf{g}) = \mathbf{G}\sigma_g^2$, in which σ_g^2 is the additive genetic variance; and $\mathbf{G}$ is the realized genomic relationship matrix created using the method described in VanRaden (2008). That is, $\mathbf{G} = (\mathbf{M} - \mathbf{P})(\mathbf{M} - \mathbf{P})' / \sum_{i=1}^m 2p_i(1 - p_i)$, where $\mathbf{M}$ is an $n \times m$ matrix (number of animals x number of loci) with SNP coded 0, 1 and 2 for genotypes A1A1, A1A2 and A2A2, respectively. $\mathbf{P}$ is an $n \times m$ matrix containing twice the allele frequencies expressed as $P_i = 2p_i$ where p_i is the allele frequency of the homozygous genotype coded with 2 for all genotyped individuals at locus i . It is assumed that $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$, where σ_e^2 is the residual variance. All analyzes were conducted under a REML approach, using BLUPF90 family software (MIZSTAL et al., 2002).

Inside each replicate and for each scenario proposed, the direct genomic value (DGV) was calculated for all animals in the validation population as:

$$DGV = X'\hat{\beta}, \quad [7]$$

where X is the matrix of marker genotypes for each animal in the validation set and $\hat{\beta}$ is the vector of marker effects estimated based using information on the respective reference population.

3.2.7 Accuracy, prediction error and bias

As TBV were known for all animals in the simulation, the Pearson product-moment correlation coefficient (ρ) and the mean squared error (MSE) between the TBV and the DGV of animals from the validation population were used to assess the accuracy and prediction error in genomic evaluations for each selective genotyping strategy proposed. The slope (b) of the regression of TBV on DGV for the validation individuals, that is, $b = cov(TBV, DRP)/\sigma_{TBV}^2$ (OLSON et al., 2011) and the respective determination coefficient (R^2) were considered as a measure of bias in genomic evaluations. Means and respective standard deviations (from the five simulation replicates) were calculated for each proposed parameter in the strategies proposed. The Student's T test was used to investigate statistical differences when superiority in performance (for ρ , MSE, b and R^2) was observed for community-based designs in detriment to their traditional versions. Additionally, the Student's T test was also performed between all community-based designs and the RND design.

Figure 3.1. Illustrative scheme describing the simulation process and genotyping strategies considered in the present study.

Figure 3.2. Illustrative scheme describing the proportional sampling for cow genotyping considering pedigree graph communities as a tool to account for population partition.

3.2.8 Mimicking selection for an expensive- or difficult-to-measure trait in dairy cattle

Here we aimed to emulate a situation where selection is intended for a low-to-medium heritability difficult- or expensive-to-measure (sex limited) trait. In those cases, the number of available records is often low, resulting in low accuracy EBV even for bulls. Therefore, a reference population containing only cows is of use (PRYCE and DAETWYLER et al., 2012), while cow's yield deviation will mostly be the (pseudo)

phenotype of choice. For those analysis, yield-deviations (YD; VANRADEN and WIGGANS, 1991) were calculated using a combination of phenotype and a residual component for every candidate cow, and thus, used as dependent variable in the genomic evaluation. The YD were then used to calculate daughter yield deviations (DYD; VANRADEN and WIGGANS, 1991) for the progeny-tested bulls, which were used as pseudo-phenotypes in the scenario considering only bulls in the reference. We speculated that, in this situation, a small reference population containing only cows proportionally sampled based on individuals in a validation population containing individuals from specific communities could support obtaining sufficient accuracy for genomic selection.

To accomplish that, cows from ten randomly chosen communities were subset from the complete dataset, forming a sub-dataset containing only individuals from those communities from generations 16 through 20 of the recent contemporary population. This procedure was replicated five times, based on a randomly chosen simulation replicate. Different sizes for the sampled communities were allowed and the only criteria to be met was that the total number of individuals had to be higher than 3,000. Two main scenarios were then proposed: a reference population of 1,000 cows selected by distinct methods from generations 16 through 19 within the ten communities and a validation from generation 20 of animals from within the ten communities (In_Ref); and a reference population of 1,000 cows selected by distinct methods from all individuals in generations 16 through 19 and a validation considering all possible candidate cows from generation 20 (All_Ref). Both In_Ref and All_Ref scenarios were replicated considering three different cow selective genotyping methods, RND_{YD} (A), T_{YD} (B) and EX_{YD} (C), which were

similar to the previously presented strategies, but considering yield-deviations (YD) of candidate cows as dependent variable in the model, using a community-based proportional sampling approach. The accuracy, prediction error and bias in genomic predictions were calculated as in previous analysis in this study.

3.3 RESULTS

3.3.1 Population structure description

The pedigree used for community detection contained all animals from generations 12 through 20 ($n = 180,000$). There was a total of 9 generations, containing 2,005 sires and 68,111 dams. The mean inbreeding coefficient for the total population and for inbred individuals was 0.47% and 1.50%, respectively.

The amount of linkage disequilibrium, here measured as r^2 (HILL and ROBERTSON, 1968), found in chromosomes 1 and 29 for and 764 progeny-test bulls and cows in generations 16 through 20 ($n = 50,764$) is shown in Figure 3.3. The average r^2 for adjacent SNP markers was 0.24 ± 0.25 and 0.21 ± 0.23 for chromosomes 1 and 29, respectively

Figure 3.3. Linkage-disequilibrium decay for chromosomes 1 (left) and 29 (right) considering all genotyped cows from generations 16 through 20 and the progeny-tested bulls ($n = 50,764$).

3.3.2 Community detection and population partitioning

The total number of communities detected for the five simulation replicates was 362, 365, 362, 360 and 364 when considering a lower (0.5) RsPr value; 134, 129, 132, 134 and 132 for the medium (1.0) RsPr value; and 82, 85, 82, 84, 83 for the higher (1.5) RsPr value. As described in the methods section, as the value of RsPr increased, the number of detected communities decreased.

Figure 3.4 shows the frequency of progeny-tested bulls included in the reference population (independent of the scenario) and cows from generations 16 through 19 (candidates to get genotyped) inside each detected community for one given (randomly chosen) simulation replicate. Figure 3.5 shows the same information but for animals from generation 20 (validation population). Figures 3.4 and 3.5 also show results for the frequency of animals inside communities when considering low, medium and high RsPr values. Lower values for the RsPr were associated to a less variable frequency of individuals inside the detected communities. Increasing the RsPr (from 0.5 to 1.0 and 1.5) resulted in an increase in the variation of the frequency of animals between detected

communities. This changes in frequency of individuals inside communities was identical in all simulation replicates (results not shown).

Figure 3.4. Frequency of progeny tested bulls ($n = 764$, left), from generations 15 through 18, and cows from generations 16 through 19 ($n = 40,000$, right) inside each detected community for the scenarios when setting a low (A), medium (B) and high (C) resolution parameter value for the community detection algorithm.

The mean of the average relationship (calculated using a pedigree-based relationship matrix) for animals inside and outside their community considering distinct values of RsPr were 0.08% and 0.002% (low); 0.04% and 0.005% (medium), 0.03% and

0.006% (high), respectively. This is a strong indication that the community detection algorithm has successfully grouped individuals more genetically connected among themselves than to the rest of the population.

Figure 3.5. Frequencies of male ($n = 10,000$, left) and female ($n = 10,000$, right) animals from generation 20 inside each detected community for the scenarios when setting a low (A), medium (B) and high (C) resolution parameter value for the community detection algorithm.

3.3.3 Accuracy of genomic evaluations

When including only the progeny-tested bulls with more than 40 daughters from generations 15 through 18 in the reference population, the average ρ between TBV and DGV of the validation animals (in five simulation replicates) was 0.30 ± 0.03 . As the main objective of the present study was to investigate the inclusion of phenotyped and genotyped cows in the reference population by various methods, this accuracy value will be considered as a basis for comparison throughout this document and will be referred to as SireRP.

Figure 3.6 shows the average (and respective standard deviation) ρ values for each proposed design and reference population size studied. Overall, the accuracy was strongly dependent on both the genotyping design and size of the reference population. As expected, as the size of the reference population increased accuracies also increased, independent of the strategy adopted. Community-based methods exhibited accuracies varying from slightly inferior to slightly superior than traditional methods. Subtle differences were detected within community-based designs when comparing low, medium and high RsPr values. In general, when the RsPr deviated from the medium value (1.0), it resulted in a decrease in accuracy. Standard deviations for the mean accuracy within simulation replicates varied from 0.03 to 0.08, where the highest value was observed for the scenario including 2,000 cows considering the $T_{dEBV}C_H$ method, and the lowest for the scenario including 1,000 cows considering the RND method.

Strategies based on selecting cows with high dEBV values (T_{dEBV} , $T_{dEBV}C_L$, $T_{dEBV}C_M$ and $T_{dEBV}C_H$) to be included in the reference population resulted in the lower increase in accuracy. The scenario including 1,000 cows via T_{dEBV} method has achieved

only minor increase when compared to SireRP. For this method, when including a higher number of cows (5,000), the maximum increase in accuracy was +0.06. When high-dEBV cows were sampled from inside communities (T_{dEBVC_L} , T_{dEBVC_M} and T_{dEBVC_H}), a higher increase in accuracy was obtained only for T_{dEBVC_M} when compared to its traditional method. When compared to traditional T_{dEBV} method, the increase in accuracy varied from +0.01 ($T_{dEBVC_M}/1000$) to +0.04 ($T_{dEBVC_M}/5000$), depending on the value set for the RsPr.

The RND method showed intermediate accuracy when compared to dEBV-dependent methods (top or extreme). With the increase in the number of cows in the reference population, the increase in accuracy of RND varied from +0.09 to +0.23 (compared to SireRP) for scenarios considering 1,000 and 5,000 cows.

Accuracy-based designs (T_{EBVAC} , T_{EBVACC_L} , T_{EBVACC_M} and T_{EBVACC_H}) had intermediate performance on accuracy when compared to other proposed scenarios, yielding similar results to the RND design. For the inclusion of 5,000 cows, the highest increase in accuracy obtained by including cows in the reference population from accuracy-based designs was for T_{EBVACC_L} and T_{EBVACC_M} , being +0.21 when compared to SireRP.

Figure 3.6. Average (and respective standard deviations)¹ accuracy, mean squared error (MSE), slope (b) and coefficient of determination (R^2) for the linear regression between true breeding value and direct genomic value for animals in the validation population for each selective scenario² and genotyping strategy³ tested.

¹ Mean and standard deviations obtained from five replicates.

² 1000/2000/5000 represent the number of cows included in the reference population with 764 bulls. The total number of individuals is 1764/2764/5764.

³ See Table 3.1 for description of genotyping strategies.

Result for SireRP was $\rho = 0.30 \pm 0.03$, $MSE = 10.88 \pm 0.04$, $b = 0.78 \pm 0.10$, $R^2 = 0.08 \pm 0.03$.

RND = at random; TdEBV = top de-regressed proof values; EXdEBV = extreme de-regressed proof values;

Best results in accuracy were obtained for genotyping designs considering cows with extreme dEBV values (EX_{dEBV}, EX_{dEBV}C_L, EX_{dEBV}C_M, EX_{dEBV}C_H). Both traditional and community-based designs delivered similar results, varying from +0.15 (EX_{dEBV}C_L/1000) to +0.30 (EX_{dEBV}C_M/5000) in accuracy when compared to SireRP. Directly comparing among extreme dEBV designs, a slight superiority was observed for results obtained from EX_{dEBV}C_M, which had from 0.01 (1,000 cows) to 0.02 (5,000 cows) higher accuracy than EX_{dEBV}.

Table 3.2. Results of the Student's T test (P-value) for the difference in genomic prediction accuracy between traditional and community-based cow genotyping designs considering a medium resolution parameter for the graph-community detection algorithm.

Design ¹	Size of the RP ²	P-Value versus community-based ³	P-value COM x RND ⁴
		Medium RsPr	
T _{dEBV}	1000	0.396	0.017*
	2000	0.024*	0.004**
	5000	0.013*	0.003**
EX _{dEBV}	1000	0.344	0.001**
	2000	0.045*	<0.001**
	5000	0.029*	<0.001**
T _{dEBV} ACC	1000	0.472	0.013*
	2000	0.092	0.332
	5000	0.355	0.026*

¹ See Table 1 for description of genotyping strategies

² 1000/2000/5000 represent the number of cows included in the reference population with 764 bulls. The total number of individuals is 1764/2764/5764

³ P-value for the Student's Test between the respective traditional design and its community-based medium RsPr version.

⁴ P-value for the Student's Test between the community-based design and traditional RND cow genotyping design.

T_{dEBV} = top de-regressed proof values; EX_{dEBV} = extreme de-regressed proof values;

T_{dEBV}ACC = top de-regressed proof accuracy; *Differ at P < 0.05, **Differ at P < 0.01.

Table 3.2 shows P-values for the Student's T test for the mean p (obtained from 5 replicates) between traditional (T_{dEBV} , EX_{dEBV} and $T_{dEBVACC}$) and medium RsPr community-based cow genotyping strategies. Significant differences were detected between T_{dEBV} , EX_{dEBV} and their community-based versions when including 2,000 and 5,000 cows in the reference population, but no difference was detected for $T_{dEBVACC}$. All community-based designs exhibited significant difference when compared to RND.

3.3.4 Prediction error and bias

The SireRP design resulted in an average MSE of 10.88 ± 0.04 , regression slope (b) of 0.78 ± 0.10 and a regression R^2 of 0.08 ± 0.03 between TBV and DGV for the validation individuals. In general, the increase in the number of cows included in the reference population resulted in reduction of the prediction error and bias (Figure 3.6).

When compared to traditional methods, community-based genotyping designs yielded a slight reduction in MSE. Within community-based strategies, the increase in the RsPr caused an increase in MSE values. When compared to SireRP, best results were observed for EX_{dEBVC_M} which yielded a reduction of 1.43, 1.91 and 2.51 in MSE, respectively, for the inclusion of 1,000, 2,000 and 5,000 cows in the reference population.

An increase on the MSE was observed when considering the inclusion of cows in the reference population by traditional high-dEBV method (T_{dEBV}) when compared to the SireRP. No consistent reduction in the prediction error was observed when increasing the number of cows to be included in the reference population by T_{dEBV} . For the inclusion of 5,000 cows considering T_{dEBV} , the MSE was still higher than the SireRP, which

contained only 764 bulls. When under a community-based approach, high-dEBV-based designs ($T_{dEBV}C_L$, $T_{dEBV}C_M$ and $T_{dEBV}C_H$) showed better results than its traditional version (T_{dEBV}). Overall, the high-dEBV-based strategies exhibited more prediction error (MSE) than other proposed methods, even when under a community-based approach.

In general, the regression coefficient (b) of TBV on DGV for the validation animals was closer to 1.00 with the inclusion of cows in the SireRP, the only exception occurred for the T_{dEBV} based designs (traditional and community-based), in which the inclusion of cows resulted in more bias ($b > 1.00$). Actually, top-dEBV methods resulted in the highest increase in b values when compared to SireRP and, therefore, in the highest amount of deflation of breeding values. RND and $T_{dEBV}ACC$ methods had intermediate results, which were lower than the obtained for T_{dEBV} but higher than EX_{dEBV} strategies. Despite yielding the lowest amount of prediction error in all scenarios, genotyping designs based on extreme dEBV values (both traditional and community-based) showed the highest amount of inflation ($b < 1.00$) of breeding values.

As the number of cows included in the reference population increased, the R^2 obtained from the regression of TBV on DGV in the validation population also increased (Figure 3.6). The R^2 values obtained for extreme-dEBV designs were the highest among all strategies proposed, while the RND method resulted in superior values than obtained for top-dEBV, but lower than obtained for extreme-dEBV designs. The lowest increase was observed for top-dEBV genotyping designs, of which T_{dEBV} also obtained the lowest absolute R^2 values, varying from 0.13 to 0.17 when including 1,000 and 5,000 cows in the reference population, respectively. Community-based high-dEBV designs had better results than traditional T_{dEBV} for all reference population size scenarios. The same pattern

observed for the accuracy and MSE was detected for the R^2 values. When comparison is made within community-based methods, the deviation from 1.00 for the RsPr caused a slight decrease in R^2 .

Table 3.3. Results of the Student's T test (P-value) for the difference in mean squared error (MSE), slope (b) and coefficient of determination (R^2) for the regression between true breeding value and direct genomic value for animals in the validation population (n = 20,000) between traditional and community-based cow genotyping designs considering a medium resolution parameter in graph-community detection algorithm.

Design ¹	Cows in the RP ²	P-Value versus community-based (Medium RsPr) ³		
		MSE	b	R^2
T _{dEBV}	1000	0.097	0.479	0.531
	2000	0.037*	0.184	0.436
	5000	0.007**	<0.001**	0.014*
EX _{dEBV}	1000	0.489	0.012*	0.558
	2000	0.504	0.011*	0.473
	5000	<0.001**	0.005**	0.446
T _{dEBV} ACC	1000	0.251	0.372	0.369
	2000	0.432	0.490	0.490
	5000	0.297	0.003**	0.455

¹ See Table 3.1 for description of genotyping strategies

² 1000/2000/5000 represent the number of cows included in the reference population with 764 bulls. The total number of individuals is 1764/2764/5764

³ P-value for the Student's Test between the respective traditional design and its community-based version.

T_{dEBV} = top de-regressed proof values; EX_{dEBV} = extreme de-regressed proof values;

T_{dEBV}ACC = top de-regressed proof accuracy; *Differ at P < 0.05, **Differ at P < 0.01.

Table 3.3 shows P-values obtained from the Student's T test for the MSE, b and R^2 (obtained from 5 replicates) between traditional (T_{dEBV}, EX_{dEBV} and T_{dEBV}ACC) and medium RsPr community-based cow genotyping strategies. For MSE, significant

differences were observed between $T_{dEBV}/2000$, $T_{dEBV}/5000$ and $EX_{dEBV}/5000$ and their community-based versions. For b , significant P-values were observed for $T_{dEBV}/5000$, all EX_{dEBV} scenarios and $T_{dEBV}ACC/5000$. The only significant difference detected for R^2 between traditional and community-based designs was in the $T_{dEBV}/5000$ scenario.

3.3.5 Community-based genotyping design for a reference population of cows

When considering SireRP (using DYD as dependent variable) to predict breeding values for a validation population formed by individuals inside the respective chosen communities, obtained results were $\rho = 0.28 \pm 0.04$; $MSE = 10.79 \pm 0.18$; $b = 0.86 \pm 0.12$ and $R^2 = 0.11 \pm 0.04$.

Table 3.4 shows results for accuracy, prediction error and bias when considering a reference population of cows chosen from inside the communities of animals in the validation population and from outside those communities by the methods previously proposed, but using YD as dependent variable in the models.

In general, results were markedly superior when genotyped cows were chosen from inside the chosen communities (In_Ref) than when chosen from outside those communities (All_Ref). The highest increase in performance between All_Ref and In_Ref was observed for the T_{YD} design, where differences for the parameters was +0.22 for ρ , +0.55 for MSE, -2.66 for b and +0.08 for R^2 . Results obtained for all designs under the In_Ref scenario outperformed the results of accuracy obtained for SireRP in all scenarios.

As observed in the previous analyzes in the present study (which considered de-regressed proofs as dependent variable), extreme-YD (EX_{YD}) methods showed the best,

while high-YD (T_{YD}) presented the worst results for prediction accuracy. Likewise, highest bias was observed for extreme-YD methods (EX_{YD}).

Table 3.4. Average (and respective standard deviation)¹ for accuracy (ρ), mean squared error (MSE), slope (b) and coefficient of determination (R^2) of the regression between the true breeding value and predicted direct genomic values for animals in the validation population for the genotyping strategies² tested for cow-based reference populations.

Design	Parameter	In_Ref	All_Ref
RND _{YD}	ρ	0.34 ± 0.05	0.26 ± 0.03
	MSE	10.36 ± 0.09	10.42 ± 0.08
	b	1.02 ± 0.04	0.81 ± 0.05
	r^2	0.12 ± 0.03	0.08 ± 0.02
T_{YD}	ρ	0.31 ± 0.07	0.09 ± 0.02
	MSE	10.50 ± 0.11	11.05 ± 0.18
	b	0.84 ± 0.07	3.45 ± 0.28
	r^2	0.09 ± 0.03	0.01 ± 0.01
EX_{YD}	ρ	0.46 ± 0.07	0.37 ± 0.06
	MSE	8.38 ± 0.05	9.42 ± 0.08
	b	0.45 ± 0.06	0.58 ± 0.06
	r^2	0.19 ± 0.06	0.12 ± 0.05

¹ Mean and standard deviations obtained from five replicates.

² See Table 3.1 for description of genotyping strategies.

Student's T test between In_Ref and All_Ref methods was significant ($P < 0.05$) for all studied parameters (ρ , MSE, b and R^2);

Results for SireRP (using DYD as dependent variable) were $\rho = 0.28 \pm 0.04$; MSE = 10.79 ± 0.18 ; $b = 0.86 \pm 0.12$ and $R^2 = 0.11 \pm 0.04$.

In_Ref = reference population of 1,000 cows selected by distinct methods from generations 16 through 19 within the ten communities and a validation from generation 20 of animals from within the ten randomly chosen communities;

All_Ref = reference population of 1,000 cows selected by distinct methods from all individuals in generations 16 through 19 and a validation from generation 20 of animals from within the ten randomly chosen communities.

RND_{YD} = at random; T_{YD} = top yield deviation; EX_{YD} = extreme yield deviation values;

3.4 DISCUSSION

Here we investigated, via simulation, the impact of accounting for population structure when defining selective cow genotyping strategies on the performance of genomic evaluations. The simulation was conducted to mimic selection considering a medium-to-low heritability trait for a small dairy cattle population, in which only few progeny-tested bulls had sufficient reliable dEBV and the reliability of dEBV for the candidate cows to be included in the reference population was plausible for such a population, ranging from 0.30 to 0.65. The observed levels (and decay) of LD and inbreeding in our simulations were also in accordance to the literature for dairy cattle populations (DE ROOS et al., 2008; HABIER et al., 2010).

The heritability of a trait has major effect on the reliability of genomic evaluations (ZHOU et al., 2014) and it would be plausible to investigate the efficiency of considering population community structure for different heritability values. However, as the focus of this study was to comprehend how the community detection algorithm would perform to assess population partitioning and its effects on informativeness of the reference population, we limited the analyzes to a fixed heritability value ($h^2 = 0.15$). It is expected, though, that highly heritable traits will possibly yield higher improvements than observed in the present study.

Results suggest that the inclusion of cows in the reference population, even in moderate numbers, seems to be a valuable strategy for improving the performance of genomic selection in small cattle populations. Similar results were observed in the literature for real (GAO et al., 2015; KOIVULA et al., 2016; JENKO et al., 2017; UEMOTO et al., 2017) and simulated (MCHUGH et al., 2011; JIMÉNEZ-MONTERO et al., 2012;

THOMASEN et al., 2014; PLIESCHKE et al., 2017) data. In general, selective genotyping designs based on high-dEBV cows performed poorly when compared to other proposed designs. Random and high-EBV-accuracy-based strategies yielded medium prediction ability and extreme-dEBV methods had the best results in accuracy. Including extreme cow (pseudo) phenotypes in the reference population seems useful to more accurately predict which are the most influential SNPs in genome wide association studies, but with the cost of inflation of the genomic estimated breeding values for selection candidates.

When comparison between non-contemporaneous individuals is required, high-EBV-accuracy-based designs seem to be the most indicated to choose cows to be included in the reference population for its observed reduced bias. However, it must be emphasized that in the considered simulated population, EBV accuracy for cows which were candidates to the reference population ranged from 0.30 to 0.65 and, thus, populations where female's EBV accuracy vary from those values may probably deviate from the obtained results here.

Overall, adopting “proportional sampling” within communities to choose cows to be included in the reference population was beneficial to improve accuracy (Figure 3.6 and Table 3.2) and reduce bias (Figure 3.6 and Table 3.3) mainly for some dEBV-based (T_{dEBV} and EX_{dEBV}), but not for accuracy-based ($T_{dEBV/ACC}$) scenarios. Considering sampling of cows from distinct sparsely connected groups in the population probably result in an indirect sampling of representative haplotypes which are present in a validation population level, improving prediction ability on selection candidates. Rincent et al. (2017) has found promising results when considering population structure to improve accuracies in plant breeding based on individual's genomic information to perform this task. The

method proposed in the present study uses exclusively genealogical data for designating an optimum set of cows to be included in the reference population. This approach, when under the scope of an animal breeding program, is of extreme value as pedigree information has been routinely recorded for decades. In this manner, the application of such procedure in animal breeding programs would result in no extra costs. Community “proportional sampling” could, therefore, easily be introduced in both new and already established animal breeding programs as a tool to designate an informative set of cows to be genotyped and included in the reference population.

Despite the continuous reduction of genotyping costs, in some situations it is still too expensive to perform massive genotyping of females. Also, in practical terms for animal breeding programs, it is really difficult to implement actual random genotyping. This means that breeders will most likely to genotype their high-merit cows (PRYCE AND DAETWYLER et al., 2012). The increase in accuracy and reduction in bias obtained for T_{dEBVC_M} (including 5,000 cows in the reference population) compared to its traditional version (T_{dEBV}) may indicate a path to overcome this situation in small animal breeding programs, in countries where genotyping costs are still high or when random cow genotyping is unfeasible. Choosing high-yield cows to be included in the reference population by proportional sampling within detected communities could help to obtain more value from cow's genotypes and phenotypes when an animal breeding program is running under logistical problems or reduced budget situations.

The RsPr was demonstrated to have some impact on both accuracy and bias, as values departing from 1.00 resulted in a slight decrease in performance of genomic selection. However, it is important to emphasize that the simulated population exhibited

a simple linear structure, in which the same number of bulls and cows were selected and culled each generation. Real cattle populations will often exhibit a much more complex structure within and between families and, thus, the efficiency of the methods here presented must be investigated in such datasets. It is intuitive to infer that as the population structure increases in complexity, the “real” underlying community partitioning will get similarly complex and values different from 1.00 for the RsPr may yield better results than observed in this study.

The inclusion of information (de-regressed proofs) of genotyped cows in a reference population of bulls reduced EBV inflation/deflation in the majority of scenarios, except for T_{dEBV} and EX_{dEBV} , which led to evident deflation and inflation, respectively. The increase in bias was also observed by Jiménez-Montero et al. (2012) when investigating reference populations formed by genotyped high- and extreme-yield cows. The authors have observed that cow reference populations, independent of the method considered for choosing those females, resulted in more bias than bull reference populations. However, when under a community-based proportional sampling approach (T_{dEBV_L} , T_{dEBV_M} , T_{dEBV_H}), this problem seems to be considerably reduced for T_{dEBV} , as values of b obtained from sampling cows to be genotyped by such community-based methods were far closer to 1.00. In this case, even when genotyping top-yield individuals, bias was relatively controlled in genomic predictions. It is important, however, to emphasize that in real situations the inclusion of high-merit cows coming from elite herds may generate bias in genomic evaluations due to possible preferential treatment (DASSONNEVILLE et al., 2012; DEHNAVI et al., 2017). In the present study, no effect of preferential treatment was

present in the simulated data, thus the impacts of proportional sampling under the presence of such situation are yet to be investigated.

The applicability of reference populations formed exclusively by cows have been investigated in the literature. Ding et al. (2013), while studying genomic selection for milk traits in the Chinese Holstein population, considered a reference population of 3,087 cows and obtained from 13% to 33% of increase in accuracy when compared to the pedigree index. Promising results of reference population formed exclusively by cows in small cattle populations were also presented via simulation by Jiménez-Montero et al. (2012), in which sampling cows by extreme yield deviation values to include in the reference population resulted in better accuracy than a reference population of bulls. Egger-Danner et al. (2014) have studied the economic viability of genotyping cows when selection is performed over health traits in Simmental cattle. The authors have stated that the high amount of information required to obtain a reliable accuracy of selection would impair the implementation of genomic selection for such traits. Genotyping cows from few specific highly reliable herds and, thus, limiting genotyping and phenotyping to individuals considerably connected with the phenotyped reference cows could help to overcome this obstacle (PRYCE and DAETWYLER, 2012; EDEL et al., 2016).

When under the scope of proportional sampling presented in this study, genotyping individuals from certain communities (instead of actual herds) could help to improve selection accuracy. Improvements were obtained when accounting for population structure for building a cow reference population to predict breeding values of individuals from specific subpopulations (Table 3.4). Sampling cows to be included in the reference population from inside communities proportional to their representation in the validation

population yielded considerable improvements in accuracy and prediction error when compared to sampling cows from outside those communities by the same genotyping designs. Results obtained for the In_Ref (using cow YD as dependent variable) method were even better than the obtained for SireRP (using DYD as dependent variable). This approach may be attractive when selection is intended for traits such as health disorders (BUCH et al., 2011; VUKASINOVIC et al., 2017) and hormone profiles (TENGHE et al., 2014), in which cow's phenotype is the only information available and data collection is often limited due to high costs or logistical problems. A constraining factor to this approach is that validation bulls would come from specific groups in the population and, therefore, maintenance of genetic diversity could be a concern. The same procedure may also be of use in smaller animal breeding programs, where the validation population is represented by individuals from specific sub-groups of a bigger population (DING et al., 2013; ANDONOV et al., 2016; UEMOTO et al., 2017).

There has been considerable concern on controlling inbreeding in cattle populations in the genomics era (HOWARD et al., 2017). Doekes et al. (2018) have investigated genetic diversity trends in the Dutch-Flamish Holstein-Friesian bull population, and reported that after the implementation of GS, inbreeding and kinship have increased substantially. Especially in small populations, the loss in genetic variance stand as an obstacle to improve genetic gains when performing genomic selection for many generations (JANNINK et al., 2010). Community "proportional sampling" could stand as a tool to prevent loss of genetic variance imposed by the increase on inbreeding caused by intense selection. In a long-term, including individuals from distinct well-connected groups in the reference population can result in higher reliabilities for animals not strongly

connected to the selected population and, thus, could help to identify high merit bulls/cows to be used in mating systems aiming to prevent the increase of kinship and inbreeding. Although convincing breeders to use a wider panel of genomically evaluated bulls could be an obstacle (BOUQUET and JUGA, 2013).

As GS establishes as a consolidated method to improve genetic gains, strategies to maintain its efficiency in the long-term must be considered. In that sense, updating the reference population adequately is crucial to maintain a constant level of accuracy through generations of GS (WU et al., 2015; CALUS, 2016). The proportional sampling procedure here presented can be used for choosing the optimal set of cows to supplement the reference population over generations in order to maximize the estimated accuracy of selection candidates. At each generation (or year), the new selection candidates (forming the current validation population) would be included in the complete pedigree, requiring the assessment of their community and, afterwards, new phenotyped cows could be genotyped and included in the reference population based on the proportion of individuals in each community inside the current validation population. Further studies would help to quantify long-term impacts of sampling cows to supplement the reference population by the methods here presented over population's inbreeding, homozygosity and genetic gains.

3.5 CONCLUSION

An unquestionable increase in accuracy was observed when cows were included in the reference population of bulls. Community based proportional cow sampling have slightly overcome traditional genotyping strategies, yielding higher accuracy and lesser

bias on genomic evaluations. The proposed methods also resulted in a set of cows less connected within the reference population and between reference and validation populations. Community-based cow sampling showed potential as a tool for building cow-only reference populations for maximizing SNP marker value in situations such as an expensive to measure trait.

3.6 REFERENCES

- ANDONOV, S., LOURENCO, D. A. L., FRAGOMENI, B. O., MASUDA, Y., POCRINIC, I., TSURUTA, S., & MISZTAL, I. Accuracy of breeding values in small genotyped populations using different sources of external information – A simulation study. **Journal of Dairy Science**, 100:1-7, 2017.
- BLONDEL, V. D., GUILLAUME, J. L., LAMBIOTTE, R., & LEFEBVRE, E. Fast unfolding of communities in large networks, 1–12. **Journal of Statistical Mechanics: Theory and Experiment**, 10, 10008, 2008.
- BOUQUET, A., & JUGA, J. Integrating genomic selection into dairy cattle breeding programs: a review. **Animal**, 7:5, 705-713, 2013.
- BUCH, L.H., SØRENSEN, A.C., LASSEN, J., BERG, P., ERIKSSON, J.A., JAKOBSEN, J.H., & SØRENSEN, M.K. Hygiene-related and feed-related hoof diseases show different patterns of genetic correlations to clinical mastitis and female fertility. **Journal of Dairy Science**, 94, 1540-1551, 2011.
- CALUS, M. P. L., De HAAS, Y., PSZCZOLA, M., & VEERKAMP, R. F. Predicted accuracy of and response to genomic selection for new traits in dairy cattle. **The Animal**

Consortium, 7(2), 183–191, 2013

CALUS, M. P. L. Editorial: Genomic selection with numerically small reference populations. **Animal**, 10:6, pp 1016-1017, 2016.

CLARK, S. A., HICKEY, J.M., DAETWYLER, H.D., & VAN DER WERF, J.H.J. 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 schemes. **Genetics Selection Evolution**, 44:4, 2012.

COLE, J. Pypedal: a computer program for pedigree analysis. **Computer and Electronics in Agriculture**, 51(1), 107-113, 2007.

DASSONNEVILLE, R., BAUR, A., FRITZ, S., BOICHARD, D., & DUCROCQ, V. Inclusion of cow records on genomic evaluations and impact on bias due to preferential treatment. **Genetics Selection Evolution**, 44:40, 2014.

De ROOS, P. W., HAYES, B. J., SPELMAN, R. J., & GODDARD, M. E. Linkage disequilibrium and persistence of phase in Holstein Friesian, Jersey and Angus cattle. **Genetics**, 179:1503–1512, 2008.

DING, X., ZHANG, Z., LI, X., WANG, S., WU, X., SUN, D., YU, Y., LIU, J., WANG, Y., ZHANG, Y., ZHANG, S., ZHANG, Y., & ZHANG, Q. Accuracy of genomic prediction for milk production traits in the Chinese Holstein population using a reference population consisting of cows. **Journal of Dairy Science**, 96:5315-5323, 2013.

DOEKES, H. P., VEERKAMP, R. F., BIJMA, P., HIEMSTRA, S. J., & WINDIG, J. J. Trends in genome-wide and region-specific genetic diversity in the Dutch-Flemish

Holstein-Friesian breeding program from 1986 to 2015. **Genetics Selection Evolution**, 2013, 50.15, 2013.

EDEL, C., PIMENTEL, E. C. G., PLIESCHKE, L., EMMERLING, R., & GÖTZ, K. U. Short Communication: The effect of genotyping cows to improve the reliability of genomic predictions for selection candidates. **Journal of Dairy Science**, 99:1-6, 2016.

EGGER-DANNER, C., SCHWARZENBACHER, H., & WILLIAN, A. Short communication: Genotyping of cows to speed up availability of genomic estimated breeding values for direct health traits in Austrian Fleckvieh (Simmental) cattle – Genetic and aspects. **Journal of Dairy Science**, 97:7, 4552-4556, 2014.

GAO, H., LUND, M. S., ZHANG, Y. & SU, G. Accuracy of genomic prediction using different models and response variables in the Nordic Red cattle population. **Journal of Animal Breeding and Genetics**, 130:333-340, 2013.

GARCIA-RUIZ, A., COLE, J. B., VAN RADEN, P. M., WIGGANS, G. R., RUIZ-LOPES, F. J., & VAN TASSEL, C. P. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. **Proceedings of the National Academic Science USA**, Jul 12, 113(28):E3995-4004, 2016.

GARRICK, D. J., TAYLOR, J. F., & FERNANDO, R.L. De-regressing estimated breeding values and weighting information for genomic regression analysis. **Genetics Selection Evolution**, 41:55, 2009.

GIRVAN, M., & NEWMAN, M. E. J. Community structure in social and biological networks, 99(12), 2002.

- HABIER, D., FERNANDO, R. L., & DEKKERS, J. C. M. The impact of genetic relationship on genome-assisted breeding values. **Genetics**, 177:2389-2397, 2007.
- HABIER, D., TETENS, J., SEEFRIED, F. R., LICHTNER, P., & THALLER, G. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. **Genetics Selection Evolution**, 42:5, 2010.
- HAYES, B. J., BOWMAN, P. J., CHAMBERLAIN, A. J., & GODDARD, M. E. Genomic selection in dairy cattle: progress and challenges. **Journal of Dairy Science**, 92(2), 433–443, 2009.
- HENDERSON, C. R. Best Linear Unbiased Estimation and Prediction under a Selection Model. **Biometrics**, 31(2), 423–447, 1975.
- HENRYON, M., BERG, P., SORENSEN, A. C., OSTERSEN, T., & NIELSEN, B. Most of the benefits from genomic selection can be realised by genotyping a proportion of selection candidates. **Journal of Animal Science**, 90, 4681–4689, 2015.
- HILL, W.G. & ROBERTSON A. Linkage disequilibrium in finite populations. **Theoretical and Applied Genetics**, 38(6):226-31, 1968.
- HOWARD, J. T., PRYCE, J. E., BAES, C., & MALTECCA, C. Invited review: Inbreeding in the genomics era: Inbreeding, inbreeding depression, and management of genetic diversity. **Journal of Dairy Science**, 100(8):6009-6024, 2017.
- ISIDRO, J., JANNINK, J.L., AKDEMIR, D., POLAND, J., HESLOT, N., & SORRELLS, M. E. Training set optimization under population structure in genomic selection. **Theoretical and Applied Genetics**, 128(1), 145–158, 2014.

- JENKO, J., WIGGANS, G. R., COOPAR, T. A., EAGLEN, S. A. E., LUFF, W. G. L., BICHARD, M., PONG-WONG, R., & WOOLIAMs J. A. **Journal of Dairy Science**, 100(1):439-452, 2017.
- JIMÉNEZ-MONTERO, J. A., GONZÁLEZ-RECIO O., & ALENDA R. Genotyping strategies for genomic selection in small dairy cattle populations. **Animal**, 6, 1216–1224, 2012.
- KOIVULA, M., STRANDÉN, I., AAMAND, G. P., & MÄNTYSAARI, E. A. Effect of cow reference group on validation reliability of genomic evaluation. **Animal**, 10:1021–6, 2016.
- LAMBIOTTE, R., BLONDEL, V., de KERCHOVE, C., HUENS, E., PRIUER, C., SMOREDA, S., & VAN DOOREN P. Geographical dispersal of mobile communication networks. **Physica A: Statistical Mechanics and its Applications**, 387:21,5317-5325, 2008.
- MCHUGH, N., MEUWISSEN, T. H. E., CROMIE, A. R., & SONESSON, A. K. Use of female information in dairy cattle genomic breeding programs. **Journal of Dairy Science**, 94:4109-4118, 2011.
- MEUWISSEN, T. H. E., HAYES, B. J., & GODDARD, M.E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, 157(4), 1819–1829, 2001.
- MISZTAL, I., TSURUTA, S., LOURENCO, D., AGUILLAR, I., LEGARRA, A. & VITEZICA, Z. Manual for BLUPF90 family programs. Accessed July, 20, 2017.
- MOGHADDAR, N., GORE, K. P., DAETWYLER, H. D., HAYES, B. J., & VAN DER WERF,

- J. H. J. Accuracy of genotype imputation based on random and selected reference sets in purebred and crossbred sheep populations and its effect on accuracy of genomic prediction. **Genetics, Selection, Evolution : GSE**, 47, 97, 2015.
- OLSON, K. M., VANRADEN, P. M., TOOKER, M. E., & COOPER, T. A. Differences among methods to validate genomic evaluations for dairy cattle. **Journal of Dairy Science**, 94, 2613-2620, 2011.
- PLIESCHKE, L., EDEL, C., PIMENTEL, E. C. G., EMMERLING, R., BENNEWITZ, J., & GOTZ K. U. Systematic genotyping of groups of cows to improve genomic estimated breeding values of selection candidates. **Genetics Selection Evolution**, 48:73, 2016.
- PRYCE, J.E., & DAETWYLER, H. D. Designing dairy cattle breeding schemes under genomic selection: a review of international research. **Animal Production Science**, 52, 107-114, 2012.
- PSZCZOLA, M., STRABEL, T., MULDER, H. A., & CALUS, M. P. L. Reliability of direct genomic values for animals with different relationships within and to the reference population. *Journal of Dairy Science*, 95(1), 389–400, 2012.
- RINCENT, R., CHARCOSSET, A., & MOREAU, L. Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. **Theoretical and Applied Genetics**, 130(11):2231-2247, 2017.
- SARGOLZAEI, M., & SCHENKEL, F. S. QMSim: a large-scale genome simulator for livestock. **Bioinformatics**, 25:680–1, 2009.

- SCHÖPKE, K. & SWALVE, H. H. Review: Opportunities and challenges for small populations of dairy cattle in the era of genomics. **Animal**, 10:6, 1050-1060, 2016.
- THENGHE, A. M. M, BAOUWMAN, A. C., BERGLUND, B., de KONING, D. J., VEERKAMP, R.R. Improving accuracy of bull's predicted genomic breeding values for fertility using daughters' milk progesterone profiles. **Journal of Dairy Science**, in press, 2018
- THOMASEN, J. R., SORENSEN, A. C., LUND, M. S., & GULDBRANDSEN B. Adding cows to the reference population makes a small dairy population competitive. **Journal of Dairy Science**, 97(9):5822-32, 2014.
- UEMOTO, Y., OSAWA, T., & SABURI, J. Effect of genotyped cows in the reference population on the genomic evaluation of Holstein cattle. **Animal**, 11:3, 382-393, 2017.
- VENTURA, R. V., LARMER, S., SCHENKEL, F. S., MILLES, S. P., SULLIVAN, P. Genomic clustering helps to improve prediction in a multibreed population. **Journal of Animal Science**, 94(5):1844-56, 2016.
- VUKASINOVIC, N., BACCIU, N., PRZYBYLA, C. A., BODDHIREDDY, P., & DENISE, S. K. Development of genetic and genomic evaluation for wellness traits in US Holstein cows. **Journal of Dairy Science**, 100:1, 428-438, 2017.
- ZHOU, L., HERINGSTAD, B., Su, G., GULDBRANDSEN, B., MEUWISSEN, T. H. E., SVENDSEN, M., GROVE, H., NIELSEN, U. S., & LUND, M.S. Genomic predictions based on a joint reference population for the Nordic Red Cattle breeds. **Journal of Dairy Science**, 97, 4485-4496, 2014.

WIENTJES, Y. C. J., CALUS, M. P. L., GODDARD, M. E., & HAYES, B. J. Effect of genetic architecture of multi breed genomic prediction. In **Proceedings of the 10th WCGALP**, Vancouver, BA, Canada, 17 to 22 August 2014.

WU, X., LUND, M. S., SUN, D., ZHANG, Q., & SU, G. Impact of relationship between test and training animals and among training animals on reliability of genomic prediction. **Journal of Animal Breeding and Genetics**, 132, 366-375, 2015.

CHAPTER 4. ASSESSING GENOTYPE AND PHENOTYPE INFORMATION VALUE WHEN CONSIDERING DISTINCT WELL-CONNECTED GROUPS OF COWS OVER GENOME-WIDE ASSOCIATION STUDIES RESULTS FOR LOWLY HERITABLE TRAITS.

Bruno da Costa Perez¹, Ricardo Vieira Ventura², Gerson Antonio Oliveira Junior³,
Fernando de Oliveira Bussiman², Julio Cesar de Carvalho Balieiro²

¹ *Department of Animal Science, College of Animal Science and Food Engineering,
University of São Paulo (FZEA/USP), Pirassununga, 13630-000 São Paulo, Brazil*

² *College of Veterinary Medicine and Animal Science, University of São Paulo
(FMVZ/USP), Pirassununga, 13630-000 São Paulo, Brazil*

³ *Centre for Genetic Improvement of Livestock, Department of Animal Biosciences,
University of Guelph, Guelph, ON, Canada.*

ABSTRACT The objective of the present study was to investigate the impact of considering the complete population structure when choosing the set of genotyped individuals over results obtained from genome-wide association studies (GWAS). We also aimed to test if the inclusion of information from (genotyped and non-genotyped) cows by different criteria could bring value to the models; and to evaluate the adaptability and efficiency of different models (GBLUP, ssGBLUP and Bayes R) for a low heritability trait. A small dairy cattle population was simulated and, based on three traditional genotyping designs (random, top-yield and extreme-yield cows), different numbers (1,000 and 5,000) of cows were sampled and included in the genotypic dataset previously containing progeny tested bulls. The traditional designs were replicated considering or not population structure and compared considering their ability to detect quantitative trait loci via GWAS by different genomic evaluation models. The inclusion of genotypes from phenotyped

cows in the GWAS analysis improved the ability to detect the existing QTLs. Community-based genotyping designs showed slight superior ability to detect influential genomic regions and the magnitude of the genetic variance present in these regions when compared to traditional methods. In general, Bayes R and ssGBLUP (third iteration) detected higher amounts of genetic variance explained by SNP windows than GBLUP and ssGBLUP (first iteration) models. The inclusion of phenotypic information from ungenotyped cows via ssGBLUP model has yielded a slight superiority when compared to GBLUP, but showed similar results to the Bayes R model. Results found in the present study suggest that considering population partition for choosing cows to be genotyped by extreme phenotype values may provide the best resolution in GWAS for detecting QTL for a sex-limited low heritability trait.

Keywords: genetic architecture, dairy cattle, genotyping strategies, graph community, quantitative trait loci

4.1 INTRODUCTION

Genome-wide association studies (GWAS) have been widely used for QTL detection using single-nucleotide polymorphism (SNP) marker data in animals, plants and humans (HIRSCHHORN AND DALY., 2005; HAYES AND GODDARD., 2010; HUANG AND HAN, 2014). It plays an important role as a tool to better comprehend the genetic architecture permeating complex traits. Despite the large number of methods available (MEUWISSEN et al., 2001; GIANOLA et al., 2009; GARRICK et al., 2014), assessing regions on the genome which influence the genetic variance present in traits

of economic interest, especially in low heritability or difficult- or- expensive to measure (often sex-limited) traits, is still a difficult task (HAYES AND GODDARD., 2010; BLASCO AND TORO, 2014). For such traits, the availability of records is often limited, especially when considering phenotypes only from genotyped animals.

GWAS have been primarily performed using Bayesian multiple-regression models (MEUWISEN et al., 2001; HABIER et al., 2011; FERNANDO AND GARRICK, 2013; VAN DEN BERG et al., 2013). Such models are flexible, allowing prior knowledge about the distribution of SNP marker effects to be considered, which is an advantage specially when modeling traits in which variance is affected by QTL of major effect. Moser et al. (2015) have presented and tested a Bayesian mixture model (Bayes R) that assumes the distribution of SNP effects comes from more than one component with a mass density in zero. The authors obtained superior results for mapping influential SNPs and estimate the genetic variance explained genomic regions collectively for human diseases (and via simulation) than other traditional methods.

The single-step genomic best linear unbiased prediction (ssGBLUP, MISZTAL et al., 2009; CRISTENSEN & LUND, 2010) method is an elegant strategy (using a traditional GBLUP approach) to include both genotyped and non-genotyped animals in genomic analyzes and overcome some challenges imposed by situations where records are difficult or expensive to obtain, or when the “genotyped and phenotyped” population is small. It was first applied to genomic selection, but further implementations weighting marker effects (WANG et al., 2012) turned possible to consider such method for QTL mapping studies (WANG et al., 2014, TIEZZI et al., 2015, HAN AND PEÑAGARICANO,

2016) by shrinking effects of non-influential SNPs in order to evidence genomic regions strongly associated to traits of interest.

The structure of the set of genotyped individuals is crucial for the results obtained from genomic analyzes and a wide range of studies have assessed its impact over accuracy of genomic selection in real and simulated data (JIMENEZ-MONTERO et al., 2012, GAO et al., 2015; JENKO et al., 2017; PLIESCHKE et al., 2017; UEMOTO et al., 2017) in the scope of genomic selection. Despite genomic selection being related to GWAS in a plethora of aspects, there are crucial differences, as the main objective of GWAS is to investigate the genetic architecture of quantitative traits whereas in genomic selection the aim is to predict breeding values to support selection (BLASCO AND TORO., 2014; MEUWISSEN et al., 2016). In this sense, despite some studies have focused in assessing the efficiency of selective genotyping for GWAS (BOVENHUIS AND SPELMAN, 2000; GALLAIS et al., 2007), not much effort was put into investigating impacts of considering population structure to obtain the genotype dataset over results of QTL fine-mapping through GWAS.

The objective of the present study was to investigate, via simulation: 1) the impact of considering the complete population structure when choosing the set of genotyped individuals to be included for genome-wide association studies, 2) test if the inclusion of information from (genotyped and non-genotyped) cows by different criteria could bring value to the models, and 3) to evaluate the adaptability and efficiency of different models (GBLUP, ssGBLUP and Bayes R) over each proposed scenario for a low heritability complex trait.

4.2 MATERIAL AND METHODS

4.2.1 Simulated population and genome

We used the free access software QMSim (SARGOLZAEI AND SCHENKEL, 2009) to perform the proposed simulation scenarios, considering five replicates. The aim was to simulate a small dairy cattle population based on previous results presented by Jimenez Montero et al. (2012) and Plieschke et al. (2016). A historical population initially containing 1,500 unrelated individuals with balanced sex ratio was created, in which animals were randomly mated for 1,000 generations. From that point, the historical population size was reduced from 1500 to 400 animals in 20 generations. A second population was created in which founders consisted of 100 males and 300 females randomly sampled from the last generation of the historical population (generation 1020). This expansion population was maintained for 60 generations considering a random mating system and growth rate of 1.2x for both sexes by generation. A recent population was created by sampling 400 males and 20,000 females from the last generation of the expansion population. The recent population was run for 20 overlapping generations. In each generation, 10,000 male and 10,000 female offspring were simulated. Replacement rates were 30% for dams and 50% for sires. Genealogical information was recorded for all individuals from the recent population. Breeding animals were selected based on their BLUP-based breeding values (here we used EBV) and culled by age. Cows from generations 16 to 20 and sires from generations 15 to 18 had genotypic data recorded. Bulls with more than 40 daughters were considered as progeny tested individuals ($n = 764$) and assigned to the reference set. Animals from generation 20 were considered as validation population.

The simulated genome had 29 pairs of autosome chromosomes with total length of 2333 cM, in which each chromosome's length mimicked the bovine genome without sex chromosomes. The intent was to generate a more plausible scenario when considering real distances between markers and QTL loci. A total of 50,000 bi-allelic SNP markers were evenly distributed throughout the genome, in which the number of SNP per chromosome was proportional to its size. The additive genetic component was determined by 750 randomly distributed QTL along the genome, which effects were sampled from a gamma distribution with a shape parameter of 0.4 (HAYES AND GODDARD, 2001). The number of QTL (750) was defined aiming to assure a polygenic nature for the simulated trait. The mutation rate was identical (2.5×10^{-5}) for both markers and QTL (JIMENEZ-MONTERO et al., 2012). The rate of missing marker genotypes was 0.01 and the rate of marker genotyping error was 0.005.

4.2.2 Phenotypes

A single sex-limited trait with heritability of 0.15 and phenotypic variance of 1.0 was simulated. Yield-deviations (YD; VANRADEN AND WIGGANS, 1991) were calculated using a combination of the TBV and a residual component for every candidate cow, and thus, used as dependent variable in the GWAS. The YD were then used to calculate daughter yield deviations (DYD; VANRADEN AND WIGGANS, 1991) for the progeny-tested bulls. All phenotypic information included as dependent variable in the model was weighted by its reliability in order to account for different variances between YD and DYD.

4.2.3 Graph community detection and population partitioning

A “pedigree graph” was generated by transforming genealogical information into a matrix (adjacency matrix) of zeros and ones, with dimensions $n \times n$, where n is the total number of individuals in the pedigree, that represent direct connections between individuals (parent-progeny). In practical terms, this basic approach consists of transforming the numerator relationship matrix (A ; HENDERSON, 1975) into a graph adjacency matrix where the following criteria must be met: elements of the pedigree graph input adjacency matrix receive the value 1 when the respective A_{ij} element is 0.50, and 0 otherwise. A more detailed description of this method is found in Chapter 3 (Material and Methods section).

4.2.4 Selective cow genotyping designs

Cows from generations 16 through 19 of the recent population represented a contemporary overlapping active population of 40,000 individuals. From them, 1000 and 5000 were sampled as female candidates to be genotyped, based on the methods described further in this section.

For all reference population sizes proposed (1000 and 5000 cows), a total of 7 different genotyping designs were considered. Of these, four traditional designs were included as a base for comparison, being: 1) ϕ females randomly selected (RND), 2) ϕ females with the highest YD (T_{YD}), and 3) $\phi/2$ females with the highest and $\phi/2$ females with the lowest YD (EX_{YD}), where ϕ was 1000 or 5000, depending on the reference population scenario.

Other designs were proposed as a way to investigate both efficiency and adaptability of the community partitioning algorithm (BLONDEL et al., 2008). To accomplish that, a genealogical dataset containing all individuals from generations 12 to 20 ($n = 180,000$) of the recent contemporary population was built and analyzed as a graph. Animals from generation 12 had their parents omitted (set as “unknown”) to represent founder individuals of the pedigreed population of a relatively small cattle population. The algorithm designated a community for every animal in the pedigree. Community-based designs were primarily aimed to integrate communities and the traditional “high or extreme” YD. The representativeness of each detected community in the validation population was considered when sampling cows to be included into the reference population in each situation. These cow genotyping designs were elaborated as follows: 4) α females with the highest YD inside each detected community (T_{YDC}); 5) $\alpha/2$ females with the highest and $\alpha/2$ females with the lowest YD inside each detected community (EX_{YDC}); where α is the representativeness of the respective community in the total number of individuals defined as the validation population, calculated as $\alpha = (N_{ref_i} * N_{smp_l_j}) / n_{total}$, where N_{ref_i} is the size of reference population for the i^{th} scenario (1000 or 5000), $N_{smp_l_j}$ is the size of the j^{th} community and n_{total} is the total number of candidate cows for the reference population. Therefore, the number of cows sampled from each community was directly linked to the proportion of individuals from this same community in the validation population. For further discussion on this document, this procedure will be referred at as “proportional sampling”. All proposed scenarios are described in Table 4.1.

Table 4.1 Summary of genotyping scenarios and number of genotypes for bulls and cows in the genotype dataset for each design proposed.

Design	Summary of genotype dataset
RND	764 progeny tested bulls + ϕ cows randomly genotyped.
T _{YD}	764 progeny tested bulls + ϕ cows with highest YD genotyped.
EX _{YD}	764 progeny tested bulls + $\phi/2$ cows with highest and $\phi/2$ cows with lowest YD genotyped.
T _{YDC}	764 progeny tested bulls + $\alpha/2$ females with the highest and $\alpha/2$ females with the lowest YD values from within each detected community.
EX _{YDC}	764 progeny tested bulls + α females with the highest YD values from within each detected community.

ϕ = depended on the scenario considered (1000 or 5000); $\alpha = (Nref_i * Nsmpl_j) / n_{total}$, where $Nref_i$ is the size of the genotype dataset for the i^{th} scenario (1000 or 5000), $Nsmpl_j$ is the size of the j^{th} community and n_{total} is the total number of candidate cows from generations 16 through 19 of the recent population.
RND = at random; T_{YD} = top yield deviation values; EX_{YD} = extreme yield deviation values;
T_{YDC} = top de-regressed proof values inside communities;
EX_{YDC} = extreme de-regressed proof values inside communities.

4.2.5 Quality control and genome-wide association study

Genotype data quality control was performed prior to all GWAS analyzes in the present study. In all proposed scenarios, criteria were: SNPs with minor allele frequency lower than 0.02 and call rate lower than 0.95, as well as samples with call rate lower than 0.90 were excluded from analysis. After quality control, around 43,500 SNPs remained, with slight variations depending on the scenario considered and the simulation replicate.

Three distinct statistical methods were used to perform the proposed association analysis: GBLUP (VAN RADEN, 2008), Bayes R (MOSER et al., 2015) and WssGBLUP (WANG et al., 2012).

4.2.5.1 Genomic best linear unbiased prediction (GBLUP)

The GBLUP model was implemented following the general equation:

$$y = \mu 1 + Zg + e, \quad [6]$$

where y is the vector of the dEBV for the genotyped individuals in the reference population; μ is the general mean (no fixed effects were simulated); $\mathbf{1}$ is a vector of ones; $\mathbf{Z}$ is an incidence matrix allocating records to breeding values; $\mathbf{g}$ is a vector of GEBV with $\text{var}(\mathbf{g}) = \mathbf{G}\sigma_g^2$, in which σ_g^2 is the additive genetic variance; and $\mathbf{G}$ is the realized genomic relationship matrix created using the method described in VanRaden (2008). That is, $\mathbf{G} = (\mathbf{M} - \mathbf{P})(\mathbf{M} - \mathbf{P})' / \sum_{i=1}^m 2p_i(1 - p_i)$, where $\mathbf{M}$ is an $n \times m$ matrix (number of animals $\times$ number of loci) with SNP coded 0, 1 and 2 for genotypes A1A1, A1A2 and A2A2, respectively. $\mathbf{P}$ is an $n \times m$ matrix containing twice the allele frequencies expressed as $P_i = 2p_i$ where p_i is the allele frequency of the homozygous genotype coded with 2 for all genotyped individuals at locus i . It is assumed that $e \sim N(0, \sigma_e^2)$, where σ_e^2 is the residual variance. All analyzes were conducted under a REML approach, using BLUPF90 family software (MIZSTAL et al., 2017).

4.2.5.2 Weighted single-step GBLUP

The WssGBLUP method considered was based on the following model: $\mathbf{y} = \mu\mathbf{1} + \mathbf{Z}_a\mathbf{a} + \mathbf{e}$, where y is the vector of YD/DYD; μ is the general mean (no fixed effects were simulated); $\mathbf{1}$ is a vector of ones; $\mathbf{Z}_a$ is an incidence matrix that relates animals to phenotypes; $\mathbf{a}$ is the vector of direct additive genetic effects and $\mathbf{e}$ is the vector of random residuals.

Assuming a matrix form for the model above, the covariance matrix for the random terms is given by:

$$\text{var} \begin{bmatrix} \mathbf{a} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} H\sigma_a^2 & \mathbf{0} \\ \mathbf{0} & I\sigma_e^2 \end{bmatrix} \quad (2)$$

where H is the combined pedigree-genomic relationship matrix; its inverse is given by:

$$H^{-1} = \begin{bmatrix} A_{11}^{-1} & A_{12}^{-1} \\ A_{21}^{-1} & G^{-1} - A_{22}^{-1} \end{bmatrix} \quad (3)$$

where A_{11}^{-1} is the inverse of the first quadrant of the pedigree relationship matrix (animals without genotype); A_{12}^{-1} is the inverse of the second quadrant of the pedigree relationship matrix (animals without genotype); A_{21}^{-1} is the inverse of the third quadrant of the pedigree relationship matrix (animals without genotype) and $G^{-1} - A_{22}^{-1}$ is the difference between the realized (G^{-1}) and expected (A_{22}^{-1}) relationship matrices for the genotyped animals. In this case, G was constructed according to VanRaden (2008), and was assumed the allele frequencies of the population under analyses and adjusts for compatibility with A_{22} , as proposed by Chen et al. (2011).

To calculate SNP effects ($\hat{l}$), it was considered a weighted genomic relationship matrix G^* through the equation: $\hat{l} = DZ'[ZDZ']^{-1}\hat{a}$ following the steps described by Wang et al. (2012, 2014). A total of three iterations (w1, w2 and w3) for the WssGBLUP were performed for each scenario, resulting in an increasing shrinkage from w1 to w3 for the SNPs explaining lower variance and, consequently, in an increasing proportion of variance being explained by the remaining markers. Thus, for this paper, the notation WssGBLUPw1 refers to the scenario using weight w1 (first iteration), WssGBLUPw2 to scenario using w2 (second iteration), and WssGBLUPw2 refers to the scenario using weight w3 (third iteration).

To assess the value of phenotypic information from non-genotyped cows for QTL mapping (via GWAS), the ssGBLUP models (w1, w2 and w3) were applied considering

all available phenotypes of candidate cows from generations 16 to 19 of the recent simulated population (N = 40,000).

4.2.5.3 Bayes R

For the Bayes R method, phenotypes (YD and DYD) were related to markers with a standard linear regression model:

$$y = \mu \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where $\mathbf{y}$ is a n -dimensional vector of phenotypes, $\mathbf{1}_n$ is a n -dimensional vector of ones, μ is the general mean, $\mathbf{X}$ is an $n \times p$ matrix of genotypes encoded as 0, 1 or 2 copies of a reference allele. The vector $\boldsymbol{\beta}$ is a p -dimensional vector of SNP effects and $\mathbf{e}$ is a n -dimensional vector of residuals, $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$ with $\mathbf{I}$ being a $n \times n$ identity matrix.

It was assumed that SNP effects come from a finite mixture of K components so that the probability of the $\boldsymbol{\beta}$ effects conditional on the variance of the components $\sigma^2 = (\sigma_1^2, \dots, \sigma_K^2)$ and the mixture proportions $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ which are constrained to be positive and to sum to unity:

$$p(\boldsymbol{\beta}|\boldsymbol{\pi}, \sigma^2) = \sum_{k=1}^K \pi_k N(\boldsymbol{\beta}|0, \sigma_k^2),$$

where $N(\boldsymbol{\beta}|0, \sigma_k^2)$ denotes the density function of the univariate normal distribution with mean 0 and variance σ_k^2 .

The Bayesian approach requires the assignment of prior distributions to all unknowns in the model. We followed Erbe et al. (2012) and *a priori* assumed a mixture of

four zero mean normal distributions, where the relative variance for each mixture component is fixed:

$$p(\beta|\pi, \sigma_g^2) = \pi_1 \times N(0, 0 \times \sigma_g^2) + \pi_2 \times N(0, 10^{-4} \times \sigma_g^2) + \\ \pi_3 \times N(0, 10^{-3} \times \sigma_g^2) + \pi_4 \times N(0, 10^{-2} \times \sigma_g^2),$$

where σ_g^2 is the additive genetic variance explained by SNPs. For the present study, we considered $\pi_1 = 0.999$, $\pi_2 = 0.0005$, $\pi_3 = 0.0003$ and $\pi_4 = 0.0002$. Sparseness is considered into the model by setting the effect and variance of the first mixture component to zero. Analyses for the Bayes R model were conducted in the GCTB software (ZENG et al., 2018).

4.2.6 Comparison of models and scenarios

Based on the proposed by Melo et al. (2016) for assessing the efficiency of GWAS methods via simulation, the GWAS results were compared based on estimates of the proportions of genetic variance explained by SNPs within consecutive non-overlapping 1-Mb windows. The genetic variance explained by each locus was computed as follows (FALCONER AND MACKAY, 1996): $\alpha_j = 2p_j(1 - p_j)\beta_j^2$, where p_j is the minor allele frequency of the j^{th} SNPs and β_j its estimated effect. The α values for all locus were then grouped in 1Mb SNP windows to calculate the proposed parameters, as follows: number of QTLs explaining 0.5% or more of the genetic variance (topQTL); number of top 1-Mb marker windows accounting for the highest proportion of variance explained by markers (topMRKw) – this number was set equal to topQTL so that the different methods could be

compared on the same basis; sum of the percentages of genetic variances explained by the topQTL set (Pvar_topQTL) and top marker windows (Pvar_topMRKw); highest percentage of genetic variance explained by a topQTL (Pvar_1stQTL) and a top marker window (Pvar_1stMRKw), and the estimated number of topQTL detected by a topMRKw located no more than 1Mb from a true QTL position.

4.3 RESULTS

4.3.1 Genomic data structure description and population partition

The total number of communities detected for the five simulation replicates was 134, 129, 132, 134 and 132. The mean of the average relationship (calculated using a pedigree-based relationship matrix) for animals inside and outside their community was 0.04% and 0.005%, respectively. Therefore, there are indications that the community detection algorithm has successfully grouped individuals more genetically connected among themselves than to the rest of the population.

In Figure 4.1 it is shown a brief description of the observed minor allele frequency (MAF) before quality control checks and linkage disequilibrium, measured by r^2 (Hill and Robertson, 1968) for genotyped cows from generations 16 through 20.

The observed average minor allele frequency distribution exhibited a high frequency of MAF < 0.02 and a relatively constant frequency of MAF values for the interval between 0.02 and 0.50. Linkage disequilibrium found for adjacent SNP markers varied from 0.247 (BTA 6) to 0.206 (BTA 29), depending on the chromosome.

Figure 4.1 Average minor allele frequency distribution (red dotted line represents the quality control threshold set) (top) and the average linkage disequilibrium (r^2) for adjacent SNP (bottom) observed for cows from generations 16 through 20.

4.3.2 Parameters of comparison

The simulation process resulted in an average number of QTLs explaining 0.5% or more of the genetic variance (topQTL) of 13.6. Altogether, the topQTL explained 34.1% of the genetic variance, with the most important QTL explaining on average 7.54%.

In Figure 4.2 it is shown the average $Pvar_topMRKw$ for all genotyping designs proposed. Except for EXT_{YD} under Bayes R and WssGBLUPw3 models, the top marker windows detected a smaller proportion of the genetic variance when compared to the topQTL. Community-based genotyping designs resulted in slightly higher values of $Pvar_topMRKw$ detected when compared to their traditional versions, independent of the model considered.

The average $Pvar_1^{st}MRKw$ found for the simulated population is shown in Figure 4.3. Independent of the scenario proposed, lower amounts of genetic variance explained by top markers was detected for models that assume a simple Gaussian distribution for the effect of SNPs (GBLUP and WssGBLUPw1), being followed by the WssGBLUPw2 and, finally, the Bayes R and WssGBLUPw3 models. It was also observed a higher value of $Pvar_1^{st}MRKw$ when community-based selective genotyping of cows was considered despite the model considered and the size of the genotyped dataset.

Overall, when comparing traditional genotyping methods (RND, T_{YD} and EX_{YD}), the ability to detect topQTL was higher for EX_{YD} (Figure 4.4), in which the average trueQTL varied from 4 to 6 and from 5 to 8 depending on the method, respectively, for including genotypes from 1000 and 5000 cows. The lowest average trueQTL was estimated using the T_{YD} method, varying from 1 to 4 and 2 to 5 depending on the model, respectively, for including 1000 and 5000 cows. Results for RND were placed in between the observed for EX_{YD} and T_{YD} (Figure 4.4), yielding an average trueQTL varying from 2 to 6 and from 3 to 7 depending on the model, respectively, for including genotypes for 1000 and 5000 cows.

Increasing the number of cows with genotypes included in the GWAS analyzes from 1,000 to 5,000 has slightly improved the ability to detect influential genomic regions (Figures 4.2, 4.3 and 4.4). For the trueQTL parameter, observed differences between 1,000 to 5,000 cows varied from +1 to +3, +1 and from +2 to +3 respectively, for RND, T_{YD} and EX_{YD} strategies. When considering community-based cow genotyping strategies, differences between 1,000 and 5,000 cow genotypes for the trueQTL parameter were, on average, of +2.

When genotyping cows considering a proportional sampling approach ($T_{YD}C$ and $E_{YD}C$), a larger portion of variance explained by top markers was captured when compared to their traditional versions (for T_{YD} and EX_{YD} , Figure 4.3). The inclusion of updated SNP weights in the ssGBLUP method ($W_{ssGBLUPw2}$ and $W_{ssGBLUPw3}$) resulted in a slightly superior ability to map QTLs for the $W_{ssGBLUPw3}$ model. When under a community-based genotyping approach, analyzes using the $W_{ssGBLUPw3}$ model capture the highest amount of genetic variance among all models tested.

Figure 4.2 Average^a (and respective standard deviation) for the genetic variance (%) explained by the sum of variances accounted by top marker windows ($Pvar_topMRKw$) of the simulated population for all genotyping designs proposed.

^aThe averages are expressed over five simulation replicates; Red line = True values for simulated data were $Pvar_topQTL$ (%) = 30.18 ± 5.95 ; RND = at random; T_{YD} = top yield deviation values; EX_{YD} = extreme yield deviation values; T_{YDC} = top yield deviation values inside communities; EX_{YDC} = extreme yield deviation values inside communities.

Figure 4.3 Average^a (and respective standard deviation) for the maximum genetic variance (%) explained by a top marker window ($Pvar_{1stMRKw}$) of the simulated population for all genotyping designs proposed.

^aThe averages are expressed over five simulation replicates; Red line = True values for simulated data were $Pvar_{1stQTL}$ (%) = 7.54 ± 2.77 ; RND = at random; T_{YD} = top yield deviation values; EX_{YD} = extreme yield deviation values; T_{YDC} = top yield deviation values inside communities; EX_{YDC} = extreme yield deviation values inside communities.

Figure 4.4 Average^a (and respective standard deviation) for the number of N_{topQTL} s identified by a top marker window distant no more than 1 Mb from a N_{topQTL} ($N_{trueQTL}$) of the simulated population for all genotyping designs proposed.

^aThe averages are expressed over five simulation replicates; Red line = True values for simulated data were $N_{topQTL} = 13.65 \pm 3.12$; RND = at random; T_{YD} = top yield deviation values; EX_{YD} = extreme yield deviation values; T_{YDC} = top yield deviation values inside communities; EX_{YDC} = extreme yield deviation values inside communities; N_{topQTL} = estimated number of true QTLs explaining 1 % or more of the genetic variance.

4.4 DISCUSSION

In this study we aimed to investigate the impacts of including phenotypes and genotypes from cows chosen by different methods over the performance of GWAS for detecting QTLs for a sex-limited complex trait. For such task we simulated a small dairy cattle population and phenotypic data for a lowly heritable trait (0.15), which is a common magnitude of genetic variance for quantitative traits of interest (BUCH et al., 2011; TENGHE et al., 2014; VUKASINOVIC et al., 2017). Although promising results were observed, it needs to be cleared that the performance of the proposed genotyping strategies may be restricted to the proposed situation. Thus, further investigations considering real dairy cattle populations would help to validate such methods. It is important to emphasize that due to the linear aspect for the population structure obtained from simulations, results obtained in real dairy cattle may differ due to distinct grades of complexity present in pedigreed livestock.

As expected, the variance explained by the top marker windows and by the window explaining the highest amount of genetic variance were higher for Bayes R and WssGBLUPw3 models. This happens because of SNP effect shrinkage imposed by the SNP weighting strategy applied in the case of WssGBLUPw3 and for the very restrictive prior set for the number of SNPs with no effect ($\pi_1 = 0.999$) considered for the Bayes R model. The variance explained by top markers was similar between the GBLUP and WssGBLUPw1 models, which indicates that the sole inclusion of non-genotyped cows was not capable of improving the ability to capture the real genetic variance explained by genomic regions in GWAS analyzes. Similar behavior was previously observed by Melo

et al. (2016) when investigating the inclusion of information from non-genotyped animals in GWAS analyzes.

The increase in the number of genotypes from cows have improved the ability to detect QLTs in almost every model and scenario proposed, except for top-yield deviation strategies, for which differences in trueQTL between 1,000 and 5,000 for T_{YD} were minimal. For such designs, only a slight superiority was observed for trueQTL when genotypes from cows were included considering a community-based design (T_{YDC}) when compared to its traditional version (T_{YD}). This information is of major concern when considering that, in practical situations, the most common genotyping strategies usually favor the high yield cows (PRYCE AND DAETWYLER, 2012). When genotypes of cows included in GWAS analyzes are obtained from such strategies, information value is very limited even with a large increase (from 1,000 to 5,000) in the number of individuals.

The inclusion of phenotyped and non-genotyped cows in the dataset used for analysis (by using the single-step method) have slightly overcome results of QTL detection from GBLUP in almost all scenarios studied. This result suggests that the inclusion of information from pedigreed cows may improve the ability to detect QTLs in GWAS. Melo et al. (2016) have indicated that the inclusion of phenotypic information from non-genotyped animals provided small improvement in the detection of QTLs in GWAS of low heritability complex traits. However, differences observed by the later authors was less pronounced than the observed in the present study. The number of QTL set for the present study was much lower (750) than proposed in Melo et al. (2016; 7,000), and this less polygenic aspect of the trait may have some influence in the ability to detect the most influential QTL by the models, since the presence of more QTLs with lower effects may

result in the inclusion of more noise. There is also the possibility that the population structure (and thus, the grade of linkage disequilibrium) obtained from different simulation setups could impair the similarity of the results.

Results observed for extreme-yield cow genotyping strategies (EX_{YD} and EX_{YDC}) were promising for improving the ability to detect QTL in fine-mapping studies using SNP data, specially when under a community-based approach (EX_{YDC} , Figure 4.3). The superiority of the traditional two-tailed extreme phenotype selective genotyping has been observed by Bovenhuis and Spelman (2000). It is important, though, to emphasize that such designs although feasible in a scientific investigation scope, may be extremely difficult to implement in an animal breeding program environment, since convincing breeders to genotype low yield cows can become a challenge in most situations (BOUQUET AND JUGA, 2013, PRYCE AND DAETWYLER, 2012).

The observed results suggest that considering population structure, via pedigree graph community detection, for including genotyped cows in GWAS analyzes may be an effective strategy for improving the ability to detect QTLs influencing quantitative traits in bovine populations. This information may support further investigations of influential genomic regions for traits in which phenotypic information is expensive or difficult to establish routine recording, such as hormone profiles (TENGHE et al., 2014), health disorders (BUCH et al., 2011; VUKASINOVIC et al., 2017) and feed efficiency (HARDIE et al., 2017). It also has applications when the GWAS analysis is intended in more commonly recorded traits, but for small dairy cattle populations or when animals consist of a subset from a bigger population (JIANG et al., 2010), such as Holstein populations in countries where animal breeding programs were recently implemented.

4.5 CONCLUSION

The inclusion of genotypic information from cows sampled from distinct communities in the population improves the ability to detect QTLs in genome-wide association studies for low heritability sex-limited traits. Genotyping and including cows by extreme-yield designs can consistently improve QTL detection ability in GWAS analyzes. Considering phenotypes from non-genotyped cows via weighted single-step method may yield better results in QTL mapping studies.

4.6 REFERENCES

AGUILAR I.; MISZTAL I.; JOHNSON D.L.; LEGARRA A., TSURUTA S., LAWLOR T.J.

Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal of Dairy Science**; 93:743–52, 2010.

ANDONOV, S.; LOURENCO, D. A. L.; FRAGOMENI, B. O.; MASUDA, Y., POCRINIC, I.,

TSURUTA, S., & MISZTAL, I. Accuracy of breeding values in small genotyped populations using different sources of external information – A simulation study. **Journal of Dairy Science**, 100:1-7, 2017.

BLASCO, A.; & TORO, M.A. A short critical history of the application of genomics to animal breeding. **Livestock Science**, 2014.

BLONDEL, V.D.; GUILLAUME, J.L.; LAMBIOTTE, R.; & LEFEBVRE, E. Fast unfolding of communities in large networks, 1–12. **Journal of Statistical Mechanics: Theory**

and Experiment, 10, 10008, 2008.

BOUQUET, A., & JUGA, J. Integrating genomic selection into dairy cattle breeding programs: a review. **Animal**, 7:5, 705-713, 2013.

BOVENHUIS, H., & SPELMAN, R. J. Selective Genotyping to Detect Quantitative Trait Loci for Multiple Traits in Outbred Populations. **Journal of Dairy Science**, 83(1), 173–180, 2000.

CHRISTENSEN O.F. & LUND MS. Genomic prediction when some animals are not genotyped. **Genetics Selection Evolution**. 42:2, 2010.

ERBE M, HAYES BJ, MATUKUMALLI LK, GOSWAMI S, BOWMAN PJ. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. **Journal of Dairy Science**, 95: 4114–4129, 2012.

FERNANDO RL. & GARRICK D. Bayesian methods applied to GWAS. In: GONDRO C, VAN DER WERF J, HAYES B. Genome-wide association studies and genomic predictions. Springer New York, Heidelberg, Dordrecht, London: Humana Press. 2013, chap.10, p. 237–274.

GALLAIS, A., MOREAU, L., & CHARCOSSET, A. Detection of marker–QTL associations by studying change in marker frequencies with selection. **Theoretical and Applied Genetics**, 114(4), 669–681, 2006.

- GAO H., MADSEN P., NIELSEN U.S., AAMAND G.P., SU G., BYSKOV K., & JENSEN J. Including different groups of genotyped females for genomic prediction in a Nordic Jersey population. **Journal of Dairy Science**, 98(12):9051-9, 2015.
- GARRICK, D., DEKKERS, J. & FERNANDO, R. The evolution of methodologies for genomic prediction. **Livestock Science**, [Volume 166](#), August 2014, Pages 10-18, 2014.
- GIANOLA D, DE LOS CAMPOS G, HILL WG, MANFREDI E, FERNANDO R. Additive Genetic Variability and the Bayesian Alphabet. **Genetics**, 183:347–63, 2009
- GIRVAN, M., & NEWMAN, M. E. J. Community structure in social and biological networks, 99(12), 2002.
- HABIER, D., FERNANDO, R., KIZILKAYA, K & GARRICK, D.J. Extension of the Bayesian alphabet for genomic selection. **BMC Bioinformatics**, 12:186, 2011.
- HAYES B.J. & GODDARD M.E. The distribution of the effects of genes affecting quantitative traits in livestock. **Genetics Selection Evolution**, 33:209–29, 2001.
- HAYES, B.J. AND GODDARD, M. Genome-wide association and genomic selection in animal breeding. **Genome**, 2010, 53:876-883, 2010.
- HARDIE, L.C., VANDEHAAR, M.J., TEMPELMAN, R.J., WEIGEL, K.A., ARMENTANO, L.E., WIGGANS, G.R., VEERKAMP, R.F., DE HAAS, Y., COFFEY, M.P., CONNOR, E.E., HANIGAN, M.D., STAPLES, C., WANG, Z., DEKKERS, J.C.M. & SPURLOCK, D.M. The genetic and biological basis of feed efficiency in mid-lactation Holstein dairy cows. **Journal of Dairy Science**, 100:9061–9075, 2017

- HILL, W.G. & ROBERTSON A. Linkage disequilibrium in finite populations. **Theoretical and Applied Genetics**, 38(6):226-31, 1968.
- HIRSCHHORN, J.N.; & DALY, M.J. Genome-wide association studies for common diseases and complex traits. **Nature Reviews Genetics**, 6, pages 95–108, 2005.
- HUANG, X. & HAN, B. Natural Variations and Genome-Wide Association Studies in Crop Plants. **Annual Review of Plant Biology**, 65:1, 531-551, 2014.
- JENKO, J., WIGGANS, G. R., COOPAR, T. A., EAGLEN, S. A. E., LUFF, W. G. L., BICHARD, M., PONG-WONG, R., & WOOLIAMs J. A. **Journal of Dairy Science**, 100(1):439-452, 2017.
- JIANG, L., LIU, J., SUN, D., MA, P., DING, X., YU, Y., & ZHANG, Q. Genome Wide Association Studies for Milk Production Traits in Chinese Holstein Population. **PLOS ONE**, 5(10): e13661, 2010.
- JIMÉNEZ-MONTERO, J. A., GONZÁLEZ-RECIO O., & ALENDA R. Genotyping strategies for genomic selection in small dairy cattle populations. **Animal**, 6, 1216–1224, 2012.
- MISZTAL, I., TSURUTA, S., LOURENCO, D., AGUILLAR, I., LEGARRA, A. & VITEZICA, Z. Manual for BLUPF90 family programs. Accessed July, 20, 2017. http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all2.pdf
- MOSER, G., LEE, H.S., HAYES, B.J., GODDARD M.E., WRAY, N.R. & VISSCHER P.M. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. **PLoS Genetics**, 11: e1004969, 2015.

- MEUWISSEN, T.H.E., HAYES, B.J. & GODDARD, M.E. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. **GENETICS** *April 1, 2001 vol. 157 no. 4 1819-1829*, 2001.
- MEUWISSEN, T., HAYES, B., & GODDARD, M. Genomic selection: A paradigm shift in animal breeding. **Animal Frontiers**, 6(1), 6–14, 2016.
- PLIESCHKE, L., EDEL, C., PIMENTEL, E. C. G., EMMERLING, R., BENNEWITZ, J., & GOTZ K. U. Systematic genotyping of groups of cows to improve genomic estimated breeding values of selection candidates. **Genetics Selection Evolution**, 48:73, 2016.
- PRYCE, J.E., & DAETWYLER, H. D. Designing dairy cattle breeding schemes under genomic selection: a review of international research. **Animal Production Science**, 52, 107-114, 2012.
- SARGOLZAEI M, & SCHENKEL FS. QMSim: User's Guide. Centre for Genetic Improvement of Livestock. Department of Animal and Poultry Science. Guelph, Canada: University of Guelph, 2013.
- TIEZZI F.; PARKER-GADDIS K.L.; COLE J.B.; CLAY JS, MALTECCA C. A Genome-Wide Association Study for Clinical Mastitis in First Parity US Holstein Cows Using Single-Step Approach and Genomic Matrix Re-Weighting Procedure. **PLOS ONE** 10(2): e0114919, 2015.
- VAN DEN BERG I, FRITZ S, BOICHARD D. QTL fine mapping with Bayes C (π): a simulation study. **Genetic Selection Evolution**, 45:19, 2013.

ZENG, J., DE VLAMING, R ; WU, Y., ROBINSON, M.R., LLOYD-JONES, L.R., YENGO, L., YAP, C.X., XUE, A.L., SIDORENKO, J., MCRAE, A.F., POWELL, J.E., MONTGOMERY, G.W., METSPALU, A., ESKO, T., GIBSON, G., WRAY, N.R., VISSCHER, P.M. & YANG, J. Signatures of negative selection in the genetic architecture of human complex traits. **Nature Genetics**, 50, 5, Page 746, 2018.

WANG H., MISZTAL I., AGUILAR I., LEGARRA A. & MUIR WM. Genome-wide association mapping including phenotypes from relatives without genotypes. **Genetic Research**, 94:73–83, 2012.

WANG H, MISZTAL I, AGUILAR I, LEGARRA A, FERNANDO RL, VITEZICA Z, et al. Genome-wide association mapping including phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body weight in broiler chickens. **Frontiers in Genetics**, 5:134, 2014.

CHAPTER 5. IMPACTS OF SELECTIVE GENOTYPING OVER RESULTS OF GENOMIC EVALUATIONS UNDER THE ASSUMPTION OF PREFERENTIAL TREATMENT OF COWS IN SMALL DAIRY CATTLE POPULATIONS.

Bruno da Costa Perez¹, Ricardo Vieira Ventura², Gerson Antonio Oliveira Junior³,
Rafael Guimaraes Vizona², Julio Cesar de Carvalho Balieiro²

¹ *Department of Animal Science, College of Animal Science and Food Engineering, University of São Paulo (FZEA/USP), Pirassununga, 13630-000 São Paulo, Brazil*

² *College of Veterinary Medicine and Animal Science, University of São Paulo (FMVZ/USP), Pirassununga, 13630-000 São Paulo, Brazil*

³ *Centre for Genetic Improvement of Livestock, Department of Animal Biosciences, University of Guelph, Guelph, ON, Canada.*

ABSTRACT Preferential treatment is defined all management practices that increase production from specific groups of animals in detriment of others. In small dairy cattle populations, where a reduced number of progeny-tested bulls is available, the impact of preferential treatment for cows is stronger than when a higher number of bulls with accurate EBV is available. The objective of the present study was to investigate the impact of preferential treatment of cows and 2) the effect of different cow genotyping strategies over bias and accuracy of genomic evaluations for a small dairy cattle population. To perform the proposed analyzes, a small dairy cattle population was simulated, considering a lowly heritable sex-limited trait. Based on three cow genotyping designs (random, top-yield, extreme-yield cows), different numbers (1,000 and 5,000) of cows were sampled and included in the reference population containing progeny tested bulls. Different percentages of cows from the contemporary generations were considered

to be preferentially treated before estimation of their (pedigree-based) breeding values, consisting in scenarios with 0%, 10%, 25% and 50% of cows with an artificially increased phenotypic observation. Impacts of preferential treatment of cows over accuracy and bias of genomic evaluations was detected. The magnitude of impacts was proportional to the proportion of cows preferentially treated and included in the reference population. Best results were obtained for extreme-yield cow genotyping strategies. Drastic decrease in accuracy and increase in bias was observed when genotyping only the best cows. Results observed indicate the importance of quantifying the frequency and magnitude of preferential treatment of cows in animal breeding programs of dairy cattle populations of reduced size.

Keywords: accuracy, bias, genotyping designs, genomic selection, phenotyping

5.1 INTRODUCTION

The use of genome-wide marker information has become a standard procedure for estimating breeding values in dairy cattle breeding programs (GODDARD AND HAYES, 2009; KÖNIG AND SWALVE, 2009; LOURENÇO et al., 2014). The inclusion of such source of information in genetic analyzes may provide accurate estimates early in the animal's life (VAN DER WERF., 2010; DE ROOS et al., 2011), reducing generation interval and increasing genetic improvement rates (SCHAEFFER, 2006; GARCIA-RUIZ et al., 2016). A diversity of effects has been reported to influence accuracy and bias in genomic evaluations and, from these, the size (LUND et al., 2011, JIMÉNEZ-MONTERO et al., 2012) and structure (PSZCZOLA et al., 2012; RINCENT et al., 2017) of the

reference population; and connectiveness between the reference population and the selection candidates (CLARK et al., 2012, WU et al., 2015, VENTURA et al., 2016) are the main aspects considered in literature for the success of genomic selection. However, the quality of phenotypic records must be considered as a relevant factor (GODDARD, 2009; BOUQUET AND JUGA, 2016) influencing results.

As genomic evaluations often consider deregressed estimated breeding values (dEBV) as a pseudo-phenotype, progeny-tested bulls were the main candidates to be genotyped and integrate the reference population (PATRY AND DUCROQ, 2011; PRYCE AND DAETWYLLER, 2011; WIGGANS et al., 2011) for having more reliable information than cows. However, in small dairy cattle populations, or when selection is intended for an expensive-to-measure sex limited trait, phenotypic data is limited and, therefore, even estimated breeding values for bulls are low in reliability (BOUQUET AND JUGA, 2013; CALUS, 2016). To overcome that, in many situations phenotypic and genotypic information from cows have been included in the reference population (PRYCE AND DAETWYLLER, 2012; DING et al., 2013; THOMASEN et al., 2014; GAO et al., 2015; PLIESCHKE et al., 2016, JENKO et al., 2017; UEMOTO et al., 2017).

Preferential treatment (PT) can be defined as management practices that increase production (WIGGANS et al., 2011; DASSONEVILLE et al., 2012) from specific groups of animals in detriment of others. It is of common sense that PT is more often performed in elite herds and for bull-dams (KUHN, 1994; DEHNAVI et al., 2018), however, the real extension of the magnitude and prevalence of PT in elite and commercial herds is not precisely known. The inclusion of such inflated phenotypic records has been reported to increase bias in genetic evaluations (BURNSIDE AND MEYER, 1988; FERRIS, 1991;

KUHN, 1994; 1995; TSURUTA, 2000). In this manner, as traditional pedigree-based methods are needed to generate the dEBV used for genomic evaluations, PT for cows will not only influence the estimation of their own genomic EBV, but also the estimation for bulls (DASSONEVILLE et al., 2012; DEHNAVI et al., 2018).

Bias arising from PT was first reported by comparing the parent average breeding value of specific bulls with the phenotypic performance of their daughters (VAN VLECK, 1987). Tsuruta et al. 2000 have investigated the inclusion of Holstein cows treated with Bovine Somatotropin (bST) in genetic evaluations for milk production, reporting bias of small magnitude. However, the number of cows treated with bST considered in the study represented only 5% (on average) of the total number of phenotypic records. It is plausible to assume that the magnitude of PT can be much higher in other breeds. Kuhn et al., 1994 indicated, via simulation, that the magnitude of bias in genetic evaluations will depend on the amount of relative information that had been inflated by PT. There are not many studies in literature accounting for impacts of PT for cows in results of genomic analyzes and, therefore, of genomic selection programs. Dassoneville et al. (2012) demonstrated that including own records from cows of elite herds results in biased estimation of genomic breeding values. Dehnavi et al. (2018) detected, via simulation, that PT performed over top-merit cows would not only increase bias, but decrease accuracy of genomic evaluations. The authors also observed that the impacts of PT would decrease as the number of bulls increased in the reference population.

In small dairy cattle populations, where the number of progeny-tested bulls is limited, the impact of PT for cows is probably stronger than when a higher number of bulls with accurate EBV is available to get included in the reference population. Also, to the

best of our knowledge, there are no studies in literature investigating how different cow genotyping designs would impact the effects of PT. Therefore, the objective of the present study was to investigate the impact of PT of cows and 2) the effect of different cow genotyping strategies over bias and accuracy of genomic evaluations for a small dairy cattle population.

5.2 MATERIAL AND METHODS

5.2.1 Simulated population

The open access software QMSim (SARGOLZAEI AND SCHENKEL, 2009) was considered to run the simulations for the present study, considering five replicates. We simulated a relatively small dairy cattle population, similarly to previous simulation studies (JIMÉNEZ MONTERO et al., 2012; PLIESCHKE et al., 2016, ANDONOV et al., 2016). To accomplish that, a historical population consisting of 1,500 unrelated individuals with balanced sex ratio was created. Those animals were randomly mated for 1,000 generations. A bottleneck was introduced from generation 1,001, reducing the historical population size from 1500 to 400 animals in 20 generations. From the historical population, a second population, here referred as expansion population, was created in which founders were 100 males and 300 females randomly sampled from the last generation of historical population (1020). This expansion population was ran for 60 generations in which animals were also randomly mated, considering a growth rate of 1.2x for both sexes by generation. From the expansion population, a recent population was created by sampling 400 males and 20,000 females from the last generation of the expansion population. The recent population was run for 20 overlapping generations. In

each generation, 10,000 male and 10,000 female offspring were simulated. At the same time, 30% of the dams and 50% of the sires were replaced. Genealogical information was recorded for all individuals from the recent population. Breeding animals were selected based on their BLUP-based breeding values (here we used EBV) and culled by age. Cows from generations 16 to 20 and sires from generations 15 to 18 had genotypic data recorded. Bulls with more than 40 daughters were considered as progeny tested individuals ($n = 764$) and assigned to the reference set. Animals from generation 20 were considered as validation population.

5.2.2 Genome simulation

The simulated genome had 29 pairs of autosome chromosomes with total length of 2333 cM, in which each chromosome's length mimicked the bovine genome without sex chromosomes. The intent was to generate a more plausible scenario when considering real distances between markers and QTL loci. A total of 50,000 bi-allelic SNP markers were evenly distributed throughout the genome, in which the number of SNP per chromosome was proportional to its size. The additive genetic component was determined by 750 randomly distributed QTL along the genome, which effects were sampled from a gamma distribution with a shape parameter of 0.4 (HAYES AND GODDARD, 2001). The number of QTL (750) was defined aiming to assure a polygenic nature for the simulated trait. The mutation rate was identical (2.5×10^{-5}) for both markers and QTL (JIMENEZ-MONTERO et al., 2012). The rate of missing marker genotypes was 0.01 and the rate of marker genotyping error was 0.005.

5.2.3 Phenotypic observations and preferential treatment simulation

A single sex-limited trait with heritability of 0.15 and phenotypic variance of 1.0 was simulated. Preferential treatment was simulated in cows from generation 19 of the recent population by adding a residual term to their phenotypic observation. The residual term was sampled from a random Gaussian distribution with mean equal to 30% of the phenotypic mean for generation 19 and a standard deviation of one fifth of this mean. This artificial increase in the phenotypic performance was added to 10, 25 and 50% randomly chosen cows, which represented approximately 1,000; 2,500 and 5,000 individuals, respectively, from the total number of cows in generation 19.

Pedigree-based EBV were estimated for all individuals from generations 12 to 20 of the recent population, including the preferentially treated cows for each extension of PT (10%, 25% and 50%). The analysis was carried out under a restricted maximum likelihood (REML) approach using an animal model. Implementation was performed in the BLUPF90 family software (MIZSTAL et al., 2014). The model was as follows:

$$\mathbf{y} = \mu \mathbf{1} + \mathbf{Z}\mathbf{a} + \mathbf{e}, \quad [1]$$

where $\mathbf{y}$ is the vector of phenotypes for cows; μ is the general mean (no fixed effects were simulated); $\mathbf{1}$ is a vector of ones; $\mathbf{Z}$ is an incidence matrix allocating records to breeding values; $\mathbf{a}$ is a vector of EBV with $var(\mathbf{a}) = A\sigma_a^2$, where $\mathbf{a} \sim N(0, A\sigma_a^2)$, in which σ_a^2 is the additive genetic variance and A (HENDERSON, 1975) is the pedigree-based relationship matrix containing the expectancy of genetic connection between all pedigreed individuals, and $\mathbf{e}$ is the random residual effect, where $\mathbf{e} \sim N(0, I\sigma_e^2)$ in which σ_e^2 is the residual variance.

The EBV accuracies were calculated as $r = \sqrt{1 - PEV/\sigma_a^2}$, where PEV is the prediction error variance obtained from the estimate's standard error. The EBV and respective reliabilities were then used to obtain de-regressed proofs (dEBV) for all individuals following (GARRICK et al., 2009), which were then used as pseudo-phenotypes in the genomic evaluation. The pseudo-phenotypic information included as dependent variable in the model was weighted by its reliability. As different dEBV accuracies for individuals in the reference population may influence performance of genomic evaluations for distinct scenarios tested, the mean (and standard deviation) for accuracy of cows included in the reference population was reported for each scenario to check if results are being compared in equality.

The true breeding value (TBV) for each animal was calculated as the sum of the QTL additive effects, as $TBV_k = \sum_{j=1}^{qtl} \beta_j \cdot Q_{kj}$, where qtl is the total number of QTL, β_j is the additive effect of QTL genotype (j) and Q_{kj} is the QTL genotype at locus j , coded as 0,1 or 2, representing the number of copies of the QTL allele that an animal (k) carries.

5.2.4 Selective genotyping designs

First, scenarios not including PT (PT_0) containing only progeny-tested bulls (Bull_PT0), only cows (Cow_PT0) and progeny-tested bulls combined with cows (BxC_PT0) in the reference population were proposed as a base result to be compared with scenarios including PT. All base scenarios were replicated considering the different extensions of PT proposed (PT10, PT25 and PT50). As the inclusion of cows in the reference population under the assumption of PT was one of the main objectives of this study, different numbers of cows were tested. Cows from generation 19 of the recent

population represented a contemporary overlapping active population of approximately 10,000 individuals. From them, 1,000 and 5,000 were sampled as female candidates to be genotyped and included in the reference population. Therefore, each scenario (with exception of the bull-based reference populations) were replicated considering these two quantities of females, sampled according to three distinct strategies: 1) δ females randomly selected (RND), 2) δ females with the highest dEBV (TOP) and 3) $\delta/2$ females with the highest and $\delta/2$ females with the lowest dEBV (EXT), where δ was 1,000 or 5,000, depending on the scenario. A description of all proposed scenarios is presented in Table 5.1.

Table 5.1 Summary of genotyping scenarios and number of genotypes for bulls and cows in the reference population for each method proposed.

Design	Summary of reference population
Bull	764 progeny-tested bulls
RND _{COW}	1000 or 5000 randomly sampled cows
TOP _{COW}	1000 or 5000 cows with the highest values of dEBV
EXT _{COW}	1000 or 5000 cows with extreme values of dEBV
RND _{BxC}	764 progeny-tested bulls + 1000 or 5000 randomly sampled cows
TOP _{BxC}	764 progeny-tested bulls + 1000 or 5000 cows with the highest values of dEBV
EXT _{BxC}	764 progeny-tested bulls + 1000 or 5000 cows with extreme values of dEBV

5.2.5 Genotype quality control and genomic analysis

Prior to all genomic evaluations, genotype data quality control was. For all proposed scenarios, SNPs with minor allele frequency lower than 0.02 and call rate lower than 0.95, as well as samples with call rate lower than 0.90 were excluded from analysis. After quality control, around 43,500 SNPs remained, with slight variations depending on the reference population considered, the simulation replicate and the scenario proposed.

The genomic evaluation was carried out using a best linear unbiased prediction (GBLUP) (VANRADEN, 2008). The general equation is as follows:

$$\mathbf{y} = \mu \mathbf{1} + \mathbf{Z}\mathbf{g} + \mathbf{e}, \quad [2]$$

where $\mathbf{y}$ is the vector of the dEBV for the genotyped individuals in the reference population; μ is the general mean (no fixed effects were simulated); $\mathbf{1}$ is a vector of ones; $\mathbf{Z}$ is an incidence matrix allocating records to breeding values; $\mathbf{g}$ is a vector of GEBV with $\text{var}(\mathbf{g}) = \mathbf{G}\sigma_g^2$, in which σ_g^2 is the additive genetic variance; and $\mathbf{G}$ is the realized genomic relationship matrix created using the method described in VanRaden (2008). That is, $\mathbf{G} = (\mathbf{M} - \mathbf{P})(\mathbf{M} - \mathbf{P})' / \sum_{i=1}^m 2p_i(1 - p_i)$, where $\mathbf{M}$ is an $n \times m$ matrix (number of animals $\times$ number of loci) with SNP coded 0, 1 and 2 for genotypes A1A1, A1A2 and A2A2, respectively. $\mathbf{P}$ is an $n \times m$ matrix containing twice the allele frequencies expressed as $P_i = 2p_i$ where p_i is the allele frequency of the homozygous genotype coded with 2 for all genotyped individuals at locus i . It is assumed that $\mathbf{e} \sim N(0, \sigma_e^2)$, where σ_e^2 is the residual variance. All analyzes were conducted under a REML approach, using BLUPF90 family software (MIZSTAL et al., 2014).

The direct genomic value (DGV) was calculated for all animals in the validation population as:

$$DGV = X'\hat{\beta}, \quad [3]$$

where X is the matrix of marker genotypes for each animal in the validation set and $\hat{\beta}$ is the vector of marker effects estimated based using information on the respective reference population.

5.2.6 Assessing accuracy and bias

As TBV were known for all animals in the simulation, the Pearson product-moment correlation coefficient (ρ) between the TBV and the DGV of animals from the validation population were used to assess the accuracy and prediction error in genomic evaluations for each selective genotyping strategy proposed. The slope (b) of the regression of TBV on DGV for the validation individuals, that is, $b = cov(TBV, DRP) / \sigma_{TBV}^2$ (OLSON et al., 2011). Means and respective standard deviations (from the five simulation replicates) were calculated for each proposed parameter in the strategies proposed.

5.3 RESULTS

5.3.1 Accuracy of de-regressed estimated breeding values for genotyped cows

As the reliability of phenotypes (or pseudo-phenotypes) for individuals in the reference population may affect results of genomic evaluations, the mean accuracy of dEBV for cows included in the reference was assessed and is shown in Figure 5.1. There was observed no major variation of this parameter when comparing animals included by the various scenarios proposed, indicating that results observed were not influenced by different magnitude of accuracies for reference animals from distinct genotyping strategies.

Figure 5.1 Average (and respective standard deviation) obtained in five replicates for the accuracy of de-regressed estimated breeding values of cows included in the reference population for genotyping designs and scenarios proposed.

RND_1000 and RND_5000 = 1000 and 5000 cows included randomly; TOP_1000 and TOP_5000 = 1000 and 5000 cows with the highest de-regressed EBV; EXT_1000 and EXT_5000 = 1000 and 5000 cows with extreme de-regressed EBV. Scenarios are valid for cow-only and bull + cows reference populations.

5.3.2 Performance of genomic evaluations in the absence of preferential treatment

In Figure 5.2 there are shown results for the Pearson correlation coefficient between TBV and DGV for validation individuals considering different reference population structure scenarios proposed in the present study considering the presence and absence of PT.

The accuracy of genomic evaluations observed for the bull-only reference population in the absence of PT was 0.296 ± 0.03 (Bull_PT0). Results observed for cow-only reference populations considering no PT varied from 0.075 ± 0.02 to 0.521 ± 0.06 , depending on the number of cows included in the reference population and the

genotyping strategy considered. Also, accuracy from genomic evaluations obtained from reference populations combining bull and cow information yielded the best results when compared to bull-only and respective cow-only scenarios.

When considering cow-only reference populations, results varied from lower to higher than the accuracy observed for bull-only reference population (Bull_PT0), depending on the cow genotyping strategy and the number of cows included. The highest negative and positive differences between accuracies obtained from cow-only and bull-only reference populations was observed for TOP_{COW_1000_PT0} (-0.221) and EXT_{COW_5000_PT0} (+0.249).

The increase in accuracy of genomic evaluations when adding cows to the reference population was proportionally higher for cow-only scenarios than observed for combined (bull + cows) scenarios. Differences on the observed accuracy for the inclusion of 1,000 and 5,000 cows in the reference population varied from +0.084 to +0.303, respectively, for TOP_{BXC} and TOP_{COW} scenarios.

Figure 5.2 Average (and respective standard deviations)¹ accuracy between true breeding value and direct genomic value for animals in the validation population for each preferential treatment scenario² and genotyping strategy³ tested (above); and the difference in accuracy between all preferential treatment scenarios when compared to the absence of preferential treatment (below). PT_0 = no preferential treatment for cows; PT_10 = 10% of reference cows preferentially treated; PT_25 = 25% of reference cows preferentially treated; PT_50 = 50% of reference cows preferentially treated.

5.3.3 Accuracy and bias of genomic evaluations in the presence of preferential treatment

The inclusion of PT has negatively influenced the accuracy obtained from genomic predictions in all scenarios proposed proportionally to the extension of PT considered. In Figure 5.2 it is also shown the difference in accuracy obtained between scenarios including PT (PT10, PT25 and PT50) and the scenario where PT was absent (PT0). The highest decrease in accuracy caused by the increase in the extent of PT in reference cows was observed for scenarios considering the inclusion of cows based on high dEBV values (TOP_{COW} and TOP_{BxC}). At the same time, the lowest decrease in accuracy for the increase in the extent of PT was found for genotyping of cows based on extreme dEBV values (EXT_{COW} and EXT_{BxC}). Differences in accuracy obtained for cow-only reference population scenarios were higher than the observed for combined reference populations scenario.

Figure 5.3 shows the amount of bias in genomic predictions for all scenarios proposed as the distance between the regression coefficient obtained and the value 1, which would indicate the result in which there is no inflation or deflation of EBV. The bull-only reference population have resulted in values of b lower than 1, despite the amount of PT included, being respectively 0.788 ± 0.08 , 0.764 ± 0.07 , 0.527 ± 0.05 and 0.442 ± 0.05 , respectively for PT0, PT10, PT25 and PT50.

Figure 5.3 Average (and respective standard deviations)¹ bias (linear regression coefficient) between true breeding value and direct genomic value for animals in the validation population for each preferential treatment scenario² and genotyping strategy³ tested.

The increase in the number of preferentially treated cows have caused a decrease in values of b when compared to PT0 scenarios, independent of the structure or genotyping strategy adopted. Overall, cow genotyping strategies based on high dEBV values have yielded mainly positive b values ($b > 1$), except for TOP_{cow}_1000_PT25 and TOP_{cow}_1000_PT50. Results obtained for b when considering cow genotyping strategies based on extreme dEBV values were all negative ($b < 1$). The highest amount of deflation was detected for TOP_{cow}_1000_PT0 ($b = +1.852 \pm 0.21$) while the highest amount of deflation was observed for EXT_{cow}_5000_PT50 ($b = -0.589 \pm 0.05$). When randomly sampling cows to be included in either cow-only or combined (bulls and cows) reference populations yielded inconsistent results, with no clear decrease in b values in contrary to the observed for other genotyping designs, except for the extreme PT scenario (PT50).

5.4 DISCUSSION

In the present study, we hypothesized that since the number of progeny-tested bulls is limited in dairy cattle breeds of reduced size, genotypic/phenotypic information from cows is of greater relevance for the results of genomic evaluations. Therefore, not only the structure of the set of animals included in the reference population, but also the quality of their phenotypic observations will have pronounced impact on the success of genomic selection in small dairy cattle breeds. In the presence of PT, which can be defined as a deviation of the phenotypic observation from its real (unknown) value and, therefore, an imputed error, distinct selective genotyping strategies may behave differently on accuracy and bias yielded.

As expected, due to the limited number of progeny-tested bulls present in a reduced number population environment, accuracies obtained from bull-only reference populations were lower than observed by other authors in larger populations (VANRADEN et al., 2009; LUND et al., 2011; GAO et al., 2015; DEHNAVI et al., 2018). Jiménez-Montero et al. (2012) have found accuracies of genomic evaluations varying from 0.48 to 0.55 when considering a bull-only reference population ($N = 996$), respectively, for low and medium heritability values. However, in the present study individuals considered as progeny-tested bulls were around 23% lower in numbers ($N = 764$) than considered in the mentioned study. Results here reported for bull-based reference populations explicit the conditions described in literature for the implementation of genomic selection programs in dairy breeds of reduced size, in which genotypic and phenotypic information from progeny-tested bulls is often insufficient for a robust

estimation of breeding values (PRYCE AND DAETWYLLER, 2012; MCHUGH et al., 2011; BOUQUET AND JUGA; 2013).

Reference populations containing only cows were reported in the literature as a viable strategy to overcome obstacles imposed by reduced size populations, when phenotype recording is difficult or too expensive to be implemented in a wide scale (PRYCE AND DAETWYLLER et al., 2012; CALUS et al., 2013; BOUQUET AND JUGA, 2016; UEMOTO et al., 2017). The results of accuracy found for the scenario with no PT in the present study corroborate the idea that cow-only reference populations may yield reasonable accuracy in genomic evaluations, especially when selective genotyping strategies are feasible in routine cow genotyping (JIMÉNEZ-MONTERO et al., 2012). However, in the presence of PT, the cow-only reference populations have suffered the highest decrease in accuracy when compared to bull-only or combined (cows and bulls) reference populations. This is an indication that the decision on establishing a cow-only reference population must be conditional to an estimation of the magnitude and extension of PT present in the evaluated individuals for the given trait.

The inclusion of females in the reference population have been reported to increase accuracy of genomic evaluations (MCHUGH et al., 2011; JIMÉNEZ-MONTERO et al., 2012; THOMASEN et al., 2014; GAO et al., 2015; KOIVULA et al., 2016; PLIESCHKE et al., 2016; JENKO et al., 2017; UEMOTO et al., 2017). Dehnavi et al. (2018) have observed higher improvements in accuracy of genomic evaluations when the same number of cows were included in reference populations containing lower number of bulls. Considering that the size of the bull-only reference population in the present study was much lower than the considered in the above-mentioned study, results on the

increase in accuracy by the inclusion of cows in a reference population of bulls were proportionally in line with its findings. Therefore, it is plausible to assume that the including information from cows in genomic evaluations may be of extreme value for dairy breeds of reduced size (THOMASEN et al., 2014; EDEL et al., 2016; JENKO et al., 2017).

When under the presence of PT, the observed impacts over accuracy and bias by the inclusion of cow genotypes/phenotypes in a reference population previously composed by progeny-tested bulls were reduced when compared to cow-only reference populations. Similar results were previously observed by Dehnavi et al. (2018) although in smaller magnitude probably because of the larger number of bulls included in the reference population on such study. As suggested by the authors, as the number of progeny-tested bulls in the reference population increase, the impacts of including information from cows will be proportionally reduced. This happens because of the higher reliabilities for bull's EBV when compared to EBV for cows. Therefore, when PT is practiced in small dairy cattle breeds, even with the reduced numbers of bulls having reliable pseudo-phenotypic information available it is of extreme value to consider genotypes/phenotypes from males in order to prevent major impacts of preferentially treated cow's phenotypes over the accuracy and bias of genomic predictions.

The inclusion of bias in genomic evaluations has been a concern since mass genotyping of cows became economically feasible (LUND et al., 2011; WIGGANS et al., 2011). In general, authors have suggested an inflation of the genomic EBV when information from cows are included in the reference population (DASSONEVILLE et al., 2014; DEHNAVI et al., 2018). McHugh et al. (2011) have found considerable decrease in the regression coefficient obtained from the linear regression of deregressed proofs on

GEBV of validation individuals for milk production, milk content, body conformation, milking speed and health traits when including cows in the reference population of Nordic Jersey bulls. Jiménez-Montero et al. (2012) have observed deflation of genomic EBV when cows included in the reference population were sampled from the top-yield individuals in the population. This result is in line to our findings for the TOP_{COW} and TOP_{BxC} scenarios, in which estimated b was predominantly higher than 1. As also observed by Dehnavi et al. (2018), results presented in this study indicate that randomly sampling cows from a preferentially treated contemporary population had no major effects on avoiding impacts of PT over bias and accuracy of genomic evaluations.

For computational purposes, only cows from the latest phenotyped population (generation 19) were submitted to preferential treatment in the present study. However, it is important to emphasize that in real situations the extension of PT is certainly not restricted to the contemporary generation (KUHN et al., 1994; TSURUTA et al., 2000). This implies that EBV from older cows and bulls have suffered from bias caused by PT practiced in older generations, in which impacts are probably cumulative for the younger animals in recent generations. Since notification on PT is often difficult to be implemented and, in many situations, unfeasible, its effects over accuracy and bias of genetic and genomic evaluations using real data are difficult to estimate precisely. Therefore, in such situations, simulation studies are of extreme value. Further studies would help to quantify possible long-term impacts of PT in both genetic and genomic evaluations.

5.5 CONCLUSION

Preferential treatment of cows influences both accuracy and bias of genomic evaluations and therefore, must be taken into account in animal breeding programs. Negative effects of preferential treatment are proportional to the number of cows included in the reference population. Selective cow genotyping strategies have remarkable impacts over the performance of genomic selection under a multi-step procedure in the presence of preferential treatment of cows. When preferential treatment is assumed to exist in a population, genotyping only the best cows will drastically impact the estimates of genomic breeding values.

5.6 REFERENCES

- ANDONOV, S., LOURENCO, D. A. L., FRAGOMENI, B. O., MASUDA, Y., POCRINIC, I., TSURUTA, S., & MISZTAL, I. Accuracy of breeding values in small genotyped populations using different sources of external information – A simulation study. **Journal of Dairy Science**, 100:1-7, 2017.
- BLONDEL, V. D., GUILLAUME, J. L., LAMBIOTTE, R., & LEFEBVRE, E. Fast unfolding of communities in large networks, 1–12. **Journal of Statistical Mechanics: Theory and Experiment**, 10, 10008, 2008.
- BOUQUET, A., & JUGA, J. Integrating genomic selection into dairy cattle breeding programs: a review. **Animal**, 7:5, 705-713, 2013.
- CALUS, M. P. L., De HAAS, Y., PSZCZOLA, M., & VEERKAMP, R. F. Predicted accuracy

- of and response to genomic selection for new traits in dairy cattle. **The Animal Consortium**, 7(2), 183–191, 2013
- CALUS, M. P. L. Editorial: Genomic selection with numerically small reference populations. **Animal**, 10:6, pp 1016-1017, 2016.
- CLARK, S. A., HICKEY, J.M., DAETWYLER, H.D., & VAN DER WERF, J.H.J. 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 schemes. **Genetics Selection Evolution**, 44-4, 2012.
- DASSONNEVILLE, R., BAUR, A., FRITZ, S., BOICHARD, D., & DUCROCQ, V. Inclusion of cow records on genomic evaluations and impact on bias due to preferential treatment. **Genetics Selection Evolution**, 44:40, 2014.
- DE ROOS APW, SCHROOTEN C, VEERKAMP RF, VAN ARENDONK JAM. Effects of genomic selection on genetic improvement, inbreeding, and merit of young versus proven bulls. **Journal of Dairy Science** 94, 1559–1567, 2011.
- DEHNAVI, E., MAHYARI, S. A., SCHENKEL, F. S., & SARGOLZAEI, M. The effect of using cow genomic information on accuracy and bias of genomic breeding values in a simulated Holstein dairy cattle population. *Journal of Dairy Science*, 101(6), 5166–5176, 2018.
- DING, X., ZHANG, Z., LI, X., WANG, S., WU, X., SUN, D., YU, Y., LIU, J., WANG, Y., ZHANG, Y., ZHANG, S., ZHANG, Y., & ZHANG, Q. Accuracy of genomic prediction for milk production traits in the Chinese Holstein population using a reference population consisting of cows. **Journal of Dairy Science**, 96:5315-5323, 2013.

- EDEL, C., PIMENTEL, E. C. G., PLIESCHKE, L., EMMERLING, R., & GÖTZ, K. U. Short Communication: The effect of genotyping cows to improve the reliability of genomic predictions for selection candidates. **Journal of Dairy Science**, 99:1-6, 2016.
- GAO, H., LUND, M. S., ZHANG, Y. & SU, G. Accuracy of genomic prediction using different models and response variables in the Nordic Red cattle population. **Journal of Animal Breeding and Genetics**, 130:333-340, 2013.
- GARCIA-RUIZ, A., COLE, J. B., VAN RADEN, P. M., WIGGANS, G. R., RUIZ-LOPES, F. J., & VAN TASSEL, C. P. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. **Proceedings of the National Academic Science USA**, Jul 12, 113(28):E3995-4004, 2016.
- GARRICK, D. J., TAYLOR, J. F., & FERNANDO, R.L. De-regressing estimated breeding values and weighting information for genomic regression analysis. **Genetics Selection Evolution**, 41:55, 2009.
- GODDARD, M. E. (2009). Genomic selection: Prediction of accuracy and maximization of long-term response. *Genetica* 136:245–257. <https://doi.org/10.1007/s10709-008-9308-0>.
- HAYES, B. J., BOWMAN, P. J., CHAMBERLAIN, A. J., & GODDARD, M. E. Genomic selection in dairy cattle: progress and challenges. **Journal of Dairy Science**, 92(2), 433–443, 2009.
- HENDERSON, C. R. Best Linear Unbiased Estimation and Prediction under a Selection Model. **Biometrics**, 31(2), 423–447, 1975.

- JENKO, J., WIGGANS, G. R., COOPAR, T. A., EAGLEN, S. A. E., LUFF, W. G. L., BICHARD, M., PONG-WONG, R., & WOOLIAMs J. A. **Journal of Dairy Science**, 100(1):439-452, 2017.
- JIMÉNEZ-MONTERO, J. A., GONZÁLEZ-RECIO O., & ALENDA R. Genotyping strategies for genomic selection in small dairy cattle populations. **Animal**, 6, 1216–1224, 2012.
- KÖNIG S.; & SWALVE H.H. Application of selection index calculations to determine selection strategies in genomic breeding programs. *Journal of Dairy Science* 92, 5292–5303, 2009.
- KUHN, M. T., BOETTCHER, P. J., & FREEMAN, A. E. Potential Biases in Predicted Transmitting Abilities of Females from Preferential Treatment. **Journal of Dairy Science**, 77(8), 2428–2437, 1994.
- MCHUGH, N., MEUWISSEN, T. H. E., CROMIE, A. R., & SONESSON, A. K. Use of female information in dairy cattle genomic breeding programs. **Journal of Dairy Science**, 94:4109-4118, 2011.
- MISZTAL, I., TSURUTA, S., LOURENCO, D., AGUILLAR, I., LEGARRA, A. & VITEZICA, Z. Manual for BLUPF90 family programs. Accessed July, 20, 2017.
- OLSON, K. M., VANRADEN, P. M., TOOKER, M. E., & COOPER, T. A. Differences among methods to validate genomic evaluations for dairy cattle. **Journal of Dairy Science**, 94, 2613-2620, 2011.

- PLIESCHKE, L., EDEL, C., PIMENTEL, E. C. G., EMMERLING, R., BENNEWITZ, J., & GOTZ K. U. Systematic genotyping of groups of cows to improve genomic estimated breeding values of selection candidates. **Genetics Selection Evolution**, 48:73, 2016.
- PRYCE, J.E., & DAETWYLER, H. D. Designing dairy cattle breeding schemes under genomic selection: a review of international research. **Animal Production Science**, 52, 107-114, 2012.
- PRYCE, J. E.; HAYES, B. J.; & GODDARD M.E. Genotyping dairy females can improve the reliability of genomic selection for young bulls and heifers and provide farmers with new management tools. Proc 38th ICAR Conf, Cork, Ireland, 2012.
- PSZCZOLA, M., STRABEL, T., MULDER, H. A., & CALUS, M. P. L. Reliability of direct genomic values for animals with different relationships within and to the reference population. *Journal of Dairy Science*, 95(1), 389–400, 2012.
- RINCENT, R., CHARCOSSET, A., & MOREAU, L. Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. **Theoretical and Applied Genetics**, 130(11):2231-2247, 2017.
- SARGOLZAEI, M., & SCHENKEL, F. S. QMSim: a large-scale genome simulator for livestock. **Bioinformatics**, 25:680–1, 2009.
- SCHAEFFER L.R. Strategy for applying genome-wide selection in dairy cattle. **Journal of Animal Breeding and Genetics** 123, 218–223, 2006.

- SCHÖPKE, K. & SWALVE, H. H. Review: Opportunities and challenges for small populations of dairy cattle in the era of genomics. **Animal**, 10:6, 1050-1060, 2016.
- THOMASEN, J. R., SORENSEN, A. C., LUND, M. S., & GULDBRANDSEN B. Adding cows to the reference population makes a small dairy population competitive. **Journal of Dairy Science**, 97(9):5822-32, 2014.
- TSURUTA, S.; KEOWN, JEFFREY F.; VAN VLECK, L. DALE; AND MISZTAL, I. Bias in Genetic Evaluations by Records of Cows Treated with Bovine Somatotropin. **Faculty Papers and Publications in Animal Science**, 150, 2000.
- UEMOTO, Y., OSAWA, T., & SABURI, J. Effect of genotyped cows in the reference population on the genomic evaluation of Holstein cattle. **Animal**, 11:3, 382-393, 2017.
- VENTURA, R. V., LARMER, S., SCHENKEL, F. S., MILLES, S. P., SULLIVAN, P. Genomic clustering helps to improve prediction in a multibreed population. **Journal of Animal Science**, 94(5):1844-56, 2016.
- VAN DER WERF, J. Genomic Selection in Animal Breeding Programs. *Genome-Wide Association Studies and Genomic Prediction*, 543–561, 2013.
- VUKASINOVIC, N., BACCIU, N., PRZYBYLA, C. A., BODDHIREDDY, P., & DENISE, S. K. Development of genetic and genomic evaluation for wellness traits in US Holstein cows. **Journal of Dairy Science**, 100:1, 428-438, 2017.

WU, X., LUND, M. S., SUN, D., ZHANG, Q., & SU, G. Impact of relationship between test and training animals and among training animals on reliability of genomic prediction.

Journal of Animal Breeding and Genetics, 132, 366-375, 2015.