SÃO PAULO STATE UNIVERSITY

JABOTICABAL CAMPUS

STRATEGIES FOR GENOMIC PREDICTIONS OF AN INDICINE MULTIBREED POPULATION USING SINGLE-STEP GBLUP

Marisol Londono Gil

Animal Scientist

MSc in Science - Biotechnology

2025

T
E
S
E

/

L
O
N
D
O
Ñ
O

G
I
L

M.

2
0
2
5

SÃO PAULO STATE UNIVERSITY

JABOTICABAL CAMPUS

STRATEGIES FOR GENOMIC PREDICTIONS OF AN INDICINE MULTIBREED POPULATION USING SINGLE-STEP GBLUP

Student: Marisol Londono Gil

Advisor: Prof. Dr. Fernando Baldi

Coadvisor: Prof. Dra. Daniela Lourenco

Doctoral Thesis presented to the Animal Science Graduate Program of the Faculty of Agricultural and Veterinary Sciences – UNESP, Jaboticabal Campus, as part of the requirements for obtaining the title of Doctor in Animal Breeding and Genetics.

2025

CATALOG CARD

L847s

Londoño Gil, Marisol

Strategies for genomic predictions of an indicine multibreed population using single-step GBLUP / Marisol Londoño Gil. -- Jaboticabal, 2025

136 p. : il., tabs.

Tese (doutorado) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal

Orientadora: Fernando Sebastian Baldi Rey

Coorientadora: Daniela Andressa Lino Lourenço

1. Accuracy. 2. Metafounders. 3. Reference population. 4. SNP effects. 5. Multi-breed evaluation. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

POTENTIAL IMPACT OF THIS RESEARCH

Brazil is home to one of the largest cattle populations in the world, with more than 238 million animals spread across the country. Among the most important zebu breeds are Nellore, Guzerat, Tabapua, and Brahman. These breeds are well adapted to tropical climates and perform well even under challenging pasture conditions. However, when it comes to applying genomic tools, that is, predicting an animal's potential based on its DNA, there is a significant imbalance. Nellore has a big amount of genomic and performance data available, while the other breeds lag far behind, limiting the accuracy and reliability of genetic evaluations for young animals that could become the next generation of sires.

This research addresses that gap by proposing a multi-breed genomic evaluation strategy. Instead of evaluating each breed separately, the study integrates data from all four breeds into a single, unified evaluation model. Using a method known as metafounders, the genetic differences between breeds are respected and accounted for, while still allowing the information to be shared across populations. This makes genetic predictions more accurate and stable, especially for underrepresented breeds like Guzerat and Tabapua.

Another major contribution of this work is the use of indirect predictions (IP), a method that estimates the genetic merit of young animals without performance or pedigree records. These predictions are calculated using information from older, well-evaluated animals and allow for early selection decisions. This is particularly valuable in commercial herds, where rapid and cost-effective decision-making is critical. IP enables producers to identify promising animals at an earlier age, reducing feeding and management costs, and accelerating the pace of genetic improvement. Moreover, the method simplifies the use of genomic selection in routine evaluations. Since IP relies on previously estimated SNP effects, there is no need to recalculate entire genomic evaluations each time a new animal is genotyped. This greatly reduces computational costs and makes genomic tools more accessible to a wider range of breeders.

The real-world impact of this research is already visible. The 2024 Sire Report published by ANCP (National Association of Breeders and Researchers) officially adopted the multi-breed evaluation strategy developed in this study. This has given breeders access to more reliable and science-based tools for selection, even for breeds that previously lacked robust genomic information. Furthermore, countries that collaborate with ANCP and work with Brahman cattle, for example, can now use indirect predictions for selection, even if they don't yet have performance data, as long as genomic data is available.

Looking ahead, this research has the potential to transform how genetic improvement is approached in tropical cattle systems worldwide. As more countries and organizations adopt multi-breed evaluations and indirect prediction strategies, we can expect a wider and fairer distribution of the benefits of genomics, particularly in low- and middle-income countries where data collection remains a challenge.

The methodology developed here also opens the door to international genetic evaluations between indicine breeds, promoting genetic exchange and collaboration across countries. This could increase the genetic diversity and resilience of cattle populations, supporting long-term sustainability in the face of climate change and evolving market demands.

Additionally, by enabling faster and earlier selection, this approach can significantly reduce the generation interval, making it possible to achieve genetic gains more quickly and efficiently. This is a game changer not just for elite breeding programs, but also for small and medium producers who can now access high-impact genetic tools without needing complex infrastructure.

Ultimately, this research provides a scalable and inclusive framework for implementing genomic selection in real-world settings. As technology becomes more accessible and costs continue to drop, the strategies proposed here can serve as a model for other livestock species and agricultural sectors aiming to increase productivity, sustainability, and food security through science-based innovation.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: STRATEGIES FOR GENOMIC PREDICTIONS OF AN INDICINE MULTIBREED POPULATION USING SINGLE-STEP GBLUP

AUTORA: MARISOL LONDOÑO GIL
ORIENTADOR: FERNANDO SEBASTIAN BALDI REY
COORIENTADORA: DANIELA ANDRESSA LINO LOURENÇO

Aprovada como parte das exigências para obtenção do Título de Doutora em Genética e Melhoramento Animal, pela Comissão Examinadora:

Prof. Dr. FERNANDO SEBASTIAN BALDI REY (Participaçao Virtual)
Departamento de Zootecnia / FCAV UNESP Jaboticabal

gov.br Documento assinado digitalmente
FERNANDO SEBASTIAN BALDI REY
Data: 12/05/2025 06:22:37-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. JORGE ANGEL HIDALGO MORENO (Participaçao Virtual)
Departament of Animal and Dairy Science / University of Georgia (UGA) - Athens/GA/EUA

Prof. Dr. DANISIO PRADO MUNARI (Participaçao Virtual)
Departamento de Ciencias Exatas / FCAV UNESP Jaboticabal

gov.br Documento assinado digitalmente
DANISIO PRADO MUNARI
Data: 09/05/2025 20:15:37-0300
Verifique em https://validar.iti.gov.br

Pesquisador Dr. IGNACIO AGUILAR (Participaçao Virtual)
Instituto Nacional de Investigación Agropecuaria (INIA) / Tacuarembó/Uruguay

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

Pós-doutorando LÚCIO FLÁVIO MACÊDO MOTA (Participaçao Virtual)
Departamento de Zootecnia / FCAV UNESP Jaboticabal

Documento assinado digitalmente
LUCIO FLAVIO MACEDO MOTA
Data: 10/05/2025 14:03:01-0300
Verifique em https://validar.iti.gov.br

Jaboticabal, 09 de maio de 2025

AUTHOR'S CURRICULAR DATA

Marisol Londoño Gil was born on August 31, 1993, in Entrerríos, Antioquia, Colombia, daughter of Francisco Rodrigo Londoño Londoño and Luz Marina de Jesús Gil Lopera. She earned her bachelor's degree in Animal Science from the Universidad Nacional de Colombia, Medellín campus, in May 2016. During her undergraduate studies, she gained practical experience working as an Animal Scientist at Ganadería Rusticana from January 2016 to January 2018. In February 2018, she began her master's degree in Science – Biotechnology at the same institution, under the supervision of Prof. Dr. Luis Gabriel González Herrera and Prof. Dr. Albeiro López Herrera. She completed her master's in June 2020, earning meritorious distinction for her dissertation.

In August 2020, she started her Ph.D. in Animal Breeding and Genetics at São Paulo State University "Júlio de Mesquita Filho" (UNESP), under the supervision of Prof. Dr. Fernando Sebastián Baldi Rey and Prof. Dr. Daniela Andressa Lino Lourenço. Her doctoral studies were initially funded by a CAPES scholarship, later replaced by a FAPESP scholarship (FAPESP process 2020/14275-5). As part of her Ph.D. research, she conducted an internship at the University of Georgia (UGA) in the United States in 2023, supported by a FAPESP BEPE scholarship (FAPESP process 2022/15168-3), under the guidance of Prof. Dr. Daniela Andressa Lino Lourenço. Throughout her doctoral studies, she actively participated in various research projects, conferences, courses, and symposiums, further strengthening her expertise in animal breeding and genetics.

“Remember to look up at the stars and not down at your feet. Try to make sense of what you see and wonder about what makes the universe exist. Be curious. And however difficult life may seem, there is always something you can do and succeed at.

It matters that you don't just give up”.

Stephen Hawking

With heartfelt gratitude, I dedicate this work to my unwavering pillars of strength, my beloved parents, Luz Marina Gil Lopera and Rodrigo Londoño Londoño. Your enduring support has illuminated even the darkest paths of my journey.

ACKNOWLEDGMENTS

To the graduate program in Animal Breeding and Genetics at the Faculty of Agricultural and Veterinary Sciences of the Sao Paulo State University (FCAV/UNESP) for the opportunity to complete this stage of my life.

To the São Paulo Research Foundation FAPESP for the scholarships during my doctoral degree (grant # 2020/14275-5) and my internship abroad (grant # 2022/15168-3) which allowed me great professional growth. And also, the grant that allowed to have more genotyped animals for the analysis (grant # 2021/04807-2).

To the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) which provided funds to have more data for the research (grant # 402313/2021-6).

To the National Association of Breeders and Researchers (ANCP), for providing the data needed for this research.

To the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior -Brazil (CAPES – PROEX) who partially funded my studies at the Sao Paulo State University.

To my advisor Prof. Dr. Fernando Baldi for the time, trust, patience and advice and his example of a professional who encouraged me to pursue my goals every day during my learning and PhD process.

To my co-advisor Prof. Dr. Daniela Lourenco for accepting me to do my internship in the Department of Animal and Dairy Science at University of Georgia, teaching and giving me so many opportunities that help me to achieve my doctoral degree.

To the University of Georgia for accepting me as Research Visitor for 15 months.

To the members of the thesis defense committee Danisio Munari, Jorge Hidalgo, Lúcio Mota, and Ignacio Aguilar, Fernando Baldi, thank you for your valuable contributions, for helping me improve my thesis, and for granting me the approval of my Ph.D.

To my friends and colleges at Sao Paulo State University, such as Ana Maria Fernanda Ramirez, Daniel Cardona, Joao Barbosa, Roney Teixeira, Gabriel Gubiani, and Larissa Bordin Temp, who were really important and helpful during my doctorate journey.

To my colleges at the University of Georgia, such as Zuleica Trujano, Arielly Garcia, Sergio Sanchez, Fernando Bussman and Jorge Hidalgo, whom I'm really grateful by their support during my internship and my doctoral projects.

To my family, especially my mom, Luz Marina Gil Lopera; my dad, Rodrigo Londoño Londoño; my brother, Andrés Londoño Gil; my sisters, Natalia Andrea Londoño Gil and María Londoño Gil; my nephew, Jerónimo Rodrigues Londoño; and my nieces, Juliana Rodrigues Londoño and Hannah Cabrera Londoño, whom I love with all my heart and who are the reason I'm here today.

SUMMARY

STRATEGIES FOR GENOMIC PREDICTIONS OF AN INDICINE MULTIBREED POPULATION USING SINGLE-STEP GBLUP

ABSTRACT – Genomic selection has changed animal breeding by enabling the early identification of superior individuals based on their genetic merit. However, this approach faces considerable challenges when applied to indicine cattle, particularly populations such as Nellore, Brahman, Guzerat, and Tabapua. Brahman, Guzerat, and Tabapua present small reference populations, making genomic selection difficult and inaccurate. A multi-breed evaluation could solve this problem, but it must consider the heterogeneity in allele frequencies among breeds and the incompatibility between genomic (**G**) and pedigree (**A**) relationship matrices. This work investigated strategies, feasibility, and effectiveness for genomic prediction in indicine cattle by focusing on a multi-breed population that includes Nellore, Brahman, Guzerat, and Tabapua using the single-step genomic best linear unbiased predictor (ssGBLUP). Data for approximately 4.2 M pedigree records, 329 K phenotypic observations, and 63 K genotyped animals were used. The traits analyzed were adjusted weight at 210 days (W210), adjusted weight at 450 days (W450), and scrotal circumference at 365 days (SC365). Different scenarios were evaluated, comparing single-breed and multi-breed approaches. Multi-breed evaluations were conducted both with and without the inclusion of metafounders (MF) to account for genetic differences across breeds. At the same time, without MF, the inclusion of breed-specific allele frequencies was analyzed or not adjusted at all. Comparative analyses of the different scenarios were performed using the linear regression (LR) method to assess bias, dispersion, and accuracy. In addition to improving direct genomic evaluations, this work addressed the challenge of obtaining genomic predictions for young, genotyped animals lacking phenotypic data or pedigree information by computing indirect predictions (IP). These IP were derived as the sum of SNP effects weighted by gene content. The evaluation of the different scenarios was made using the LR method. Moreover, the accuracy of IP (ACC_{IP}) was estimated from the prediction error covariance (PEC) of SNP effects and validated through correlations with the accuracy of genomic estimated breeding values ($ACCG_{EBV}$). The results of this work showed that multi-breed evaluations significantly improved prediction accuracy compared to single-breed approaches,

particularly helping breeds with smaller genomic datasets such as Guzerat, and Tabapua. Incorporating MF reduced bias and mitigated under- or over-dispersion in the predictions, with a modest increase in IP accuracy. The single-breed evaluations frequently showed lower accuracy, especially for Guzerat and Tabapua which have limited genomic data. The study demonstrated that structured multi-breed genomic evaluations are a robust and effective strategy for improving genomic selection in Brazilian indicine cattle. While using genomic information across multiple breeds, this approach improved the accuracy of predictions for traits of economic importance and allowed rapid IP for young animals. Consequently, Brazilian indicine multi-breed evaluations can help selection decisions and potentially accelerate genetic progress, especially in breeds with constrained reference populations.

Keywords: Accuracy, allele frequencies, Brahman, genomic selection, Guzerat, metafounders, multi-breed evaluation, Nellore, reference population, SNP effects, Tabapua.

ESTRATÉGIAS PARA PREDIÇÃO GENÔMICA EM POPULAÇÕES MULTIRRACIAIS DE ZEBUÍNOS APLICANDO O MÉTODO DO PASSO ÚNICO GENÔMICO BLUP

RESUMO – A seleção genômica transformou a seleção de animais ao permitir a identificação precoce de indivíduos superiores com base em seu mérito genético. Contudo, essa abordagem enfrenta desafios consideráveis quando aplicada ao gado zebuíno, especialmente em populações como Nelore, Brahman, Guzerá e Tabapuã. Brahman, Guzerá e Tabapuã apresentam populações de referência pequenas, o que torna a sua seleção genômica difícil e imprecisa. Uma avaliação multirracial poderia resolver esse problema, mas deve considerar a heterogeneidade nas frequências alélicas que podem existir entre as raças e a incompatibilidade entre as matrizes de relacionamento genômico (**G**) e de pedigree (**A**). Este trabalho pesquisou estratégias para as predições genômicas de gado zebuíno, em uma população multirracial que incluiu o Nelore, Brahman, Guzerá e Tabapuã, utilizando o preditor linear não viesado genômico de uma única etapa (ssGBLUP). Foram utilizados dados de aproximadamente 4,2 milhões de registros de pedigree, 329 mil observações fenotípicas e 63 mil animais genotipados. As características analisadas foram o peso ajustado aos 210 dias (W210), o peso ajustado aos 450 dias (W450) e a circunferência escrotal aos 365 dias (SC365). Diferentes cenários foram avaliados, comparando abordagens de raça única e multirracial. As avaliações multirraciais foram conduzidas tanto com a inclusão de metafundadores (MF), para levar em conta as diferenças e similitudes genéticas entre as raças, quanto sem eles. Sem MF, foi analisada o ajuste ou não de frequências alélicas específicas de cada raça. Análises comparativas dos diferentes cenários foram realizadas utilizando o método de regressão linear (LR) para avaliar viés, dispersão e acurácia. Além das avaliações genômicas diretas, este trabalho abordou o desafio de obter predições genômicas para animais jovens genotipados que carecem fenótipos ou pedigree, calculando predições indiretas (IP). Essas IP foram derivadas como a soma dos efeitos dos SNPs ponderados pelo conteúdo gênico. A avaliação dos diferentes cenários foi realizada utilizando o método LR. Ademais, a acurácia das IP (ACC_{IP}) foi estimada a partir da covariância do erro de predição (PEC) dos efeitos dos SNPs e validada através de correlações com a acurácia dos valores genômicos estimados (ACC_{GEBV}). Os resultados deste trabalho

demonstraram que as avaliações multirraciais melhoraram significativamente a precisão das predições em comparação com as abordagens de raça única, beneficiando particularmente as raças com conjuntos de dados genômicos menores, como Guzerá e Tabapuã. A inclusão de MF reduziu o viés e mitigou a sub ou superdispersão nas predições, com um modesto aumento na acurácia das IP. As avaliações de raça única frequentemente apresentaram menor precisão, especialmente para Guzerá e Tabapuã, que possuem dados genômicos limitados. O estudo demonstrou que as avaliações genômicas multirraciais estruturadas são uma estratégia robusta e eficaz para melhorar a seleção genômica no gado zebuíno brasileiro. Ao utilizar informações genômicas de múltiplas raças, essa abordagem melhorou a precisão das predições para características de importância econômica e permitiu a rápida obtenção de IP para animais jovens. Consequentemente, as avaliações multirraciais de gado zebuíno brasileiro podem auxiliar nas decisões de seleção e potencialmente acelerar o progresso genético, especialmente em raças com populações de referência pequenas.

Palavras-chave: Acurácia, avaliação multirracial, Brahman, efeitos dos SNPs, frequências alélicas, Guzerá, metafundadores, Nelore, população de referência, seleção genômica, Tabapuã.

LISTS OF ABBREVIATIONS

A: Pedigree-based relationship matrix

ACC_{GEBV}: Accuracy of GEBV

ACC_{IP}: Accuracy of IP

ANCP: National Association of Breeders and Researchers *Associação Nacional de Criadores e Pesquisadores*

BLUP: Best linear unbiased prediction

C: Coefficient matrix of the mixed model equations

EBV: Estimated breeding value

G: Genomic relationship matrix.

GBLUP: Genomic best linear unbiased prediction

GEBV: Genomic estimated breeding value

H: Combined genomic and pedigree-based relationship matrix

IBGE: Brazilian Institute of Geography and Statistics

IP: Indirect predictions

LD: Linkage disequilibrium

LHS: Left hand side

LR: Linear regression

MAF: Minor allele frequency

MAS: Marker assisted selection

MF: Metafounders

MME: Mixed model equations

N: Number of records

PCA: Principal component analysis

PEC: Prediction error covariance

PEV: Prediction error variance

QTL: Quantitative trait loci

SC365: Fcrotal circumference at 365 days

SD: standard deviation

SNP: Single nucleotide polymorphism

ssGBLUP: Single step genomic best linear unbiased prediction

SVD: Singular value decomposition

UPG: Unknown parent groups

USA: United States of America

USDA: United States Department of Agriculture

W210: Adjusted weight at 210 days

W450: Adjusted weight at 450 days

CHAPTER 1: GENERAL CONSIDERATIONS

1.1 INTRODUCTION

Indicine Beef cattle breeds (*Bos taurus indicus*) are socially and economically important worldwide, dominating meat production systems in tropical and subtropical regions. They serve as a crucial resource for food security in communities within developing countries (Thornton, 2010). The dominance of Zebu breeds is mainly due to their superior ability to tolerate high temperatures and humidity, which are common in these regions. Additionally, these cattle exhibit greater resistance to ticks and other environmental challenges than taurine breeds (*Bos taurus taurus*), typically found in temperate regions(Utsunomiya et al., 2019).

According to the United States Department of Agriculture (USDA, 2024), Brazil is the leading beef exporter and producer worldwide, with production systems predominantly relying on indicine breeds (*Bos taurus indicus*) adapted to tropical conditions. The Nellore breed, which accounts for around 80% of Brazil's approximately 200 million cattle, is the most prominent, followed by Brahman, Guzerat, and Tabapua(Cunha et al., 2023; de Souza et al., 2021). Most technological developments and applications in genomic evaluations have focused on Nellore. Despite these efforts, the other breeds are still in the early stages of genomic evaluations, limiting their potential gains from genomic selection, mainly due to the small size of their reference populations.

A key challenge in these breeds is the limited size of their reference populations, markedly when pedigree, phenotypic, and genomic data are scarce. This limitation makes it difficult to develop a sufficiently large reference population with high accuracy. One strategy to improve genomic evaluations is to combine these breeds into a metapopulation, which may increase the predictive power without additional data recording costs (Steyn et al., 2019). However, integrating multiple breeds may introduce compatibility issues between genomic (**G**) and pedigree (**A**) relationship matrices, as allele frequencies used for scaling **G** are typically estimated based on genotyped individuals across all populations.

To mitigate these compatibility issues in multi-breed evaluations, Legarra et al. (2015) proposed the use of metafounders (MF) in single-step genomic best linear unbiased prediction(ssGBLUP). MF are pseudo-individuals that act as founders of base populations, which can be considered simultaneously as the sire and dam of the base population. This base population becomes related and can be inbred (Legarra et al., 2015), ensuring compatibility between **G** and **A** matrices in genomic evaluations. Therefore, this methodology has been recommended when combining genomic and pedigree information for individuals, as in ssGBLUP, and when several base populations are combined (Legarra et al., 2015), as in the multi-breed evaluations.

Another strategy for improving the compatibility is to include breed-specific allele frequencies when constructing **G.** This is important because genomic regions inherited from different breeds may show unique allele frequencies (Lourenco et al., 2016). The genome of multi-breed populations combines genomic regions inherited from their parental breeds. Thus, the frequency and effects of single nucleotide polymorphism (SNP) alleles may vary significantly across different breeds, depending on their genetic diversity and history (Sevillano et al., 2019). By considering these differences, it may be possible to enhance the accuracy of genomic predictions in multi-breed populations.

While incorporating multi-breed evaluations, MF, and breed-specific allele frequencies can improve compatibility and the accuracy of genomic evaluations for small breeds, a significant challenge remains in accurately predicting young animals. For closely related breeds such as indicine cattle, across-breed genomic predictions are expected to be highly accurate (Steyn et al., 2019), meaning that merging them in a single evaluation should not compromise prediction power. However, young animals without phenotypes or progeny data are often excluded from official evaluations (Tsuruta et al., 2021), despite many commercial farms increasingly genotyping them for herd management. Including these animals in the evaluation could reduce the accuracy of genomic breeding values (GEBVs) and may introduce bias (Lourenco et al., 2018), highlighting the growing need for reliable indirect predictions (IP) in these individuals (Londoño-Gil et al., 2023; Tsuruta et al., 2021).

IP can help identify superior young candidates early, especially in regions with

limited genomic evaluations. They are calculated as the sum of SNP effects weighted by gene content and can be derived from GEBVs. In genomic BLUP (GBLUP) or ssGBLUP, SNP effects are not directly available but can be backsolved from GEBVs using formulas (Strandén & Garrick, 2009; VanRaden, 2008). Once SNP effects are computed, IP for young animals are obtained as described before. This approach is commonly used in dairy cattle genomic predictions to release fast interim predictions based on IP, though SNP effects are typically estimated through multi-step procedures.

The accuracy of predicted breeding values, and therefore IP, is also important for better selection decisions. Henderson (1984) showed that accuracies of EBVs can be calculated based on the prediction error variance (PEV) by inverting the coefficient matrix of the BLUP mixed model equations (MME). Similarly, Strandén & Christensen (2011) and Tier et al. (2018) showed how to calculate the accuracy of IP using the SNP prediction error covariance (PEC). Calculating PEC for SNP effects in ssGBLUP requires considering the number of genotyped animals (Gualdrón Duarte et al., 2014). In principle, the exact computation of SNP PEC requires obtaining the full PEC matrix for all genotyped animals from the inverse of the MME(Garcia et al., 2022). Thus, the accuracy of IP is also feasible to obtain in the multi-breed scenario.

In the context of multi-breed evaluations, particularly relevant for Brazil's livestock sector, improving the accuracy of genomic predictions is essential. Evaluating multiple breeds together may increase the precision of genetic predictions, providing more accurate estimates for selection decisions. Precise genomic evaluations improve genomic breeding values (GEBVs) and may provide more accurate indirect predictions (IP), which is important to identify young superior animals earlier, especially in regions with less established genomic infrastructure. The different strategies to do so, therefore, help to increase evaluation accuracies, making these methods important for future breeding decisions and genetic progress for these populations, especially the ones of small size.

1.2 OBJECTIVES

1.2.1 General objective

To evaluate different strategies for genomic evaluation of a metapopulation composed of the breeds Nellore, Brahman, Guzerat, and Tabapua using the single-step genomic BLUP method for growth traits such as weaning weight (W210) and weight adjusted at 15 months of age (W450), as well as for the reproductive trait scrotal circumference adjusted to 365 days of age (C365).

1.2.2 Specific objectives

- To evaluate the feasibility of a multi-breed single-step genomic BLUP (ssGBLUP) evaluation composed of Nellore, Brahman, Guzerat, and Tabapua breeds, comparing the performance of multi-breed ssGBLUP evaluations with or without specific compatibility treatments to the single-breed evaluations.

- To compare compatibility between **G** and **A** matrices using different strategies, including: **1**) varying numbers and definitions of metafounders (MF), and **2**) implementing breed-specific allele frequencies in the construction of **G**.

- To compute indirect predictions (IP) for young genotyped animals from Nellore, Brahman, Guzerat and Tabapua using single-breed reference populations and a multi-breed reference population.

- To determine the minimum number of genotyped animals required in the official Nellore evaluation to estimate accurate SNP effects and IP.

- To compute the individual accuracy of IP by back-solving the prediction error covariance (PEC) of genomic estimated breeding values (GEBVs) from ssGBLUP into the PEC of SNP effects, and to investigate the feasibility of this method for calculating IP accuracy in a multi-breed evaluation.

1.3 LITERATURE REVIEW

1.3.1 Beef production and the role of Indicine breeds in Brazil

The beef production system of Brazil has experienced a remarkable transformation in recent decades. With an estimated 238.6 million cattle in 2023, according to the Brazilian Institute of Geography and Statistics (IBGE, 2024), the herd has more than doubled since the 1970s, driving significant productivity improvements. This growth enabled Brazil to shift from being a meat importer to achieving self-sufficiency in beef production. In 2022, 72.1% of production was consumed domestically, contributing to a per capita consumption of 36.7 kg annually, one of the highest globally. The remaining 27.9% was exported (Cunha et al., 2023), strengthening the position of Brazil as the leading beef exporter in the world since 2004, with exports reaching nearly 3 million tons in 2023 (USDA, 2024). Genetics and management innovations have played a key role in this success, transforming farms into profitable, quality-focused rural enterprises.

Indicine cattle breeds (*Bos taurus indicus*) play an important role in Brazil's beef production. Zebu cattle were first introduced in the country during the 19th and 20th centuries through imports of nearly 7,000 animals of Indian origin (Santana et al., 2016). These breeds exhibit strong adaptation to high temperatures, as in tropical countries, and low-quality pastures and are remarkably resistant to parasites and diseases. Among them, the Nellore breed dominates the national herd, representing ~80% of all beef cattle (USDA, 2021). However, other breeds such as Guzerat, Gir, Brahman, and Tabapua also play significant roles in Brazil’s beef production system (Ferraz & Felício, 2010).

The Nellore originated from the Ongole cattle of India. The Ongole cattle have a history of over 2,000 years of adaptation to harsh Indian conditions. Over time, this breed, later known as Nellore in Brazil, developed traits such as a large hump, pigmented skin, distinct ear shape, robust body structure, and a protective coat, all of which proved advantageous in tropical climates. Nellore cattle were first introduced to Brazil in 1868. Their population expanded significantly with the establishment of the Nellore Herd Book in 1938, followed by the last major importation of Ongole cattle in

1962. Today, Brazil is the largest producer of Nellore cattle, and the breed has been exported to several countries worldwide, especially in Latin America. Nellore cattle are known for their adaptability to tropical conditions, resistance to heat and insects, metabolic efficiency, and strong maternal traits. Due to its economic importance, the breed has been intensively selected in Brazil, with genetic improvement supported by artificial insemination and embryo transfer technologies (Ferraz & Felício, 2010; https://breeds.okstate.edu/cattle/, 2025).

Brahman cattle were created in the USA through crossbreeding in the 19th century, using bulls from the Guzerat, Nellore (Ongole), Krishna, and Gir breeds, along with local American *Bos taurus* cattle (Koufariotis et al., 2018). Breeding programs developed a population that retained the adaptability of its Indian ancestors while improving traits relevant to beef production, such as growth rate and carcass quality, through the introgression of *Bos taurus* genetics (Koufariotis et al., 2018). Brahman cattle exhibit exceptional heat tolerance, parasite resistance, and efficient feed utilization. The breed is characterized by a prominent hump, loose skin, and large ears that assist in thermoregulation. Furthermore, its capability to sweat and release oils helps repel insects and increase their adaptation to warm and tropical temperatures. The breed also demonstrates high fertility, strong maternal instincts, and resistance to common bovine diseases, making them valuable in beef production (Briggs & Briggs, 1980).

The Guzerat breed, also known as Guzera, Gujera, Gujrati, Gusera, and Guzerath, was developed from the Kankrej breed imported from Gujarat, India, between the 1870s and 1920s. Characterized by a large frame, gray color, and long, lyre-shaped horns, they are prized as draft animals and moderate milk producers. The breed, which ranges from light gray to black, significantly contributed to the development of the American Brahman. Guzerat cattle are valued for their adaptability to tropical climates, insect tolerance, disease resistance, and strong maternal traits (Pires et al., 2020). Although their carcass traits are generally lower in quality than *Bos taurus*, they excel in growth and feed efficiency. They are known for their docile temperament and fertility under adverse conditions. The breed is versatile, hardy, and well adapted to tropical environments, with females showing excellent maternal ability

and males exhibiting fast weight gain (Brito et al., 2020). The breed has been selected for beef, milk, and more recently, dual-purpose production (Peixoto et al., 2021).

Tabapua was Brazil's first polled Zebu breed, originating from a cross of polled Nellore, Guzerat, and Gyr - breeds highly adapted to Brazilian tropical conditions (Bernardes et al., 2016). The polled phenotype likely comes from Nellore or domestic cattle with European cattle introgression(Filho et al., 2012; Vercesi Filho et al., 2002). The Tabapua is fertile, early-maturing, and heavy at weaning. The breed is widely used for beef production. It has excellent carcass yield rates, and its females exhibit great maternal ability, as evidenced by the quality of the weaned calves. Therefore, the goal of the breeding program for this breed is to improve productive and reproductive performance (Bernardes et al., 2015).

Genetic selection in Brazil has primarily focused on the Nellore breed, leading to significant advancements in genetic improvement. However, other breeds such as Brahman, Guzerat, and Tabapua also play a key role in national beef production and have been developing their genetic improvement programs. In this way, the primary goal of genetic selection in zebu cattle has been to enhance beef production and profitability by identifying and breeding genetically superior animals. Selection programs have traditionally prioritized growth traits such as weaning weight and yearling weight for both males and females, as they show moderate to high heritability (Chiaia et al., 2015; Lemos et al., 2015; Torres-Vázquez & Spangler, 2016). These traits are relatively easy to measure in selection candidates and serve as strong indicators of slaughter weight.

Maternal ability, which includes milk production and the care of the calving until weaning, is also of great economic importance. However, the heritability of maternal effects on weaning weight is typically low (Kluska et al., 2018; Martínez et al., 2016), limiting genetic gains from direct selection. Moreover, those effects are challenging to improve since a sire's maternal effect is only expressed in the weaning weight of his grandsons, resulting in a delayed response. Despite these challenges, maternal effects (both genetic and permanent maternal environmental components) explain a significant proportion of the phenotypic variation in weaning weight, comparable to the contribution of direct additive genetic variance. Therefore, incorporating these effects

into genetic evaluation models is essential to obtaining reliable and unbiased estimates of genetic parameters.

Given the genetic potential of Nellore, Brahman, Guzerat, and Tabapua cattle, further genetic and genomic research is essential to expanding the genetic base of Brazilian beef production. This will optimize selection strategies and enhance the efficiency and sustainability of tropical beef systems. In recent years, genetic programs for these breeds have expanded, focusing on improving key traits, such as adaptability, feed efficiency, and maternal ability, to ensure long-term productivity and resilience.

1.3.2 History of genetic evaluations

At the beginning of the 20th century, a theoretically infinite number of additive unlinked loci with small additive effects (the infinitesimal model) were reported to be responsible for variations in phenotypic traits that exhibit continuous variation (Fisher, 1918). Lush (1935) was one of the pioneers in genetic evaluation, as he developed the foundations of progeny testing by using the additive genetic fraction of the total phenotypic variance – heritability – to conduct his evaluations. Subsequently, animal genetic improvement gained significant momentum in the 1940s, based on research by Fisher (1936), Smith (1936) and Hazel (1943), who developed multiple selection methodologies, such as selection indices. However, it was Lush in 1947 who introduced the use of different sources of pedigree information among individuals in a population to estimate genetic merit.

During the 1950s and 1960s, methods for estimating variance components were developed based on least squares expectations, with their most important property being their unbiasedness(Harvey, 1960; Henderson, 1953). Today, the least squares-based methods have evolved into minimum variance quadratic unbiased estimation methods (Rao & Edwards, 1971) and maximum likelihood methods and their variants (Hartley & Rao, 2014; Patterson & Thompson, 1971).

Another significant contribution to genetic evaluation was made by Henderson (1963, 1973, 1976, 1984) and Henderson & Quaas (1976), who outlined and implemented complex mathematical models known as univariate and multivariate linear mixed models in animal breeding. They also introduced the Best Linear

Unbiased Predictor (BLUP), which enabled the estimation of solutions for the random effects in mixed models, using the pedigree-based relationship matrix (**A**). These methods significantly improved genetic evaluation programs and contributed to genetic progress in domestic species, particularly in cattle. As a result, BLUP became widely disseminated and adopted as the standard method for predicting genetic values in plant and animal breeding programs. This led to significant impacts of quantitative genetics on animal genetic improvement programs worldwide throughout the 1980s and 1990s (Delgado et al., 1995; Tirados, 2001). Major advances were made over the years in different countries, with the United States pioneering these genetic evaluation and selection procedures.

After these years, there was a growing interest in the use of molecular markers to accelerate genetic gain in both males and females, including young animals. That strategy was also important for identifying genes and polymorphisms associated with economically important traits across various livestock breeds and was called Marker assisted selection (MAS) (Goddard et al., 2011). As a result, the discovery of quantitative trait loci (QTL) in several domestic species spurred the development of new methodologies that integrated molecular marker information into genetic evaluation models (Georges et al., 1995; Crawford 2001), which led to new methods for estimating breeding values using information from molecular markers and QTL (Meuwissen et al., 2001).

Initially, microsatellites were used as genome wide markers with potential to detect QTLs through linkage analysis in full – and half-sib families. Typically, panels of 100-200 microsatellites were employed to identify QTLs regardless of their chromosomal location. However, their application was limited due to challenges in precisely mapping QTLs and the weak linkage disequilibrium (LD) between microsatellites and QTLs. In practice, the linkage phase had to be determined for each family before marker-assisted selection could be applied, resulting in modest gains and minimal practical implementation (Goddard et al. 2011)

These limitations led Meuwissen et al. (2001) to propose an innovative methodology that revolutionized traditional genetic selection. This approach integrated genomic information from thousands of single nucleotide polymorphisms (SNPs)

across the genome into conventional genetic evaluation, leading to what is now known as genomic selection, by using the genomic-based relationship matrix (**G**). Genomic selection, therefore, is the selection based on genomic estimated breeding values (GEBV), which uses the linkage disequilibrium between SNPs and QTL. Today, it is one of the most advanced technologies in animal breeding and is implemented in most of the livestock species (Misztal et al., 2021; Schefers & Weigel, 2012), and several types of methodologies for its use were implemented (Aguilar et al., 2010; Habier et al., 2011; Legarra et al., 2009; Meuwissen et al., 2001; VanRaden, 2008). Consequently, genomic selection has been widely adopted in breeding programs for both taurine and zebu beef cattle breeds (Berry et al., 2016; Mrode et al., 2019). By incorporating genomic information, GEBVs have been predicted with reasonable accuracy even without phenotyping selection candidates, increasing the selection accuracy of young animals (Carvalho et al., 2019) and improving genetic progress for traits that are difficult or costly to measure, have low heritability, or require progeny testing for reliable genetic predictions (Garrick, 2011).

1.3.3 Single-step genomic best linear unbiased predictor (ssGBLUP)

With the advent of genomic selection, multiple methodologies have been developed to enhance the accuracy of genetic predictions (Aguilar et al., 2010; Habier et al., 2011; Legarra et al., 2009; Meuwissen et al., 2001; VanRaden, 2008). However, in most evaluation models, not all animals included in a genomic evaluation are genotyped. In fact, only a small proportion of the population typically has genotypic data. Additionally, not all animals have phenotypic records, such as young individuals or those with sex-limited traits. To address the limitations of multi-step procedures, Aguilar et al. (2010), Legarra et al. (2009), and Christensen & Lund (2010) proposed combining pedigree (**A**) and genomic (**G**) relationship matrices into a single **H** matrix and using this matrix as the covariance structure in the mixed model equations. This allowed all available sources of information about an animal to be used in a single step.

The covariance structure among animals in ssGBLUP was defined as:

$$\mathbf{H^{-1}} = \mathbf{A^{-1}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G^{-1}} - \mathbf{A_{22}^{-1}} \end{bmatrix},$$

where **G** is constructed as in VanRaden (2008), that is, $\mathbf{G} = \mathbf{ZZ'}$, where:

$$\mathbf{Z} = (\mathbf{M} - \mathbf{P}) / \left[2 \sum_{j=1}^{n} p_j (1 - p_j) \right]^{1/2},$$

with $\mathbf{M}$ being the matrix of n SNP genotypes for each animal, with elements set to 0, 1, and 2 for the homozygote, heterozygote, and the other homozygote, respectively, which represent the count of the second allele; $\mathbf{P}$ is the matrix of twice the frequency of the second allele p at locus j (p_j).

This approach reduces computational costs and increases the accuracy of evaluations compared to models that rely solely on phenotypic information (Legarra et al., 2014). One of the key advantages of the ssGBLUP method is the ease of inverting the **H** matrix, as well as the ability to jointly evaluate genotyped and non-genotyped animals. This is particularly important since genotyping the entire population is often not feasible. Therefore, GEBVs can be easily obtained for all the animals included in the evaluation.

1.3.4 Genomic evaluation challenges in small and multi-breed populations

Genomic selection has significantly accelerated genetic progress in various livestock populations, particularly in breeds with large, well-established reference populations (Guinan et al., 2023; Scott et al., 2021). A good example is the Holstein breed, where genomic selection has been highly effective due to the large availability of genotyped and phenotyped animals (VanRaden et al., 2009). However, implementing genomic selection in other populations presents challenges, especially in multi-breed and crossbred populations where historical reference datasets, such as those available for dairy cattle, are not as readily accessible (Hayes et al., 2023).

Multi-breed genomic evaluations are desirable in livestock breeding programs, as they enable the selection of sires and dams across breeds, crossbreds, and composites, increasing selection intensity. The benefits of multi-breed evaluations for improving the accuracy of genomic predictions, especially for breeds with smaller reference populations, have been demonstrated in dairy cattle (Khansefid et al., 2020), sheep (Brito et al., 2017; Moghaddar et al., 2014), and beef cattle (Hayes et al., 2019). Different methodologies have been proposed to optimize multi-breed genomic

predictions, including: (i) using a single genomic relationship matrix that includes all breeds and breed groups (Hozé et al., 2014; Steyn et al., 2019), (ii) adjusting for breed-specific allele frequencies while accounting for pedigree-based proportions (Makgahlela et al., 2014), (iii) incorporating breed-specific effects or considering the breed origin of alleles (Karaman et al., 2021; Lopes et al., 2017), and (iv) considering linkage disequilibrium (LD) patterns.

A joint multi-breed genomic evaluation assumes a single genetic effect for all genotyped animals, either by applying a unified additive genetic model, where all individuals share a common genomic relationship matrix, or by estimating the same SNP effects across breeds (Hozé et al., 2014; Olson et al., 2012; Steyn et al., 2019). However, the success of this approach depends on two key factors: (i) the degree to which QTL effects remain consistent across breeds, and (ii) whether the SNP density used provides sufficient genome-wide coverage for all populations.

The persistence of LD phase across breeds plays a crucial role in the accuracy of genomic predictions in multi-breed populations. Studies in pigs have shown that inconsistent LD phase can explain why SNP markers associated with major effects in one breed may not be effective in another (Veroneze et al., 2014). Similarly, Hozé et al. (2014) demonstrated that the accuracy gains from a multi-breed reference population were significantly higher for animals without their sire in the reference population (12–16%) compared to those with a known sire (1–2%). Additionally, research in beef cattle by Chen et al. (2013) suggested that the advantage of a combined reference population over a single-breed reference is more pronounced when within-breed relationships between reference and validation animals are strong.

This challenge is particularly relevant for indicine cattle, including Nellore, Brahman, Guzerat, and Tabapua, where differences in breed-specific adaptation, genetic diversity, and evolutionary history may affect LD persistence and the transferability of genomic prediction models. To optimize multi-breed evaluations for these breeds, further research is needed to refine reference population composition and identify key causative mutations influencing economically important traits.

Nonetheless, multi-breed genomic evaluations are especially beneficial for smaller breeds like Brahman, Guzerat, and Tabapua, which have limited population

sizes. For these breeds, building a large reference population with a substantial number of phenotyped and genotyped animals, and thus achieving accurate genetic evaluations, remains a challenge. In Brazil, significant progress has been made in Nellore breeding programs, with large-scale genotyping efforts focusing on key sires, making genomic evaluations a well-established component of genetic improvement strategies. However, despite initiatives by breed associations and genetic programs, genomic evaluations for Brahman, Guzerat, and Tabapua remain in their early stages or involve relatively small numbers of genotyped individuals. This limitation constrains the genetic gains achievable through genomic selection, underscoring the need for expanded genotyping efforts and optimized multi-breed evaluation strategies to maximize the benefits of genomic selection in indicine cattle.

1.3.5 Strategies used in multi-breed genomic predictions

Metafounders

Legarra et al. (2015) proposed the concept of metafounders (MF) in the context of ssGBLUP to ensure compatibility between $\mathbf{G}$ and $\mathbf{A}_{22}$ in the presence of multiple base populations. In some cases, these matrices may be incompatible because not all animals in the pedigree are genotyped. The base population for the pedigree-based relationship matrix consists of the founders of the pedigree. In contrast, the base population for the genomic relationship matrix consists of the population of genotyped animals. This difference in base populations leads to the incompatibility between these matrices (Legarra et al., 2014, 2015). Furthermore, pedigree-based relationships assume that all base individuals are unrelated and come from an infinite population size. However, the knowledge about them is limited due to the starting point at which the data collection begins, and that was the reason the MF theory was developed (Legarra, Bermann, et al., 2024b; Legarra et al., 2015)

Metafounders are pseudo-individuals acting as founders of the populations, which can be considered simultaneously as the sire and dam of the base population. This allows the base population to be related and potentially inbred (Legarra et al., 2015). Relationships across the MF are captured in the $\boldsymbol{\Gamma}$ matrix, which models the covariances between population means, levels of homozygosity, and overall similarity.

The MF approach defines a reference point as an ideal population with allele frequencies 0.5 at biallelic markers, ensuring maximum heterozygosity under Hardy-Weinberg equilibrium. This reference also aligns with genomic relationships, using the same 0.5 allele frequency assumption (Legarra et al., 2024).

Consequently, the **Γ** matrix represents the average genetic relationships among an unobserved pool of founder gametes, which are referred to as metafounders (Legarra, Bermann, et al., 2024a). When several populations are considered simultaneously, genetic relationships across them can also be accounted for. In this context, MF are related through an additive relationship matrix, **A**(**Γ**), given by a positive definite matrix **Γ**, as follows:

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma^{\mathrm{A}} & \gamma^{\mathrm{A,B}} & \cdots & \gamma^{\mathrm{A,N}} \\ \gamma^{\mathrm{B,A}} & \gamma^{\mathrm{B}} & \cdots & \gamma^{\mathrm{B,N}} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma^{\mathrm{N,A}} & \gamma^{\mathrm{N,B}} & \cdots & \gamma^{\mathrm{N}} \end{bmatrix},$$

where γ^{A} and γ^{B} are the relationships within MF A and B, respectively, up to N MF, and $\gamma^{\mathrm{A,B}}$ are the coefficients across MF (Legarra et al., 2015).

When using the ssGBLUP method, the inverse of $\mathbf{H}^{\Gamma-1}$ matrix is adjusted to be incorporated to the mixed model equations, as follows:

$$\mathbf{H}^{\Gamma-1} = \mathbf{A}^{\Gamma-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{05}^{\Gamma-1} - \mathbf{A}_{22}^{\Gamma-1} \end{bmatrix},$$

where $\mathbf{H}^{\Gamma-1}$ is the inverse matrix that combines the genomic and pedigree matrix considering MF; $\mathbf{A}^{\Gamma-1}$ is an inverse matrix of additive relationships with various MF; $\mathbf{G}_{05}^{\Gamma-1}$ is the inverse of the genomic relationship matrix of genotyped animals and $\mathbf{A}_{22}^{\Gamma-1}$ is the matrix of additive relationships between genotyped animals, considering MF.

The MF approach has been recommended when combining genomic and pedigree information for individuals, as in ssGBLUP, and when several base populations are combined (Legarra et al., 2015), as in the multi-breed evaluations.

Research has demonstrated that in genetic evaluations comprising multiple breeds, the use of MF can increase the prediction accuracy while also may reduce the bias (Granado-Tajada et al., 2020; Junqueira et al., 2020; Kudinov et al., 2022;

Londoño-Gil et al., 2024a; López-Correa et al., 2021; Macedo et al., 2022; van Grevenhof et al., 2019)

Adjusts in the genomic relationship matrix

Another strategy used when evaluating multi-breed populations while improving the compatibility between **G** and $\mathbf{A}_{22}$ was introduced by Lourenco et al. (2016).This approach involves using breed-specific allele frequencies when constructing **G**. In multi-breed populations, the genome is a combination of genomic regions inherited from their parental breeds. Hence, the frequency, linkage disequilibrium, and effects of SNP alleles can vary noticeably depending on the breed that they originated from (Sevillano et al., 2019). Since **G** reflects population or breed stratification, different breeds may have different allele frequencies. Therefore, accounting for breed-specific allele frequencies may be beneficial for genomic predictions(Lourenco et al., 2016). To address this, Lourenco et al. (2016) proposed to center and to scale the **G** matrix based on the breed allele frequencies. Considering that frequencies are different across breeds and $\mathbf{G} = \mathbf{Z}\mathbf{Z}'$, $\mathbf{Z}$ can be created as follows:

$$\mathbf{Z} = (\mathbf{M} - \mathbf{P}_{\mathrm{K}}) / \left[2 \sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{p}_{\mathrm{jK}} \left(1 - \mathrm{p}_{\mathrm{jK}}\right) \right]^{1/2},$$

where $\mathbf{P}_{\mathrm{K}}$ contains twice p_{jK}, the frequency of the second allele at the j locus, specific for the Kth breed.

As result, this matrix is centered to fix the mean values of the allele effects as zero, and it is adjusted to the specific allele frequency of the founding breed, providing greater consistency between the coefficients of relationship of genotyped and non-genotyped individuals (Makgahlela et al., 2013).

However, this approach presents certain challenges. Some studies have shown that constructing **G** using breed-specific allele frequencies was not always advantageous on crossbred or multi-breed evaluations but for accounting for compatibility of pedigree and genomic information was crucial (Kluska et al., 2021; Lourenco et al., 2016; Makgahlela et al., 2013). The implementation of this strategy in multi-breed beef evaluations has been scarcely explored in the literature. Nevertheless, these approaches primarily focus on enhancing the robustness and

predictive accuracy of genomic information in genetic evaluations. For example, accounting for different allele frequencies was not found to be effective.

Indirect Predictions (IP)

Indirect predictions (IP) are genomic predictions obtained as the sum of SNP effects weighted by gene content (Tsuruta et al., 2021). These SNP effects are derived from genomic estimated breeding values (GEBV) computed in genomic evaluations. Different to direct genomic predictions, which require animals to have recorded phenotypes, IP allow for the estimation of genomic merit in animals that are genotyped but lack phenotypic records or progeny information(Tsuruta et al., 2021). This approach is particularly useful for young selection candidates or animals in commercial herds that will not be included in official genetic evaluations (Londoño-Gil et al., 2023; Tsuruta et al., 2021).

To obtain IP, GEBV must be first calculated, followed by the computation of SNP effects, which are used to estimate IP for young genotyped animals. The SNP effects needed for IP are calculated as proposed by Wang et al. (2012) using the equation:

$$\hat{\mathbf{a}}|\hat{\mathbf{u}} = (1-\alpha)b\frac{1}{k}\mathbf{Z}'\mathbf{G}^{\text{tunned}-1}\hat{\mathbf{u}},$$

where $\hat{\mathbf{a}}$ is a vector of SNP effects, and $\hat{\mathbf{u}}$ is a vector of GEBV for the old genotyped animals. $\mathbf{G}^{\text{tunned}} = b\mathbf{G}^{\text{blended}} + a$, $\mathbf{G}^{\text{blended}} = (1-\alpha)\mathbf{G} + \alpha\mathbf{A_{22}}$, **G** is blended to be invertible with a small percentage ($\alpha = 0.05$) of the pedigree relationship matrix among genotyped animals, a and b are the tuning parameters, calculated as in Christensen et al. (2012) and solved the system of equations $\text{Avg}\left(\text{diag}\left(\mathbf{G}^{\text{blended}}\right)\right)b + a = \text{Avg}\left(\text{diag}(\mathbf{A}_{22})\right)$, $\text{Avg}\left(\text{offdiag}\left(\mathbf{G}^{\text{blended}}\right)\right)b + a = \text{Avg}(\text{offdiag}(\mathbf{A}_{22}))$, where a adds the extra average relationship and at the same time models the difference in μ from the pedigree base to the genomic base; and the b considers the reduction in variance due to drift. Additionally, *k* is = $2\sum_{j=1}^{n} p_j(1-p_j)$.

Once the SNP effects are estimated, IP are obtained as in Lourenco et al.

(2018), with $\hat{\mathbf{u}}_{\mathrm{IP}} = \hat{\mu}_g + \mathbf{Z}_{\mathrm{v}}\hat{\mathbf{a}}$, where $\hat{\mathbf{u}}_{\mathrm{IP}}$ is a vector of estimated IP, $\mathbf{Z}_{\mathrm{v}}$ is the matrix of SNP content for the young genotyped animals, centered by the same allelic frequencies as the original **Z**, $\hat{\mathbf{a}}$ is a vector of SNP effects, and $\hat{\mu}_g$ is the weighted average of GEBV for the old genotyped animals, computed as: $\hat{\mu}_g = \mathbf{1}'\mathbf{G}^{-1}\hat{\mathbf{u}}$, where $\mathbf{1}$ is a vector of 1's. Adding $\hat{\mu}_g$ is necessary to guarantee that the IP is on the same scale as the GEBV form the official evaluations, considering the tuning of **G** to align with **A** (Lourenco et al., 2018).

However, if the model includes MF, no tunning is required and the mean $\hat{\mu}_g = 0$ (Legarra et al., 2024)

IP offer several advantages in genomic selection, particularly in large-scale evaluations and breeding programs with limited phenotypic data. First, IP help reduce the computational burden, as many genotyped animals are young and do not have phenotypic records, meaning they do not need to be included in official genomic evaluations. Excluding these animals reduces computational costs and the complexity of routine evaluations (Tsuruta et al., 2021). Furthermore, IP facilitate the early identification of superior candidates, as genomic predictions for young animals become available before they accumulate phenotypic records or progeny data, enabling earlier selection decisions. This is particularly beneficial in populations where genomic evaluations are conducted infrequently.

Including animals with incomplete pedigrees in official evaluations can decrease accuracy and increase inflation and bias in GEBV (Lourenco et al., 2018). If IP are used instead, genomic predictions can be provided without negatively affecting the accuracy and stability of GEBV (Lourenco et al., 2018). In regions or breeds where the number of phenotyped and genotyped animals is limited; IP provide the genomic selection benefits to animals that would otherwise be excluded from evaluations.

An important aspect of implementing IP effectively is the establishment of a robust reference population, which must represent all the chromosome segments segregating within the target population (Lourenco et al., 2015;Tsuruta et al., 2021). This is particularly important in breeds with limited phenotypic and genotypic data, such as Brahman, Guzerat, and Tabapua, where constructing adequate reference

populations presents substantial challenges (Hayes et al., 2023) Therefore, acquiring IP for these breeds through a broader reference population, as in the multi-breed approach may improve the accuracy of genomic selection for breeds with limited data.

1.3.6. Accuracy of genomic predictions and IP:

In animal breeding programs, the accuracy or credibility of the predicted breeding values is essential for selection decisions. Henderson (1984) demonstrated that the accuracy of estimated breeding values (EBVs) can be determined by the prediction error variance (PEV), which is typically calculated by inverting the coefficient matrix (**C**) of the BLUP mixed model equations (MME). If **C** is known, PEV is calculated directly from the diagonal elements (c^{ii}) of the inverse $\mathbf{C}^{-1}$. This can be done using pedigree information or genomics (Garcia et al., 2022).

The accuracy of the EBV for an individual i is calculated as in Gorjanc et al. (2015) :

$$r_i = \sqrt{1 - \frac{Var(a_i - \hat{a}_i)}{\sigma_{\hat{u}}^2}},$$

with $Var(a_i - \hat{a}_i)$ as the variance of prediction errors of the EBV of animal i ($c^{ii}\sigma_e^2$ or PEV), and σ_u^2 is the additive genetic variance. This accuracy measure quantifies the magnitude of prediction error variance (PEV) relative to the base additive genetic variance. It is commonly employed to assess how EBVs might change as more information becomes available. This is especially important in breeding programs with overlapping generations, such as in cattle (Gorjanc et al., 2015). This accurate value is an expected theorical value.

A similar approach has been proposed for the accuracy of indirect predictions (IP). Strandén & Christensen, (2011) showed that IP accuracies can be calculated by evaluating the prediction error covariances (PEC) of SNP effects, and this method ensures that the accuracies of both GEBVs and IP align correctly when back-solving SNP PEC under ssGBLUP. In this case, from **C** is extracted a submatrix that corresponds to the genotyped animals $\mathbf{C}^{\mathbf{u}_2\mathbf{u}_2}$, which contains the PEC of their EBVs, $\hat{\mathbf{u}}_2$, Var($\mathbf{u}$ - $\hat{\mathbf{u}}_2$). Then after back-solving SNP effects estimates as described before,

the prediction error variances of SNP estimates are obtained as in Aguilar et al. (2019) and Garcia et al. (2022) as:

$$\mathrm{Var}(\hat{a}) = (1-\alpha)b\,\frac{1}{2\sum p_i(1-p_i)}\left(\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z}\sigma_u^2 - \mathbf{Z}'\mathbf{G}^{-1}\mathbf{C}^{\mathbf{u}_2\mathbf{u}_2}\mathbf{Z}\right)\frac{1}{2\sum p_i(1-p_i)}b(1-\alpha),$$

Once this variance is available, accuracy for IP for animal j ($\mathrm{ACC}_{\mathrm{IP}_j}$) can be calculated as:

$$\mathrm{ACC}_{\mathrm{IP}_j} = \sqrt{\frac{\mathbf{z}_j \mathrm{Var}(\hat{a})\,\mathbf{z}_j'}{\sigma_u^2}}$$

This theoretical accuracy is important in animal breeding and genetics because it measures the potential for response to selection in a breeding program. Furthermore, it is crucial to obtain accurate predictions in Brazilian beef production systems and to account for this when a multi-breed approach is used. Ensuring high accuracy can enhance selection decisions for breeders, optimize genetic progress, and improve overall production efficiency.

1.4 REFERENCES

Aguilar, I., Legarra, A., Cardoso, F., Masuda, Y., Lourenco, D., & Misztal, I. (2019). Frequentist p-values for large-scale-single step genome-wide association, with an application to birth weight in American Angus cattle. **Genetics Selection Evolution**, 51(1). https://doi.org/10.1186/s12711-019-0469-3

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., & Lawlor, T. J. (2010). 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(2), 743–752. https://doi.org/10.3168/jds.2009-2730

Bermann, M., Aguilar, I., Lourenco, D., Misztal, I., & Legarra, A. (2023). Reliabilities of estimated breeding values in models with metafounders. **Genetics Selection Evolution**, 55(1). https://doi.org/10.1186/s12711-023-00778-2

Bernardes, P. A., Grossi, D. A., Savegnago, R. P., Buzanskas, M. E., Ramos, S. B., Romanzini, E. P., Guidolin, D. G. F., Bezerra, L. A. F., Lôbo, R. B., & Munari, D. P. (2016). Population structure of Tabapuã beef cattle using pedigree analysis. **Livestock Science**, 187, 96–101. https://doi.org/10.1016/j.livsci.2016.03.002

Bernardes, P. A., Grossi, D. A., Savegnago, R. P., Buzanskas, M. E., Urbinati, I., Bezerra, L. A. F., Lôbo, R. B., & Munari, D. P. (2015). Estimates of genetic parameters and genetic trends for reproductive traits and weaning weight in Tabapuã cattle. **Journal of Animal Science**, 93, 5175–5185. https://doi.org/10.2527/jas2015-9212

Berry, D. P., Garcia, J. F., & Garrick, D. J. (2016). Development and implementation of genomic predictions in beef cattle. **Animal Frontiers**, 6(1), 32–38. https://doi.org/10.2527/af.2016-0005

Brito, L. C., Peixoto, M. G. C. D., Carrara, E. R., Fonseca e Silva, F., Ventura, H. T., Bruneli, F. A. T., & Lopes, P. S. (2020). Genetic parameters for milk, growth, and reproductive traits in Guzerá cattle under tropical conditions. **Tropical Animal Health and Production**, 52(5), 2251–2257. https://doi.org/10.1007/s11250-020-02255-0

Brito, L. F., Clarke, S. M., McEwan, J. C., Miller, S. P., Pickering, N. K., Bain, W. E., Dodds, K. G., Sargolzaei, M., & Schenkel, F. S. (2017). Prediction of genomic breeding values for growth, carcass and meat quality traits in a multi-breed sheep population using a HD SNP chip. **BMC Genetics**, 18(1). https://doi.org/10.1186/s12863-017-0476-8

Callister, A. N., Bermann, M., Elms, S., Bradshaw, B. P., Lourenco, D., & Brawner, J. T. (2022). Accounting for population structure in genomic predictions of Eucalyptus globulus. **G3: Genes**, Genomes, Genetics, 12(9). https://doi.org/10.1093/g3journal/jkac180

Carvalho, M. E., Baldi, F. S., Alexandre, P. A., Santana, M. H. A., Ventura, R. V., Bueno, R. S., Bonin, M. N., Rezende, F. M., Coutinho, L. L., Eler, J. P., & Ferraz, J. B. S. (2019). Genomic regions and genes associated with carcass quality in Nelore cattle. **Genetics and Molecular Research**, 18(1), 1–15. https://doi.org/10.4238/gmr18226

Chen, L., Schenkel, † F, Vinsky, M., Crews, D. H., & Li, C. (2013). Accuracy of predicting genomic breeding values for residual feed intake in Angus and Charolais beef cattle. **Journal of Animal Science**, 91, 4669–4678. https://doi.org/10.2527/jas2013-5715

Chiaia, H. L. J., de Lemos, M. V. A., Venturini, G. C., Aboujaoude, C., Berton, M. P., Feitosa, F. B., Carvalheiro, R., Albuquerque, L. G., de Oliveira, H. N., & Baldi, F. (2015). Genotype × environment interaction for age at first calving, scrotal circumference, and yearling weight in Nellore cattle using reaction norms in multitrait random regression models. **Journal of Animal Science**, 93(4), 1503–1510. https://doi.org/10.2527/jas.2014-8217

Christensen, O. F., Madsen, P., Nielsen, B., Ostersen, T., & Su, G. (2012). Single-step methods for genomic evaluation in pigs. **Animal**, 6(10), 1565–1571. https://doi.org/10.1017/S1751731112000742

Christensen, O., & Lund, M. (2010). Genomic relationship matrix when some animals are not genotyped. **Genetics Selection Evolution**, 42(2), 1–8.

Crawford, A. M. (2001). A Review of QTL experiment in sheep. Proc. Assoc. **Advmt. Breed. Genet.,** 14, 33–38.

Cunha, G., Paulo, M., & Nogueira Biscola, H. (2023). Anuário CiCarne da cadeia produtiva da carne bovina - 2023. www.embrapa.br/fale-conosco/sac

de Souza, A. C. B., Do Egito, A. A., Peripolli, V., & McManus, C. M. (2021). Bovine landscape genetics in brazil. **Scientia Agricola**, 79(1), 1–8. https://doi.org/10.1590/1678-992x-2020-0142

Delgado, J. V., Molina, A., Camacho2, M. E., & Rodero, A. (1995). EVOLUCION DE LA METODOLOGIA DE LA MEJORA GENETICA EN EL VACUNO DE CARNE. **Archivos de Zootecnia**, 44, 111–121.

Ferraz, J. B. S., & Felício, P. E. de. (2010). Production systems - An example from Brazil. In **Meat Science** (Vol. 84, Issue 2, pp. 238–243). https://doi.org/10.1016/j.meatsci.2009.06.006

Filho, E. da S., da Silva, M. H., Campelo, J. E. G., Harada, M. L., & DeRosia, M. R. (2012). Genetic characterization of a bovine breed (Tabapua) under artificial selection located in the Amazon region. **American Journal of Animal and Veterinary Sciences**, 7(4), 226–232. https://doi.org/10.3844/ajavssp.2012.226.232

Fisher, R. A. (1918). The correlation between Relatives on the Supposition of Mendelian Inheritance. **Transations of the Royal Society of Edinburg**, 52(1918), 399–433.

FISHER, R. A. (1936). the Use of Multiple Measurements in Taxonomic Problems. **Annals of Eugenics**, 7(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Garcia, A., Aguilar, I., Legarra, A., Tsuruta, S., Misztal, I., & Lourenco, D. (2022). Theoretical accuracy for indirect predictions based on SNP effects from single-step GBLUP. **Genetics Selection Evolution**, 54(1), 66. https://doi.org/10.1186/s12711-022-00752-4

Garcia-Baccino, C. A., Legarra, A., Christensen, O. F., Misztal, I., Pocrnic, I., Vitezica, Z. G., & Cantet, R. J. C. (2017). Metafounders are related to F st fixation indices and reduce bias in single-step genomic evaluations. **Genetics Selection Evolution**, 49(1), 1–14. https://doi.org/10.1186/s12711-017-0309-2

Garrick, D. J. (2011). The nature, scope and impact of genomic prediction in beef cattle in the United States. **Genetics Selection Evolution**, 43(1), 17.

Georges, M., Nielsen, D., Mackinnon, M., Mishra, A., Okimoto, R., Pasquino, A. T., … Hoeschele, I. (1995). **Mapping Quantitative Trait Loci Controlling Milk Production in Dairy Cattle by Exploiting Progeny Testing. Genetics Society of America**, 139 (February), 907–920.

Goddard, M. E., Hayes, B. J., & Meuwissen, T. H. E. (2011). Genomic selection in livestock populations. **Genet. Res., Camb**., (2010), 413–421. https://doi.org/10.1017/S0016672310000613

Gorjanc, G., Bijma, P., & Hickey, J. M. (2015). Reliability of pedigree-based and genomic evaluations in selected populations. **Genetics Selection Evolution**, 47(1). https://doi.org/10.1186/s12711-015-0145-1

Granado-Tajada, I., Legarra, A., & Ugarte, E. (2020). Exploring the inclusion of genomic information and metafounders in Latxa dairy sheep genetic evaluations. **Journal of Dairy Science**, 103(7), 6346–6353. https://doi.org/10.3168/jds.2019-18033

Gualdrón Duarte, J. L., Cantet, R. J., Bates, R. O., Ernst, C. W., Raney, N. E., & Steibel, J. P. (2014). Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. **BMC Bioinformatics**, 15(1), 246. https://doi.org/10.1186/1471-2105-15-246

Guinan, F. L., Wiggans, G. R., Norman, H. D., Dürr, J. W., Cole, J. B., Van Tassell, C. P., Misztal, I., & Lourenco, D. (2023). Changes in genetic trends in US dairy cattle since the implementation of genomic selection. **Journal of Dairy Science**, 106(2), 1110–1129. https://doi.org/10.3168/jds.2022-22205

Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection**. BMC Bioinformatics**, 12(186), 1–12. https://doi.org/10.1186/1471-2105-12-186

Hartley, H. O., & Rao, J. N. K. (2014). Maximum-Likelihood Estimation for the Mixed Analysis of Variance Model. **Biometrika**, 54(1), 93. https://doi.org/10.2307/2333854

Harvey, W. (1960). Least squares of data with unequal subclass numbers. (USDA, Ed.; 1 ed). **National Agricultural Library**.

Hayes, B. J., Copley, J., Dodd, E., Ross, E. M., Speight, S., & Fordyce, G. (2023). Multi-breed genomic evaluation for tropical beef cattle when no pedigree information is available. **Genetics Selection Evolution**, 55(1). https://doi.org/10.1186/s12711-023-00847-6

Hayes, B. J., Corbet, N. J., Allen, J. M., Laing, A. R., Fordyce, G., Lyons, R., McGowan, M. R., & Burns, B. M. (2019). Towards multi-breed genomic evaluations for female fertility of tropical beef cattle. **Journal of Animal Science**, 97(1), 55–62. https://doi.org/10.1093/jas/sky417

Hazel, L. N. (1943). **The genetic basis for constructing selection indexes**. Iowa State College, November, 476–490.

Henderson, C. R. (1953). Estimation of Variance and Covariance Components. **Biometrics**, 9(2), 226–252.

Henderson, C. R. (1963). Selection Index and Expected Genetic Advance. Statical **Genetics And Plant Breeding**, 141–163.

Henderson, C. R. (1973). Sire evaluation and genetic trends**. Journal of Animal Acience**, 1973(1), 10–41.

Henderson, C. R. (1976). A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. **Biometrics**, 32(1), 69–83.

Henderson, C. R. (1984). **Applications of Linear Models in Animal Breeding** (1st ed.). Guelph University.

Henderson, C. R., & Quaas, R. L. (1976). Multiple trait evaluation using relatives' records. **Journal of Animal Acience**, 43(6), 1188–1197.

Hidalgo, J., Lourenco, D., Tsuruta, S., Bermann, M., Breen, V., & Misztal, I. (2023). Derivation of indirect predictions using genomic recursions across generations in a broiler population. **Journal of Animal Science**, 101. https://doi.org/10.1093/jas/skad355

Hozé, C., Fritz, S., Phocas, F., Boichard, D., Ducrocq, V., & Croiseau, P. (2014). Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population. **Journal of Dairy Science**, 97(6), 3918–3929. https://doi.org/10.3168/jds.2013-7761

https://breeds.okstate.edu/cattle/. (2025). **Breeds of Livestock**. https://breeds.okstate.edu/cattle/

IBGE Brazilian Institute of Geography and Statistics. (2024). Rebanho de Bovinos (Bois e Vacas). **Rebanho de Bovinos (Bois e Vacas)** . https://www.ibge.gov.br/explica/producao-agropecuaria/bovinos/br

Junqueira, V. S., Lopes, P. S., Lourenco, D., Silva, F. F. e., & Cardoso, F. F. (2020). Applying the Metafounders Approach for Genomic Evaluation in a Multibreed Beef Cattle Population. **Frontiers in Genetics**, 11(December), 1–10. https://doi.org/10.3389/fgene.2020.556399

Karaman, E., Su, G., Croue, I., & Lund, M. S. (2021). Genomic prediction using a reference population of multiple pure breeds and admixed individuals. **Genetics Selection Evolution**, 53(1). https://doi.org/10.1186/s12711-021-00637-y

Khansefid, M., Goddard, M. E., Haile-Mariam, M., Konstantinov, K. V., Schrooten, C., de Jong, G., Jewell, E. G., O'Connor, E., Pryce, J. E., Daetwyler, H. D., & MacLeod, I. M. (2020). Improving Genomic Prediction of Crossbred and Purebred Dairy Cattle. **Frontiers in Genetics**, 11. https://doi.org/10.3389/fgene.2020.598580

Kluska, S., Masuda, Y., Ferraz, J. B. S., Tsuruta, S., Eler, J. P., Baldi, F., & Lourenco, D. (2021). Metafounders May Reduce Bias in Composite Cattle Genomic Predictions. **Frontiers in Genetics**, 12(August). https://doi.org/10.3389/fgene.2021.678587

Kluska, S., Olivieri, B. F., Bonamy, M., Chiaia, H. L. J., Feitosa, F. L. B., Berton, M. P., Peripolli, E., Lemos, M. V. A., Tonussi, R. L., Lôbo, R. B., Magnabosco, C. de U., Di Croce, F., Osterstock, J., Pereira, A. S. C., Munari, D. P., Bezerra, L. A., Lopes, F. B., & Baldi, F. (2018). Estimates of genetic parameters for growth, reproductive, and carcass traits in Nelore cattle using the single step genomic BLUP procedure. **Livestock Science**, 216, 203–209. https://doi.org/10.1016/j.livsci.2018.08.015

Koufariotis, L., Hayes, B. J., Kelly, M., Burns, B. M., Lyons, R., Stothard, P., Chamberlain, A. J., & Moore, S. (2018). Sequencing the mosaic genome of Brahman cattle identifies historic and recent introgression including polled. **Scientific Reports**, 8(1), 17761. https://doi.org/10.1038/s41598-018-35698-5

Kudinov, A. A., Koivula, M., Aamand, G. P., Strandén, I., & Mäntysaari, E. A. (2022). Single-step genomic BLUP with many metafounders. **Frontiers in Genetics**, 13. https://doi.org/10.3389/fgene.2022.1012205

Legarra, A., Aguilar, I., & Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. **Journal of Dairy Science**, 92(9), 4656–4663. https://doi.org/10.3168/jds.2009-2061

Legarra, A., Bermann, M., Mei, Q., & Christensen, O. F. (2024a). Estimating genomic relationships of metafounders across and within breeds using maximum likelihood, pseudo-expectation–maximization maximum likelihood and increase of relationships. **Genetics Selection Evolution**, 56(1). https://doi.org/10.1186/s12711-024-00892-9

Legarra, A., Bermann, M., Mei, Q., & Christensen, O. F. (2024b). Redefining and interpreting genomic relationships of metafounders. **Genetics Selection Evolution**, 56(1). https://doi.org/10.1186/s12711-024-00891-w

Legarra, A., Christensen, O. F., Aguilar, I., & Misztal, I. (2014). Single Step, a general approach for genomic selection. **Livestock Science**, 166(August 2014), 54–65. https://doi.org/10.1016/j.livsci.2014.04.029

Legarra, A., Christensen, O. F., Vitezica, Z. G., Aguilar, I., & Misztal, I. (2015). Ancestral Relationships Using Metafounders: Finite Ancestral Populations and Across Population Relationships. **Genetics**, 200(2), 455–468. https://doi.org/10.1534/genetics.115.177014

Legarra, A., Lourenco, D., & Vitezica, Z. (2024). **Bases for Genomic** Prediction (pp. 1–167).

Legarra, A., & Reverter, A. (2018). Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. **Genetics Selection Evolution**, 50(1), 53. https://doi.org/10.1186/s12711-018-0426-6

Lemos, M. V. A., Chiaia, H. L. J., Berton, M. P., Feitosa, F. L. B., Aboujaoude, C., Venturini, G. C., Oliveira, H. N., Albuquerque, L. G., & Baldi, F. (2015). Reaction norms for the study of genotype-environment interaction for growth and indicator traits of sexual precocity in Nellore cattle. **Genetics and Molecular Research**, 14(2), 7151–7162. https://doi.org/10.4238/2015.June.29.9

Londoño-Gil, M., Cardona-Cifuentes, D., Espigolan, R., Peripolli, E., Lôbo, R. B., Pereira, A. S. C., Aguilar, I., & Baldi, F. (2023). Genomic evaluation of commercial herds with different pedigree structures using the single-step genomic BLUP in Nelore cattle. **Tropical Animal Health and Production**, 55(2). https://doi.org/10.1007/s11250-023-03508-4

Londoño-Gil, M., López-Correa, R., Aguilar, I., Magnabosco, C. U., Hidalgo, J., Bussiman, F., Baldi, F., & Lourenco, D. (2024). Strategies for genomic predictions of an indicine multi-breed population using single-step GBLUP. **Journal of Animal Breeding and Genetics**. https://doi.org/10.1111/jbg.12882

Lopes, M. S., Bovenhuis, H., Hidalgo, A. M., Van Arendonk, J. A. M., Knol, E. F., & Bastiaansen, J. W. M. (2017). Genomic selection for crossbred performance accounting for breed-specific effects. **Genetics Selection Evolution**, 49(1). https://doi.org/10.1186/s12711-017-0328-z

López-Correa, R. D., Legarra, A., & Aguilar, I. (2021). Modelling missing pedigree with metafounders and validating ssGBLUP evaluation for Uruguayan Holstein dairy cattle R.D. **World Congress on Genetics Applied to Livestock Production** (WCGALP 2022), April, 1–4.

Lourenco, D. A. L., Legarra, A., Tsuruta, S., Moser, D., Miller, S., & Misztal, I. (2018). Tuning Indirect Predictions Based on SNP Effects from Single-Step GBLUP. **Interbull Bulletin**, 53, 48–53. https://journal.interbull.org/index.php/ib/article/view/1448/1511

Lourenco, D. A. L., Tsuruta, S., Fragomeni, B. O., Chen, C. Y., Herring, W. O., & Misztal, I. (2016). Crossbreed evaluations in single-step genomic best linear unbiased predictor using adjusted realized relationship matrices. **Journal of Animal Science**, 94(3), 909–919. https://doi.org/10.2527/jas2015-9748

Lourenco, D. A. L., Tsuruta, S., Fragomeni, B. O., Masuda, Y., Aguilar, I., Legarra, A., Bertrand, J. K., Amen, T. S., Wang, L., Moser, D. W., & Misztal, I. (2015). Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. **Journal of Animal Science**, 93(6), 2653–2662. https://doi.org/10.2527/jas.2014-8836

Lourenco, D., Legarra, A., Tsuruta, S., Moser, D., Miller, S., & Misztal, I. (2018). Tuning indirect predictions based on SNP effects from ssGBLUP.

Lourenco, D., Tsuruta, S., Masuda, Y., Berman, M., Legarra, A., & Misztal, I. (2022, July 3). Recent updates in the BLUPF90 software suite. **World Congress on Genetics Applied to Livestock Production**: 3 - 8 July.

Lush, J. L. (1935). The inheritance of productivity in farm livestock Pt. V. Discussion of preceding contributions. **Emp. J. Exp. Agric**, 3, 25–30.

Lush, J. L. (1947). **Family Merit and Individual Merit as Bases for Selection**. Part I. The American Naturalist, 81(799), 241–261. https://doi.org/10.1086/681988

Macedo, F. L., Astruc, J. M., Meuwissen, T. H. E., & Legarra, A. (2022). Removing data and using metafounders alleviates biases for all traits in Lacaune dairy sheep predictions. **Journal of Dairy Science**, 105(3). https://doi.org/10.3168/jds.2021-20860

Makgahlela, M. L., Strandén, I., Nielsen, U. S., Sillanpää, M. J., & Mäntysaari, E. A. (2013). The estimation of genomic relationships using breedwise allele frequencies among animals in multibreed populations. **Journal of Dairy Science**, 96(8), 5364–5375. https://doi.org/10.3168/jds.2012-6523

Makgahlela, M. L., Strandén, I., Nielsen, U. S., Sillanpää, M. J., & Mäntysaari, E. A. (2014). Using the unified relationship matrix adjusted by breed-wise allele frequencies in genomic evaluation of a multibreed population**. Journal of Dairy Science**, 97(2), 1117–1127. https://doi.org/10.3168/jds.2013-7167

Martínez, R. A., Dassonneville, R., Bejarano, D., Jimenez, A., Even, G., Mészáros, G., & Sölkner, J. (2016). Direct and maternal genetic effects on growth, reproduction, and ultrasound traits in zebu Brahman cattle in Colombia. **Journal of Animal Science**, 94(7), 2761–2769. https://doi.org/10.2527/jas.2016-0453

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, 157(4), 1819–1829. https://doi.org/11290733

Misztal, I., Aguilar, I., Lourenco, D., Ma, L., Steibel, J. P., & Toro, M. (2021). Emerging issues in genomic selection. **Journal of Animal Science**, 99(6). https://doi.org/10.1093/jas/skab092

Misztal, I., Steyn, Y., & Lourenco, D. A. L. (2022). Genomic evaluation with multibreed and crossbred data. **JDS Communications**, 3(2), 156–159. https://doi.org/10.3168/jdsc.2021-0177

Misztal, I., Tsuruta, S., Lourenco, D. a L., Aguilar, I., Legarra, A., & Vitezica, Z. (2015). **Manual for BLUPF90 family of programs**. 125.

Misztal, I., Tsuruta, S., Lourenco, D., Aguilar, I., Legarra, A., & Vitezica, Z. (2014). **BLUPF90 family of programs**. In Athens: University of Georgia (p. 125). http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all2.pdf

Moghaddar, N., Swan, A. A., & Van Der Werf, J. H. J. (2014). Genomic prediction of weight and wool traits in a multi-breed sheep population. **Animal Production Science**, 54(5), 544–549. https://doi.org/10.1071/AN13129

Mrode, R., Ojango, J. M. K., Okeyo, A. M., & Mwacharo, J. M. (2019). Genomic Selection and Use of Molecular Tools in Breeding Programs for Indigenous and Crossbred Cattle in Developing Countries: Current Status and Future Prospects. **Frontiers in Genetics**, 9(JAN). https://doi.org/10.3389/fgene.2018.00694

Olson, K. M., VanRaden, P. M., & Tooker, M. E. (2012). Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. **Journal of Dairy Science**, 95(9), 5378–5383. https://doi.org/10.3168/jds.2011-5006

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. **Biometrika**, 58(3), 545–554. https://doi.org/10.1093/biomet/58.3.545

Peixoto, M. G. C. D., Carvalho, M. R. S., Egito, A. A., Steinberg, R. S., Bruneli, F. Â. T., Machado, M. A., Santos, F. C., Rosse, I. C., & Fonseca, P. A. S. (2021). Genetic diversity and population genetic structure of a guzerá (Bos indicus) meta-population. **Animals**, 11(4). https://doi.org/10.3390/ani11041125

Pires, B. V., de Freitas, L. A., da Silva, G. V., Mendonça, G. G., Savegnago, R. P., de Lima, M. L. P., El Faro, L., dos Santos Gonçalves Cyrillo, J. N., & de Paz, C. C. P. (2020). Maternal-offspring behavior of guzerat beef cattle. **Pesquisa Agropecuaria Brasileira**, 55.

R Core Team. (2023). **R: A Language and Environment for Statistical Computing** (4.3.0). R Foundation for Statistical Computing. https://www.r-project.org

Rao, D. P., & Edwards, L. L. (1971). On the Diffusion Model o f Mao and Toor A n Experimental Investigation of a Porous Medium Model with Nonuniform Pores. **AIChE Journal**, 17(5), 2–3.

Santana, M. L., Pereira, R. J., Bignardi, A. B., Ayres, D. R., Menezes, G. R. O., Silva, L. O. C., Leroy, G., Machado, C. H. C., Josahkian, L. A., & Albuquerque, L. G. (2016). Structure and genetic diversity of Brazilian Zebu cattle breeds assessed by pedigree analysis. **Livestock Science**, 187, 6–15. https://doi.org/10.1016/j.livsci.2016.02.002

Schefers, J. M., & Weigel, K. A. (2012). Genomic selection in dairy cattle: Integration of DNA testing into breeding programs. **Animal Frontiers**, 2(1), 4–9. https://doi.org/10.2527/af.2011-0032

Scott, B. A., Haile-Mariam, M., Cocks, B. G., & Pryce, J. E. (2021). How genomic selection has increased rates of genetic gain and inbreeding in the Australian national herd, genomic information nucleus, and bulls. **Journal of Dairy Science**, 104(11), 11832–11849. https://doi.org/10.3168/jds.2021-20326

Sevillano, C. A., Bovenhuis, H., & Calus, M. P. L. (2019). Genomic Evaluation for a Crossbreeding System Implementing Breed-of-Origin for Targeted Markers. **Frontiers in Genetics**, 10(MAY), 1–11. https://doi.org/10.3389/fgene.2019.00418

Smith, H. F. (1936). A discrimant function for plant selection. **Annals of Eugenics**, 7, 240–250. https://doi.org/10.1111/j.1469-1809.1936.tb02143.x

Steyn, Y., Lourenco, D. A. L., & Misztal, I. (2019). Genomic predictions in purebreds with a multibreed genomic relationship matrix. **Journal of Animal Science**, 97(11), 4418–4427. https://doi.org/10.1093/jas/skz296

Strandén, I., & Christensen, O. F. (2011). Allele coding in genomic evaluation. **Genetics Selection Evolution**, 43(1). https://doi.org/10.1186/1297-9686-43-25

Strandén, I., & Garrick, D. J. (2009). Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. **Journal of Dairy Science**, 92(6), 2971–2975. https://doi.org/10.3168/jds.2008-1929

Thornton, P. K. (2010). Livestock production: recent trends, future prospects. Philosophical **Transactions of the Royal Society B: Biological Sciences**, 365(1554), 2853–2867. https://doi.org/10.1098/rstb.2010.0134

Tier, B., Meyer, K., & Swan, A. A. (2018). On implied genetic effects, relationships and alternate allele coding. **Proceedings of the World Congress on Genetics Applied to Livestock Production**, 11, 127.

Tirados, S. P. F. (2001). La mejora genética animal en la segunda mitad del siglo XX. **Archivos De Zootecnia**, 50, 517–546.

Tonussi, R. L., Londoño-Gil, M., de Oliveira Silva, R. M., Magalhães, A. F. B., Amorim, S. T., Kluska, S., Espigolan, R., Peripolli, E., Pereira, A. S. C., Lôbo, R. B., Aguilar, I., Lourenço, D. A. L., & Baldi, F. (2021). Accuracy of genomic breeding values and predictive ability for postweaning liveweight and age at first calving in a Nellore cattle population with missing sire information. **Tropical Animal Health and Production**, 53(4), 432. https://doi.org/10.1007/s11250-021-02879-w

Torres-Vázquez, J. A., & Spangler, M. L. (2016). Genetic parameters for docility, weaning weight, yearling weight, and intramuscular fat percentage in Hereford cattle. **Journal of Animal Science**, 94(1), 21–27. https://doi.org/10.2527/jas.2015-9566

Tsuruta, S., Lourenco, D. A. L., Masuda, Y., Lawlor, T. J., & Misztal, I. (2021). Reducing computational cost of large-scale genomic evaluation by using indirect genomic prediction. **JDS Communications**, 2(6), 356–360. https://doi.org/10.3168/jdsc.2021-0097

USDA United States Department of Agriculture Foreign Agricultural Service. (2021). **The Brazilian Bovine Genetics Market and US Exports**. https://usdabrazil.org.br/wp-content/uploads/2021/03/The-Brazilian-Bovine-Genetics-Market-and-US-Exports_Brasilia_Brazil_03-01-2021-1.pdf

USDA United States Department of Agriculture Foreign Agricultural Service. (2024). **Livestock and Poultry: World Markets and Trade**. https://public.govdelivery.com/accounts/USDAFAS/subscriber/new

Utsunomiya, Y. T., Milanesi, M., Fortes, M. R. S., Porto-Neto, L. R., Utsunomiya, A. T. H., Silva, M. V. G. B., Garcia, J. F., & Ajmone-Marsan, P. (2019). Genomic clues of the evolutionary history of Bos indicus cattle. **Animal Genetics**, 50(6), 557–568. https://doi.org/10.1111/age.12836

van Grevenhof, E. M., Vandenplas, J., & Calus, M. P. L. (2019). Genomic prediction for crossbred performance using metafounders. **Journal of Animal Science**, 97(2), 548–558. https://doi.org/10.1093/jas/sky433

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. **Journal of Dairy Science**, 91(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., & Schenkel, F. S. (2009). Invited Review: Reliability of genomic predictions for North American Holstein bulls. **Journal of Dairy Science**, 92(1), 16–24. https://doi.org/10.3168/jds.2008-1514

Vercesi Filho, A. E., Faria, F. J. C., Madalena, F. E., & Josahkian, L. A. (2002). Estrutura populacional do rebanho Tabapuã registrado no Brasil. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, 54(6), 609–617. https://doi.org/10.1590/S0102-09352002000600009

Veroneze, R., Bastiaansen, J. W., Knol, E. F., Guimarães, S. E., Silva, F. F., Harlizius, B., Lopes, M. S., & Lopes, P. S. (2014). Linkage disequilibrium patterns and persistence of phase in purebred and crossbred pig (Sus scrofa) populations. **BMC Genetics**, 15(1), 126. https://doi.org/10.1186/s12863-014-0126-3

Wang, H., Misztal, I., Aguilar, I., Legarra, A., & Muir, W. M. (2012). Genome-wide association mapping including phenotypes from relatives without genotypes. **Genetics Research**, 94(2), 73–83. https://doi.org/10.1017/S0016672312000274

CHAPTER 2: STRATEGIES FOR GENOMIC PREDICTIONS OF AN INDICINE MULTI-BREED POPULATION USING SINGLE-STEP GBLUP

Londoño-Gil, Marisol[1,2*]; López-Correa, Rodrigo[3]; Aguilar, Ignacio[4]; Magnabosco, Claudio U[5]; Hidalgo, Jorge[2]; Bussiman, Fernando[2], Baldi, Fernando[1]; Lourenco, Daniela[2]

[1] Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista Júlio de Mesquita Filho, Via de Acesso Prof. Paulo Donato Castellane s/n, 14884-900 Jaboticabal, SP, Brazil
[2] Department of Animal and Dairy Science, University of Georgia, 30602 Athens, GA, USA
[3] Facultad de Veterinaria, Departamento de Genética y Mejoramiento Animal, Universidad de la República, Montevideo, Uruguay.
[4] Instituto Nacional de Investigación Agropecuaria (INIA), 11500 Montevideo, Uruguay.
[5] Embrapa Cerrados, BR-020, 18, Sobradinho, 70770-901 Brasilia, DF, Brazil

Published in the Journal of Animal Breeding and Genetics.
Doi: 10.1111/jbg.12882
Received: 22 March 2024, Accepted: 15 May 2024

ABSTRACT

Brazilian livestock breeding programs strive to enhance the genetics of beef cattle, with a strong emphasis on the Nellore breed, which has an extensive database and has achieved significant genetic progress in the last years. There are other indicine breeds that are economically important in Brazil; however, these breeds have more modest sets of phenotypes, pedigree, and genotypes, slowing down their genetic progress as their predictions are less accurate. Combining several breeds in a multi-breed evaluation could help enhance predictions for those breeds with less information available. This study aimed to evaluate the feasibility of multi-breed, single-step genomic best linear unbiased predictor genomic evaluations for Nellore, Brahman, Guzerat, and Tabapua. Multi-breed evaluations were contrasted to the single-breed

ones. Data were sourced from the National Association of Breeders and Researchers of Brazil and included pedigree (4,207,516), phenotypic (328,748), and genomic (63,492) information across all breeds. Phenotypes were available for adjusted weight at 210 and 450 days of age, and scrotal circumference at 365 days of age. Various scenarios were evaluated to ensure pedigree and genomic information compatibility when combining different breeds, including metafounders (MF) or building the genomic relationship matrix with breed-specific allele frequencies. Scenarios were compared using the linear regression method for bias, dispersion, and accuracy. The results showed that using multi-breed evaluations significantly improved accuracy, especially for smaller breeds like Guzerat and Tabapua. The validation statistics indicated that the MF approach provided accurate predictions, albeit with some bias. While single-breed evaluations tended to have lower accuracy, merging all breeds in multi-breed evaluations increased accuracy and reduced dispersion. This study demonstrates that multi-breed genomic evaluations are proper for indicine beef cattle breeds. The MF approach may be particularly beneficial for less-represented breeds, addressing limitations related to small reference populations and incompatibilities between **G** and $\mathbf{A}_{22}$. By leveraging genomic information across breeds, breeders and producers can make more informed selection decisions, ultimately improving genetic gain in these cattle populations.

Keywords: beef cattle, compatibility, LR method, metafounders, reference population.

2.1 INTRODUCTION

Brazil is the leading beef exporter and producer worldwide (USDA, 2022). The production systems usually comprise indicine breeds (*Bos taurus indicus*) adapted to tropical conditions to thrive in distinct ecosystems. The Nellore breed is the dominant beef cattle breed, making up ~80% of the roughly 200 million heads in the national herd (ANUALPEC, 2021), followed by other breeds such as Brahman, Guzerat, and Tabapua. All major breeds are part of breeding programs; however, the primary efforts have been made in the Nellore cattle, a breed with extensive recording of phenotypes, genotypes, and pedigree information. Despite the efforts made by the Brahman,

Guzerat, and Tabapua breed associations, the genomic evaluations are still in the initial phase or have a limited number of animals with genomic information, minimizing the possible gains promoted by genomic selection.

For the lesser representative breeds, i.e., those with less pedigree, phenotypic, and genomic information, it is challenging to build a large reference population with a reasonable number of high-accuracy genotyped animals (Hayes et al., 2009; Thomasen et al., 2014; VanRaden et al., 2009). A feasible alternative to increase the value of genomic information in breeds with small reference populations is to group closely related breeds into a metapopulation without adding extra costs for data recording (Steyn et al., 2019). However, incompatibilities between the genomic (**G**) and pedigree (**A**) relationship matrices can be intensified in a multi-breed population as the allele frequencies utilized to center and scale **G** are typically based on means of genotyped individuals across all populations.

Legarra et al. (2015) proposed the concept of metafounders (MF) in the context of single-step genomic best linear unbiased prediction (ssGBLUP) to ensure compatibility between **G** and **A** in the presence of multiple base populations. Metafounders are pseudo-individuals acting as founders of the populations, which can be considered simultaneously as the sire and dam of the base population. Then, this base population becomes related and can be inbred (Legarra et al., 2015). Therefore, MF has been recommended when combining genomic and pedigree information for individuals, as in ssGBLUP, and when there are several base populations combined (Legarra et al., 2015), as in the multi-breed evaluations. Another strategy to improve compatibility between **G** and A in multi-breed populations was applied by Lourenco et al. (2016), which is to have breed-specific allele frequencies when constructing **G**. The genome of multi-breed populations is a combination of genomic regions inherited from their parental breeds. Therefore, the frequency and effects of single nucleotide polymorphism (SNP) alleles can vary considerably depending on the breed that they originated from (Sevillano et al., 2019).

For closely related breeds, such as indicine beef cattle breeds (Gibbs et al., 2009), it would be expected to have high-accuracy across-breed predictions, i.e., when the predicted breed is different from the reference breed (Steyn et al., 2019). Likewise,

combining indicine breeds in a single evaluation should not decrease the prediction power because of the high level of similarity. Therefore, our objective was to evaluate the feasibility of a multi-breed ssGBLUP evaluation composed of Nellore, Brahman, Guzerat, and Tabapua breeds. To ensure the compatibility between **G** and **A** in the evaluation, we used 1) different numbers/definitions of metafounders and 2) **G** with breed-specific allele frequencies. Those two approaches were compared to a multi-breed evaluation without any specific compatibility treatment and to single-breed evaluations.

2.2 MATERIAL AND METHODS

No Animal Care and Use Committee approval was necessary for this study as the dataset was acquired from an existing database.

2.2.1 Data

The dataset, including pedigree (4,207,516), phenotypic (328,748), and genomic (63,492) information was provided by the National Association of Breeders and Researchers (*Associação Nacional de Criadores e Pesquisadores*– ANCP, Ribeirão Preto - SP, Brazil), which controls the breeding programs of Nellore (Nellore Brazil program), Brahman (Brahman Brazil program), Guzerat (Brahman Brazil program), and Tabapua (Tabapua Brazil program). Commercial and registered herds are allowed in the programs, and Brahman records are available from herds in Brazil and Bolivia. Generally, in the ANCP evaluations, these Brahman animals are evaluated together, but the ranking is done separately.

Records for adjusted weight at 210 (W210) and 450 (W450) days of age, and for scrotal circumference at 365 (SC365) days of age were used (Table 1). Phenotypic records deviating from the mean of contemporary groups ± 3 standard deviations, and contemporary groups with less than five records were removed. Contemporary groups for W210 and W450 were based on breed, farm, year and season of birth, sex, and management group classes. Contemporary groups for SC365 were defined using breed, farm, year of birth, and weaning management group classes. A pedigree depth of three was used, for a total of 15 generations, which comprehended the years from

1940 until 2022.

Nellore, Brahman, Guzerat, and Tabapua animals were genotyped with the ZBN commercial panel from Zoetis (ZBN, Zoetis, Kalamazoo) chip with 74,653 SNP (Table 2.1). The criteria for the quality control of SNP were based on minor allele frequency (MAF), call rate, and Mendelian conflicts. Thus, markers with Mendelian conflicts (1), MAF < 0.05 (29,342), call rate < 0.90 (0), sexual or mitochondrial (9,217), and monomorphic (4,098) were excluded. Quality control was performed using the preGSF90 software from the BLUPF90 family of programs (Misztal et al., 2015). Principal component analysis (PCA) based on SNP was also performed using the preGSF90 software to investigate population structure. After quality control, 36,093 SNP were used for the multi-breed evaluations.

Table 2.1. Descriptive data for adjusted weight at 210 (W210) and 450 (W450) days of age, and scrotal circumference at 365 (SC365) days of age for the Nellore Brahman, Guzerat, and Tabapua Breeds.

Trait		Nellore	Brahman	Guzerat	Tabapua	Total
W210 (kg)	N	204,019	88,099	26,632	9,998	328,748
	Mean	204.81	182.93	188.73	191.19	197.23
	SD	31.43	31.38	36.52	32.25	33.38
	Genotypes	58,574	3,102	1,389	427	63,492
	Pedigree	398,585	112,747	40,593	18,486	570,459
	Pheno & Geno	57,448	2,327	1,026	99	60,900
W450 (kg)	N	150,384	40,923	21,797	6,699	219,803
	Mean	319.29	271.24	290.95	289.64	306.63
	SD	58.21	56.62	65.33	60.29	61.85
	Genotypes	46,162	3,102	1,389	427	51,080
	Pedigree	315,097	68,119	34,432	14,086	431,173
	Pheno & Geno	44,435	1,686	985	166	47,272
SC365 (cm)	N	100,513	16,536	7,164	2,168	126,381
	Mean	22.43	20.17	21.03	21.69	22.04
	SD	2.68	2.59	2.49	2.39	2.77
	Genotypes	26,703	3,102	1,389	427	31,621
	Pedigree	248,120	40,101	17,976	7,377	313,300
	Pheno & Geno	25,206	423	342	102	26,073

N: Number of records, SD: standard deviation, Total: Descriptive statistics for the four breeds combined. Genotypes: Number of genotyped individuals. Pedigree: Number of animals in pedigree. Pheno & Geno: Number of animals with both phenotypes and genotypes.

2.2.2 Variance components and heritabilities estimation

The animal model used to estimate variance components and heritabilities for W210, W450, and SC365 through the Average Information Restricted Maximum Likelihood method is presented below:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z_1 u} + \mathbf{Z_2 m} + \mathbf{Z_3 c} + \mathbf{e},$$

where $\mathbf{y}$, $\boldsymbol{\beta}$, $\mathbf{u}$, $\mathbf{m}$, $\mathbf{c}$, and $\mathbf{e}$ are the vectors of phenotypes, fixed effects (contemporary group and age of dam classes), direct additive genetic, maternal additive genetic, maternal permanent environmental, and residual effects, respectively, and $\mathbf{X}$, $\mathbf{Z_1}$, $\mathbf{Z_2}$ and $\mathbf{Z_3}$ are the incidence matrices for the effects contained in $\boldsymbol{\beta}$, $\mathbf{u}$, $\mathbf{m}$, and $\mathbf{c}$. Maternal genetic and permanent maternal environmental effects were considered only for W210. Direct and maternal genetic effects were not correlated. Heritabilities were estimated using the best linear unbiased predictor (BLUP). Variance components were estimated without genomic information using the BLUPF90+ software from the BLUPF90 suite of programs (Lourenco et al., 2022; Misztal et al., 2014). Thus, these were estimated for each breed and the multi-breed when merging all breeds. Since it did not include genomic information, there were no estimates based on other scenarios or with MF since MF required scaled variance components to be comparable with different models.

2.2.3 Scenarios for the estimation of genomic breeding values

Different scenarios were proposed to evaluate the feasibility of multi-breed ssGBLUP evaluations for Nellore, Brahman, Guzerat, and Tabapua, including MF or building **G** with breed-specific allele frequencies. All scenarios are described in Table 2.2.

Table 2.2. Scenarios tested in the multi-breed evaluation of the Nellore, Brahman, Guzerat, and Tabapua populations from Brazil

Scenario	Description
G_0	**G** default with the current allele frequencies, constructed as VanRaden (2008)
G_{breed}	**G** matrix was centered and scaled by breed-specific allele frequencies as in Lourenco et al. (2016)
MF_2	Two metafounders: one for Nellore and one for Brahman, Guzerat, and Tabapua
MF_4	Four metafounders: Nellore, Brahman, Guzerat, and Tabapua
MF_6	Six metafounders: Commercial Nellore, Brazilian Brahman, Bolivian Brahman, Registered Nellore, Guzerat, and Tabapua
Single breed	ssGBLUP for each breed, individually

Without MF, the covariance structure among animals in ssGBLUP (Legarra et al., 2009) was defined as in Aguilar et al. (2010):

$$\mathbf{H^{-1}} = \mathbf{A^{-1}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G^{-1}} - \mathbf{A_{22}^{-1}} \end{bmatrix},$$

where **G** was constructed as in VanRaden (2008), that is, $\mathbf{G} = \mathbf{ZZ}'$, where:

$$\mathbf{Z} = (\mathbf{M} - \mathbf{P}) / \left[2 \sum_{j=1}^{n} p_j (1 - p_j) \right]^{1/2},$$

with $\mathbf{M}$ being the matrix of n SNP genotypes for each animal, with elements set to 0, 1, and 2 for the homozygote, heterozygote, and the other homozygote, respectively, which represent the count of the second allele; $\mathbf{P}$ is the matrix of twice the frequency of the second allele p at locus j (p_j). When this **G** was used for genomic evaluations, the scenario was named $\mathbf{G}_0$.

For the scenario where **G** accounted for breed-specific allele frequency ($\mathbf{G}_{breed}$), $\mathbf{G}_{breed}$ was constructed according to Lourenco et al. (2016). Considering that frequencies are different across breeds, **Z** for each breed was created as follows:

$$\mathbf{Z} = (\mathbf{M} - \mathbf{P}_{\mathrm{K}}) / \left[2 \sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{p}_{\mathrm{jK}} (1 - \mathrm{p}_{\mathrm{jK}}) \right]^{1/2},$$

where $\mathbf{P}_{\mathrm{K}}$ contains twice p_{jK}, the frequency of the second allele at the j locus, specific for the Kth breed.

The MF approach proposed by Legarra et al. (2015) was used for the further multi-breed scenarios. For this, three scenarios that considered MF were constructed based on the genomic PCA (Figure 2.1) and according to the breed (MF_2, MF_4, and MF_6, Table 2.1). The MF were related by a matrix of additive relationships $\mathbf{A}(\boldsymbol{\Gamma})$ given by a positive definite matrix $\boldsymbol{\Gamma}$, as follows:

$$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma^{\mathrm{A}} & \gamma^{\mathrm{A,B}} & \cdots & \gamma^{\mathrm{A,N}} \\ \gamma^{\mathrm{B,A}} & \gamma^{\mathrm{B}} & \cdots & \gamma^{\mathrm{B,N}} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma^{\mathrm{N,A}} & \gamma^{\mathrm{N,B}} & \cdots & \gamma^{\mathrm{N}} \end{bmatrix},$$

where, γ^{A} and γ^{B} were the relationships within MF A and B, respectively, up to N MF, and $\gamma^{\mathrm{A,B}}$ were the coefficients across MF. Those coefficients within and across base populations (MF) were estimated using genomic information from their offspring and the pedigree. The generalized least squares (GLS) method for multiple populations was used as proposed by Garcia-Baccino et al. (2017), using the GAMMAF90 program from the BLUPF90 suite of programs (Lourenco et al., 2022; Misztal et al., 2014). Then, the inverse relationship matrix with pedigree, genomics, and MF information $\mathbf{H}^{\Gamma-1}$ was calculated using the BLUP90IOD3 program from the BLUPF90 suite of programs (Lourenco et al., 2022; Misztal et al., 2014), as follows:

$$\mathbf{H}^{\Gamma-1} = \mathbf{A}^{\Gamma-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{05}^{\Gamma-1} - \mathbf{A}_{22}^{\Gamma-1} \end{bmatrix},$$

where $\mathbf{H}^{\Gamma-1}$ is the inverse matrix that combines the genomic and pedigree matrix considering MF; $\mathbf{A}^{\Gamma-1}$ is an inverse matrix of additive relationships with various MF; $\mathbf{G}_{05}^{\Gamma-1}$ is the inverse of the genomic relationship matrix of genotyped animals and $\mathbf{A}_{22}^{\Gamma-1}$ is the matrix of additive relationships between genotyped animals, considering MF.

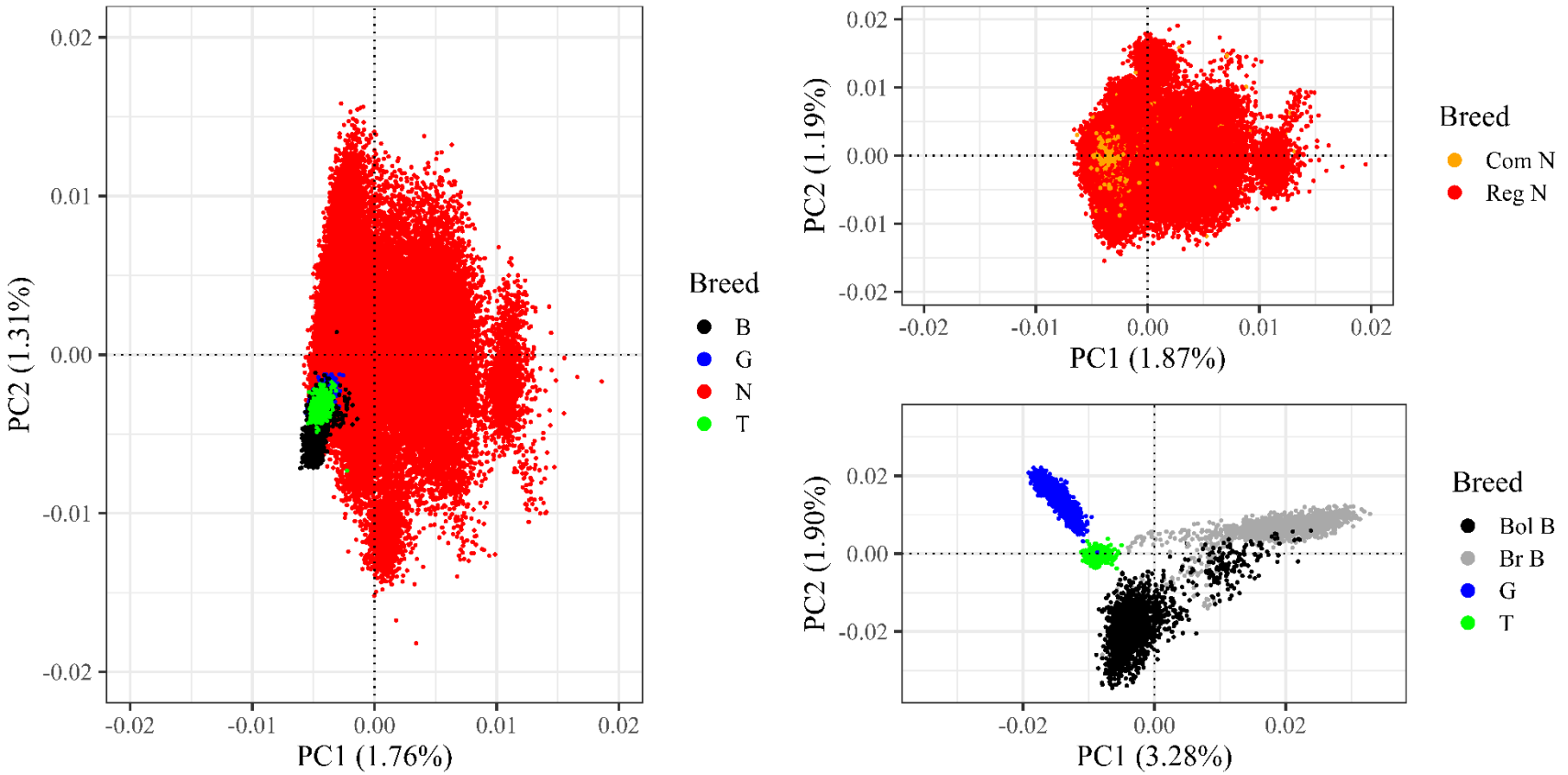


Figure 2.1 Principal components plot for the multi-breed population composed of Nellore, Brahman, Guzerat and Tabapua breeds. On the left side, the four breeds are in the same PCA; at the top of the right side, the genotyped commercial and registered Nellore animals are plotted, and finally, at the bottom of the right side are the lesser representative breeds, Brahman – from Brazil and Bolivia -, Guzerat and Tabapua. B: Brahman, G: Guzerat, N: Nellore, T: Tabapua, Com: Comercial, Reg: Registered, Bol: Bolivia, Br: Brazil

Single-breed ssGBLUP evaluations were run as an additional scenario.

2.2.4 Comparison of scenarios

All scenarios were compared using the linear regression method (LR) proposed by Legarra and Reverter (2018). The "whole" dataset included all sources of information for the animals, whereas the "partial" dataset omitted phenotypic information from 2021 onwards for W210 or 2020 for W450 and SC365. The validation animals were young, genotyped individuals without their own or progeny records in the partial set but with records in the whole set, as shown in Table 2.3. The LR estimators of accuracy, bias, and dispersion for the validation animals were computed as follows (Legarra and Reverter, 2018): $\text{accuracy} = \widehat{\text{acc}} = \sqrt{\frac{\text{cov}(\hat{u}_w,\hat{u}_p)}{(1-\bar{F})\widehat{\sigma_u^2}}}$; $\text{bias} = \hat{\Delta}_{wp} = \frac{\bar{\hat{u}}_p - \bar{\hat{u}}_w}{\sigma_u}$; $\text{dispersion} = \hat{b}_{w,p} = \frac{\text{cov}(\hat{u}_p,\hat{u}_w)}{\text{var}(\hat{u}_p)}$; where $\hat{u}$ ($\bar{\hat{u}}$) is the GEBV (average GEBV) from the partial (p) or whole (w) datasets, $\bar{F}$ is the mean inbreeding coefficient of the validation genotyped animals, $\widehat{\sigma_u^2}$ is the genetic variance for the trait, and σ_u is the square root of the genetic variance.

Table 2.3. Number of records in the whole and partial datasets and validation animals used in Linear Regression method to assess statistics for the different scenarios for adjusted weight at 210 (W210), and 450 (W450) days of age, and scrotal circumference at 365 (SC365) days of age.

Trait		Nellore	Brahman	Guzerat	Tabapua	4 Breeds
W210 (kg)	**Whole**	204,019	88,099	26,632	9,998	328,748
	Partial	187,268	84,475	26,131	9,895	307,769
	Validation	12,936	330	151	61	*
W450 (kg)	**Whole**	150,384	40,923	21,797	6,699	219,803
	Partial	126,987	37,039	21,168	6,558	191,752
	Validation	14,558	521	279	102	*
SC365 (cm)	**Whole**	100,513	16,536	7,164	2,168	126,381
	Partial	83,472	14,606	6,794	2,089	106,961
	Validation	8,615	271	152	61	*

4 breeds: All four breeds in the multi-breed evaluation. * The validation animals depended on the breed evaluated but are the sum of the four breeds.

In this study, the objective of having different scenarios was to improve the compatibility between **G** and $\mathbf{A_{22}}$ when combining multiple breeds in a single evaluation. Therefore, statistics from these two matrices and the regression coefficients of **G** on $\mathbf{A_{22}}$ (b0 and b1) were computed using in-house software programmed in Fortran 90 using Intel MKL routines. Additionally, Pearson correlations between GEBV from the single-breed ssGBLUP and all other scenarios were computed for all animals but broken into genotyped and non-genotyped animals using the R program (R Core Team, 2023).

2.3 RESULTS AND DISCUSSION

Principal component plots based on genomic information are shown in Figure 2.1. There are minimal, subtle distinctions among Brahman, Guzerat, and Tabapua. In particular, Brahman stands out, displaying a discernible separation based on its geographic origin, Brazilian or Bolivian, and was the second breed in the number of genotyped animals. Consequently, Brahman diverged more than Guzerat and Tabapua, being the latter in the middle. Although there is separation among the smaller indicine breeds, they are all close to Nellore, the most prominent breed, with over 58k genotyped individuals and a considerable variation. Although very variable, there is no separation between commercial and registered animals within the Nellore breed. This may be due to using registered bulls as parents of commercial cattle (Londoño-Gil et al., 2023).

It is well known that Brahman and Tabapua breeds originated as crossbreeds. In the case of the Brahman, some *Bos taurus indicus* cattle breeds, such as the Ongole, Guzerat, Nellore, Gyr, and Krisna, were used in repeated crossbreeding with the local American *Bos taurus* cattle (Briggs & Briggs, 1980; Koufariotis et al., 2018). On the other hand, Tabapua was the first polled zebu breed formed in Brazil that originated as a cross of the polled Nellore, Guzerat, and Gyr. The polled phenotype likely comes from Nellore or domestic cattle with European cattle introgression (Vercesi Filho et al., 2002). However, strong claims cannot be made based on these results due to the limited population size.

The variance component estimates are presented in Table 2.4. The estimates

are remarkably close among breeds for W210, except for Brahman, in which the maternal effect is larger and almost equally important as the direct effect for weaning weight. For the traits and breeds, the variance components and heritabilities estimates were similar to what is found in the literature for the *Bos taurus indicus* cattle when using single-breed or multi-breed populations (Carrara et al., 2023; da Silva Neto et al., 2020; da Silva Neto et al., 2023; Kluska et al., 2018; Londoño-Gil et al., 2022), and shown that the selection for these traits is feasible.

Table 2.4. Variance components estimates and heritabilities for the traits adjusted weight at 210 (W210) and 450 (W450) days of age, and scrotal circumference at 365 (SC365) days of age for the single- or multi-breed populations.

Breed	σ^2_d	σ^2_m	σ^2_{mpe}	σ^2_e	h^2_d	h^2_m
			W210			
Nellore	100.59	49.01	87.53	272.50	0.20 ± 0.01	0.10 ± 0.01
Brahman	70.10	57.70	72.12	337.66	0.13 ± 0.01	0.11 ± 0.01
Guzerat	107.76	21.26	55.91	376.38	0.19 ± 0.02	0.04 ± 0.01
Tabapua	98.68	49.08	50.02	300.32	0.20 ± 0.03	0.10 ± 0.02
Multi-breed	93.12	49.17	74.87	303.06	0.18 ± 0.01	0.09 ± 0.00
			W450			
Nellore	301.95	-	-	546.26	0.36 ± 0.01	-
Brahman	250.89	-	-	595.23	0.30 ± 0.02	-
Guzerat	443.62	-	-	609.97	0.42 ± 0.02	-
Tabapua	338.53	-	-	584.89	0.37 ± 0.04	-
Multi-breed	305.69	-	-	562.65	0.35 ± 0.01	-
			SC365			
Nellore	1.60	-	-	2.18	0.42 ± 0.01	-
Brahman	1.22	-	-	2.10	0.37 ± 0.03	-
Guzerat	1.05	-	-	1.98	0.35 ± 0.04	-
Tabapua	1.61	-	-	1.84	0.47 ± 0.08	-
Multi-breed	1.48	-	-	2.19	0.40 ± 0.01	-

Multi-breed: Nellore, Brahman, Guzerat, and Tabapua evaluated together in the traditional BLUP; σ^2_d: Additive genetic variance for direct effects; σ^2_m: Additive genetic variance for maternal effects; σ^2_{mpe}: Variance due to maternal permanent environmental effects; σ^2_e: Residual error variance; h^2_d: Direct heritability; h^2_m: Maternal heritability.

2.3.1 Relationships within and between Metafounders

The relationships within MF (Diagonal of **Γ**) were smaller than one, and those

between MF (off-diagonals of **Γ**) were different from zero in the three scenarios (MF_2, MF_4, and MF_6) as shown below.

$$\boldsymbol{\Gamma}_{\mathrm{MF}_2} = \begin{bmatrix} 0.71 & 0.67 \\ 0.67 & 0.69 \end{bmatrix}$$

$$\boldsymbol{\Gamma}_{\mathrm{MF}_4} = \begin{bmatrix} 0.71 & 0.66 & 0.68 & 0.67 \\ 0.66 & 0.69 & 0.68 & 0.66 \\ 0.68 & 0.68 & 0.76 & 0.69 \\ 0.67 & 0.66 & 0.69 & 0.73 \end{bmatrix}$$

$$\boldsymbol{\Gamma}_{\mathrm{MF}_6} = \begin{bmatrix} 0.72 & 0.65 & 0.67 & 0.70 & 0.68 & 0.67 \\ 0.65 & 0.71 & 0.66 & 0.64 & 0.67 & 0.65 \\ 0.67 & 0.66 & 0.70 & 0.66 & 0.68 & 0.66 \\ 0.70 & 0.64 & 0.66 & 0.71 & 0.68 & 0.67 \\ 0.68 & 0.67 & 0.68 & 0.68 & 0.76 & 0.69 \\ 0.67 & 0.65 & 0.66 & 0.67 & 0.69 & 0.73 \end{bmatrix}$$

When insufficient historical pedigree information is available, it is often assumed that the founders are unrelated and, therefore, originated from an infinite population. However, this assumption is flawed (Legarra et al., 2015). The gamma matrix is a mathematical description of the relationship between all the MF in a population. This matrix is typically estimated using a statistical method called maximum likelihood, which maximizes the probability of detecting the observed genotypes in the base population (Garcia-Baccino et al., 2017). In our study, relationships within MF ranged from 0.69 (Brahman) to 076 (Guzerat) and across MF from 0.64 (Brazilian Brahman with Registered Nellore) to 0.70 (Commercial Nellore with Registered Nellore). Hence, it was noticed that the relationships within breeds are comparable to the between breeds, showing that the base animals, within and across breeds, are related. In other words, the populations have some ancestors in common, and the founders were connected, which is expected for the *Bos taurus indicus* cattle. It was evident that the relationships of Brahman within and across MF were slightly lower, which makes sense according to the breed formation and the PCA (Figure 2.1).

According to Masuda et al. (2022), diagonal elements of **Γ** should be between 0 and 2, and off-diagonal elements should be between −1 and 1. The study results are within these ranges, and convergence issues did not occur. In addition, the matrices are similar across the different scenarios, indicating how closely related the breeds'

founders were. Similar results were reported by Kudinov et al. (2022) and Junqueira et al. (2020) using many MF in dairy and beef cattle populations.

2.3.2 Comparison of scenarios

The accuracy, bias, and dispersion of genomic predictions for the validation animals in each scenario and trait are shown in Figure 2.2. In general, when the breeds were evaluated in the single-breed approach, the accuracy was lower, and the dispersion was worse (i.e., further away from the ideal value of 1.0), especially for the Guzerat and Tabapua breeds, which have less data. In the multi-breed approach, merging all breeds increased the accuracy for most breeds and scenarios; the dispersion was closer to one, and the bias was nearer to its optimal value of zero. However, the improvement was inversely proportional to the number of genotypes and phenotypes within a breed, being trait and breed dependent. For example, for Brahman, the second breed with more data, the multi-breed evaluation was less beneficial than for Guzerat and Tabapua.

Wicki et al. (2023) observed that combining genetically close populations (same breed or related breeds) in the reference population increased the prediction accuracy as the reference population size increased when using MF. Similarly, in our study, higher prediction accuracies were obtained when combining breeds with or without MF approach (see Figure 2.2). Especially for the small breed Tabapua, the MF approach gave higher accuracies across the traits. Accuracy values with MF ranged from 0.54 (Tabapua) to 0.77 (Nellore) for W210, 0.43 (Guzerat) up to 0.71 (Brahman) for W210m, 0.59 (Tabapua) to 0.71(Nellore) for W450, and 0.39 (Guzerat) to 0.82 (Nellore) for SC365. Several studies have demonstrated the advantages of using MF in genomic predictions. For instance, Bradford et al. (2019) compared the accuracy of simulated genomic predictions in dairy cattle using MF and traditional methods to account for the base population like unknown parent groups (UPG). Metafounders improved the accuracy compared to UPG (over 5% higher accuracies), especially for low heritability traits. Kudinov et al. (2022) and Macedo et al. (2020) provide further support by demonstrating the effectiveness of MF in improving genomic prediction accuracies in red dairy cattle and sheep, respectively. These studies underline the versatility of MF across different livestock species and genetic traits.

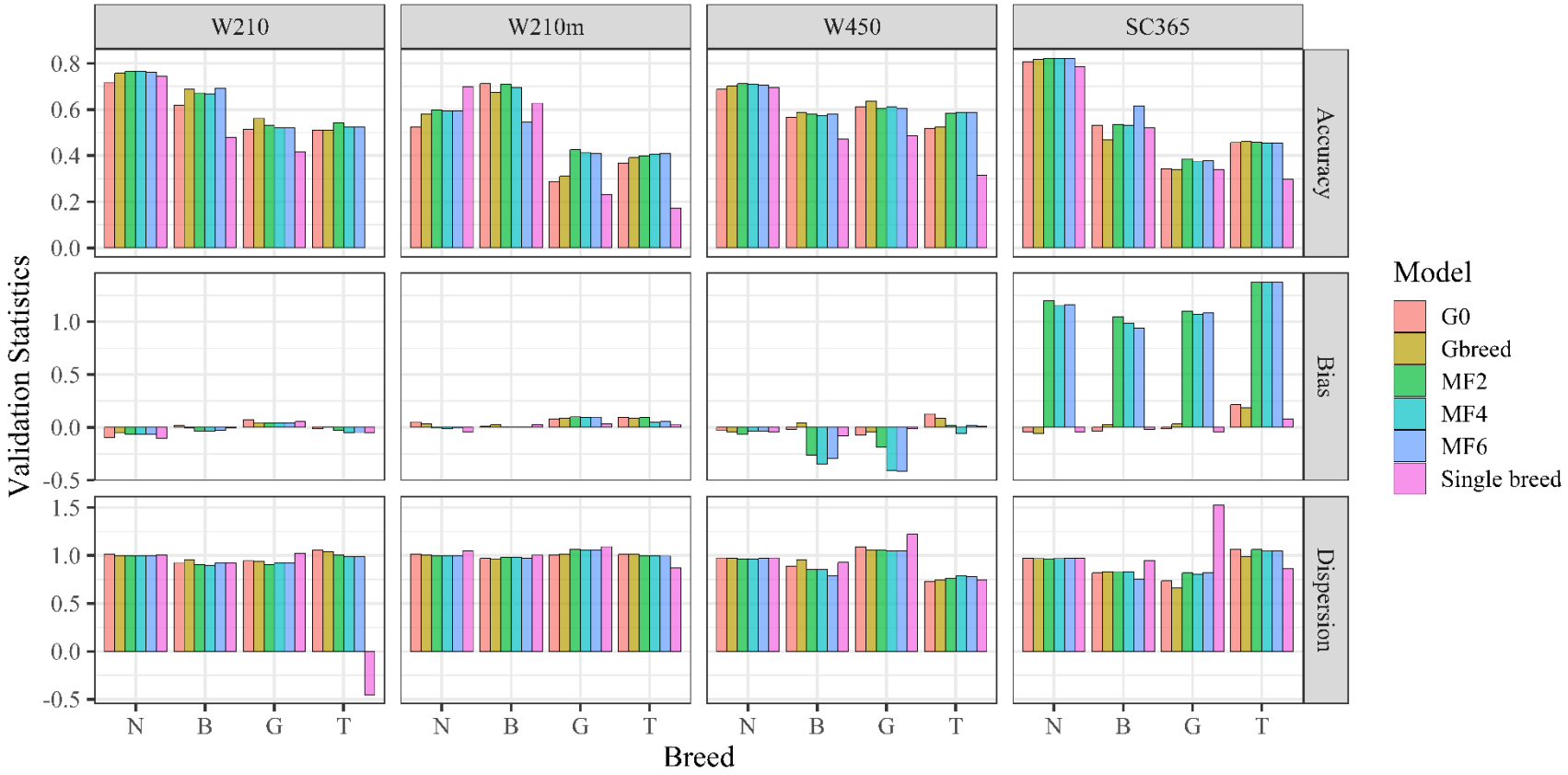


Figure 2.2. Accuracy, bias and dispersion for genomic predictions of the validation animals in each scenario and trait. N: Nellore, B Brahman, G: Guzerat, T: Tabapua. W210, W210m, W450: adjusted weight at 210 direct and maternal, and 450 days of age, respectively. SC365: scrotal circumference at 365 days of age. Tabapua accuracy for W210 was not possible to estimate.

In general, all the scenarios investigated in this study produced biased predictions (i.e., $\widehat{\Delta}_{\mathrm{wp}} \neq 0$). The expected value of bias is 0 if the model is correct; its importance resides in the fact that the estimated genetic progress is true only when predictions are unbiased (Macedo et al., 2020). With the MF approach, there was more bias for W450 in Brahman and Guzerat and for the SC365 in all breeds. It could be explained by insufficient genotyped animals with phenotypes connected with MF to estimate the MF effects properly. When data are limited, biased predictions can arise. These biases can significantly impact the perceived genetic progress, rendering them untrustworthy. Interestingly, this finding contradicts Bradford et al. (2019), who reported that using more information with UPG but not MF led to more consistent results. This suggests that the data needed for accurate prediction might vary depending on the specific method and breed employed. In this case, the $\mathbf{G}_0$ and $\mathbf{G}_{\text{breed}}$ methods showed better bias in Brahman and Guzerat for W450 and the four breeds in SC365. For the other traits, the bias was similar between the methods, and sometimes, either the $\mathbf{G}_{\text{breed}}$ or the MF approach was closer to zero. Similarly, Makgahlela et al. (2014) demonstrated that considering the ancestral origins of alleles within the population led to comparable validation biases and reliabilities, contrasting assumptions of a uniform structure within a mixed population. Further research is necessary to fully understand this interplay between data availability, methodology choice, and the resulting accuracy of genetic predictions.

In single-breed evaluations, the dispersion (i.e., slope) for W210 in Tabapua, and for SC365 in Guzerat deviated significantly from 1.0. However, when multi-breed evaluations were used, predictions were less dispersed, i.e., with estimates around one. The best estimates for small breeds, especially Guzerat and Tabapua, were obtained with the MF scenarios (MF_2, MF_4, or MF_6). Therefore, by using multi-breed evaluations, like MF, the dispersion of genomic predictions within the examined population significantly improved compared to single scenarios, as reported by several authors (Bradford et al., 2019; Fu et al., 2021; Junqueira et al., 2020; Kluska et al., 2021; Kudinov et al., 2020, 2022; Macedo et al., 2022; Masuda et al., 2021). Corresponding to these findings, Makgahlela et al. (2014) showed that when using genomic relationship matrices constructed ignoring or not, the allele frequencies (like $\mathbf{G}_0$ or $\mathbf{G}_{\text{breed}}$) across breeds gave about the same validation statistics. According to

Junqueira et al. (2020), fitting MF to account for the base population relationships for a multi-breed evaluation can contribute to estimating less under- or over-dispersed breeding values (slope close to one). Moreover, Macedo et al. (2022) declared that the reduction of the dispersion can be associated with a decrease in MF estimates, which results in a decrease in genetic trends.

In our study, the estimated genetic trends were similar across all traits for most models and scenarios (Figure 2.3, supp. Fig. 2.1-3). However, they were shifted by a constant when applying MF in the multi-breed evaluation, as reported by Macedo et al. (2022). The slight differences may indicate less overprediction in the MF models than in the other single-step models, as Koivula et al. (2021) suggested. When considering the relationship between base populations, the genetic trend for MF scenarios is lower because it is adjusted for the estimate of the MF effect, which may avoid bias in the trends (Meyer et al., 2018). Usually, these trends may be skewed and contrasted with the expectations of gene frequencies, how they are estimated, or how compatible **G** and $\mathbf{A}_{22}$ are (Meyer et al., 2018). However, Kudinov et al. (2020) found that the validation results and genetic trends when applying the MF model to a multi-breed Nordic Red Dairy cattle population were almost identical to those with other methodologies.

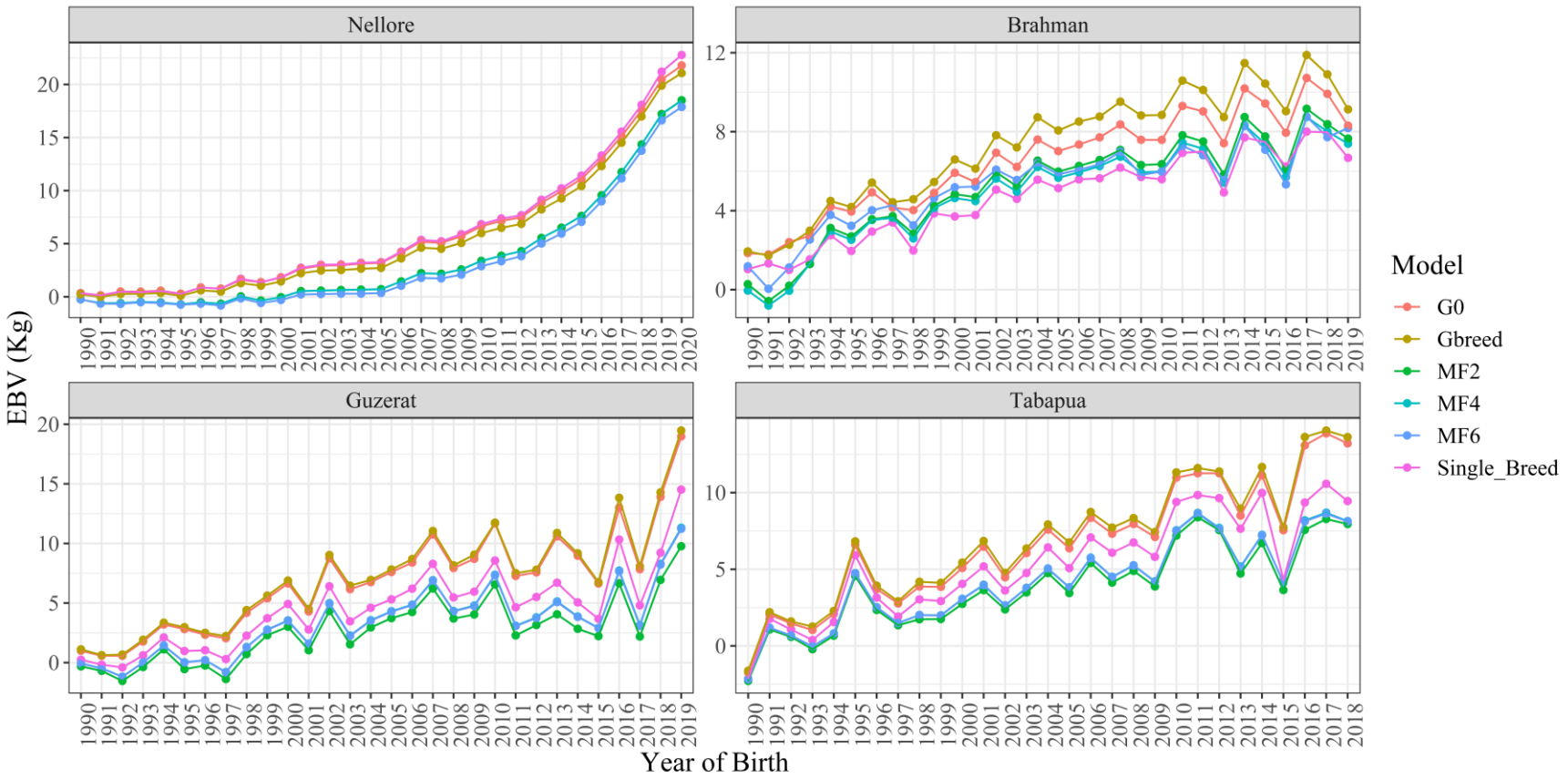


Figure 2.3. Genetic trends for the trait adjusted weight at 210 (W210) days of age for Bulls of the Breeds Nellore, Brahman, Guzerat, and Tabapua, in the six models evaluated

2.3.3 Compatibility between **G** *and* $\mathbf{A}_{22}$

When having a multi-breed population, animals from different breeds may be related through the pedigree; however, because each breed association keeps separate records, relationships across breeds are not considered or stored. Based on genomic relationships, which are identical by state, all genotyped animals share at least a few alleles; therefore, they may be all related (Misztal et al., 2021). If animals are related through **G** but not through $\mathbf{A}_{22}$, there may be some incompatibilities between these two matrices, causing bias and losses in accuracy.

The compatibility statistics for all the scenarios and breeds are shown in Figure 2.4 and Table 2.5. The extent of missing pedigree information determines the level of compatibility. In these situations, scaling $\mathbf{A}_{22}$ and **G** matrices can be problematic due to $\mathbf{A}_{22}$ being sparser and less informative than **G**. Table 2.5 shows that, in general, for b_1 (optimal value 1) and the correlation of the diagonal (optimal value 1) and off-diagonal (optimal value 1) elements of **G** and $\mathbf{A}_{22}$, the best scenarios are those with MF, which agrees with the theory presented by Legarra et al. (2015). This was valid for all the breeds and statistics, except by Brahman and Guzerat in b_1. Conversely, the correlation of the off-diagonal elements when analyzed within breeds (Figure 2.4) was low in these scenarios. Hence, there are still compatibility issues between the pedigree and genomic relationship matrices, not only with the MF approach but also with the $\mathbf{G}_{breed}$ and $\mathbf{G}_0$ scenarios.

The objective of the MF is to make $\mathbf{A}_{22}$ compatible with **G**, which theoretically eliminates the need for alignment, i.e., tuning (Masuda et al., 2022). According to Masuda et al. (2022), there is likely an underlying reason for the incompatibility, possibly linked to the calculation of base allele frequencies. A more thorough determination of these frequencies is probably required due to the uneven distribution of MF among genotypes (Kudinov et al., 2020). In fact, most individuals who underwent genotyping were Nellore.

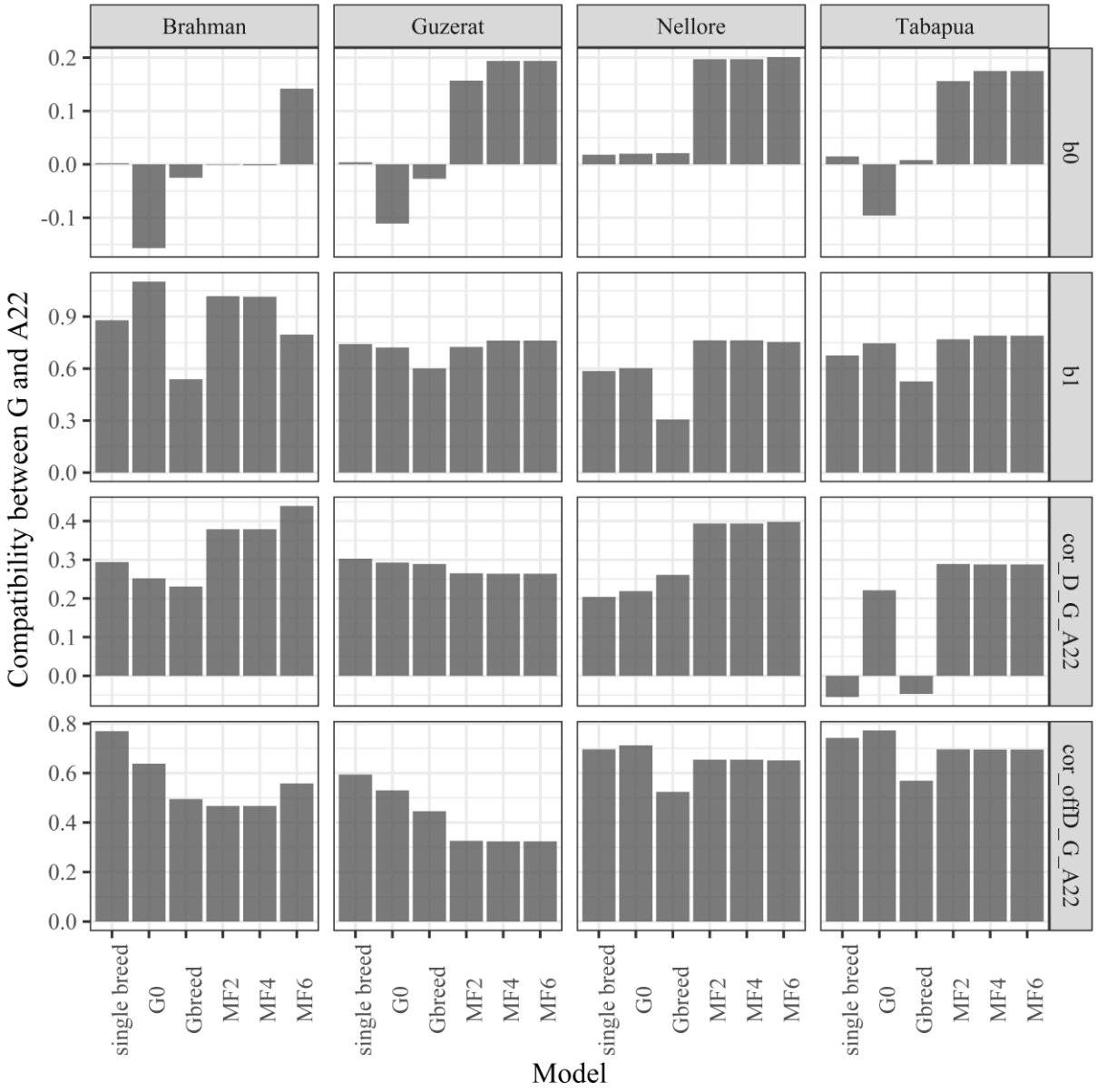


Figure 2.4. Within breed compatibility statistics for the **G** and $\mathbf{A}_{22}$ matrices for the six methodologies and breeds, B: Brahman, G: Guzerat, N: Nellore, and T: Tabapua.

Table 2.5. Compatibility statistics between **G** and $\mathbf{A_{22}}$ for single- and multi-breed evaluations of Nellore, Brahman, Guzerat, and Tabapua in different scenarios.

Model	Breed	b_0	b_1	Correlation diagonal elements G & A	Correlation Off-Diagonal elements G & A
$\mathbf{G_0}$	Four Breeds	0.02	0.54	0.06	0.71
$\mathbf{G_{breed}}$	Four Breeds	0.02	0.27	0.24	0.52
$\mathbf{MF_2}$	Four Breeds	0.22	0.72	0.29	0.73
$\mathbf{MF_4}$	Four Breeds	0.22	0.72	0.38	0.74
$\mathbf{MF_6}$	Four Breeds	0.23	0.71	0.41	0.75
Single breed	Nellore	0.02	0.59	0.20	0.70
Single breed	Brahman	0.00	0.88	0.29	0.77
Single breed	Guzerat	0.00	0.74	0.30	0.59
Single breed	Tabapua	0.02	0.68	-0.06	0.74

$\mathbf{G}_0$: **G** default with the current allele frequencies**;** $\mathbf{G}_{breed}$: **G** centered and scaled by breed-specific allele frequencies; MF_2: Two metafounders: one for Nellore and one for Brahman, Guzerat, and Tabapua; MF_4: Four metafounders: Nellore, Brahman, Guzerat, and Tabapua; MF_6: Six metafounders: Commercial Nellore, Brazilian Brahman, Bolivian Brahman, Registered Nellore, Guzerat, and Tabapua; Single Breed: Single step genomic best linear unbiased predictor for each breed separately.

Pearson's correlations for non-genotyped and genotyped animals between the GEBV from single-breed evaluations and all the other scenarios are in Table 2.6. There were small changes for the non-genotyped individuals, with correlations ranging from 0.92 up to 1 (except for Tabapua in the scenario $\mathbf{G}_{breed}$ with a correlation of 0.87). This is because changes in **G** and $\mathbf{A}_{22}$ affect those individuals indirectly based on the propagation of genomic information to non-genotyped animals in **H**. As the propagation relies on pedigree relationships between genotyped and non-genotyped individuals, fewer changes are expected because of limited connections among animals in **A**.

Table 2.6. Pearson's correlation of non-genotyped and genotyped animals comparing the single-breed ssGBLUP predictions versus all the multi-breed scenarios for the traits adjusted weight at 210 (W210) and 450 (W450) days of age, and scrotal circumference at 365 (SC365) days of age.

Single breed vs	Non-Genotyped animals				Genotyped animals			
	Nellore	Brahman	Guzerat	Tabapua	Nellore	Brahman	Guzerat	Tabapua
				W210				
$\mathbf{G_0}$	1.00	0.98	0.97	0.99	0.99	0.81	0.80	0.58
$\mathbf{G_{breed}}$	1.00	0.97	0.97	0.98	0.99	0.78	0.82	0.59
$\mathbf{MF_2}$	0.97	0.98	0.97	0.98	0.99	0.80	0.76	0.61
$\mathbf{MF_4}$	0.97	0.98	0.98	0.99	0.99	0.80	0.78	0.61
$\mathbf{MF_6}$	0.97	0.92	0.98	0.99	0.99	0.79	0.78	0.61
				W450				
$\mathbf{G_0}$	1.00	0.99	0.98	0.99	0.99	0.79	0.86	0.64
$\mathbf{G_{breed}}$	1.00	0.97	0.98	0.98	0.99	0.70	0.86	0.63
$\mathbf{MF_2}$	0.96	0.98	0.99	0.99	0.99	0.77	0.83	0.67
$\mathbf{MF_4}$	0.96	0.98	0.99	0.99	0.99	0.76	0.83	0.67
$\mathbf{MF_6}$	0.96	0.98	0.99	0.99	0.99	0.78	0.83	0.67
				SC365				
$\mathbf{G_0}$	1.00	0.98	0.97	0.98	1.00	0.72	0.67	0.69
$\mathbf{G_{breed}}$	0.99	0.92	0.93	0.87	1.00	0.57	0.65	0.68
$\mathbf{MF_2}$	0.98	0.95	0.98	0.98	0.99	0.73	0.67	0.67
$\mathbf{MF_4}$	0.98	0.98	0.98	0.98	0.99	0.70	0.68	0.67
$\mathbf{MF_6}$	0.98	0.98	0.98	0.97	0.99	0.70	0.67	0.66

$\mathbf{G_0}$: **G** default with the current allele frequencies**;** $\mathbf{G_{breed}}$: **G** centered and scaled by breed-specific allele frequencies; MF_2: Two metafounders: one for Nellore and one for Brahman, Guzerat, and Tabapua; MF_4: Four metafounders: Nellore, Brahman, Guzerat, and Tabapua; MF_6: Six metafounders: Commercial Nellore, Brazilian Brahman, Bolivian Brahman, Registered Nellore, Guzerat, and Tabapua

For the genotyped individuals, even though there was an increase in accuracy, fluctuations in GEBV are expected because any change in **G** affects relationships among all individuals with genomic information (Misztal et al., 2020, 2021). In general, there were more changes for the less representative breeds, with correlations ranging from 0.57 up to 0.81 for Brahman, 0.65 up to 0.86 for Guzerat, and 0.58 up to 0.69 for Tabapua, compared to the 0.99 up to 1 for Nellore, respectively. These correlations were similar among all the multi-breed scenarios, meaning that changes will happen when moving from single- to multi-breed evaluations, and reranking of animals is expected. Comparable results were reported by Kluska et al. (2021), reinforcing that changes and re-ranking of genotyped animals are rather natural in multi-breed

evaluations because of changes in **G**.

The results of our work show the importance of genomic predictions in modern livestock breeding programs that often have multi-breed populations. In this case, using genomic information from a breed with a large reference population as Nellore, with extensive genomic data and well-established genetic evaluations, provides valuable contributions to the smaller breeds, resulting in more accurate and less biased predictions for them. This allows breeders to make informed decisions to improve traits of economic importance, even for populations with limited information. However, the feasibility is directly proportional to the connectedness among breeds.

2.4. CONCLUSIONS

This work is a steppingstone toward more precise and efficient beef cattle breeding practices in programs that have indicine breeds with varying amounts of information. Combining such breeds in a single evaluation requires a proper methodology to avoid loss of precision of breeding values. Therefore, multi-breed genomic evaluations are proper for indicine beef cattle. The MF approach may be particularly beneficial for less-represented breeds, addressing limitations related to small reference populations and limited phenotyping/genotyping, and also addressing issues of compatibility between **G** and $\mathbf{A}_{22}$. By leveraging genomic information across breeds, breeders and producers can make more informed selection decisions, ultimately improving genetic gain in these cattle populations.

2.5. REFERENCES

ANUALPEC. (2021). **Anuário da Pecuária** Brasileira (IHS MARKIT - Agribusiness Intelligence, Ed.).

Bradford, H. L., Masuda, Y., VanRaden, P. M., Legarra, A., & Misztal, I. (2019). Modeling missing pedigree in single-step genomic BLUP. **Journal of Dairy Science**, 102(3), 2336–2346. https://doi.org/10.3168/jds.2018-15434

Briggs, H. M., & Briggs, D. M. (1980). **Modern Breeds of Livestock** (Macmillan, Ed.; 4th Edition). The Macmillian Publishing Company.

da Silva Neto, J. B., Peripolli, E., da Costa e Silva, E. V., Espigolan, R., Neira, J. D. R., Schettini, G., da Costa Filho, L. C. C., Barbosa, F. B., Macedo, G. G., Costa-Brunes, L., Lobo, R. B., Pereira, A. S. C., & Baldi, F. (2020). Genetic correlation estimates between age at puberty and growth, reproductive, and carcass traits in young Nelore bulls. **Livestock Science**, 241(October 2019), 104266. https://doi.org/10.1016/j.livsci.2020.104266

da Silva Neto, J. B., Peripoli, E., Pereira, A. S. C., Stafuzza, N. B., Lôbo, R. B., Fukumasu, H., Ferraz, J. B. S., & Baldi, F. (2023). Weighted genomic prediction for growth and carcass-related traits in Nelore cattle. **Animal Genetics**. https://doi.org/10.1111/age.13310

Fu, C., Ostersen, T., Christensen, O. F., & Xiang, T. (2021). Single-step genomic evaluation with metafounders for feed conversion ratio and average daily gain in Danish Landrace and Yorkshire pigs. **Genetics Selection Evolution**, 53(1), 79. https://doi.org/10.1186/s12711-021-00670-x

Garcia-Baccino, C. A., Legarra, A., Christensen, O. F., Misztal, I., Pocrnic, I., Vitezica, Z. G., & Cantet, R. J. C. (2017). Metafounders are related to F st fixation indices and reduce bias in single-step genomic evaluations. **Genetics Selection Evolution**, 49(1), 1–14. https://doi.org/10.1186/s12711-017-0309-2

Gibbs, R. A., Taylor, J. F., Van Tassell, C. P., Barendse, W., Eversole, K. A., Gill, C. A., Green, R. D., Hamernik, D. L., Kappes, S. M., Lien, S., Matukumalli, L. K., McEwan, J. C., Nazareth, L. V., Schnabel, R. D., Weinstock, G. M., Wheeler, D. A., Ajmone-Marsan, P., Boettcher, P. J., Caetano, A. R., … Dodds, K. G. (2009). Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. **Science**, 324(5926), 528–532. https://doi.org/10.1126/science.1167936

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E. (2009). Invited review : Genomic selection in dairy cattle : Progress and challenges. **Journal of Dairy Science**, 92(2), 433–443. https://doi.org/10.3168/jds.2008-1646

Junqueira, V. S., Lopes, P. S., Lourenco, D., Silva, F. F. e., & Cardoso, F. F. (2020). Applying the Metafounders Approach for Genomic Evaluation in a Multibreed Beef Cattle Population. **Frontiers in Genetics**, 11(December), 1–10. https://doi.org/10.3389/fgene.2020.556399

Kluska, S., Masuda, Y., Ferraz, J. B. S., Tsuruta, S., Eler, J. P., Baldi, F., & Lourenco, D. (2021). Metafounders May Reduce Bias in Composite Cattle Genomic Predictions. **Frontiers in Genetics**, 12(August). https://doi.org/10.3389/fgene.2021.678587

Kluska, S., Olivieri, B. F., Bonamy, M., Chiaia, H. L. J., Feitosa, F. L. B., Berton, M. P., Peripolli, E., Lemos, M. V. A., Tonussi, R. L., Lôbo, R. B., Magnabosco, C. de U., Di Croce, F., Osterstock, J., Pereira, A. S. C., Munari, D. P., Bezerra, L. A., Lopes, F. B., & Baldi, F. (2018). Estimates of genetic parameters for growth, reproductive, and carcass traits in Nelore cattle using the single step genomic BLUP procedure. **Livestock Science**, 216, 203–209. https://doi.org/10.1016/j.livsci.2018.08.015

Koivula, M., Strandén, I., Aamand, G. P., & Mäntysaari, E. A. (2021, April 26). Meta-model for genomic relationships of metafounders applied on large scale single-step random regression test-day model. **INTERBULL BULLETIN NO. 56**., 56, 1–6. https://journal.interbull.org/index.php/ib/article/download/70/70/63

Koufariotis, L., Hayes, B. J., Kelly, M., Burns, B. M., Lyons, R., Stothard, P., Chamberlain, A. J., & Moore, S. (2018). Sequencing the mosaic genome of Brahman cattle identifies historic and recent introgression including polled. **Scientific Reports**, 8(1), 17761. https://doi.org/10.1038/s41598-018-35698-5

Kudinov, A. A., Koivula, M., Aamand, G. P., Strandén, I., & Mäntysaari, E. A. (2022). Single-step genomic BLUP with many metafounders. **Frontiers in Genetics**, 13. https://doi.org/10.3389/fgene.2022.1012205

Kudinov, A. A., Mäntysaari, E. A., Aamand, G. P., Uimari, P., & Strandén, I. (2020). Metafounder approach for single-step genomic evaluations of Red Dairy cattle. **Journal of Dairy Science**, 103(7), 6299–6310. https://doi.org/10.3168/jds.2019-17483

Legarra, A., Christensen, O. F., Vitezica, Z. G., Aguilar, I., & Misztal, I. (2015). Ancestral Relationships Using Metafounders: Finite Ancestral Populations and Across Population Relationships. **Genetics**, 200(2), 455–468. https://doi.org/10.1534/genetics.115.177014

Legarra, A., & Reverter, A. (2018). Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. **Genetics Selection Evolution**, 50(1), 53. https://doi.org/10.1186/s12711-018-0426-6

Londoño-Gil, M., Cardona-Cifuentes, D., Espigolan, R., Peripolli, E., Lôbo, R. B., Pereira, A. S. C., Aguilar, I., & Baldi, F. (2023). Genomic evaluation of commercial herds with different pedigree structures using the single-step genomic BLUP in Nelore cattle. **Tropical Animal Health and Production**, 55(2). https://doi.org/10.1007/s11250-023-03508-4

Londoño-Gil, M., Cardona-Cifuentes, D., Rodríguez, J. D., Brunes, L. C., Magnabosco, C. U., Pereira, A. S. C., Peripolli, E., Lôbo, R. B., & Baldi, F. (2022). Heritability and genetic correlations between marbling in longissimus dorsi muscle and conventional economic traits in Nellore beef cattle. **Tropical Animal Health and Production**, 54(5), 274. https://doi.org/10.1007/s11250-022-03293-6

Lourenco, D. A. L., Tsuruta, S., Fragomeni, B. O., Chen, C. Y., Herring, W. O., & Misztal, I. (2016). Crossbreed evaluations in single-step genomic best linear unbiased predictor using adjusted realized relationship matrices. **Journal of Animal Science**, 94(3), 909–919. https://doi.org/10.2527/jas2015-9748

Lourenco, D., Tsuruta, S., Masuda, Y., Berman, M., Legarra, A., & Misztal, I. (2022, July 3). Recent updates in the BLUPF90 software suite. **World Congress on Genetics Applied to Livestock Production**: 3 - 8 July.

Macedo, F. L., Astruc, J. M., Meuwissen, T. H. E., & Legarra, A. (2022). Removing data and using metafounders alleviates biases for all traits in Lacaune dairy sheep predictions. **Journal of Dairy Science**, 105(3). https://doi.org/10.3168/jds.2021-20860

Macedo, F. L., Christensen, O. F., Astruc, J.-M., Aguilar, I., Masuda, Y., & Legarra, A. (2020). Bias and accuracy of dairy sheep evaluations using BLUP and SSGBLUP with metafounders and unknown parent groups. **Genetics Selection Evolution**, 52(1), 47. https://doi.org/10.1186/s12711-020-00567-1

Makgahlela, M. L., Strandén, I., Nielsen, U. S., Sillanpää, M. J., & Mäntysaari, E. A. (2014). Using the unified relationship matrix adjusted by breed-wise allele frequencies in genomic evaluation of a multibreed population. **Journal of Dairy Science**, 97(2), 1117–1127. https://doi.org/10.3168/jds.2013-7167

Masuda, Y., Tsuruta, S., Bermann, M., Bradford, H. L., & Misztal, I. (2021). Comparison of models for missing pedigree in single-step genomic prediction. **Journal of Animal Science**, 99(2), 1–10. https://doi.org/10.1093/jas/skab019

Masuda, Y., VanRaden, P. M., Tsuruta, S., Lourenco, D. A. L., & Misztal, I. (2022). Invited review: Unknown-parent groups and metafounders in single-step genomic BLUP. **Journal of Dairy Science**, 105(2), 923–939. https://doi.org/10.3168/jds.2021-20293

Meyer, K., Tier, B., & Swan, A. (2018). Estimates of genetic trend for single-step genomic evaluations. **Genetics Selection Evolution**, 50(1). https://doi.org/10.1186/s12711-018-0410-1

Misztal, I., Aguilar, I., Lourenco, D., Ma, L., Steibel, J. P., & Toro, M. (2021). Emerging issues in genomic selection. **Journal of Animal Science**, 99(6). https://doi.org/10.1093/jas/skab092

Misztal, I., Tsuruta, S., Lourenco, D. a L., Aguilar, I., Legarra, A., & Vitezica, Z. (2015). **Manual for BLUPF90 family of programs**. 125.

Misztal, I., Tsuruta, S., Lourenco, D., Aguilar, I., Legarra, A., & Vitezica, Z. (2014). **BLUPF90 family of programs**. In Athens: University of Georgia (p. 125). http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all2.pdf

Misztal, I., Tsuruta, S., Pocrnic, I., & Lourenco, D. (2020). Core-dependent changes in genomic predictions using the Algorithm for Proven and Young in single-step genomic best linear unbiased prediction. **Journal of Animal Science**, 98(12). https://doi.org/10.1093/JAS/SKAA374

R Core Team. (2023). R: A Language and Environment for Statistical Computing (4.3.0). **R Foundation for Statistical Computing.** https://www.r-project.org

Sevillano, C. A., Bovenhuis, H., & Calus, M. P. L. (2019). Genomic Evaluation for a Crossbreeding System Implementing Breed-of-Origin for Targeted Markers. **Frontiers in Genetics**, 10(MAY), 1–11. https://doi.org/10.3389/fgene.2019.00418

Steyn, Y., Lourenco, D. A. L., & Misztal, I. (2019). Genomic predictions in purebreds with a multibreed genomic relationship matrix. **Journal of Animal Science**, 97(11), 4418–4427. https://doi.org/10.1093/jas/skz296

Thomasen, J. R., Sørensen, A. C., Lund, M. S., & Guldbrandtsen, B. (2014). Adding cows to the reference population makes a small dairy population competitive. **Journal of Dairy Science**, 97(9), 5822–5832. https://doi.org/10.3168/jds.2014-7906

USDA. (2022). **Livestock and poultry: World Markets and Trade**. United States Department of Agriculture and Foreign Agricultural Service, April 8. https://www.fas.usda.gov/data/livestock-and-poultry-world-markets-and-trade

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. **Journal of Dairy Science**, 91(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., & Schenkel, F. S. (2009). Invited Review: Reliability of genomic predictions for North American Holstein bulls. **Journal of Dairy Science**, 92(1), 16–24. https://doi.org/10.3168/jds.2008-1514

Vercesi Filho, A. E., Faria, F. J. C., Madalena, F. E., & Josahkian, L. A. (2002). Estrutura populacional do rebanho Tabapuã registrado no Brasil. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, 54(6), 609–617. https://doi.org/10.1590/S0102-09352002000600009

Wicki, M., Raoul, J., & Legarra, A. (2023). Effect of subdivision of the Lacaune dairy sheep breed on the accuracy of genomic prediction. **Journal of Dairy Science**, 106(8), 5570–5581. https://doi.org/10.3168/jds.2022-23114

CHAPTER 3 - INDIRECT GENOMIC PREDICTIONS FOR INDICINE CATTLE BREEDS WITH SNP EFFECTS FROM A MULTI-BREED GENOMIC EVALUATION

Marisol Londoño-Gil[1,2], Jorge Hidalgo[2], Andres Legarra[2,3], Claudio U. Magnabosco[4], Fernando Baldi[1], Daniela Lourenco[2]

[1] Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista Júlio de Mesquita Filho, Via de Acesso Prof. Paulo Donato Castellane s/n, 14884-900 Jaboticabal, SP, Brazil

[2] Department of Animal and Dairy Science, University of Georgia, Athens, GA, 30602, USA

[3]Council on Dairy Cattle Breeding, Bowie, MD, 20716, USA

[4] Embrapa Cerrados, BR-020, 18, Sobradinho, 70770-901 Brasilia, DF, Brazil

Submitted to the Journal of Animal Breeding and Genetics on February 22, 2025

ABSTRACT

Indirect predictions (IP) are used for young genotyped animals that lack phenotypes (on their own or from progeny) or are from commercial herds. The former can be left behind because they do not contribute to the official genomic evaluations. The latter are often excluded from the evaluations because they are not registered and may not have pedigree information. Including such animals could result in inflated and biased genomic breeding values (GEBV). In Brazil, information is scarce for important breeds like Brahman, Guzerat, and Tabapua. Indirect predictions for young animals of these breeds based on a larger reference population could enhance genomic selection accuracy. Our objective in this study was to compute IP for young genotyped Nellore, Brahman, Guzerat, and Tabapua animals using single- and multi-breed analyses, with or without metafounders (MF) to model genetic differences across breeds. Records from the four breeding programs of the National Association of Breeders and Researchers (ANCP—Ribeirão Preto, SP, Brazil) were used. Data included pedigree (4.2M), phenotypes (329K), and genotypes (63.5K) across all breeds. The number of genotyped animals within each breed was 58,574, 3,102, 1,389, and 427 for Nellore, Brahman, Guzerat, and Tabapua, respectively. The analyzed traits were adjusted

weight at 210 (W210) and 450 (W450) days of age and the scrotal circumference at 365 (SC365) days of age. IP were derived as the sum of the SNP effects weighted by the gene content using different reference populations: multi-breed with or without MF, Nellore, or within-breed. Scenarios were compared using the linear regression (LR) method for bias, dispersion, and accuracy. Adding MF helped to decrease bias with values ≤ 0.13, 0.18, 0.28, and 0.55 for W210, W210m, W450, and SC365, respectively and avoid under or over-dispersion (values closer to 1), with slightly increased accuracy of IP. Combining breeds increased the accuracy of IP, mainly benefiting breeds with a small number of genotypes compared to their single breed evaluations. These findings suggest that when young genotyped animals are not included in an official multi-breed evaluation, robust IP can be obtained with proper modeling, regardless of the breed. This helps obtain fast genomic predictions for young animals without overwhelming the evaluation system.

Keywords: genomics, metafounders, reference population, selection, young genotyped animals.

3.1 INTRODUCTION

Indirect predictions (IP) can help offload the computational burden of large-scale evaluations. This is because many genotyped individuals are young, meaning they have no phenotypes or progeny phenotypes. These animals must not be included in the official evaluations (Tsuruta et al., 2021). Additionally, commercial farms are genotyping animals for herd management, increasing the need for IP as those animals will never take part in an official genomic evaluation (Londoño-Gil et al., 2023; Tsuruta et al., 2021). Through IP, superior young selection candidates can be identified earlier, especially in places where genomic evaluations are not frequent. Furthermore, if these animals have incomplete pedigrees, excluding them from official evaluations may prevent a decline in accuracy and an increase in inflation and bias in genomic breeding values (GEBV) (Lourenco et al., 2018). IP are obtained as the sum of the SNP effects weighted by the gene content. These SNP effects can be calculated based on the GEBV from the official genomic evaluations. Hence, these IP provide quick genomic predictions for newly genotyped young animals without affecting routine evaluations or compromising the accuracy and stability of GEBV (Tsuruta et al., 2021).

To obtain reliable predictions, it is essential to calculate the SNP effects using a large group of phenotyped and genotyped animals encompassing all the different chromosome segments segregating in the population (Lourenco et al., 2015;Tsuruta et al., 2021). In Brazil, the number of phenotyped and genotyped animals is limited for breeds such as Brahman, Guzerat, and Tabapua, generating challenges in enhancing the accuracy of genomic selection for these breeds. The difficulty lies in constructing a large reference population within each single cattle breed, as Hayes et al. (2023) indicated. Moreover, the independent chromosome segments comprise high linkage disequilibrium blocks with low recombination rates, indicating a gradual formation of new segments, as evidenced by Hidalgo et al. (2023). Therefore, acquiring IP for these breeds through a broader reference population, as in the multi-breed approach, could significantly improve the accuracy of genomic selection for breeds with limited data.

A multi-breed evaluation uses a single genetic effect for all the genotyped animals and a single genomic relationship matrix (**G**), resulting in the same SNP effects among breeds (Misztal et al., 2022). However, there are more appropriate approaches than this since the multi-breed model should include effects that account for breed differences while allowing for the compatibility of genomic and pedigree relationships (Misztal et al., 2022). Otherwise, the genetic predictions may be compromised (Steyn et al., 2019). Efficiently, the distinct genetic bases across breeds can be modeled with Metafounders (Bermann et al., 2023; Legarra et al., 2015).

Metafounders (MF) serve as virtual ancestors, simultaneously encapsulating the genetic contributions from paternal and maternal lines or breeds. This establishes relationships and facilitates considering inbreeding within and across base populations (Legarra et al., 2015). Consequently, they play a crucial role in addressing genetic disparities between breeds and refining the accuracy of genomic selection. By integrating MF into the genomic selection process, breeders can more precisely forecast individuals' genetic potential, regardless of their diverse backgrounds. Moreover, MF are instrumental in delineating the origins of cattle populations, offering insights into the biologically significant connections among distinct breeds shaped by evolutionary forces and geographical or genetic differences (Callister et al., 2022). Additionally, MF may enable a more precise representation of genetic diversity and

relationships within a multi-breed population, indispensable for breeding programs aiming to enhance specific traits across different breeds.

Therefore, this work aimed to compute IP for young genotyped Nellore, Brahman, Guzerat, and Tabapua animals using single-breed reference populations or a multi-breed reference population in the presence or absence of metafounders to model the genetic differences across breeds. A secondary objective was determining the minimum number of genotyped animals in the official Nellore evaluation to estimate accurate SNP effects and IP.

3.2 MATERIAL AND METHODS

This study did not require Animal Care and Use Committee approval, as the dataset was acquired from an existing database.

3.2.1 Data

The National Association of Breeders and Researchers (*Associação Nacional de Criadores e Pesquisadores* - ANCP, Ribeirão Preto - SP, Brazil) provided the data, which comprised pedigree (4,207,516), phenotypic (328,748), and genomic (63,492) information. This association runs the breeding programs of Nellore (Nellore Brazil program), Brahman (Brahman Brazil program), Guzerat (Brahman Brazil program), and Tabapua (Tabapua Brazil program). The dataset contained phenotypes for adjusted weight at 210 (W210) and 450 (W450) days of age and for scrotal circumference adjusted at 365 (SC365) days of age. Animals were genotyped with the ZBN Zoetis chip with 74,653 SNP (Table 3.1).

Phenotypic records deviating from the mean of contemporary groups ± 3 standard deviations, and contemporary groups with less than five records were removed. Contemporary groups for W210 and W450 were based on breed, farm, year and season of birth, sex, and management group classes. Contemporary groups for SC365 were defined using breed, farm, year of birth, and weaning management group classes. A pedigree depth of three was used, for a total of 15 generations, which comprehended the years from 1940 until 2022. The criteria for the quality control of SNP were based on minor allele frequency (MAF), call rate, and Mendelian conflicts.

Thus, markers with MAF < 0.05, call rate < 0.90, and monomorphic were excluded. Quality control was performed using the preGSF90 software from the BLUPF90 family of programs (Misztal et al., 2015).

Table 3.1 Descriptive data for adjusted weight at 210 (W210) and 450 (W450) days of age, and scrotal circumference at 365 (SC365) days of age for the Nellore Brahman, Guzerat, and Tabapua Breeds.

Trait		Nellore	Brahman	Guzerat	Tabapua	Total
W210 (kg)	N	204,019	88,099	26,632	9,998	328,748
	Mean	204.81	182.93	188.73	191.19	197.23
	SD	31.43	31.38	36.52	32.25	33.38
	Genotypes	58,574	3,102	1,389	427	63,492
	Pedigree	398,585	112,747	40,593	18,486	570,459
	Pheno & Geno	57,448	2,327	1,026	99	60,900
W450 (kg)	N	150,384	40,923	21,797	6,699	219,803
	Mean	319.29	271.24	290.95	289.64	306.63
	SD	58.21	56.62	65.33	60.29	61.85
	Genotypes	46,162	3,102	1,389	427	51,080
	Pedigree	315,097	68,119	34,432	14,086	431,173
	Pheno & Geno	44,435	1,686	985	166	47,272
SC365 (cm)	N	100,513	16,536	7,164	2,168	126,381
	Mean	22.43	20.17	21.03	21.69	22.04
	SD	2.68	2.59	2.49	2.39	2.77
	Genotypes	26,703	3,102	1,389	427	31,621
	Pedigree	248,120	40,101	17,976	7,377	313,300
	Pheno & Geno	25,206	423	342	102	26,073

N: Number of records, SD: standard deviation, Total: Descriptive statistics for the four breeds combined. Genotypes: Number of genotyped individuals. Pedigree: Number of animals in pedigree. Pheno & Geno: Number of animals with both phenotypes and genotypes.

The model used is presented below:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z_1 u} + \mathbf{Z_2 m} + \mathbf{Z_3 c} + \mathbf{e},$$

where $\mathbf{y}$, $\boldsymbol{\beta}$, $\mathbf{u}$, $\mathbf{m}$, $\mathbf{c}$, and $\mathbf{e}$ are the vectors of phenotypes, fixed effects (contemporary group and age of dam classes), direct additive genetic, maternal additive genetic, maternal permanent environmental, and residual effects, respectively, and $\mathbf{X}$, $\mathbf{Z_1}$, $\mathbf{Z_2}$,

and $\mathbf{Z_3}$ are the incidence matrices for the effects contained in $\boldsymbol{\beta}$, $\mathbf{u}$, $\mathbf{m}$, and $\mathbf{c}$. Maternal genetic and permanent maternal environmental effects were considered only for W210. The correlation between direct and maternal genetic effects was considered as zero as in Londoño-Gil et al. (2024).

3.2.2 Analyses

Four different genomic analysis scenarios were performed using ssGBLUP (Aguilar et al., 2010) to predict the GEBV and to compute SNP effects and, therefore, IP for young genotyped Nellore, Brahman, Guzerat, and Tabapua animals. These young animals were born after 2021 (W210) or 2020 (W450 and SC3650). The SNP effects needed for IP were estimated based on the following scenarios:

1) Multi-breed ssGBLUP with no metafounders (**no MF**): **G** was constructed as in VanRaden (2008) based on multi-breed population current allele frequencies.

2) Multi-breed ssGBLUP with four metafounders, one for each breed (**MF**).

3) Single-breed Nellore ssGBLUP (**Nellore-based**): SNP effects based on Nellore single breed evaluation used for IP of small breeds.

4) Single-breed ssGBLUP (**Single Breed**): Each breed was evaluated separately.

Without MF (no MF, Nellore-based, and Single breed scenarios), the covariance structure among animals in ssGBLUP (Legarra et al., 2009) was defined as in Aguilar et al. (2010) and constructed using the BLUP90IOD3 program from the BLUPF90 suite of programs (Lourenco et al., 2022; Misztal et al., 2014):

$$\mathbf{H^{-1}} = \mathbf{A^{-1}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G^{-1}} - \mathbf{A_{22}^{-1}} \end{bmatrix},$$

where **G** was constructed as in VanRaden (2008), that is, $\mathbf{G} = \mathbf{ZZ}'$, where:

$$\mathbf{Z} = \frac{(\mathbf{M} - \mathbf{P})}{\sqrt{k}},$$

with $\mathbf{M}$ being the matrix of n SNP genotypes for each animal, with elements indicating the gene content (i.e., copies of the reference allele) and set to 0, 1, 2, and 5 for the homozygote, heterozygote, alternative homozygote, and missing genotypes,

respectively; $\mathrm{p_j}$ corresponds to the reference allele frequency at locus *j*. The $\mathbf{P}$ matrix contains elements equal to $2p_j$ to center elements of **Z** by the mean of gene content, that is, twice the frequency of the reference allele at locus j. Additionally, *k* is $2\sum_{\mathrm{j=1}}^{\mathrm{n}}\mathrm{p_j}(1-\mathrm{p_j})$. Allele frequencies were calculated based on the current genotyped population (single or multi-breed).

For the MF scenario, the metafounders were related by a matrix of additive relationships **A**(**Γ**) given by a positive definite matrix that was invertible, **Γ**, as follows:

$$\boldsymbol{\Gamma} = \begin{bmatrix} 0.71 & 0.66 & 0.68 & 0.67 \\ 0.66 & 0.69 & 0.68 & 0.66 \\ 0.68 & 0.68 & 0.76 & 0.69 \\ 0.67 & 0.66 & 0.69 & 0.73 \end{bmatrix}$$

where the relationships within MF (Diagonal of $\boldsymbol{\Gamma}$) and between MF (off-diagonals of $\boldsymbol{\Gamma}$ belonged to Nellore, Brahman, Guzerat, and Tabapua) (Londoño-Gil et al., 2024b). This matrix was computed using the GAMMAF90 program from the BLUPF90 suite of programs (Lourenco et al., 2022; Misztal et al., 2014). Then, the inverse relationship matrix with pedigree, genomics, and MF information $\mathbf{H}^{\Gamma-1}$ was calculated using the BLUP90IOD3 program from the BLUPF90 suite of programs (Lourenco et al., 2022; Misztal et al., 2014), as follows:

$$\mathbf{H}^{\Gamma-1} = \mathbf{A}^{\Gamma-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{05}^{\Gamma-1} - \mathbf{A}_{22}^{\Gamma-1} \end{bmatrix},$$

where $\mathbf{H}^{\Gamma-1}$ is the inverse matrix that combines the genomic and pedigree matrix considering MF; $\mathbf{A}^{\Gamma-1}$ is an inverse matrix of additive relationships with MF; $\mathbf{G}_{05}^{\Gamma-1}$ is the inverse of the genomic relationship matrix of genotyped animals, with allele frequencies $\mathrm{p}_j = 0.5$, and $\mathbf{A}_{22}^{\Gamma-1}$ is the matrix of additive relationships between genotyped animals, considering MF. $\mathbf{G}_{05} = \mathbf{ZDZ}' = \frac{2}{m}\mathbf{ZZ}'$, with **Z** coded as before, and m number of markers.

For all the scenarios and to avoid singularity problems, **G** was blended to be invertible with a small percentage ($\alpha = 0.05$) of the pedigree relationship matrix among genotyped animals, which corresponds to adding a residual polygenic effect as follows:

- MF scenario: $\mathbf{G}_{05}^{\mathrm{blended}} = (1-\alpha)\mathbf{G}_{05} + \alpha\mathbf{A}_{\Gamma 22} = (1-\alpha)\left(\frac{2}{m}\mathbf{ZZ}'\right) + \alpha\mathbf{A}_{\Gamma 22}$

- Other scenarios (without metafounders): $\mathbf{G}^{\text{blended}} = (1-\alpha)\mathbf{G} + \alpha\mathbf{A_{22}}$

In the scenarios without MF, $\mathbf{G}^{\text{blended}}$ was tuned to have the same means of diagonals and off-diagonals as $\mathbf{A}_{22}$, therefore:

$$\mathbf{G}^{\text{tunned}} = b\mathbf{G}^{\text{blended}} + a,$$

where a and b are the tuning parameters, calculated as in Christensen et al. (2012) and solved the system of equations $\text{Avg}\left(\text{diag}\left(\mathbf{G}^{\text{blended}}\right)\right)b + a = \text{Avg}\left(\text{diag}(\mathbf{A}_{22})\right)$, $\text{Avg}\left(\text{offdiag}\left(\mathbf{G}^{\text{blended}}\right)\right)b + a = \text{Avg}(\text{offdiag}(\mathbf{A}_{22}))$, where a adds the extra average relationship and at the same time models the difference in μ from the pedigree base to the genomic base; and the b considers the reduction in variance due to drift.

After obtaining the GEBV in each of the four scenarios, SNP effects were computed using POSTGSF90 (Lourenco et al., 2022; Misztal et al., 2014) as proposed by Wang et al. (2012) for the approaches without MF:

$$\hat{\mathbf{a}}|\hat{\mathbf{u}} = (1-\alpha)b\frac{1}{k}\mathbf{Z}'\mathbf{G}^{\text{tunned}-1}\hat{\mathbf{u}},$$

and as proposed by Legarra et al. (2024) in the scenario with MF:

$$\hat{\mathbf{a}}|\hat{\mathbf{u}} = (1-\alpha)\mathbf{Z}'\frac{2}{m}\mathbf{G}^{\text{blended}-1}\hat{\mathbf{u}},$$

where $\hat{\mathbf{a}}$ is a vector of SNP effects, and $\hat{\mathbf{u}}$ is a vector of GEBV for the old genotyped animals. For the MF scenario, there was no tuning in **G**, since the changes to ensure the compatibility between $\mathbf{A}_{22}$ and **G** were made in $\mathbf{A}_{22}$ and **A** but not in **G**.

Once the SNP effects were calculated, the IP were obtained via PREDF90 (Lourenco et al., 2022; Misztal et al., 2014) as in Lourenco et al. (2018):

$$\hat{\mathbf{u}}_{\text{IP}} = \hat{\mu}_g + \mathbf{Z}_{\text{v}}\hat{\mathbf{a}}$$

where $\hat{\mathbf{u}}_{\text{IP}}$ is a vector of estimated indirect predictions for the validation animals, $\mathbf{Z}_{\text{v}}$ is the matrix of SNP content for the validation animals, centered by the same allelic frequencies as the original **Z**, $\hat{\mathbf{a}}$ is a vector of SNP effects, and $\hat{\mu}_g$ is the weighted average of GEBV for genotyped animals (without including the validation animals), computed as:

$$\hat{\mu}_g = \mathbf{1}'\mathbf{G}^{-1}\hat{\mathbf{u}},$$

where **1** is a vector of 1's. Adding $\hat{\mu}_g$ is necessary to guarantee that the IP is on the same scale as the GEBV form the official evaluations, considering the tuning of **G** to align with **A** (Lourenco et al., 2018).

For the MF method $\hat{\mu}_g$ is always equal to zero. Since in the MF scenario, there is no tuning, the mean is not added.

3.2.3 Validation

All scenarios were compared using the linear regression method (LR) proposed by Legarra and Reverter (2018). The "whole" dataset included all available sources of information for the reference and validation animals (GEBVs - $\hat{u}_w$), whereas the "partial" dataset omitted all the sources of information for the validation animals, i.e., young genotyped animals born from 2021 onwards for W210 or 2020 for W450 and SC365, for which we obtained the $\hat{\mathbf{u}}_{IP}$ (IP). Table 3.2 presents the number of records in the whole and partial datasets. The LR estimators (Legarra & Reverter, 2018) of accuracy, bias, and dispersion for the validation animals were computed as follows: $\text{accuracy} = \widehat{acc} = \sqrt{\frac{\text{cov}(\hat{u}_w,\hat{u}_{ip})}{(1-\bar{F})\widehat{\sigma^2_u}}}$; $\text{bias} = \hat{\Delta}_{w,ip} = \frac{\bar{\hat{u}}_{ip}-\bar{\hat{u}}_w}{\sigma_u}$; $\text{dispersion} = \hat{b}_{w,ip} = \frac{\text{cov}(\hat{u}_{ip},\hat{u}_w)}{\text{var}(\hat{u}_{ip})}$; where $\hat{u}$ ($\bar{\hat{u}}$) refers to the predicted breeding values (mean predicted breeding values), and the subscript indicates the whole (w, i.e., GEBV) or partial (i.e., IP) datasets. $\bar{F}$ is the mean inbreeding coefficient of the validation animals.

Table 3.2. Number of records in the whole and partial dataset and validation animals for whom the indirect predictions (IP) were estimated to validate using the Linear Regression method to assess statistics for the different scenarios for adjusted weight at 210 (W210), and 450 (W450) days of age, and scrotal circumference at 365 (SC365) days of age.

Trait		Nellore	Brahman	Guzerat	Tabapua	4 Breeds
W210 (kg)	Whole	204,019	88,099	26,632	9,998	328,748
	Partial	187,268	84,475	26,131	9,895	307,769
	IP	12,936	330	151	61	*
W450 (kg)	Whole	150,384	40,923	21,797	6,699	219,803
	Partial	126,987	37,039	21,168	6,558	191,752
	IP	14,558	521	279	102	*
SC365 (cm)	Whole	100,513	16,536	7,164	2,168	126,381
	Partial	83,472	14,606	6,794	2,089	106,961
	IP	8,615	271	152	61	*

4 breeds: All four breeds in the multi-breed evaluation. * The validation animals depended on the breed evaluated but are the sum of the four breeds.

We computed Pearson correlations (using R; R Core Team, 2023) among SNP effects from the different scenarios to study the agreement across models. In addition, to see the minimum number of genotyped animals necessary to obtain reliable IP for the Nellore breed, we estimated Pearson correlations among SNP effects and IP derived from GEBV of a variable number of genotyped animals. The latter analysis was performed after running the PREGSF90 software (Lourenco et al., 2022; Misztal et al., 2014) to conduct singular value decomposition (SVD) of $\mathbf{Z}$ to estimate the number of the largest eigenvalues explaining 90%, 95%, 98%, and 99% of genetic variance in $\mathbf{G}$. The number of largest eigenvalues explaining an x% of the variance in $\mathbf{G}$ was used as the number of Nellore genotyped animals included in predicting GEBV. The GEBV were used to derive the SNP effects and IP, which were compared against the SNP effects and IP obtained using all the available Nellore genotypes, i.e., 50k, representing our benchmark.

3.3 RESULTS AND DISCUSSION

3.3.1 SNP Effects

Figure 3.1 and Supplementary Figures 3.1-3 depict the correlations among SNP effects through the different scenarios. High positive correlations (r close to 1) among multi-breed scenarios (MF/no MF) and between multi-breed and the single-breed Nellore suggests they encapsulate similar SNP additive genetic effects. Given its substantial representation in genotyped samples, this reinforces the reliance on multi-breed scenarios on Nellore data. However, lower positive correlations (≤ 0.30) between MF/no MF and single-breed Brahman, Guzerat, and Tabapua indicate the existence of some shared genetic effects with discernible breed-specific differences potentially due to insufficient genotyped animals and limited phenotypic information within the smaller breeds to estimate SNP effects accurately.

Another possible explanation is that the assumption of shared SNP effects across breeds does not hold, i.e., the quantitative trait loci (QTL) affecting the traits may differ across breeds; using Nellore as a base to predict the SNP effects resulted in near-nonexistent correlations, ranging from 0.02 to 0.03 (Figure 3.1).

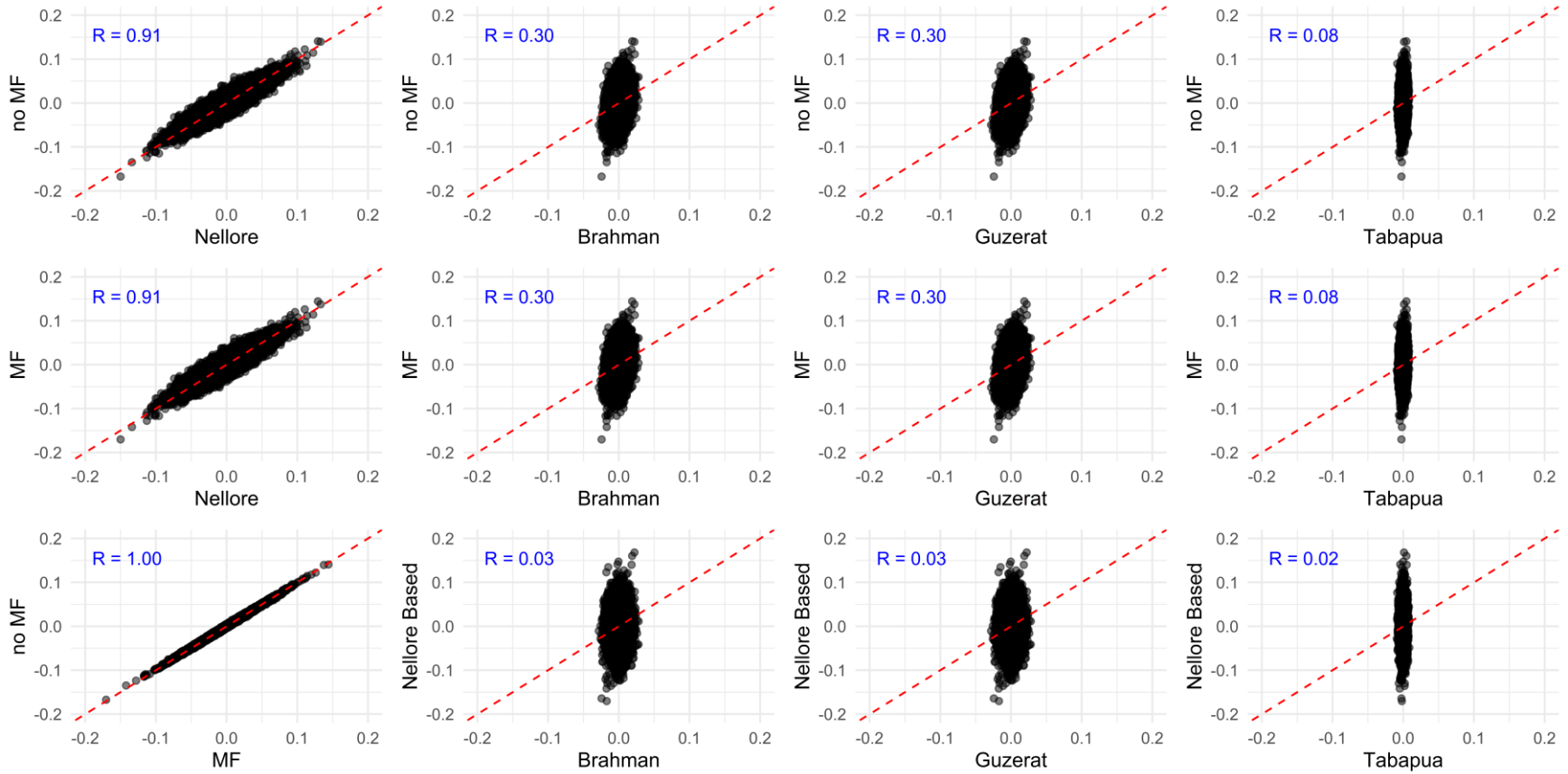


Figure 3.1. SNP effects correlations (R) between scenarios for the trait adjusted weight at 210 days of age (W210). No MF: SNP effects based on multi-breed GEBV without metafounders and using the default **G**, i.e., with current allele frequencies. MF: SNP effects based on GEBV obtained when using four metafounders, one for each breed. Nellore based: based on the Nellore GEBV for the small breeds Brahman, Guzerat and Tabapua. Single Breed (breed name is shown): based on each breed GEBV, when evaluated as a single breed. MF and no MF were both multi-breed scenarios.

When multiple breed populations are combined into a reference population, the SNP effects estimates tend to be dominated by the breeds contributing more data, primarily capturing effects of SNPs in linkage disequilibrium (LD) with QTL across all breeds or within the largest population (Karaman et al., 2021). The SNP effects are better estimated for the dominant breed due to the higher reference population, which has more information and dominates the allele frequency and substitution SNP effects. In this study, we used the MF approach to model efficiently the breed origin, adjusting SNP effects and pedigree relationships. With MF, the models can account for each breed's inherent differences and contributions (Christensen et al., 2012; Legarra et al., 2015). Therefore, accounting for genetic relationships among Nellore, Brahman, Guzerat, and Tabapua when incorporating MF in the multi-breed evaluation bridged these gaps, which were prominent when SNP effects were derived only from the Nellore reference population, resulting in very low correlations.

Steyn et al. (2019) showed the potential negative effect on prediction accuracy if SNP effects differ among breeds, especially if the scaling between $\mathbf{A}_{22}$ and **G** is incorrect and the number of SNPs is insufficient, leading to wrongly estimated SNP effects. Adjusting **G** elements by scaling diagonal and off-diagonal values to match the averages observed in $\mathbf{A}_{22}$ may mitigate the bias due to those differences (Bermann et al., 2021; Vitezica et al., 2011). A constant, $\hat{\mu}_g$, is used to address variations in the mean breeding values when using pedigree and genomic information, representing the mean breeding value for genotyped animals in the absence of selection pressure, regardless of allele frequencies (Bermann et al., 2021).

Nevertheless, when employing MF, pedigree relationships derived from related base populations alter $\mathbf{A}_{22}$ and **A**, with the overall means absorbing the covariances across individuals (i.e., covariance in $\boldsymbol{\Gamma}$). Since animals are likely related in these populations, and because the populations are under selection, the mean of the breeding values may deviate from zero. However, this deviation is adjusted for by the general mean of the model, thereby scaling the breeding values across the entire pedigree and ensuring compatibility between **G** and $\mathbf{A}_{22}$ (Legarra et al., 2015), which could improve the SNP effects estimation and the prediction accuracy.

Furthermore, we evaluated the impact of the number of genotyped animals used to estimate breeding values on SNP effects and IP accuracy. Based on the SVD of $\mathbf{Z}$ for the Nellore breed, the number of the largest eigenvalues explaining 90%, 95%, 98%, and 99% of the genetic variance was 5,443, 8,770, 13,641, and 17,289, respectively.

Figures 3.2 and 3.3 show the correlations among SNPs and IP. Figure 3.2 shows that in this Nellore population, more than 17k (number of the largest eigenvalues explaining 99% of the variance in $\mathbf{G}$) genotyped animals are needed to estimate SNP effects as accurately as using all (50k) available genotypes. The correlations were ≥ 0.90 using over 35k genotyped animals and peaked at 1.0 with 45k animals. Nevertheless, Figure 3.3 illustrates that the impact on the IP is less significant, as the correlations are higher. For example, the correlation between 17k-50k genotyped animals was 0.71 for SNP effects and 0.90 for IP. This is because IP gather the combined effect of all SNP effects, and many possible combinations can yield the same IP.

To better understand why reductions down to 9k genotyped animals still yielded a high IP correlation (0.85) with the 50k reference, we examined their distribution in the genomic PCA (results not shown), as presented in Londono-Gil et al., (2024). In the Nellore part of that PCA, the animals from the 9k subset were spread across the entire genomic space of the population, rather than being clustered. This indicates that the subset captured the overall genetic variability of the population, which helps explain the high IP accuracy despite the substantial reduction in both the number of genotyped animals and the SNP effect correlations.

These findings showed the importance of having enough genotyped animals to estimate SNP effects accurately. In our study, we kept the available phenotypic information constant in all the scenarios, and increasing the number of genotypes showed the value of realized genomic relationships, which were needed to harness the phenotypic information in the most optimum way, connecting more animals through their shared alleles or genomic similarity. Therefore, future studies should focus on optimizing the number of genotyped animals and the phenotypic records to balance accuracy and efficiency in estimating SNP effects.

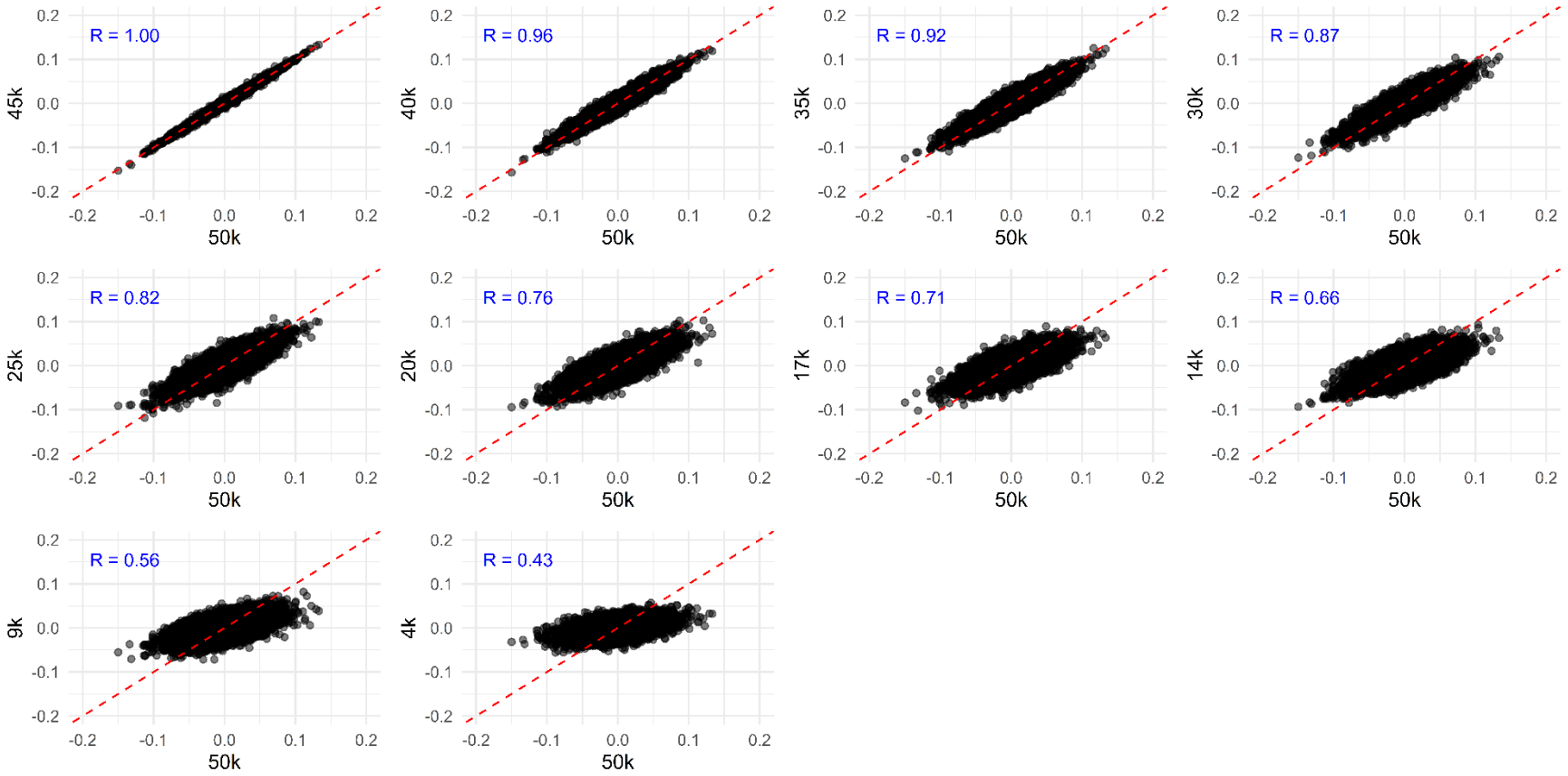


Figure 3.2. Pearson's correlations (R) among adjusted weight at 210 days of age SNP effects derived from different numbers of genotyped animals in the Nellore breed.

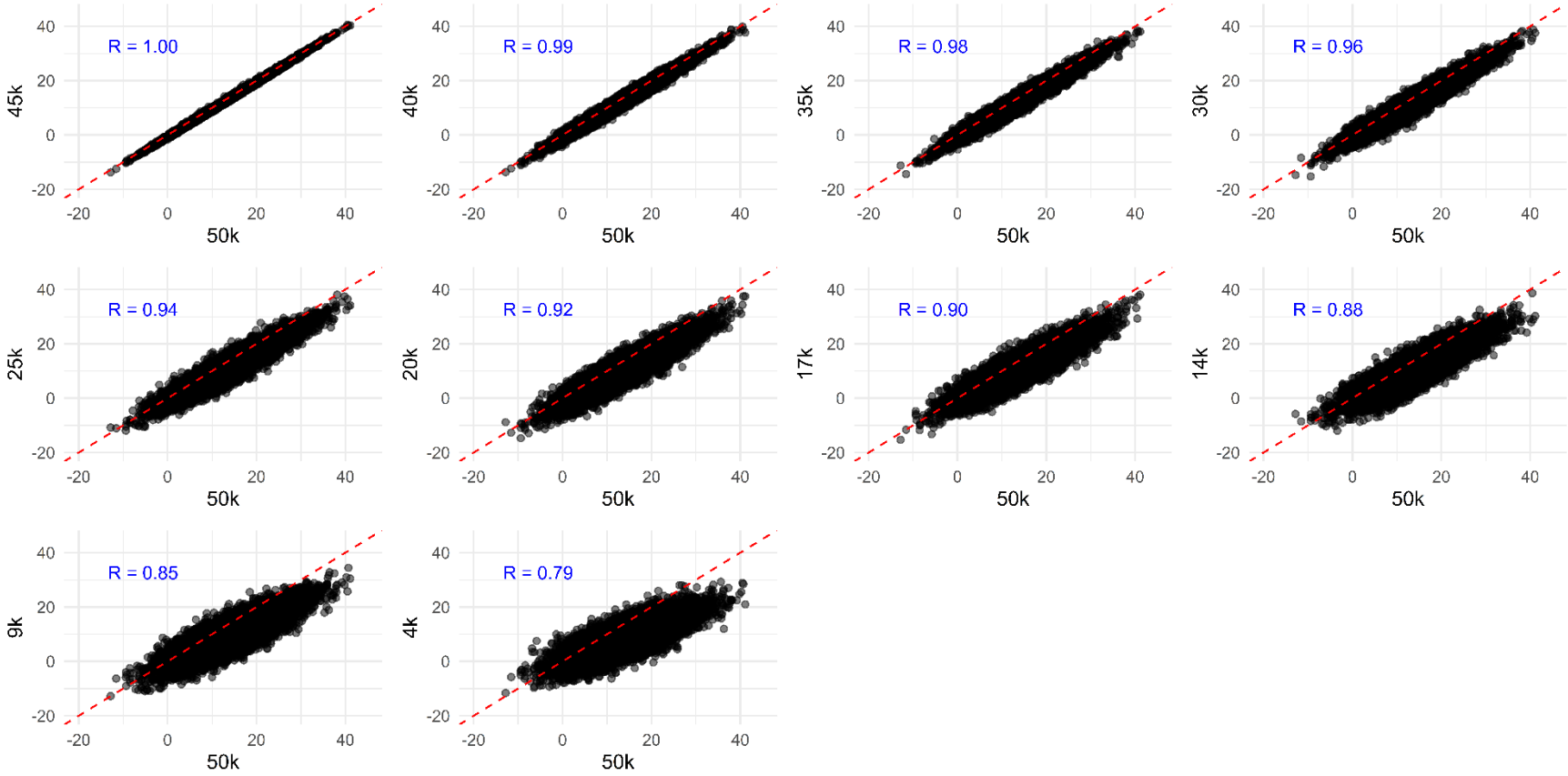


Figure 3.3. Pearson's correlations (R) among adjusted weight at 210 days of age indirect predictions derived from different numbers of genotyped animals in the Nellore breed.

This observation is consistent with previous studies demonstrating the importance of reference population size for accurate genomic predictions, particularly for multi-breed populations (Daetwyler et al., 2010; Hayes et al., 2009). For instance, Takeda et al. (2021) explored the size of the reference population for the expected accuracy of genomic prediction in Japanese Black cattle. They demonstrated that a reference population comprising more than 5,000 animals with phenotypes and genotypes was necessary for accurate genomic predictions in traits with varying heritability

Hidalgo et al. (2023) showed the value of a large (154k genotypes and 820k phenotypes for body weight) reference population in broilers; these authors demonstrated that when ample phenotypic and genomic data is used in the prediction of breeding values, these are predicted with high accuracy; therefore, the SNP effect derivation can be done from a subset of GEBV with a size equivalent to the number of largest eigenvalues explaining 99% of the variation in the genomic relationship matrix.

In our case, the combined genotype count of small breeds amounts to less than 5,000 animals, which is insufficient for accurately estimating SNP effects or representing independent chromosome segments segregating in the populations. The SVD showed that 6,445, 10,089, 15,095, and 18,726 animals are needed to explain 90%, 95%, 98%, and 99% of the genetic variance in the multi-breed population. This emphasizes the importance of using a big reference population such as Nellore to account for breed-specific effects that cannot be readily estimated when evaluated as a single breed.

3.3.2 Evaluation of Indirect Predictions

Validation statistics for the IP are shown in Figure 3.4. Multi-breed scenarios consistently achieved the highest accuracy across most traits and breeds (W210, W450, SC365), with the MF scenario achieving the highest accuracy, closely followed by the no-MF scenario. This supports the hypothesis that breed-specific adjustments are beneficial. In fact, Londoño-Gil et al. (2024) found that multi-breed evaluations benefit from incorporating metafounders or unknown parent groups compared to ignoring breed differences.

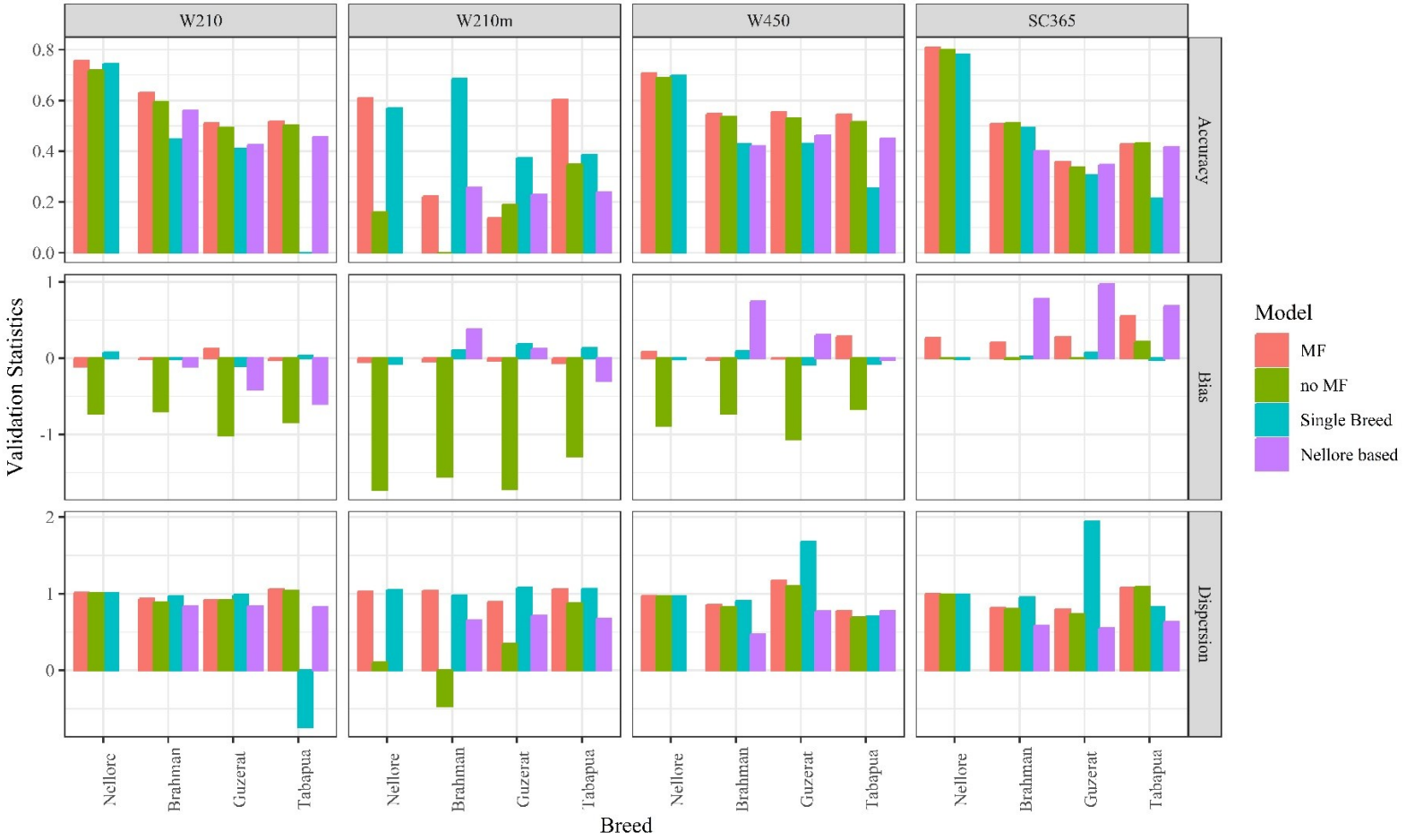


Figure 3.4. Accuracy, bias, and dispersion for indirect predictions of the validation animals in each scenario and trait. W210, W210m, W450: adjusted weight at 210 direct and maternal, and 450 days of age, respectively. SC365: scrotal circumference at 365 days of age. MF: SNP effects and IP based on GEBV obtained when using four metafounders, one for each breed. no MF: based on **G** default, with the current allele frequencies. Nellore based: based on the Nellore GEBV for the small breeds Brahman, Guzerat and Tabapua. Single Breed: based on each breed GEBV, when evaluated as a single breed. MF and no MF were both multi-breed scenarios

For smaller breeds, accuracy was notably higher when using multi-breed populations. This emphasizes the importance of large populations to estimate SNP effects (and derive IP), while ensuring that the breed of interest is adequately represented. However, the comparison across scenarios reveals important nuances. While MF improves accuracy and reduces bias, the no-MF scenario still performs well in some cases, particularly when large datasets and strong pedigree connections are available.

The Nellore-based and single-breed scenarios generally exhibited lower accuracies. While applying Nellore-based GEBVs to small breeds demonstrated some utility, it did not accomplish the same level of accuracy as breed-specific evaluations in many cases, highlighting the limitations of across-breed genomic predictions. For instance, Nellore-based accuracy was higher than single-breed accuracy for W210 across all breeds; however, the opposite was observed forW210m. This discrepancy may indicate challenges in accurately estimating maternal effects, which require sufficient generations of data and multiple parities per cow. Furthermore, maternal effects skip one generation, which complicates their estimation, particularly when validations are limited or unevenly distributed by sex. Future studies should assess the distribution of validation animals (e.g., males vs females) to ensure robust validation statistics for W210m.

The superiority of the MF scenario suggests that the remaining scenarios may not fully capture the comprehensive information encapsulated by the multi-breed approach, particularly for traits like W210 and W450, where the accuracy gap between MF and other scenarios was particularly pronounced. Accurate and unbiased prediction of breeding values for selection relies on phenotypic data connected through pedigree and genomic relationships of animals. The genetic similarity among animals, measured as the proportion of shared alleles, represents a main source of information in predicting breeding values because it connects phenotypic information across animals.

Contrary to traditional pedigrees that assume all the individuals in the base population or those not sharing known common ancestors as unrelated, genotypes

have the potential to reveal existing relationships within and across the base populations (or breeds) and individuals unrelated in the pedigree. This concept is exploited in the Metafounders theory (Legarra et al., 2015), which considers each ancestral population (or MF) as a finite-sized pool of gametes, with relationships within and across them. In addition to creating links between the populations, the Metafounders approach decomposes the additive genetic relationships into breed-specific terms for each breed and a segregation term for the pair of breeds. It also accounts for different genetic base populations due to selection (changes in the mean of breeding values and additive variance over time in a population) or crossbreeding (differences coming from different breeds), naturally overcoming possible compatibility issues between $\mathbf{A}_{22}$ and $\mathbf{G}$ matrices.

When metafounders are used, the overall mean breeding value ($\widehat{\mu_g}$) is absorbed into the model's general mean because MF accounts for ancestral relationships and adjusts the genetic base populations. This ensures compatibility between $\mathbf{G}$ and $\mathbf{A}_{22}$, aligning the scales of genomic and pedigree relationships. In scenarios without MF, $\widehat{\mu_g}$ is explicitly required to adjust for differences between genomic and pedigree-based breeding values, particularly when breed-specific SNP effects differ or when allele frequencies are inconsistent across populations (Steyn et al., 2019; Bermann et al., 2021).

These results indicate that using a multi-breed scenario can significantly improve accuracy of indirect predictions. These findings are consistent with previous studies (Brito et al., 2011; Clasen et al., 2023; Daetwyler et al., 2010; Karaman et al., 2021) that reported improved accuracy and better calibration of genomic predictions by accounting for breed differences and using multi-breed reference populations. Furthermore, Berman et al. (2023), Junqueira et al. (2020), and Legarra et al. (2015) demonstrated the utility of MF in accounting for breed-specific ancestral relationships and improving genomic predictions.

Regarding bias, the no-MF and Nellore-based approaches exhibited some bias, deviating strongly from zero for W210, W210m, and W450 (Figure 3.4), suggesting that suboptimal selection decisions will result when using these scenarios. The MF and

single-breed scenarios across traits and breeds presented lower bias values, which were ≤ 0.13, 0.18, 0.28, and 0.55 for W210, W210m, W450, and SC365, respectively (Figure 3.4).

Dispersion, measured as the deviation of b1 from 1, reflects the spread of predictions around their expected values. Values closer to 1 indicate a desirable spread, which was predominantly observed across traits and breeds in the MF and no MF scenarios (Figure 3.4). In contrast, the Nellore-based and Single Breed scenarios tended to deviate from this ideal value, suggesting either over- or under-dispersion (Figure 3.4). This emphasizes the importance of within-breed performance in multi-breed scenarios, in yielding more consistently spread predictions around the desirable values. Therefore, incorporating multi-breed evaluations and include MF could help to mitigate IP biases while adjusting the breed differences (Legarra et al., 2014; Luan et al., 2012).

Some studies, like Garcia et al. (2022) have shown the importance of having enough genotypes in the genomic evaluation when computing the GEBV used to backsolve SNP effects. These authors reported a significant drop in the correlation between GEBV and IP, from 0.88 to 0.82, when reducing the reference population from 10k to 2K animals. This finding highlights the critical role of a large and diverse population in ensuring the accuracy of IP, particularly for small breeds. In our study, the use of single-breed populations (less than 5k animals in small breeds when combined), resulted in lower accuracies, increased bias, and greater overdispersion.

These observations align with those of by Himmelbauer et al. (2024), who simulated a crossbred dairy population with ~15% missingness in the pedigrees of grandparents and great-grandparents. While they found no differences in the accuracy of predictions with or without MF, they did report improvements in dispersion (0.93 vs. 1.00) and bias (0.20 vs. 0.0 SD units), demonstrating the value of leveraging multi-breed genomic evaluations to address challenges related to incomplete or limited pedigree data.

The results of this work demonstrated that accurate indirect predictions based on a multi-breed indicine population can be obtained and are valuable for predicting

the performance of young genotyped animals. This has significant implications, particularly in countries such as Brazil, where the rapid adoption of advanced tools and technologies in livestock breeding programs has gained momentum. For smaller breeding programs, such as those for Guzerat and Tabapua, precise IP enables faster and more accurate genetic merit estimation, even in the absence of extensive phenotypic data. By ensuring adequate representation of the breed of interest within multi-breed reference populations, breeders can achieve higher accuracies and more reliable predictions, ultimately improving genetic progress and operational efficiency.

3.4 CONCLUSIONS

Incorporating multiple breeds into a large joint indicine reference population yielded enhanced accuracy, bias, and dispersion of genomic indirect predictions, especially for breeds with relatively few genotyped and phenotyped individuals. Combining breeds allowed to use of larger datasets, capturing broader genetic diversity and enabling more precise estimation of marker effects, thereby increasing the accuracy of IP. While including MF in the multi-breed evaluation showed some benefits in mitigating biases and improving dispersion, accurate IP can still be achieved without MF by leveraging well-structured multi-breed datasets. When young genotyped animals are excluded from the official multi-breed evaluation, robust IP can still be obtained with proper modeling and a sufficiently large and diverse reference population. However, relying on a single breed with limited data or making predictions based solely on another breed results in lower accuracy of IP, emphasizing the importance of multi-breed approaches for smaller breeding programs. These findings highlight the value of multi-breed genomic evaluations in achieving reliable genetic predictions and advancing genetic progress across breeds.

3.5 REFERENCES

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., & Lawlor, T. J. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J**ournal of Dairy Science**, 93(2), 743–752. https://doi.org/10.3168/jds.2009-2730

Bermann, M., Aguilar, I., Lourenco, D., Misztal, I., & Legarra, A. (2023). Reliabilities of estimated breeding values in models with metafounders. **Genetics Selection Evolution**, 55(1). https://doi.org/10.1186/s12711-023-00778-2

Bermann, M., Lourenco, D., & Misztal, I. (2021). Technical note: Automatic scaling in single-step genomic BLUP. **Journal of Dairy Science**, 104(2), 2027–2031. https://doi.org/10.3168/jds.2020-18969

Brito, F. V., Neto, J. B., Sargolzaei, M., Cobuci, J. A., & Schenkel, F. S. (2011). Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle. **BMC Genetics**, 12. https://doi.org/10.1186/1471-2156-12-80

Callister, A. N., Bermann, M., Elms, S., Bradshaw, B. P., Lourenco, D., & Brawner, J. T. (2022). Accounting for population structure in genomic predictions of Eucalyptus globulus. **G3: Genes, Genomes, Genetics,** 12(9). https://doi.org/10.1093/g3journal/jkac180

Christensen, O. F., Madsen, P., Nielsen, B., Ostersen, T., & Su, G. (2012). Single-step methods for genomic evaluation in pigs. **Animal**, 6(10), 1565–1571. https://doi.org/10.1017/S1751731112000742

Clasen, J. B., Fikse, W. F., Su, G., & Karaman, E. (2023). Multibreed genomic prediction using summary statistics and a breed-origin-of-alleles approach. **Heredity**, 131(1), 33–42. https://doi.org/10.1038/s41437-023-00619-4

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., & Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. **Genetics**, 185(3), 1021–1031. https://doi.org/10.1534/genetics.110.116855

Garcia, A., Aguilar, I., Legarra, A., Tsuruta, S., Misztal, I., & Lourenco, D. (2022). Theoretical accuracy for indirect predictions based on SNP effects from single-step GBLUP. **Genetics Selection Evolution**, 54(1), 66. https://doi.org/10.1186/s12711-022-00752-4

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E. (2009). Invited review : Genomic selection in dairy cattle : Progress and challenges. **Journal of Dairy Science**, 92(2), 433–443. https://doi.org/10.3168/jds.2008-1646

Hayes, B. J., Copley, J., Dodd, E., Ross, E. M., Speight, S., & Fordyce, G. (2023). Multi-breed genomic evaluation for tropical beef cattle when no pedigree information is available. **Genetics Selection Evolution**, 55(1). https://doi.org/10.1186/s12711-023-00847-6

Hidalgo, J., Lourenco, D., Tsuruta, S., Bermann, M., Breen, V., & Misztal, I. (2023). Derivation of indirect predictions using genomic recursions across generations in a broiler population. **Journal of Animal Science**, 101. https://doi.org/10.1093/jas/skad355

Himmelbauer, J., Schwarzenbacher, H., Fuerst, C., & Fuerst-Waltl, B. (2024). Exploring unknown parent groups and metafounders in single-step genomic BLUP: Insights from a simulated cattle population. **Journal of Dairy Science**. https://doi.org/10.3168/jds.2024-24891

Junqueira, V. S., Lopes, P. S., Lourenco, D., Silva, F. F. e., & Cardoso, F. F. (2020). Applying the Metafounders Approach for Genomic Evaluation in a Multibreed Beef Cattle Population. **Frontiers in Genetics**, 11(December), 1–10. https://doi.org/10.3389/fgene.2020.556399

Karaman, E., Su, G., Croue, I., & Lund, M. S. (2021). Genomic prediction using a reference population of multiple pure breeds and admixed individuals. **Genetics Selection Evolution**, 53(1). https://doi.org/10.1186/s12711-021-00637-y

Kluska, S., Masuda, Y., Ferraz, J. B. S., Tsuruta, S., Eler, J. P., Baldi, F., & Lourenco, D. (2021). Metafounders May Reduce Bias in Composite Cattle Genomic Predictions. **Frontiers in Genetics**, 12(August). https://doi.org/10.3389/fgene.2021.678587

Legarra, A., Aguilar, I., & Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. **Journal of Dairy Science**, 92(9), 4656–4663. https://doi.org/10.3168/jds.2009-2061

Legarra, A., Christensen, O. F., Aguilar, I., & Misztal, I. (2014). Single Step, a general approach for genomic selection. **Livestock Science,** 166(August 2014), 54–65. https://doi.org/10.1016/j.livsci.2014.04.029

Legarra, A., Christensen, O. F., Vitezica, Z. G., Aguilar, I., & Misztal, I. (2015). Ancestral Relationships Using Metafounders: Finite Ancestral Populations and Across Population Relationships. **Genetics**, 200(2), 455–468. https://doi.org/10.1534/genetics.115.177014

Legarra, A., Lourenco, D., & Vitezica, Z. (2024). **Bases for Genomic Prediction** (pp. 1–167).

Legarra, A., & Reverter, A. (2018). Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. **Genetics Selection Evolution**, 50(1), 53. https://doi.org/10.1186/s12711-018-0426-6

Londoño-Gil, M., Cardona-Cifuentes, D., Espigolan, R., Peripolli, E., Lôbo, R. B., Pereira, A. S. C., Aguilar, I., & Baldi, F. (2023). Genomic evaluation of commercial herds with different pedigree structures using the single-step genomic BLUP in Nelore cattle. **Tropical Animal Health and Production**, 55(2). https://doi.org/10.1007/s11250-023-03508-4

Londoño-Gil, M., López-Correa, R., Aguilar, I., Magnabosco, C. U., Hidalgo, J., Bussiman, F., Baldi, F., & Lourenco, D. (2024). Strategies for genomic predictions of an indicine multi-breed population using single-step GBLUP. **Journal of Animal Breeding and Genetics**. https://doi.org/10.1111/jbg.12882

Lourenco, D. A. L., Legarra, A., Tsuruta, S., Moser, D., Miller, S., & Misztal, I. (2018). Tuning Indirect Predictions Based on SNP Effects from Single-Step GBLUP. **Interbull Bulletin**, 53, 48–53. https://journal.interbull.org/index.php/ib/article/view/1448/1511

Lourenco, D. A. L., Tsuruta, S., Fragomeni, B. O., Masuda, Y., Aguilar, I., Legarra, A., Bertrand, J. K., Amen, T. S., Wang, L., Moser, D. W., & Misztal, I. (2015). Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. **Journal of Animal Science**, 93(6), 2653–2662. https://doi.org/10.2527/jas.2014-8836

Lourenco, D., Legarra, A., Tsuruta, S., Moser, D., Miller, S., & Misztal, I. (2018). Tuning indirect predictions based on SNP effects from ssGBLUP.

Lourenco, D., Tsuruta, S., Masuda, Y., Berman, M., Legarra, A., & Misztal, I. (2022, July 3). Recent updates in the BLUPF90 software suite. **World Congress on Genetics Applied to Livestock Production**: 3 - 8 July.

Luan, T., Woolliams, J. A., Degård, J., Dolezal, M., Roman-Ponce, S. I., Bagnato, A., & Meuwissen, T. H. E. (2012). The importance of identity-by-state information for the accuracy of genomic selection. **Genetics Selection Evolution**, 44(1). https://doi.org/10.1186/1297-9686-44-28

Misztal, I., Steyn, Y., & Lourenco, D. A. L. (2022). Genomic evaluation with multibreed and crossbred data. **JDS Communications**, 3(2), 156–159. https://doi.org/10.3168/jdsc.2021-0177

Misztal, I., Tsuruta, S., Lourenco, D. a L., Aguilar, I., Legarra, A., & Vitezica, Z. (2015). **Manual for BLUPF90 family of programs**. 125.

Misztal, I., Tsuruta, S., Lourenco, D., Aguilar, I., Legarra, A., & Vitezica, Z. (2014). **BLUPF90 family of programs**. In Athens: University of Georgia (p. 125). http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all2.pdf

R Core Team. (2023). **R: A Language and Environment for Statistical Computing** (4.3.0). R Foundation for Statistical Computing. https://www.r-project.org

Steyn, Y., Lourenco, D. A. L., & Misztal, I. (2019). Genomic predictions in purebreds with a multibreed genomic relationship matrix. **Journal of Animal Science**, 97(11), 4418–4427. https://doi.org/10.1093/jas/skz296

Takeda, M., Inoue, K., Oyama, H., Uchiyama, K., Yoshinari, K., Sasago, N., Kojima, T., Kashima, M., Suzuki, H., Kamata, T., Kumagai, M., Takasugi, W., Aonuma, T., Soma, Y., Konno, S., Saito, T., Ishida, M., Muraki, E., Inoue, Y., … Uemoto, Y. (2021). Exploring the size of reference population for expected accuracy of genomic prediction using simulated and real data in Japanese Black cattle. **BMC Genomics**, 22(1). https://doi.org/10.1186/s12864-021-08121-z

Tonussi, R. L., Londoño-Gil, M., de Oliveira Silva, R. M., Magalhães, A. F. B., Amorim, S. T., Kluska, S., Espigolan, R., Peripolli, E., Pereira, A. S. C., Lôbo, R. B., Aguilar, I., Lourenço, D. A. L., & Baldi, F. (2021). Accuracy of genomic breeding values and predictive ability for postweaning liveweight and age at first calving in a Nellore cattle population with missing sire information. **Tropical Animal Health and Production**, 53(4), 432. https://doi.org/10.1007/s11250-021-02879-w

Tsuruta, S., Lourenco, D. A. L., Masuda, Y., Lawlor, T. J., & Misztal, I. (2021). Reducing computational cost of large-scale genomic evaluation by using indirect genomic prediction. **JDS Communications**, 2(6), 356–360. https://doi.org/10.3168/jdsc.2021-0097

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. **Journal of Dairy Science**, 91(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

Vitezica, Z. G., Aguilar, I., Misztal, I., & Legarra, A. (2011). Bias in genomic predictions for populations under selection. **Genetics Research**, 93(5), 357–366. https://doi.org/10.1017/S001667231100022X

Wang, H., Misztal, I., Aguilar, I., Legarra, A., & Muir, W. M. (2012). Genome-wide association mapping including phenotypes from relatives without genotypes. **Genetics Research**, 94(2), 73–83. https://doi.org/10.1017/S0016672312000274

CHAPTER 4 - ACCURACY OF INDIRECT PREDICTIONS FOR INDICINE CATTLE BREEDS BASED ON A MULTI-BREED GENOMIC EVALUATION

Marisol Londoño-Gil[1,2], Jorge Hidalgo[2], Claudio U. Magnabosco[3], Fernando Baldi[1], Daniela Lourenco[2]

[1] Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista Júlio de Mesquita Filho, Via de Acesso Prof. Paulo Donato Castellane s/n, 14884-900 Jaboticabal, SP, Brazil

[2] Department of Animal and Dairy Science, University of Georgia, Athens, GA, 30602, USA

[3] Embrapa Cerrados, BR-020, 18, Sobradinho, 70770-901 Brasilia, DF, Brazil

[*] Correspondence: londono.gil@unesp.br

ORCIDS and email address:

Londoño-Gil, M: https://orcid.org/0000-0001-6522-5567, londono.gil@unesp.br

Hidalgo, J: https://orcid.org/0000-0002-0783-381X, jh37900@uga.edu

Legarra, A: https://orcid.org/0000-0001-8893-7620, andres.legarra@uscdcb.com

Magnabosco, C.U: https://orcid.org/0000-0002-7274-0134, claudio.magnabosco@embrapa.br

Baldi, F: https://orcid.org/0000-0003-4094-2011, fernandobaldiuy@gmail.com

Lourenco, D: https://orcid.org/0000-0003-3140-1002, danilino@uga.edu

ABSTRACT

Indirect genomic predictions (IP) are essential for identifying genetically superior young animals, particularly in breeds with limited reference populations, such as Brahman, Guzerat, and Tabapua. This study aimed to estimate the accuracy of IP based on the prediction error covariance (PEC) of SNP effects derived from single-step GBLUP (ssGBLUP) evaluations in multi-breed indicine cattle populations. Three scenarios were analyzed: (1) multi-breed ssGBLUP without metafounders (no MF), (2) multi-breed ssGBLUP with metafounders (MF), and (3) single-breed ssGBLUP. The dataset included pedigree, phenotypic, and genomic information from the Brazilian

National Association of Breeders and Researchers (ANCP), with 4.2 M pedigree records, 328,7 K phenotypic records, and 61,7 K genotyped animals, for the traits adjusted weight at 450 days (W450) and scrotal circumference at 365 days (SC365). Genomic estimated breeding values (GEBV) were estimated, and SNP effects were backsolved to compute IP. Accuracy (ACC_{IP}) was derived from the SNP PEC and validated by correlating IP accuracy with benchmark GEBV accuracy (ACC_{GEBV}). The MF scenario provided the highest ACC_{IP} across all breeds and traits. For W450, the highest correlations between ACC_{IP} and ACC_{GEBV} were observed in Nellore (0.87), Guzerat (0.74), and Tabapua (0.55). For SC365, MF also was better than the other models, with correlations of 0.61 in Nellore and 0.57 in Tabapua. The single-breed approach showed the lowest accuracy, particularly for Brahman and Tabapua, confirming its limited effectiveness for breeds with small genomic datasets. Regression coefficients of ACC_{GEBV} on ACC_{IP} also supported the consistency of the MF model, with values being the closest to 1 across breeds, albeit with significant bias. In conclusion, the inclusion of MF in multi-breed genomic evaluations improved the accuracy of IP, particularly for breeds with smaller reference populations. These findings support the use of structured multi-breed evaluations to improve genomic selection strategies in indicine cattle.

Keywords: genomics, metafounders, reference population, selection, young genotyped animals.

4.1 INTRODUCTION

Indirect predictions (IP) play a key role in identifying genetically superior young animals early, particularly in regions with limited genomic evaluations or breeds with small reference populations, such as Brahman, Guzerat, and Tabapua. These predictions are calculated as the sum of SNP effects weighted by gene content and can be derived from genomic estimated breeding values (GEBVs). Still, in genomic BLUP (GBLUP) or single-step GBLUP (ssGBLUP), SNP effects are not directly available. Instead, they must be backsolved from GEBVs obtained from official genetic evaluations using specific formulas (Strandén & Garrick, 2009; VanRaden, 2008). Once the SNP effects are estimated, IP can be computed.

In animal breeding programs, the accuracy or credibility of predicted breeding values (and, therefore, IP) is critical for improving selection decisions. Accuracy (ACC) is an individual, model-derived metric that quantifies the precision of an estimated breeding value (EBV). It is generally defined (Van Vleck, 1993) as the correlation, under repeated hypothetical sampling, between the true breeding value (u) of an individual and its estimate ($\hat{u}$). For BLUP models, assuming the model correctly represents reality, Henderson (1984) demonstrated that $Cov(u, \hat{u}) = Var(\hat{u}) = Var(u) - PEV = Var(\hat{u} - u)$. and then, the accuracy of estimated breeding values (EBVs) can be determined by the prediction error variance (PEV), which is typically calculated by inverting the coefficient matrix (**C**) of the BLUP mixed model equations (MME), according to the definition of correlation.

If **C** is known, PEV is calculated directly from the diagonal elements (c^{ii}) of the inverse $\mathbf{C}^{-1}$ This can be done using pedigree-based or genomic information (Garcia et al., 2022). Similarly, Strandén & Christensen (2011) and Tier et al. (2018) showed methods to calculate the accuracy of IP using the SNP prediction error covariance (PEC). Further studies by Gualdrón Duarte et al. (2014) and Aguilar et al. (2019) introduced derivations to estimate SNP effects under the ssGBLUP. For this, the exact computation of SNP PEC requires obtaining the full PEC matrix for all genotyped animals from the inverse of the MME (Garcia et al., 2022). Furthermore, in the calculation of accuracy, inbreeding must be considered. Aguilar et al. (2020) demonstrated that ignoring inbreeding led to both under and overestimation of

reliability in dairy cattle, for both genotyped and non-genotyped animals. In this regard, accounting for inbreeding computation of accuracy from PEV or PEC is essential for obtaining more precise and reliable predictions.

In the context of multi-breed evaluations, particularly relevant for the livestock sector in Brazil, improving the accuracy of genomic predictions is essential. Evaluating multiple breeds together may increase the precision of genetic predictions, especially for breeds like Brahman, Guzerat, and Tabapua, which have small reference populations. These breeds can benefit from a larger reference population, such as Nellore, leading to more accurate estimates for selection decisions. In this scenario, obtaining accurate indirect predictions (IP) is crucial for identifying genetically superior young animals earlier, allowing the producers and breeding companies to make more informed selection decisions with less uncertainty.

Consequently, this work aimed to estimate individual IP accuracies based on the error covariance (PEC) of SNP effects, derived from the PEC of GEBV in ssGBLUP, for young genotyped Nellore, Brahman, Guzerat, and Tabapua animals in multi-breed populations, with or without MF.

4.2 MATERIAL AND METHODS

This study did not require Animal Care and Use Committee approval, as the dataset was acquired from an existing database.

4.2.1 Data

The National Association of Breeders and Researchers (*Associação Nacional de Criadores e Pesquisadores* - ANCP, Ribeirão Preto - SP, Brazil) provided the data, which comprised pedigree (4,207,516), phenotypic (328,748), and genomic (61,732) information. This association runs the breeding programs of Nellore (Nellore Brazil program), Brahman (Brahman Brazil program), Guzerat (Brahman Brazil program), and Tabapua (Tabapua Brazil program). The dataset contained phenotypes for adjusted weight at 450 (W450) days of age and for scrotal circumference adjusted at 365 (SC365) days of age. Animals were genotyped with the ZBN Zoetis chip with

74,653 SNP (Table 4.1). The criteria for the quality control of SNP were based on minor allele frequency (MAF), call rate, and Mendelian conflicts. Thus, markers with MAF < 0.05, call rate < 0.90, and monomorphic were excluded. Quality control was performed using the preGSF90 software from the BLUPF90 family of programs (Misztal et al., 2015). After quality control 36,116 SNPs remained for the analysis.

Table 4.1. Descriptive data for adjusted weight at 450 (W450) days of age, and scrotal circumference at 365 (SC365) days of age for the Nellore Brahman, Guzerat, and Tabapua Breeds.

Trait		Nellore	Brahman	Guzerat	Tabapua	Total
W450 (kg)	N	150,384	40,923	21,797	6,699	219,803
	Mean	319.29	271.24	290.95	289.64	306.63
	SD	58.21	56.62	65.33	60.29	61.85
SC365 (cm)	N	100,513	16,536	7,164	2,168	126,381
	Mean	22.43	20.17	21.03	21.69	22.04
	SD	2.68	2.59	2.49	2.39	2.77
Both	Genotypes	56,814	3,102	1,389	427	61,732

N: Number of records, SD: standard deviation, Total: Descriptive statistics for the four breeds combined. Genotypes: Number of genotyped individuals.

The model used was as in Londoño-Gil et al (2024) and is presented below:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e},$$

where $\mathbf{y}$, $\boldsymbol{\beta}$, $\mathbf{u}$, and $\mathbf{e}$ are the vectors of phenotypes, fixed effects (contemporary group and age of dam classes), direct additive genetic and residual effects, respectively, and $\mathbf{X}$ and $\mathbf{Z}$, and are the incidence matrices for the effects contained in $\boldsymbol{\beta}$ and $\mathbf{u}$.

4.2.2 Scenarios

Three different genomic analysis scenarios were performed using ssGBLUP (Aguilar et al., 2010) to predict the GEBV and their accuracies and to compute SNP effects and, therefore, IP and their accuracies for young genotyped Nellore, Brahman, Guzerat, and Tabapua animals. These young animals were born after 2020 for both W450 and SC365. The SNP effects needed for IP were estimated based on the following scenarios:

1) Multi-breed with no metafounders (**no MF**)

2) Multi-breed ssGBLUP with four metafounders, one for each breed (**MF**)
3) Single-breed ssGBLUP (**Single Breed**).

Without MF (no MF, and Single breed scenarios), the covariance structure among animals in ssGBLUP (Legarra et al., 2009) was defined as in Aguilar et al. (2010) and constructed using the BLUPF90+ program from the BLUPF90 suite of programs (Lourenco et al., 2022; Misztal et al., 2014):

$$\mathbf{H^{-1}} = \mathbf{A^{-1}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G^{-1}} - \mathbf{A_{22}^{-1}} \end{bmatrix},$$

where **G** was constructed as in VanRaden (2008), that is, $\mathbf{G} = \mathbf{ZZ}'$, with:

$$\mathbf{Z} = \frac{(\mathbf{M} - \mathbf{P})}{\sqrt{k}},$$

and $\mathbf{M}$ being the matrix of n SNP genotypes for each animal, with elements indicating the gene content (i.e., copies of the reference allele) and set to 0, 1, 2, and 5 for the homozygote, heterozygote, alternative homozygote, and missing genotypes, respectively; $\mathrm{p_j}$ corresponds to the reference allele frequency at locus *j*. The $\mathbf{P}$ matrix contains elements equal to $2p_j$ to center elements of **Z** by the mean of gene content, that is, twice the frequency of the reference allele at locus j. Additionally, *k* is $2\sum_{\mathrm{j=1}}^{\mathrm{n}} \mathrm{p_j}(1 - \mathrm{p_j})$. Allele frequencies were calculated based on the current genotyped population.

For the MF scenario, the four metafounders were related by a matrix of additive relationships **A**(**Γ**) as in Londoño-Gil et al. (2024). Then, the inverse relationship matrix with pedigree, genomics, and MF information $\mathbf{H}^{\Gamma-1}$ was calculated using the BLUPF90+ program from the BLUPF90 suite of programs (D. Lourenco et al., 2022; Misztal et al., 2014), as follows:

$$\mathbf{H}^{\Gamma-1} = \mathbf{A}^{\Gamma-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{05}^{\Gamma-1} - \mathbf{A}_{22}^{\Gamma-1} \end{bmatrix},$$

where $\mathbf{H}^{\Gamma-1}$ is the inverse matrix that combines the genomic and pedigree matrix considering MF; $\mathbf{A}^{\Gamma-1}$ is an inverse matrix of additive relationships with MF; $\mathbf{G}_{05}^{\Gamma-\mathbf{1}}$ is the inverse of the genomic relationship matrix of genotyped animals, with allele frequencies $\mathrm{p}_j = 0.5$, and $\mathbf{A}_{22}^{\Gamma-1}$ is the matrix of additive relationships between

genotyped animals, considering MF. $\mathbf{G}_{05} = \mathbf{ZDZ}' = \frac{2}{m}\mathbf{ZZ}'$, with **Z** coded as before, and m number of markers.

For all the scenarios and to avoid singularity problems, **G** was blended to be invertible with a small percentage ($\alpha = 0.05$) of the pedigree relationship matrix among genotyped animals, which corresponds to adding a residual polygenic effect as follows:

- MF scenario: $\mathbf{G}_{05}^{\text{blended}} = (1-\alpha)\mathbf{G}_{05} + \alpha\mathbf{A}_{\Gamma 22} = (1-\alpha)\left(\frac{2}{m}\mathbf{ZZ}'\right) + \alpha\mathbf{A}_{\Gamma 22}$
- Other scenarios (without metafounders): $\mathbf{G}^{\text{blended}} = (1-\alpha)\mathbf{G} + \alpha\mathbf{A_{22}}$

In the scenarios without MF, $\mathbf{G}^{\text{blended}}$ was tuned to have the same means of diagonals and off-diagonals as $\mathbf{A}_{22}$, therefore:

$$\mathbf{G}^{\text{tunned}} = b\mathbf{G}^{\text{blended}} + a,$$

where a and b are the tuning parameters, calculated as in Christensen et al. (2012) and solved the system of equations $\text{Avg}\left(\text{diag}\left(\mathbf{G}^{\text{blended}}\right)\right)b + a = \text{Avg}\left(\text{diag}(\mathbf{A}_{22})\right)$, $\text{Avg}\left(\text{offdiag}\left(\mathbf{G}^{\text{blended}}\right)\right)b + a = \text{Avg}(\text{offdiag}(\mathbf{A}_{22}))$, where a adds the extra average relationship and at the same time models the difference in μ from the pedigree base to the genomic base; and the b considers the reduction in variance due to drift.

Once $\mathbf{H^{-1}}$ or $\mathbf{H}^{\Gamma-1}$ was built, the ssGBLUP MME were:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R^{-1}}\mathbf{X} & \mathbf{X}'\mathbf{R^{-1}}\mathbf{W} \\ \mathbf{W}'\mathbf{R^{-1}}\mathbf{X} & \mathbf{W}'\mathbf{R^{-1}}\mathbf{W} + \mathbf{H}^{*}\sigma_u^2 \end{bmatrix}\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R^{-1}}\mathbf{y} \\ \mathbf{W}'\mathbf{R^{-1}}\mathbf{y} \end{bmatrix},$$

where $\mathbf{R} = \mathbf{I}\sigma_e^2$ is the residual variance and σ_u^2 the additive genetic variance and $\mathbf{H}^{*}$ was the $\mathbf{H^{-1}}$ or $\mathbf{H}^{\Gamma-1}$ matrix, either not including or including MF.

4.2.3 Benchmark GEBV and accuracy

ssGBLUP evaluations using complete data (i.e., including validation animals) was run to obtain the benchmark GEBV accuracy (ACC_{GEBV}) for validation animals. The ACC_{GEBV} for an animal i was calculated based on its PEV ($Var(a_i - \hat{a}_i)$) from the inverse of the LHS of MME, as follows (Wang et al., 2012):

$$\text{Let } \mathbf{C^{-1}} = \begin{bmatrix} \mathbf{C^{\beta\beta}} & \mathbf{C^{\beta u}} \\ \mathbf{C^{\beta u}} & \mathbf{C^{uu}} \end{bmatrix}, \text{ then, } \text{ACC}_{\text{GEBV}_i} = r_i = \sqrt{1 - \frac{Var(a_i - \hat{a}_i)}{(1+F_i)\sigma_u^2}},$$

where the PEV_i is the diagonal element i in the prediction error variance matrix $\mathbf{C}^{\mathbf{uu}}$.

4.2.4 Indirect predictions and accuracy

After obtaining the GEBV in each of the three scenarios, SNP effects and SNP PEC were computed using POSTGSF90 and the OPTION snp_var (Lourenco et al., 2022; Misztal et al., 2014). The SNP effects were calculated as proposed by Wang et al. (2012) for the approaches without MF:

$$\hat{\mathbf{a}}|\hat{\mathbf{u}} = (1-\alpha)b\frac{1}{k}\mathbf{Z}'\mathbf{G}^{\text{tunned}-1}\hat{\mathbf{u}},$$

and as proposed by Legarra et al. (2024) in the scenario with MF:

$$\hat{\mathbf{a}}|\hat{\mathbf{u}} = (1-\alpha)\mathbf{Z}'\frac{2}{m}\mathbf{G}^{\text{blended}-1}\hat{\mathbf{u}},$$

where $\hat{\mathbf{a}}$ is a vector of SNP effects from a ssGBLUP evaluation, and $\hat{\mathbf{u}}$ is a vector of GEBV for the old genotyped animals (reduced data). For the MF scenario, there was no tuning in **G**, since the changes to ensure the compatibility between $\mathbf{A}_{22}$ and **G** were made in $\mathbf{A}_{22}$ and **A** but not in **G**.

The PEC of SNPs was calculated as in Gualdrón Duarte et al. (2014), Aguilar et al. (2019), and Garcia et al. (2022), as:

$$\text{Var}(\hat{\text{a}}) = (1-\alpha)b\,\frac{1}{2\sum \text{p}_\text{i}(1-\text{p}_\text{i})}(\mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z}\sigma^2_\text{u} - \mathbf{Z}'\mathbf{G}^{-1}\mathbf{C}^{\mathbf{u}_2\mathbf{u}_2}\mathbf{Z})\,\frac{1}{2\sum \text{p}_\text{i}(1-\text{p}_\text{i})}b(1-\alpha),$$

with α and b as blending and tuning parameters, as described before, accounted for in PEC computations, and $\mathbf{C}^{\mathbf{u}_2\mathbf{u}_2}$ is part of the inverse of the LHS of MME corresponding to genotyped animals.

In the computation of PEC, inbreeding was accounted for through both pedigree and genomic relationships. Specifically, **A** included inbreeding coefficients, and **G** was constructed following the method of VanRaden (2008). As a result, inbreeding present in the population was properly incorporated into the mixed model equations (MME) and consequently into the PEC matrix derived from their inverse. This approach helped avoid underestimation of uncertainty in SNP effects, particularly in small breeds.

Once the SNP effects and SNP PEC were calculated, the IP and their accuracies were obtained via PREDF90 with the addition of the –acc argument

(Lourenco et al., 2022; Misztal et al., 2014) as in Lourenco et al. (2018) and Garcia et al. (2022), the IP were calculated as:

$$\hat{\mathbf{u}}_{\mathrm{IP}} = \hat{\mu}_g + \mathbf{Z}_{\mathrm{v}}\hat{\mathbf{a}}$$

where $\hat{\mathbf{u}}_{\mathrm{IP}}$ is a vector of estimated indirect predictions for the validation animals, $\mathbf{Z}_{\mathrm{v}}$ is the matrix of SNP content for the validation animals, centered by the same allelic frequencies as the original **Z**, $\hat{\mathbf{a}}$ is a vector of SNP effects, and $\hat{\mu}_g$ is the weighted average of GEBV for genotyped animals (without including the validation animals), computed as:

$$\hat{\mu}_g = \mathbf{1}'\mathbf{G}^{-1}\hat{\mathbf{u}},$$

where $\mathbf{1}$ is a vector of 1's. Adding $\hat{\mu}_g$ is necessary to guarantee that the IP is on the same scale as the GEBV form the official evaluations, considering the tuning of **G** to align with **A** (Lourenco et al., 2018). For the MF method $\hat{\mu}_g$ is always equal to zero. Since in the MF scenario, there is no tuning, the mean is not added (Legarra et a. 2015).

The accuracy for IP for animal i ($\mathrm{ACC}_{\mathrm{IP}_i}$) was computed as in Garcia et al. (2022), as:

$$\mathrm{ACC}_{\mathrm{IP}_i} = \sqrt{\frac{\mathbf{z}_i \mathrm{Var}(\hat{\mathrm{a}})\, \mathbf{z}_i'}{\sigma^2_{\mathrm{u}}}}$$

with $\mathbf{z}_i$ as the row vector from the $\mathbf{Z}$ matrix, that contains the centered genotypes for the animal i.

4.2.5 Validation of IP accuracies

The scenarios no MF, MF, and Single Breed were used to compute the SNP PEC and, consequently, the accuracies of IP from ssGBLUP. In all these scenarios, the ssGBLUP analysis included all available data except for the validation individuals.

The validation set consisted of young genotyped animals born from 2020 onwards for W450 and SC365, for which we estimated $\hat{\mathrm{u}}_{\mathrm{IP}}$ (IP)This setup simulates a real-life scenario where GEBVs are available for older animals, allowing the estimation of IP by back-solving SNP effects based on these GEBVs. Table 4.2 presents the number

of records in the benchmark, reduced, and validation (IP) datasets, for which GEBV and IP accuracies were estimated

For all scenarios, the $\mathrm{ACC_{IP}}$ for validation animals were compared with the benchmark $\mathrm{ACC_{GEBV}}$ evaluated when the validation animals were included in the ssGBLUP evaluation. To verify the quality of the $\mathrm{ACC_{IP}}$, Pearson correlations were computed between the $\mathrm{ACC_{IP}}$ and $\mathrm{ACC_{GEBV}}$. Additionally, a regression model was applied as $\mathrm{ACC_{GEBV}} = b_0 + b_1\mathrm{ACC_{IP}}$ to explore the presence of dispersion in the $\mathrm{ACC_{IP}}$ estimate. Finally, the average and maximum absolute differences between the $\mathrm{ACC_{GEBV}}$ and $\mathrm{ACC_{IP}}$ were computed to assess the proximity between the accuracies obtained by the two approaches.

Table 4.2. Number of records in the benchmark, reduced, and validation (IP – Indirect Predictions) datasets, for which GEBV and IP accuracies were estimated across different scenarios for adjusted weight at 450 days (W450) and scrotal circumference at 365 days (SC365).

Trait		Nellore	Brahman	Guzerat	Tabapua	4 Breeds
W450 (kg)	**Benchmark**	150,384	40,923	21,797	6,699	219,803
	Reduced	126,987	37,039	21,168	6,558	191,752
	IP (validation)	14,556	519	277	100	*
SC365 (cm)	**Benchmark**	100,513	16,536	7,164	2,168	126,381
	Reduced	83,472	14,606	6,794	2,089	106,961
	IP (validation)	8,614	270	151	60	*

4 breeds: All four breeds in the multi-breed evaluation. * The validation animals depended on the breed evaluated but are the sum of the four breeds

4.3 RESULTS AND DISCUSSION

The accuracy of IP was evaluated using Pearson correlations between the $\mathrm{ACC_{GEBV}}$ and $\mathrm{ACC_{IP}}$ across different scenarios. Figures 4.1 and 4.3 illustrate these relationships for Nellore, Brahman, Guzerat, and Tabapua breeds under three evaluation scenarios: MF, no MF, and Single Breed.

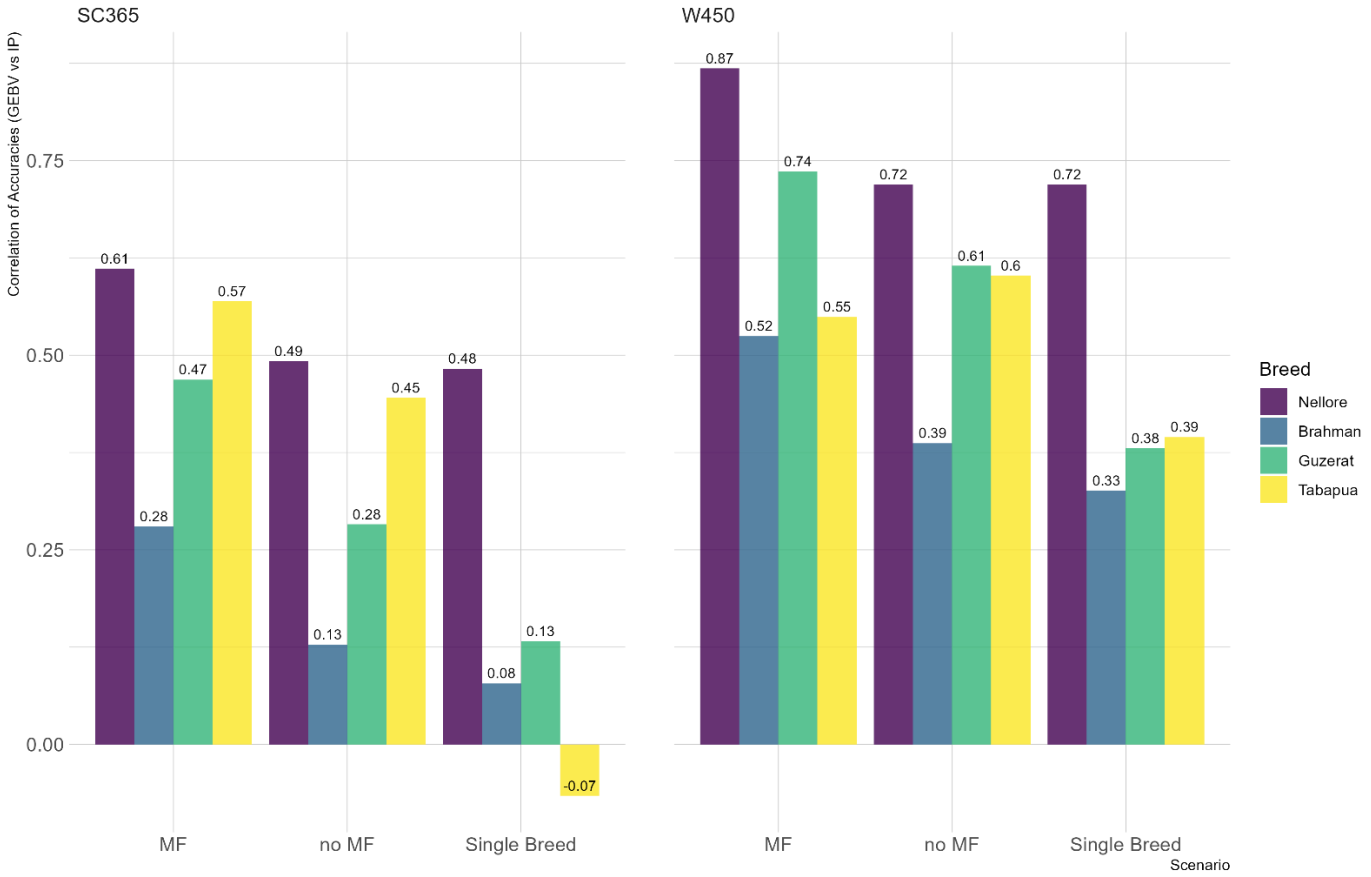


Figure 4.1. Correlation between the accuracy of genomic estimated breeding values ($\mathrm{ACC}_{\mathrm{GEBV}}$) and the accuracy of indirect predictions $\mathrm{ACC}_{\mathrm{IP}}$, for the Nellore Brahman, Guzerat, and Tabapua breeds, across different scenarios: Multi-breed ssGBLUP with metafounders (MF), Multi-breed ssGBLUP with no metafounders (no MF), and Single-breed ssGBLUP (Single Breed), for the traits scrotal circumference at 365 (SC365) days of age and adjusted weight at 450 (W450) days of age.

ACC_{IP} was the highest among the multi-breed model with metafounders (MF), pointing out the role of genetic structure in improving prediction accuracy. For W450, the MF scenario showed the strongest correlations, particularly in Nellore (0.87), Guzerat (0.74), and Tabapua (0.55). The no MF scenario performed slightly worse but still maintained relatively high correlations, especially for Nellore (0.72) and Guzerat (0.61). The Single Breed approach produced the lowest values, notably for Brahman (0.33), suggesting that single-breed models may be less effective for breeds with limited data

For SC365, the MF scenario again resulted in the highest correlations across most breeds, particularly for Nellore (0.61) and Tabapua (0.57). The no MF model showed slightly lower values, with Nellore (0.49) and Tabapua (0.45) still presenting relatively good correlations. However, Brahman and Guzerat had notably weaker correlations. The Single Breed approach had the poorest correlations, especially for Tabapua (-0.07) and Brahman (0.08), reinforcing the limitations of single breed evaluations for smaller populations.

Including MF in the multi-breed analysis consistently improved ACC_{IP} across all breeds, mainly benefiting Guzerat and Tabapua, which gained accuracy from a larger reference pool. Removing metafounders (no MF) led to decreased correlations but still performed better than single-breed models. Nellore consistently showed the highest correlations across all scenarios, likely due to a larger dataset and well-connected pedigree and genomic information. These results are similar to other multi-breed studies, where larger and more diverse reference populations improved accuracy, particularly for minor breeds (Jin et al., 2024; Takeda et al., 2021).

Single-breed reference populations often struggle to capture genetic variation in small populations, reducing prediction accuracy (Brito et al., 2017; Khansefid et al., 2020). The MF approach showed higher accuracies for the IP across all breeds, supporting previous studies that found MF models help correct biases caused by population stratification (Anglhuber et al., 2024). The increased accuracy for smaller breeds like Guzerat and Tabapua suggests that multi-breed reference populations help mitigate accuracy loss in populations with fewer genotyped individuals, as also reported in dairy cattle (Cole et al., 2021; Hayes et al., 2009).

Regression coefficients (b_1) of ACC_{GEBV} on ACC_{IP} (Figure 4.2) provide an idea of the potential biases in the IP accuracy estimates for the indicine breeds, especially when the data sets are small. A b_1 coefficient close to 1 indicates a reliable estimation of ACC_{IP}, whereas deviations suggest over- or underestimation. Table 4.3 summarizes the descriptive statistics of ACC_{GEBV} and ACC_{IP} across breeds, traits, and evaluation scenarios. This table further confirms that multi-breed models with metafounders (MF) are the most effective approach, leading to higher and more stable ACC_{IP} values across all breeds and traits, which was consistent with the regression coefficient.

Even though the correlations were generally high, especially in Nellore under the MF scenario (0.87), the wide range of regression coefficients suggested that the ACC_{IP} might be skewed or miscalibrated and that the IP scale did not match the GEBVs well. Animals with inflated IP may be given preference over those with lower actual genetic merit, which can result in mistakes in selection decisions. Given that genomic predictions are more sensitive to the size and composition of the reference population in small or less connected populations, these disparities are particularly significant. For all breeds, but especially for Nellore and Brahman, the MF model showed regression coefficients that were closest to 1, whereas the single-breed and no MF models tended to underestimate ACC_{IP}. For most of the scenarios and models, however, the ACC_{IP} was greatly overestimated. The notable bias found in these accuracy estimates could point to difficulties exactly forecasting the PEV in indicine breeds.

Ignoring genetic structure may lead to underestimated accuracy, particularly for smaller populations, according to prior research on PEV-based accuracy estimations (Aguilar et al., 2020; Gorjanc et al., 2015). This is particularly crucial for indicine cattle, as genetic variation between breeds and variations in population size can impact estimates of prediction error variance. Under multi-breed evaluations, breeds like Guzerat and Tabapua had the biggest increases in ACC_{IP}, especially in the MF scenario, which modeled the genetic variations between the breeds. This is in line with research showing that in breeds with few reference populations, adding larger reference populations improved genomic prediction accuracy (Rahimi et al., 2020; Xu et al., 2019). These enhancements were probably made possible by the utilization of Nellore genetic data, which makes up the majority of the dataset (Table 4.1).

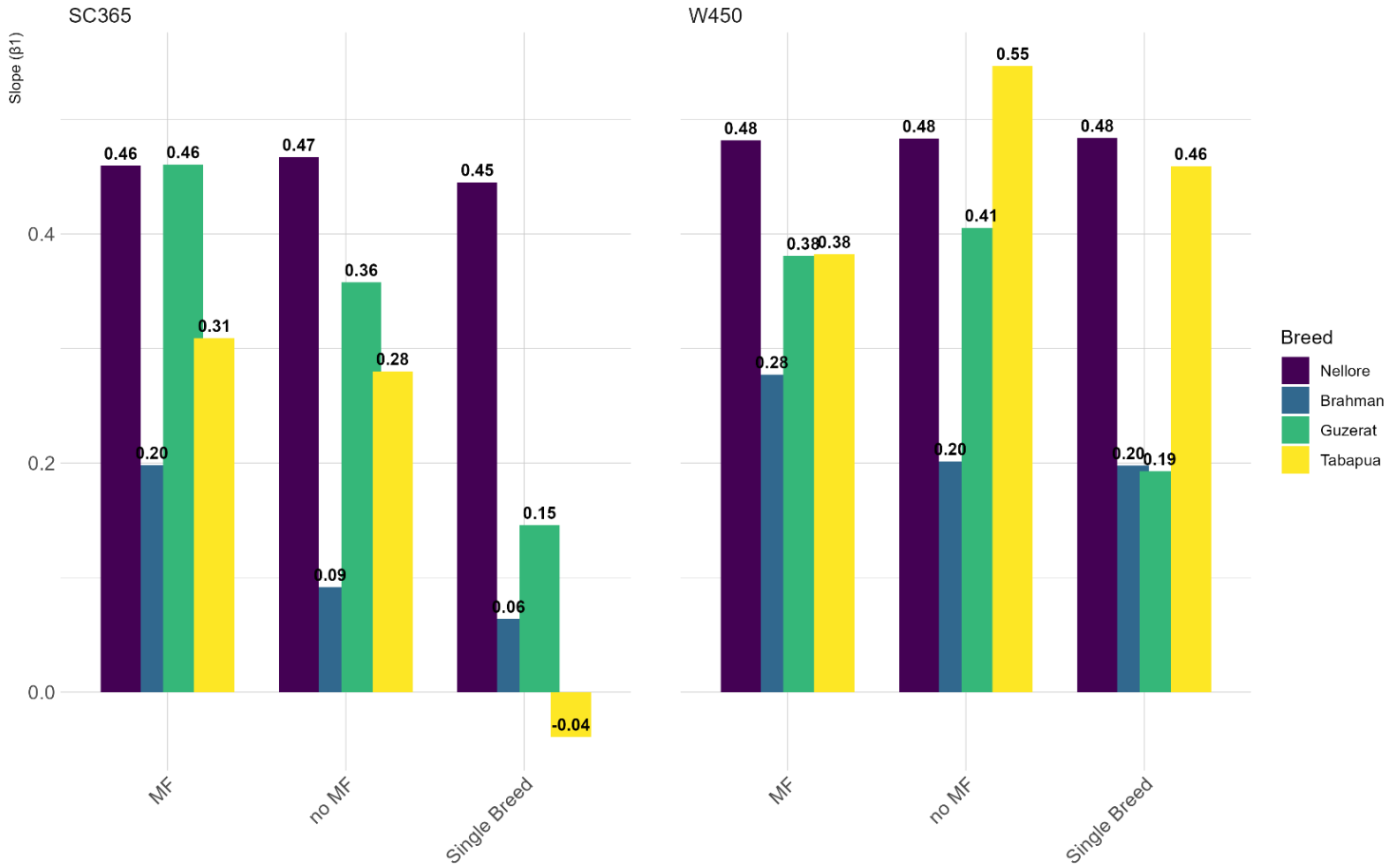


Figure 4.2. Regression coefficient (b_1) of the regression of the accuracy of genomic estimated breeding values ($\mathrm{ACC_{GEBV}}$) on the accuracy of indirect predictions $\mathrm{ACC_{IP}}$, for the Nellore Brahman, Guzerat, and Tabapua breeds, across different scenarios: Multi-breed ssGBLUP with metafounders (MF), Multi-breed ssGBLUP with no metafounders (no MF), and Single-breed ssGBLUP (Single Breed), for the traits scrotal circumference at 365 (SC365) days of age and adjusted weight at 450 (W450) days of age.

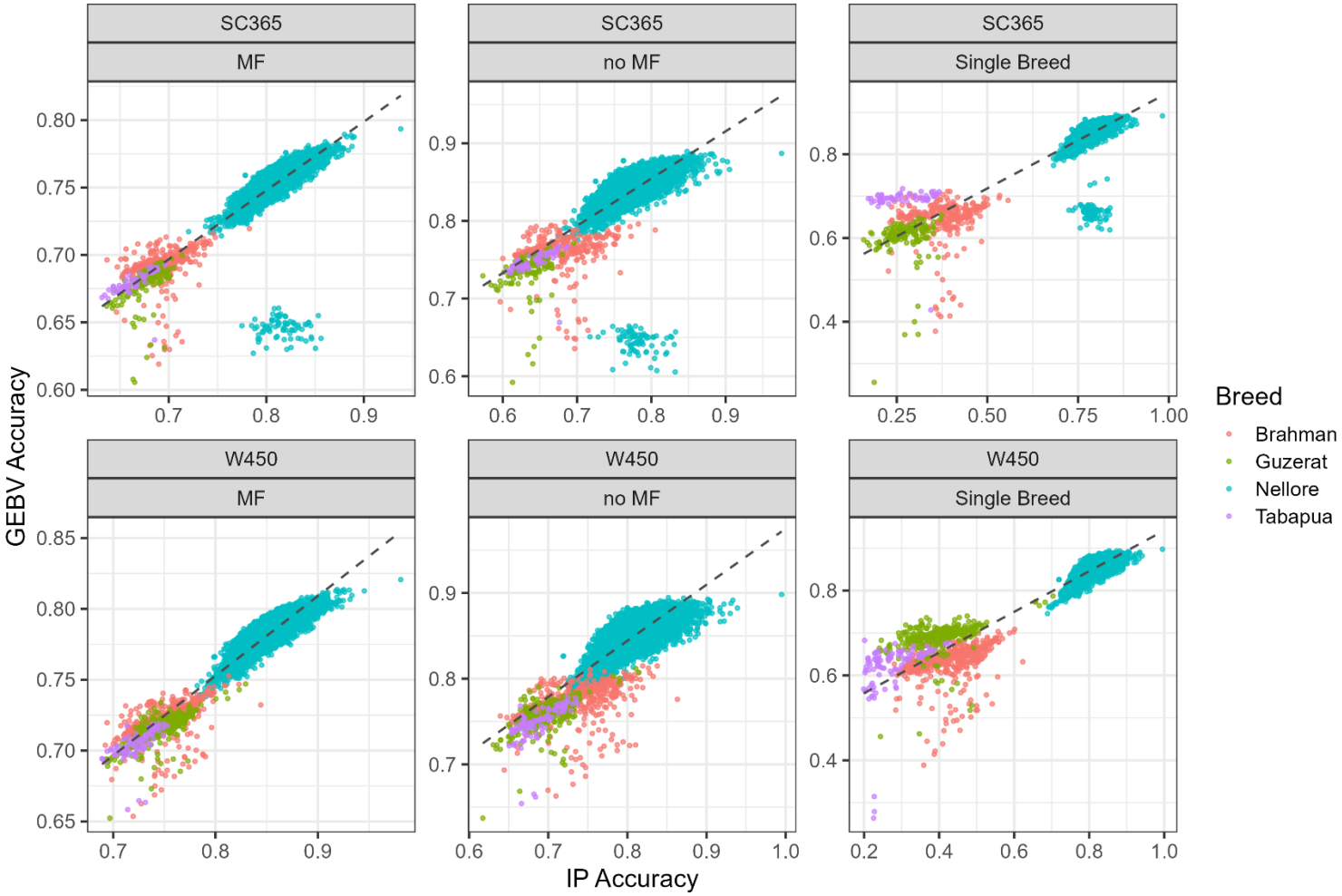


Figure 4.3. Scatter plots of the accuracy of genomic estimated breeding values ($\mathrm{ACC}_{\mathrm{GEBV}}$, y-axis) versus the accuracy of indirect predictions ($\mathrm{ACC}_{\mathrm{IP}}$, x-axis) for Nellore, Brahman, Guzerat, and Tabapua across three ssGBLUP scenarios: Multi-breed with metafounders (MF), Multi-breed with no metafounders (no MF), and Single-breed (Single Breed), for the traits scrotal circumference at 365 (SC365) days of age and adjusted weight at 450 (W450) days of age. Each point represents an individual, color-coded by breed.

Table 4.3. Descriptive statistics of $\mathrm{ACC_{GEBV}}$ and $\mathrm{ACC_{IP}}$ across breeds, traits, and evaluation scenarios

Breed	Trait	Scenario	$\overline{\mathrm{ACC_{GEBV}}}$	Min. $\mathrm{ACC_{GEBV}}$	Max. $\mathrm{ACC_{GEBV}}$	$\overline{\mathrm{ACC_{IP}}}$	Min. $\mathrm{ACC_{IP}}$	Max. $\mathrm{ACC_{IP}}$	Absolute difference	
									Mean	Max.
Nellore	W450	MF	0.79	0.73	0.82	0.86	0.76	0.98	0.07	0.16
		no MF	0.85	0.75	0.90	0.81	0.69	0.99	0.04	0.09
		Single Breed	0.85	0.75	0.90	0.81	0.69	0.99	0.04	0.09
	SC365	MF	0.76	0.63	0.79	0.81	0.72	0.94	0.05	0.15
		no MF	0.84	0.61	0.89	0.78	0.67	0.98	0.06	0.09
		Single Breed	0.85	0.62	0.89	0.79	0.68	0.98	0.06	0.09
Brahman	W450	MF	0.72	0.65	0.75	0.75	0.69	0.84	0.03	0.09
		no MF	0.77	0.66	0.81	0.73	0.64	0.86	0.04	0.05
		Single Breed	0.63	0.39	0.71	0.42	0.27	0.62	0.21	0.09
	SC365	MF	0.69	0.62	0.72	0.70	0.64	0.78	0.01	0.06
		no MF	0.76	0.64	0.80	0.68	0.60	0.81	0.08	0.01
		Single Breed	0.64	0.38	0.71	0.38	0.23	0.56	0.26	0.15
Guzerat	W450	MF	0.71	0.65	0.75	0.74	0.69	0.83	0.03	0.08
		no MF	0.76	0.64	0.81	0.70	0.62	0.81	0.06	0.00
		Single Breed	0.69	0.46	0.79	0.40	0.24	0.70	0.29	0.09
	SC365	MF	0.68	0.61	0.70	0.68	0.64	0.71	0.00	0.01
		no MF	0.74	0.59	0.77	0.64	0.57	0.70	0.10	0.07
		Single Breed	0.60	0.26	0.65	0.27	0.16	0.37	0.33	0.28
Tabapua	W450	MF	0.70	0.66	0.72	0.72	0.69	0.75	0.02	0.03
		no MF	0.62	0.26	0.68	0.27	0.20	0.43	0.35	0.25
		Single Breed	0.75	0.65	0.78	0.69	0.65	0.74	0.06	0.04
	SC365	MF	0.68	0.64	0.69	0.67	0.63	0.70	0.01	0.01
		no MF	0.75	0.67	0.77	0.65	0.61	0.69	0.10	0.08
		Single Breed	0.69	0.43	0.72	0.27	0.17	0.37	0.42	0.35

$\overline{\mathrm{ACC_{GEBV}}}$: Mean accuracy of genomic estimated breeding values; $\overline{\mathrm{ACC_{IP}}}$ mean accuracy of indirect predictions $\mathrm{ACC_{IP}}$; MF: Multi-breed ssGBLUP with metafounders; no MF: Multi-breed ssGBLUP with no metafounders; Single Breed: Single-breed ssGBLUP; W450: adjusted weight at 450 days of age; SC365: scrotal circumference at 365 days of age.

The findings of this study imply that indicine genomic selection programs ought to give preference to multi-breed reference populations, especially for small breeds that have trouble with sparse genotypic data. The use of structured multi-breed evaluations to increase the accuracy of genetic selection was further supported by the improved accuracy of IPs under multi-breed ssGBLUP with MF (Brito et al., 2017; Lourenco et al., 2018).

Additionally, ACCIP computation using SNP-derived PEC estimates has gained recognition as a reliable method for assessing accuracy in genomic selection models. Consistent with the findings of Li et al. (2023), the correlations found between ACCIP and ACCGEBV verify that SNP PEC-based accuracy estimations closely adhere to theoretical expectations. A number of studies have pointed out how important it is to appropriately account for genetic relationships in PEV-based accuracy estimates because PEV computation errors can result in systematic biases in GEBV (Dekkers et al., 2021; van den Berg et al., 2019). Our results demonstrate that, especially in multi-breed evaluations, taking SNP PEC into account when calculating IP accuracy results in more consistent and trustworthy accuracy estimates.

Prior studies have explored indirect prediction accuracy in various contexts. Wang et al. (2012) reported that SNP-based indirect predictions achieved correlations above 0.80 with GEBV in well-structured datasets. In the present study, correlation patterns varied by breed and scenario but remained generally consistent with these findings. Garcia et al. (2022) reported even higher correlations (≥0.99), demonstrating that IP accuracies are compatible with the GEBV accuracies. However, their study was limited to a single-breed evaluation and trait. Additionally, they focused on American Angus, a *Bos taurus* breed for which most genotyping chips are optimized. This likely contributed to more precise SNP effect estimation, enhancing the accuracy of their predictions.

The intricacy of defining accuracy in breeding schemes was brought to light by Bijma (2012), who emphasized that accuracy has distinct functions based on whether it gauges EBV stability or selection response. EBV stability is associated with an animal's standard error, whereas selection response accuracy represents the

relationship between true breeding values and EBVs at the population level. Because selection criteria and intensity can affect observed accuracies, this distinction becomes especially important when thinking about indirect predictions. Although multi-breed models with MF increase IP accuracy in the context of indirect predictions, selection dynamics within breeds may still have an impact on their correlation with the GEBVs. This might occur because the way selection pressures influence genetic variation in various populations affects prediction accuracy in addition to model structure.

Additionally, though it was not our case, ignoring inbreeding in PEV-based accuracy estimates has been shown to cause over- and underestimations, particularly in small populations (Aguilar et al., 2020). Given that inbreeding levels are increasing in indicine cattle breeding programs, further research should explore the direct impact of inbreeding on genomic prediction accuracy, particularly in small and genetically diverse populations.

The findings from this study highlight the importance of multi-breed genomic evaluations, particularly for breeds with limited reference populations. MF improved ACC_{IP} and provided a more stable and reliable framework for estimating genetic relationships. The findings of this study have significant implications for genomic selection programs in indicine cattle in Brazil. The results suggest that multi-breed genomic evaluations, particularly those incorporating MF, should be prioritized in genetic evaluations. These findings contribute valuable insights into optimizing indirect predictions for young genotyped animals, ensuring that breeding programs can achieve higher genetic gains through more reliable selection decisions. Overall, the results suggest that indirect predictions can be a viable tool for genomic selection, provided that appropriate methodologies, such as multi-breed ssGBLUP with MF, are employed.

4.4 CONCLUSION

This study demonstrated that incorporating MF in multi-breed genomic evaluations significantly improved the accuracy of IP in indicine cattle breeds with limited reference populations. This scenario provided the highest accuracy estimates, particularly benefiting smaller breeds like Guzerat and Tabapua by using the larger

reference population of Nellore. However, some bias remains in the accuracy estimation of IP, which requires further investigation.

Additionally, using SNP-derived PEC estimates to compute IP accuracy proved to be a reliable approach, reinforcing the importance of multi-breed evaluations in genomic selection. These results suggest that indirect predictions can be a valuable tool for genomic selection when appropriate methodologies, such as multi-breed ssGBLUP with MF, are applied. Integrating these strategies into breeding programs can enhance selection efficiency, reduce uncertainty, and accelerate genetic progress in the livestock industry.

4.5 REFERENCES

Aguilar, I., Fernandez, E. N., Blasco, A., Ravagnolo, O., & Legarra, A. (2020). Effects of ignoring inbreeding in model-based accuracy for BLUP and SSGBLUP. **Journal of Animal Breeding and Genetics**, 137(4), 356–364. https://doi.org/10.1111/jbg.12470

Aguilar, I., Legarra, A., Cardoso, F., Masuda, Y., Lourenco, D., & Misztal, I. (2019). Frequentist p-values for large-scale-single step genome-wide association, with an application to birth weight in American Angus cattle. **Genetics Selection Evolution**, 51(1). https://doi.org/10.1186/s12711-019-0469-3

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., & Lawlor, T. J. (2010). 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(2), 743–752. https://doi.org/10.3168/jds.2009-2730

Anglhuber, C., Edel, C., Pimentel, E. C. G., Emmerling, R., Götz, K. U., & Thaller, G. (2024). Definition of metafounders based on population structure analysis. **Genetics Selection Evolution**, 56(1). https://doi.org/10.1186/s12711-024-00913-7

Bijma, P. (2012). Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations. **Journal of Animal Breeding and Genetics**, 129(5), 345–358. https://doi.org/10.1111/j.1439-0388.2012.00991.x

Brito, L. F., Clarke, S. M., McEwan, J. C., Miller, S. P., Pickering, N. K., Bain, W. E., Dodds, K. G., Sargolzaei, M., & Schenkel, F. S. (2017). Prediction of genomic breeding values for growth, carcass and meat quality traits in a multi-breed sheep population using a HD SNP chip. **BMC Genetics**, 18(1). https://doi.org/10.1186/s12863-017-0476-8

Christensen, O. F., Madsen, P., Nielsen, B., Ostersen, T., & Su, G. (2012). Single-step methods for genomic evaluation in pigs. **Animal**, 6(10), 1565–1571. https://doi.org/10.1017/S1751731112000742

Cole, J. B., Dürr, J. W., & Nicolazzi, E. L. (2021). Invited review: The future of selection decisions and breeding programs: What are we breeding for, and who decides? **Journal of Dairy Science**, 104(5), 5111–5124. https://doi.org/10.3168/jds.2020-19777

Dekkers, J. C. M., Su, H., & Cheng, J. (2021). Predicting the accuracy of genomic predictions. **Genetics Selection Evolution**, 53(1). https://doi.org/10.1186/s12711-021-00647-w

Garcia, A., Aguilar, I., Legarra, A., Tsuruta, S., Misztal, I., & Lourenco, D. (2022). Theoretical accuracy for indirect predictions based on SNP effects from single-step GBLUP. **Genetics Selection Evolution**, 54(1), 66. https://doi.org/10.1186/s12711-022-00752-4

Gorjanc, G., Bijma, P., & Hickey, J. M. (2015). Reliability of pedigree-based and genomic evaluations in selected populations. **Genetics Selection Evolution**, 47(1). https://doi.org/10.1186/s12711-015-0145-1

Gualdrón Duarte, J. L., Cantet, R. J., Bates, R. O., Ernst, C. W., Raney, N. E., & Steibel, J. P. (2014). Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. **BMC Bioinformatics**, 15(1), 246. https://doi.org/10.1186/1471-2105-15-246

Hayes, B. J., Bowman, P. J., Chamberlain, A. C., Verbyla, K., & Goddard, M. E. (2009). Accuracy of genomic breeding values in multi-breed dairy cattle populations. **Genetics Selection Evolution**, 41(1), 1–9. https://doi.org/10.1186/1297-9686-41-51

Henderson, C. R. (1984). **Applications of Linear Models in Animal Breeding** (1st ed.). Guelph University.

Jin, L., Xu, L., Jin, H., Zhao, S., Jia, Y., Li, J., & Hua, J. (2024). Accuracy of Genomic Predictions Cross Populations with Different Linkage Disequilibrium Patterns. **Genes**, 15(11). https://doi.org/10.3390/genes15111419

Khansefid, M., Goddard, M. E., Haile-Mariam, M., Konstantinov, K. V., Schrooten, C., de Jong, G., Jewell, E. G., O'Connor, E., Pryce, J. E., Daetwyler, H. D., & MacLeod, I. M. (2020). Improving Genomic Prediction of Crossbred and Purebred Dairy Cattle. **Frontiers in Genetics**, 11. https://doi.org/10.3389/fgene.2020.598580

Legarra, A., Aguilar, I., & Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. **Journal of Dairy Science**, 92(9), 4656–4663. https://doi.org/10.3168/jds.2009-2061

Legarra, A., Lourenco, D., & Vitezica, Z. (2024). **Bases for Genomic Prediction** (pp. 1–167).

Li, L., Gurman, P. M., Swan, A. A., & Tier, B. (2023). Approximating prediction error variances and accuracies of estimated breeding values from a SNP–BLUP model for genotyped individuals. **Animal Production Science**, 63(11), 1086–1094. https://doi.org/10.1071/an23027

Londoño-Gil, M., López-Correa, R., Aguilar, I., Magnabosco, C. U., Hidalgo, J., Bussiman, F., Baldi, F., & Lourenco, D. (2024). Strategies for genomic

predictions of an indicine multi-breed population using single-step GBLUP. **Journal of Animal Breeding and Genetics**. https://doi.org/10.1111/jbg.12882

Lourenco, D. A. L., Legarra, A., Tsuruta, S., Moser, D., Miller, S., & Misztal, I. (2018). Tuning Indirect Predictions Based on SNP Effects from Single-Step GBLUP. **Interbull Bulletin**, 53, 48–53. https://journal.interbull.org/index.php/ib/article/view/1448/1511

Lourenco, D., Legarra, A., Tsuruta, S., Moser, D., Miller, S., & Misztal, I. (2018). Tuning indirect predictions based on SNP effects from ssGBLUP.

Lourenco, D., Tsuruta, S., Masuda, Y., Berman, M., Legarra, A., & Misztal, I. (2022, July 3). Recent updates in the BLUPF90 software suite. **World Congress on Genetics Applied to Livestock Production**: 3 - 8 July.

Misztal, I., Tsuruta, S., Lourenco, D. a L., Aguilar, I., Legarra, A., & Vitezica, Z. (2015). **Manual for BLUPF90 family of programs**. 125.

Misztal, I., Tsuruta, S., Lourenco, D., Aguilar, I., Legarra, A., & Vitezica, Z. (2014). **BLUPF90 family of programs**. In Athens: University of Georgia (p. 125). http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all2.pdf

Mohammad Rahimi, S., Rashidi, A., & Esfandyari, H. (2020). Accounting for differences in linkage disequilibrium in multi-breed genomic prediction. **Livestock Science**, 240. https://doi.org/10.1016/j.livsci.2020.104165

Strandén, I., & Christensen, O. F. (2011). Allele coding in genomic evaluation. **Genetics Selection Evolution**, 43(1). https://doi.org/10.1186/1297-9686-43-25

Strandén, I., & Garrick, D. J. (2009). Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. **Journal of Dairy Science**, 92(6), 2971–2975. https://doi.org/10.3168/jds.2008-1929

Takeda, M., Inoue, K., Oyama, H., Uchiyama, K., Yoshinari, K., Sasago, N., Kojima, T., Kashima, M., Suzuki, H., Kamata, T., Kumagai, M., Takasugi, W., Aonuma, T., Soma, Y., Konno, S., Saito, T., Ishida, M., Muraki, E., Inoue, Y., … Uemoto, Y. (2021). Exploring the size of reference population for expected accuracy of genomic prediction using simulated and real data in Japanese Black cattle. **BMC Genomics**, 22(1). https://doi.org/10.1186/s12864-021-08121-z

Tier, B., Meyer, K., & Swan, A. A. (2018). On implied genetic effects, relationships and alternate allele coding. **Proceedings of the World Congress on Genetics Applied to Livestock Production**, 11, 127.

van den Berg, I., Meuwissen, T. H. E., MacLeod, I. M., & Goddard, M. E. (2019). Predicting the effect of reference population on the accuracy of within, across, and multibreed genomic prediction. **Journal of Dairy Science**, 102(4), 3155–3174. https://doi.org/10.3168/jds.2018-15231

Van Vleck, L. D. (1993). **Selection index and introducion to mixed model methods for genetic improvement of animals: The green book** (SF105.26.V). CRC Press, Inc.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. **Journal of Dairy Science**, 91(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

Wang, H., Misztal, I., Aguilar, I., Legarra, A., & Muir, W. M. (2012). Genome-wide association mapping including phenotypes from relatives without genotypes. **Genetics Research**, 94(2), 73–83. https://doi.org/10.1017/S0016672312000274

Xu, L., Wang, Z., Zhu, B., Liu, Y., Li, H., Bordbar, F., Chen, Y., Zhang, L., Gao, X., Gao, H., Zhang, S., Xu, L., & Li, J. (2019). Theoretical evaluation of multi-breed genomic prediction in Chinese indigenous cattle. **Animals**, 9(10). https://doi.org/10.3390/ani9100789

CHAPTER 5 – FINAL CONSIDERATIONS

Animal breeding has gone through a revolution thanks to the development of genomic selection, especially for economically important livestock like indicine cattle. Focusing on the breeds Nellore, Brahman, Guzerat, and Tabapua, this thesis presents an in-depth investigation of genomic predictions made using the single-step genomic best linear unbiased predictor (ssGBLUP) in a multi-breed context. The results demonstrated the possibility of improved genetic evaluation accuracy, especially for breeds with sparse genomic information.

5.1 KEY FINDINGS

Comparing multi-breed genomic evaluations to single-breed evaluations, this study demonstrated that the former greatly increased the accuracy of breeding value estimates. Among the main conclusions were:

- **Improved Accuracy**: For smaller breeds, especially Guzerat and Tabapua, which historically have limited reference populations, the use of multi-breed evaluations and the addition of metafounders (MF) led to higher prediction accuracy. This improvement is particularly noticeable in indirect prediction (IP) accuracy, where reliability is increased by combining MF with a multi-breed model. Because Guzerat and Tabapua gain predictive power by using data from breeds with larger reference populations, such as Nellore, the advantages of this approach are most noticeable in their context.
- **Mitigation of Bias**: The use of metafounders contributed to reducing the bias of the genomic predictions, leading to more consistent and reliable estimations across breeds. However, despite these increases, some residual bias remained, particularly in breeds with smaller datasets, indicating the need for further methodological refinement and additional research.
- **Feasibility of Indirect Predictions**: The study confirmed the effectiveness of using indirect predictions for young genotyped animals that lack phenotypic data or that come from commercial herds. This is especially important for the commercial programs, where early selection decisions can be made with confidence thanks to the improved accuracy provided by the multi-breed

ssGBLUP model. Moreover, the application of SNP-based prediction error covariance (PEC) showed to be a good strategy for estimating accuracy in these scenarios, supporting its use as a standard tool in future evaluations.

5.2 IMPLICATIONS FOR BREEDING PROGRAMS

Beyond scholarly contributions, this work has practical tools and insights for livestock breeding programs, especially in Brazil, the largest beef producer globally. The adoption of multi-breed genomic evaluations and the further use of metafounders is recommended as a standard strategy in breeding programs for the following reasons:

- **Enhanced genetic progress:** Multi-breed evaluations contribute to faster genetic gain since allow more precise selection decisions. This is particularly crucial for economically significant traits in underrepresented breeds. In addition to improving prediction accuracy, the combination of ssGBLUP and MF lowers genetic estimate uncertainty, enabling more intelligent and efficient selection.
- **Increased efficiency:** The selection process and computational demands become more efficient when young animals without phenotypic records are excluded from the official genomic evaluation and their genetic merit is instead estimated through IP based on prior evaluations. This strategy is particularly valuable in countries like Brazil, where routine national genomic evaluations are limited. The IP can reduce the computational costs while enable accurate early selection, accelerating the identification of superior candidates and enhancing the overall efficiency of breeding programs by optimizing both time and resource allocation.
- **Support for small breeds:** The data gap for smaller breeds is filled by utilizing larger reference populations, like the Nellore population. In order to preserve biodiversity and ensure the long-term viability of the beef production industry, this encourages the conservation and genetic improvement of breeds such as Guzerat and Tabapua.

Importantly, this work has already been translated into practice. The ANCP (Associação Nacional de Criadores e Pesquisadores) has adopted the findings and incorporated them into the 2024 Sire report (available on May 20, 2025 in:

https://www.ancp.org.br/wp/wp-content/uploads/2024/08/Sum%C3%A1rio-ANCP-2024_FINAL_09082024_compressed-2.pdf). This report allows producers, especially those working with smaller breeds, to make more informed and precise selection decisions, reinforcing the practical relevance and real-world impact of this research.

5.3 FUTURE RESEARCH DIRECTIONS

To build upon these results and address remaining challenges, future research should explore the following areas:

- **Broader applications**: Expanding multi-breed genomic evaluations to other livestock breeds and production systems could provide a more comprehensive understanding of the potential of genomic selection in the country.
- **Longitudinal studies**: Long-term evaluations of how multi-breed genomic selection affects genetic improvement, and economic gain would provide crucial information about how sustainable these strategies are.
- **Integration of advanced technologies**: New tools such as machine learning or artificial intelligence may help future prediction models, increasing accessibility for small breeds, and allow for adaptation to diverse breeding environments.
- **Focus on environmental adaptability:** With growing concerns about climate change, there is a big need to focus on traits associated with environmental sustainability and resilience. Future genomic evaluations should be focused on these traits to improve future livestock systems. However, there might be limitations due to the data records that could be not balanced or present for different traits in the different breeds.

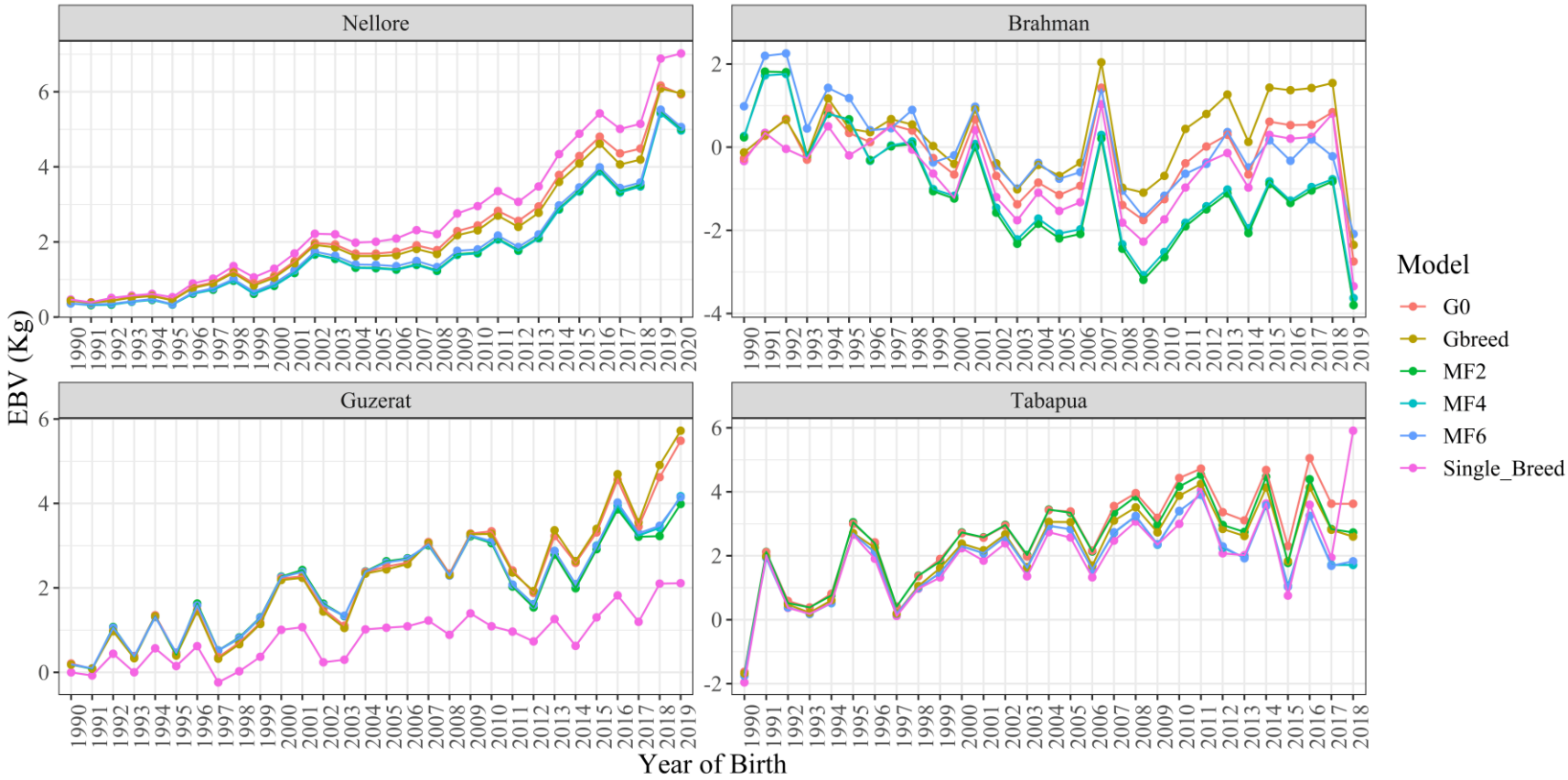
5.4 CONCLUSION

In conclusion, this thesis demonstrated the potential improvement of genomic selection in indicine cattle breeding programs of Brazil using multi-breed strategies and innovative tools like metafounders and IP. Through these strategies, it is possible to achieve more accurate and less biased genetic evaluations. This will not only

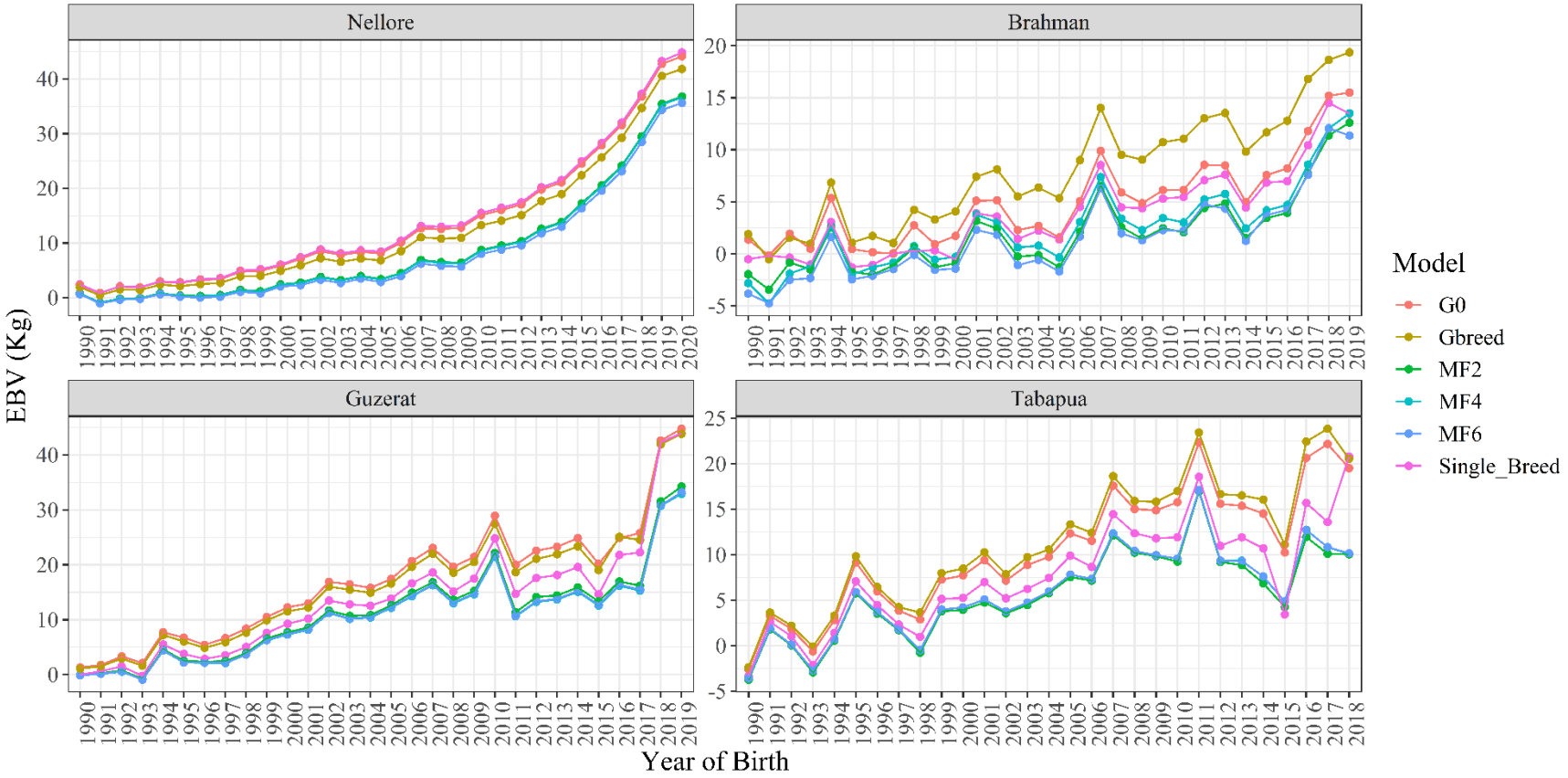
accelerate genetic progress but will ensure more sustainable animal breeding, with greater profitability.

Most significantly, this research has already influenced real-world practice, as shown by its implementation in the 2024 ANCP Sire Report. The results of this study act as a catalyst for innovation, closing the gap between research and on-farm decision-making, by offering useful, scientifically based solutions, particularly for breeds with smaller datasets.

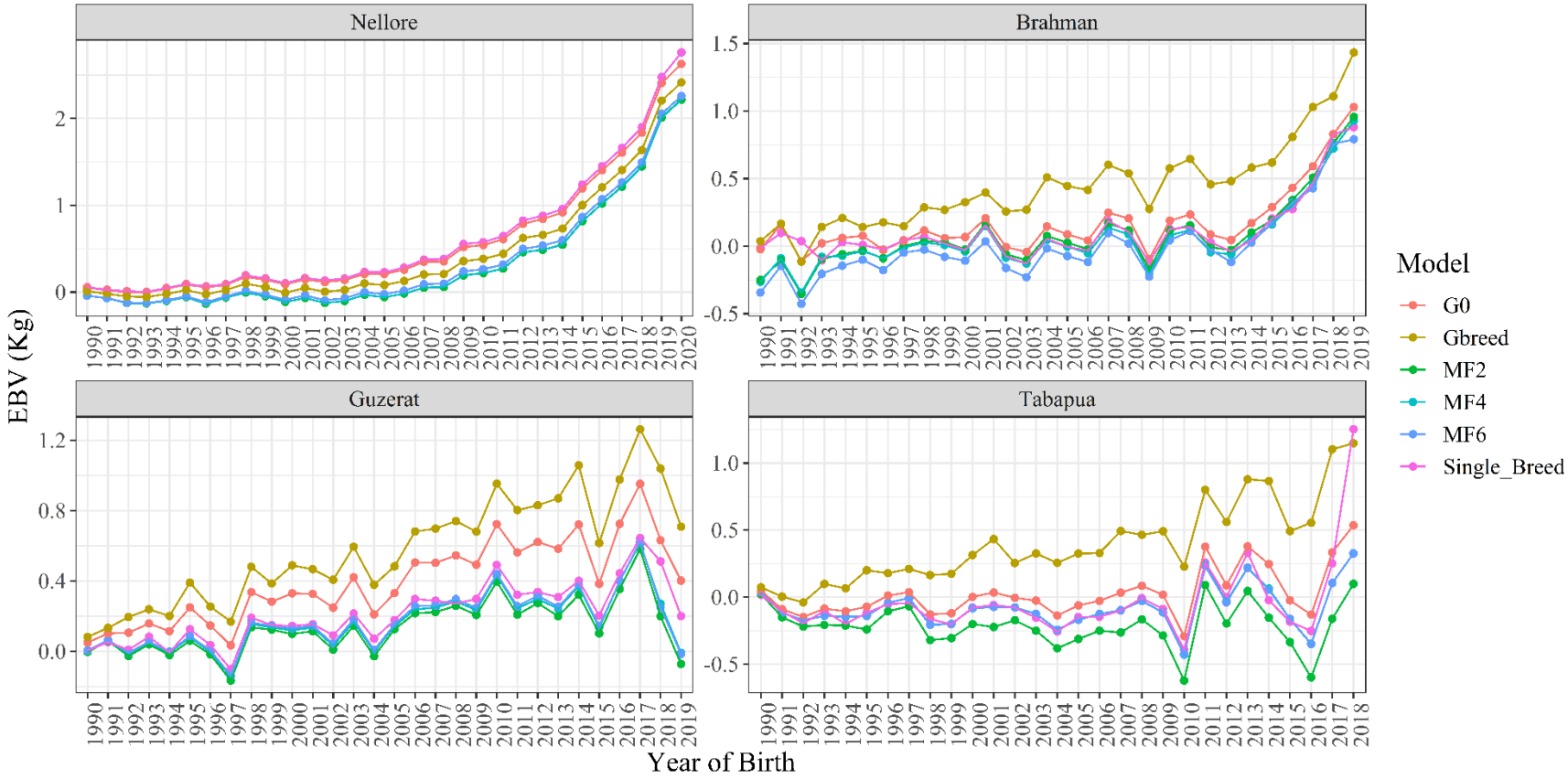
APPENDIX



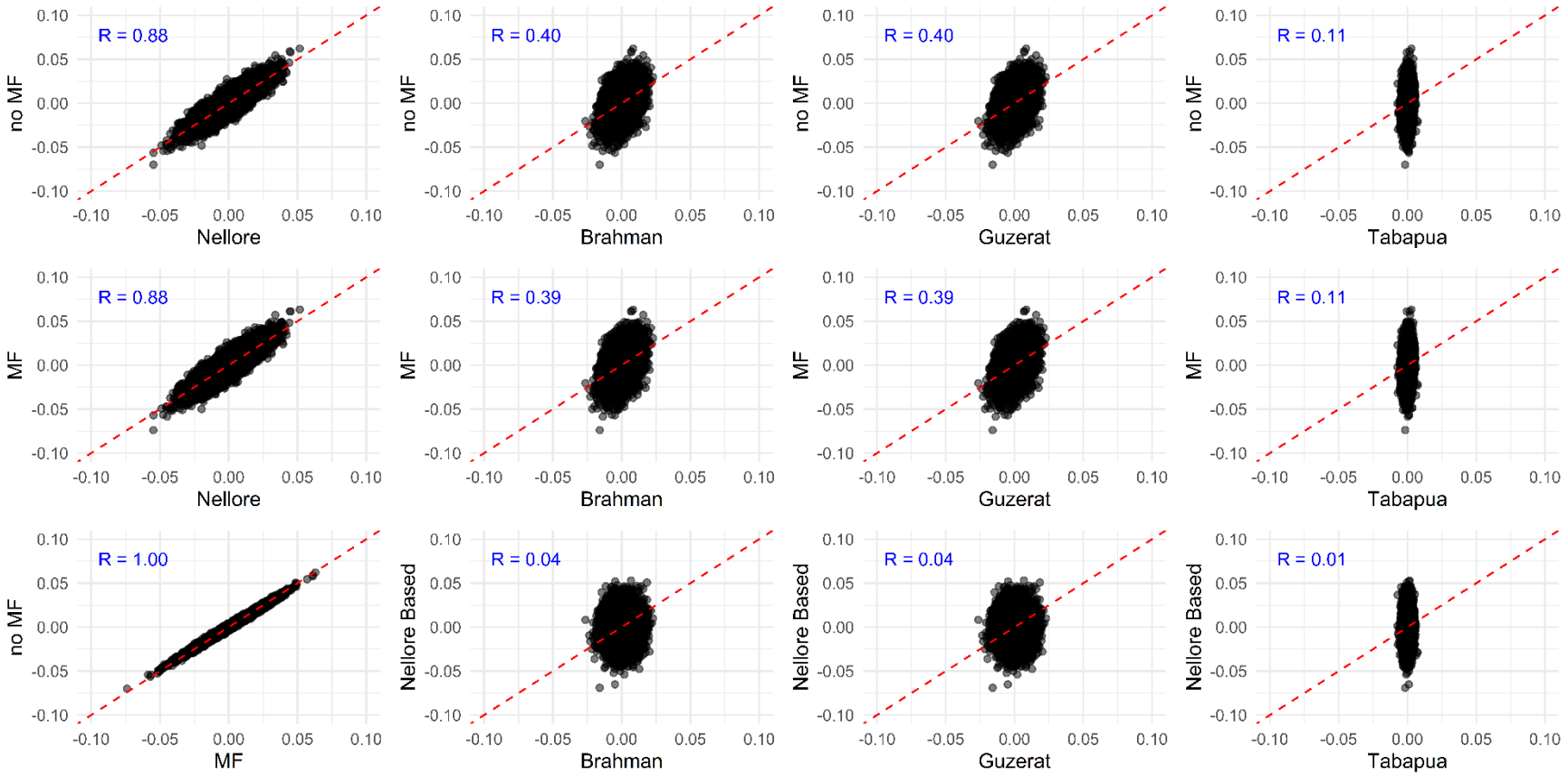
Supp. Figure 2.1. Maternal genetic trends for the trait adjusted weight at 210 (W210) days of age for Bulls of the Breeds Nelore, Brahman, Guzerat and Tabapua, in the six models evaluated.



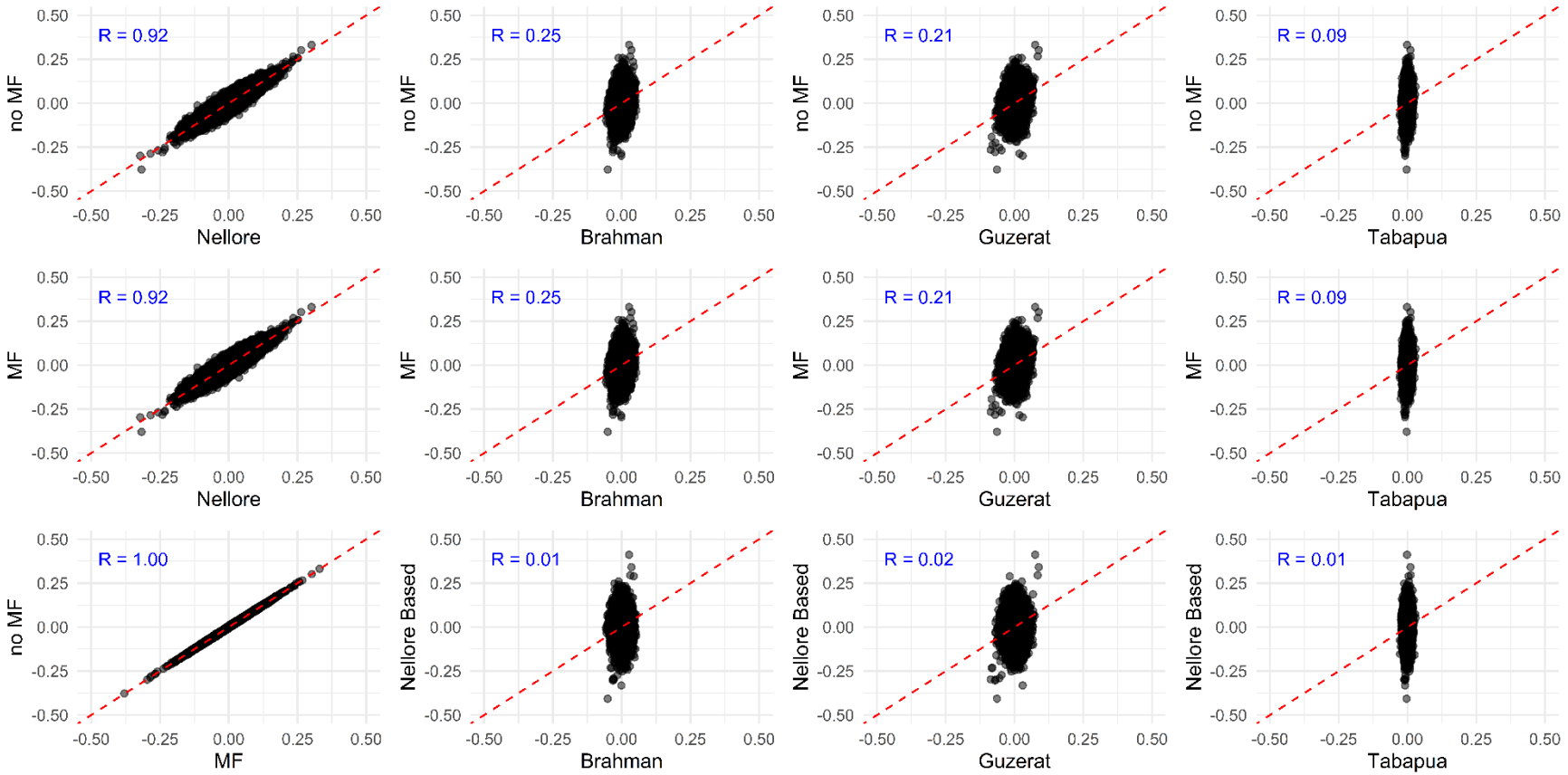
Supp. Figure 2.2. Genetic trends for the trait adjusted weight at 450 (W450) days of age for Bulls of the Breeds Nelore, Brahman, Guzerat and Tabapua, in the six models evaluated.



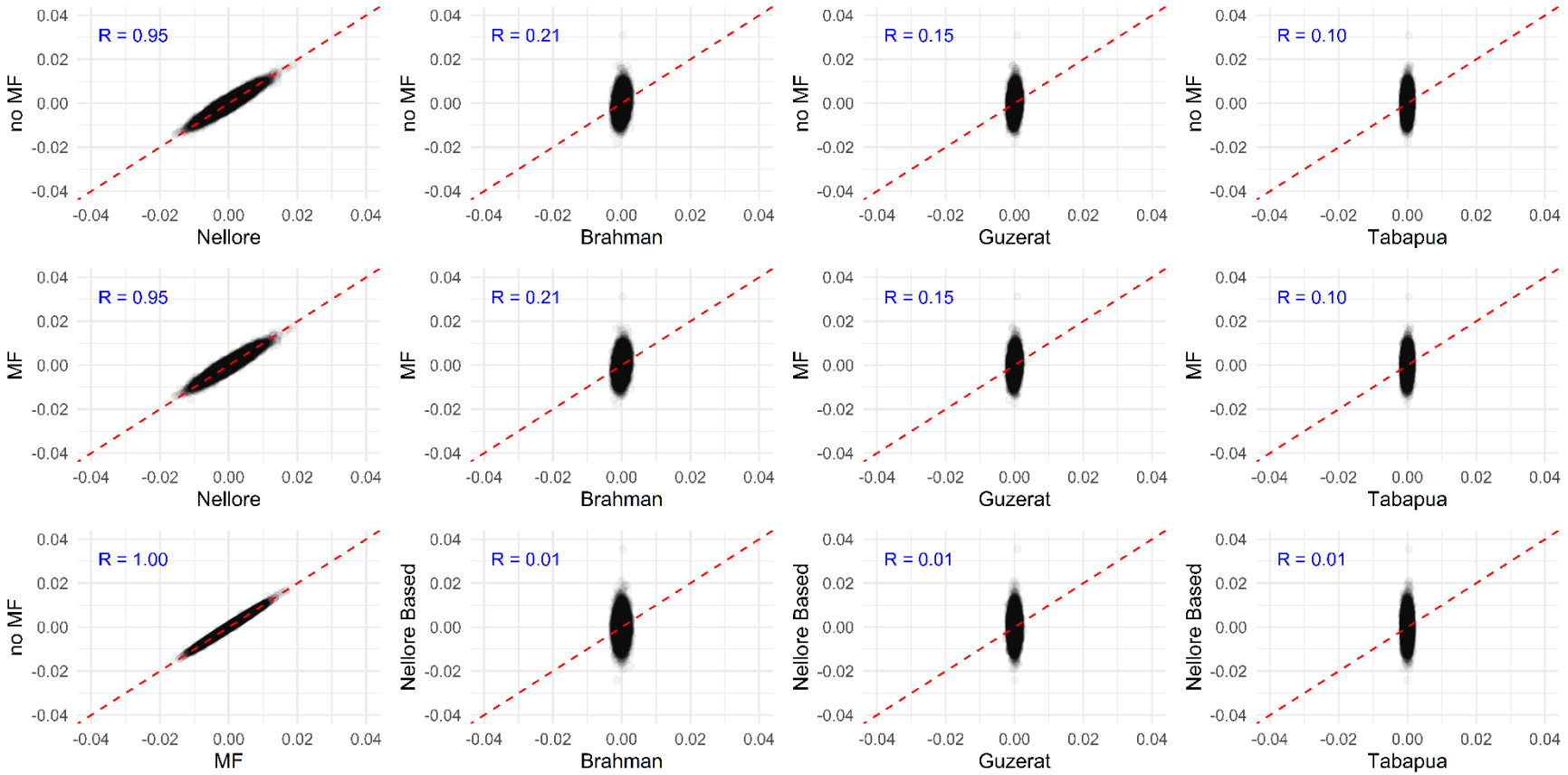
Supp. Figure 2.3. Genetic trends for the trait scrotal circumference at 365 (SC365) days of age for Bulls of the Breeds Nelore, Brahman, Guzerat and Tabapua, in the six models evaluated.



Supp. Figure 3.1. SNP effects correlations (R) between scenarios for the trait maternal adjusted weight at 210 days of age (W210m). No MF: SNP effects based on multi-breed GEBV without metafounders and using the default **G**, i.e., with current allele frequencies. MF: SNP effects based on GEBV obtained when using four metafounders, one for each breed. Nellore based: based on the Nellore GEBV for the small breeds Brahman, Guzerat and Tabapua. Single Breed (breed name is shown): based on each breed GEBV, when evaluated as a single breed. MF and no MF were both multi-breed scenarios.



Supp. Figure 3.2. SNP effects correlations (R) between scenarios for the trait adjusted weight at 450 days of age (W450). No MF: SNP effects based on multi-breed GEBV without metafounders and using the default **G**, i.e., with current allele frequencies. MF: SNP effects based on GEBV obtained when using four metafounders, one for each breed. Nellore based: based on the Nellore GEBV for the small breeds Brahman, Guzerat and Tabapua. Single Breed (breed name is shown): based on each breed GEBV, when evaluated as a single breed. MF and no MF were both multi-breed scenarios.



Supp. Figure 3.3. SNP effects correlations (R) between scenarios for the trait scrotal circumference at 365 days of age (SC365). No MF: SNP effects based on multi-breed GEBV without metafounders and using the default **G**, i.e., with current allele frequencies. MF: SNP effects based on GEBV obtained when using four metafounders, one for each breed. Nellore based: based on the Nellore GEBV for the small breeds Brahman, Guzerat and Tabapua. Single Breed (breed name is shown): based on each breed GEBV, when evaluated as a single breed. MF and no MF were both multi-breed scenario