THE UNIVERSITY *of* EDINBURGH

College of Medicine and Veterinary Medicine
The Roslin Institute and Royal (Dick) School of Veterinary Studies
University of Edinburgh

This thesis is presented for the degree of

Doctor of Philosophy

Methods for utilising genomic diversity in tropical dairy breeding

submitted by

Gabriela Mafra Fortuna
MSc Animal Breeding and Genetics
(Georg-August-Universität Göttingen,
Swedish University of Agricultural Sciences)
BVetMed (Universidade Federal Fluminense)

Supervisor: **Prof. Gregor Gorjanc**

August, 2025

Declaration

I declare that this thesis and the data presented in it are original and my own work, unless otherwise specified. This work has not been submitted for any other degree or professional qualification except as specified. I have read and understood The University of Edinburgh guidelines on Plagiarism and declare that this written dissertation is all my own work except where I indicate otherwise by proper use of quotes and references. This thesis is an account of work conducted by me whilst studying for the degree of Doctor of Philosophy at the University of Edinburgh.

Gabriela Mafra Fortuna
August 2025

List of publications

Conference attendance

Mafra Fortuna, G., Gorjanc, G., Tolhust, D. The importance of trait stability in crossbreeding: uncovering the impacts of genotype-by-environment interaction. In: 76th Annual Meeting of the European Federation of Animal Science (EAAP), Innsbruck, AUS, August 25-29th, 2025. Presentation.

Mafra Fortuna, G., Obšteter, J., Moškrič, A., Gorjanc, G. Estimating haplotype and mutation effects in the context of genome sequence via ancestral recombination graphs. In: 7th International Conference of Quantitative Genetics (ICQG), Vienna, AUS, July 22-26th, 2024. Poster.

Mafra Fortuna, G., Obšteter, J., Moškrič, A., Gorjanc, G. Estimating haplotype and mutation effects in the context of genome sequence via ancestral recombination graphs. In: 75th Annual Meeting of the European Federation of Animal Science (EAAP), Florence, ITA, September 1-5th, 2024. Presentation.

Houaga, I., **Mafra Fortuna, G.**, Obšteter, J., Kambal, S., Tijjani, A., Hanotte, O., Pocrnic, I., Gorjanc, G. African indigenous cattle genealogy revealed by whole-genome tree sequences. In: 75th Annual Meeting of the European Federation of Animal Science (EAAP), Florence, ITA, September 1-5th, 2024. Presentation.

Mafra Fortuna, G., Obšteter, J., Kranis, A., Gorjanc, G. Exploring global cattle genealogy with tree sequence. In: 74th Annual Meeting of the European Federation of Animal Science (EAAP), Lyon, FRA, August 26th-September 1st, 2023. Presentation.

Gorjanc, G., Obšteter, J., **Mafra Fortuna, G.**, Ros-Freixedes, R., Johnsson, M., Pocrnic, I. Storing and analysing a million genomes on a desktop computer. In: 74th

Annual Meeting of the European Federation of Animal Science (EAAP), Lyon, FRA, August 26th-September 1st, 2023. Presentation.

Mafra Fortuna, G., Obšteter, J., Kranis, A., Gorjanc, G. Studying global cattle genealogy using tree sequence. In: British Society of Animal Science Annual Conference (BSAS), Birmingham, United Kingdom, March 28-30th, 2023. Key-note speaker.

Mafra Fortuna, G., Zumbach, B.J., Johnsson, M., Pocrnic, I., Gorjanc, G. Accounting for nuclear- and mito-genome in genetic evaluation and breeding of dairy cattle. In: Proceedings of the 12th World Congress on Genetics Applied to Livestock Production (WCGALP), Rotterdam, The Netherlands, July 3-8th, 2022. Presentation.

Abstract

Dairy production plays a crucial role in addressing food insecurity and poverty in tropical regions by providing high-value protein and generating income for millions of smallholder farmers. It also empowers women and girls, strengthening rural communities, and contributing to sustainable economic development.

Tropical dairy production commonly relies on crossbreeding environmentally-adapted local *Bos indicus* (indicine) breeds with high-yielding exotic breeds, usually from *Bos taurus* (taurine) origin. This strategy aims to increase productivity by leveraging breed complementarity and heterosis. However, crossbred performance is highly variable, to the point of instability. This instability undermines the success of crossbreeding by posing short-term challenges to production management and long-term challenges to the optimisation of selective breeding.

This thesis explores three fundamental aspects behind the challenges of tropical crossbreeding strategies: (i) the genetic distance between environmentally-adapted local breeds and high-yielding exotic breeds, (ii) the lack of statistical methods tailored to dealing with such genetic distance, and (iii) the genotype-by-environment (G×E) interaction that underlies the instability in crossbred performance across environments. These fundamental aspects address the long-standing objective of breeders to effectively utilise genetic variation in tropical dairy breeding. The work presented here seeks to introduce novel methods for uncovering and leveraging the genetic variation available in tropical dairy systems. The thesis is structured as follows.

Chapter 1 provides a review of the evolutionary and breeding history of cattle. It outlines the origins of genetic divergence between populations adapted to the tropical climates of the Global South and those favoured in the Global North. The chapter also introduces the concepts of Ancestral Recombination Graphs (ARG) and multiplicative models; two methodological approaches used throughout the thesis to analyse genetic

diversity, the effects of genetic diversity on phenotypic performance, and the effects of G×E interaction. Particular focus is paid to the inherent differences between pedigree and genomic-based models, and the benefits of ARG-based inference. The chapter concludes by stating the research objectives and the thesis structure.

Chapter 2 investigates global patterns of genetic diversity and population structure in cattle. This chapter demonstrates how the fast-evolving field of ARG reconstruction can benefit livestock genomics, providing an information-rich new format for storing and analysing genomic data. The results show that tree-sequence-based ARGs capture fine-scale population structure across cattle populations worldwide. In particular, local ancestry inference with ARGs enables the assignment of breed-of-origin that can inform crossbred genomic evaluation, especially in contexts of complex ancestry composition.

Chapter 3 introduces a novel ARG-based statistical model for estimating haplotype and ancestry-specific mutation effects considering the genome sequence context.

The genomic distance between indicine and taurine cattle breeds underlie variation in their adaptation and performance. The distinct selective pressures experienced by these populations over time have shaped the genomic context in which mutations occur, resulting in ancestry-specific mutations and their effects. These differences limit the predictive accuracy of crossbreeding and multi-breed genetic evaluations, as current methods often assume mutations and their effects are shared across populations.

By capturing the historical recombination, mutation, and coalescence processes that shape genetic variation, ARGs offer a biologically informed basis for modelling mutation effects. The proposed model, initially developed for a single, non-recombining genomic region, achieves predictive accuracy comparable to standard SNP-based approaches while reducing computational demands and providing additional biological insights. This work lays the foundation for future expansion to more complex scenarios including multiple recombining regions.

Another limitation addressed in this thesis is the instability of crossbred performance across environments arising from G×E interaction, which is often ignored in tropical dairy systems. G×E interaction reflects the variable genotype performance between environments, and when not managed, reduces the predictability of the deviation in trait expression. This phenomenon increases genetic variance and alters genotype ranking across environments, hindering the optimisation of production.

Chapter 4 develops a novel framework for decomposing genetic variance that exposes

the variation due to G×E interaction. The framework leverages a rotated multiplicative model and subsequent visualisation approaches commonly used in plant breeding but rarely explored by animal breeders. The outputs of this analysis express the genetic variance in a way that is more accessible and actionable for breeding decisions, providing the means to identify broadly and specifically adapted individuals.

Using stochastic simulations, the chapter demonstrates how G×E interaction affects crossbreeding outcomes and how adaptability patterns of underutilised genetic diversity in the exotic population can improve crossbreeding responses to specific environments. This result indicates that the correlation between environments secures genetic resources despite different breeding objectives in these environments. The chapter serves as a recommendation on how to apply the framework to inform breeding decisions, focusing on leveraging useful genetic diversity to produce crossbred that are stable yet responsive to changes in environment.

Overall, my thesis introduces new methodologies to the context of dairy breeding that contribute to the analysis and understanding of genetic diversity and provides practical tools for uncovering useful genetic variation and analysing its effect on adaptation and performance in tropical environments. **Chapter 5** provides a general discussion of the opportunities and challenges for future work in tropical dairy breeding.

Lay summary

Dairy farming has a significant socioeconomic value worldwide, being especially relevant in tropical countries, where it contributes to tackling poverty by providing good-quality food and generating income for millions of smallholder farmers. In these regions, dairy farming is often a family activity, with women and girls playing a central role, which makes the activity fundamental to their empowerment. However, climate and economic development pose challenges to tropical dairy production.

Globally, there are two major types of cattle: indicine (*Bos indicus*), which are adapted to the heat and humidity of tropical climates and can sustain milk production even when resources are scarce, but produce low volume of milk; and taurine (*Bos taurus*), which produce high volumes of milk, but depend on intensive farming conditions, often indoors with controlled temperature and high-quality feed. To boost productivity, tropical dairy farmers commonly cross the indicine and taurine cattle. This strategy is known as crossbreeding, and the expected outcome is a cow that produces more milk than indicine cows while still being able to cope with the tropical environment. However, the performance of crossbreds is often variable; different progeny from the same cross perform very differently, deviating from the expected outcome. Further generations of crossing deepen this variation. This variability challenges farming and breeding decisions, compromising the long-term success of tropical dairy.

This thesis explores three fundamental aspects behind the challenges of tropical crossbreeding strategies:

1. **The genetic distance between indicine and taurine breeds:** Indicine and taurine cattle have evolved under different environmental pressures for thousands of years, undergoing separate domestication and breeding events. These differences have imposed distinct pressure on their DNA, resulting in differences in their genetic makeup. When crossbreeding, these differences are exposed as in-

creased genetic diversity, which has implications for how these animals express desirable traits, but also provides a rich genetic resource for selective breeding.

2. **The lack of statistical methods tailored to dealing with such genetic distance:** Current statistical methods for estimating the genetic merit of animals based on their ability to perform, known as breeding values, often assume that the genetic makeup is comparable across populations. Specifically, the methods assume mutations occur in all populations with the same effects on the expression of traits. This leads to inaccurate predictions of breeding values for crossbred animals which inherit DNA from parental populations with different genetic makeups.

3. **The genotype-by-environment (G×E) interaction that underlies the instability in crossbred performance across environments:** G×E interaction is the phenomenon in which the environment where the animal is placed affects how its DNA is expressed, for example, how much milk they produce or how well they cope with heat stress. This means that the performance of the same animal can vary greatly across different environments; additionally, these responses to environments can vary between animals. This source of variation challenges breeding decisions because it is hard to predict how crossbred animals will perform in a specific environment.

These fundamental aspects address the long-standing objective of breeders to effectively utilise the genetic variation within and outwith tropical dairy systems. This thesis introduces novel methods for uncovering and leveraging the genetic variation available to cattle breeders in tropical dairy systems. The thesis is structured as follows.

Chapter 1 provides a review of the history of cattle, outlining its origins and the consequences of the differences in genetic makeup between indicine and taurine cattle and the importance and challenges of tropical dairy crossbreeding. It also introduces methodological concepts used throughout the thesis to analyse (i) genetic diversity, (ii) the effects of genetic diversity on performance, and (iii) the effects of GxE interaction on performance and breeding decisions.

Chapter 2 investigates the genetic variation across cattle populations from around the world. The chapter applies a new methodology for reconstructing historical paths of DNA inheritance, providing local genealogical trees along segments of DNA for each individual. This output is encoded as an Ancestral Recombination Graph (ARG) object and its operational data format, called a tree sequence (denoting a sequence of

genealogical trees along DNA). This methodology offers important insights into the origin of DNA segments while providing a parsimonious way to store and analyse genomic data. The results capture fine-scale population structure and enable tracing the origin of mutations and how they segregate in populations over time. This information is especially relevant in contexts where many breeds are used for crossbreeding, or the origin of individuals is unknown.

Chapter 3 focuses on the lack of statistical methods that can analyse the genetic differences between indicine and taurine cattle. The chapter introduces a new statistical model that leverages the genealogical trees described above for individual DNA segments and the rich information they provide to estimate breeding values and mutation effects, which can differ between populations used in crossbreeding.

Chapter 4 develops a framework to disentangle the effects of GxE interaction on crossbred performance. Using state-of-the-art simulations, the chapter demonstrates the impact of GxE on crossbreeding outcomes and uncovers patterns of adaptability to tropical environments in the taurine population that are otherwise underutilised. The outputs of the framework convey the genetic variation resulting from the GxE interaction in a more informative and actionable way for breeding decisions, providing the means for identifying potential crosses of individuals that are more likely to generate offspring with high yield and stable performance across environments.

Overall, my thesis introduces new animal breeding methodologies that contribute to the analysis and understanding of genetic diversity across populations. It also provides practical tools for uncovering useful genetic variation and analysing its effect on adaptation and performance in different environments. **Chapter 5** provides a general discussion of the opportunities and challenges for future work in tropical dairy breeding.

Acknowledgments

This PhD was a journey that started way before being awarded the position and the scholarship. For getting me here, and getting me through it, I acknowledge several people. Firstly, I am immensely grateful to my parents **Angela** and **Ricardo** and my sister **Giovana**, who nursed my curiosity and interest in science since I was a child. Thank you for supporting my international endeavors and for always being there for me.

I sincerely thank my supervisors, **Prof. Gregor Gorjanc**, **Prof. James Prendergast** and **Dr. Andreas Kranis**. Gregor, thank you for your trust, guidance and for always reminding me that we are pushing the boundaries of science! James, thank you for always being available to help. Andreas, thank you for your guidance, support and trust in my work.

Many thanks to **EastBio** and **Genus/ABS** for funding my doctoral studies and to **Carolina Campos** from **Genus/ABS** for her patience, kindness and support. For putting up with me and our endless attempts to ship samples around the world.

To all of the Highlander Lab members, thank you for making it a great environment for learning and growing. Special thanks to **Dr. Ivan Pocrnic** for his assistance whenever I needed it and for the confidence he has placed in me; and **Dr. Daniel Tolhurst** for his excitement, encouragement, and the time he gave me. I am also grateful to **Dr. Jana Obsteter**, who has guided and supported me throughout my studies, with whom I had discussions and exchanges that were fundamental to the completion of this work. Thank you for all the effort and friendship.

Importantly, I thank my partner **Rebecka** for her love, for keeping me sane in the hardest times, for her patience and understanding.

List of abbreviations

1KB	1000 Bull Genomes Project
ARG	ancestral recombination graph
BLUP	best linear unbiased prediction
BOA	breed-of-origin of alleles
EBV	estimated breeding value
FA	factor analytic
FA-LMM	factor analytic linear mixed model
GNN	genealogical nearest neighbour
GRM	genomic relationship matrix
GS	genomic selection
G×E	genotype-by-environment
HOL	holstein
LMM	linear mixed model
mtDNA	mitochondrial DNA
SNP	single nucleotide polymorphism
TBV	true breeding value
WGS	whole-genome sequence

À minha filha, Inês. Que esse trabalho te inspire a
seguir firme a sua verdade

Contents

List of Figures

List of Tables

1 General introduction

Promoting resilient dairy production in the Global South[1] is imperative, considering the rapid population growth and the vulnerability of the region to climate change. While intensification is the ideal solution to increase production, this approach carries ethical and environmental concerns regarding the replacement of traditional practices, the displacement of populations, misuse of land and so on. Crossbreeding, however, is a sustainable and pragmatic alternative for improving tropical dairy systems. Securing the success of crossbreeding strategies, however, depends on strategic, science-driven implementations tailored for local conditions.

In 2023, the global dairy herd consisted of approximately 292 million dairy cows. Of those, near 83% were located in the Global South (FAOSTAT, 2025). Despite the great concentration of milking cows, only two countries from the region, India and Brazil, consistently rank among the top five global milk producers. Brazil, a continental country with diverse agro-ecological zones and heterogeneous production systems, has an estimated national herd of 17 million cows, expected to produce 25.4 million tons of milk in 2025 (Castro and Degreenia, 2024). Notably, only 7.4% of producers manage herds that yield more than 200 liters of milk per day, concentrating over half of the total national production. Meanwhile, the majority of Brazilian dairy farms produce less than 200 liters of milk per day, with a substantial proportion of farmers struggling to reach 10 l/day (Embrapa Gado de Leite, Accessed: 27 May 2025). India, the largest milk producer in the world, has a dairy herd projected to reach 62 million cows by

[1]Throughout this work I refer to Global South/North as a categorization of the world based on GDP (and power), rather than geography. The terminology became increasingly adopted since the Cold War to refer to post-colonial nations and regions marginalized within the new capitalist order (Mahler, 2017). In this context the Global South refers to countries in Latin America, Africa and parts of Asia, while Global North refers to North America, Europe, Australia, and New Zeeland. The division is relevant to this thesis as tropical dairy systems are predominantly encompassed within the Global South and, generally, the challenges discussed here are a product of the historical economic and social disadvantages shared by the countries within the category.

2025 (Bhogal and Brown, 2024). Similar to Brazil, Indian dairy is predominantly based on smallholder systems producing in very diverse circumstances. More than 70% of dairy operations are extensive or semi-extensive, and nearly 75 million farmers manage fewer than five milking cows (Kona et al., 2025). Together, this data reflects a highly fragmented and low-efficiency production base, a critical profile of dairy production in the Global South.

In both countries, growth in dairy production has been driven by governmental development policies focused on improving veterinary and extension services, enhancing feed and nutrition practices, and addressing sanitary status (Kona et al., 2025; Vilela et al., 2017). Central to these efforts has been the widespread adoption of reproductive technologies such as artificial insemination and embryo transfer, and especially the use of crossbreeding. In Brazil, crosses between the locally adapted Gir and the high-yielding exotic Holstein breed, formalised as the Girolando breed, dominate milk production. Similarly in India, local breeds from *Bos indicus* origin are crossed with high-yielding exotic breeds such as Holstein, Jersey, and Brown-Swiss (Kona et al., 2025). Despite significant progress and increased global representation, productivity in both countries remains low relative to the Global North, which reflects structural inefficiencies common to many production systems in the Global South.

In the Global North, long-standing herdbooks dating back to the 18^{th} century and extensive use of progeny testing enabled the development of intensively specialized, high-yielding dairy breeds (Weigel et al., 2017). Advances in genetics and breeding, combined with improved management and nutritional strategies, have transformed dairy production into an industrialised, high-input and high-output system. Importantly, the success of these systems has been contingent on controlled, resource-intensive environments to which trait expression of improved breeds depends on. An alarming drawback from the industrialisation of dairy is the reduced resilience particularly in terms of fertility, health, longevity, and environmental sensitivity (Brito et al., 2021). In contrast, dairy systems of the Global South lack the infrastructural and economic conditions necessary to support such high-input production. Constraints such as high feed costs, limited access to veterinary care, and underdeveloped infrastructure for data collection and genetic research and have prevented the development of breeding programs for locally adapted breeds (Michael et al., 2022).

Notably, productivity disparities persist even as herd sizes grow. For example, while the United States produced 15% of the global milk (99.5 billion liters) in 2024 with a declining herd of 9.4 million dairy cows (3% of the global dairy herd), India contributed

roughly 30% of the global milk output using around 22% of the global milking cows (Bhogal and Brown, 2024). Additionally, a trend of growing herd sizes is increasingly observed across East Asia and sub-Saharan Africa (FAO and GDP, 2018). These figures highlight the reliance of many Global South countries on expanding herd size, as a short cut for increasing their milk production.

Alongside the challenges of low productivity, dairy farming in the Global South faces growing risk due to climate change. Climate change is one of the defining challenges of the 21^{st} century (FAO and GDP, 2018), and dairy farming configures one of the most susceptible agricultural sectors to the impacts of increase in temperature and changes in precipitation patterns (Guzmán-Luna et al., 2022; Gauly and Ammer, 2020). This occurs directly, due to heat stress, or indirectly, through reduced availability of feed and water and rise in prices of inputs. One way or another, farmers are affected by the increasing instability in global climate (Godde et al., 2021). Tropical dairy farming has a higher risk and lower ability to mitigate impacts. This has to do with inherent characteristics of tropical dairy farming such as the use of extensive systems and a greater dependency on grazing, but also on the fact that tropical regions have less resources and infrastructure to deal with the impacts of climate change.

In the face of such challenges, the sustainability of agricultural systems requires a balanced approach that integrates environmental, economic, and social dimensions, ensuring long-term resilience and growth. The model of expansion-based growth used so far compromises the long-term resilience of dairy systems in the Global South. In contrast, breeding and genetics offer a better avenue. This thesis aims at addressing the issue of low productivity per cow in resource-constrained settings via the strategic use of genetic diversity.

In this chapter, I first review the historical context of cattle, focusing on the divergence between indicine and taurine cattle and the role the two subspecies play in shaping diary production systems of the Global North and of the Global South (Section 1.1). I then provide an overview of crossbreeding as an alternative for tropical dairy production systems, highlighting the challenges and opportunities associated with this strategy (Section 1.2). In the following section (Section 1.3) I discuss advances in genomic analysis and statistical methods that can benefit tropical dairy breeding. In the last section, Section 1.4, I outline the general and specific research objectives of the thesis.

1.1 The historical context of cattle

1.1.1 Evolution from the wild aurochs to domesticated cattle

The domestication of cattle was a pivotal process in human history, significantly contributing to the rise of settled agricultural societies (Ajmone-Marsan et al., 2010). This process began with the domestication of *Bos primigenius*, the wild aurochs. Aurochs likely originated in South Asia, dominating fertile temperate zones, and played a key role in maintaining ecosystems. Their grazing and browsing behaviours helped shape landscapes, promoting a mixture of grasslands and forests. Genetic diversity in these populations was shaped by environmental challenges and has been observed in Europe, North Africa, North Asia, South Asia, and Southwest Asia (Rossi et al., 2024). Today, two primary subspecies of domestic cattle are critical for global food production: *Bos taurus* and *Bos indicus*. The first is also referred to as taurine cattle, predominantly found in the Global North but with important representation in West-Africa. The later is commonly known as indicine cattle, a subspecies well adapted to the climates and conditions of the Global South.

The origins and domestication processes of these two groups have been the subject of extensive debate. While there is general agreement about the existence of two major domestication centres, one in the Near East and another in Southwest Asia (Pitt et al., 2018; Beja-Pereira et al., 2006), a third potential domestication centrer in Africa has been proposed (Brass, 2021; Dunne et al., 2012). The hypothesis of three domestication centres, however, remains controversial due to the lack of archaeological evidence and disagreements regarding the definition of domesticated cattle (Brass, 2021).

A central study by Loftus et al. (1994) gives support for the hypothesis of distinct origins of taurine and indicine cattle. The authors identified the two subspecies as carriers of different mitochondrial DNA (mtDNA) haplogroups. Taurine cattle was found to be carriers of the T haplogroup, while indicine cattle showed variations of the I haplogroup. Since mtDNA is maternally inherited, it is commonly used to trace evolutionary lineages. Using a molecular clock approach, the authors estimated the start of the two lineages at between 200,000 and 1 million years ago, when they diverged from a common ancestor. *Bos primigenius nomadicus*, the ancestors of modern indicine cattle populated tropical South Asia after diverging from *Bos primigenius*, the ancestors of taurine cattle, which initially populated Southwest Asia (Rossi et al., 2024).

Bradley et al. (1996) examined the T mitochondrial haplogroup in European and

African cattle. Their results suggest that European and African cattle may have originated from different subspecies of *Bos primigenius*, separated approximately 22,000 years ago. The authors also indicate African taurine cattle undergoing an independent domestication event. The hypothesis of a third domestication centre and the distinct origins of European and African taurine cattle is indicated by evidence that the T1 haplogroup, found in many North African taurine cattle, is rare in the Middle East and Anatolia, the centre of European taurine cattle domestication. Recent genomic research has further confirmed the existence of a preglacial auroch population in northern Africa, which likely carried the T1 mtDNA haplogroup (Rossi et al., 2024).

The domestication of European taurine cattle dates back approximately 10,000 years ago, around the time of the Neolithic Revolution in the Fertile Crescent. This region, which includes parts of modern-day Iraq, Syria, Lebanon, Israel, and Jordan, was a key site for the transition of humans from hunter-gatherer lifestyles to settled agricultural societies (Scanes, 2018). The process is thought to have been largely female-mediated, due to the ease of handling smaller animals. The presence of fertile river valleys and grasslands provided abundant resources for wild aurochs, making them an attractive target for early farmers that would capture mostly female individuals and their calves (Arbuckle and Kassebaum, 2021). Modern taurine cattle also show signs of early male-mediated introgression from North Asian, North African, and European ancestries (Rossi et al., 2024).

Bos indicus were domesticated independently roughly 2,000 years later (8,000 years ago), in the Indus Valley (present-day India and Pakistan) (Chen et al., 2010). Like the Near East, this region also experienced a Neolithic Revolution, where domesticated crops from the Near East influenced the domestication of local livestock, including cattle. Due to the warmer climate of the region, indicine cattle developed distinct morphological characteristics, marked by a dorsal hump. Nonetheless, archaeological evidence suggests that humped (modern indicine) and humpless cattle coexisted in the region around 5,000 years ago (Felius et al., 2014), indicating that the adaptation could have also been promoted by artificial selection.

The domestication in Africa most likely occurred in the northeastern part of the continent. Recently domesticated populations were further developed by later introgression with European taurine migrating from the Near East. The presence of the T1 mtDNA haplogroup in Franco-Iberian, Italian, and Greek cattle populations suggests that domesticated African taurine cattle were also introduced into Europe through the Mediterranean Sea (Beja-Pereira et al., 2006).

Regardless of the exact number of centres, cattle domestication had profound effects on human societies, providing a reliable source of food, labor, and power for agriculture and transportation. As humans migrated and spread across the globe, cattle were dispersed widely, resulting in the development of a variety of breeds with distinct phenotypes and genetic profiles (Xia et al., 2023). This domestication process not only transformed agriculture but also shaped the economies and cultures of societies around the world.

1.1.2 Migration patterns and environmental adaptation

The domestication of cattle, alongside other agricultural species, represented a new era of human-nature interaction. It catalyzed demographic expansion, habitat fragmentation, shifts in ecological relationships, and transformations in global biodiversity. The introduction of domesticated species into Europe represents the first documented case of intentional species translocation. This process transformed humans social structures, health and nutrition (McClure, 2013). Following their domestication, cattle became integral to human societies, providing essential resources such as meat, milk, and labor. The global spread of cattle mirrored patterns of human migration and expansion into new territories (Ajmone-Marsan et al., 2010).

The initial spread of cattle populations began in the Fertile Crescent, where early Neolithic farmers moved domesticated taurine cattle into Europe and Africa. Migration routes likely passed through Antolia, into the Balkans, and across the Mediterranean Sea, reaching southern Europe (Scanes, 2018; Pitt et al., 2018). The dissemination into the African continent occurred through a northern route, via Egypt, from where animals spread along the North African coast and into West Africa. Evidence also suggests that cattle from North Africa were moved into the Iberian Peninsula via the Mediterranean sea (Beja-Pereira et al., 2006). Between 5,000 and 4,000 years ago, cattle were moved northeastward, eventually adapting to the cold climates of Scandinavia and Siberia (Xia et al., 2023). During the spread of taurine cattle throughout Europe and Africa, admixture with wild aurochs likely occurred as domesticated and wild herds coexisted until the 17^{th} century, when the last known European auroch was extinct (Götherström et al., 2005). These movements also led to the introduction of taurine cattle into northern China (Rossi et al., 2024; Xia et al., 2023). As this migratory flow progressed, an early interaction between taurine and indicine cattle occurred, which had lasting impacts on the genomic and phenotypic development of indigenous Chinese cattle populations (Xia et al., 2023).

Indicine cattle also spread through human migration, occupying much of Southeast Asia and China (Chen et al., 2010). Their dispersal began around 4,500 years ago, with a migratory wave moving east and south across the Indian subcontinent (Felius et al., 2014). As indicine populations adapted to the tropical and subtropical climates, they faced high temperatures and disease pressures. These environmental challenges led to the development of characteristic morphological adaptations, including the formation of a dorsal hump, large ears, and excess skin, particularly around the neck, chest, and navel. In addition, genetic selection for resistance to pathogens and parasites contributed to their survival and success in these regions (Xia et al., 2023).

Pastoralist communities served an important role introducing indicine cattle into East Africa, primarily through the Horn of Africa, from where they spread southward and into central Africa (Hanotte et al., 2002). Modern African cattle populations have been significantly influenced by male-mediated indicine introgression, raising concerns about the genetic integrity of indigenous African taurine cattle (Kim et al., 2020; MacHugh et al., 1997).

These complex migratory pathways and regional adaptations laid the genetic foundation for the diversity observed in modern cattle populations. Understanding these origins is particularly relevant to understanding the challenges faced by indicine-taurine crossbreeding programmes in tropical environments.

1.1.3 The role of cattle in colonisation and settlement

The expansion of cattle into new territories was significantly shaped by European colonisation of the Americas and Australia. In the 16^{th} century, Spanish and Portuguese fleets first introduced taurine cattle to the Americas. These animals, originating from the Iberian Peninsula, Cape Verde, and the Canary Islands, contributed to the foundation of local livestock populations (Ginja et al., 2019; Felius et al., 2014). In the Americas, cattle encountered a range of environments, from semi-arid regions to tropical forests that imposed considerable selective pressure. Populations that successfully adapted gave rise to Creole cattle breeds, which are now widespread across the continent (Ginja et al., 2019; O'Neill et al., 2010). Their adaptability is largely attributed to African taurine ancestry, evidenced by the predominance of the T1 mtDNA haplogroup and Y chromosome lineages associated with African cattle (Ginja et al., 2019).

During the colonial period, cattle continued to advance into new ecological frontiers,

fulfilling critical roles in the colonial economy and profoundly impacting local ecosystems. The widespread cultural practice of allowing feral cattle to roam freely in large herds led to the degradation of native vegetation, transforming landscapes to support livestock production and agriculture (Sluyter, 2023).

From the 19^{th} century, as the American colonies gained independence and their economies diversified, urbanisation and technological advances drove the intensification of livestock production. This period saw the importation of specialised European breeds, such as Holstein, Hereford and Aberdeen Angus, to meet the growing demand of the new agricultural economy. It was also during this time that indicine cattle were introduced into the continent (Sluyter, 2023; Willham, 1982).

Initially imported to Brazil, the indicine breeds Gir, Nelore, and Guzerá were highly desirable due to ability to adapt to diverse climatic regions. Their utility in tropical and subtropical regions led to their rapid adoption to Brazilian cattle production. Ranchers in the United States soon followed, importing indicine cattle from India and from Brazil interested in their rusticity and adaptability (Hunnicutt, 1915). This marked the primordial use of indicine-taurine crossbreeding aimed at improving disease resistance and adaptation to southern climates (Hunnicutt, 1915; Willham, 1982).

By the mid 20^{th} century, cattle breeding in the Americas had become highly diversified. Breeds from various origins were used, along with the development of new breeds designed to meet local environmental challenges. Crossbreeding between indicine and taurine cattle had become a widespread strategy.

1.1.4 Examples of specialisation and typification

As cattle dispersed across diverse environments, local breeds began to emerge shaped by specific needs of communities in different regions of the world. This led to the first significant divergence in cattle populations, resulting in the development of major continental breeds, e.g., Charolais and Simmental, British breeds, e.g., Hereford and Red Poll, and northern breeds, e.g., Yakutian and Swedish mountain cattle (Xia et al., 2023; O'Neill et al., 2010). Further specialisation occurred with the differentiation of cattle into functional types, notably dairy breeds (e.g., Holstein and Jersey) and beef breeds (e.g., Angus and Limousin) (Felius et al., 2014).

Selective breeding in specific environments caused a reduction in the diversity in production systems, as a few breeds were preferred for their ability to deliver higher yields or better quality products. Among dairy breeds, the Holstein-Friesian breed became

the dominant due to its exceptional milk yield. Originating from black-and-white cattle of the Rhine Delta region, located along the present-day border between Germany and the Netherlands, the breed has been selectively bred for milk production for over two millennia (Lush et al., 1936). Holstein-Friesians were first introduced to North America by Dutch settlers in 1631. Their ability for high milk production quickly made them popular and by 1873, the first herdbook for Holstein-Friesians was published in the United States (Weigel et al., 2017).

The Holstein-Friesian Association of America was established in 1885, followed by the Canadian in 1891. In the 1960s, North American Holstein began reintroduction to European herds to boost productivity. By 1985, they constituted over 95% of Canadian dairy herd (Lush et al., 1936) and in 1994, the US Association was renamed Holstein Association USA, reflecting global reach of the breed. Today, American Holsteins average over 11,000 kg per lactation (Holstein Association USA, Inc., 2021), placing the United States as second-largest milk producer in the world (OECD/Food and Agriculture Organization of the United Nations, 2015).

For *Bos indicus* breeds, the Gir breed stands out for its resilience, longevity, and milk yield. Native to the Gir Forest in Gujarat, India, the breed thrives in tropical environments. They are valued for their ability to produce milk on pasture with minimal nutritional input and their remarkable heat and parasite tolerance (Santana Jr et al., 2014). These characteristics have led to their wide adoption in tropical regions and made them instrumental to crossbreeding programmes. Gir has been involved in the development of synthetic breeds, such Brahman in the United States and Girolando in Brazil (Maiorano et al., 2018).

Despite its significance, the Gir breed is currently threaten in India. Indiscriminate breeding practices, crossbreeding with exotic breeds, and limited access to artificial insemination infrastructure risk the genetic integrity of the breed. Since 2012, conservation efforts have been in place with progeny testing programmes aimed at increasing genetic gain (Board, 2017). Nonetheless, the recovery of the breed is constrained by long generation intervals and low selection intensity.

Conversely, the Gir breed flourished in South America, where it became the primary breed for milk production in the tropical centers of the continent. Between 1870 and 1962, 146 animals were introduced to Brazil were they adapted well to similar environmental conditions to those of the Kathiawar Peninsula (Gujarat state) in India (Reis Filho et al., 2010). Formal breeding efforts began in 1938 with the establish-

ment of a herdbook, and by 1985, the Brazilian Dairy Gir National Breeding Programme (PNMGL) was launched. Over 36 years, the programme has achieved consistent progress, with genetic gains of 1% per year for milk yield. By 2018, the programme had incorporated genomic data for over 14,000 animals, becoming the first indicine breeding programme to implement genomic selection. These advancements doubled milk production in participating herds, while also improving traits such as age at first calving and milk composition (Panetto et al., 2021).

The divergent domestication histories of indicine and taurine cattle have resulted in differences in morphology, temperament, productivity, and adaptability to environmental pressures. These differences made each subspecies vital to different production systems across variable global contexts. Crossbreeding efforts that combine the resilience of indicine breeds with the productivity of taurine breeds form the foundation of tropical dairy systems. However, the long evolutionary distance between the two subspecies and the variability of tropical environments poses challenges to the success of such strategies.

1.2 Crossbreeding in tropical dairy systems

1.2.1 The historical context of dairy breeding

Dairy farming has long been at the forefront of breeding innovation. In the 19^{th} century, breeding societies were established in Europe and North America aiming at standardising populations, often by selection on type traits. Pedigree and performance recording followed, and by the 1930s, methods for genetic evaluation of dairy sires based on the phenotype of their daughters were available (Weigel et al., 2017).

The Third Industrial Revolution promoted the intensification of agriculture in the Global North, prioritising efficiency and revenue. Artificial insemination, coupled with statistical methods for herdmate comparisons, accelerated genetic gain. Selection for milk yield and composition dominated breeding programmes and improvements in milking techniques promoted healthier dairy products (Clay et al., 2020). The rise of genetics and technology was accompanied by improvements in management and nutrition, culminating in high-yielding specialised breeds. Today, taurine breeds such as Holstein, Jersey, and Brown Swiss account for over 95% of milk production in intensified systems (Brito et al., 2021). The industrialisation of dairy saw continuous increases in milk production and the decline in number of cows and herds. Smallholders were progressively replaced by larger, more competitive technological farms, capable of

thriving in the new capitalist global market. But the industrialisation of dairy came at the cost of resilience. High-yielding cows, especially Holsteins, faced declining fertility, reduced longevity, and increased metabolic diseases (Brito et al., 2021). In response to these issues, breeding indexes were reformulated in the early 2000s to include functional traits, promoting more balanced breeding goals (Miglior et al., 2005).

In contrast, the progress of dairy production in the Global South was restricted by socioeconomic disadvantages, feed costs and limited infrastructure for data collection and genetic research. Throughout the years, tropical dairy systems remained largely extensive, with disconnected production chains and smallholder farms that struggle to expand or invest in improved genetic resources (Michael et al., 2022). Cattle in these systems faced challenging environmental conditions, needing to adapt to high temperatures, humidity, poor feed quality, and the high incidences of diseases and parasites. Government investments were scarce and, when present, insufficient to drive the necessary changes for the sector to become globally competitive. Breeding programmes for locally adapted *Bos indicus* breeds have only emerged in recent years (Santana Jr et al., 2014) and in contexts where there is strong government support and economic progress.

To address these challenges, crossbreeding environmentally adapted indicine breeds with intensive taurine cattle became a commonly adopted strategy. Crossbreeding aims to increase production and accelerate genetic progress by exploiting breed complementarity and heterosis, the improved mean performance of the crosses by combining production and adaptability of their parents (McDowell, 1985).

1.2.2 Genetics of crossbreeding

Historically, dairy breeding has relied primarily on additive genetic variance to promote population improvement. In the Global North, crossbreeding between high-yielding taurine breeds rarely produced offspring that outperformed the best parental breed for production traits. As a result, breeding efforts focused on within-breed additive genetic variance, and non-additive genetic variance was left not utilised (William and Pollak, 1985). With the decline in fertility and health traits, especially in American Holsteins, rotational crossbreeding among high-yielding taurine breeds became more common (Hazel et al., 2020). The extensive dissemination of genomic prediction and the use of sexed semen facilitated the implementation of crossbreeding schemes, making it more profitable.

Non-additive genetic variance, resulting from dominance and epistatic interactions, differs across traits of economic interest. Production traits typically have moderate heritability and are mainly governed by additive genetic effects, with expected low non-additive variance. In contrast, fitness traits have low heritability and are expected to exhibit higher non-additive genetic and environmental variance (William and Pollak, 1985; Falconer and Mackay, 1996). This reflects differences in trait architecture and selective pressure.

Fitness traits directly influence survivor and reproduction, involving many biological pathways. As a result, these traits are governed by a great number of loci creating opportunities for interaction effects, which increases dominance and epistatic variance. Moreover, fitness traits are under strong natural selection, which depletes additive genetic variance faster than dominance genetic variance. Consequently, dominance and epistatic effects contribute proportionally more than additive effects to the genetic architecture of fitness traits than production traits, resulting in lower heritability (Merilä and Sheldon, 1999).

In tropical dairy breeding, both production and fitness traits are targets in crossbreeding schemes, creating greater opportunities to exploit additive and non-additive genetic variance.

Crossbreeding capitalises on two genetic mechanisms: **breed complementarity**, the combination of desirable traits from different breeds, such as milk yield from the taurine breed and heat tolerance from the indicine breed, and **heterosis**, the average superiority of crossbred individuals over the average of the parental breeds. When both mechanisms are effectively exploited, the crossbreds express superiority across multiple traits, hence increase in profit. This is particularly relevant in tropical dairy breeding where the genetic distance between indicine and taurine breeds create opportunities for selection through increased genetic variance. Since breed complementarity is essentially averaging of parental genetic values, the following will describe only the genetic basis of heterosis.

The genetic basis of heterosis

Heterosis results from the interaction between alleles at the same locus (dominance) or across loci (epistasis). In a single locus model, heterosis can be expressed as a function of the squared allele frequency difference between parental breeds and the dominance effect. Heterosis due to dominance can then be expressed as the sum across all loci:

$$\overline{F1} - \overline{P} = \sum_i \Delta_i^2 \times d_i, \tag{1.1}$$

where $\overline{F1}$ is the mean performance of the F1 cross, $\overline{P}$ is the mean performance of the parental breeds, Δ_i is the difference in allele frequencies between the two parental breeds, and d_i is the dominance effect at the locus. This dominance effect represents the average deviation of the heterozygote genetic value from the average of the homozygote genetic effect (Falconer and Mackay, 1996). Maximum heterosis will occur when opposite alleles are fixed in each parental breed. Meanwhile, little to no heterosis will be observed when the breeds are closely related and no allelic differences exist (William and Pollak, 1985; Falconer and Mackay, 1996).

Limitations

While heterosis is at its maximum in the F1 cross, the subsequent generations face a progressive decline in heterotic effects. This decline stems from inbreeding, which reestablishes homozygosity reducing dominance deviation. It is important to emphasise that there can also be recombination loss in crossbreds compared to purebreds. During meiosis, recombination events might disrupt linkage between loci, compromising epistatic interactions (Dickerson, 1973).

Without strategic management, such as structured rotational crossbreeding schemes or synthetic breed formation, it is impossible to sustain heterosis. Rotational crossbreeding uses planned mating schemes in which two or more parental breeds are rotated across generations, varying the alleles introduced into the population, and thus recovering the level of heterosis. In synthetic breed formations, the population is closed for introgression after achieving certain breed composition, leading to a new, stable population. But even deciding on the crossbreeding scheme is challenging and context-dependent. Syrstad (1989) argues that, because of recombination loss, the use of rotational crossing schemes would be more beneficial than the development of synthetic breeds as a crossbreeding strategy. On the other hand, William and Pollak (1985) argues that, when environmental conditions do not favour dairy production, the development of tailored synthetic breeds may be the ideal strategy.

In rare cases, crossbreeding between highly divergent populations may result in hybrid incompatibility. While less common in intraspecies crosses, mitonuclear incompatibility has been observed in taurine-indicine crosses. For example, Ward et al. found

that African crossbreds carrying taurine mitochondrial DNA and indicine nuclear genes involved in mitochondial function exhibited reduced fertility and mitochondrial malfunction.

Determinants of crossbreeding outcomes

Thus, the magnitude and persistence of crossbreeding benefits depend on:

1. **Genetic distance** between the purebreds used in the system, which determines differences in allele frequency (Δ). The more distant the breeds, greater will be the heterosis expressed in the crossbreds, but not distant enough to cause incompatibility between allelic combinations.

2. **Genetic architecture** of the traits, meaning the degree of dominance and epistasis associated with the traits of interest. If dominance deviates in different directions at distinct loci, their effects could cancel each other resulting in no heterosis, despite the dominance effect at individual loci;

3. **Individual crossbred genotype**, which determines the extent to which favourable allelic combinations will occur and, thus the amount of heterosis expressed;

4. **Production environment**, which can interact with genotype to alter the magnitude and direction of heterosis, resulting in genotype-by-environment interaction.

1.2.3 Genotype-by-environment interaction and trait stability

Among the factors influencing the success of crossbreeding is genotype-by-environment (G×E) interaction. In tropical dairy systems, production environments can vary in climate, feed availability, health and management practices. This heterogeneity makes G×E interaction a critical factor shaping the performance of crossbreds. G×E interaction reflects the differential response of genotypes to changes in their environment, causing variation in trait expression and trait stability (Kang, 1997). It is important to recognise that the same trait measured in different environments may reflect distinct but correlated genetic architectures (Falconer, 1952).

G×E interaction manifests broadly in two forms (Fig. 1.1). In non-crossover interaction there is change in the scale of response of genotypes between environments, but their rank remains the same. In contrast, crossover interaction involves re-ranking of genotypes across environments (Beker, 1988; Gail and Simon, 1985). In reality, both forms of G×E interaction are entangled, posing challenges to breeding programmes,

since animals selected as superior in one environment may underperform elsewhere. Ignoring G×E interaction can therefore lead to wrong selection decisions and delayed genetic progress.

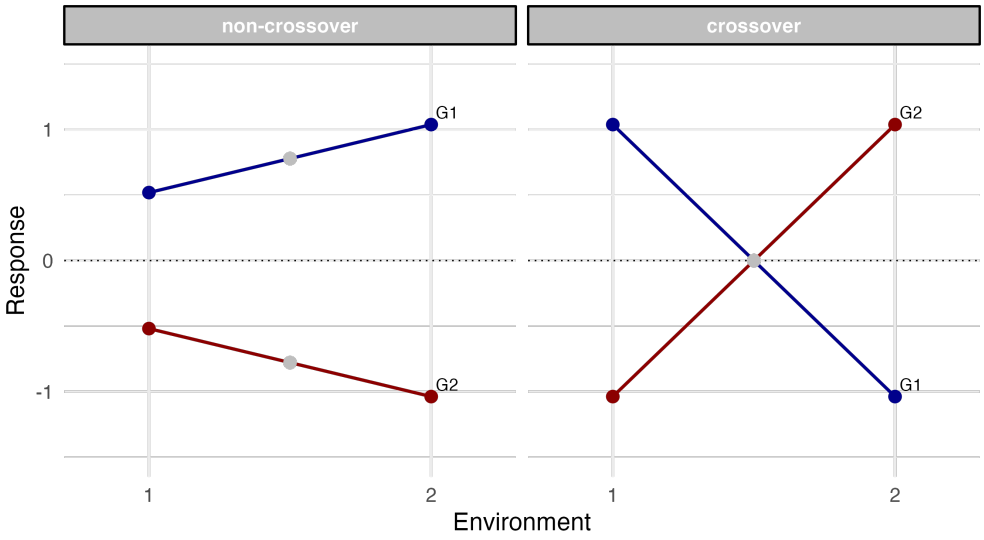

Figure 1.1: Illustration of non-crossover and crossover G×E interaction. Response of individuals G1 and G2 to change from environment E1 to E2 for a single continuous trait. Grey dots represent the average performance of the individual between environments.

Dairy breeding programmes in the Global North have historically minimised environmental variability through optimised management practices (Tiezzi and Maltecca, 2022). By reducing the influence of environmental factors such as temperature, desease load and feed quality, on performance, these systems have reduced the need to account for G×E interaction and favoured animals adapted to uniform, controlled conditions. However, this approach does not translate to the heterogeneous production systems of the Global South, where environmental variation is intrinsic to dairy farming.

Evidence from multiple species highlight the environment-dependent nature of heterosis. In plant breeding, dominance is well recognised a component of the G×E interaction, with purebreds and hybrids responding differently to environmental factors (Hunt et al., 2020). In aquaculture, Wang et al. (2024a) reported F1 catfish crosses, whose fitness advantages reversed under different environmental conditions. Similarly, in tropical dairy systems, G×E interaction is expected to reduce the superiority of the crossbred over local breeds as environmental stress intensifies (Bunning et al., 2018; Leroy et al., 2016).

These challenges highlight the importance of trait stability, the capacity of genotypes to

maintain performance and health across environments (Sánchez-Molano et al., 2023). Animal breeders usually refer to this concept as resilience, but an important definition tied to the impacts of G×E interaction comes from plant breeding. There, stability is classified as either static, where the average performance of a population remains constant regardless of environmental changes, or dynamic, where performance varies in a predictable manner with environmental changes (Becker and Leon, 1988). While static stability is often emphasised, it leaves no variation for genetic improvement since it involves the population mean. Dynamic stability, on the other hand, is valuable, desirable, and exploitable for tropical dairy crossbreeding, as predictable changes in performance are easier to manage than unstable responses while still promoting genetic improvement. Selecting individuals with adaptative responses and stable performance can improve mating decisions and contribute to the long-term success of crossbreeding programmes in tropical regions.

Combined with the frequent lack of well-defined crossbreeding policies, general recommendations and performance monitoring systems in many tropical countries, these factors make response to crossbreeding highly variable. This variability highlights the urgent need for breeding strategies tailored to local conditions; strategies that integrate the genetic principles of breed complementarity and heterosis, G×E interaction, and trait stability, with the socioeconomic and environmental realities of tropical dairy farming.

1.2.4 Empirical examples of crossbreeding implementation

Crossbreeding locally-adapted indicine with exotic high-yielding breeds is used as a strategy to overcome the intrinsic limitations of dairy production in the Global South. It is widely used to address the late development of tropical dairy. However, the implementation of crossbreeding strategies is not straightforward, especially when production systems and climates are highly variable. The absence of nation- or even region-wide polices and structured strategies to guide crossbreeding efforts delay genetic progress and the success of such practices. To address this gap the Brazilian National Dairy Cattle Research Centre and the Food and Agricultural Organization (FAO) of the United Nations, conducted a comprehensive trial comparing crossbreeding strategies. It was the first of its kind aimed at defining breeding recommendations for crossbreeding in tropical and sub-tropical regions (Madalena et al., 1990). The trial compared five strategies:

1. A replacement system of F1 females: only first-generation indicine × Holstein

are used as milking cows, offspring are sold or culled;

2. Upgrading to Holstein (exotic genetics): indicine and crossbred are crossed to Holstein sires every breeding cycle increasing the exotic genetic proportion of the herd to 100% Holstein;

3. The development of a new synthetic breed: crossbreeding is performed until reaching expected genetic composition of 5/8 Holstein and 3/8 indicine. The synthetic breed is developed from their interse crossing;

4. Rotational crossing: alternate use of Holstein and indicine sires every breeding cycle; and

5. Modified rotational crossing: use of Holstein sires for two generations followed by indicine sire for one generation.

Results from the trial indicated the F1 replacement system as the most profitable strategy across production systems. In contrast, upgrading to Holstein (exotic genetics) was detrimental in low management conditions. The synthetic breed strategy also performed poorly, requiring a cohesive breeding system in place to maintain the benefits of heterosis through strong selection practices. The findings aligned well with theoretical expectations, emphasizing that sustained benefits from crossbreeding depend on matching breeding methods with adequate production environment.

Despite the results of the trial, Brazil has focused crossbreeding efforts on the synthetic Girolando breed, a product of the systematic Gir × Holstein cross (Canaza-Cayo et al., 2016). In 1996, the Girolando breed was formalised within the Ministry of Agriculture and the Breeders Association was established, aimed at maintaining the Girolando herdbook. Officially, the breed is the product of the inter se crossing of 5/8 Holstein individuals (Silva et al., 2025), however, in practice a larger range of breed compositions are recognised. While the breed has achieved important success and enjoys policy support, the production system remains diverse, making other breed compositions resulting from various crossbreeding strategies attractive. This reflects the challenges of coordinating breeding strategies even when government policies are in place.

More recent studies mirrored the lessons from Madalena et al. (1990). For smallholder farmers in tropical regions, the most cost-effective strategy is to improve local breeds through crossbreeding with exotic genetics. However, for it to work it is necessary to simultaneously invest in better management practices and infrastructure (Roschinsky et al., 2015; Marshall et al., 2019; Hunde et al., 2024).

1.2.5 Challenges for a sustainable tropical dairy

These results indicate that adopting crossbreeding in tropical dairy systems implies the intensification of production at the farm level. This is the case due to the environmental requirements of exotic breeds, such as higher energy demands from feed, less resilience to temperature fluctuations and diseases. It is important to note that the intensification increases pressure on resources and labor, especially on women. Women make up two thirds of livestock keepers in low-income settings and as the systems intensify, they usually loose access to decision-making power due to gendered dynamics (Food and Agriculture Organization of the United Nations (FAO), 2013). Additionally, while intensification may lead to a slight increase in output and a reduction in herd size, these benefits are accompanied by greater susceptibility to disease and higher input costs.

In practice, there is a prevailing assumption that crossbreeding will always outperform native breeds. This misconception rises from a misunderstanding the genetic and environmental factors affecting crossbred performance. Farmers may expect a simple linear relationship between breed composition and performance. Such assumptions can results in inadequate management of F1 calves, the adoption of suboptimal mating strategies and profound economic and welfare consequences. It erodes the long-term stability of crossbreeding systems (McDowell, 1985).

Also problematic is the idea that one top performing bull suits all production systems. It overlooks the wide variation in production and ecological systems across tropical countries. Each of these systems presents unique challenges, requiring breeding strategies designed specifically for local conditions.

1.3 Novel genomic analysis and statistical methods

Improving crossbreeding outcomes in tropical dairy systems is a multi-faceted issue. Breeding and genetics can contribute to the development of crossbreeding strategies better-suited to local conditions. Especially, there are three fundamental aspects hindering the success of tropical crossbreeding that can be addressed with novel genomic analysis and statistical methods: (i) the genetic distance between tropical-adapted local breeds and high-yielding exotic breeds, (ii) the lack of statistical methods tailored to dealing with such genetic distance, and (iii) the G×E interaction that underlies the instability in crossbred performance across environments.

Addressing these challenges requires tools that can explicitly account for the evolutionary history of indicine and taurine breeds, account for ancestry-specific genetic effects that arise from their divergence, and capture performance variation across environments.

Recent advancements in Ancestral Recombination Graph (ARG) inference for large genomic datasets provide new opportunities for exploring the genetic distance separating indicine and taurine cattle and leveraging ancestry-specific mutations and their effects to the benefit of tropical dairy crossbreeding. At the same time, statistical approaches such as multiplicative models, largely adopted in plant breeding but generally overloked by animal breeders, offer a powerful way to model G×E facilitating the interpretation of environmental similarities and the adaptation pattern of individuals.

The following sections introduce these two methodologies. ARGs provide the biological and evolutionary tools for understanding breed divergence, while multiplicative models offers the statistical framework to translate environmental disturbance into practical breeding decisions.

1.3.1 Ancestral Recombination Graphs (ARGs)

The ARG, introduced by Griffiths and Marjoram (1997), models the ancestry of a sample of haplotypes in the presence of recombination. It expands on the coalescent theory (**?**), which describes the stochastic process that traces the genealogical history of a sample of haplotypes backwards in time. In the most basic coalescent model, two sample haplotypes coalesce into one ancestral haplotype, which then recursively coalesce until reaching the most recent common ancestor (MRCA). Thinking and modelling with such DNA trees has been central to biology since the work of Darwin (Darwin, 1859). While coalescent theory provides a mathematical treatment of DNA trees, it ignores a major evolutionary process, recombination.

Recombination is one of the most important evolutionary forces shaping the genome and genetic diversity. It is responsible for breaking up linkage disequilibrium, preventing the accumulation of deleterious mutations, freeing alleles to evolve independently, and increasing genetic diversity accumulated through selection and drift (Felsenstein, 1974; Otto and Payseur, 2019).

Until recently, full understanding of patterns of DNA transmission from ancestors to descendants was limited to the availability of pedigree and genomic data. While pedigree can provide expected estimates of ancestry, it is not possible to determine the

exact inheritance pattern for any particular region of the genome. Genomic data can provide a more precise account of estimates of ancestry at the genomic level, but full understanding of the origins and patterns of transmission of specific genes is still hard to obtain. Such full description of the transmission of DNA between generations is given by ARGs (Arenas, 2013).

Consider Figure 1.2, which represents the demographic history of cattle, with taurine and indicine populations diverging from a common ancestor and evolving independently towards recent time. Suppose we sample five hypothetical individuals from the recent taurine population, and three from the recent indicine population, assuming each genome is represented as a continuous 500 base-pair DNA sequence. The ancestral relationship between these individuals can be described using an ARG.

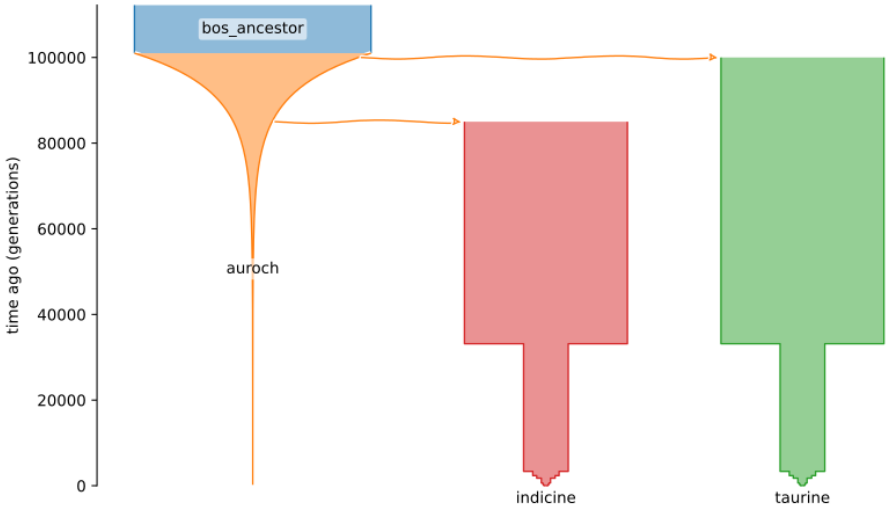

Figure 1.2: Demographic history of cattle. The diagram illustrates the divergence of taurine and indicine cattle from a common ancestor, followed by independent evolution into distinct modern populations.

At the base of the ARG (Figure 1.3), each sampled individual is represented as a pair of nodes, one for each haplotype (cattle being a diploid organism). The branches of the ARG connect the sample nodes to their ancestral nodes, with edges where coalescent or recombination events occurred. All nodes in the ARG converge to a single MRCA at the root of the graph for this small region. While the ARG provides the genetic history for a set of individuals along the entire length of the genome, it can be decomposed and simplified into local trees.

Local trees are defined by the recombination breakpoints along the genome. In Figure

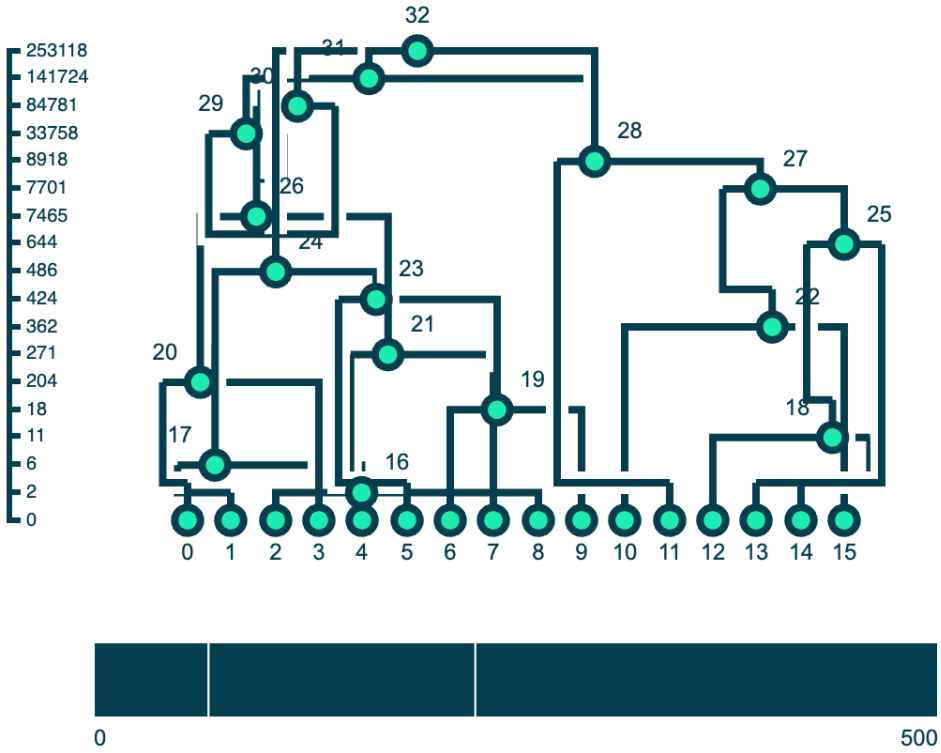

Figure 1.3: Hypothetical ancestral recombination graph (ARG). The graph describes the genetic ancestry of eight sampled individuals, five from the taurine population (nodes 0-9) and three from the indicine population (nodes 10-15), along a 500 bp genomic region, tracing recombination and coalescent events back to the most recent common ancestor (MRCA).

1.4, the hypothetical ARG is composed of three local trees, the first between positions 0-68, the second from 68-226, and the last from 226-500. Recombination events can cause the topology of the trees to change along the genome. In this example, the tree topology reveals that nodes 0, 3, 5, and 7 retain the same ancestry across all recombination breakpoints, while nodes 1, 2, and 8 experience a shift in ancestry between positions 68 and 226.

This example demonstrates that, at the genome level, ancestry varies due to recombination, which breaks up and modifies the inheritance of pieces of DNA and thus the genetic history of a sample. Arenas (2013) highlighted the difficulties involved in building ARGs, the importance of inferring these graphs, and the need for a standard inference format to facilitate communication among different software tools. However, the application of ARGs was confined to theoretical biology. The inference process was both complex and computationally expensive, compounded by the intricate representation of ARGs, not as phylogenetic networks, but embedding the evolutionary history and ancestral material for each node. Less than a decade later, new methodological

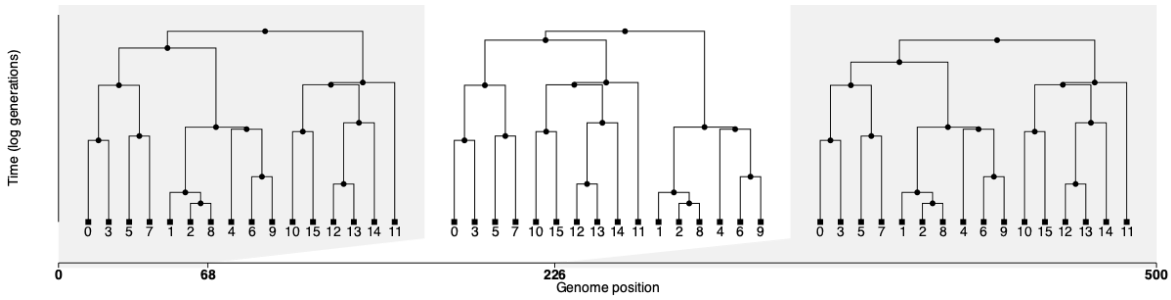

Figure 1.4: Decomposition of the hypothetical ARG into local trees. Recombination breakpoints segment the genome into intervals where a new genealogical tree is formed, representing the local ancestry of the sampled individuals.

approaches have led to the development of efficient algorithms for ARG inference from genomic data, as reviewed by Lewanski et al. (2024).

Building on the advancements of ARG inference, Wong et al. (2024) proposed a distinction between the ARG as a stochastic process and the ARG as a practical and computationally efficient data structure. The authors then introduced a new formalism, the genome ARG (gARG), where nodes represent haploid genomes, and branches represent genetic transmission between the genome across generations, as shown in Fig. 1.3. Recombination events are succinctly described through the annotation of genome coordinates on the branches. Mutation are also accounted and annotated to the branches. The gARG formalism thus generalises the classical ARG from Griffiths and Marjoram (1997), serving as a standard format to facilitate communication across softwares.

Applications to breeding and genetics

ARGs have recently become a powerful tool for population genetics, enabling the reconstruction of complex evolutionary histories and the inference of demographic events. They have been used to study population structure and migration patterns, improve the detection of selection signatures and the study of the effects of selection on genetic variation (Stern et al., 2019; Speidel et al., 2019; Schraiber et al., 2024; Nielsen et al., 2025).

As the field of ARG inference quickly progress, research grows into its application to quantitative genetics. The ARG-based genomic relationship matrix (GRM) is expected to better capture the structure of a population than the pedigree or the genomic GRMs. Lehmann et al. (2025) developed a tree sequence-based algorithm to compute relatedness without explicitly building a dense GRM, while Lee et al. (2025); Zhu et al.

(2024); Zhang et al. (2023) integrated the ARG topology into mixed models to study and model the evolution of complex traits. Central to the method of Lehmann et al. (2025) is the understanding that branches in the local trees are better descriptors of relatedness, and the ARG itself is a sparse representation of the genotype matrix.

For tropical dairy crossbreeding, ARGs could enable explicitly distinguishing between indicine and taurine ancestry, their unique mutations, and their distinct effects on performance. This information can be integrated with linear mixed models to incorporate ancestry-specific effects, account for genetic distance between breeds and improve prediction accuracy.

1.3.2 Linear mixed models

In animal breeding, linear mixed models (LMMs) provide flexibility to model a continuous response in terms of fixed effects shared by individuals in the population, and random sources of variation (Henderson, 1984).

For a trait measured once per individual, the model can be expressed as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\epsilon}, \tag{1.2}$$

where $\mathbf{y}$ is a vector of phenotypic observations, $\mathbf{b}$ is a vector of unknown fixed effects with design matrix $\mathbf{X}$, $\mathbf{u}$ is a vector of unknown random effects with design matrix $\mathbf{Z}$, and $\boldsymbol{\epsilon}$ is a vector of random residuals.

We assume the random effects and residuals are distributed as:

$$\mathbf{u} \sim \mathcal{N}(0, \mathbf{G}), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{R}), \tag{1.3}$$

where $\mathbf{u}$ and $\boldsymbol{\epsilon}$ are independent with covariance matrices $\mathbf{G}$ and $\mathbf{R}$, assumed to be positive (semi)- definite. The covariance of $\mathbf{y}$ is therefore:

$$Cov(\mathbf{y}) = \mathbf{ZGZ}^\top + \mathbf{R}. \tag{1.4}$$

When $\mathbf{u}$ represents genetic effects, the covariance matrix $\mathbf{G}$ can be expressed as:

$$\mathbf{G} = \sigma_g^2 \mathbf{A}, \tag{1.5}$$

where σ_g^2 is the additive genetic variance and $\mathbf{A}$ is the additive relationship matrix. Relationships can be determined from pedigree (Wright, 1921), from genomic data

(VanRaden, 2008), or a combination of both information (Misztal et al., 2009). Note that other (non-additive) relationships can be included where appropriate (Vitezica et al., 2013).

1.3.3 Modelling genotype-by-environment interaction

Incorporating G×E effects into LMMs can be a major challenge, as the number of effects to be estimated grows with the number of environments. Many approaches have been proposed to model G×E interaction, but the challenge is to find a solution able to impose structured covariance matrices that parsimoniously describe heterogeneity in variance and covariance across environments, while retaining interpretability (Meyer, 2009b).

Traditional models for G×E interaction

In the simplest setting, G×E interaction is implicitly integrated into the linear model through the residual as:

$$y_{ij} = \mu + g_i + e_j + \epsilon_{ij}, \tag{1.6}$$

where y_{ij} is the phenotypic record for individual i in environment j, μ is the overall population mean, g_i is the main genetic effect of individual i, e_j is the main effect for environment j, and ϵ is the random residual. This model assumes independence between genetic and environmental main effects, and constant genetic variance across environments. G×E interaction effects, ge_{ij}, are captured within the residuals, ϵ_{ij}.

A more general form, explicitly dissociating between the interaction effects and the residual is given by:

$$y_{ik} = \mu + g_i + e_j + ge_{ij} + \epsilon_{ij}. \tag{1.7}$$

The limitation with this formulation is that, for the separation of ge_{ij} and ϵ_{ij} to be possible, appropriate population structure and/or a large number of records is required, given the infeasibility of replicating animals across environments. Thus, accounting for the interaction leads to substantial increase in the number of effects to be estimated. This means that for n individuals in p environments, there are $n \times (p-1)$ extra effects to be estimated. The covariance structure is typically given by:

$$\mathbf{G} = \mathbf{G}_e \otimes \mathbf{A} \tag{1.8}$$

where $\mathbf{G}_e$ is the $p \times p$ genetic covariance matrix between environments, $\otimes$ denotes the Kronecker product and $\mathbf{A}$ is the $n \times n$ additive relationship matrix between individuals (and ancestors) from Eq. 1.5.

Multi-trait and random regression models

Multi-trait models treat the performance in each environments as different, correlated traits (Falconer, 1952). They are especially useful when the number of environments is small, as they allow for the estimation of genetic correlations between environments and direct quantification of non-crossover and crossover G×E interaction. These models are based on an unstructured genetic covariance matrix that is fully parametrised, with the genetic variance of each environment and the genetic covariance between each pair of environments estimated from the data. The unstructured genetic covariance matrix between environments is expessed as:

$$\mathbf{G}_e = \begin{bmatrix} \sigma^2_{g_1} & \sigma_{g_{12}} & \cdots & \sigma_{g_{1j'}} \\ \sigma_{g_{21}} & \sigma^2_{g_2} & \cdots & \sigma_{g_{2j'}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{g_{j1}} & \sigma_{g_{j2}} & \cdots & \sigma^2_{g'_j} \end{bmatrix}, \tag{1.9}$$

where $\sigma^2_{g_j}$ is the genetic variance for environment j and $\sigma_{g_{jj'}}$ is the genetic covariance between environments j and j'.

It is important to note that multi-trait models rely on the genetic connectivity between environments, depending almost exclusively on additive relationships. Additionally, the number of unique variance and covariance parameters to be estimated, $k(k+1)/2$, becomes computationally prohibitive as the number of environment, k, becomes large.

Random regression models instead treat the environment as a continuous covariate and regress individual performance on this scale (Schaeffer, 2004; Manuck, 2009). The model takes the form:

$$y_{ij} = \mu + g_i + (1 + \beta_i)e_j + \epsilon_{ij}, \tag{1.10}$$

where μ is the population mean, g_i in the main genetic effect of individual i, β_i is the regression coefficient for individual i and e_j is the environmental covariate for

environment j. The covariance structure is then expressed as:

$$\mathbf{G}_e = \begin{bmatrix}\mathbf{1} & \mathbf{e}\end{bmatrix} \begin{bmatrix}\sigma_g^2 & \sigma_{12} \\ \sigma_{21} & \sigma_\beta^2\end{bmatrix} \begin{bmatrix}\mathbf{1} & \mathbf{e}\end{bmatrix}^\top \tag{1.11}$$

where $\mathbf{1}$ is a vector of ones, $\mathbf{e}$ is a vector of environmental covariates, σ_g^2 is the genetic variance of the main effects (intercepts), σ_β^2 is the genetic variance of the slopes, and $\sigma_{12} = \sigma_{21}$ is their covariance. While this structure is more parsimonious, it relies on the use of a single known, measured environmental variable, which can oversimplify the complexity of G×E problems if there are other important unknown drivers of environmental variation (Waters et al., 2023; Tolhurst, 2024).

Recent extension of the random regression partitions the slopes into non-crossover and crossover type components (Waters et al., 2023; Tolhurst, 2024). This approach improved the understanding of underlying mechanisms and the interpretability of G×E patterns. However, the limitation of depending on predefined covariates remains.

Multiplicative models

Multiplicative models, widely used in plant breeding, capture both non-crossover and crossover G×E interaction, using a reduced number of multiplicative terms. This class of methods decompose G×E interaction into products of genetic effects and latent environmental covariates (Meyer, 2009b). These terms are estimated directly from the data, without the use of constraints, an advantage over random regression models.

The main genetic effects (g_i) and interaction effects (ge_{ij}) are combined, and reformulated as:

$$g_i + ge_{ij} = \sum_{r=1}^{p} \lambda_{rj} f_{ri} \tag{1.12}$$

where λ_{rj} is the covariate of environment j for term r and f_{ri} is the corresponding slope of animal i. The form of Eq. 1.12 can be obtained by applying a singular value decomposition to the G×E table constructed from the genetic effects (or breeding values) of all individuals in each environment.

The model generates a set of environmental covariates responsible for changes in genetic effects between environments; they focus on identifying commonality between environments responsible for the G×E patterns. Notably, the use of unobserved environmental covariates can represent a complication as they may not be easily interpreted.

The genetic covariance matrix between environments is now expressed as:

$$\mathbf{G}_e = \mathbf{\Lambda D \Lambda}^\top \tag{1.13}$$

where $\mathbf{\Lambda}$ is a matrix of latent environmental covariates and $\mathbf{D}$ is a diagonal matrix comprising the variances of each multiplicative term.

Multiplicative models therefore provide a robust way to model G×E interaction, while preserving the covariance structure of the data. The association of such models with principal component rotations provide tools for better understanding the latent environmental covariates and how they impact trait expression across environments (Smith and Cullis, 2018). For tropical dairy systems, where environmental variability is high, these resources can be of great value in identifying stable genetic effects and exploiting adaptive responses in crossbred populations.

1.4 Thesis objectives

This thesis is structured around three related research chapters, each addressing a key challenge in leveraging genetic diversity to promote tropical dairy breeding. Combined, these chapters address the thesis main objective, to demonstrate how underutilized genetic diversity can be leveraged to improve the performance of crossbred cattle across diverse tropical environments.

Despite major advances in genomic technologies and breeding methodologies, important challenges remain in the effective application of these innovations to diverse cattle populations, particularly in the Global South. In tropical dairy systems, crossbreeding locally adapted with high-performing exotic breeds holds promise to improve productivity and environmental robustness. However, the success of these systems remain hindered in inconsistent crossbred performance, under-characterized genetic diversity, and the lack of tools tailored to admixed populations and diverse environments.

This thesis aims at addressing these gaps through the following work, each covering one specific objective:

1. Chapter 2: **Explore the genetic diversity within and between cattle populations**. This chapter presents the first large-scale application of Ancestral Recombination Graphs (ARGs) in cattle, using tree-sequence reconstruction based on the 1000 Bull Genome Project. The goal is to explore the genetic diversity across global cattle populations. We demonstrate how ARGs can provide scalable,

fine-resolution insight into population structure, local ancestry, and evolutionary history, offering a more informative and efficient alternative to conventional methods.

2. Chapter 3: **Develop a new statistical methodology that models ancestry-specific effects in diverse populations.** Genomic selection (GS) has significantly advanced livestock breeding, especially in dairy cattle, by improving selection accuracy and reducing generation intervals; however, it faces challenges such as computational complexity, reduced genetic variance over time, and limited prediction accuracy across diverse or admixed populations. Current models struggle to capture ancestry-specific mutation effects and sequence context dependencies, and while whole-genome sequencing (WGS) offers potential improvements, its benefits have been inconsistent due to modeling limitations. This chapter proposes a statistical model using ancestral recombination graphs (ARGs) to estimate ancestry-specific mutation effects within the sequence context, offering a scalable, biologically informed alternative to traditional genomic models.

3. Chapter 4: **Provide a framework for analysing genotype-by-environment interactions and uncover underutilized genetic diversity to improve crossbreeding strategies in tropical dairy systems.** Crossbreeding is a key strategy in tropical dairy systems, but its outcomes are highly variable due to genotype-by-environment (G×E) interaction and genetic divergence between the breeds. These factors reduce trait stability and complicate selection, particularly as G×E interaction can cause performance rankings to shift across environments. In this chapter, we use stochastic simulations and multiplicative models to provide a framework for identifying adapted and resilient genotypes, exploiting the genetic diversity that arises from G×E interaction to inform more effective breeding strategies for variable tropical environments.

In Chapter 5, I summarize the main findings and discuss their implications for tropical dairy systems. I also outline future research directions, emphasizing the need for further exploration of genetic diversity in under-characterized cattle populations to support the sustainable development of dairy breeding programmes in the Global South.

2 Global cattle genealogy inferred from ancestral recombination graphs

This chapter presents the manuscript *Global Cattle Genealogy Inferred from Ancestral Recombination Graphs*, which investigates global patterns of genetic diversity and population structure in cattle and demonstrates the application of ARGs to livestock genomics.

As the scale of genomic data in livestock continues to expand, conventional matrix-based format increasingly limits data storage and analysis. This manuscript addresses this limitation and provide additional information by applying Ancestral Recombination Graph (ARGs) inference. This is a novel, powerful approach for representing the shared ancestry among individuals in a population that captures the full history of coalescence, mutations and recombination across genomes. Specifically, the study focuses on reconstructing tree-sequence-based ARGs from the 1000 Bull Genomes Project public dataset to explore fine-scale population structure, local ancestry and demographic history across 109 cattle populations worldwide, focusing on indicine-taurine divergence. The results demonstrate how ARGs can reveal detailed evolutionary and population-level insights, even in admixed or under characterized populations.

This work is a collaboration between Gabriela Mafra Fortuna[1], Jana Obšteter[1,2], Hannes Becher[1], and Gregor Gorjanc[1].

[1] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Edinburgh EH25 9RG, Midlothian, United Kingdom; [2] Agricultural Institute of Slovenia, Hacquetova ulica 17, 1000 Ljubljana, Slovenia.

Manuscript Status: In preparation.

Core ideas

- ARGs inference, via the tree-sequence data format, offer a powerful and scalable way to reconstruct cattle genome evolution by encoding past coalescence, recombination and mutation events that generated observed genomic sequences, allowing for efficient storage and biologically-informed evolutionary inference.
- Genomic differences between *Bos taurus* and *Bos indicus* are a result of deep evolutionary divergence and domestication history, which affect the expression of traits of interest and crossbreeding outcomes in tropical dairy systems.
- Admixed populations show high genetic and phenotypic diversity, complicating genetic improvement and requiring advanced methods that account for differences in genetic architecture.

Abstract

Obtaining whole-genome sequence genotypes has never been more accessible. This affordability is allowing the discovery of a great number of new genetic variants across species. At the same time, storing and analysing the data is becoming challenging. Ancestral Recombination Graphs (ARG), inferred in the form of tree-sequences surge as a powerful tool to efficiently store and analyse genomic data. In the past decade, ARGs have transitioned from a theoretical concept, to practical, computationally efficient methods. Regarded as the ideal representation of DNA variation in populations, an ARG captures the entire history of inheritance and genetic variation. Although the method is actively used in human population genomics, its application to agricultural species, especially livestock, is still limited. Here we infer tree-sequences for a cattle whole-genome sequence dataset comprising 1,832 individuals across 109 populations from diverse and complex ancestries. The raw data is 874 GB in size and after filtering and removing non-biallelic SNP and structural variants 8.4 GB. We obtained 26.9 million time-resolved tree-sequences across 28.4 million variant sites, composed of 74 million nodes, 1 billion branches, and 28 million mutations, all stored in 9.9 GB. We show how the methodology can contribute to the understanding of the genetic diversity and population structure of cattle. While our results illustrate the potential of ARG to livestock genomics, they also highlight the challenges of applying the method to agricultural species. Thorough data processing and inference optimisation is required to obtain accurate tree-sequences.

2.1 Introduction

Over the past two decades, genotyping, whole-genome sequencing and resequencing have become progressively more accessible (Wetterstrand, 2023). Across fields, these technologies have uncovered billions of DNA variants. In human genetics, biobanks have integrated genome-wide data from millions of individuals alongside detailed phenotypic and medical records. These resources are fuelling the discoveries of causative loci and transforming healthcare through precision medicine (Gallagher et al., 2025). Despite the availability of genomic information, there is still a bias of Western European ancestry, while efforts in countries such as Brazil and Mexico provide insights into historical patterns of genetic evolution and admixture of underrepresented populations (Nunes et al., 2025; Sohail et al., 2023).

In agriculture, large-scale SNP array genotyping and sequencing efforts are transforming livestock breeding. For example, the US National Cooperator Database (NCD) maintains the largest animal dataset worldwide for genetic evaluation, with more than nine million cattle with SNP array genotypes, 100 million individuals in pedigree, and 100 million lactation records (Miles et al., 2025). Additionally, commercial efforts have already imputed nearly a million pig haplotypes using low-coverage sequencing and pedigree information (Ros-Freixedes et al., 2022b).

As the volume of data continues to increase, conventional genotype matrix-based storage and analysis methods do not scale. A promising solution is inference of Ancestral Recombination Graphs (ARG) (Nielsen et al., 2025). An ARG encodes the genealogical relationships between DNA segments by tracing the evolutionary events of coalescence, mutations and recombination (See Fig. 1.3; Griffiths and Marjoram, 1997). While the ARG encodes the ancestral history of entire genome coalescence, they can be decomposed and simplified into local trees (Figure 1.4). Local trees are defined by recombination break points and represent the ancestry at a locus. The local tree is formed of nodes (the sampled haploid genomes and their ancestors) and branches (the ancestor-descendant connection), and annotated with mutations (Griffiths and Marjoram, 1997; Wong et al., 2024). A particularly efficient data format of ARG is the tree-sequence structure, which stores shared haplotypes between ancestors and their immediate descendants, achieving substantial gains in both storage and analytical efficiency (Wong et al., 2024; Kelleher et al., 2019).

Besides computational efficiency, ARGs enable a range of powerful evolutionary inferences. They provide encoding of the recombination history that allows resolving

fine-scale population structure, detecting local ancestry, estimating divergence times, and reconstructing demographic histories (Arenas, 2013). Moreover, ARGs have an unexplored potential for application in quantitative genetics. By replacing genotype matrices with the structured full genealogical context of each genetic variant, ARGs can support biologically grounded models of trait heritability, genetic values, and selection, especially in populations with complex evolutionary histories.

Among livestock species, dairy cattle represent one of the most successful cases of such genomic applictions. In American Holsteins, the use of genomic information for estimating breeding values nearly doubled the genetic gain for milk yield in less than a decade (García-Ruiz et al., 2016). Today, genotyping young selection candidates is routine in major dairy breeding programmes of the Global North, leading to improvements in productivity, health, and welfare.

In contrast, the availability of genomic data and the implementation of genomic selection in the Global South are limited due to economic constraints, infrastructure barriers, and inadequate data recording systems. Production in these regions often relies on locally adapted populations, mostly *Bos indicus* (indicine) in Latin America and Asia, *Bos taurus* (taurine) in Africa, or their crosses with intensively selected breeds from the Global North. Consequently, these populations are under-characterised, with less information on population structures, genetic differentiation, and breeding potential (Maiorano et al., 2018).

The 1000 Bull Genome Project (Hayes and Daetwyler, 2019) has played a pivotal role in developing genomic resources for cattle, unveiling the extent of genetic diversity and helping to link phenotypic variation with variation at the DNA level. Despite the significant impact of this work, there is a sampling bias toward taurine breeds from the Global North. The 1000 Bull Genome Project identified that while the sample size seemed sufficient to discover the many genetic variants in taurine genomes, the number of variants identified increased with the introduction of new indicine samples, highlighting the high level of genetic diversity in this cattle subspecies.

The taurine sampling bias in current public datasets and their use in population and statistical genetics pose challenges to the characterisation and improvement of indicine populations. Current methods may fail to capture the full complexity of indicine genomes, hiding important variation (Ogunbawo et al., 2024; Talenti et al., 2022). Thus, there is an urgent need for analytical frameworks that can accommodate complex population histories, high levels of admixture, and rich evolutionary dynamics. In

this study, we demonstrate how ARGs can be used to address this gap. We used the 1000 Bull Genome dataset comprising taurine and indicine cattle breeds worldwide to demonstrate how ARG inference with tree-sequences can manage storage requirements, resolve population structure, and enable local ancestry inference and demographic analysis in complex evolutionary and admixture contexts. Our work is the first to apply ARG inference to livestock genomes on a large scale and illustrates the versatility and power of ARG-based methods in this context, highlighting their potential for both research and breeding applications. This work aims to present the concept of ARGs to the livestock breeding community, demonstrate the inference procedure, and underscore both the promises and the current limitations, setting the stage for a broader application of these methods in livestock genomics.

2.2 Materials and methods

We inferred ARG in the form of time-resolved tree-sequences for the public dataset of the 1000 Bull Genomes Project Run8 (Hayes and Daetwyler, 2019), aligned against the ARS-UCD1.2 bovine genome assembly. Tree sequence inference was tested using a subset of the data and simulated data, comparing four inference parameter configurations to address an observed excess of mutations per site with real data. The established optimum configuration was applied to the final results. After tree-sequence inference, we performed population structure analyses and demographic inferences.

2.2.1 Whole-genome sequence genotype

The data were obtained from the European Nucleotide Archive (accession number PRJEB42783). Variant sites were filtered using GATK software (McKenna et al., 2010), retaining only autosomal biallelic SNPs that passed variant quality score recalibration (VQSR). Genotypes were phased with SHAPEIT5 (Hofmeister et al., 2023), using the recombination map from Ma et al. (2015) lifted to the ARS-UCD1.2 assembly of the bovine genome.

We will refer to this dataset throughout the text as **1KB**. It comprised 1,832 cattle samples from 109 breeds worldwide. To facilitate downstream analyses, we classified these breeds into five higher systematic groups: taurine (*Bos taurus*), indicine (*Bos indicus*), African, Primigenius (*Bos primigenius*), and Crossbred. For simplicity and interpretability, we restricted most analysis to a subset of six breeds focal breeds representative of the higher systematic groups taurine, indicine, and African cattle. These focal breeds span contrasting production systems (dairy, beef and dual-purpose) and evolutionary histories, thereby capturing key dimensions of global cattle diversity. An overview of the focal breeds including the number of individuals analysed (sample size) is provided in Table 2.1.

2.2.2 ARG inference

We inferred ARG, represented as tree-sequences, using *tsinfer* 0.3.1 (Kelleher et al., 2019) with default parameters. Inference parameter optimisation was performed and a detailed description of this process is given in the Supplementary Material Section A.1. *Tsinfer* uses a Hidden Markov Model (HMM) to infer the genealogies that describe the ancestry of the sampled chromosomal haplotypes in a dataset and was our method of choice due to its capacity to handle large sample sizes. Inference was performed

Table 2.1: Focal cattle breeds analysed in this study, grouped by higher systematic category, production type, and sample size.

Systematic group	Focal breed	Type	Individuals
Taurine (*Bos taurus*)	Holstein	Dairy	389
	Angus	Beef	121
Indicine (*Bos indicus*)	Gir	Dual-purpose	1
	Nelore	Beef	5
African	Ankole	African indicine	10
	N'Dama	African taurine	10

on phased autosome using ancestral allele information from Talenti et al. (2025). To estimate the age of ancestral haplotypes across inferred genealogies, we used *tsdate* 0.2.1 (Wohns et al., 2022), assuming a constant mutation rate of 1×10^{-8} per base pair per generation. The final tree-sequences were compressed using *tszip* 0.2.5 (Wong et al., 2024). After inference, we summarized information regarding the number of nodes, edges, trees, sites, mutation, and file sizes for each chromosome. We used *tskit* 0.6.3 for analysis of the inferred ARGs.

We used genealogical nearest neighbours (GNN) and branch-based genomic relationship matrix (GRM) to examine the population structure within the 1KB dataset. GNN quantifies, for each focal node, the proportion of its nearest neighbours in the local genealogical tree that belong to each reference set specified by the user (Kelleher et al., 2019). The GRM quantifies the genome-wide genetic similarity between and within pairs of individuals based on their shared haplotype history. The matrix is obtained by calculating the total area of shared branches across the chromosome ancestral to the sampled individuals in the focus set (Lehmann et al., 2025). Both matrices were generated for all individuals per chromosome and added across chromosomes with their lengths as weights. For visualisation of both the GNN and the GRM matrix, individuals were hierarchically clustered using the method *average* in the package *scipy* (Virtanen et al., 2020) and coloured according to their population. If the population was not one of the focal breeds, the individual was coloured according to the higher systematic group to which it belongs.

In addition, GNN was used to assign the breed composition of 100 individuals of unknown origin. We established a two-step approach. First, we defined the reference panel by calculating the GNN using the 100 individuals as focal set and samples from the taurine breeds as reference sets. We then calculated the contribution of each taurine breed by summing their GNN proportion per individual and selected the 10 most

representative. This process resulted in the following breeds with top GNN proportions: Holstein, Simmental, Braunvieh, Angus, Charolais, Brown-Swiss, Limousin, Hereford, Jersey, and Belgian-Blue. We then repeated the same procedure with samples from the indicine breeds as reference sets. The Brahman breed showed high representation, leading to its replacement of the Belgian-Blue in the final reference panel. In the second step, we recalculated GNN proportions using only the selected breeds as the reference set. In addition, we randomly selected an individual from the 100 to evaluate its GNN composition along each haplotype on chromosome 25, and estimated window-based time to the most recent common ancestor (TMRCA) between the two haplotypes.

The branch-based principal component analysis (PCA) (Lehmann et al., 2025) was performed for three regions of functional interest: the bovine leucocyte antigen (BoLA) gene on chromosome 23 [28720499, 28724294], associated with the immune response and known for its preferred heterozygous state (Giovambattista et al., 2020, 2013; Miyasaka et al., 2012; Takeshima and Aida, 2006); the DGAT1 gene on chromosome 14 [7110087, 7266726], associated with milk production traits in taurine breeds (Grisart et al., 2002; Winter et al., 2002); and the ZFAT gene, also on chromosome 14 [604179, 614155], associated with survival (Jenko et al., 2019).

We evaluated TMRCA between focal breeds using the branch-based `divergence` method. Pairwise divergence values were halved to reflect split times and hierarchical clustering was used to visualize the results (Virtanen et al., 2020).

Finally, we investigated the demographic history of focal breeds by computing coalescence rates over time using the `pair_coalescence_rates` method. These were calculated across 100 uniformly spaced time intervals from the present to the oldest node time in the tree-sequence. We used the inverse of the coalescence rate as an estimate of the effective population size (N_e).

To ensure a robust estimation of N_e, we implemented several optimisation strategies in the tree-sequence inference process (see Section A.1 for details). In early versions, sites with excessive number of mutations (>100 per site) were observed (Figure S1), leading to inaccurate tree-sequence dating and overestimated N_e, particularly near the present (Figure S3). This artifact was exacerbated in inference runs using the recombination_rate parameter during the tree-sequence inference. By removing sites with more than two mutations and reinfering the tree-sequence, with the method simplify from *tskit* with `reduce_site_topology=True` to eliminate fixed or uninformative sites, we were able to recover estimates consistent with the reference model previous published

by MacLeod et al. (2013).

Scripts for all the analyses steps are available at:
GitHub repository `https://github.com/gmafrafortuna/tree-sequence-tests`.

2.3 Results

We inferred genome-wide tree-sequences for 1,832 diploid individuals classified into 109 breeds. The final inference yielded 26.9 million time-resolved local trees across 28.4 million variant sites on 28 autosomal chromosomes. The trees contained 74 million nodes, 1 billion branches and 28 million inferred mutations, and required 46 GB of storage space. Using *tszip*, the required space for compressed tree-sequences was reduced to 9.9 GB. Table S1 summarizes the statistics per chromosome, including tree count, node count, edge count, and compressed file sizes.

2.3.1 Population structure from genealogical relationships

We characterised broad and fine-scale genetic structure using two genealogically informed measures: the GNN matrix and the branch-derived GRM. Both approaches revealed clear clustering by breed and higher order group. In addition to expected clusters between populations, stratification within populations is observed, which may indicate line formation within the same breed, for example populations from different countries (Figure 2.1).

Taurine and indicine animals formed distinct clusters, with African individuals positioned mostly within the indicine cluster. Holstein and Angus individuals formed large, internally structured clusters, while Nelore individuals showed a more diffuse distribution. Limited sample sizes restricted the resolution for Gir, N'Dama, and Ankole. Some individuals labelled as indicine were found clustering with taurine samples and a large block of taurine samples was observed in the indicine cluster. This could either indicate a mislabelling issue or admixture.

We used GNN proportions to infer ancestral contributions in randomly selected 100 individuals labelled only as taurine (Figure 2.2a). All individuals exhibited some degree of Holstein ancestry, ranging from 0.85 to 0.03, alongside frequent contributions from Brown-Swiss and Jersey. Shared ancestry with one or more taurine beef breeds (Hereford, Charolais, Limousin, Angus, or Simmental) was also predominant. Overall contribution from each breed was: Holstein 0.27, Simmental 0.06, Braunvieh 0.40, An-

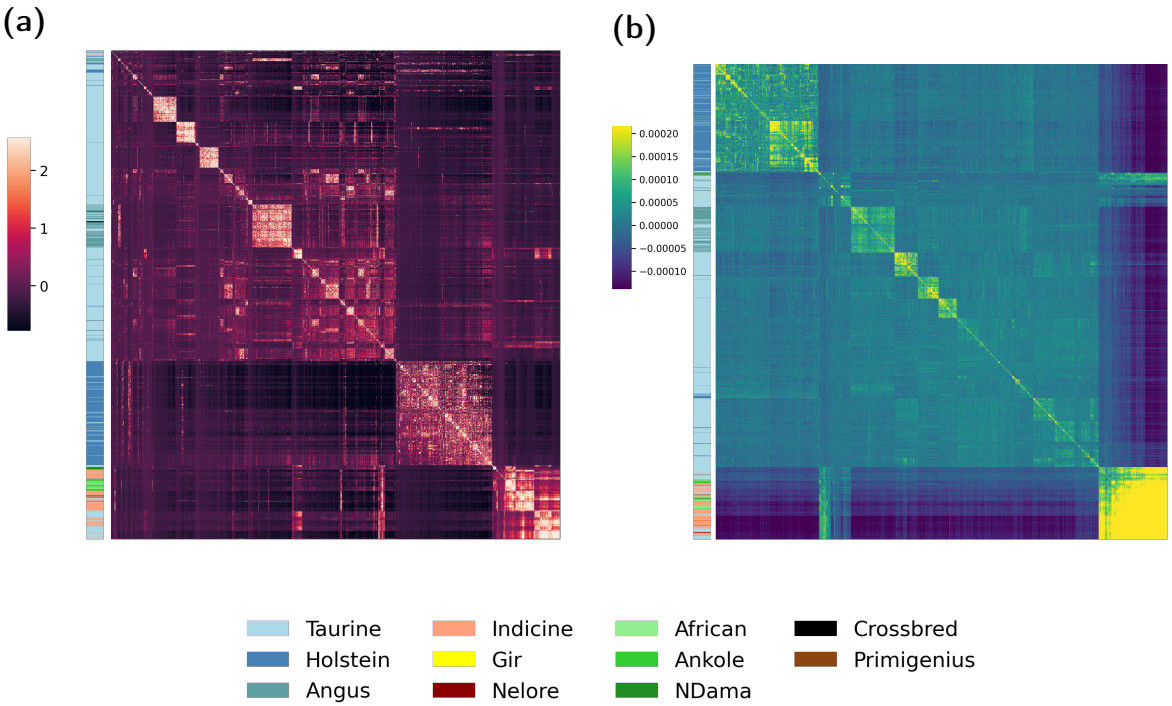

Figure 2.1: Genetic structuring of individuals and populations. A) heat-map of the genealogical nearest neighbour matrix. B) Heat-map of the genetic relationship matrix. Higher values indicating closer relationships. Colour bar highlights the breeds (Holstein, Angus, Gir, Nelore, N'Dama, and Ankole) and higher systematic group (Taurine, Indicine, African, Primigenius, or Crosbred).

gus 0.40, Charolais 0.03, Brown-Swiss 0.03, Limousin 0.03, Hereford 0.02, Jersey 0.02. Brahman ancestry was also detected in several individuals.

To assess local ancestry resolution, we analysed both haplotypes of chromosome 25 of a randomly selected individual ('SAMEA5714973'). GNN proportions varied over chromosomes and between haplotypes (Figure 2.2b). We identified three blocks of distinct ancestry in haplotype 1. One presumably associated with the Hereford breed, another with Simmental, and the last with Angus. Haplotype 2 showed a predominant association with the Angus breed. The TMRCA between the two haplotypes varied along the genome, in line with the differences in breed GNN proportions.

2.3.2 Population divergence from TMRCA estimates

Pairwise TMRCA estimates across focus populations further quantified divergence times (Figure 2.3). The indicine breeds, Gir and Nelore, formed a consistent clade, diverging from the other populations around 20,000 generations ago, while their divergence was also deep (around 16,000 generations ago). The taurine breeds Holstein and

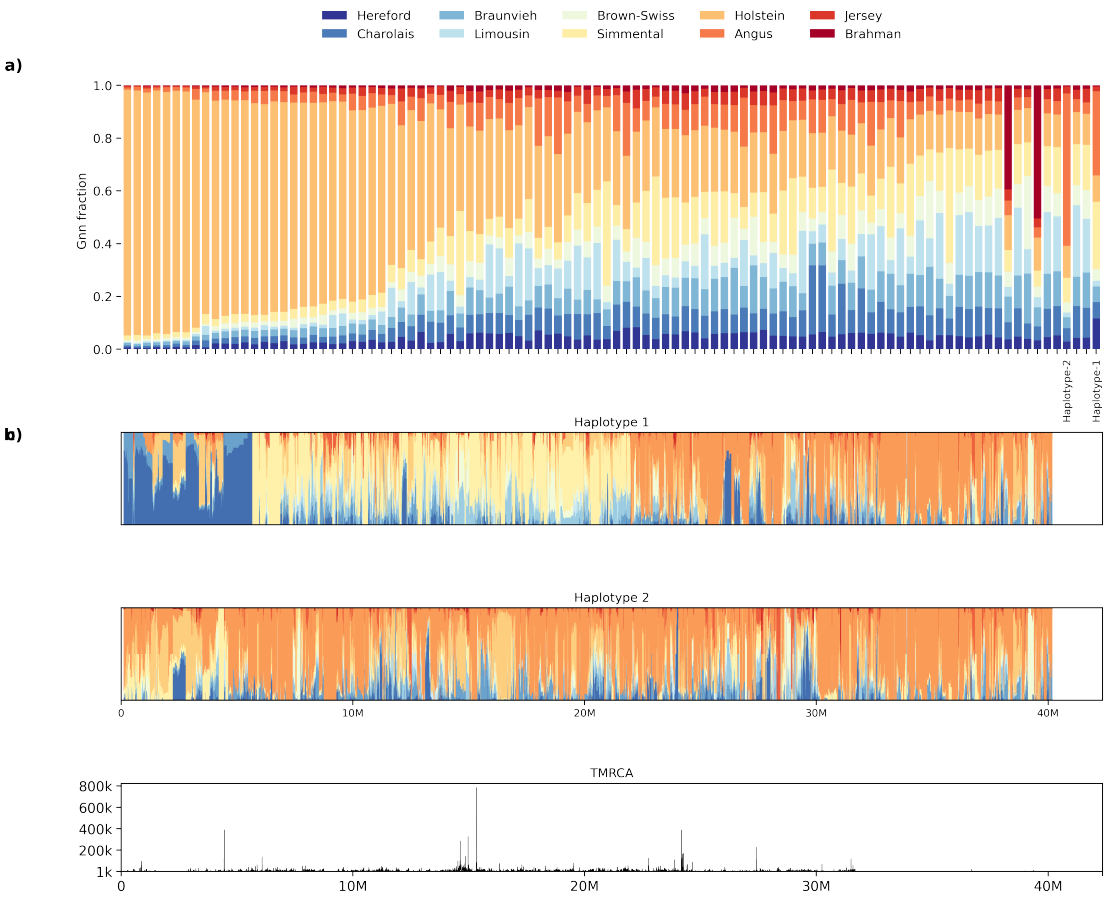

Figure 2.2: Genealogical nearest neighbor proportions of reference populations for 100 taurine individuals at chromosome 25. (A) Stacked barplot for individuals and their GNN proportions of Hereford, Charolais, Braunvieh, Limousin, Brown-Swiss, Simmental, Holstein, Angus, Jersey, and Brahman. (B) Stacked barplot for two haplotypes of the selected individual ('SAMEA5714973'), and time to most recent common ancestor (TMRCA) in generations between the two haplotypes (bottom panel).

Angus diverged from each other in more recent times, around 5,000 generations ago, and from the African taurine N'Dama around 7,500 generations ago. Ankole showed the deepest divergence within taurine breeds, with a TMRCA to the others close to 12,000 generations ago.

Coalescence rate distributions

The time-resolved tree-sequence provides a high-resolution view into the genealogy of sampled haplotypes. We used the inverse coalescence rate (ICR) as an estimate for the effective population size (Ne) of the six focal breeds over 1,000,000 generations (Figure 2.4). The demographic model of cattle from MacLeod et al. (2013) served as a reference model for comparison.

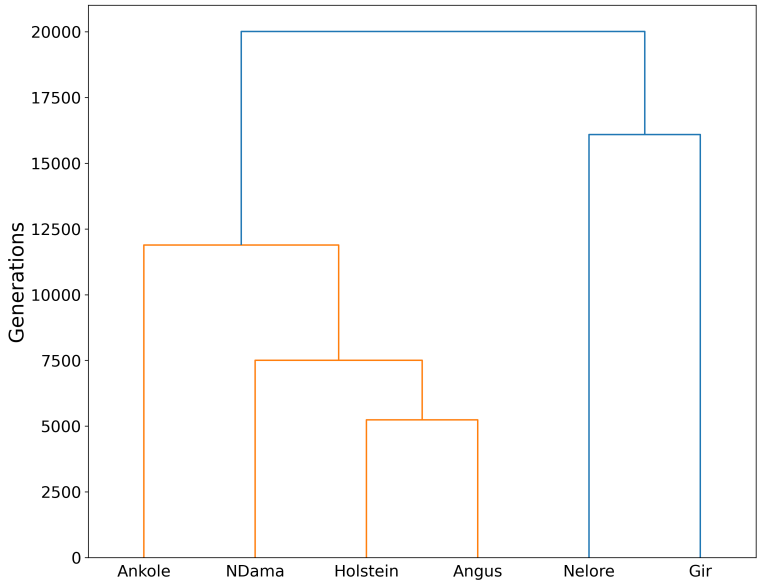

Figure 2.3: Dendrogram for focal breeds for chromosome 25. Dendrogram is based on time to most recent common ancestor (TMRCA) between individuals. The dendrogram shows the relationship between the six focal breeds.

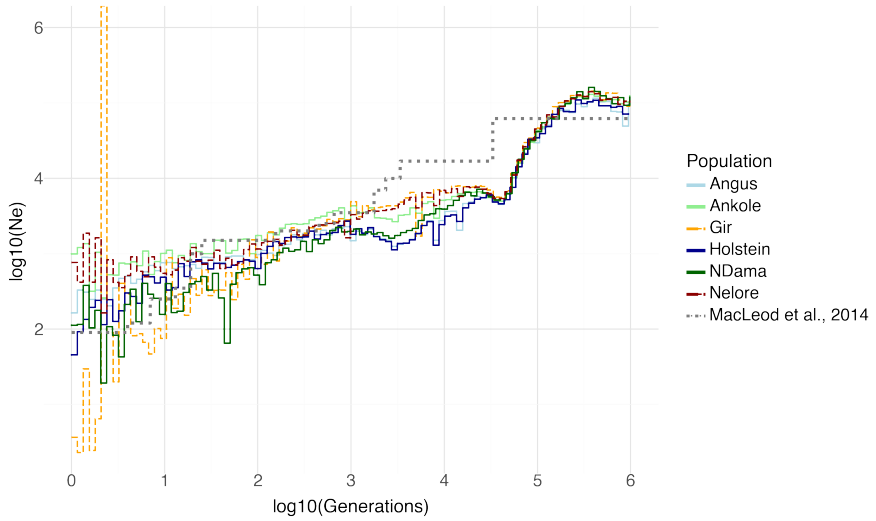

Figure 2.4: Inverse coalescence rate estimates for focal breeds for chromosome 25. Effective population size as the inverse of the coalescence rate for the six focal breeds (Holstein, Angus, Gir, Nelore, Ankole and N'Dama). The demographic model of cattle from MacLeod et al. (2013) is shown in grey as reference. Time is shown in the x-axis in generations. Both axes use logarithmic scale.

The Holstein population showed a consistent pattern with the reference demographic model, and served as a validation of our inference. Although tree sequence-based estimates generally agreed with the reference demographic model, a mismatch in N_e was observed for the period between 1,000 and 100,000 generations ago. Nevertheless,

both approaches converged on a present-day estimation of approximately 100.

Across breeds, similar historical trends were observed, with N_e dropping from 100,000 to around 3,000 in the period between approximately 100,000 and 10,000 generations ago. After that we observe population differentiation that continued until the present day. Nelore and Ankole had the highest current N_e estimates, close to 1,000. Estimates for Holstein, Angus, and N'Dama were around 100. The Gir population showed the lowest estimates, close to 10.

Locus-specific ancestry analysis

Figure 2.5 shows the BoLA region (first column) with weak clustering structure across principal components. In contrast, DGAT1 and ZFAT regions (second and third columns) show clear separation between taurine and indicine populations in the PC1 vs. PC2 space. Higher-order PCs further resolved the within-taurine population structure, with Holstein individuals showing a distinct direction for the DGAT1 gene.

2.4 Discussion

2.4.1 Tree sequence inference enables efficient encoding of evolutionary events, with caveats

The tree-sequence data format takes advantage of the correlation between genealogical trees along the genome, based on coalescence, mutations and recombination events, to optimize the storage of large-scale genomic data. When Kelleher et al. (2019) introduced the algorithm for inferring ARGs for million of samples, the promise was to enable population-scale genealogy reconstruction with computational and storage efficiency. Our study applies this methodology to a diverse livestock dataset, showcasing the utility of the algorithm in species with complex population structure and varying demographic histories.

We successfully inferred genome-wide tree-sequences for nearly 2,000 diploid individuals across 109 populations and more than 28 million segregating sites. This generate 73,954,207 nodes, of which 73,950,207 are ancestral nodes, indicating a high level of haplotype diversity in the dataset. No data size reduction was observed, with change from 8.4 G with the VCF format to 9.9 GB using *tsinfer* and *tszip* (Table S1). This is expected given the dimensions we are working with. According to Kelleher et al. (2019), the size reduction benefits start to show when initial file sizes are closer to 1 GB.

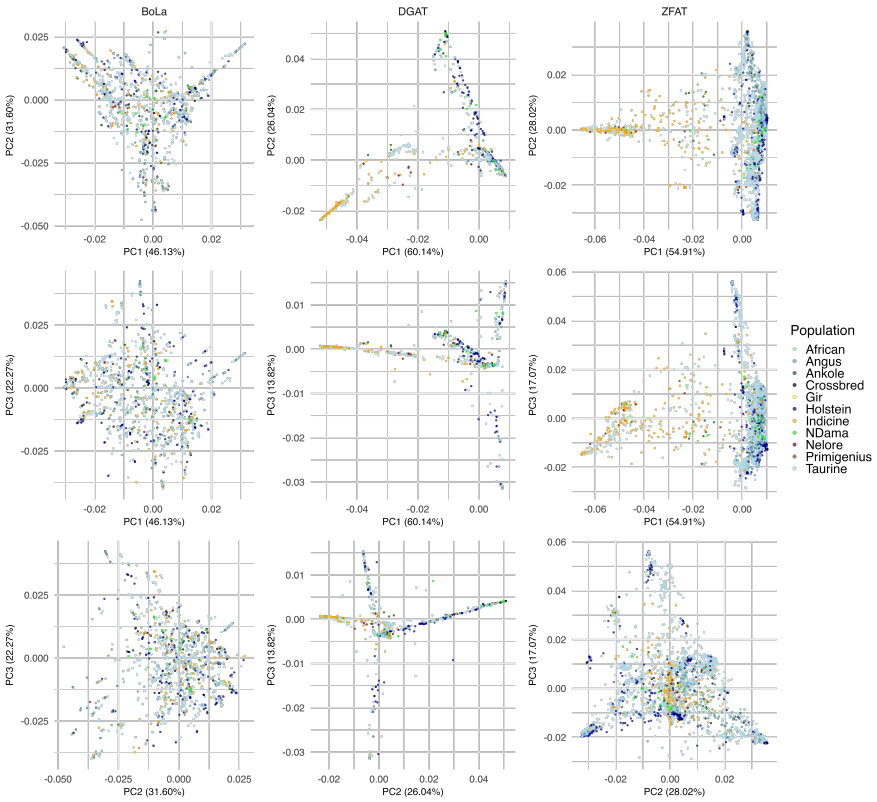

Figure 2.5: Local principal component analysis of three cattle loci. Tree sequence branch-derived principal components for chromosomes 23 at BoLa region (left panel) and chromosome 14 at DGAT1 region (middle panel) and ZFAT region (right panel). The first two principal components are shown in the x-axis and y-axis, respectively. Rows show different principal components with PC1 vs PC2 in the first row, PC1 vs PC3 in the second row, and PC2 vs PC3 in the third row. The colours indicate the classification of the individuals based on target populations (Holstein, Angus, Gir, Nelore, Ankole, NDama, taurine, indicine, Crossbred, and Primigenius).

Contrasting with the conventional matrix-based variant call format (VCF), the tree-sequence relies on six tables to represent nodes, edges, mutations, sites, and individuals. Increasing the number of samples and variants in a VCF scales quadratically, quickly creating constraints of storage and analyses. The tree-sequence storage format, on the other hand, scales linearly, facilitating work with large genomic datasets.

Despite these advantages, the inference process remains computationally demanding. For example, inferring the tree-sequence for the 1,832 samples at chromosome 1, the largest bovine autosome, required more than 26 hours with a maximum RAM requirement of 150 GB (Table S4). These steps can become even more demanding for larger datasets with more complex demographies, which may pose practical limitations for users without access to high-performance computing resources.

It is important to note that since the completion of this work, the software has been updated and improvement in performance is expected. Nonetheless, our findings highlight the current trade-off between data compression and computational efficiency in the ARG inference process, especially for complex genomic datasets. There are also other software available for inferring ARGs that differ in computational requirements such as Relate (Speidel et al., 2019), ARGNeedle (Zhang et al., 2023), and SINGER (Deng et al., 2024), but a general trend suggests a trade-off between accuracy of inference and computational efficiency or capacity to process high numbers of samples.

2.4.2 Resolving population structure and breed composition in admixed cattle

In the context of animal breeding, a key application of the tree-sequence is to assign the breed (proportions) and to determine the breed of origin of alleles (BOA). The way tree-sequences encode genomic information provides a scalable and informative framework to investigate ancestry at both genome-wide and local levels. These insights are critical for understanding and managing genetic diversity and for genetic evaluations, especially in multibreed and crossbreeding contexts.

Conventional breed assignment or BOA methods generally do not scale well to large datasets and require assumptions of breed composition, which can be limiting when working with highly admixed or understudied populations (Eiríksson et al., 2021; Vandenplas et al., 2016). In contrast, tree-sequence-based statistics, such as the genealogical nearest neighbour (GNN), enable ancestry inference without prior assumptions about the constitution of the population in study. The only requirement is a reference

set of purebred individuals, which can be easily obtained from publicly available or curated datasets, and further validated with the GNN itself.

In our study, we demonstrated the effectiveness of the GNN method in resolving ancestry in populations with unknown or mixed origins by implementing a two-step GNN-based admixture analysis that enabled us to derive more interpretable and stable estimates of both whole-genome ancestry (Figure 2.2A) and local ancestry along the genome (Figure 2.2B). This two-step analysis is particularly valuable in settings with extensive admixed or limited documentation. The context of African cattle gives good examples where taurine-indicine introgression happens for generations without clear records of the original ancestry (Gebrehiwot et al., 2020). Our use of the publicly available 1KB genomes dataset demonstrates how existing resources can be repurposed for ancestry estimation when paired with tree-sequence methods, without requiring an priori breed knowledge.

The ability to conduct ancestry inference at the haplotype level along the genome mirrors traditional BOA approaches, but free from assumptions regarding breed composition. For instance, in one representative animal, we identified distinct ancestry blocks on chromosome 25, with ancestry contributions from Hereford, Simmental, and Angus in one haplotype, while the other was predominantly associated with the Angus population. This fine-scale resolution, paired with the local TMRCA offers insight into both the origin and timing of admixture events, and the inheritance pattern of genomic regions.

While GNN-based population inference offers resolution and flexibility, it is sensitive to the quality of phasing, the accuracy of tree-sequence inference, and the level of diversity in the reference dataset. Admixed individuals with ancestry missing from the reference panel might be misclassified and assigned to genetically similar but incorrect populations. Future work should evaluate the robustness of this approach under varying levels of admixture and population structure.

A natural extension of this work is to explore whether GNN-based proportions can serve as proxies for breed composition in genomic prediction models. Since these proportions are derived directly from the shared genealogical history encoded in the ARG, they may offer meaningful covariates for use in crossbred evaluations. This application could help address challenges associated with ancestry-specific marker effects that complicate multibreed or crossbred genomic predictions.

2.4.3 Model assumptions and sensitivity of coalescence patterns

As established earlier, tree-sequences open up powerful opportunities for reconstructing ancestral genealogies and inferring aspects of population history. Our results demonstrate how this representation can be leveraged to estimate population parameters and even infer demographic events, such as admixture events, divergence times, and changes in effective population size. However, these insights are only as robust as the underlying ARG inference. The ARG inferrence is shaped by both biological and technical assumptions. These assumptions directly affect inference quality and thus are worth discussing, especially as we intend to use tree-sequence output to inform breeding decisions.

By default, the inference algorithm behind *tsinfer* assumes the infinite-sites mutation model (Kimura, 1969), which assumes that new mutations occur at a previously unmutated site, preventing recurrent or back mutations at the same site. While this assumption simplifies the inference, it is likely violated. Factors such as large cenzus population sizes, strong selection on linked sites, and complex population structure, all of which are characteristic of artificially bred populations, can increase the occurrence of recurrent mutations, back mutations and (unmodeled) structural variation (Avalos-Pacheco et al., 2024; Conrad and Hurles, 2007). These processes may distort the inferred genealogies by affecting tree topology and consequently the estimates of coalescence times.

Recent work in human populations further underscores the limitations of the infinite-site model assumption. Gao et al. (2023) demonstrated that in large-scale datasets, recurrent mutations at hypermutable sites (CpG dinucleotides) can lead to biases in ARG-based inference methods. These sites are often excluded during the inference due to ambiguity in their placement on the tree, which in turn reduces variant density and alters the apparent distribution of the age of mutation. This distorts demographic inference and may mask recent evolutionary events.

In our study, we observed similar patterns with *tsinfer*. Some sites in the HOL dataset accumulated over 100 mutations, strongly suggesting violation of the infinite-site model. These multi-hit sites inflated edge counts (Table S3) and distorted coalescence time estimates. Using inverse coalescence rates, we were only able to recover trends consistent with the known cattle demographic model after removing sites in the data with more than two mutations, rephasing and re-inferring the ARG (Figure S3). While this fil-

tering enforced compliance with the model assumptions, it is not ideal.

A more principled alternative is to adapt the inference parameters. In *tsinfer*, this means adjusting the `mismatch_ratio` parameter. Although we did not explore this further than using the default value 1, the current recommendation from software authors is to reduce even further this parameter during the `match_samples` step. This may improve robustness to recurrent mutations and improve inference outcomes. Future work should investigate parameter tuning strategies more systematically and evaluate their impact on ARGs accuracy across different population structures and data qualities.

Concluding remarks

In conclusion, this study demonstrates the practical application of tree-sequence inference for compressing genomic data and resolving complex population structures in livestock genomics. By applying the method to diverse and admixed cattle populations, we illustrate how tree-sequences can promote ancestry-aware analyses at both local and genome-wide scales without relying on rigid or predefined assumptions about ancestry composition.

The explicit representation of genealogical relationships along the genome brings important implications to breeding and even population management. Tree-sequences provide direct analysis of relatedness, allowing the identification of extended runs of homozygosity (ROH), and the genealogical origin of homozygous segments. Such information can be used to monitor levels of inbreeding, distinguish ancient and recent ROH, and to design mating strategies to minimise the accumulation of deleterious segments and maintain genetic diversity.

Despite the benefits of the tree sequence methodology, however, inferring ARGs remains computationally demanding, especially for large and structurally complex datasets. Another important limitation to the wide application of the method in livestock genomics and breeding settings is its inability to model structural variation, as it plays a critical role in the differentiation of populations. Importantly, ARG inference methodology is being developed focusing mostly on human and model-organism data. Agricultural datasets present different challenges that require adjustments in methodology. Nevertheless, ARG inference is in constant development and presents great potential for breeding applications.

Code availability

Scripts for tree-sequence inference, optimisation and analyses presented in this manuscript are available at:
GitHub repository `https://github.com/gmafrafortuna/tree-sequence-tests`.

ORCID of Authors

Gabriela Mafra Fortuna 0000-0001-8921-642X
Jana Obšteter 0000-0003-1511-3916
Hannes Becher 0000-0003-3700-2942
Gregor Gorjanc 0000-0001-8008-2787

Funding

GMF and GG acknowledge funding from BBSRC DTP (EASTBio) CASE PhD studentship with Genus, BBSRC Institute Strategic Programme funding to The Roslin Institue (BBS/E/D/30002275, BBS/E/RL/230001A), BBSRC grants BB/T014067/1 and BB/M009254/1, and The University of Edinburgh.

3 Estimating haplotype values and mutation effects in the context of a local DNA tree

This chapter presents the manuscript *Estimating haplotype values and mutation effects in the context of a local DNA tree.* The study introduces a new statistical model to estimate haplotype and ancestry-specific mutation effects considering the context in which the mutation occurs.

This work focuses on two limitations of performing genomic prediction using whole-genome sequence data, (i) the computational burden of dense relationship matrices and (ii) the instability of current methods to capture the context of mutations effects. The ancestral recombination graph (ARG) is used in replace of the dense genomic relationship matrix to capture the intricate inheritance patterns of DNA across generations.

As proof-of-concept, the model is applied to cattle mitochondrial DNA, a non-recombining region that yields a single ARG representing the genealogical history of the entire genome. A sparse (co)variance matrix is constructed using the conditional distributions of ancestor-descendent haplotype values within the ARG.

The results suggest the model can achieve accurate predictions while significantly reducing the computational burden associated with genomic prediction. The model serves as an introduction to the application of ARG to quantitative genetics, particularly in the context of genomic prediction and breeding programs.

The manuscript is a joint work of Gabriela Mafra Fortuna[1], Jana Obšteter[2], Ajda Moškrič[2], Gregor Gorjanc[1].

[1] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Edinburgh EH25 9RG, Midlothian, United Kingdom; [2] Agri-

cultural Institute of Slovenia, Hacquetova ulica 17, 1000 Ljubljana, Slovenia.

Core ideas

- Ancestral recombination graphs (ARGs) can be leverage to perform biologically informed genomic prediction by encoding shared ancestry and sequence context, instead of relying on standard allele dosages.
- In a non-recombining genome, interpreting haplotype values as cumulative mutation effects along ARG branches, it is possible to estimate ancestry-specific effects and local epistasis, offering greater biological resolution.
- The recursive nature of the ARG-based haplotype relationship and sparse matrix operations enables efficient computation, however, extending this to recombining genomes is computationally challenging.

Abstract

Genomic prediction with whole-genome sequences has challenges and opportunities. High computational cost and the seemingly limited increase in accuracy are significant challenges. Opportunities include capturing causal mutations and estimating their effects in the context of genome sequence to boost biological discovery. We present a novel statistic model aiming to address these. Our model leverages local tree structure in ancestral recombination graphs (ARG), which describe genetic variation in the presence of mutation, recombination and coalescence, to efficiently predict haplotype and ancestry-specific effects. We assume the difference among ancestor-descendent haplotype values is the effect of the mutation(s) occurring when haplotypes are transmitted between generations. In an ARG, this process is represented by branches connecting both haplotypes. Thus, haplotype values are the sum of all branch effects from it to the tree root. By estimating the haplotype effects from observed phenotype values and haplotypes embedded in a tree, we estimate branch/mutation effects that can vary depending on the sequence context. This enables estimation of ancestry-specific effects and local epistasis, e.g. due to variation in a codon or other functional DNA elements. The tree structure generates a recursive and sparse model via conditional distributions of ancestor-descendant haplotype values, allowing efficient calculations via generalised Cholesky decomposition of variance and precision matrices between haplotype values. The model is demonstrated via a sample of cattle mtDNA and associated real/simulated phenotypes. Since mtDNA is non-recombinant, we work with a single ARG spanning the entire genome. Results show how our mutation-event-aware model differs

from the standard variant-allele-dosage model and generates more accurate estimates of mutation effects and haplotype values in different genetic backgrounds. Extending the model to recombining genomes still imposes a significant computational challenge and is ongoing research.

3.1 Introduction

Genomic selection has revolutionised animal breeding. The use of a large number of genome-wide SNP markers combined with powerful statistical models has enabled more informed quantitative genetic analyses and more accurate or earlier breeding decisions compared to traditional pedigree-based approaches (Meuwissen et al., 2001, 2016; Johnsson, 2023b). One of the most successful implementations of genomic selection has been in dairy cattle. There, the use of genomic predictions more than halved the generation interval for the male path of selection and, consequently, doubled the rate of genetic gain per unit of time (García-Ruiz et al., 2016; Wiggans et al., 2017). Similar success stories have also been reported for the implementation of genomic selection in other species (Bagnato and Rosati, 2012; Hickey et al., 2017).

Despite the undeniable power of genomic data for quantitative genetics and animal breeding applications, challenges and opportunities for further improvement remain. One such challenge is that mutations and their effects at causal loci can differ across genetic distances (between generations or populations), but most quantitative genomic models assume constant effects. Some of the causes for different mutations and their effects across genetic distances are allelic heterogeneity, dominance and epistatic interaction effects, genotype-by-environment interaction effects, and population genetic processes that generate sufficient allele and genotype frequency variation at sites contributing to these effects (Fisher, 1919; Falconer and Mackay, 1996; Mackay, 2014; Jones et al., 2014; Hormozdiari et al., 2017; Wainschtein et al., 2022; Boye et al., 2024; Herrera-Luis et al., 2024). Theoretical modelling and simulations show how mutation effects change due to interactions and population genetic processes (Mäki-Tanila and Hill, 2014; Jones et al., 2014; Legarra et al., 2021; Wientjes et al., 2022; Oman et al., 2022), which is corroborated with empirical results from a range of species (Park et al., 2022; Ling et al., 2020; Di Bari et al., 2024; Tang et al., 2023). The term epistatic drift has been proposed to describe these changes in mutation effects (Park et al., 2022). However, two recent studies of human admixed populations found very high correlation between effects sizes from multiple continental ancestral populations (Hou et al., 2023; Hu et al., 2025). Perhaps this is due to the fact that the majority of the genetic

variance for traits is statistically additive (Hill et al., 2008) and very large datasets are required to estimate the interaction effects (Hivert et al., 2021; Sandler and York, 2025).

Another challenge is that most applications use marker genotype data from SNP arrays. Many of these SNP markers were chosen because they were polymorphic across several populations (meaning that they largely predate population splits) and were close to uniformly covering the genome. While this is an effective strategy to capture substantial amounts of genetic variation between and within populations, many of the rare or recent mutations are not captured (Young, 2022). Since many causal loci with larger mutation effects are expected to host rare or recent mutations and many causal loci with smaller mutation effects are expected to host older mutations (Gibson, 2012), these effects are captured through an intricate linkage-disequilibrium between causal and marker loci (Meuwissen et al., 2001; Qanbari et al., 2010), but also between the loci and population or pedigree structure (Habier et al., 2013; Cuyabano et al., 2024; Boichard et al., 2025; Wicki et al., 2025). Because linkage-disequilibrium is dynamic across generations of a population and across populations, so are estimates of allele substitution effects at the marker loci with respect to the true mutation effects at the causal loci. This dynamic is driven by the genetic distances between genotyped individuals (and the associated time for recombinations between marker and causal loci) (Habier et al., 2013; Cuyabano et al., 2024; Boichard et al., 2025; Habier et al., 2007; Clark et al., 2012; Hickey et al., 2014; Scutari et al., 2016; Privé et al., 2022; Ding et al., 2023), though this relationship varies between traits and contexts, also indicating the importance of trait evolution, genetic architecture, and contexts in which data are collected (Hou et al., 2024; Wang et al., 2024b).

The challenges with rare mutations and drifting mutation effects and estimating them via linked SNP markers are particularly relevant for genomic predictions across generations within a population and across populations, where the genetic distance between training and prediction sets is large. Such genomic predictions are particularly important for numerically small breeding programmes with limited training set sizes (Raymond et al., 2018) and in crossbreeding programs, where it is relevant to predict across populations (Hayes et al., 2023). The accuracy of such predictions is moderate, although expanding the genomic prediction model into main and population-specific effects has been shown to increase the accuracy somewhat, with mixed results across studies (Hayes et al., 2023; Sevillano et al., 2017; Junqueira et al., 2020; Karaman et al., 2021; Eiríksson et al., 2021; Guillenea et al., 2023; Tabet et al., 2025; Londoño-Gil et al., 2025). These results mirror the results from human genetics, where transferability of

polygenic risk scores is a challenge due to potential differences in genetic architecture across populations, past demography and selection effects, and other factors (Ding et al., 2023; Martin et al., 2017; Durvasula and Lohmueller, 2021; Yair and Coop, 2022; Wang et al., 2022; Kachuri et al., 2023).

Theoretically, some of these challenges could be addressed by replacing SNP arrays with whole-genome sequence (WGS). The advantages of WGS would come from capturing more genome variation, avoiding the ascertainment bias of SNP array data, and encompassing both marker and causal loci, which would enable more accurate estimates between generations as well as across larger genetic distances (Meuwissen and Goddard, 2010; Hickey, 2013; Goddard, 2017; Johnsson, 2023a). However, genomic prediction with WGS has shown marginal improvements in accuracy over SNP arrays with inconsistent results between studies (see review from (Ros-Freixedes, 2024)), even with the largest WGS efforts to date (Ros-Freixedes et al., 2022a). One reason for this is that these studies mostly used biallelic SNPs, ignoring structural variants, and significantly leveraged imputation to increase the training set size. The benefit of WGS over SNP arrays is also often context-dependent, varying by trait, population, and computational methods (Lin et al., 2025). Computational effort required for working with WGS is also a major challenge in itself.

One way forward with WGS data that is gaining traction in population and human genetics, is to use Ancestral Recombination Graphs (ARGs) to efficiently represent, store, and analyse large-scale WGS datasets (Kelleher et al., 2019; Ralph et al., 2020; Zhang et al., 2023; Wong et al., 2024; Zhu et al., 2024; Gunnarsson et al., 2024). ARGs provide a comprehensive way to represent observed variation in a sample of genomes through past branching/coalescence, mutation, and recombination events (**?**Griffiths and Marjoram, 1997; Wong et al., 2024). Recent introductions to ARGs are provided by (Harris, 2023; Brandt et al., 2024; **?**; **?**), while (Wong et al., 2024) focuses on how to encode an ARG. An ARG can be represented as a graph, where haplotypes are treated as nodes and connections between haplotypes as edges (also called branches). The haplotypes can span whole chromosomes. The branches represent the transmission of DNA between immediate ancestor and descendant haplotypes, across one or many generations. Mutations occurring along the branches change the sequence of descendant haplotypes, while recombinations positionally change their ancestors. At a specific position in the genome, an ARG can be represented as a (local) DNA tree, where the root haplotype is the most recent common ancestor of all sampled haplotypes. While there is now an active development of quantitative genetic methods to work with ARGs and phenotypes (Ralph et al., 2020; Zhang et al., 2023; Zhu et al., 2024; Rebollo et al.,

2025; Lehmann et al., 2025; Lee et al., 2025; Christ et al., 2025), it is still unclear how this tree-based modelling benefits the estimation of rare and drifting mutation effects.

The aim of this contribution is to study how a tree-based model can be used to estimate the effect of mutation events and understand how they drive variation in haplotype values and downstream phenotype values. To this end, we organise this contribution in three parts. First, we study a statistical model of haplotype values on a local DNA tree with branch effects representing the sum of all mutation effects along the branch in that specific genome sequence context and hence also generation or population context. We demonstrate this model with a small example and highlight its key features. Second, we extend the demonstration with a simulation study based on the tree of mitochondrial DNA (mtDNA) from the 1000 Bull Genomes Project data (Hayes and Daetwyler, 2019; Dorji et al., 2022), where we consider multiple scenarios of mutation effects and study how well we can estimate them. Finally, we apply the model to a real mtDNA and phenotype dataset (Brajkovic et al., 2025). Throughout, we focus on a simplified case of a non-recombining region of the genome and ignore recombination, which is a focus of other work (Lehmann et al., 2025; Rebollo et al., 2025; Lee et al., 2025). Results show that leveraging information from a local DNA tree can improve the estimation accuracy, provide more interpretable results, and speeds up the computations.

3.2 Materials and methods

In this section we develop theory for a statistical model of phenotype, genetic, and haplotype values, and branch and mutation effects on a local DNA tree, and demonstrate it with a small example. We then evaluate the model on simulation data based on the mtDNA tree from the 1000 Bull Genomes Project (Hayes and Daetwyler, 2019; Dorji et al., 2022) and real mtDNA and phenotype data (Brajkovic et al., 2025).

3.2.1 Theory

Phenotype value model

Following Fisher (1919), we model the observed phenotype value of an individual, y_i, as a linear combination of the population mean μ, genetic value a_i for the genome region of interest, and residual $e_i \sim \mathcal{N}(0, \sigma_e^2)$ with σ_e^2 the residual variance:

$$y_i = \mu + a_i + e_i.$$

We focus on additive genetic values only and call them genetic values for brevity, but note that these are statistical parameters that depend on additive and non-additive genetic effects, and allele and genotype frequencies in a population (Fisher, 1919; Falconer and Mackay, 1996; Mäki-Tanila and Hill, 2014). For a whole data set of n_y phenotype records, the model in matrix form is:

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

where $\mathbf{1}$ is an $n_y \times 1$ design vector of ones for the population mean, $\mathbf{Z}$ is an $n_y \times n_i$ design matrix for genetic values, $\mathbf{a}$ is an $n_i \times 1$ vector of genetic values, $\mathbf{e}$ is an $n_y \times 1$ vector of residuals $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_e^2)$, and $\mathbf{a}$ and $\mathbf{e}$ are assumed to be uncorrelated.

Genetic value model

Let $\mathbf{x}_i$ be an $1 \times n_l$ genotype vector for individual i in the genome region of interest spanning n_l segregating biallelic loci, encoded as the dosage of alternative allele 1 compared to the reference allele 0, giving $\mathbf{x}_i \in \{0, 1, 2\}^{n_l}$. For n_i individuals, the corresponding $n_i \times n_l$ genotype matrix is $\mathbf{X} \in \{0, 1, 2\}^{n_i \times n_l}$. The genetic value of an individual (a_i) and the whole population of individuals ($\mathbf{a}$) are a linear combination of individuals' genotypes and corresponding allele substitution effects α (Fisher, 1919;

Falconer and Mackay, 1996):

$$a_i = \mathbf{x}_i \alpha, \qquad \mathbf{a} = \mathbf{X}\alpha. \tag{3.1}$$

By convention, genetic values are expressed as deviations from the population mean μ (Fisher, 1919), which is achieved by centring the genotypes in Eq. 3.1 around their expected dosage at each locus. We omit this centring from the equations for brevity and note that it does not change the estimable contrasts between model parameters.

The allele substitution effects and corresponding genetic values are estimated from observed phenotype and genotype data, commonly using the SNP-BLUP/GBLUP approach (Meuwissen et al., 2001; VanRaden, 2008). To facilitate the estimation due to a large number of segregating loci and linkage-disequilibrium between them, SNP-BLUP/GBLUP assume a normal prior distribution for the allele substitution effects:

$$\alpha \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\sigma_\alpha^2\right), \tag{3.2}$$

with σ_α^2 the variance of allele substitution effects. This implies a normal distribution for genetic values given the genotype matrix:

$$\mathbf{a}|\mathbf{X} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{X}\mathbf{X}^T\sigma_\alpha^2\right). \tag{3.3}$$

This approach assumes that allele substitution effects are constant across generations of a population and/or across populations. We aim to relax this assumption using information from a local DNA tree of haplotypes.

Haplotype value model

The genetic value of an individual for a genome region of interest is a sum of its two haplotype values $h_{i,1}$ and $h_{i,2}$:

$$a_i = h_{i,1} + h_{i,2}, \qquad \mathbf{a} = \mathbf{K}\mathbf{h},$$

where $\mathbf{K}$ is an $n_i \times 2n_i$ design matrix between genetic values $\mathbf{a}$ and their underlying haplotype values $\mathbf{h}$. Following Eq. 3.1 we have for $k = 1, 2$-nd haplotype:

$$\begin{aligned} h_{i,k} &= \mathbf{x}_{i,k}\alpha, \\ \mathbf{h} &= \mathbf{X}_h\alpha. \end{aligned} \tag{3.4}$$

with haplotype vector $\mathbf{x}_{i,k} \in \{0,1\}^{n_l}$ and haplotype matrix $\mathbf{X}_h \in \{0,1\}^{2n_i \times n_l}$. Unconditionally on the haplotype vectors, the haplotype values are normally distributed as:

$$\begin{aligned} h_{i,k} &\sim \mathcal{N}\left(0, \sigma_h^2\right), \\ \mathbf{h} &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\sigma_h^2\right), \end{aligned}$$

with σ_h^2 the variance of haplotype values. Conditionally on the haplotype vectors, the haplotype values are normally distributed as:

$$\begin{aligned} h_{i,k}|\mathbf{x}_{i,k} &\sim \mathcal{N}\left(0, \mathbf{x}_{i,k}\mathbf{x}_{i,k}^T\sigma_\alpha^2\right), \\ \mathbf{h}|\mathbf{X}_h &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{X}_h\mathbf{X}_h^T\sigma_\alpha^2\right). \end{aligned} \tag{3.5}$$

Haplotype value model on a local DNA tree

To expand the models Eq. 3.4-Eq. 3.5 such that it could account for rare and drifting mutation effects across generations or populations, we consider a local DNA tree of haplotypes that are separated by mutations and corresponding changes in haplotype values due to mutation effects (Selle et al., 2021; Link et al., 2023). A hypothetical example of a local DNA tree is shown in Figure 3.1.

This aim is driven by the fact that a local DNA tree is the generative model of haplotypes in line with the biology of DNA (Kingman, 1982; Griffiths and Marjoram, 1997). We would like to evaluate the usefulness of this generative model also for haplotype values that underlie genetic values. To this end, we introduce four changes. First, we switch from 0/1 encoding of reference and alternative alleles to 0/1 encoding of ancestral and derived (mutation) alleles, implying we know the ancestral allele. Second, we assume we know the local DNA tree of haplotypes, which we will use in the following. Third, we allow for multiple mutation events at the same locus, such as recurring mutations between two alleles and/or as mutations to multiple alleles, and we store these events by expanding the haplotype vector. This extended haplotype vector will have a length of $n_l + (n_m - n_l)$ for n_l loci and n_m mutation events, which have generated

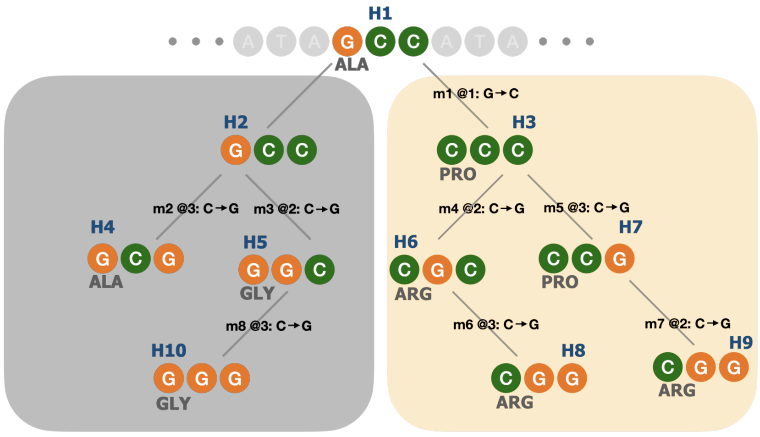

Figure 3.1: A small hypothetical local DNA tree. Haplotypes (H1-H10) span a codon in a protein-coding DNA sequence from two clades (shown as coloured boxes representing two populations) connected via the most recent common ancestor (root) haplotype (H1). Mutations between haplotypes are represented by the letter 'm' and sequential number, position in the codon, and nucleotide substitution; m1@1: G →C is the first mutation, it occurred at position 1, and nucleotide G mutated to C. The mutations change the codon sequence and corresponding amino acid as shown (ALA - alanine, PRO - proline, GLY - glycine, and ARG - arginine).

variation across the haplotypes at these loci. To avoid confusion, we call this this encoding as *mutation haplotype vector* ($\mathbf{w}_{i,k}$) and *mutation haplotypes matrix* ($\mathbf{W}$) with mutation event dosages, as compared to *allele haplotype vector/matrix* ($\mathbf{x}_{i,k}/\mathbf{X}_h$) and *allele genotype vector/matrix* ($\mathbf{x}_i/\mathbf{X}$) with allele dosages. Also, for simplicity we refer to the n-th mutation haplotype vector as $\mathbf{w}_n$ and to its value as h_n. All these three extensions require additional inferences from genomic data (ancestral alleles and local DNA tree with corresponding mutation events). Fourth, we assign effects to mutations and assume they are normally and independently distributed:

$$m_j \sim \mathcal{N}\left(0, \sigma_m^2\right),$$
$$\mathbf{m} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\sigma_m^2\right),$$

with σ_m^2 the variance of mutation effects. This assumption is the "house of cards" model (Kingman, 1977), where each mutation event has an independent effect of the other mutations (at the same or different loci). Other assumptions are possible (See Selle et al., 2021; Zeng and Cockerham, 1993). This setup gives a similar SNP-BLUP/GBLUP

setting as in Eq. 3.1-Eq. 3.5:

$$
\begin{aligned}
h_n &= \mathbf{w}_n\mathbf{m}, \\
h_n|\mathbf{w}_n &\sim \mathcal{N}\left(0, \mathbf{w}_n\mathbf{w}_n^T\sigma_m^2\right), \\
\mathbf{h} &= \mathbf{W}\mathbf{m}, \\
\mathbf{h}|\mathbf{W} &\sim \mathcal{N}\left(0, \mathbf{W}\mathbf{W}^T\sigma_m^2\right), \\
\mathbf{a} &= \mathbf{K}\mathbf{W}\mathbf{m}, \\
\mathbf{a}|\mathbf{W} &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{K}\mathbf{W}\mathbf{W}^T\mathbf{K}^T\sigma_m^2\right),
\end{aligned} \tag{3.6}
$$

but with the mutation haplotype matrix from a local DNA tree of individuals' haplotypes ($\mathbf{W}$) instead of the matrix of "flattened" allele haplotype/genotype matrix from genotyped individuals without reference to the local DNA tree ($\mathbf{X_h}, \mathbf{X}$). We refer to this new approach as TBLUP (tree BLUP) due to leveraging the local DNA tree structure.

The mutation haplotype matrix $\mathbf{W}$ in TBLUP stores information on which haplotypes have inherited which mutations. Because mutations are hierarchically inherited between haplotypes, we can leverage the local DNA tree embedded in $\mathbf{W}$ to gain further insights and facilitate more efficient computations. Let haplotype 2 be a descendant of haplotype 1. Considering their values independently and as a function of haplotype 1 value we have:

$$
\begin{aligned}
h_1 &= \mathbf{w}_1\mathbf{m}, \\
h_2 &= \mathbf{w}_2\mathbf{m}, \\
h_2 &= \mathbf{w}_2\mathbf{m} + \mathbf{w}_1\mathbf{m} - \mathbf{w}_1\mathbf{m} = \mathbf{w}_1\mathbf{m} + (\mathbf{w}_2 - \mathbf{w}_1)\,\mathbf{m}, \\
&= \mathbf{w}_1\mathbf{m} + \eta_2,
\end{aligned} \tag{3.7}
$$

where $\eta_2 = (\mathbf{w}_2 - \mathbf{w}_1)\,\mathbf{m}$ is the change in h_2 compared to h_1 due to mutations that occured between haplotypes 1 and 2. In case of a single mutation, this change is simply the mutation effect m_j, while in case of multiple mutations, this change is the sum of corresponding mutation effects. We will refer to this change in haplotype values also as the branch effect. Thence, the conditional expected value of haplotype 2 is the value of haplotype 1, while the conditional variance is proportional to the number of mutations between the two haplotypes:

$$
h_2|\mathbf{w}_1, \mathbf{w}_2 \sim \mathcal{N}\left(\mathbf{w}_1\mathbf{m}, (\mathbf{w}_2 - \mathbf{w}_1)\,(\mathbf{w}_2 - \mathbf{w}_1)^T\,\sigma_m^2\right).
$$

Leveraging this hierarchical local DNA tree structure and conditional independence of

mutation effects also allows us to construct a sparse precision matrix for the haplotype values $Var\left(\mathbf{h}|\mathbf{W}\right)^{-1} = \left(\mathbf{W}\mathbf{W}^T\sigma_m^2\right)^{-1} = \mathbf{Q}_h\sigma_m^{-2}$ directly from the local DNA tree as shown by Selle et al. (2021) and demonstrated in Supplement B.1. In this case, we don't centre the mutation dosages and hence the haplotype values are expressed as deviations from the root haplotype value. Working with the sparse precision matrix $\mathbf{Q}_h$ in TBLUP is computationally advantageous over the dense precision matrix from the allele haplotype matrix $\left(\mathbf{X}_h\mathbf{X}_h^T\right)^{-1}$ or the allele genotype matrix $\left(\mathbf{X}\mathbf{X}^T\right)^{-1}$.

We demonstrate the different approaches using allele, mutation, and local DNA tree information with the small example in Supplement B.1.

Table 3.1: Correlation between true and estimated haplotype values for the small example with different information and approaches. TBLUP uses precision matrix $\mathbf{Q}_h$ directly from the local DNA tree, avoiding the inversion and numerical errors.

Information	Approach	Correlation
Allele dosage $\mathbf{X}_h$	SNP-BLUP	0.71
	GBLUP	0.71
Mutation dosage $\mathbf{W}$	SNP-BLUP	0.79
	GBLUP	0.79
	TBLUP	0.82

The demonstration is summarised with accuracy as a correlation between the true and estimated haplotype values in Table 3.1, showing that modelling with the mutation haplotype matrix ($\mathbf{W}$) can improve the accuracy compared to modelling with the allele haplotype matrix ($\mathbf{X}_h$).

When is modelling with the mutation haplotype matrix ($\mathbf{W}$) on a local DNA tree expected to be "helpful" compared to modelling with the allele haplotype matrix ($\mathbf{X}_h$) or allele genotype matrix ($\mathbf{X}$)? If we consider only biallelic loci, that each allele arises from a single past mutation event, and that there are no untyped loci/variants in the genome region of interest, then the matrices $\mathbf{W}$, $\mathbf{X}_h$, and $\mathbf{X}$ and their corresponding models work with the equivalent information. Vice versa, allele dosages in $\mathbf{X}_h$ and $\mathbf{X}$ can also be polarised into ancestral and derived alleles, expanded with multiallelic loci and recurring mutations, which would provide equivalent information as $\mathbf{W}$. Hence the main advantage of modelling with $\mathbf{W}$ over $\mathbf{X}_h$ or $\mathbf{X}$ is that the underlying hierarchical generative model of haplotypes on the local DNA tree i) provides further insights into how different mutations affect traits and ii) facilitates faster computations with sparse matrices ($\mathbf{Q}_h$). However, critically, depending on the structure of the observed

genomic and phenotypic data, we might have limited ability to estimate the mutation effects $\mathbf{m}$ (either with $\mathbf{W}$ or with $\mathbf{X}_h$ or $\mathbf{X}$). To see this, note that a branch effect between haplotypes represents the sum of all mutation effects on the branch, including typed and untyped loci/variants. Hence, we can only estimate individual mutation effects, if we observe haplotypes that are separated by a single mutation event and if we have associated phenotypes with these haplotypes. This is similar to estimating the Mendelian sampling term of each individual in the pedigree-based model (Mrode and Pocrnic, 2023). There are two edge cases that need consideration when modelling haplotype values on a local DNA tree with TBLUP. First, the root haplotype of the local DNA tree is the most recent common ancestor of all sampled haplotypes and has by definition ancestral alleles 0 at all loci, giving a vector $\mathbf{w}_0$ of zeros. This means that prior distribution Eq. 3.6 for the root haplotype value is "degenerate" with all the probability point mass at 0, because $\mathbf{w}_0\mathbf{w}_0^T = 0$. In that case the haplotype value h_0 is alliased with the population mean μ in the model and can hence be dropped from the complete model, which then estimates other haplotype values as deviations from the root haplotype value. This is equivalent to other random walk models (Rue and Held, 2005), including those used in phylogenetics (Hadfield and Nakagawa, 2010; Harmon, 2019). Second, we can observe two haplotypes in two individuals that are in ancestor-descendant relationship, but the haplotypes are identical to each other due to a lack of mutations or due to untyped loci/variants. This last case of untyped loci/variants is relevant when we work with marker genotypes from SNP arrays or reduced-representation WGS, provided that we can estimate the local DNA tree of haplotypes from the data. In this case, the descendant haplotype is an identical copy of the ancestral haplotype and its conditional prior distribution is also "degenerate" with all the probability point mass at the value of the ancestral haplotype. This is also shown as a reduced rank of the matrix $\mathbf{W}$. We can address this situation in two ways. One way is to reassign any phenotype value from the descendant haplotype to the ancestral haplotype (or vice versa) and remove one haplotype from the local DNA tree and the model. Another way is to add a noise term ϵ with a very small variance to the descendant haplotype value (or to all haplotype values), indicating expected variation due to untyped loci/variants. This would change Eq. 3.7 to $h_1 = \mathbf{w}_1\mathbf{m} + \epsilon_1$ and $h_2 = h_1 + \epsilon_2$. The addition of the noise term also allows keeping the root haplotype in the model, because its prior variance becomes σ^2_ϵ.

3.2.2 Application

We tested the application of the above described models and souces of information using cattle mitochondrial DNA (mtDNA). The cattle mitochondrial genome is $\sim$ 16 Kbp long, non-recombining, maternally inherited, and haploid. Its small size allows fast analysis and the absence of recombination means that one local DNA tree captures the genealogical history of the entire mtDNA and hence serves as a good example for demonstration. Also, being haploid, it is de-facto phased, reducing the need for phasing and limiting errors in the inference of the local DNA tree. We describe in the following two different mtDNA datasets, how we inferred ancestral alleles and the local DNA tree, simulation and empirical data analysis, finally followed by evaluation of model fits from the SNP-BLUP approach on an allele haplotype matrix and the TBLUP approach on a local DNA tree. All the computations were performed on the University of Edinburgh High-Performance Computing cluster (Eddie). Scripts used in this study are available at GitHub repository `https://github.com/HighlanderLab/gmafrafortuna_quangen_seq`, utilising Python for infering and working with the mtDNA tree using *tsinfer* (Kelleher et al., 2019; Wohns et al., 2022) and *tskit* (Ralph et al., 2020; tskit developers, 2025) packages, and utilising R for data manipulation, simulation, and analysis using *Matrix* (Bates et al., 2025), *pedigreem* (Vazquez et al., 2010), *lubridate* (Grolemund and H, 2011), and *tidyverse* (Wickham et al., 2019) packages. All models were fitted using the R package *INLA* (Håvard et al., 2009; Rue et al., 2017).

Datasets

The first dataset was from the public-available 1000 Bull Genomes Project Run8 (Hayes and Daetwyler, 2019) (available at `https://www.ebi.ac.uk/ena/browser/view/PRJEB42783`). A full description of the dataset of relevance for this study is given by (Dorji et al., 2022). The dataset contained mtDNA whole-genome-sequences for 1,842 animals of *Bos taurus* and *Bos indicus* origin. The mtDNA had 4,540 sites, of which 3,518 were biallelic SNPs. No phenotypes associated with these genotypes are available in this dataset. We used this dataset for inferring ancestral alleles and the local mtDNA tree, and for evaluating the models via simulations. Throughout the text, we refer to this dataset as *1KB*.

The second dataset was from (Brajkovic et al., 2025). It contained data on 359 SNPs for 96 unique mtDNA sequences of Croatian Holstein cattle, as well as the pedigree and phenotypic data for milk production traits. Due to the maternal inheritance and

absence of recombination, the authors linked the 96 unique mtDNA sequences through a pedigree of 6,336 animals to 3,040 females and their 7,576 lactation records. We used this dataset for evaluating the models via empirical data analysis. Throughout the text, we refer to this dataset as *CRO*.

Ancestral allele inference

To infer ancestral alleles for the cattle mtDNA, we obtained complete mtDNA sequences for seven reference species from the GenBank repository (NCBI): *Bos taurus* (V00654), *Bos indicus* (NC_005971), *Bos primigenious* (NC_013996), *Bos grunniens* (yak, KR011113), *Bison bison* (bison, NC_014044), *Ovis aries* (sheep, KR868678), and *Camelus bactrianus ferus* (camel, EF212038). We aligned the sequences with *Mafft* (Katoh et al., 2002) using default settings. A poorly aligned portion of the D-loop region was discarded, giving an alignment with 16,882 bp. For alignment refinement, further exclusion of sites was performed using Gblocks 0.91b (Castresana, 2000). Parameters were set for the least stringent selection. The final alignment with 16,327 bp was retained for further analyses.

Cattle ancestral alleles were inferred using the software *est-sfs* (Keightley and Jackson, 2018). Bison, yak, and camel served as the chosen outgroups due to their evolutionary relationships with cattle (the focal group). Yak and bison are closely related outgroup ruminant species, while camel, an artiodactyl, represents a more distant outgroup (Zhang et al., 2020). Allele counts for cattle were extracted from the *1KB* dataset, while for each outgroup were extracted from the alignment. For each site, we obtained the probability of the major allele in the focal species being the ancestral allele. The major allele was considered ancestral if this probability was greater than 0.9, while the probability less than 0.1 indicated the minor allele as ancestral. For sites with the probability between 0.1 and 0.9, we assumed the ancestral allele to be the major allele in the outgroups. We were able to determine the ancestral state for 3,257 polymorphic sites in the mtDNA.

mtDNA tree inference

We inferred the mtDNA tree using the Python package *tsinfer* (Kelleher et al., 2019; Wohns et al., 2022). We set the *recombination_rate* parameter to $1e-20$ (effectively zero) and *mismatch_ratio* parameter to $1e18$. The mismatch ratio argument controls whether a conflict is resolved via recurrent mutation (high value, > 1) or recombination (low value, < 1). Setting the mismatch ratio to $1e18$ forced the inference to resolve all conflicts through mutations instead of recombinations, as expected for the non-

recombining DNA, resulting in a single tree and recurrent mutations, as opposed to the *tsinfer* default infinite-sites model. *tsinfer* can generate mutations above the root, especially after tree simplification that, among others, removes unary nodes (Wong et al., 2024). For the purpose of model on a local DNA tree, we created a new node with all ancestral alleles and set it as an ancestor to the inferred root haplotype.

During tree sequence inference for the *1KB* dataset, polymorphic sites were removed. Furthermore, we observed some branches with an excess of mutations (in hundreds). This could indicate, among other things, low-quality sequence data. To address this, we removed samples whose incoming branch had more than 100 mutations. This led to the removal of 214 samples and additional sites. The final tree had 1,684 nodes (1669 sampled mtDNA sequences and 13 ancestors), 1,682 edges (branches), 1,029 sites, and 9,598 mutations. For the *CRO* dataset, the final mtDNA tree had 22 nodes (96 samples and 126 ancestors), 124 edges (branches), 357 sites, and 411 mutations.

Simulation

We used the inferred mtDNA tree from the *1KB* dataset for a simulation study. Using R, we simulated haplotype values across the 1,683 nodes in the *1KB* mtDNA tree. For sample nodes (animals) we also simulated phenotype values. We evaluated 36 scenarios with different assumptions about mutation effects and phenotyping strategies. Each scenario was replicated 100 times to assess variance in results. The haplotype values represent the cumulative effect of all mutations along the path from the root to a haplotype.

To simulate the effect of mutations we considered two primary scenarios: all mutations were causal, labeled **ALL**, or only 10% of the mutations were causal, labelled **FEW**. We then considered three sub-scenarios to determine the effect of the mutations: (i) all mutations had unique effects (**DIFF**), (ii) all mutations of the same type (such as, A $\rightarrow$ T) shared the same effect (**SAME**) and (iii) symmetric mutation types (such as, A $\rightarrow$ T and A $\leftarrow$ T) had the same absolute effect but opposite signs (**SYM**). In all cases, mutation effects were drawn from a normal distribution $\mathbf{m} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_m^2)$ with $\sigma_m^2 = 0.5$. This led to six distinct mutation scenarios: ALL-DIFF, ALL-SAME, ALL-SYM, FEW-DIFF, FEW-SAME, and FEW-SYM. Having simulated mutation effects, we calculated branch effects as the sum of mutation effects along the branch between two haplotypes. Haplotype values (h_n) were obtained by setting the root haplotype to zero $h_0 = 0$ and recursively accumulating branch effects $h_n = h_{p(n)} + \eta_n$, where $a(n)$ is the immediate ancestor haplotype of haplotype n and η_n the branch effect.

To assess the impact of phenotyping strategy, we considered three scenarios: (i) all 1,669 animals were phenotyped, (ii) only a quarter of the animals (417) were phenotyped, and (iii) only 70 animals were phenotyped. Phenotype values were generated by adding up a fixed population mean (10), the haplotype values, and a random residual from $\mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_e^2)$ with $\sigma_e^2 = 1$.

The simulated data was first fitted in the SNP-BLUP approach on column-centred *allele haplotype matrix* ($\mathbf{X}_h$) Eq. 3.5 to obtain estimates of marker effects, and corresponding estimates of haplotype values (in results we refer to this approach as SNP-BLUP for simplicity). We then fitted also the TBLUP approach on *mtDNA tree* with sparse precision matrix ($\mathbf{Q}_h$) Eq. 3.6 to obtain estimates of haplotype (node) values, and corresponding estimates of branch effects, mutation effect, and marker effects. Refer to Supplement B.1 for demonstration of these approaches.

We evaluated the performance of the approaches by comparing the accuracy calculated as the correlation between true and estimated effects or values. We evaluated the accuracy of haplotype values (separately for sampled haplotypes and for sampled and ancestral haplotypes), branch effects, mutation effects, and marker effects. Haplotype values for ancestors, branch effects, and mutation effects were available only with the TBLUP approach. Node effects included the effect of all haplotypes (sampled and ancestral). Branch effects were estimated by multiplying the TBLUP estimates of haplotype values by $\mathbf{T}^{-1}$ (see Eq. B.1.5 in Supplement B.1), where $\mathbf{T}$ is the incidence matrix connecting haplotypes (nodes) to branches. Since we can not estimate the effect of individual mutations on a branch, we naïvely estimated their effect by dividing an estimated branch effect by the number of mutations on the branch. In the opposite direction, we calculated marker effects by summing the mutation effects at a locus - we did this both for the true and estimated mutation effects. We also evaluated the elapsed time for each model in both the simulation and the empirical data analysis.

Empirical data analysis

We refitted the base model from (Brajkovic et al., 2025) by using the SNP-BLUP approach on column-centred *allele haplotype matrix* Eq. 3.5 and the TBLUP approach on *mtDNA tree* Eq. 3.6 on the *CRO* dataset and focused on the estimation of variance components for various quantities. Both approaches used the repeatability model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_c\mathbf{c} + \mathbf{Z}_i(\mathbf{a} + \mathbf{x} + \mathbf{pe} + \mathbf{m}) + \mathbf{e}, \tag{3.8}$$

where $\mathbf{y}$ is the vector of milk yield records, $\mathbf{X}$ is the design matrix for the fixed effect - age at first calving with $\mathbf{b}$ the corresponding regression coefficient, $\mathbf{Z}_c$ the design matrix for the random effect of contemporary group, defined as hear-year-season effects modelled as $\mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_c^2)$, $\mathbf{Z}_i$ the design matrix for individual animal effects, $\mathbf{a}$ the vector of genetic values for the autosomal DNA modelled as $\mathcal{N}(\mathbf{0}, \mathbf{A}_a\sigma_a^2)$ using autosomal pedigree relationship matrix $\mathbf{A}_a$, $\mathbf{x}$ the vector of genetic values for the X chromosome modelled as $\mathcal{N}(\mathbf{0}, \mathbf{A}_x\sigma_x^2)$ using X chromosome pedigree relationship matrix $\mathbf{A}_x$, $\mathbf{m}$ the vector of genetic values for the mtDNA modelled with SNP-BLUP or TBLUP approaches, $\mathbf{pe}$ is the vector of permanent environment effects modelled as $\mathbf{N}(\mathbf{0}, \mathbf{I}\sigma_{pe}^2)$, and $\mathbf{e}$ is the vector of residuals modelled as $\mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_e^2)$.

From the fitted models Eq. 3.8 we obtained estimates of ("standard") variance components listed above for SNP-BLUP and TBLUP approaches: σ_c^2, σ_a^2, σ_x^2, σ_{pe}^2, and σ_e^2. With SNP-BLUP we also obtained a direct estimate of the variance of marker effects σ_α^2. With TBLUP we also obtained a direct estimate of the variance of mutation effects σ_m^2. Using the `inla.posterior.sample` function from the *INLA* package, we obtained 1,000 Monte Carlo samples from the posterior distribution of marker effects α for the SNP-BLUP approach and mutation effects $\mathbf{m}$ for the TBLUP approach. With these samples, we also obtained samples of haplotype values, node values, and branch effects in line with the theory and simulation subsections, and calculated their variance for each sample, giving us posterior estimates of corresponding ("derived") variance components: σ_h^2 is the variance of mtDNA haplotype values (sample haplotypes), σ_n^2 is the variance of mtDNA haplotype/node values (sample and ancestral haplotypes), and σ_b^2 is the variance of branch effects. For each variance we reported the posterior mean and the posterior standard deviation.

3.3 Results

We compared the performance of the TBLUP approach on a local DNA tree with corresponding mutation events against the SNP-BLUP approach on an allele haplotypes matrix in a simulation and with an empirical data analysis. In the following, we present the results for each separately, showing that TBLUP tends to outperform SNP-BLUP in estimation accuracy as the number of phenotyped animals increases, reduces the computational demand, and allows the determination more information from the WGS data.

3.3.1 Simulation

Estimation of haplotype values

Overall, both TBLUP and SNP-BLUP approaches showed comparable performance in the accuracy of estimated haplotype values Figure 3.2, with average accuracy across replicates and scenarios ranging between 0.70 and 0.99.

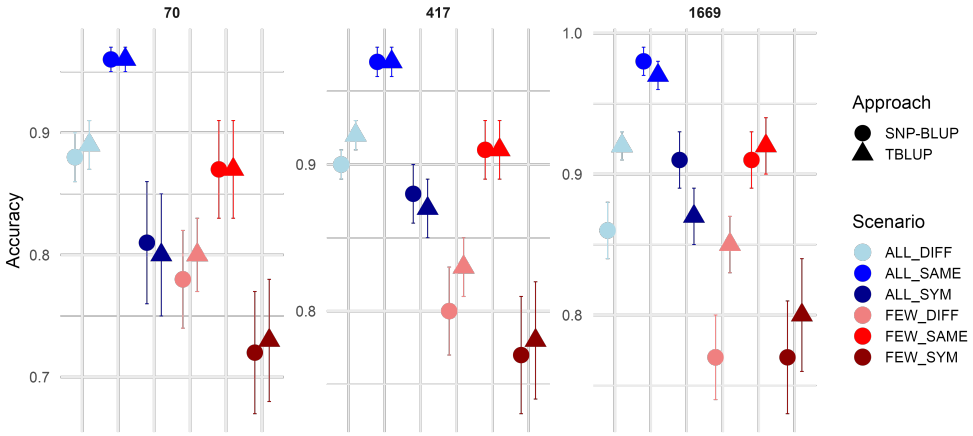

Figure 3.2: Average (± standard error) accuracy of estimated haplotype effects with TBLUP and SNP-BLUP approaches in simulation with different scenarios. The x-axis lists different mutation scenarios, while accuracy is on the y-axis. Symbol colors and shapes distinguish mutation scenarios and the approaches: blue for all mutations having an effect (ALL), red for few mutations having an effect (FEW), lighter colours for all mutations having different effects (DIFF), medium for the same type of mutations have the same effect (SAME), and darkest for added symmetry to SAME for reverse mutations (SYM). Circles represent SNP-BLUP and triangles represent TBLUP results. The three panels represent the number of phenotypes used in the analysis.

Significant differences between the approaches were observed only in the scenario with 1,669 phenotypes and for the mutation scenarios ALL-DIFF, ALL-SYM, and FEW-DIFF. Specifically, TBLUP outperformed SNP-BLUP by 6.52% in the ALL-DIFF scenario and by 9.41% in the FEW-DIFF scenario. In the ALL-SYM scenario, SNP-BLUP outperformed TBLUP by 4.6%. Nonetheless, we observed a general tendency of higher accuracies with TBLUP, particularly with more phenotyped animals.

Phenotyping scenarios Increasing the number of phenotyped animals improved the accuracy for TBLUP and SNP-BLUP. The largest improvement was observed in the ALL-SYM scenario, where accuracy increased by 0.07 (8.75%) for TBLUP and 0.1 (12.35%) for SNP-BLUP. Notably, when only 70 animals were phenotyped, standard

errors were expectedly larger than with more phenotyped animals.

Mutation scenarios The ALL-SAME scenario had the highest accuracy across all phenotyping scenarios. Specifically, TBLUP and SNP-BLUP had both accuracy 0.99 ± 0.01 with 70 animals phenotyped and 0.97 ± 0.01 with 417 animals phenotyped. With 1,669 animals phenotyped, the TBLUP had accuracy of 0.97 ± 0.01 and the SNP-BLUP had accuracy of 0.98 ± 0.01, without significant difference. In contrast, the FEW-SYM scenario had the lowest accuracy across all phenotyping scenarios. Accuracy was 0.73 ± 0.05 for TBLUP and 0.72 ± 0.05 for SNP-BLUP with 70 animals phenotyped, 0.78 ± 0.04 and 0.77 ± 0.04 with 417 animals, and 0.80 ± 0.04 and 0.77 ± 0.04 with 1,669 animals. None of the differences was significant. Overall, within each mutation scenario (ALL or FEW), scenarios with the same effects for a mutation type (ALL-SAME and FEW-SAME) consistently performed better, while scenarios with symmetric mutation effects (ALL-SYM and FEW-SYM) proved more challenging.

Elapsed time

We evaluated elapsed time to run TBLUP and SNP-BLUP across the scenarios (Figure 3.3.

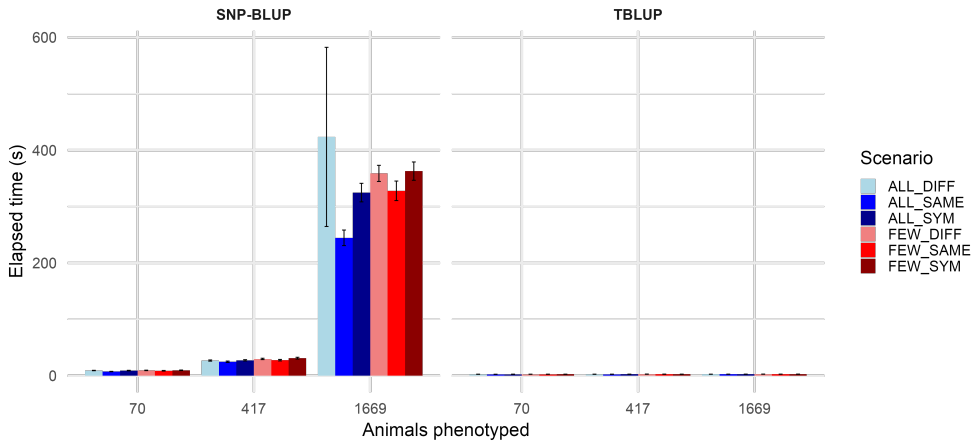

Figure 3.3: Average ($\pm$ standard error) elapsed time in seconds for TBLUP and SNP-BLUP approaches across scenarios. The x-axis lists different number of phentyped animals, while elapsed time is on the y-axis. Colors distinguish mutation scenarios: blue for all mutations having an effect (ALL), red for few mutations having an effect (FEW), lighter colours for all mutations having different effects (DIFF), medium for the same type of mutations have the same effect (SAME), and darkest for added symmetry to SAME for reverse mutations (SYM).

Overall, the TBLUP was fitted faster than SNP-BLUP with *INLA*. TBLUP elapsed

time was less than 3 seconds for all scenarios, while SNP-BLUP elapsed time ranged from 7.5 seconds with 70 phenotyped animals to more than 400 seconds with 1,669 phenotyped animals. There was some variation across mutation scenarios for SNP-BLUP with 1,669 phenotyped animals; with large average elapsed time and its standard error, as well as indication of larger elapsed time when few mutations were causal.

Estimation of node values and branch, mutation, and marker effects

Finally, in the simulation we also assessed the accuracy of estimating node values (haplotype values for sampled and ancestral haplotypes) and decomposing these into branch, mutation, and marker effects with TBLUP (Figure 3.4).

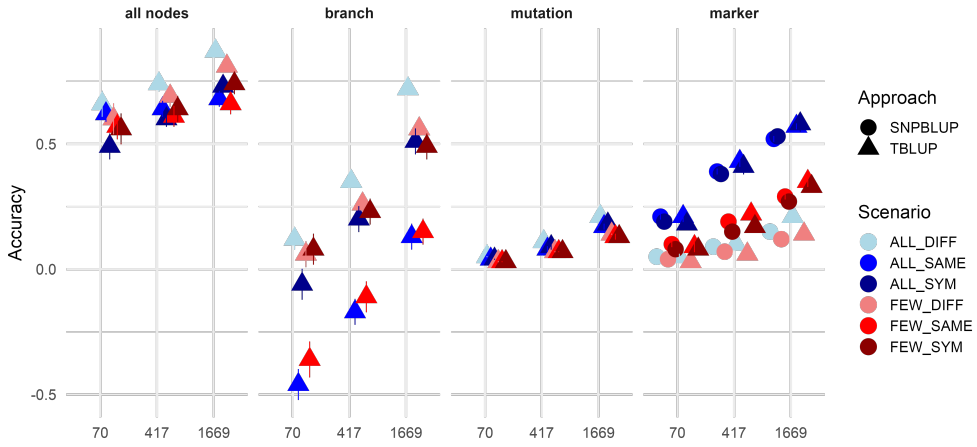

Figure 3.4: Average (± standard error) accuracy of estimated node, branch, mutation nd marker effects. The x-axis details different sample sizes; correlation values are on the y-axis. Color coding and shapes are used to distinguih mutation scenarios and models: Blue for all mutations having an effect (ALL), red for few mutations having an effect (FEW), lighter colours for DIFF scenarios, medium for SAME and darkest for SYM. Circles represent TBLUP results, and triangles represent SNP-BLUP results.

Accuracy for these estimates generaly reduced as we decomposed the node values into branch effects and then into mutation effects, while marker effects had intermediate accuracy between the branch and mutation effects. Node, branch, and mutation estimates were available only with the TBLUP approach, so comparison with SNP-BLUP was only possible for the marker effects.

Node effects The accuracy of node value estimates ranged from 0.49 ± 0.05 (for the FEW-SYM scenario with 70 animals) to 0.87 ± 0.02 (for the ALL-DIFF scenario with 1,669 animals). As in Figure 3.2, accuracy generally increased with more phenotyped

animals, however the rate of increase differed between mutation scenarios. Specifically, when we increased the number of phenotyped animals from 70 to 1,669 in the ALL-DIFF scenario, accuracy increased by 31.8%. In the ALL-SYM scenario, accuracy increased by 49.0%. In the FEW-DIFF scenario, accuracy increased by 35.0% In the FEW-SYM scenario accuracy increased by 32.1%.

Branch effects The accuracy of estimated branch effects was the most sensitive to phenotyping scenarios and benefited the most from increasing the number of phenotyped animals across all mutation scenarios. The greatest improvement was observed when increaseing the number of phenotyped animals from 70 (0.12 ± 0.04) to 1,669 phenotypes (0.72 ± 0.02) in the ALL-DIFF scenario. The ALL-DIFF scenario also had the highest accuracy for branch effects, followed by FEW-DIFF, FEW-SYM, and ALL-SYM scenarios, while the ALL-SAME and FEW-SAME had the lowest. The later two even had negative accuracies with small numbers of phenotyped animals.

Mutation effects The accuracy of estimated mutation effects was the least impacted by changes in phenotyping scenarios and generally the most challenging effect to increase its accuracy of estimation. This is expected since we are not able to discriminate the effect of individual mutations on a branch - assuming the naïve estimate of dividing the branch effect by the number of mutations. The highest accuracy was obtained for the ALL-DIFF scenario with 1,669 phenotypes (0.21 ± 0), which is in line with the high accuracy of branch effects in this scenario.

Marker effects In the addition of comparing the accuracy of estimated haplotype values in Figure 3.2, we compared TBLUP and SNP-BLUP based the accuracy of estimated marker effects. Increasing the number of phenotyped animals improved the accuracy of marker effects for both TBLUP and SNP-BLUP; with the highest accuracy of 0.58 ± 0.03 for ALL-SYM scenario with 1,669 phenotypes. The accuracy increased the most for the ALL-SAME and ALL-SYM scenarios, followed by FEW-SAME and FEW-SYM, with the lowest increases for ALL-DIFF and FEW-DIFF. There was no significant difference between the two approaches, with slight tendency of TBLUP to have higher accuracy estimates.

3.3.2 Empirical analysis

We compared the TBLUP and SNP-BLUP approaches for the estimation of variance components using the real *CRO* dataset (Table 3.2). Both approaches estimated similar variance components with some minor differences. SNP-BLUP approach estimated the

Table 3.2: Estimated variance components from the *CRO dataset*. The "standard" components are: σ_c^2 is the contemporary group (herd-year-season) variance, σ_a^2 is the genetic variance for autosomal DNA, σ_x^2 is the genetic variance for X chromosome, σ_{pe}^2 is the permanent environment variance, σ_e^2 is the residual variance. The SNP-BLUP and TBLUP specific effect components are: σ_α^2 is variance or marker effects and σ_m^2 is the variance of mutation effects. The "derived" components are: σ_h^2 is the variance of mtDNA haplotype values (sample haplotypes), σ_n^2 is the variance of mtDNA haplotype/node values (sample and ancestral haplotypes), and σ_b^2 is the variance of branch effects.

Variance components	SNP-BLUP	TBLUP
	"Standard"	
σ_c^2	0.096 ± 0.008	0.095 ± 0.009
σ_a^2	0.324 ± 0.010	0.377 ± 0.010
σ_x^2	0.000 ± 0.000	0.000 ± 0.000
σ_{pe}^2	0.061 ± 0.018	0.057 ± 0.017
σ_e^2	0.376 ± 0.028	0.330 ± 0.027
	SNP-BLUP and TBLUP specific	
σ_α^2	0.002 ± 0.005	NA
σ_m^2	NA	0.004 ± 0.001
	"Derived"	
σ_h^2	0.074 ± 0.013	0.065 ± 0.012
σ_n^2	NA	0.065 ± 0.011
σ_b^2	NA	0.063 ± 0.012

genetic variance for autosomal DNA σ_a^2 at 0.324 ± 0.010, while TBLUP estimated it at a larger 0.377 ± 0.010. Correspondingly, SNP-BLUP approach estimated the residual variance σ_e^2 at 0.376 ± 0.028, while TBLUP estimated it at smaller 0.330 ± 0.027. Variance of marker effects σ_α^2 was estimated at 0.002 ± 0.005 with SNP-BLUP, while the variance of mutation effects σ_m^2 was estimated to be larger at 0.004 ± 0.001 with TBLUP. The variance of haplotype values σ_h^2 was estimated at 0.074 ± 0.013 with SNP-BLUP, and at slightly smaller 0.065 ± 0.012 with TBLUP, which was also the estimate of the variance of all haplotype (node) values σ_n^2 with TBLUP. Variance of branch effects σ_b^2 was estimated at 0.063 ± 0.012 with TBLUP. At this scale of the data (359 SNPs for 3,040 animals) and this model complexity, the elapsed time to fit the TBLUP was on average 74.7 seconds, while for SNP-BLUP it was 77.9 seconds.

3.4 Discussion

Our results show that, in a non-recombining DNA region, the presented TBLUP approach tends to outperform SNP-BLUP in terms of computational effort and informativeness of results. TBLUP also has the potential to generate higher estimation accu-

racy, but predicting mutation effects is a challenge. Computational efficiency comes from the generative model of haplotype values on a local DNA tree and associated conditional distributions for ancestor–descendant haplotype values with a sparse precision matrix. Using a local DNA tree with associated branches and mutations, the TBLUP approach also provides additional information about branch effects and mutation effects, shedding light on the estimability of gene flow and causal mutations. Moreover, the new approach opens up new opportunities for understanding how these effects drive variation in haplotype and phenotype values. Generally, estimation accuracy tended to be higher with the TBLUP approach, however, this benefit depends on the additional information of mutation dosages compared to allele dosages. Importantly, accuracy of predictions, especially for mutation effects, is limited by the data structure with respect to phenotype distribution across the sampled and ancestral haplotypes in the local DNA tree. In the following, we discuss three critical points that arise from this work: (i) novel contributions and the advantages of the TBLUP approach; (ii) why are mutation effects difficult to estimate; (iii) opportunities and future directions with local DNA trees and ARGs.

Novel contributions and the advantages of TBLUP

The novelty of the TBLUP approach lies in its ability to leverage the topology of a local DNA tree and the detailed ancestral information encoded within it. This provides a rich framework for modelling genetic variation in a non-recombining genomic region, leading to two main advantages.

First, the TBLUP approach offers computational efficiency. By incorporating the structure of a local DNA tree into the BLUP framework, we replace the dense genomic relationship matrix between individuals/haplotypes with a sparse graph representation of DNA inheritance between generations. As genomic datasets grow, computing, storing and inverting genomic relationship matrices becomes increasingly limiting, although various approaches have been proposed to overcome this problem (Calus et al., 2015; Misztal et al., 2014; Misztal, 2015; Tsuruta et al., 2021; Nilforooshan, 2024). The TBLUP approach bypasses the need to setup and invert the relationship matrix, by directly constructing the sparse inverse (precision), $\mathbf{Q}_h$, analogous to the pedigree-based model (Henderson, 1976). We have previously worked on this problem (see Selle et al., 2021) as have others in the context of phylogeny (see Lynch, 1991; Hadfield and Nakagawa, 2010). While the phylogenetic models work with the whole genome and are used to model variation between (sub-)populations, our work is for a region of DNA and models variation between individuals (within or across populations). Here, we have

now evaluated the benefits of the generative modelling approach more extensively, showing the potential and possibility of estimating context-dependent effects (see also Link et al., 2023). This work is part of a greater effort to leverage ARGs in quantitative genetics (Ralph et al., 2020; Zhang et al., 2023; Zhu et al., 2024; Rebollo et al., 2025; Lehmann et al., 2025; Lee et al., 2025), which leverages the efficiency of computing with large-scale genomic datasets stored in the tree sequence data format (Kelleher et al., 2019; Ralph et al., 2020; Wong et al., 2024). Second, the TBLUP approach is based on a richer model. It provides estimates for the effect of sampled haplotypes, as well as for the ancestral haplotypes, branches between all the haplotypes, and associated mutations. These estimates provide a more detailed view of the genetic architecture underlying trait variation and how it evolved over time. This makes our approach similar to other quantitative genomic models that aim to capture recent mutations in addition to past mutations (Casellas and Varona, 2011; Casellas et al., 2013). Also, the branch effect variance is proportional to the mutational variance (Lynch and Hill, 1986; Lee et al., 2025).

Why are mutation effects difficult to estimate

One of the hypotheses of this study was that the TBLUP approach could accurately estimate mutation effects. Our results show that this remains challenging. Two main factors justifying this limitation are: (i) small size of our dataset (although we focused on just one local DNA tree) and (ii) complete linkage inherent to a non-recombining genome region. Under these conditions, mutations share similar inheritance patterns, which diversify over generations, but slowly. This condition makes it difficult to distinguish mutation effects, particularly if a group of mutations is observed on the same branch of a local DNA tree. In our study, we simply approximated mutation effects by dividing the estimated branch effect by the number of mutations observed on it. Furthermore, we had more mutations than phenotypes (9,598 vs 1,669), so our statistical power was limited. These results echo the warnings about the interpretation of past genomic trends (Novembre and Barton, 2018). They also highlight a fundamental constraint of working with non-recombining genome regions and motivate future work with recombining genome regions. Since recombinations reduce the linkage-disequilibrium between sites and therefore mutations, we expect to be able to better disentangle the effects of mutations in recombining regions.

Opportunities and future directions with local DNA trees and ARGs

Despite the advantages and novelty of this study, several challenges and opportunities remain. The TBLUP approach requires an accurate reconstruction of a local DNA tree and ultimately complete ARGs to work across the whole genome. Errors in this reconstruction may propagate into the statistical estimation of the effects. The field of ARG inference is advancing rapidly, with many scalable algorithms now available (Gunnarsson et al., 2024; Deng et al., 2024; Wohns et al., 2022; Kelleher et al., 2019; Speidel et al., 2019; Rasmussen et al., 2014). As these methods improve, so will the applicability of TBLUP and similar approaches. Although the sparse precision matrix provides computational benefits, the magnitude of these gains depends on software implementation. Here we have used R-INLA (Håvard et al., 2009; Rue et al., 2017) to fit the models because it is optimised for work with sparse matrices, but other software optimised for dense matrices may perform better. Nevertheless, setting up a sparse precision matrix remains an advantage compared to calculating and inverting a dense genomic relationship matrix. For TBLUP to outperform SNP-BLUP in terms of accuracy, ancestral nodes within the local DNA tree must be phenotyped. This represents an important practical limitation. Since mutations are rare, the chances of observing phenotyped individuals that directly inherit causal mutations are very small. The accuracy of estimated mutation effects will improve by working with the full ARG that encompasses all local DNA trees, and hence recombinations between these. There is an active topic of research (Zhang et al., 2023; Zhu et al., 2024; Rebollo et al., 2025; Lehmann et al., 2025; Lee et al., 2025). While we focused on a non-recombining region of genome, this study demonstrates the potential and benefits of leveraging the generative model for haplotype effects in quantitative genetics.

3.5 Chapter conclusion

In summary, the TBLUP approach of leveraging information from a local DNA with associated mutation events provides a promising extension of the current SNP-BLUP and GBLUP approaches in terms of computational speed, information content (estimation of haplotype values, branch effects, and mutation effects), and estimation accuracy. As such, it estimates the effects of mutations in the context of a local DNA tree enabling their estimation at the point of occurrence, with a possibility to capture non-additive effects if the same type of mutation occurs in different parts of the tree. While we studied the TBLUP approach in a non-recombining region of the genome, its potential

is even greater when applied across recombining regions of the genome, which is the focus of other research (Rebollo et al., 2025; Lehmann et al., 2025; Lee et al., 2025; Christ et al., 2025). Further extension of the approach and its evaluation in these more challenging settings will be crucial to fully understand its capabilities and limitations.

Funding

GMF and GG acknowledge funding from BBSRC DTP (EASTBio) CASE PhD studentship with Genus, BBSRC Institute Strategic Programme funding to The Roslin Institue (BBS/E/D/30002275, BBS/E/RL/230001A), BBSRC grants BB/T014067/1 and BB/M009254/1, and The University of Edinburgh. JO and AM acknowledge the core financing of the Slovenian Research and Innovation Agency (programme P4-0133 "Sustainable agriculture").

Availability of data and materials

All scripts for ARG inference, simulations and real data analysis are available at GitHub repository `https://github.com/HighlanderLab/gmafrafortuna_quangen_seq`.

4 The importance of trait stability in crossbreeding: uncovering the impacts of genotype-by-environment interaction

This chapter presents the manuscript *The importance of trait stability in crossbreeding: uncovering the impacts of genotype-by-environment interaction*, which introduces a new framework for investigating genotype-by-environment (G×E) interaction to uncover useful genetic variation within tropical crossbreeding programmes.

This manuscript focuses on the challenges of crossbreeding for dairy production in tropical regions, with extensive genetic diversity, variable production systems, and changing environments making the response of crossbreeding highly variable. The study looks at the role played by G×E interaction in the instability of crossbred response, highlighting the importance of trait stability in crossbreeding programmes. A new genetic simulation with ancestry-specific effect is develop and a new analytical framework is proposed which has been adapted and expanded from plant breeding methods, for exploring G×E interaction in animal breeding. The genetic simulation includes deep split between indicine and taurine cattle populations, with corresponding additive and dominance effects that can be ancestry-specific, driving opportunity for improving crossbreeding. The G×E framework uses multiplicative models to decompose the genotype response into a series of multiplicative terms which capture meaningful components of genetic and G×E variance. It then redefines G×E relative to the breeding objectve to capture meaningful components of the correlated response to selection. This identifies genotypes and environments of particular importance to crossbreeding objectives. Finally, it provides a set of tools to visually analyse and interpret the results for improving breeder's selection decisions in tropical systems.

The framework was applied to a simulated tropical breeding programme and efficiently outlines genetic and G×E patterns which, if strategically used, can help breeders (i) make better use of underutilised genetic diversity, (ii) design climate-resilient breeding programs, and (iii) improve genetic gain and trait stability.

The manuscript is a joint work between Gabriela Mafra Fortuna[1], Gregor Gorjanc[1] and Daniel Tolhurst[1].

[1] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Edinburgh EH25 9RG, Midlothian, United Kingdom.

Manuscript Status: In preparation.

Core ideas

- *Bos taurus* and *Bos indicus* cattle breeds originate from a deep divergence from a common ancestor, and distinctive evolutionary events and breeding practices, leading to the development of unique genetic archtecture. The genetic differences between the two subpopulations pose challanges and opportunities for tropical dairy breeders.
- Crossbreeding for tropical dairy production faces unique biological and environmental complexities, operating under variable environmental conditions and structural constraints. Conventional breeding practices designed for controlled, intensive systems, fail to address these complexities and this limits the success of crossbreeding strategies.
- Genotype-by-environment (G×E) interaction is commonly ignored or over-simplified, despite playing a critical role in tropical dairy production. This leads to unstable performance and underutilised genetic diversity.
- Strategic use of genetic diversity requires tools and resources that simplify the genomic complexities, allowing breeders to make informed decisions that align breeding goals with the realities of tropical agriculture.

Abstract

Intensive dairy farming has evolved to maximise milk production under optimised conditions. However, in tropical systems, such optimal conditions are often unavailable due to challenges like heat stress, resource variability, and climate instability.

These factors limit response to selection, delaying genetic progress. Crossbreeding local environmentally-adapted indicine breeds and exotic, high-yielding taurine breeds is a common breeding strategy designed to overcome the challenges inherent to tropical systems. The goal is to produce crossbred animals that, on average, outperform the mean of the parents. In practice, response to crossbreeding is highly inconsistent, which can be attributed to the great genetic distance between indicine and taurine breeds, and primary focus on mean performance, while genotype-by-environment interaction (G×E) is often overlooked. Despite evidence of the role of G×E on performance, few solutions address this source of variation, reflecting a gap in understanding how to optimise crossbreeding effectively. Given the increasing unpredictability of tropical climates driven by climate change, understanding how G×E affects dairy production in these areas is crucial. Our research aims to uncover the impacts of G×E on tropical dairy production and demonstrate the advantages of pursuing trait stability- the consistent performance of a genotype across environments- for the success of crossbreeding strategies. Using stochastic simulations, we compare three breeding programmes: a purebred intensive programme (e.g., Holstein), a purebred tropical programme (e.g., locally adapted indicine breeds), and a crossbreeding programme. The intensive programme is tested in a single favourable environment, while the tropical purebred and crossbred programmes are evaluated in five environments representing varying tropical conditions. Environments are correlated, with stronger correlations between neighboring environments. We evaluate response to selection for the six environments, examining the impacts of crossbreeding and G×E interaction. Our results demonstrate the reality of crossbreeding expectations in light of G×E interaction, the increase instability with loss in performance across environments reflected by the increase in genetic variance over time. This work offers a framework based on a multiplicative model for G×E interaction for understanding the impacts of environmental variability on dairy production and provides a novel tool for designing breeding strategies better suited to these conditions, supporting the sustainable future of tropical dairy farming.

4.1 Introduction

Crossbreeding high performing *Bos taurus* breeds with environmentally adapted *Bos indicus* is a widely used strategy in tropical dairy systems, aiming to increase production and accelerate genetic progress by producing crossbred animals which, on average, are expected to outperform the mean of their parents (McDowell, 1985). In practice, the genetic response to crossbreeding is highly inconsistent, with crossbred performance varying depending on breed compositions and environmental conditions

(Marshall et al., 2019; Bunning et al., 2018; Leroy et al., 2016; Madalena et al., 1990). This inconsistency can be attributed to two main factors: (i) the genetic divergence between *Bos taurus* and *Bos indicus*, which can lead to unfavourable allelic combinations complicating breeding strategies, and (ii) genotype-by-environment (G×E) interaction, which affects the stability of traits across different environments.

G×E interaction plays a critical role in the success of crossbreeding, as genotypes exhibit a change in their response to a change in their environment. G×E interactions can manifest as either non-crossover or crossover interaction, reflecting changes in the magnitude of genotype response between environments or re-rankings of genotypes (see Fig. 1.1; Gail and Simon, 1985). Specifically, non-crossover G×E interaction reflects changes in the magnitude of genotypic response without re-rankings between environments, resulting in a correlation of +1 between environments. In contrast, crossover G×E interaction reflects re-rankings of genotypes, resulting in a correlation of less than +1 between environments and posing important challenges to breeders as genotype performance is environment-specific (de Leon et al., 2016). In practice, both non-crossover and crossover interaction is present, which further complicates breeders decisions. The framework developed in this chapter leverages non-crossover and crossover interaction for selection, with particular focus on applications to tropical crossbreeding systems.

In tropical crossbreeding systems, the potential of heterosis to increase overall performance is often compromised by ignoring G×E interaction. One important consequence of this is a reduction in trait stability, i.e., the ability of a genotype to maintain consistent performance across diverse environmental conditions. Trait stability can be broadly categorised into two distinct concepts: static stability, where environmental variation does not affect trait expression, and dynamic stability, where the trait exhibits a predictable response to environmental changes (Becker and Leon, 1988). In such contexts, predicting the value of a genotype becomes more challenging, as performance in one environment may not be a good indicative of performance in another. The reduction in stability can result in unfavourable correlations between performance under specific environmental conditions and overall mean performance. This, in turn, limits realised response to selection and complicates efforts to achieve consistent breeding progress (Mulder, 2017).

The long evolutionary and breeding distance between *Bos taurus* and *Bos indicus* cattle, with the two subspecies diverging over 100,000 years ago (Pitt et al., 2018), has likely resulted in distinct genetic architectures. These differences manifest in unique mutations, allele frequencies, and linkage disequilibrium patterns between SNP mark-

ers and QTL. The genetic divergence may contribute to heterogeneity in allele effects between populations, causing challenges in multi-breed genetic evaluations (Rio et al., 2020; Misztal et al., 2022). Additionally, discrepancies between selection strategies for purebred versus crossbred populations may result in incompatible allelic combinations due to recombination load- the disruption of coadapted gene formations in the purebreds due to meiotic recombination in the crossbreds, and undermining heterosis, ultimately limiting the success of crossbreeding programmes (Rutledge, 2001).

The performance of taurine-indicine crossbred animals is expected to differ from that of their purebred parents due to the inherent genetic differences between *Bos taurus* and *Bos indicus*. Consequently, the G×E patterns exhibited in the crossbred animals are also expected to differ from the purebred parents (Mulder, 2017), meaning that an increase in both genetic and G×E variance is expected in the crossbreds due to new allelic combinations and greater influence of non-additive effects. It is important to note that, if well managed, the increased genetic variance and plasticity generated by crossbreeding and G×E interactions can be of great value and an advantage for the same crossbreeding programmes (Mulder, 2017, 2016; Buxadera and Mandonnet, 2006). The increase in genetic variance can serve as a resource, allowing breeders to identify animals better suited to the increasing unpredictability of tropical climates driven by climate change (Freychet et al., 2021; Rodgers et al., 2021).

Evidence shows that G×E interaction influences dairy cattle performance under diverse environmental settings (Boettcher et al., 2003; Hayes et al., 2003; Sartori et al., 2022). Although the challenges of evaluating crossbred animals has been well documented (Misztal et al., 2022), a key gap remains in understanding how G×E interaction and genetic architecture intersect to impact the performance of tropical crossbred dairy cattle, and how to effectively incorporate these factors to optimise crossbreeding strategies.

This study aims to study how to leverage genetic and G×E interaction in tropical crossbreeding, demonstrate how to model ancestry and environmental-specific QTL effects using stochastic simulations, and investigate how genetic and environmental variability influence crossbreeding performance across multiple environments. By applying methodologies from plant breeding, we developed a framework to uncover underutilised genetic diversity, providing insights that can guide the development of breeding strategies better suited to unstable environmental conditions, supporting the sustainable future of tropical dairy farming and the continued importance of crossbreeding strategies.

4.2 Materials and methods

In this study, we used stochastic simulations to evaluate the genetic and environmental factors that affect the performance of crossbred dairy cattle in tropical systems. The simulation was developed using the AlphaSimR package (Gaynor et al., 2021). We modelled both additive and dominance genetic effects, accounting for ancestry-specific QTL effects. G×E interaction was incorporated by simulating six correlated environments, with performance in each environment treated as a separate trait in AlphaSimR (Falconer, 1952). The simulation design consisted of five main components: (1) Population settings, (2) Environmental settings, (3) Breeding schemes, (4) Breeding programme evaluation, and (5) Investigating G×E interaction. Each of these components is detailed below.

4.2.1 Population settings

We simulated 2,000 founder haplotypes using the coalescent simulator MaCS (Chen et al., 2009), as implemented in AlphaSimR. The simulation used the built-in "CATTLE" parameters to model 30 diploid chromosomes, each consisting of 10^8 base pairs. The mutation rate was set at 2.5×10^{-8} per base pair per generation and the recombination rate at 1.0×10^{-8} per base pair per generation.

Following the demographic model in MacLeod et al. (2013), an evolutionary split was simulated 100,000 years ago representing the differentiation between *Bos primigenius primigenius* and *Bos primigenius nomadicus*, ancestors of modern taurine and indicine cattle. This differentiation was quantified by the correlation between allele frequencies in the two allelic pools, resulting in a correlation of approximately 0.07.

Three traits were simulated to represent milk yield in the taurine, indicine, and crossbred populations, hereafter denoted as T_T, T_I, and T_C, respectively. The traits capture the inherent differences in allele effects between populations arising from the evolutionary split. We assumed an additive genetic correlation of 0.6 between the traits and a narrow sense heritability of 0.3 for all traits. All three traits were simulated in six environments, providing the true performance of all taurine, indicine, and crossbred animals in all environments, as described in Environmental Settings.

We assigned 300 QTLs and 2,000 SNPs per chromosome for all traits, resulting in 9,000 QTLs and 60,000 SNPs in total. Additive and dominance effects were then assigned to each QTL, capturing 0.7 and 0.3 of the total genetic variance of each purebred trait

in the base population, respectively. The additive QTL effects were sampled from a multivariate normal distribution with mean determined by the trait (see Table 4.1) and target variance of $\sigma_a^2 = 18.0$. This target variance was achieved in each population by firstly setting a sufficiently high value in the combined base population prior to subsetting the indicine and taurine populations. As a result, the T_T and T_I haplotypes inherited the same QTLs but received distinct effects, representing ancestry-specific additive QTL effects from the taurine and indicine populations. The additive QTL effects were therefore calculated differently for each trait:

- The T_T and T_I purebred traits were defined as per AlphaSimR for all populations.
- The T_C crossbred trait was adjusted for the crossbred calves based on the breed of origin of the alleles. Using AlphaSimR's recombination tracking and identity-by-descent functionality, we traced the ancestry of alleles in each new set of crossbred calves and re-assigned the additive QTL effects based on whether the allele had taurine or indicine ancestry. The additive genetic effect of crossbred individual i at locus j was therefore generated as $a_{C_{ij}} = z_{T_{ij}} a_{T_j} + z_{I_{ij}} a_{I_j}$, where $z_{T_{ij}}$ is the taurine allele dosage with QTL effect a_{T_j} and $z_{I_{ij}}$ is the indicine allele dosage with QTL effect a_{I_j}. This ensured that the crossbred trait captured the true crossbred performance in the calves, and not some arbitrary trait representing the mid parent value.

The dominance QTL effects were generated by sampling dominance degree effects from a multivariate normal distribution with mean of $\mu_\kappa = 0.04$ and variance of $\sigma_\kappa^2 = 0.94$, resulting in the required proportion of dominance variance of 0.3 in the base population for each trait (Ertl et al., 2014; Aliloo et al., 2016). Additionally, for every 1% increase in inbreeding, a reduction of 0.45% was exhibited in the trait mean (Doekes et al., 2019). The dominance QTL effects were calculated differently for each trait:

- For the T_T and T_I purebred traits, the dominance QTL effect at locus j was calculated as $d_j = |a_j| \times \kappa_j$, where $|a_j|$ is the absolute value of the additive QTL effect with $a_j \sim N(\mu_a, \sigma_a^2)$ and κ_j is the dominance degree with $\kappa_j \sim N(\mu_\kappa, \sigma_\kappa^2)$. The dominance degree quantifies the extent to which the heterozygote deviates from the mid-point of the homozygotes, with a value of $\kappa_j = 0$ indicating no dominance at locus j, $0 < \kappa_j < 1$ indicating partial dominance, $\kappa_j = 1$ indicating complete dominance, and $\kappa_j > 1$ indicating overdominance (Falconer and Mackay, 1996).
- For the crossbred T_C trait, the dominance QTL effects were adjusted based on

the additive QTL effects of the purebred parents. Specifically, the absolute value of the QTL effect in the crossbred calves was defined as $|a_{C_j}| = (|a_{Tj}| + |a_{I_j}|)/2$. Note, however, that this notation is *incorrect*, as it results in $d \neq 0$ when $a_{T_j} = a_{I_j}$. See Section C.1 for further discussion. The dominance QTL effect at locus j was then calculated as $d_{Cj} = |a_{C_j}| \times \kappa_{C_j}$. This approach is an extension of Falconer and Mackay (1996) for crossbreeding settings (also see Lo et al., 1993; García-Cortés and Toro, 2006; Vitezica et al., 2013).

Note that we considered scenarios without dominance simulated, as described in Environmental Settings.

4.2.2 Environmental Settings

To generate additive and dominance G×E interaction, we simulated six correlated environments hereafter denoted as E_1 to E_6. Environment E_1 represented the climatic conditions and management practices of temperate regions with intensive production systems, while environments E_2 to E_6 represented a gradient of tropical regions, with E_2 being the most favourable and E_6 being the most challenging and resource-limited. The generative model used here follows a classical quantitative genetics approach in which milk yield for the three populations in the six environments were considered as different correlated traits (Falconer, 1952). This resulted in the simulation of 18 trait-by-environment combinations in AlphaSimR, which allowed us to map the true performance of all taurine, indicine, and crossbred animals in all environments. Phenotypes were then constructed for the observed combinations, as described in Breeding schemes.

Table 4.1: Additive genetic means for the 18 trait-by-environment combinations simulated, expressed in tonnes per lactation (t/l). Values in bold indicate observed trait-environment combination, given the environment where the population was placed.

	E_1	E_2	E_3	E_4	E_5	E_6
T_T	**10.37**	6.74	5.70	3.11	2.07	1.04
T_I	6.53	**5.54**	**4.75**	**3.57**	**2.97**	**1.59**
T_C	9.41	**8.08**	**6.93**	**5.42**	**2.76**	**1.30**

The additive genetic means for the 18 trait-by-environment combinations are presented in Table 4.1 for the base population, and ranged from 1.04 for taurine in environment E_6 to 10.37 for taurine in environment E_1. Trait means were based on the 305-day adjusted average production of the breeds Holstein (for T_T in E_1 see Holstein Association USA,

Inc., 2021), Brazilian Gir (for T_I in E_3 see Panetto et al., 2025) and Brazilian Girolando (for T_C in E_3 see Silva et al., 2024).

We assumed an additive genetic variance of 30% of the phenotypic variation for all combinations. The additive genetic correlations between environments are presented in Figure 4.1, with pair-wise correlations ranging from 0.9 to -0.3. By design, environments closer together are more similar than those further apart, with the correlations decaying as the distance between environments increased and their similarity decreased. Environment E_1 was therefore most correlated with E_2, followed by E_3, and so on until E_6. We assumed the same pair-wise correlations for all purebred and crossbred traits in the base population, representing a realistic level of additive G×E interaction.

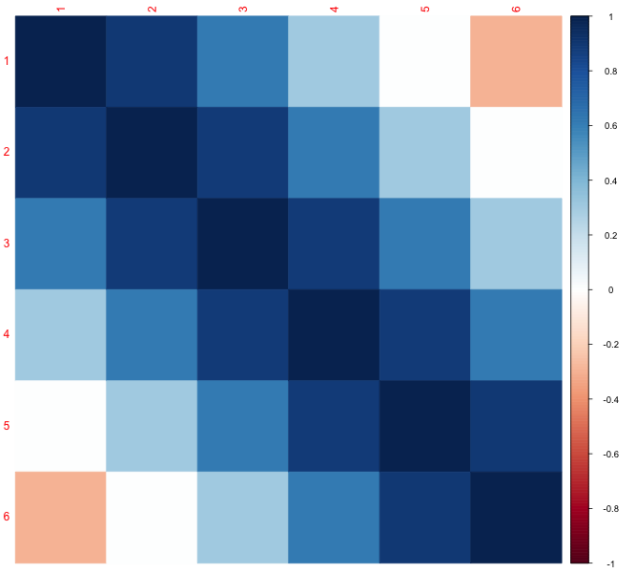

Figure 4.1: Heatmap of the additive genetic correlations between environments. The colour key ranges from +1 (perfect agreement in genotype rankings) through 0 (disagreement in rankings) to −1 (complete reversal of rankings). Note: Environment E_1 represents temperate regions with intensive production systems, while environments E_2 to E_6 represent a gradient of tropical regions with resource-limited production systems.

Three levels of dominance G×E interaction were considered; low, moderate, and high. The low level set a correlation of zero between all pairs of environments, which assumed independence between dominance degree effects in different environments. The other levels set a mean correlation of zero, but ranged from +0.5 to -0.5 for any given pair for the moderate level or +1 to −1 for the high level. These created more structured relationships between dominance degrees in different environments, impacting the expression of heterosis and its prediction from one location to the next. The scenario

without dominance simulated ignored dominance × environment interaction.

We used the R package FieldSimR (Werner et al., 2024) to generate the input parameters for AlphaSimR based on the genetic architecture and environment definitions described above.

4.2.3 Breeding schemes

We simulated two purebred breeding programmes for the taurine and indicine populations, which where then used in the crossbreeding programme as illustrated in Fig. 4.2.

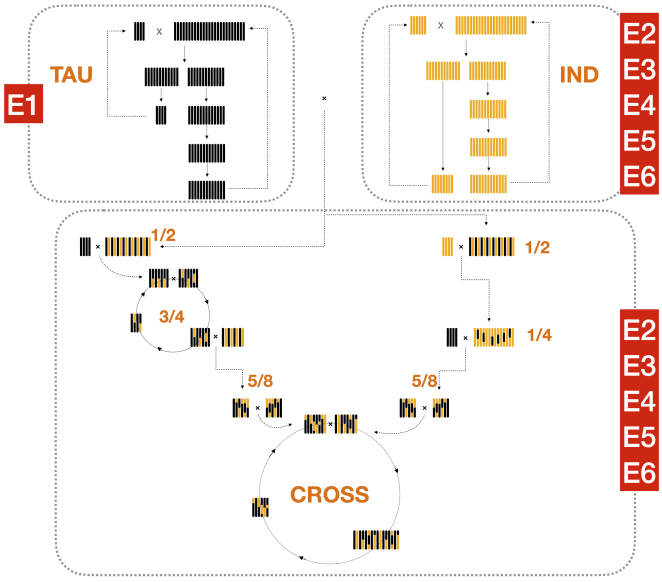

Figure 4.2: Schematic representation of the simulated breeding programmes across environments. Red labels indicate the environment where the breeding programme is placed and evaluated. Taurine (left) and indicine (right) breeding programmes are closed systems. Crossbreeding programme (bottom) is an oppen system, receiving input from purebred programmes every generation. Fractions represent the taurine proportion at each breeding cycle. Circles indicate evaluation and selection.

Breeding programmes were composed of five categories of animals: Heifers, cows, bull-calves, waiting-bulls and proven-bulls. Each category was composed of one to five contemporary groups. The total number of animals per category per breeding programme is given in Table 4.2.

Animals were randomly assigned to herds located in one of the six environments. Tau-

Table 4.2: Total animals per contemporary group for the Taurine, Indicine, and Crossbred breeding programmes, assuming a culling rate of 0.3 per generation for all female groups.

Category	Groups	Taurine	Indicine	Crossbred
Heifers	1	8,400	3,354	1,209
Cows				
– Lactation 1	1	5,880	2,348	846
– Lactation 2	1	4,116	1,644	591
– Lactation 3	1	2,881	1,151	414
– Lactation 4	1	2,017	807	291
– Lactation 5	1	1,412	565	204
	5	16,306	6510	2346
Bulls				
– Calves	1	200	100	100
– Waiting	4	80	80	80
– Proven	3	5	16	16
	8	285	196	196

rine animals were assigned to environment E_1 only (the temperate environment), while indicine animals were assigned to one of environments E_2 to E_6 (the tropical environments). Each herd varied in size, with herds in environment E_1 ranging from 40 to 400 animals with an average of 250 ± 100 animals, and herds in environments E_2 to E_6 ranging from 40 to 200 animals with an average of 70 ± 42.5 animals. Note that all calves were assigned to the same herd and environment as their mothers.

The phenotypic means for the 18 trait-by-environment combinations observed are presented in Table 4.1 for the base population, and ranged from 1.30 for the crossbred population in environment E_6 to 10.37 for the taurine population in environment E_1. Note that phenotypes were only expressed by females for the environment in which they were assigned, i.e., the taurine population only expressed in environment E_1, while the indicine and crossbred populations expressed in environments E_2 to E_6.

The phenotype of female i in environment k for lactation l was generated as:

$$y_{ikl} = g_{ik} + pe_{il} + h_{my} + e_{il} \tag{4.1}$$

where g_{ik} is the genetic effect of animal i in environment k, pe_{il} is the permanent environment effect of animal i for lactation l with $pe_{il} \sim \mathrm{N}(0, \sigma^2_{pe})$, h_{my} is the effect of herd m in year y with $h_{my} \sim \mathrm{N}(0, \sigma^2_h)$, and e_{il} is the random error with $e_{il} \sim \mathrm{N}(0, \sigma^2_e)$.

The genetic effect of an animal was generated as $g_{ik} = \sum_{j=1}^{n_q} z_{ij} a_{S_{jk}} + w_{ij} d_{S_{jk}}$, where z_{ij} is the additive QTL genotype for animal i at locus j coded as 0, 1, 2 and w_{ij} is the

dominance QTL genotype coded as 0, 1, 0 for animal i at locus j, $a_{S_{jk}}$ is the additive QTL effect of population S at locus j in environment k and $d_{S_{jk}}$ the dominance QTL effect of population S, at locus j in environment k ($a_{S_{jk}}$ and $d_{S_{jk}}$ are as defined in Population settings). Note that the genetic effect reduced to $g_{ik} = \sum_{j=1}^{n_q} z_{ij} a_{S_{jk}}$ for the scenario without dominance simulated. The variances set at the base population were defined according to the additve variance, σ_a^2, obtained after AlphaSimR rescaling and are summarised in Table 4.3.

Table 4.3: Variance components in the base population. Values are defined based on a ratio to phenotypic variance such to secure heritability $h^2 = 0.3$ for all traits and in all environments.

Component	Ratio to σ_p^2	Approx. value
Phenotypic variance (σ_p^2)	1.00	60.0
Additive genetic variance (σ_a^2)	0.3	18.0
Dominance variance (σ_d^2)	0.13	7.8
Permanent environment variance (σ_{pe}^2)	0.10	6.0
Herd variance (σ_h^2)	0.35	21.0
Residual variance (σ_e^2)	0.12	7.2
Residual variance (no dominance scenario)	0.25	15.0

Calves were generated every breeding cycle. In each cycle, heifers and bull-calves were randomly selected and kept for one breeding cycle, during which heifers were mated to young bulls. In the subsequent cycle, superior heifers were selected based on phenotypes to become cows1 (cows with one completed lactation), and the non-selected heifers were excluded from the breeding programme. A general culling rate of 30% per generation was applied across all breeding programmes. This process was repeated each breeding cycle until the female reached the cows5 category (five completed lactations), at which point all animals were culled. Superior bull-calves were selected based on true genetic values after one breeding cycle and moved to the waiting-bulls category, where young bulls were progeny tested. After four cycles, waiting-bulls were selected to become proven-bulls. Proven-bulls were used as sires for three breeding cycles, and the non-selected animals were excluded from the breeding programme.

The crossbreeding programme was initiated after 20 burn-in cycles of the purebred programmes and required 9 additional breeding cycles to populate all animal categories. In the first breeding cycle, sires from the taurine breeding programme were mated with the top indicine cows. We assumed reproductive technologies were in place so that indicine cows could mother calves in both the purebreeding and crossbreeding programmemes during the same breeding cycle. All calves produced in this cycle

were F1, with a 50% taurine and 50% indicine breed composition. From the second cycle onwards, the crossbreeding programme followed the procedure described earlier, incorporating indicine sires into the mating scheme. In the third cycle, F1 heifers were backcrossed with both parental breeds, generating animals with 3/4 and 1/4 breed compositions. We confined the expected breed compositions to 1/4, 1/2, 5/8, 3/4, and 7/8 taurine. Both expected and realized breed compositions were tracked for all animals in the crossbreeding programme.

All breeding programmes advanced for 40 breeding cycles.

Breeding objective. Across all breeding programmes, bulls were selected using a selection index (Hazel, 1943; Smith, 1936). The selection index for animal i was defined as:

$$I_i = \omega_1 a_{i1} + \omega_2 a_{i2} + \cdots + \omega_k a_{ik} = \sum_{k=1}^{p} \omega_k a_{ik} \tag{4.2}$$

where ω_k is the weight of environment k, with $\sum_{k=1}^{p} \omega_k = 1$. The index for the temperate bulls (I_{T_i}) was taken as $I_{T_i} = a_{i1}$ with $\omega_1 = 1$ and $\omega_{2:6} = 0$, reflecting that all taurine cows expressed their phenotypes exclusively in E_1. In contrast, the index for the tropical bulls (I_{I_i}) was taken as $I_{I_i} = \frac{1}{5}\sum_{k=2}^{6} a_{ik}$ with $\omega_1 = 0$ and $\omega_{2:6} = 1/5$. This reflected the fact that indicine and crossbred cows expressed phenotypes across E_2-E_6, which were given equal weight in the genetic evaluation.

4.2.4 Breeding programme evaluation

The performance of the crossbreeding programme was assessed by comparing the genetic gain in the calves produced in the last breeding cycle with those from the purebred programmes. Genetic gain was measured as the difference between the mean selection index in the first and last breeding cycles. For clearer comparison, we present the outcome of the taurine breeding programme based on the tropical selection index I_{I_i} in addition to its outcome in the temperate environment (I_{T_i}).The mean genetic value for all 18 trait-by-environment combinations over the 40 post-burnin breeding cycles was tracked as well as the genetic variance of individuals in the tropical breeding programmes (indicine and crossbreeding) to determine the impacts of the selection on mean performance on the adaptability of individuals and the progression of the population.

The metrics described above were compared across four different dominance scenarios. Scenarios are labelled as follows: A for no dominance (additive only), ADL for low

dominance G×E interaction, ADM for medium dominance G×E interaction, and ADH for high dominance G×E interaction. Each scenario was replicated 10 times to account for stochasticity in the simulation process.

4.2.5 Investigating G×E interaction

To investigate G×E interaction, we explored informative decompositions of the breeding values in each environment and developed novel metrics and visualisations, with the aim to provide breeders with a general and interpretable framework for improving selection in crossbreeding programmes. Assuming bulls have lactating daughters in all environments, we decomposed their breeding values for each environment into the product of two components: a set of (hypothetical) environmental covariates and a set of genotypic slopes. This produced a multiplicative model for G×E interaction which decomposes the genetic variance structure into environmental and genotypic contributions, facilitating interpretation of environmental similarity as well as stability and adaptability of the selection candidates. A new rotation is then presented based on the work of Tolhurst (2024) from plant breeding, which we have adapted and extended for animal breeding settings. Importantly, the extensions below also have meaningful application back to plant breeding. The methods are demonstrated below for the additive only scenario (A), where the breeding value of individual i in environment k is given by $u_{ik} = a_{ik}$, and a_{ik} is equal to the genetic effect from Eq. 4.1 minus the population mean for environment k. The population superscript T, I, or C removed below for brevity. The extension to the dominance scenarios (ADL, ADM and ADH) is straightforward by defining $u_{ik} = a_{ik} + d_{ik}$. The methods below were applied to the selection candidates from the waiting_bulls category, comprising 80 bulls from each of the taurine, indicine, and crossbreeding programmes.

Baseline model

We initially decomposed the breeding value for each individual in each environment into the sum of two components: a main effect and a G×E interaction effect. The main effect represents the average breeding value across all environments while the interaction effect represents the deviation from the average for a particular environment. Consequently, high performing and stable individuals exhibit a large positive main effect and near-zero interaction effects for all six environments, whereas specifically adapted individuals exhibit large positive interaction effects for particular environments.

Assuming breeding values are available for v individuals across p environments, with

$p = 6$ for the simulated crossbreeding programme, the breeding value of individual i in environment k can be expressed as:

$$a_{ik} = a_i + \varepsilon_{ik} \tag{4.3}$$

where a_i is the main effect of individual i given by its average breeding value across environments ($a_i = \frac{1}{6}\sum_{k=1}^{6} a_{ik}$) and ε_{ik} is the interaction effect of individual i in environment k given by $\varepsilon_{ik} = a_{ik} - a_i$. It then follows that $a_i \sim \mathrm{N}(0, \sigma_a^2)$ and $\varepsilon_{ik} \sim \mathrm{N}(0, \sigma_{\varepsilon_k}^2)$, where the additive genetic variance σ_a^2 is the main effect variance and $\sigma_{\varepsilon_k}^2$ is the interaction variance for environment k. The additive genetic variance of environment k is therefore given by $\sigma_{\varepsilon_k}^2 = \sigma_a^2 + \sigma_{\varepsilon_k}^2 + 2\sigma_{a\varepsilon_k}$, where $\sigma_{a\varepsilon_k}$ is the covariance between the main effects and interaction effects.

The model above provides a simple representation of the breeding values across different environments, and a simple way to summarise an individual's overall performance and stability. Overall performance is generally taken as an individual's main effect, a_i, while stability is taken as the variance of its interaction effects across environments, $\frac{1}{6}\sum_{k=1}^{6} \varepsilon_{ik}^2$. Note, however, this model has a key limitation in the presence of non-crossover G×E interaction because it assumes the magnitude of an individual's main effect is exactly the same in each environment, meaning that any genetic variance or covariance heterogeneity between environments is captured exclusively in the interaction effects. This leads to spurious measures of stability, with individuals capturing more heterogeneity in their breeding values being inappropriately penalised, i.e. appearing more unstable. This limitation is addressed below through the use of multiplicative models.

Multiplicative model

We then decomposed the breeding values for each individual into the product of two independent components: a set of latent (unobserved) covariates for each environment and a set of slopes for each individual. This produces a multiplicative model, which has been widely adopted in plant breeding settings for its ability to capture genetic variance and covariance heterogeneity between environments using a small number of multiplicative terms (Gollob, 1968; Mandel, 1971). The multiplicative model used here comprises $p = 6$ terms, meaning that the model has full rank and captures all genetic variance and covariance heterogeneity in the breeding values. When the number of terms is less than the number of environments, the model has reduced rank and will capture some (generally large) proportion of the total genetic variance.

The breeding value of individual i in environment k can therefore be re-expressed as:

$$a_{ik} = \lambda_{k1} f_{i1} + \lambda_{k2} f_{i2} + \cdots + \lambda_{k6} f_{i6} = \sum_{r=1}^{6} \lambda_{kr} f_{ir} \tag{4.4}$$

where λ_{kr} is the covariate of environment k for term r and f_{ir} is the corresponding slope of individual i. It is assumed that the environmental covariates are orthonormal, meaning that the they have unit length for each term while being independent across terms, i.e. $\sum_{r=1}^{6} \lambda_{kr}^2 = 1$ and $\sum_{r=1}^{6} \lambda_{kr}\lambda_{k'r} = 0$ for all r. It then follows that $f_{ir} \sim \mathrm{N}(0, d_r)$, where d_r is the variance of term r. The additive genetic variance of environment k is therefore given by $\sigma_{g_k}^2 = \sum_{r=1}^{6} \lambda_{rk}^2 d_r$.

In this study, the form of Eq. 4.5 was derived by applying a singular value decomposition to a $v \times p$ matrix, with individuals represented by rows and environments by columns (the G×E table):

$$\mathbf{E} = \mathbf{F}\mathbf{\Lambda}^{\top} \tag{4.5}$$

where $\mathbf{E} = \{a_{ik}\}$ is a $v \times 6$ matrix with columns given by the breeding values for all v waiting bulls in each of the 6 environments, $\mathbf{\Lambda} = \{\lambda_{rk}\}$ is a 6×6 matrix with columns given by the latent environmental covariates for each multiplicative term, and $\mathbf{F} = \{f_{ri}\}$ is the corresponding $v \times 6$ matrix containing the individual slopes. Applying the singular value decomposition above ensures that $\mathbf{F}$ contains the left singular vectors of $\mathbf{E}$ (scaled by the singular values) while $\mathbf{\Lambda}$ contains the right singular vectors. It also ensures that the environmental covariates are orthonormal and ordered in decreasing magnitude, meaning that the first multiplicative term captures the most genetic variance, followed by the second and so on, i.e. $d_1 > d_2 > \ldots > d_6$. This is commonly referred to as a principal component rotation in the plant breeding literature (Cullis et al., 2010; Smith et al., 2015).

The proportion of variance explained by term r is given by $v_r = d_r / \sum_{r=1}^{6} d_r$, which provides a formal way to assign importance to the genotype responses on each environmental covariate. We also examined the percentage of variance explained for individual i and environment k, given by:

$$v_{ri} = \frac{f_{ri}^2}{\sum_{r=1}^{6} f_{ri}^2} \qquad \text{and} \qquad v_{rk} = \frac{\lambda_{rk}^2 d_r}{\sum_{r=1}^{6} \lambda_{rk}^2 d_r} \tag{4.6}$$

which provide useful measures to identify terms of interest for particular individuals and environments, and to characterise interaction patterns in terms of the stability of

individuals and the discriminating ability of environments. For example, if an individuals (or environment) exhibits small amounts of variance explained across multiple terms, then its interaction pattern will be complex and its performance (or discriminating ability) will be influenced by multiple underlying environmental factors, making it more likely to be unstable across the environments under study. In contrast, if a large proportion of variance is explained by the first term, then its interaction pattern will be much simpler and its performance (or discriminating ability) will be mainly influenced by the dominant environmental factor, and therefore more likely to be stable. Animals associated with the higher-order terms also display simple interaction patterns, but these generally reflect specific adaptation to particular environmental conditions rather than broad stability. Environments associated with these terms can be used to identify and characterise specifically adapted individuals. These features will be exploited when visualising the G×E interaction patterns later in the section.

The multiplicative model provides an informative representation of the breeding values by capturing genetic variance and covariance between environments through a series of multiplicative terms. In animal and plant breeding data, the first term generally captures the discriminating ability of environments with regards to the trait under study while the higher order terms capture genotype contrasts (Meyer, 2007, 2009a; Zobel et al., 1988; Yan and Kang, 2003). These components have been historically referred to as non-crossover and crossover G×E interaction, respectively (Mulder and Bijma, 2005; Crossa et al., 1990; Gauch, 1992; Smith and Cullis, 2018). Note, however, although the singular value decomposition typically provides desirable features for modelling G×E interaction in practice, with the first multiplicative term capturing non-crossover interaction and the higher order terms capturing crossover interaction, it does not guarantee them. In fact, these features are a consequence of the data structure, rather than a consequence of an approach that seeks to find a first term that exclusively captures non-crossover interaction and higher order terms that exclusively capture crossover interaction. A new rotation is developed in the next section which addresses these objectives, exclusively placing all non-crossover variation into the first term and all crossover variation into the higher order terms.

A new informed rotation

Although numerous rotations have been used in breeding applications to help interpret the underlying structure from a G×E table, such as varimax (Kaiser, 1958), promax (Hendrickson and White, 1964), and oblimin/oblimax (Jennrich and Sampson, 1966), there is no current approach which actively disentangles non-crossover and crossover

interaction that is meaningful for breeder's selection decisions. The rotation developed in the following actively achieves this objective by anchoring the decomposition to the breeding objective, represented by a selection index (such as in Eq. 4.2). By treating I_i as the focal point, the method disentangles G×E variation into components corresponding to non-crossover (correlated to I_i) and crossover interaction (uncorrelated to I_i). The methods developed here are loosely based on Falconer (1990), which was recently revisited for reaction norm models by Waters et al. (2023). We extend this approach and develop a general framework for the class of multiplicative models. The application to reaction norm and random regression models with any number of environmental covariates is presented in Tolhurst (2024).

Assuming selection is applied to the main effect, the first multiplicative term will exclusively capture the expected (non-crossover) response of individuals in each environment due to perfect positive correlation with a_i, while the higher terms will capture the expected (crossover) response arising from a lack of correlation. The first term therefore provides a natural and meaningful way to summarise overall genetic merit and potential for broad adaptation across environments, while the higher order terms can be used to identify animals with high potential for specific adaptation in particular environments. These features are critical to obtain informative measures of overall performance, stability, and responsiveness of individuals which are meaningful for selection (Tolhurst, 2024).

The new rotation expresses the breeding value of individual i in environment k as:

$$a_{ik} = b_k a_i + \lambda_{k1} f_{i1} + \lambda_{k2} f_{i2} + \cdots + \lambda_{kp-1} f_{ip-1} = b_k a_i + \sum_{r=1}^{5} \lambda_{kr} f_{ir} \tag{4.7}$$

where a_i is the main effect of individual i from Eq. 4.3 with corresponding scaling coefficient b_k which have mean of one across environments, i.e. $\frac{1}{6}\sum_{k=1}^{6} b_k = 1$, λ_{kr} is the covariate of environment k for term r and f_{ir} is the corresponding slope of animal i from Eq. 4.5. The assumptions on λ_{kr} and f_{ir} are as for the multiplicative model above, but note that these terms now have reduced rank due to the formation of the first term involving the main effects. It is assumed that $b_k \geq 0$ for all k, so that the first term exclusively captures non-crossover interaction (see Tolhurst, 2024, for a complete discussion). All remaining assumptions have been relaxed to facilitate interpretation, i.e. the slope on b_k is exactly a_i, but note that these will be scaled to unit length for visualisation. Also note that constraints can be placed on the loadings for each term so that they are orthogonal to the scaling coefficients, such that $\sum_{k=1}^{6} b_k \lambda_{rk}$ for all r.

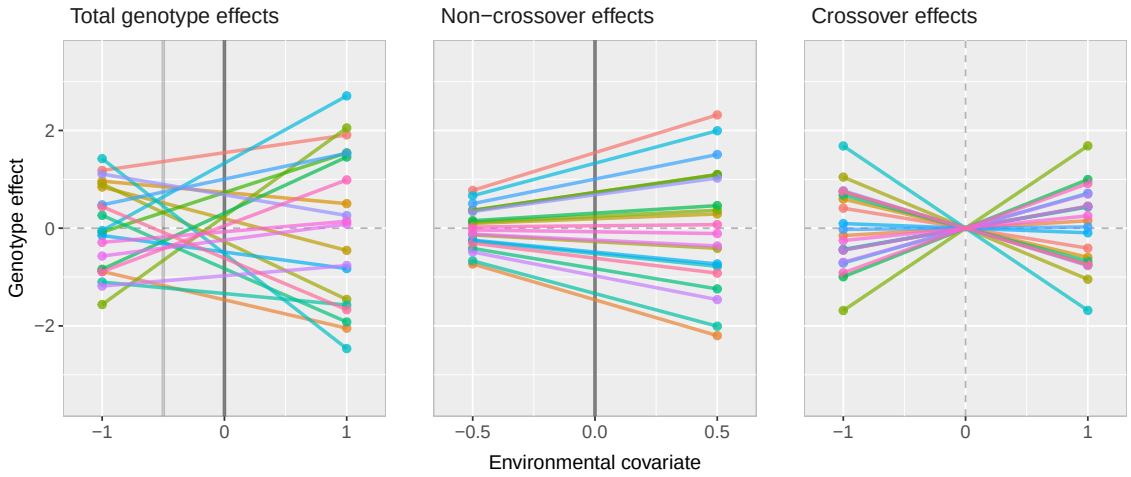

Figure 4.3: The effects of 20 genotypes in 2 environments for a hypothetical environmental covariate. The figure demonstrates the disentangling of the total genotype effects into non-crossover and crossover effects.

This is the topic of current research and practical applications.

Figure 4.3 illustrates how the new rotation disentangles non-crossover and crossover G×E interaction. The individuals' intercepts (and slopes) in the middle panel are given by the individuals main effect (a_i), while the covariates are given by $b_k = -0.5$ and 0.5. The *black lines* represent the point where selection for improving the average environment is achieved (where I_i has $\omega_1 = \omega_2$). The *grey line* represents the point where the genotype effects are independent of the slopes, i.e. where selection for improving both environments is achieved simultaneously.

The scaling coefficient b_k represents the amount of variance in the breeding values for environment k explained by the individual main effects. The coefficient is obtained by conditioning the breeding values on the main effects, producing two independent components: the non-crossover and crossover effects, which exclusively capture non-crossover and crossover interaction with reference to a_i. This process is equivalent to regressing the breeding values for environment k onto the main effects, producing a regression slope of $b_k = \sigma_{gg_k}/\sigma_a^2$, where σ_{gg_k} is the covariance between the breeding values and main effects, which is equal to $\sigma_{gg_k} = \sigma_a^2 + \sigma_{a\varepsilon_k}$ using components from Eq. 4.3, and σ_a^2 is the variance of the main effects. The non-crossover and crossover effects for individual i in environment k are therefore defined as:

$$n_{ik} = b_k a_i \qquad \text{and} \qquad c_{ik} = \sum_{r=2}^{6} \lambda_{rk} f_{ri} \tag{4.8}$$

where n_{ik} equal the fitted values along the regression line and c_{ik} equal the deviations around the regression line, meaning that both sets of effects are statistically independent. The breeding values in Eq. 4.7 can therefore be re-expressed as $a_{ik} = n_{ik} + c_{ik}$, with $n_{ik} \sim \mathrm{N}(0, \sigma^2_{n_k})$ and $c_{ik} \sim \mathrm{N}(0, \sigma^2_{c_k})$, where $\sigma^2_{n_k}$ is the non-crossover variance given by $\sigma^2_{n_k} = b^2_k \, \sigma^2_a$ and $\sigma^2_{c_k}$ is the crossover variance given by $\sigma^2_{c_k} = \sum_{r=1}^{5} \lambda^2_{rk} d_r$. Note that the non-crossover variance is equivalent to the variance in environment k explained by the main effects, i.e. $\sigma^2_{n_k} = \rho^2_{gg_k} \sigma^2_{a_k}$, while the crossover variance equals all remaining variation independent of the main effects, $\sigma^2_{c_k} = (1 - \rho^2_{gg_k}) \, \sigma^2_{a_k}$, where ρ_{gg_k} is the correlation between the breeding values and main effects given by $\rho_{gg_k} = b_k \, \sigma_a / \sigma_{a_k}$. The additive genetic variance of environment k is therefore given by $\sigma^2_{a_k} = \sigma^2_{n_k} + \sigma^2_{c_k}$. Lastly, measures of non-crossover and crossover variance were obtained, both overall and for each animal and environment.

The form of Eq. 4.7 was obtained by applying the new rotation to the G×E table, with the matrices of individual slopes and environmental covariates rotated as:

$$\mathbf{F}^* = \mathbf{F}\mathbf{V} \qquad \text{and} \qquad \mathbf{\Lambda}^* = \mathbf{\Lambda}\mathbf{V} \tag{4.9}$$

where $\mathbf{\Lambda}^* = \{b_k, \, \lambda_{rk}\}$ is a 6×6 matrix with columns given by the latent environmental covariates for each multiplicative term and $\mathbf{F}^* = \{a_i, \, f_{ri}\}$ is the corresponding $v \times 6$ matrix containing the individual slopes. It then follows that $\mathbf{E} = \mathbf{F}^* \mathbf{\Lambda}^{*\top}$ from Eq. 4.6. It is important to note that while $\mathbf{F}^*$ and $\mathbf{\Lambda}^*$ no longer contain the left and right singular vectors of $\mathbf{E}$, Eq. 4.9 does represent an orthogonal rotation, such that all multiplicative terms are orthogonal and both $\mathbf{V}^\top \mathbf{V}$ and $\mathbf{V}\mathbf{V}^\top$ are diagonal matrices. Full details on the new rotation can be found in Tolhurst (2024).

The model above provides a meaningful representation of the breeding values across different environments, and a natural way to summarise an individual's overall performance and stability which are aligned with the breeding objective. Overall performance is taken as an individual's main effect, a_i, which is equal to the average of the non-crossover effects across environments, since $\frac{1}{6} \sum_{k=1}^{6} b_k = 1$, while stability is taken as the variance of its crossover effects, $\frac{1}{6} \sum_{k=1}^{6} c^2_{ik}$. Importantly, this approach directly addresses the limitations of the measures obtained from the baseline model arising from the presence of genetic variance and covariance heterogeneity. Here, we assume the magnitude of an individual's main effect is scaled differently for each environment by b_k, meaning that any heterogeneity associated with the main effects is captured exclusively in the non-crossover effects and any remaining heterogeneity independent of the main effects is captured in the crossover effects. Consequently, high performing

and stable individuals still exhibit a large positive main effect but will have near-zero crossover effects for all six environments, whereas specifically adapted individuals exhibit large positive crossover effects for particular environments.

The new rotation and subsequent measures are particularly relevant for breeder's selection decisions because they have been built on selection index theory. Under this framework, the first multiplicative term will exclusively captures the expected *correlated* response of individuals across all environments when selection is applied to a given breeding objective, while higher-order terms capture the expected *uncorrelated* response arising from re-ranking between environments. In the initial formulation, the breeding objective was defined by the animal main effects (a_i). We subsequently redefined the breeding objective using temperate selection index, I_{T_i}, reflecting enphasis on the environment where intensive selection is applied for the taurine population. As a result, bulls were identified that not only perform well in E_1, but also perform particularly well and higher than expected in the tropical environments E_2 to E_6, as characterised by large positive crossover effects for all these environments.

The new rotation is conceptually similar to target rotation (Jennrich, 2001), where the solution is rotated toward a predefined loading structure, and to confirmatory factor analysis (Jöreskog, 1969), where structural constraints are imposed on factor loadings to match hypothesised biological patterns. The key difference here is that the first multiplicative term is completely aligned with the breeding objective, leaving the higher-order terms to capture all remaining crossover variation independent of the breeding objective. This means that selection on the non-crossover effects maximises the response to selection with respect to the breeding objective, with the non-crossover variance providing a direct measure of the expected response to selection in each environment.

Taking the individual main effects as the breeding objective, the response to selection in environment k can be expressed as:

$$R_{a_k} = i\, h_a^2\, h_{a_k}^2\, \sigma_{n_k} \tag{4.10}$$

where i is the selection intensity, h_a^2 is the narrow-sense heritability of the main effects, $h_{a_k}^2$ is the narrow-sense heritability of the breeding values, and σ_{n_k} is the square root of the non-crossover variance given by $\sigma_{n_k} = \rho_{a_k}\, \sigma_{a_k}$. The non-crossover variance therefore provides a direct measure of the expected genetic gain in each environment from selection on a_i, while the crossover variance defines any remaining potential for

genetic gain in a_{ik} specific to particular environments (or subsets there of). This means that when $\sigma_{n_k} = \sigma_{a_k}$, then $\rho_{a_k} = 1$ and environment k comprises non-crossover interaction only. However, when $\sigma_{n_k} < \sigma_{a_k}$, then $\rho_{a_k} < 1$ and environment k comprises both non-crossover and crossover interaction. The same results hold for any breeding objective, but note that the crossover effects will only sum to zero when the focus is placed at the individual main effects. These features will be exploited when visualising the G×E interaction patterns in the following section.

Visualisation and exploratory analysis

We use biplots and reaction norm plots to visualise the results from the multiplicative model with the conventional singular value decomposition and the new rotation for disentangling non-crossover and crossover interaction. Combined, these complementary tools depict the performance of animals in particular environments as well as their sensitivity to changes in the environment.

Biplots derived from multiplicative models, such as AMMI (Gauch, 1988, 1992) and GGE (Yan et al., 2000; Yan and Kang, 2003), have become a popular approach in plant breeding for visualising complex G×E interactions and illustrating the so-called "what wins where" framework (Yan and Kang, 2003) - used to identify which genotypes are best suited to which environments (Kempton, 1984). In contrast, there has been very little uptake in animal breeding, with most notable examples focusing exclusively on factor analytic modelling and principal component analysis of the genetic covariance matrix (Meyer, 2007, 2009b; Domínguez-Castaño et al., 2021). Biplots therefore present novel opportunities in animal breeding, particularly for crossbreeding settings, to explore breed complementarity and to identify particularly well-adapted purebreds and crossbreds in tropical environments, despite intensive selection applied elsewhere in temperate environments. We advocate for the use of biplots in this study. Combined with the new rotation developed above, they provide a natural platform to explore our proposed "what wins where, given selection elsewhere" framework.

The graphical components of the biplot are as follows:

- x-axis: Multiplicative term 1.
- y-axis: Multiplicative term $2, 3, \ldots, 6$.
- Projection for environment k: vector mapping the origin to $(\lambda_{1k},\ \lambda_{rk})$.
- Projection for individual i: vector mapping the origin to $(f_{1i},\ f_{ri})$.

- Key measures: proportion of variance explained by term r given by v_r and angle between terms 1 and r given by θ_{1r}.

Note that the first multiplicative term for the new rotation is rewritten as $b_k a_i = \lambda_{1k} f_{1i}$, with $\lambda_{1k} = b_k/b$ and $f_{1i} = ba_i$, where $b^2 = \sum_{k=1}^{6} b_k^2$ such that the λ_{1k} now have unit length and $f_{1i} \sim \mathrm{N}(0, d_1)$ with $d_1 = b^2\sigma_a^2$. An optional constraint may also be imposed for the new rotation, ensuring that the higher order terms become orthogonal to the first term, i.e. $\sum_{k=1} \lambda_{1k}\lambda_{rk}$ and thus $\theta_{1r} = 90$ for all $r > 1$. The biplot approach here can be used, for instance, to place the breeding objective at the individual main effects, the breeding values for E_1, or the selection index (I_i). The breeding objective is then represented by a horizontal line mapping the origin to the value of the first latent covariate, since it is fully explained by this term by design and the latent covariates are zero for all higher-order terms.

Reaction norm plots have been traditionally used in breeding and genetics applications to illustrate the response of genotypes along a continuous environmental gradient. In plant breeding, reaction norms have been widely used to identify genotypes with broad or specific adaptation, originally in the context of regressions on latent environmental indices (Finlay and Wilkinson, 1963; Eberhart and Russell, 1966) and later incorporating measured environmental variables such as temperature and rainfall (Perkins and Jinks, 1968; Knight, 1970). Similar approaches have been widely adopted in animal breeding, particularly for estimating genetic parameters to quantify plasticity, robustness, and stability in the presence of G×E interaction (Schaeffer, 2004; Calus and Veerkamp, 2003; Mulder and Bijma, 2011). Compared to biplots, reaction norm plots highlight the dynamic trajectories of genotype response across environmental gradients rather than static summaries of their interaction. We therefore utilise reaction norm plots as a complementary tool for visualising genotype-specific responses to selection and for further exploring our proposed "what wins where, given selection elsewhere" framework.

The graphical components of the reaction norm plot are as follows:

- x-axis: Environmental covariates for term r given by λ_{rk}.
- y-axis: Genotype response across environments.
- Data points: Breeding values for individual i given by a_{ik}, which are adjusted for the preceding term/s when $r > 1$, producing $a_{ik} - \sum_{h=1}^{r-1} \lambda_{hk} f_{hi}$.
- Regression lines: individual slopes for term r given by f_{ri}.

- Fitted values on the regression line given by $\lambda_{rk} f_{ri}$ and deviations around the regression line given by $\sum_{h=r+1}^{6} \lambda_{hk} f_{hi}$.
- Key measures: proportion of variance explained given by v_r.

Like above for the biplot, the reaction norm plot can also be used to examine correlated and uncorrelated response to selection with respect to the breeding objective, now represented by a vertical line for the first multiplicative term. Higher order covariates also have zero values under this breeding objective; however secondary points of selection can be chosen as required. For example, specifically adapted individuals to one or more environments can be identified after accounting for the intensive selection made in environment E_1. Such individuals are analogous to so-called "correlation breakers" in the context of multi-trait animal and plant settings (Robertson, 1959; Moll and Stuber, 1974).

While biplots have popularised the "what wins where" framework in plant breeding by identifying which genotypes are best suited for which environments, reaction norm plots extend this perspective by identifying which genotypes are best suited along explicit environmental gradients. In animal breeding, the same logic applies: biplots can illustrate which breeds or crosses perform best under different management systems, while reaction norms reveal which individuals or lines are most robust or most sensitive across production environments. Together these graphical tools provide complementary ways to visualise G×E interaction from the perspective of individuals and environments, and hence their inclusion in our new framework.

4.3 Results

Our findings demonstrate that the relative performance of crossbreeding strategies is strongly dependent on the structure of G×E interaction and the degree of dominance correlation across environments. While the indicine breeding programme appeared more beneficial on average, crossbreeding had clear advantages in specific environments, particularly those where environmental conditions were better and correlation with the taurine selection environment (E_1) was higher. Genetic gain was influenced by trait-environment correlations, with strong correlations leading to greater progress. Multi-environment breeding strategies led to increase in genetic variation, especially under scenarios with stronger dominance effects. Applying our new framework exposed distinct adaptation profiles and complex patterns of G×E interaction. Informed rotations highlighted genotype-specific adaptation of taurine bulls to different environments

indicating potential for strategical tropical crossbreeding applications.

4.3.1 Genetic gain

On average, there was no significant difference in genetic gain for the selection index in the last breeding cycle between scenarios (results not shown). Considering the dominance scenario with moderate environmental correlation (ADM), the indicine breeding programme slightly outperformed the crossbreeding programme, with an average genetic gain of 2.9 t/l compared to 2.8 t/l. The taurine breeding programme (TAU) achieved significantly higher genetic gain at the end of the simulated period when considering the selection index as I_{T_i} compared to the indicine (IND) and crossbreeding (CROSS) programmes, where selection was performed based on I_{I_i}. No significant difference was observed between the IND and CROSS programmes (Figure 4.4). When taurine animals were evaluated under tropical environments using I_{I_i} (TAU-TROP), no significant difference was observed between breeding programmes. At the end of the simulation, TAU achieved 3.3 times greater genetic gain than TAU-TROP. Among the tropical programmes, IND had the highest mean 2.9 (not significant different). F1-CROSS mean genetic gain (2.7) fell below the midparent average based on TAU (5.9) and TAU-TROP (2.8).

Figure 4.5 illustrates the trends in genetic gain over time for the three breeding programmes across environments and scenarios. Solid lines represent the traits under selection, while dotted lines track the selection index. In the taurine breeding programme, environment E_1 achieved the highest genetic gain by the end of the simulation, while environment E_6 declined in performance due to a moderate negative correlation with the selection index focused on E_1. The distance between the trait-by-environment trends shifted depending on the dominance correlation between environments but rank order remained consistent: $E_1 > E_2 > E_3 > E_4 > E_5 > E_6$.

In the indicine programme, environments E_3, E_4, and E_5 showed the strongest correlation with the selection index (the mean of $E_2 - E_6$) and exhibited the most improvement over time. The ranking of traits varied depending on dominance correlation between environments, but E_1 consistently performed the worst across scenarios, despite steady gain. Environment E_6 showed peak performance in the ADL scenario, where its correlation with the index was highest.

For the crossbreeding programme, genetic progress was driven in part by selection in the purebred populations. A notable re-ranking of environments was observed over

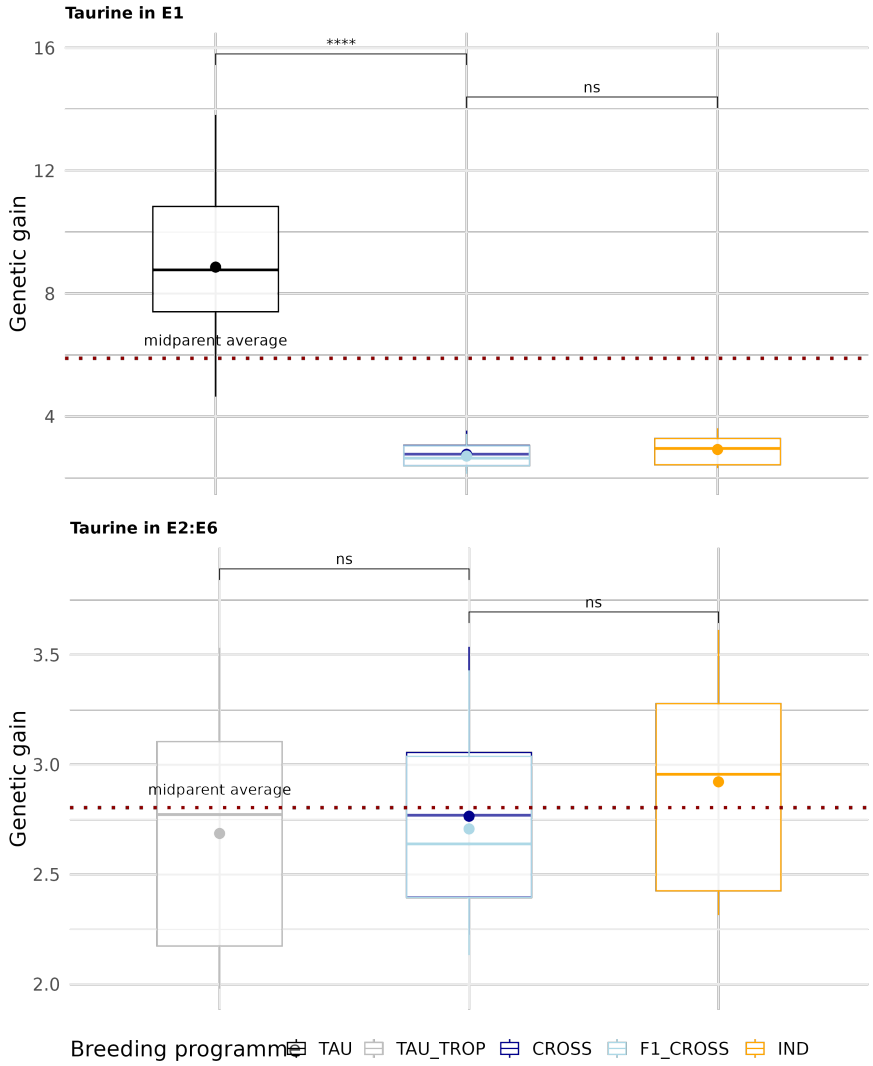

Figure 4.4: Breeding Programmes Outcomes. Genetic gain in the final year under the medium dominance correlation (ADM) scenario. Results are based on the temperate selection index (I_{T_i}) for the taurine population and tropical index (I_{I_i}) for the crossbred and indicine populations (top figure), and on I_{I_i} for the taurine population (bottom figure). First generation crossbreds (F1-CROSS) outcomes are highlighted from the crossbred population.Statistically significant differences in mean genetic gain are shown between the crossbred outcome and each purebred population.

time. For example, E_1 (not included in the selection index) performed similarly to the selection index and eventually outperformed some traits that composed the selection index. On the other hand, E_6 consistently showed the lowest genetic gain over time with its performance most impacted by increasing dominance correlation between environments.

While the crossbreeding programme did not consistently outperform the indicine breeding programme on average, it demonstrated superior genetic gain in specific environ-

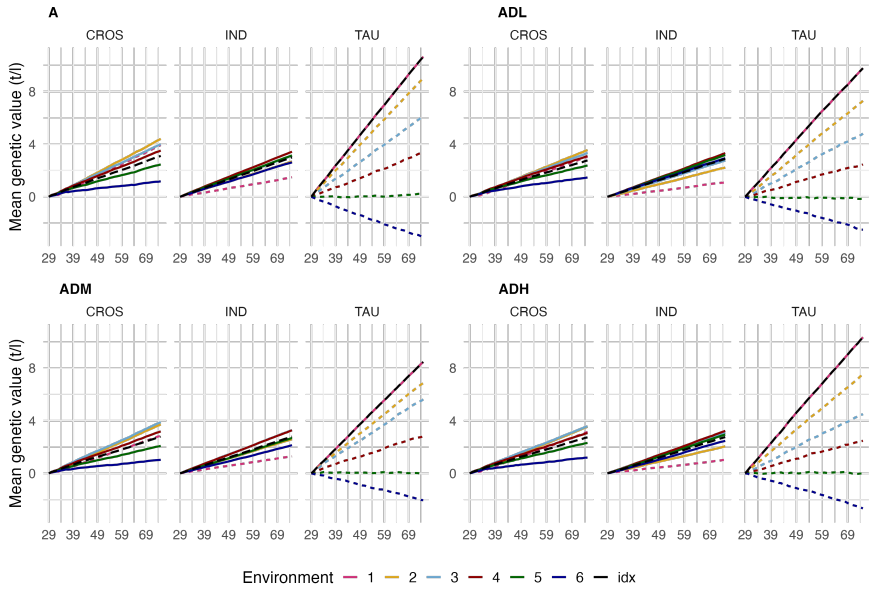

Figure 4.5: Genetic Gain Over Time Average genetic gain over time for the crossbred (CROS), indicine (IND), and taurine (TAU) breeding programmes under various scenarios. Solid lines indicate the traits under selection, while dotted lines represent the selection index. Scenarios: additive only (A) in the top left, low dominance correlation (ADL) in the top right, medium dominance correlation (ADH) in the bottom left, and high dominance correlation (ADH) in the bottom right.

ments. Across all dominance scenarios, crossbreeding outperformed indicine in E_1 through E_4, while the indicine programme consistently yielded higher gains in E_5 and E_6.

4.3.2 Genetic variance

Figure 4.6 shows the trajectory of genetic variance in both breeding programmes across all dominance scenarios, highlighting the growing variability in individual responses, especially within the crossbreeding scheme.

Selection based on the average performance across environments in the tropical breeding programmes resulted in an overall increase in individual genetic diversity. The trend was especially pronounced in the crossbreeding programme, where the average individual genetic variance rose by 4.92 t/l in the final breeding cycle compared to the first cycle in the scenario with moderate environment dominance correlation (ADM). This scenario also produced the highest variance in the indicine programme (2.41 t/l).

In the indicine breeding programme, changes in genetic variance varied across dominance scenarios: 2.11 t/l (ADH), 1.53 t/l (A), and 0.11 t/l (ADL). Similarly, in the crossbreeding programme, the observed increases were 4.21 t/l (A), 2.7 t/l (ADH), and

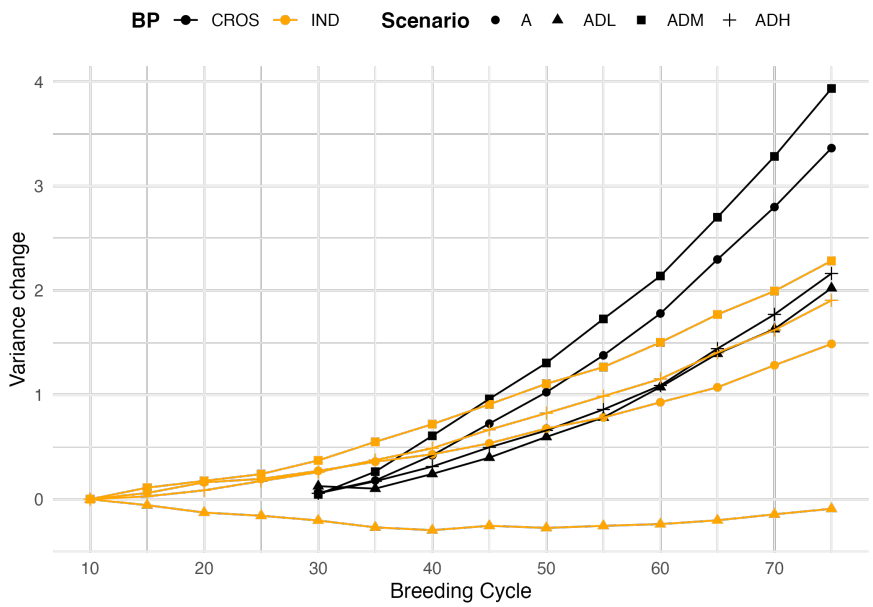

Figure 4.6: Genetic Variance Average individual genetic variance over time for the crossbred (CROS) and indicine (IND) breeding programmes across scenarios. Different shapes represent the four scenarios (A, ADL, ADM, and ADH).

2.53 t/l (ADL). These results indicate that genetic variance generally increased more in the crossbreeding programme than in the purebred indicine programme, particularly under additive or moderate dominance scenarios.

4.3.3 Genotype-by-environment framework

In this section, we apply our novel framework for investigating G×E interaction patterns. We start by describing the results of a conventional principal component (PC) rotation to illustrate its limitations. Then, using the taurine population as a case study, we demonstrate how the new *informed rotation*, which targets specific environments (E_1 in this case), improves interpretability and decision-making for crossbreeding.

Conventional principal component rotation

We began by applying a standard PC rotation to the environmental contributions obtained with the multiplicative model (Eq. 4.3). This approach aligns the environmental and genetic contributions with the directions (components) of maximal variance.

Table 4.4 summarises the variance decomposition across breeding programme and dominance scenario. For example, in the taurine population, the first two PCs explained over 96% of the total genetic variation under the no-dominance (A) scenario. However,

Table 4.4: Proportion of variance explained by each principal component (PC) in the taurine (TAU), indicine (IND), and crossbred (CROS) breeding programmes under four dominance scenarios (A, ADL, ADM, ADH).

BP	Scenario	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
TAU	A	0.86	0.10	0.02	0.01	0.00	0.00
	ADL	0.48	0.19	0.12	0.09	0.07	0.06
	ADM	0.57	0.18	0.13	0.06	0.04	0.02
	ADH	0.51	0.17	0.16	0.09	0.05	0.02
IND	A	0.63	0.28	0.06	0.03	0.00	0.00
	ADL	0.43	0.30	0.10	0.08	0.05	0.04
	ADM	0.40	0.25	0.15	0.08	0.06	0.05
	ADH	0.46	0.30	0.13	0.08	0.02	0.01
CROS	A	0.63	0.29	0.05	0.02	0.00	0.00
	ADL	0.60	0.31	0.06	0.02	0.00	0.00
	ADM	0.52	0.39	0.05	0.03	0.01	0.00
	ADH	0.70	0.25	0.04	0.01	0.00	0.00

while PC 1 explained 86% of the variance in scenario A, its contribution reduced to approximately 50% under scenarios with dominance structure (ADL, ADM, ADH). Similar trends were observed in the indicine and crossbred populations, although the reduction in the contribution of PC 1 was less accentuated (63% to 40% for indicine and 63% to 52% for crossbred populations).

Figure 4.7 shows that, for taurine bulls under scenario A, PC 1 explained most of the variation in environments E_5 (96%) and E_4 (93%), while E_1 and E_2 had greater contributions from higher order PCs (2, 3, and 4). Supplementary Table S2 provides complete observations for all breeding programmes and scenarios.

At the individual level, Supplementary Figure S7 demonstrates that bulls G_2, G_3, and G_5 have their performance largely determined by PC 1, while others such as G_1 show more balanced contributions.

Biplots further illustrate this complexity. In Figure 4.8, environmental loadings (blue) and individual scores (red) are shown across different factors, with selected bulls highlighted with bold lines. The black dotted line indicates the direction of the breeding objective- environment E_1. Panel (a) shows the first two factors, capturing the majority of the variation, but places E_1 near the origin, indicating these factors capture little variation in that environment (as shown in the previous results). Panels (b) and (d), which involve PC 3, 4, and 2, are the best at capturing the response to E_1. In these we

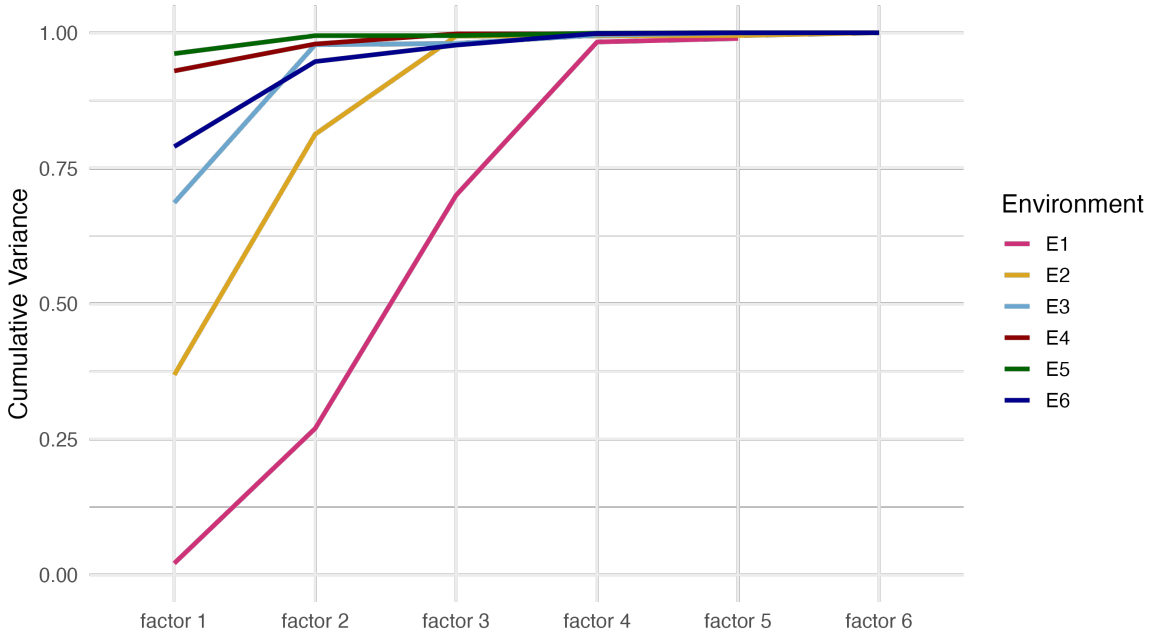

Figure 4.7: Cumulative variance explained by each principal component (PC) across environments in the taurine population under Scenario A (no dominance). Colours indicate individual environments; PC 1 explains most of the variation in E_5, followed by E_4, E_6, E_3, E_2, and E_1.

can see the selected bulls (G_1, G_2, G_3, G_4, and G_5) are well aligned with the breeding objective. Panel (c) reiterates the previous results, while also showing good responses of selected bulls to environments E_2 and E_3. Across all panels, unselected bulls (faded red lines) showed response in multiple directions, not only the direction of E_1.

Overall, standard PC rotation identified major directions of variation. However, the interpretation of these factors is limited, as they do not correspond to specific environments or breeding objectives.

Informed rotations

To better interpret response to selection in a target environment, we applied our new *informed rotation*. The approach concentrates all the variation correlated with the target environment (expressed as non-crossover G×E variance) in the first component, while uncorrelated (crossover G×E) variation is left orthogonally distributed across the higher order components.

Table 4.5 summarises the new decomposition of genetic variances with weight (1) fully placed on E_1. All non-crossover variance (4.08) is now capture by the first component. This leads to a predicted genetic gain is 2.02, which is a product of covariance with

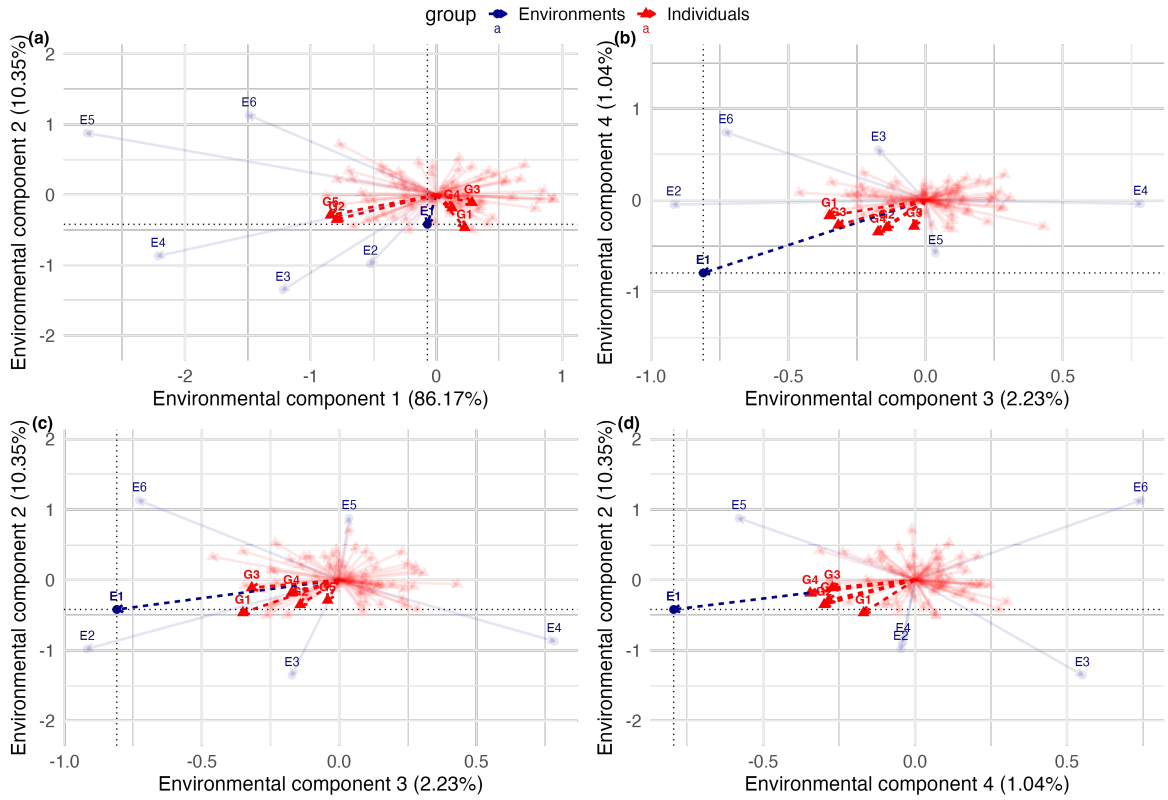

Figure 4.8: PCA bi-plots for the taurine population (Scenario A: no dominance). Bi-plots show environmental loadings (blue) and individual scores (red) across different environmental components. Dotted black line places the breeding objective. Panels (a) through (d) highlight variation involving environment E_1 and the top 5 bulls (dashed lines).

the main effects and the square root of the main effect variance.

Table 4.5: Decomposition of non-crossover and crossover variance across environments following rotation focused on E_1. Includes total variance, correlation with the main effect (E_1), implied genetic gain under selection in E_1 and weight place on each environment for rotation.

Environment	Non-cross. Var	Cross. Var	Total Var	Cor. Main Effect	Gain	Weight
E_1	4.08	0.00	4.08	1.00	2.02	1.00
E_2	6.16	6.28	12.43	0.70	2.48	0.00
E_3	4.46	31.59	36.04	0.35	2.11	0.00
E_4	2.47	85.10	87.57	0.17	1.57	0.00
E_5	0.88	132.52	133.41	0.08	0.94	0.00
E_6	0.04	46.70	46.74	-0.03	-0.20	0.00

For other environments, the correlation with the main effect is conditioned to the between-environment correlation (see Section Environmental Settings for pre-defined correlations), and so is the expected genetic gain. Thus, we observe a progressive

decay in the correlation with the main effects as we move from E_1 to E_6, with expected reduction in genetic value in E_6 due to its negative correlation with E_1. For the tropical environments E_3 to E_6, most of the total genetic variation is of the crossover type, meaning it is uncorrelated to the response to selection in E_1. The new distribution of variance across environments is given in Supplementary Figure S8.

The rotated bi-plot (Figure 4.9) reflects this structure, with E_1 fully explained by the first component (5.65% of the total variation across environments). The second component captured most of the uncorrelated variation (84.45% of the total variation). Bulls G_2 and G_5 remained strongly associated with multiple environments; other selected bulls showed limited responsiveness beyond E_1.

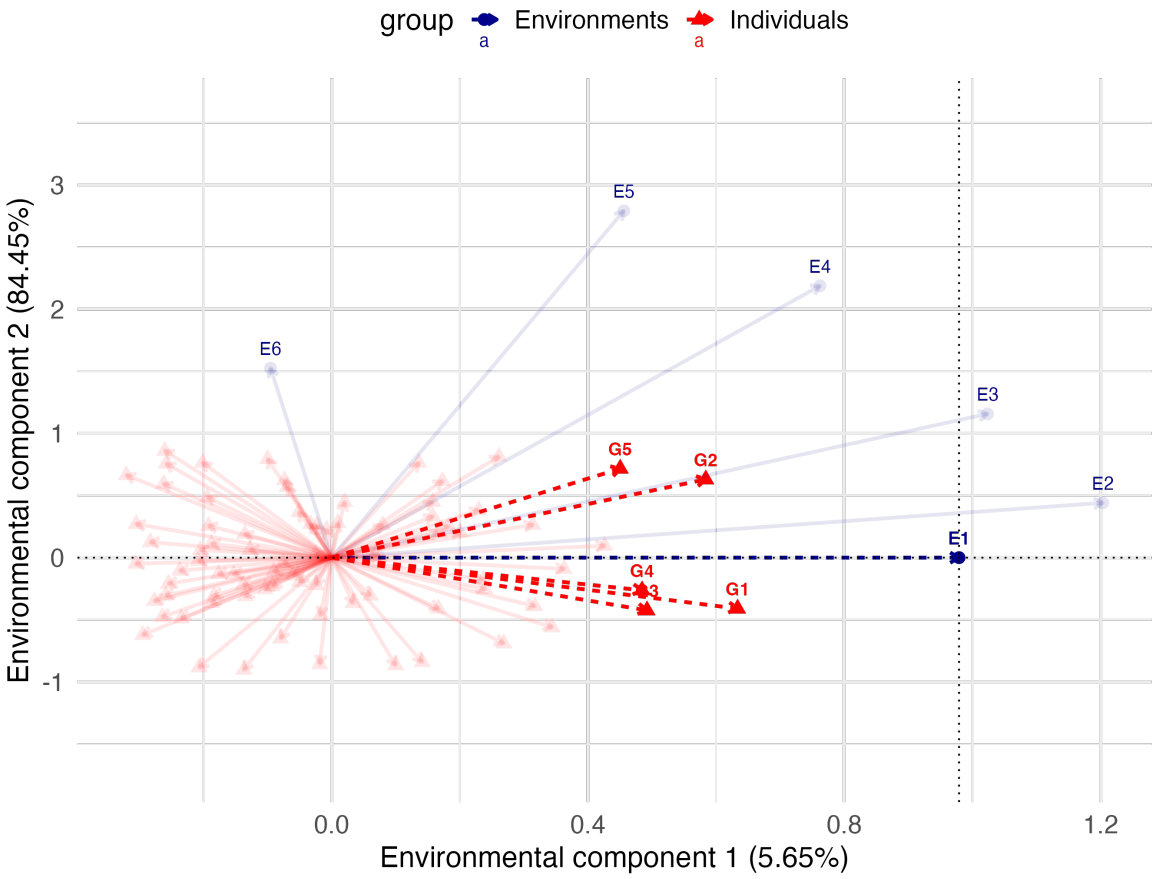

Figure 4.9: Informed rotation focused on environment E_1. Biplot of loadings and scores for taurine bulls under Scenario A. The first axis captures variance correlated with E_1; the second captures uncorrelated variation. Top 5 bulls shown in red with dashed lines.

Reaction norms (Figure 4.10), further illustrate expected and realised responses. The right panel gives environmental component 1 and shows expected responses for each

individual across the environmental gradient based on E_1-correlated variation (lines), and realised responses (points). Bulls G_2 and G_5 exhibited strong positive deviations from expected responses in most environments, while the other selected bulls had negative deviations in tropical environments.

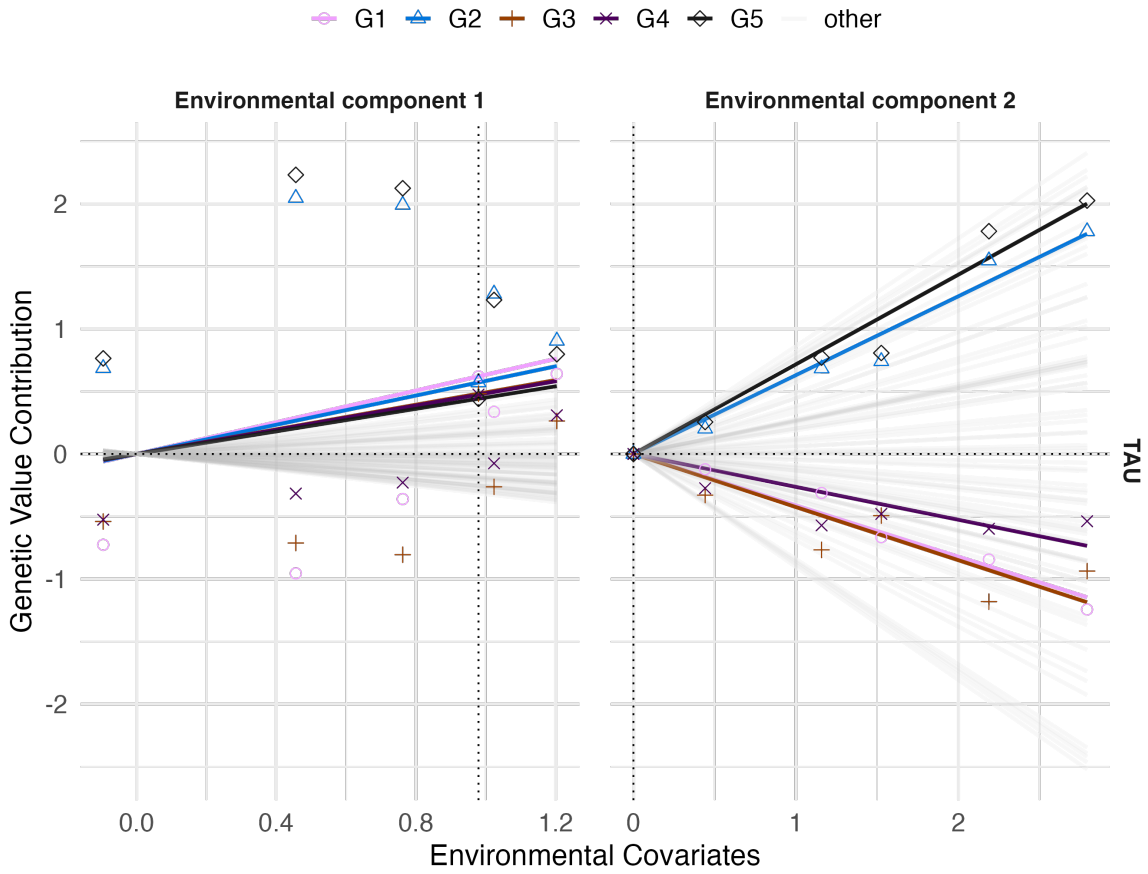

Figure 4.10: Reaction norms for the taurine bulls under E_1-focused rotation (Scenario A). Lines represent expected response across environments; coloured dots show deviations. Component 1 reflects E_1-correlated variation; component 2 captures uncorrelated (crossover) responses.

The left panel, environmental component 2, shows the uncorrelated responses, with E_1 centered at (0,0) since all variation is confined in component 1. In this panel there is a greater alignment between expected and realised responses, indicating that environmental component 2 is a better predictor of responses in tropical environments. Nonetheless, G_2 and G_5 continue to show strong positive responses, indicating they are well adapted to these environments.

The combination of the two panels, well illustrates the existence of useful genetic variation for crossbreeding within the taurine population, despite the selection focus

on E_1. Bulls G_2 and G_5 are *correlation breakers*, meaning they perform much better than expected in the tropical environments, making them ideal candidates for crossing.

These results provide a detailed account of G×E interactions, revealing how environments, genetic backgrounds, and dominance relationships shape genotype performance patterns. Genotypes selected under narrow or broad environmental objectives exhibit distinct patterns of adaptation and sensitivity, which can be effectively captured through rotated factor structures and genotype-specific reaction norms. These findings have important implications for breeding strategies, as further discussed in the next Section.

4.4 Discussion

The results of this study demonstrate that the success of tropical dairy crossbreeding programmes is highly variable and strongly influenced by environmental heterogeneity. In particular, the performance of taurine–indicine crossbreds is shaped by G×E interaction across diverse tropical production systems. Our results show that selection on mean performance, which assumes strong similarity across environments, affects optimal breed combinations and expression of heterosis. This reduces the predictability of crossbreeding outcomes and leaves genetic variation in environmental sensitivity largely untapped. Moreover, the proposed framework reframes G×E interaction relative to the breeding objective and provides interpretable measures of performance and adaptability. This perspective facilitates the identification of generalist and specialist individuals, informs the design of complementary crosses, and improves the predictability of crossbreeding responses in heterogeneous tropical systems. In the following, we discuss four main points arising from this work: (i) expectations for taurine–indicine crossbreeding under heterogeneous environments; (ii) the genetic and environmental drivers of variance in crossbred performance; (iii) strategic use of genetic variation to exploit G×E interaction; and (iv) the additional complexity introduced by dominance effects.

4.4.1 The expectations of taurine-indicine crossbreeding

When selecting a breeding strategy, farmers base their decisions on the expected improvement in performance. In tropical dairy systems, this often involves weighing the benefits of maintaining locally adapted populations against the perceived advantages of introducing exotic genetics. However, high-performing bulls from temperate environments often underperform in tropical environments (Buxadera and Mandonnet,

2006). One contributing factor is that breeding values of taurine bulls are generally not representative of their performance in tropical regions.

Falconer (1952) postulated that G×E interactions are pleiotropic effects of variants in different environments. This means that such interactions can be taken into account in genetic evaluations by treating the trait expressed in different environments as different correlated traits. This approach considers the performance of daughters of a bull in a range of environments, improving their relevance across production systems (Hayes et al., 2016). Although an international genetic evaluation for dairy cattle exists, it is largely centred in production systems of the Global North, limiting their utility for decision-making in tropical settings. Consequently, farmers in tropical regions using breeding values derived from temperate populations may operate under performance expectations that do not hold in their own production contexts. The review by Buxadera and Mandonnet (2006) demonstrates the gap between expectation and real performance. In terms of crossbreeding, this can lead to a general impression that the strategy is not beneficial.

Figure 4.4 illustrates this point by presenting the outcome of the taurine breeding programme in its original environment (TAU) and in tropical environments (TAU-TROP). The result demonstrates the common observation that crossbreeding does not meet expectations, but clarifies that this often stems from the erroneous initial assumptions that crossbred performance will lie midway between the performance of indicine (IND) and taurine (TAU) breeds. A more appropriate benchmark, however, is TAU-TROP, not TAU. When combined with the genetic gain trends presented in Figure 4.5, these results help unpack both the contribution of heterosis and the context-dependent value of crossbreeding in tropical systems. The results generally corroborate the conclusions of Kathambi et al. (2025), Marshall et al. (2019) and Madalena et al. (1990), which indicate that the benefits of crossbreeding are strongly influenced by environmental and management conditions. Under favourable conditions, crossbreeding is the best strategy because of the positive contribution of exotic genetics. In harsher environments, crossbred underperform due to the lack of fitness and adaptability of the taurine breed.

Taken together, this provides a well-supported, general understanding that crossbreeding outcomes are environment-dependent and that adoption requires careful alignment with the specific goals and constraints of the production system. However, while these are the most commonly considered aspects of crossbreeding decisions, they represent only part of the picture. In the following sections, we dive deeper into the role of G×E interaction in shaping individual variability and genetic responses, factors that further

complicate the decision-making process, but also offer new opportunities to optimise dairy breeding in tropical environments.

4.4.2 The drivers of genetic variance

Beyond mismatches in expectations, another key factor contribution to the variable success of crossbreeding strategies is an excess of genetic diversity. Crossbreeding introduces favourable alleles from exotic breeds, disrupts runs of homozygosity and reshuffles the genetic background. This process increases genetic variance, in contrast to purebreeding, which, through selection, fixes desirable alleles and thereby gradually depletes genetic variation as homozygosity increases. While the introduction of exotic genetics is a well-recognized source of diversity, our results show that it is not the only driver. In our simulations, we observed an increase in genetic variance over time in both crossbreeding and tropical indicine breeding programmes (Figure 4.6). This increase can be attributed to two mechanisms: (1) the introduction of novel alleles via crossbreeding, and (2) multi-environment selection based on mean performance.

First, the increase in genetic diversity in the crossbred population can be attributed to the disruption of linkage disequilibrium (LD). As demonstrated by Lara et al. (2022), genetic variance can be partitioned into genic variance and LD components. Specifically, LD within chromosomes and LD between chromosomes. Genic variance refers to the variance in allele substitution effects at individual loci; LD between chromosomes reflects covariances between unlinked loci, while within-chromosome LD captures covariances among (physically) linked loci. Directional selection often induces negative LD, especially between chromosomes, as certain allele combinations are favoured, reducing the efficacy of recombination and overall additive genetic variance, a phenomenon known as the Bulmer effect (Bulmer, 1971). Crossbreeding counteracts the Bulmer effect by disrupting allele associations, effectively releasing hidden genetic variance and generating positive LD.

The second driver of the increase in genetic variance observed in the crossbred and indicine breeding programmes is G×E interaction. In the presence of G×E interaction, the performance of an individual vary in different environments. When selection is based on mean performance across environments, individuals with variable responses, rather than stable performers, are inadvertently selected. Because mean performance masks variation in specific environments, this selection strategy promotes individuals with broader, more diverse responses. As a result, genetic variance increases, not only due to recombination but also due to an increase in divergence in how genotypes re-

spond to environmental variation. This culminates in the emergence of more diverse, but potentially less stable, individuals. As discussed by Mulder (2017), the excess of genetic variance can represent both an opportunity and a challenge for breeding programmes. Greater individual variation provides material for selection and may improve the adaptability of the population to changing or novel environments. However, increased genetic diversity complicates selection decisions and reduces the predictability of performance, which can be particularly problematic in resource-constrained settings, such as many tropical systems.

4.4.3 The strategic use of genetic variation

A critical constraint in such heterogeneous environments is the ability to interpret and manage genetic diversity in a meaningful way. The success of the breeding strategy depends on the ability to leverage useful variation. While in this study, we simulated six static environments, in reality, environments are dynamic. The changes can be thought of as macro-changes, going from temperate (intensive production systems) to tropical settings (semi- or extensive production systems) or micro-changes, such as seasonal changes in temperature or rainfall patterns within the farm. Or extreme environmental events becoming more frequent. Understanding the structure of G×E interaction is, therefore, essential to convert diversity into sustained genetic gain.

In this paper, we adapt tools from plant breeding to provide a framework for dissecting G×E patterns in tropical dairy cattle breeding. We use rotated factor analysis to explore genotype performance in multiple environments. Following Kempton (1984), we use biplots as an initial explanatory aid to identify the responses patterns of each individual in each environment before moving on to more formal decisions. We focus on the taurine breeding programme to understand how this framework can be used to identify useful genetic variation and improve its application within crossbreeding decisions. Nonetheless, important patterns are also observed in the indicine population as a consequence of selection in multiple environments. These results are compiled in the Supplementary Material Section C.2.

Figure 4.9 demonstrates that, despite intensive selection in environment E_1, bulls G_2 and G_5 exhibit strong adaptation to other environments, particularly environments E_2 and E_3. This suggests the presence of underutilized genetic variation that may be highly valuable for breeding programs targeting diverse production contexts.

The reaction plots in Figure 4.10 allow further interpretation of these patterns. The

first factor captures the correlated genetic response to selection in E_1. Interestingly, the genetic covariance between E_1 and environments E_2, E_3, and (less so) E_5 exceeds the genetic variance in E_1 itself. This implies that bulls G_2 and G_5 are expected to generate more genetic gain in those correlated environments than originally assumed. These individuals defy the expected pattern of performance derived from their ranking in the primary selection environment. They can be thought of as "correlation breakers", genotypes whose performance in secondary environments is not predictable from their performance in the target environment, making them prime candidates for crossbreeding in multi-environment systems.

Higher-order factors reflect uncorrelated or environment-specific responses. Here again, bulls G_2 and G_5 demonstrate advantageous profiles for tropical systems, indicating robustness not captured in single-environment evaluations. These patterns highlight the importance of decomposing G×E effects to uncover genotype-specific adaptations that are otherwise masked by aggregate performance measures and provide important insights into the potential of taurine genetics to retain substantial genetic diversity relevant to tropical environments, despite intense local selection.

The decomposition of additive G×E interaction and the integration of biplot and reaction norm analyses provide a practical framework for identifying genotypes with broad or specific adaptability. This framework offers tropical breeding programmes a means to harness latent variation and design strategies that are resilient to environmental changes. However, to fully meet the challenges of crossbreeding in tropical systems, we must also account for the additional variance structures and interaction patterns introduced by dominance as these patterns can reshape both expectations and breeding outcomes.

4.4.4 The extra complexity induced by dominance

Although we have focused our results primarily on additive genetic variation and its interaction with the environment, it is important to recognize that this is a simplistic approach. In multi-environment conditions, each variance component can independently be subject to environmental interaction (Hunt et al., 2020). In real-world crossbreeding decisions, non-additive effects such as dominance play a central role. However, dominance remains under explored in tropical dairy breeding, in part due to the assumption that dominance variance is negligible for most traits, but most importantly, due to the lack of large testing sets (i.e., large full-sib families where dominance effects can be estimated are not feasible in a species with biological limitation of one

offspring per season) (Varona et al., 2018). Thus, in dairy cattle, breeding dominance is generally excluded from genetic evaluation models as it only provides a marginal, if any, improvement in prediction accuracy in most temperate systems (Varona et al., 2018). Consequently, dominance effects are often absorbed into permanent environmental or residual components. However, our simulations suggest that this approach overlooks critical dynamics in crossbred populations and multi-environment systems, where dominance can materially influence both genetic gain and genetic variance.

Although we did not explicitly partition additive and dominance components, several patterns in our results indicate that dominance can modulate G×E interactions. For instance, differences in genetic gain across environments varied according to the degree of dominance environmental correlation (Figure 4.5). In terms of genetic variance (Figure 4.6), the scenario with a high dominance environmental correlation (ADH) led to a greater accumulation of diversity in purebred populations than in crossbred. Similarly, the structure of the principal components shifted with the inclusion of dominance effects: crossbred populations exhibited a stronger concentration of variance in the first component, while purebreds showed a more uniform distribution between components (Table S2). This happened because the different levels of dominance-by-environment correlation changed the expression of dominance degrees between environments and consequently individuals' breeding values. These findings corroborate the observation that G×E interaction influences crossbred or hybrid performance differently than in purebreds (Mulder, 2017; Betran et al., 2003).

Importantly, our results suggest that even if the dominance effect is negligible, the correlation structure between environments can lead to divergent genotype responses to changes in environment, and ultimately influence breeding outcomes. Dominance adds another layer of complexity to an already challenging decision-making landscape. In systems where G×E interaction effects already reduce predictability, dominance can further confound expectations by producing context-specific, non-linear responses. For breeders selecting high-yielding parents for crossbreeding, not accounting for these interactions can result in suboptimal breeding decisions. To address this, future work must go beyond additive models and begin to characterize the structure of the dominance component of G×E interaction in dairy systems. Doing so will require large-scale, structured datasets that integrate genotypic, phenotypic and environmental variation. Nonetheless, our results underscore the potential value of incorporating dominance-aware methods for environments where crossbreeding is prevalent and performance expectations are sensitive to local climate variability.

4.5 Concluding remarks: outcomes and implications

This study demonstrates that understanding and leveraging G×E interaction is critical for improving the predictability and effectiveness of crossbreeding in multi-environment tropical dairy systems. Selection strategies based solely on mean performance across environments obscure valuable environment-specific genetic responses, limit the effective use of existing genetic variation, and reduce the stability of crossbreeding outcomes. In contrast, the framework developed here shows that with appropriate tools it is possible to unveil otherwise overlooked variation within exotic populations and characterise how this variation contributes to performance and adaptability in tropical environments. This provides a principled basis for designing crossbreeding strategies that better exploit breed complementarity, maintaining robustness to environmental heterogeneity. The framework is especially relevant in the context of climate change, where extreme environments are becoming more common. The ability to optimise mating plans focusing on adaptability patterns support the choice of breeding strategies better aligned with current and emerging climatic conditions. Although the framework is directly relevant to breeding decisions, practical challenges for implementation must be highlighted. Application requires breeding values across environments, as well as sufficiently structured breeding programmes to connect performance records across diverse production systems. These requirements remain limiting in many tropical settings. Further work is therefore needed to integrate this framework with empirical crossbreeding data, genomic prediction and to test its utility for informed mate-allocation strategies. With this work, we hope to support a transition from reactive to proactive tropical breeding strategies; a critical shift for ensuring long-term adaptability, system resilience and sustained genetic gain, especially in the face of climate change.

Code availability

Simulation scripts for the breeding programme and the framework for investigating G×E interaction presented in this manuscript are available at GitHub repository `https://github.com/HighlanderLab/gmafrafortuna_girolando_sim`.

ORCID of Authors

Gabriela Mafra Fortuna 0000-0001-8921-642X
Gregor Gorjanc 0000-0001-8008-2787
Daniel Tolhurst 0000-0002-4787-080X

Funding

GMF and GG acknowledge funding from BBSRC DTP (EASTBio) CASE PhD studentship with Genus, BBSRC Institute Strategic Programme funding to The Roslin Institue (BBS/E/D/30002275, BBS/E/RL/230001A), BBSRC grants BB/T014067/1 and BB/M009254/1, and The University of Edinburgh. DT acknowledges funding through his Edinburgh Innovations Fellowship at The Roslin Institute.

5 General discussion

Dairy production has considerable socioeconomic importance worldwide and plays a critical role in addressing food insecurity and poverty, especially in tropical regions. However, dairy farming in these settings faces the persistent challenges of slow genetic progress and high vulnerability to climate change. To address these constraints, crossbreeding environmentally adapted local breeds with exotic intensive breeds is common practice. While crossbreeding can provide complementary advantages, outcomes are highly inconsistent. This inconsistency reduces stability of crossbred performance, complicating breeding decisions, and limiting the long-term optimisation of tropical production systems.

This thesis demonstrates, through the application and development of novel statistical methods in the context of dairy breeding, how genomic variation can be strategically leveraged to stabilise and increase performance of indicine-taurine crossbred cattle in diverse environments.

The thesis is structured around three core chapters, each addressing a distinct yet interconnected aspect of the tropical crossbreeding problem:

- Chapter 2: uses ancestral recombination graph (ARG) to encode genetic diversity in global cattle populations, demonstrating the general utility of ARGs for local ancestry inference and data compression in livestock genomics.
- Chapter 3: introduces a novel ARG-based statistical model for estimating haplotype and mutation effects within non-recombining genomic regions. This model captures ancestry-specific effects by incorporating the genealogical structure of the genome.
- Chapter 4: uses stochastic simulations to explore genotype-by-environment interaction (G×E) in tropical dairy breeding. A new framework based on multiplica-

tive models is proposed to investigate G×E for improved selection in variable environments.

The work presented addresses three fundamental challenges in tropical crossbreeding programmes: (i) the genomic distance between indicine and taurine breeds, which complicates the predictability of crossbred performance; (ii) the lack of statistical methods capable of capturing the complex genomic architecture of undercharacterised and admixed populations; and (iii) the widespread neglected impact of G×E interaction in tropical breeding schemes, which contributes to the instability of crossbred performance across environments.

The following sections build on the results from each chapter to critically examine these issues, discussing their broader implications for sustainable and climate-efficient dairy breeding in the tropics.

5.1 Genomic distance of indicine and taurine cattle

Indicine and taurine cattle diverged over thousands of years ago, followed by independent domestication processes and distinct breeding history. This deep evolutionary split has resulted in extensive genomic, physiological, and phenotypic differentiation. Crossbreeding these two subspecies introduces a large number of segregating genome variants, many of which may interact in non-additive and environment-dependent ways.

From a breeding perspective, the genomic divergence poses both challenges and opportunities. Crossbreeding creates complex mosaic patterns of ancestry due to sustained recombination, directional selection, and drift. These patterns are often non-random, reflecting natural and artificial selection for genomic segments that provide local environmental advantages (McHugo et al., 2025; Paim et al., 2020). This mosaicism enables the creation of composite animals that combine genes of interest from both ancestries, reflecting favourable phenotypic responses. Equally, the extent of divergence increases the likelihood of epistatic incompatibilities, ancestry-specific mutations effects, and G×E interactions (Vandenplas et al., 2016), complicating the prediction and selection of best performing individuals.

To better understand and manage the complexities introduced by the genomic distance between indicine and taurine cattle, **Chapter 2** explores the use of ancestral recombination graphs (ARGs) to reconstruct the genealogical history of cattle genomes. By modeling the recombination and mutation history of genomic sequences, ARGs allow

for fine-scale local ancestry inference in admixed populations (Nielsen et al., 2025; Wong et al., 2024). The results demonstrate that ARG inference via the tree sequence methodology can capture intricate patterns of ancestry, while improving data compression and computational efficiency of downstream analyses.

Nevertheless, some limitations must be acknowledged. The accuracy of demographic and ancestral inference showed sensitivity to inference quality, which is affected by assumptions often violated by the dynamics of livestock population genomics and data quality. Moreover, although tree sequences are efficient for storage and analysis, some inference steps remain computationally demanding, limiting their use. A particularly important limitation to this thesis context is the lack of ARG-inference methods modeling structural variants (SVs). In cattle, SVs are commonly associated with differences in the expression of heat tolerance, disease resistance, fertility and production traits (Naval-Sánchez et al., 2020; Braga et al., 2024). However, even when ARGs are inferred from bi-allelic SNPs, they contain signatures of SVs (Ignatieva et al., 2025).

Beyond ancestry inference, ARGs provide a foundation for modelling the evolutionary forces acting on admixed genomes. ARGs can be used to identify regions under selection, detect introgression and quantify ancestry-specific contributions to phenotypic variation. ARG-based methods offer scalable and biologically informed approaches for analysing genomic variation in crossbred populations, although its application remains largely experimental in livestock contexts.

5.2 Current statistical methods fail in accounting for the genomic distance between indicine and taurine cattle

Standard genomic prediction models rely on the assumption that SNP effects are transferable across generations and even populations, provided linkage disequilibrium (LD) persists between markers and underlying causal variants (Meuwissen et al., 2001). However, differences in population genetics processes cause LD to vary across ancestries. When coupled with non-additive genetic effects and G×E interaction, SNP effects may vary between populations, posing a challenge to multi-breed and crossbred genomic evaluations and selection.

Chapter 3 introduced a novel ARG-based method for genomic prediction that explicitly incorporates the genealogical structure embedded in whole-genome sequence

data. Unlike conventional methods that estimate SNP effects independently of their evolutionary context, the proposed model estimates these effects conditioned to their position within a local tree DNA. This informs the model of the pattern of inheritance of mutations across ancestries, capturing ancestry-specific effects.

The new approach leverages the local tree DNA topology to construct a recursive and sparse approach via conditional distributions of ancestor-descendant haplotype values. This allows efficient calculations using the generalized Cholesky decomposition of variance and precision matrices between haplotype values, significantly reducing computational demand when using genomic data. The approach was evaluated using cattle mitochondrial DNA (mtDNA), a non-recombinant genomic region, and associated real and simulated phenotypes. The results yield accurate predictions with a substantial reduction in computational demand compared to standard genomic prediction methods.

However, key limitations remain. The method is currently restricted to non-recombining genomic regions, with a single local DNA tree. Working with recombining genomic regions involves an ARG with multiple local trees along the genome and changing relationships between haplotypes. This requires a more complex approach to model the distribution of ancestral-descendent effects while retaining the sparse characteristics of the model (Lehmann et al., 2025; Lee et al., 2025).

Despite the limitations, the work presented in **Chapter 3** establishes a conceptual and computational foundation for integrating ARG into quantitative genetics and breeding, providing a proof-of-concept for leveraging WGS data and retaining their biological information. This is a current area of research that, once the computational barriers are overcome, has the potential to enable more accurate multi-breed genomic predictions.

5.3 Neglected G×E and the role it plays in the instability of crossbred performance across tropical environments

Genotype-by-environment (G×E) interaction is a well-recognised phenomenon that affects trait expression in animals and plants, reducing response to artificial selection when ignored. Despite its importance, tropical dairy breeding continues to overlook it. While indicine-taurine crossbreeding increases genetic variation at the population level through the reshuffling of alleles from distinct ancestries, unaccounted G×E interaction

further increases genetic variance across environments at the individual level. The poor management of G×E interaction leads to instability in crossbred performance, undermining the success of breeding strategies.

Chapter 4 directly addresses this limitation by challenging the conventional marginalization of G×E interaction in tropical dairy breeding. Using stochastic simulations, an effective method to compare hypotheses and assess long-term breeding strategies (Gaynor et al., 2021), the chapter demonstrates how G×E interaction contributes to the variability observed in crossbred performance, confirming its environmental dependency. Importantly, it highlights the misconception that the performance of the exotic breed in its native temperate environment sets an appropriate benchmark for tropical crossbred performance. In reality, milk yield in the tropics most likely constitutes different traits due to G×E interaction. As a result, crossbred performance should be evaluated within these local environmental contexts. This distinction is essential for grounding expectations and guiding tropical crossbreeding programmes towards more realistic solutions.

Addressing this insight involves a paradigm shift. By reframing G×E interaction as a central variable in breeding strategies and as a source of useful, untapped genetic variation, it can be leveraged to inform more resilient breeding decisions. To support this shift, the chapter provides a framework based on multiplicative models and principal component rotations (Smith et al., 2015; Smith and Cullis, 2018; Tolhurst, 2024), which allows breeders to quantify, visualise and exploit G×E patterns to promote tropical crossbreeding. The framework is summarised in the following three steps:

1. Reformulates individual genotypes into a set of multiplicative terms comprising the product of unobserved (latent) environmental covariates and genotypic slopes. This decomposition captures the shared environmental structure implicit in the environmental specific breeding values without requiring direct measurement of all between-environment covariates.

2. Rotates the environmental covariates, isolating components that correspond specifically to the environment(s) where selection was performed. This rotation is analogous to a principal component rotation, with the exception that now the first component captures all the correlated response to the selection environment(s) and the higher order components capture the uncorrelated response.

3. Uses bi-plots and reaction norm plots to interpret genotype-specific sensitivities and responses to latent environmental dimensions, enabling informed breeding

decisions. These complimentary tools provide the graphical basis to investigate our proposed "what wins where, given selection elsewhere" framework.

The application of the framework to the simulated data reveals adaptability patterns that explain why some genotypes perform unpredictably across environments. This application transforms G×E interaction from an abstract statistical complication into a strategic resource, allowing breeders to navigate performance variability and utilise specifically adapted individuals. While the framework shows strong explanatory power, its full utility in breeding programmes depends on the availability of multi-environment phenotypic data and good connectivity between individuals across environments. This is a particularly critical consideration in tropical settings where extensive pedigree and phenotype recording is often unreliable or unavailable, and the use of genomic data is still restricted.

Moreover, the chapter offers insights into the practical application of the framework by describing explanatory analysis of genomic variation with bi-plots and reaction norm plots. The primary recommendation is the use of the framework to identify complementary mating pairs. The use focuses on aiding mate allocation strategies to promote yield, trait stability, and environmental plasticity. Ongoing research is developing tools for automating this identification. This addition will equip breeders with actionable information to support the design of crosses that are both productive and stable, ultimately addressing the problem of performance instability in crossbreeding by making G×E patterns operational in breeding decisions.

5.4 Concluding remarks

Crossbreeding has the potential for supporting the sustainable development of dairy systems of the Global South. However, the success of crossbreeding programmes requires tools and strategies tailored to the specific challenges faced in these areas. This involves the promotion of under-characterised local populations and the strategic use of genomic variation available from highly productive exotic breeds. It is not about a "one fits all" solution but a "what wins where" approach.

This thesis contributes to this perspective by addressing three interconnected challenges in tropical dairy breeding: the *characterisation* of genetic variation arising from indicine–taurine divergence, the *prediction* of genetic effects in heterogeneous genomic and environmental landscapes, and the optimisation of *breeding decisions* through explicit consideration of G×E interaction. Together, these components emphasise the

importance of understanding not only how much genetic variation exists, but how it can be leveraged effectively. While the methods presented in this thesis contribute to a better understanding of genomic diversity and its strategic use in tropical dairy breeding, their transition into piratical breeding settings is not straightforward.

Ancestral recombination graphs represent a promising frontier for studying ancestry and genetic variation; however, their application in animal breeding remains constrained by inference complexity, model assumptions and data quality. Inferring ARGs for large, structured breeding datasets poses substantial computational challenges that are still to be fully resolved. Moreover, the strong population structure and intensive directional selection characteristic of agricultural populations violate many assumptions underlying current ARG models, potentially compromising inference accuracy. Although rapid methodological advances are being made, the routine adoption of ARG-based approaches in breeding programmes is likely a long-term prospect.

On the other hand, the proposed G×E interaction framework is closer to transferability to practice. Similar modelling approaches are already well established and indispensable in plant breeding, and the framework developed here complements these methods extending their applicability into animal breeding contexts. Due to climate change, traits related to heat tolerance and environmental adaptability are gaining importance. G×E interaction is increasingly recognised as a critical challenge in animal breeding, particularly in tropical dairy systems. In this setting, our framework provides a practical means to move beyond naïve crossbreeding towards the intentional pairing of complementary individuals, whether in the context of crossbreeding or other breeding practices. Nevertheless, important barriers remain. Many countries in the Global South face limited resources, and structured breeding programmes with comprehensive pedigree and performance recording are uncommon. Strengthening data collection, pedigree recording, and programme infrastructure is a necessary first step, and a prerequisite for the effective uptake of more advanced quantitative genetic tools.

Ultimately, the sustainable improvement of tropical dairy systems will depend on the ability to leverage useful genetic variation, within and between populations, in ways that align breeding objectives with local conditions and future challenges. As climates continue to change and extreme environments become more frequent, breeding strategies that anticipate environmental responses rather than react to them will be essential. The approaches developed in this thesis aim to contribute to this transition, supporting more resilient, predictable, and context-aware tropical crossbreeding programmes.

A Supplementary materials: Global cattle genealogy inferred from ancestral recombination graphs

A.1 Supplementary methods

A.1.1 Test data

A testing set containing only Holstein samples was extracted from the 1KB chromosome 25 and filtered using bcftools (Danecek et al., 2021) retaining only sites variable in the reduced dataset. We refer to this dataset as **HOL**.

We simulated genotypes for a single cattle population of 400 individuals using *stdpopsim* (Adrion et al., 2020) *msprime* (Baumdicker et al., 2022) and the cattle demographic model described by MacLeod et al. (2013). The software outputs a tree sequence object (we will refer to this output throughout the text as **SIM-TRUE**), which we converted to VCF format using *tskit* (Wong et al., 2024) and used it to evaluate tree sequences inference. We refer to the simulated VCF file as **SIM-HOL**.

Simulations with *stdpopsim* and *msprime* used the Jukes-Cantor model of mutations at discrete sites. This model assumes uniform mutation rate for all base pair changes and equal likelihood of all ancestral states. As a result, with a large number of sites, most are expected to be bi-allelic, with a single mutation per site, althought a small number of sites with multiple mutations can occur. These assumptions make SIM-TRUE approximately comparable to the filtered HOL dataset.

A.1.2 Inference parameter configuration

By default, *tsinfer* assumes the infinite-sites mutation model, meaning that sites are expected to have at most one mutation. Once the `mismatch_ratio` parameter is used, multiple mutations at a single site are permitted. We used recombination rate of 1e-8 and mismatch ratio of 1 (default when recombination rate is provided).

Tree sequence inference for the SIM-HOL, HOL, and 1KB datasets was conducted under four configurations to assess the impact of different parameters on the final output. The configurations varied in how the `recombination_rate` and `mismatch_ratio` parameters were specified during the`match_ancestors` and `match_samples` steps:

- **def:def**: default parameters in both steps,
- **def:rec**: default parameters in `match_ancestors` and `recombination_rate` of 1e-8 and `mismatch_ratio` of 1 in `match_samples`,
- **rec:def**: `recombination_rate` of 1e-8 and `mismatch_ratio` of 1 for `match_ancestors`

and default parameters in `match_samples`, and

- **rec:rec**: `recombination_rate` of 1e-8 and `mismatch_ratio` of 1 in both steps.

Depending on the used parameters, we observed inferred tree-sequences with a proportion of sites (usually $< 5\%$) with more than 100 mutations ("multi-hit" sites), which impacted downstream analyses.

To mitigate the influence of such sites, we applied two post-processing strategies. In the first strategy, we masked and remmaped multi-hit sites. For this, all sites with more than two inferred mutations in the tree sequence were masked (removed) prior to re-inference. Following inference, these sites were reintroduced using the `map_mutations` function in *tskit*, which maps mutations to a local tree branches using the Hartigan parsimony algorithm. If the remapping required more than 5 recurrent mutations to fit the tree topology, the mutation was removed to limit the reintroduction of the number of ambiguous or potentially low-quality sites. In the second strategy for limiting the number of erroneous mutations we completely remove sites with more than two mutations from the data and re-phase the cleaned dataset prior to re-inference. All strategies were evaluated on the HOL dataset by comparing the resulting number of mutations per site and instantaneous coalescence rate estimate of effective population size.

A.2 Supplementary results

A.2.1 Computational constraints

We were unable to run all configurations for all datasets due to high memory requirements. In particular, *rec:def* and *rec:rec* were computationally prohibitive when applied to the 1KB dataset.

A.2.2 Dataset summary

Table S2 summarizes the input dataset sizes, number of variants, and resulting tree sequence data sizes. Table S3 details the number of trees, nodes, edges, and mutations under each configuration and processing strategy.

Table S1: Summary of time-resolved tree sequences

	Number					Size (GB)			
chr	sites	trees	nodes	edges	mutations	ts	ts (compressed)	raw vcf (compressed)	filtered vcf (compressed)
28	**28,354,108**	**26,922,456**	**73,954,207**	**1,000,283,211**	**28,353,786**	**46.07**	**9.9**	**874**	**8.4**
1	1,859,870	1,758,514	5,004,201	69,445,661	1,859,851	3.18	0.66	58.15	0.51
3	1,515,888	1,441,262	3,794,298	49,153,010	1,515,871	2.29	0.52	43.60	0.42
4	1,355,816	1,290,485	3,375,017	41,972,515	1,355,807	1.97	0.45	45.70	0.37
5	1,269,801	1,207,212	3,161,835	39,856,521	1,269,787	1.87	0.43	43.60	0.35
6	1,349,571	1,284,488	3,362,615	42,164,682	1,349,563	1.98	0.45	45.00	0.37
7	1,234,896	1,170,415	3,363,039	48,153,477	1,234,882	2.19	0.45	38.72	0.34
8	1,294,637	1,222,991	3,492,996	48,286,922	1,294,628	2.21	0.46	40.31	0.35
9	1,347,667	1,282,485	3,339,480	41,242,012	1,347,655	1.94	0.45	39.37	0.36
10	1,301,766	1,240,657	3,274,599	41,800,689	1,301,755	1.95	0.45	37.25	0.35
11	1,320,195	1,248,729	3,560,697	50,058,856	1,320,182	2.28	0.47	37.40	0.36
12	1,110,532	1,052,308	3,005,635	43,050,525	1,110,526	1.96	0.40	36.00	0.30
13	1,066,688	1,017,328	2,697,046	34,522,158	1,066,680	1.61	0.37	29.00	0.28
14	1,068,421	1,011,526	2,885,741	39,933,483	1,068,412	1.83	0.38	29.20	0.24
15	1,108,037	1,049,619	3,019,162	43,514,473	1,108,017	1.97	0.40	33.50	0.31
16	779,328	737,549	2,110,162	29,683,732	779,318	1.35	0.28	28.90	0.21
17	865,861	817,662	2,308,794	32,698,251	865,836	1.49	0.30	27.31	0.24
18	746,654	708,012	2,047,741	30,539,726	746,643	1.38	0.30	23.60	0.21
19	781,443	746,114	2,014,200	27,915,891	781,417	1.28	0.30	21.84	0.21
20	823,793	782,723	2,042,264	25,882,294	823,789	1.21	0.27	26.63	0.22
21	835,733	789,841	2,242,505	30,987,518	835,727	1.42	0.30	25.90	0.22
22	830,972	788,258	2,249,910	31,031,312	830,964	1.42	0.30	21.40	0.22
23	625,429	595,995	1,706,435	27,099,741	625,417	1.21	0.24	25.53	0.18
24	904,150	863,059	2,259,198	28,110,851	904,143	1.32	0.30	15.43	0.24
25	576,880	550,570	1,469,260	20,479,936	576,866	0.94	0.20	15.43	0.16
26	558,009	529,603	1,500,423	20,948,979	558,005	0.96	0.20	19.60	0.15
27	611,836	584,213	1,525,006	19,220,180	611,825	0.90	0.20	19.00	0.16
28	598,906	567,105	1,601,343	22,024,497	598,901	1.01	0.20	18.00	0.16
29	611,329	583,733	1,540,605	20,505,319	611,319	0.95	0.21	20.26	0.17

Table S2: Summary of dataset properties and output size

Dataset	VCF (MB)	Compressed VCF (MB)	Variant Sites ($\times 10^3$)	Samples	Ancestors ($\times 10^3$)	SampleData (MB)	AncestorData (MB)
1KB	12,288	245	1,704	1,838	1,185	270	9,625
HOL	573	24	372	393	256	36	225
SIM-HOL	236	14	151	400	23	12	36

Table S3: Tree sequence inference summary by configuration and strategy.

Dataset	match_ancestors	match_samples	Sites ($\times 10^3$)	Trees ($\times 10^3$)	Nodes ($\times 10^3$)	Edges ($\times 10^3$)	Mutations ($\times 10^3$)	Uncompressed (MB)	Compressed (MB)
1KB8	def	def	1,704	1,212	1,978	23,996	3,245	1,229	217
	def	rec	1,704	1,012	1,560	9,529	13,364	997	129
HOL	def	def	372	265	435	3,288	600	165	28
	def	rec	372	220	325	1,297	1,904	128	18
	rec	def	372	246	431	3,221	863	170	26
	rec	rec	372	115	223	898	2,092	115	15
SIM-TRUE	–	–	151	85	65	223	152	16	4
SIM-HOL	def	def	151	25	33	72	151	14	1.1
	def	rec	151	25	33	72	151	14	1.1
	rec	def	151	12	28	55	157	14	1.1
	rec	rec	151	11	28	54	157	14	1.1
HOL_masked	def	def	356	263	467	3,574	381	190	28
HOL_remapped	def	def	364	263	467	3,574	396	190	28
HOL_rephased	def	def	337	211	361	2,616	251	138	21

A.2.3 Observations

In the HOL dataset, using the default configuration, nearly 90% of sites had a single inferred mutation. However, sites with more than 100 inferred mutations were also observed, suggesting potential data quality or model-mismatch issues. Introducing mismatch parameters (*def:rec* or *rec:rec*) reduced the proportion of single-mutation

sites as expected, but did not eliminate high-mutation outliers, and increased number of inferred edges, indicating more recombination.

Masking and remapping reduced mutation counts at multi-hit sites, but not completely. Re-phasing after completely removing sites with >2 inferred mutations reduced all mutations at sites to one and improved data compression. These results suggest that sites with multiple inferred mutations in real data likely represent errors and significantly impact inference quality.

A.2.4 Inverse coalescence rate

The inverse coalescence rate estimates under different configurations were compared to the reference cattle demographic model from (MacLeod et al., 2013). In SIM-HOL, all configurations recovered the expected demographic pattern (Figure S2), supporting the inference validity with clean data. In the HOL dataset, however, excess mutations distorted coalescence time estimates, particularly when mismatch parameter was specified (Figure S3), confirming the impacts of data errors on tree sequence inference.

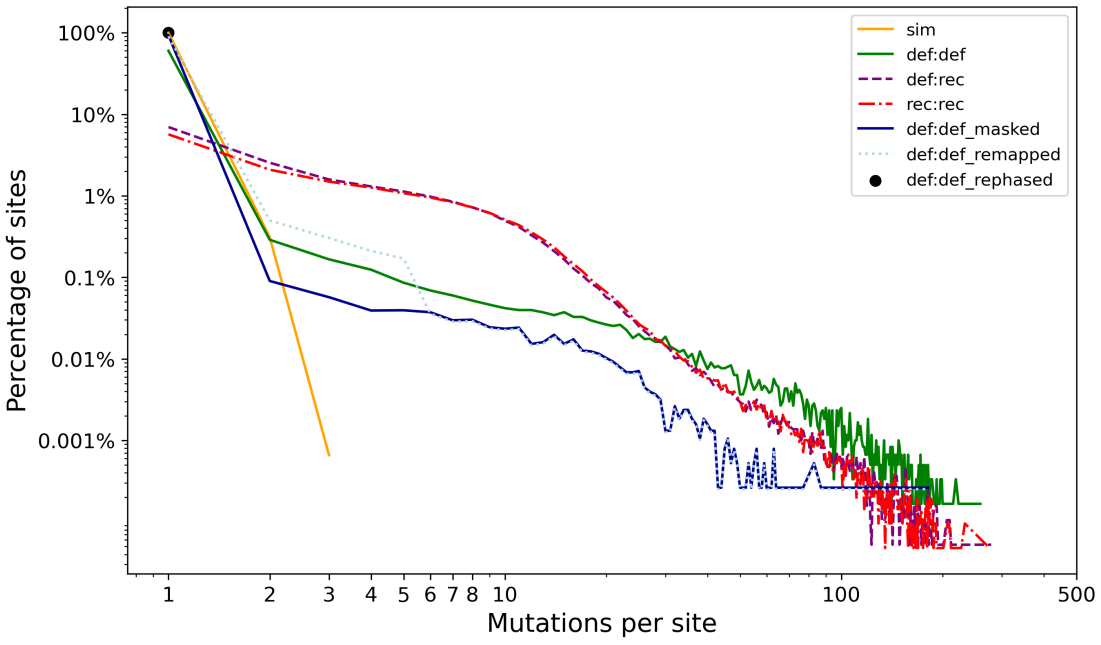

Figure S1: Distribution of mutations on the genome. Percentage of sites (y-axis) per number of mutations per site (x-axis) for the different inference scenarios (def:def, def:rec, rec:rec, def:def_masked, def:def_remmaped, and def:def_rephased) using the HOL dataset. The trend for the true simulated dataset (sim) is shown in yellow.

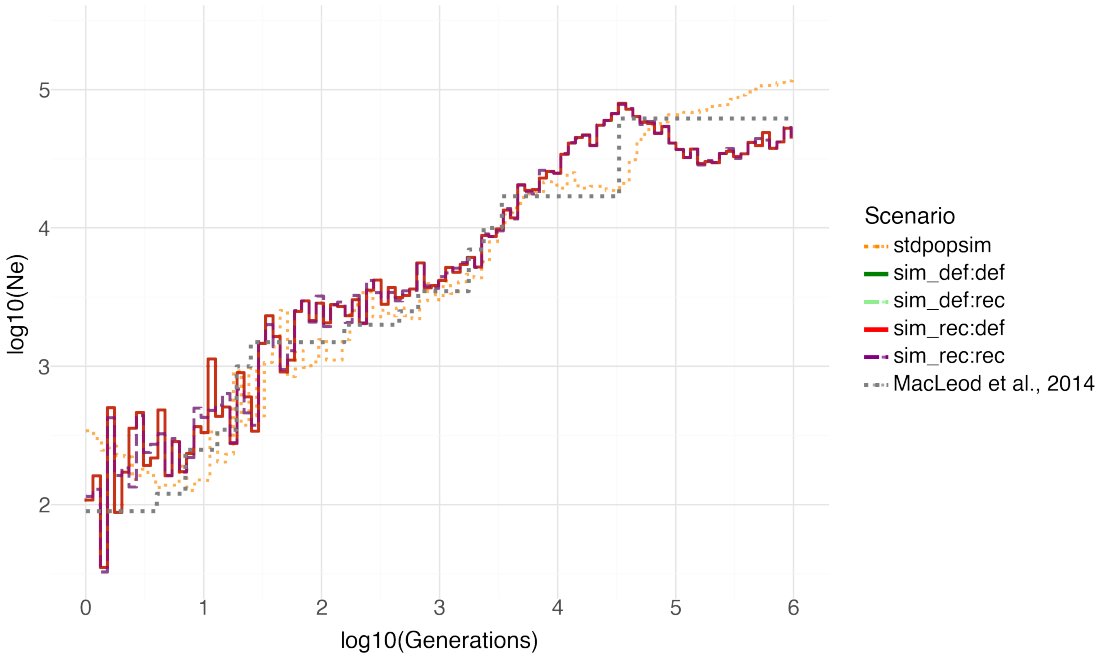

Figure S2: Inverse coalesce rate by generation for the simulated dataset SIM-HOL under different configurations. The log of the inverse coalescence rate (ICR) is shown in the y-axis and log time in generations in the x-axis. The trend for the true simulated tree sequence is shown in yellow, while the trend for the different inference scenarios are shown in grey.

Table S4: Benchmark summary for tree sequence inference for chromosome 1

Task	Runtime (h:m:s)	Max RSS (GB)	Max VMS (GB)	CPU Time (h)	Mean Load
generate_samples	1:01:19	0.51	6.44	1.02	99.97
generate_ancestors	1:03:43	18.24	1066.35	16.36	1540.15
truncate_ancestors	0:31:33	150.34	406.89	0.41	77.11
match_ancestors	13:43:57	11.99	20.21	207.99	1506.49
match_samples	13:03:12	69.08	78.04	193.59	1483.07
date	9:39:39	16.24	19.38	9.70	100.11

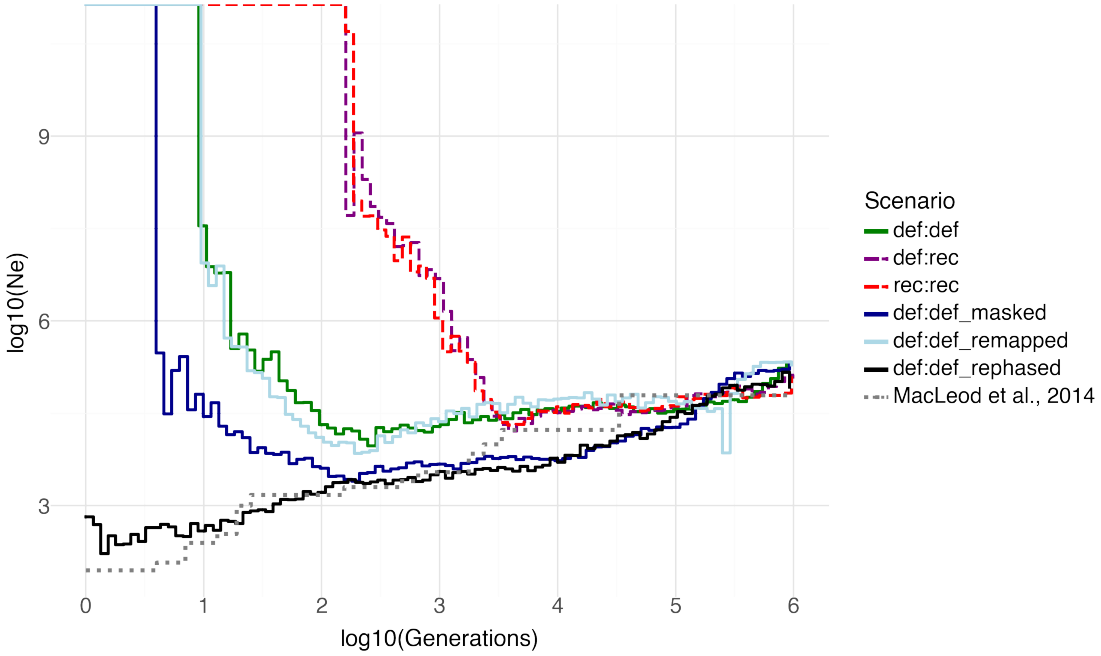

Figure S3: Inverse coalesce rate by generation as equivalent to effective population size. Colours represent the different optimization scenarios, in yellow is presented the trend for the true simulated tree sequence, in grey is the reference demographic model for cattle from (MacLeod et al., 2013) and in black is the final optimized tree sequence.

B Supplementary materials: Estimating haplotype values and mutation effects in the context of a local DNA tree

B.1 Small example and demonstration of the modelling approaches

In this section, we demonstrate the model of haplotype values on a local DNA tree with a small example. This example emphasises the potential of estimating mutation effects in their genome sequence context when we can polarise observed alleles into ancestral and derived mutations and we know the local DNA tree with corresponding mutations between haplotypes.

Consider a hypothetical local DNA tree in Figure 3.1, which describes the relationship among nine haplotypes across two clades (each denoted with a colored box representing two populations) and their most recent common ancestor (root) haplotype. These haplotypes span one codon in a protein-coding region of DNA, hence each haplotype is represented by three nucleotides that encode an amino acid. The root haplotype (H1) has the codon GCC, which encodes the amino acid alanine (ALA). The tree shows a hypothetical evolutionary history of the haplotypes over a long time period, where mutations changed the codon sequence and the corresponding amino acid. For example, H3 inherited from H1 with GCC codon (alanine), but mutation 1 at site 1 has changed the codon to CCC (proline). In total, we have ten haplotypes separated by eight mutations, and these haplotypes encode four different amino acids listed in

Table S1: Haplotype information for the small example. Information on the ten haplotypes in the local DNA tree (Haplotype) from Figure 3.1, their immediate ancestor haplotype (Ancestor), ancestral/derived allele encoding at the three nucleotides (A1, A2, and A3), amino acid (Amino Acid: ALA - alanine, PRO - proline, GLY - glycine, ARG - arginine), occurrence of a mutation since the ancestor (Mutated), mutated site (Site), mutation (Mutation), mutation effect (Effect), and haplotype value (Value).

Haplotype	Ancestor	A1	A2	A3	Amino acid	Mutated	Site	Mutation	Effect	Value
H1	/	0	0	0	ALA	0	/	/	/	0
H2	H1	0	0	0	ALA	0	/	/	/	0
H3	H1	1	0	0	PRO	1	1	m1	1	1
H4	H2	0	0	1	ALA	1	3	m2	0	0
H5	H2	0	1	0	GLY	1	2	m3	3	3
H6	H3	1	1	0	ARG	1	2	m4	1	2
H7	H3	1	0	1	PRO	1	3	m5	0	1
H8	H6	1	1	1	ARG	1	3	m6	0	2
H9	H7	1	1	1	ARG	1	2	m7	1	2
H10	H5	0	1	1	GLY	1	3	m8	0	3

(Table S1): alanine (ALA), proline (PRO), glycine (GLY), and arginine (ARG). The example is deliberately extreme with respect to the amount and type of mutations, to emphasise the potential for modelling. The 10×3 allele haplotype matrix $\mathbf{X}_h$ (columns A1, A2, and A3 in Table S1) and the 10×8 mutation haplotype matrix $\mathbf{W}$ for this example are:

$$\text{(B.1)} \qquad \mathbf{X}_h = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{W} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

For example, H3 has inherited mutation 1, H6 has inherited mutations 1 and 4, and H8 has inherited mutations 1, 4, and 6 (Figure 3.1). The relationship between $\mathbf{X}_h$ and $\mathbf{W}$ in this example is such that adding up columns of $\mathbf{W}$ for mutations at the same site gives the corresponding site columns of $\mathbf{X}_h$. For example, there is one mutation at site 1 (mutation 1), so the first columns of $\mathbf{X}_h$ and $\mathbf{W}$ are the same. There are three mutations at site 2 (mutations 3, 4, and 7), so adding up these columns of $\mathbf{W}$

gives the second column of $\mathbf{X}_h$. In the case of reverse mutations, this relationship does not hold for the used encoding. Coding reverse mutations in $\mathbf{W}$ as -1 would allow for the relationship to hold. As noted in the main text, it is common to centre the colmuns of $\mathbf{X}_h$ and $\mathbf{W}$ around expected dosage at each locus Meuwissen et al. (2001); VanRaden (2008). We omit this centring for brevity and note that it does not change the estimable contrasts between model parameters. It will also highlight the potential of working with the sparse precision matrix with the TBLUP approach for haplotype values.

We assume that each amino acid has an effect on a trait of interest and therefore each underlying haplotype has a value for the trait. We assume that the root haplotype (H1) has a value of 0 and will hence be aliased with the model intercept. The derived haplotypes are assigned different values compared to the root in line with their amino acid. We arbitrarily assigned these haplotype values to the amino acids: alanine (ALA = 0, root), glycine (GLY = 3), arginine (ARG = 2), and proline (PRO = 1), which also induces corresponding mutation effects (Table S1). Note that these mutation effects depend on the genome context (surrounding nucleotides) and hence manifest non-additive effects. For example, mutation 3 at site 2 (C →G) changed alanine to glycine with the corresponding effect of 3 units, while mutations 4 and 7 at site 2 (also C →G) changed proline to arginine with the corresponding effect of 1 unit. Allele substitution effects α at the three nucleotides calculated independently (by fitting $\mathbf{h} = \mathbf{1}\mu + \mathbf{X}_{h,k}\alpha_k + \mathbf{e}$ for each column k of $\mathbf{X}_h$) are 0.4, 2.0, and 0.4, while calculated jointly (by fitting $\mathbf{h} = \mathbf{1}\mu + \mathbf{X}_h\alpha + \mathbf{e}$) are $\sim$ 0.0, 2.0, and $\sim$ 0.0.

Based on this local DNA tree, we generated a small dataset by randomly sampling the haplotypes for nine individuals and simulating phenotypes for the individuals from $\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{h} + \mathbf{e}$, where $\mu = 10$, $\mathbf{h}$ are haplotype values from Table S1, and $\mathbf{e} \sim N(0, 1)$. The simulated data are shown in Table S2.

In the following, we demonstrate the key steps with SNP-BLUP and GBLUP approaches with allele haplotype matrix $\mathbf{X}_h$ and mutation haplotype matrix $\mathbf{W}$ (the TBLUP approach).

B.1.1 SNP-BLUP with allele dosages

Using the *SNP-BLUP approach* (3.2) with the phenotype values $\mathbf{y}$ and the design matrix $\mathbf{Z}$ from Table S2, the *allele haplotype matrix* $\mathbf{X}_h$ from (B.1), and assuming $\sigma^2_\alpha = 1$ and $\sigma^2_e = 1$, we get the following SNP-BLUP system of equations and estimates:

Table S2: Simulated data based on the small example.

Individual	Haplotype	Haplotype value	Phenotype value
1	H3	1	11.76
2	H2	0	7.78
3	H2	0	9.71
4	H5	3	12.10
5	H10	3	14.30
6	H9	2	12.49
7	H8	2	11.50
8	H6	2	11.82
9	H7	1	9.54

$$\begin{pmatrix} \mathbf{1}^T\mathbf{1} & \mathbf{1}^T\mathbf{Z}\mathbf{X}_h \\ \mathbf{X}_h^T\mathbf{Z}^T\mathbf{1} & \mathbf{X}_h^T\mathbf{Z}^T\mathbf{Z}\mathbf{X}_h + \mathbf{I}\sigma_e^2/\sigma_\alpha^2 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha \end{pmatrix} = \begin{pmatrix} \mathbf{1}^T\mathbf{y} \\ \mathbf{X}_h^T\mathbf{Z}^T\mathbf{y} \end{pmatrix},$$

$$\begin{pmatrix} 9 & 5 & 5 & 4 \\ 5 & 6 & 3 & 3 \\ 5 & 3 & 6 & 3 \\ 4 & 3 & 3 & 5 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha \end{pmatrix} = \begin{pmatrix} 101.00 \\ 57.11 \\ 62.21 \\ 47.83 \end{pmatrix}, \begin{pmatrix} \hat{\mu} \\ \hat{\alpha} \end{pmatrix} = \begin{pmatrix} 9.99 \\ 0.08 \\ 1.78 \\ 0.47 \end{pmatrix}.$$

The estimates of allele substitution effects $\hat{\alpha}$ differ from the true values above due to the small noisy dataset and penalised estimation. Applying these estimates to the allele haplotype matrix, gives estimates of haplotype values:

$$\hat{\mathbf{h}}^T = \mathbf{X}_h\hat{\alpha} = \Big(\; 0.00 \; 0.00 - 0.08 \; 0.47 \; 1.78 \; 0.85 \; 0.54 \; 2.32 \; 2.32 \; 2.24 \Big).$$

These estimates of haplotype values differ from the true values in Table S1 because we used only a small noisy dataset in the estimation (Table S2), but also because the SNP-BLUP approach uses multiple linear regression on allele dosages, whereas the mutation effects in the local DNA tree have additive and non-additive effects.

B.1.2 GBLUP with allele dosages

Using the *GBLUP approach* (3.5) with the phenotype values $\mathbf{y}$ and the design matrix $\mathbf{Z}$ from Table S2, the *allele haplotype matrix* $\mathbf{X}_h$ from (B.1), and assuming $\sigma_\alpha^2 = 1$, $\sigma_\epsilon^2 = 0.0000001$, and $\sigma_e^2 = 1$, we get the following GBLUP system of equations and

estimates:

$$\begin{pmatrix} \mathbf{1}^T\mathbf{1} & \mathbf{1}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{1} & \mathbf{Z}^T\mathbf{Z} + (\mathbf{X}_h\mathbf{X}_h^T + \mathbf{I}\sigma_\epsilon^2)^{-1}\sigma_e^2/\sigma_\alpha^2 \end{pmatrix} \begin{pmatrix} \mu \\ \mathbf{h} \end{pmatrix} = \begin{pmatrix} \mathbf{1}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{pmatrix},$$

$$\begin{pmatrix}
9 & 0e+00 & 2e+00 & 1.0e+00 & 0.0e+00 & 1.0e+00 & 1.0e+00 & 1.0e+00 & 1.0e+00 & 1.0e+00 \\
1.0e+00 & & & & & & & & & \\
0 & 1e+07 & 0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 \\
0.0e+00 & & & & & & & & & \\
2 & 0e+00 & 1e+07 & 0.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 \\
0.0e+00 & & & & & & & & & \\
1 & 0e+00 & 0e+00 & 6.4e+06 & 1.4e+06 & 1.4e+06 & -2.3e+06 & -2.3e+06 & -9.1e+05 & -9.1e+05 \\
2.7e+06 & & & & & & & & & \\
0 & 0e+00 & 0e+00 & 1.4e+06 & 6.4e+06 & 1.4e+06 & 2.7e+06 & -2.3e+06 & -9.1e+05 & -9.1e+05 \\
-2.3e+06 & & & & & & & & & \\
1 & 0e+00 & 0e+00 & 1.4e+06 & 1.4e+06 & 6.4e+06 & -2.3e+06 & 2.7e+06 & -9.1e+05 & -9.1e+05 \\
-2.3e+06 & & & & & & & & & \\
1 & 0e+00 & 0e+00 & -2.3e+06 & 2.7e+06 & -2.3e+06 & 5.5e+06 & 4.5e+05 & -1.8e+06 & -1.8e+06 \\
4.5e+05 & & & & & & & & & \\
1 & 0e+00 & 0e+00 & -2.3e+06 & -2.3e+06 & 2.7e+06 & 4.5e+05 & 5.5e+06 & -1.8e+06 & -1.8e+06 \\
4.5e+05 & & & & & & & & & \\
1 & 0e+00 & 0e+00 & -9.1e+05 & -9.1e+05 & -9.1e+05 & -1.8e+06 & -1.8e+06 & 7.3e+06 & -2.7e+06 \\
-1.8e+06 & & & & & & & & & \\
1 & 0e+00 & 0e+00 & -9.1e+05 & -9.1e+05 & -9.1e+05 & -1.8e+06 & -1.8e+06 & -2.7e+06 & 7.3e+06 \\
-1.8e+06 & & & & & & & & & \\
1 & 0e+00 & 0e+00 & 2.7e+06 & -2.3e+06 & -2.3e+06 & 4.5e+05 & 4.5e+05 & -1.8e+06 & -1.8e+06 \\
5.5e+06 & & & & & & & & &
\end{pmatrix}
\begin{pmatrix} \mu \\ \mathbf{h} \end{pmatrix} = \begin{pmatrix} 101.00 \\ 0.00 \\ 17.49 \\ 11.76 \\ 0.00 \\ 12.10 \\ 11.82 \\ 9.54 \\ 11.50 \\ 12.49 \\ 14.30 \end{pmatrix},$$

$$\hat{\mathbf{h}}^T = \Big(0.00, 0.00, 0.08, 0.47, 1.78, 1.85, 0.54, 2.32, 2.32, 2.24\Big),$$

which are the same estimates of haplotype values as with the SNP-BLUP approach with allele dosages.

B.1.3 SNP-BLUP with mutation dosages (TBLUP for mutation effects)

Using the *SNP-BLUP approach* (3.2) with the phenotype values $\mathbf{y}$ and the design matrix $\mathbf{Z}$ from Table S2, the *mutation haplotype matrix* $\mathbf{W}$ from (B.1), and assuming

$\sigma^2_m = 1$ and $\sigma^2_e = 1$, we get the following SNP-BLUP system of equations and estimates:

$$\begin{pmatrix} \mathbf{1}^T\mathbf{1} & \mathbf{1}^T\mathbf{Z}\mathbf{W} \\ \mathbf{W}^T\mathbf{Z}^T\mathbf{1} & \mathbf{W}^T\mathbf{Z}^T\mathbf{Z}\mathbf{W} + \mathbf{I}\sigma^2_e/\sigma^2_m \end{pmatrix} \begin{pmatrix} \mu \\ \mathbf{m} \end{pmatrix} = \begin{pmatrix} \mathbf{1}^T\mathbf{y} \\ \mathbf{W}^T\mathbf{Z}^T\mathbf{y} \end{pmatrix},$$

$$\begin{pmatrix} 9 & 5 & 0 & 2 & 2 & 2 & 1 & 1 & 1 \\ 5 & 6 & 0 & 0 & 2 & 2 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 1 \\ 2 & 2 & 0 & 0 & 3 & 0 & 1 & 0 & 0 \\ 2 & 2 & 0 & 0 & 0 & 3 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 2 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} \mu \\ \mathbf{m} \end{pmatrix} = \begin{pmatrix} 101.00 \\ 57.11 \\ 0.00 \\ 26.40 \\ 23.32 \\ 22.03 \\ 11.50 \\ 12.49 \\ 14.30 \end{pmatrix}, \begin{pmatrix} \hat{\mu} \\ \hat{\mathbf{m}} \end{pmatrix} = \begin{pmatrix} 10.05 \\ 0.94 \\ 0.00 \\ 1.67 \\ 0.44 \\ -0.28 \\ 0.04 \\ 0.89 \\ 1.29 \end{pmatrix}.$$

The estimates of mutation effects $\hat{\mathbf{m}}$ resemble the true effects, but ultimately differ due to the small noisy dataset. Applying these estimates to the mutation haplotype matrix, gives estimates of haplotype values:

$$\hat{\mathbf{h}}^T = \mathbf{W}\hat{\mathbf{m}} = \begin{pmatrix} 0.00 \\ 0.00 \\ 0.94 \\ 0.00 \\ 1.67 \\ 1.37 \\ 0.66 \\ 1.41 \\ 1.55 \\ 2.96 \end{pmatrix}.$$

B.1.4 GBLUP with mutation dosages (TBLUP for haplotype values)

Using the *GBLUP approach* (3.3) with the phenotype values $\mathbf{y}$ and the design matrix $\mathbf{Z}$ from Table S2, the *mutation haplotype matrix* $\mathbf{W}$ from (B.1), and assuming $\sigma^2_m = 1$, $\sigma^2_\epsilon = 1e-07$, and $\sigma^2_e = 1$, we get the following GBLUP system of equations and

estimates:

$$\begin{pmatrix} \mathbf{1}^T\mathbf{1} & \mathbf{1}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{1} & \mathbf{Z}^T\mathbf{Z} + (\mathbf{W}\mathbf{W}^T + \mathbf{I}\sigi^2_\epsilon/\sigma^2_m)^{-1}\sigma^2_e/\sigma^2_m \end{pmatrix} \begin{pmatrix} \mu \\ \mathbf{h} \end{pmatrix} = \begin{pmatrix} \mathbf{1}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{pmatrix},$$

$$\begin{pmatrix} 9 & 0e+00 & 2e+00 & 1.0e+00 & 0.0e+00 & 1.0e+00 & 1.0e+00 & 1.0e+00 & 1.0e+00 & 1.0e+00 \\ 1.0e+00 & & & & & & & & & \\ 0 & 1e+07 & 0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 \\ 0.0e+00 & & & & & & & & & \\ 2 & 0e+00 & 1e+07 & 0.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 \\ 0.0e+00 & & & & & & & & & \\ 1 & 1e+00 & 0e+00 & 4.0e+00 & 0.0e+00 & 0.0e+00 & -1.0e+00 & -1.0e+00 & -1.0e-07 & -1.0e-07 \\ 0.0e+00 & & & & & & & & & \\ 0 & 0e+00 & 0e+00 & 0.0e+00 & 1.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 \\ 0.0e+00 & & & & & & & & & \\ 1 & 1e+00 & 0e+00 & 0.0e+00 & 0.0e+00 & 3.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 \\ -1.0e+00 & & & & & & & & & \\ 1 & 1e+00 & 0e+00 & -1.0e+00 & 0.0e+00 & 0.0e+00 & 3.0e+00 & -1.0e-07 & -1.0e+00 & -1.0e-14 \\ 0.0e+00 & & & & & & & & & \\ 1 & 1e+00 & 0e+00 & -1.0e+00 & 0.0e+00 & 0.0e+00 & -1.0e-07 & 3.0e+00 & -1.0e-14 & -1.0e+00 \\ 0.0e+00 & & & & & & & & & \\ 1 & 1e+00 & 0e+00 & -1.0e-07 & 0.0e+00 & 0.0e+00 & -1.0e+00 & -1.0e-14 & 2.0e+00 & -1.0e-21 \\ 0.0e+00 & & & & & & & & & \\ 1 & 1e+00 & 0e+00 & -1.0e-07 & 0.0e+00 & 0.0e+00 & -1.0e-14 & -1.0e+00 & -1.0e-21 & 2.0e+00 \\ 0.0e+00 & & & & & & & & & \\ 1 & 1e+00 & 0e+00 & 0.0e+00 & 0.0e+00 & -1.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 & 0.0e+00 \\ 2.0e+00 & & & & & & & & & \end{pmatrix} \begin{pmatrix} \mu \\ \mathbf{h} \end{pmatrix} = \begin{pmatrix} 101.00 \\ 0.00 \\ 17.49 \\ 11.76 \\ 0.00 \\ 12.10 \\ 11.82 \\ 9.54 \\ 11.50 \\ 12.49 \\ 14.30 \end{pmatrix},$$

$$\hat{\mathbf{h}}^T = \begin{pmatrix} 0.00 \\ 0.00 \\ 0.94 \\ 0.00 \\ 1.67 \\ 1.37 \\ 0.66 \\ 1.41 \\ 1.55 \\ 2.96 \end{pmatrix},$$

which are the same estimates of haplotype values as with the SNP-BLUP approach with mutation dosages. Critically, the left-hand-side of the above system of equations is sparse, because the precision matrix for haplotype values is sparse; $\mathbf{Q}_h = (\mathbf{W}\mathbf{W}^T + \mathbf{I}\sigma^2_\epsilon/\sigma^2_m)^{-1}$, eventhough the corresponding covariance matrix $\mathbf{V}_h = \mathbf{W}\mathbf{W}^T\sigma^2_m + \mathbf{I}\sigma^2_\epsilon$ is dense.

B.1.5 TBLUP for haplotype values with the sparse precision matrix $\mathbf{Q}_h$

Instead of solving for the $\mathbf{Q}_h$ matrix, which incurs a computational cost and numerical errors, we can set it up directly from the local DNA tree Selle et al. (2021) akin to the pedigree Mrode and Pocrnic (2023). This follows from the hierarchical generative model (3.6):

$$\begin{aligned} \mathbf{h} &= \mathbf{W}\mathbf{m} + \epsilon, \\ Var\left(\mathbf{h}|\mathbf{W}\right) &= \mathbf{W}Var\left(\mathbf{m}\right)\mathbf{W}^T + Var\left(\epsilon\right), \\ \mathbf{V}_h &= \mathbf{W}\mathbf{W}^T\sigma_m^2 + \mathbf{I}\sigma_\epsilon^2, \\ &= \left(\mathbf{W}\mathbf{W}^T + \mathbf{I}\sigma_\epsilon^2/\sigma_m^2\right)\sigma_m^2, \end{aligned} \tag{B.2}$$

Generalized Cholesky decomposition of the covariance matrix $\mathbf{V}_h$ and its inverse is then:

$$\mathbf{V}_h = \mathbf{T}\mathbf{D}\mathbf{T}^T\sigma_m^2, \tag{B.3}$$

$$\mathbf{V}_h^{-1} = \left(\mathbf{T}^{-1}\right)^T\mathbf{D}^{-1}\mathbf{T}^{-1}\sigma_m^{-2} = \mathbf{Q}_h\sigma_m^{-2}. \tag{B.4}$$

For the small example the triangular matrix $\mathbf{T}^{-1}$ is:

$$\mathbf{T^{-1}} = \begin{pmatrix} 1 & . & . & . & . & . & . & . & . & . \\ -1 & 1 & . & . & . & . & . & . & . & . \\ -1 & . & 1 & . & . & . & . & . & . & . \\ . & -1 & . & 1 & . & . & . & . & . & . \\ . & -1 & . & . & 1 & . & . & . & . & . \\ . & . & -1 & . & . & 1 & . & . & . & . \\ . & . & -1 & . & . & . & 1 & . & . & . \\ . & . & . & . & . & -1 & . & 1 & . & . \\ . & . & . & . & . & . & -1 & . & 1 & . \\ . & . & . & . & -1 & . & . & . & . & 1 \end{pmatrix},$$

where diagonal elements are 1, non-zero off-diagonal elements are -1 for each pair of immediate descendant-ancestor haplotypes, and all other elements are 0, hence directly reflecting the local DNA tree. For example, H2 and H3 are immediate descendants of H1, while H4 and H5 are immediate descendants of H2, etc. The diagonal matrix $\mathbf{D}$ is conditional variance of haplotype values, given the ancestor haplotype value, hence the variance of the branch effects (the sum of mutation effects on each branch) plus

variance of additional terms added to haplotype values (ϵ). For the root haplotype, this is equal to:

$$\mathbf{D}_{1,1} = \mathbf{w}_1 \mathbf{w}_1^T + \sigma_\epsilon^2/\sigma_m^2, \tag{B.5}$$

while for any other haplotype i it is,

$$\mathbf{D}_{i,i} = n_{m(i)}\sigma_\epsilon^2/\sigma_m^2, \tag{B.6}$$

where $n_{m(i)}$ is the number of mutations that separate the haplotype from its immediate ancestor. With a way to obtain $\mathbf{T}^{-1}$ and $\mathbf{D}^{-1}$ efficiently directly from the local DNA tree, we can also efficiently setup $\mathbf{Q}_h$. For the small example, this precision matrix is:

$$\mathbf{Q}_h = \begin{pmatrix}
1e+07 & -1 & -1 & . & . & . & . & . & . & . \\
-1 & 3 & . & -1 & -1 & . & . & . & . & . \\
-1 & . & 3 & . & . & -1 & -1 & . & . & . \\
. & -1 & . & 1 & . & . & . & . & . & . \\
. & -1 & . & . & 2 & . & . & . & . & -1 \\
. & . & -1 & . & . & 2 & . & -1 & . & . \\
. & . & -1 & . & . & . & 2 & . & -1 & . \\
. & . & . & . & . & -1 & . & 1 & . & . \\
. & . & . & . & . & . & -1 & . & 1 & . \\
. & . & . & . & -1 & . & . & . & . & 1
\end{pmatrix}.$$

Using the above precision matrix $\mathbf{Q}_h$ instead of inverting the covariance matrix $\mathbf{V}_h$, produced these estimates of haplotype values:

$$\hat{\mathbf{h}}^T = \begin{pmatrix}
10.53 \\
0.00 \\
-0.61 \\
0.61 \\
-0.61 \\
1.14 \\
0.95 \\
0.24 \\
0.96 \\
1.10 \\
2.46
\end{pmatrix}.$$

B.2 Supplemental tables

C Supplementary materials: The importance of trait stability in cross-breeding: uncovering the impacts of genotype-by-environment interaction

C.1 Dominance in the crossbred population

In simulation scenarios incorporating dominance effects, the crossbred trait value (T_C) was adjusted using the additive QTL effects of the purebred parents. Our initial approach defined the crossbred additive effect as the the average of the absolute values of the purebred additive effects. However, this assumption is incompatible with the classical definition of dominance (Falconer and Mackay, 1996). As shown in Figure S1, this formulation produces a dominance effect (d) that is zero only when both purebred additive effects are zero, which is biologically inaccurate.

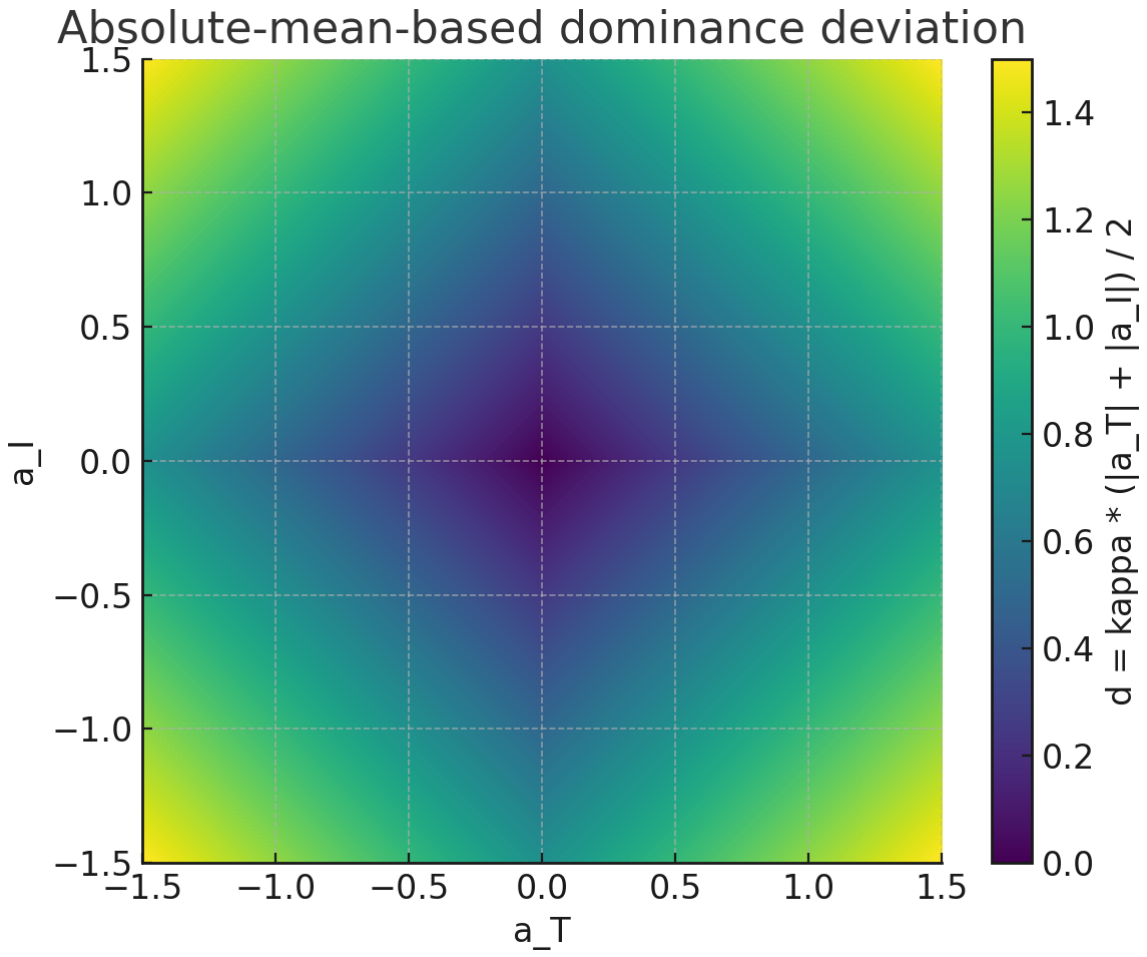

Figure S1: Absolute-mean-based dominance deviation. Heatmap of expected dominance effects (d_j) in the crossbred, given additive effects of the purebreds (a_{T_j} and a_{I_j}) ranging from -1.5 to 1.5. Under this assumption, no dominance is observed only when $a_{T_j} = a_{I_j} = 0$.

The correct formulation derives dominance from the contrast between parental additive QTL effects. Specifically, we define the crossbred additive effects as

$$|a_{C_j}| = \frac{|a_{T_j} - a_{I_j}|}{2}, \tag{C.1}$$

where $a_{T_j} - a_{I_j}$ represents the deviation between parental additive effects at any given locus. This formulation ensures that dominance reflects the heterozygote deviation and is zero whenever $a_{T_j} = a_{I_j}$, as illustrated in Figure S2.

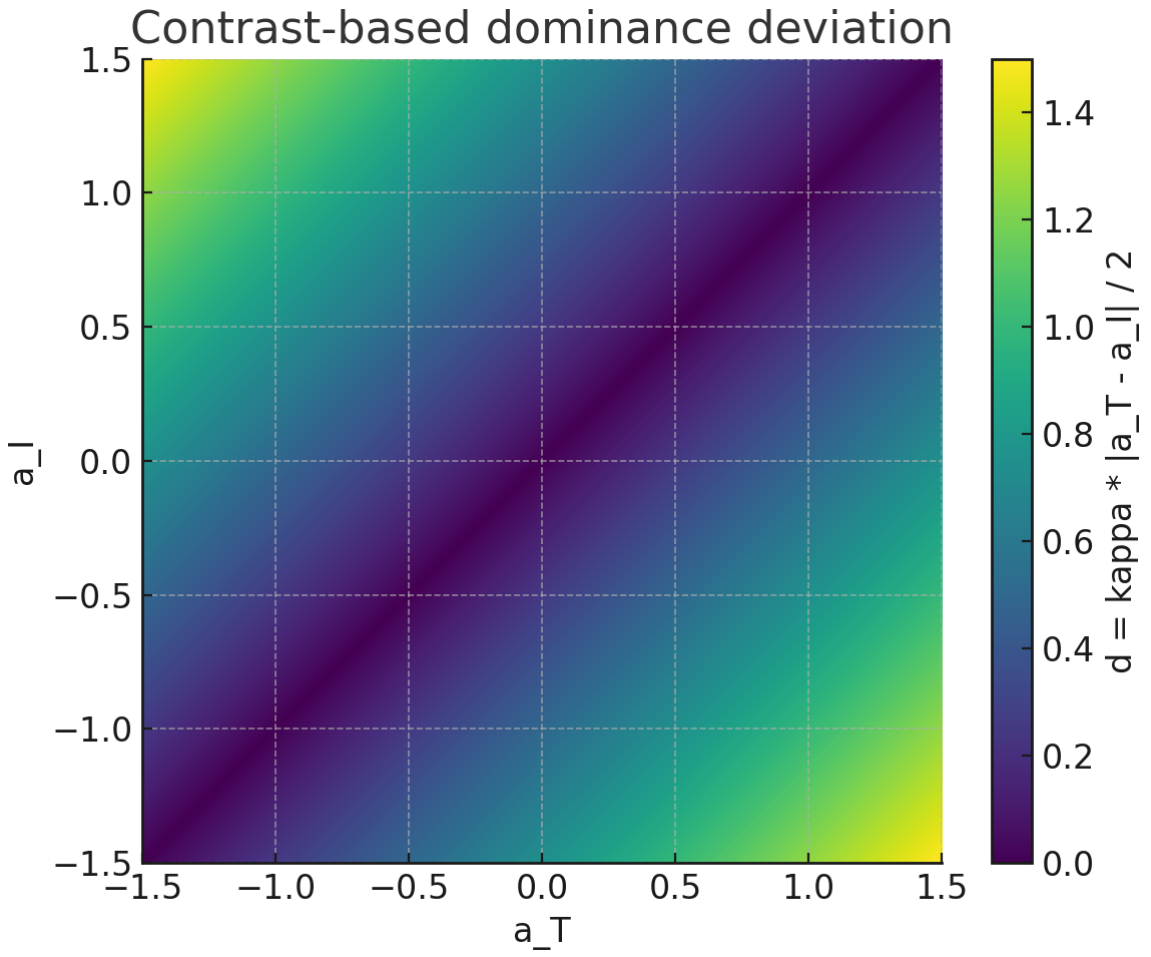

Figure S2: Contrast-based dominance deviation. Heatmap of expected dominance effects (d) in the crossbred, given additive effects of the purebreds (a_T and a_I) ranging from -1.5 to 1.5. Here, no dominance is observed when $a_T = a_I$.

C.2 G×E investigation in the indicine breeding programme

In the indicine breeding programme, where selection targets performance across multiple tropical environments (E2 to E6), the breeding objective is defined as the mean performance across these environments. This is not a rule, but an artefact of the use of (G)BLUPs without accounting for the G×E interaction structure.

The standard PC rotation shows that Factor 1 explains the majority of the variation in E2 (over 90%) and a substantial proportion in E3 (60%); the remaining environments are better captured by Factor 2 (Figure S3). The corresponded biplots in Figure S4 illustrate that, while the first two factors capture most variation, they do not clarly reflect the breeding objective.

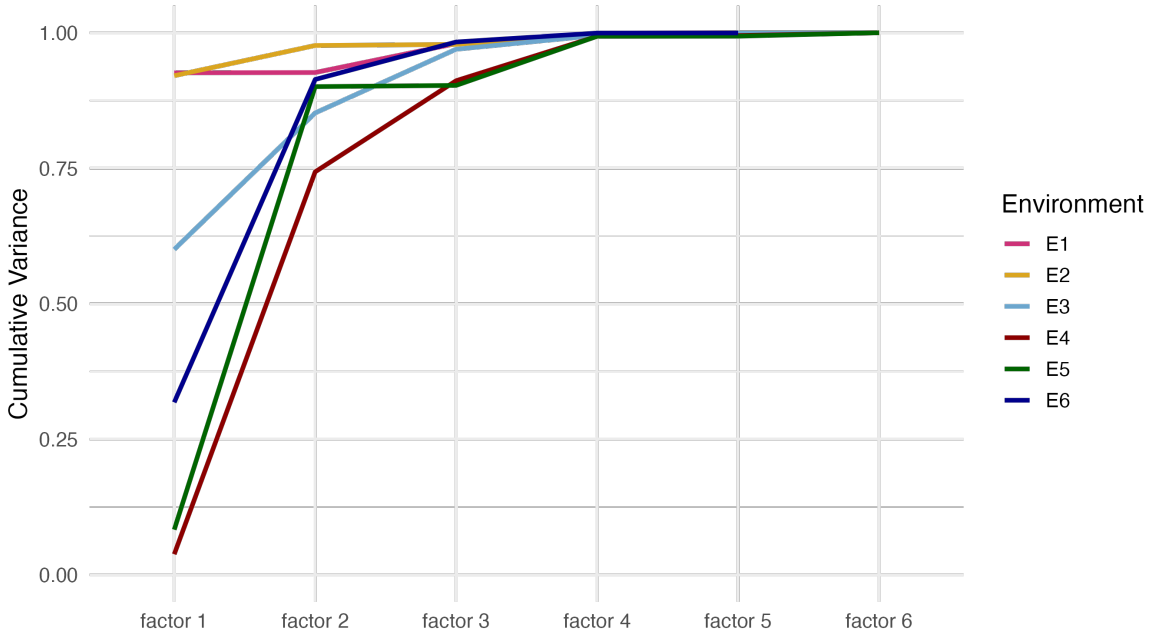

Figure S3: Cumulative variance explained by factor across environments in the indicine population under Scenario A (no dominance). Colours indicate individual environments; Factor 1 explains over 90% of the variance in E1 and E2, 60% in E3 and higher ortder factors better capture the variance in E4-E6.

Applying the informed rotation with equal focous on E2-E6 reassigns variation effectively (Table S1). Weight is 0.00 for E1, as it is not part of the breeding objective, and 0.2 for all other environments. Note that the correlation between environment and main effect is not equal across E2-E6, as it is conditioned to the correlation structure

between environments and the correlation between the environment and the breeding objective. E4, E3 and E5 have the higher correlations with the main effect (0.87, 0.78 and 0.72 respectively), while E2 and E6 have lower correlations (0.58 and 0.48 respectively). The variation in E1 is mostly of the crossover type (67.55), with a small non-crossover variance (8.16), reflecting the fact that the environment is not beeing selected for. E2 and E6 also show a high proportion of crossover variance (45.00 and 43.89 respectively).

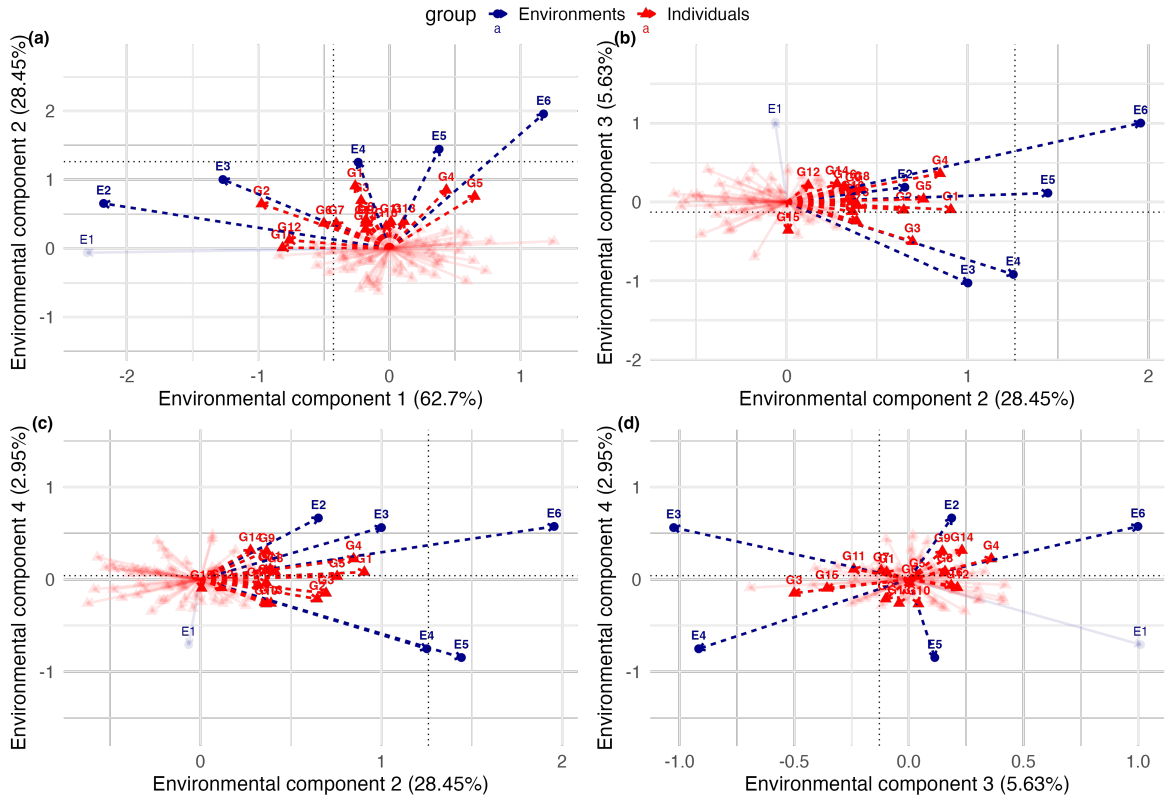

Figure S4: FAk bi-plots for the indicine population (Scenario A: no dominance). Bi-plots show environmental loadings (blue) and individual scores (red) across different environmental components (factors) combinations. Black, dotted line places the breeding objective.

The bi-plot (Figure S5) places the breeding objective, black dotted line, at the origin (0,0). The loadings for environments (red) within the breeding objective have directions that follow it, with E4 placed closely to the black dotted line, indicating it is the most correlated with the breeding objective. Selected individuals display a range of responses, with G1 and G3 well adapted to E4, G2 and G6 better adapted to E3, and G4 and G5 showing good responses to E6. G12 and G15 show good responses to E1.

The reaction norm plots (Figure S6) focusing on the top 5 selected bulls (G1, G2, G3,

Table S1: Decomposition of non-crossover and crossover variance across environments following rotation focused on E2 to E6. Includes total variance, correlation with the main effect (mean of environments E2-E6), implied genetic gain under selection in E2-E6 and weight place on each environment for rotation.

Environment	Non-cross. Var	Cross. Var	Total Var	Cor. Main Effect	Gain	Weight
E1	0.17	3.91	4.08	0.20	0.41	0.00
E2	5.77	6.66	12.43	0.68	2.40	0.20
E3	27.61	8.43	36.04	0.88	5.25	0.20
E4	82.64	4.93	87.57	0.97	9.09	0.20
E5	122.45	10.95	133.41	0.96	11.07	0.20
E6	34.78	11.96	46.74	0.86	5.90	0.20

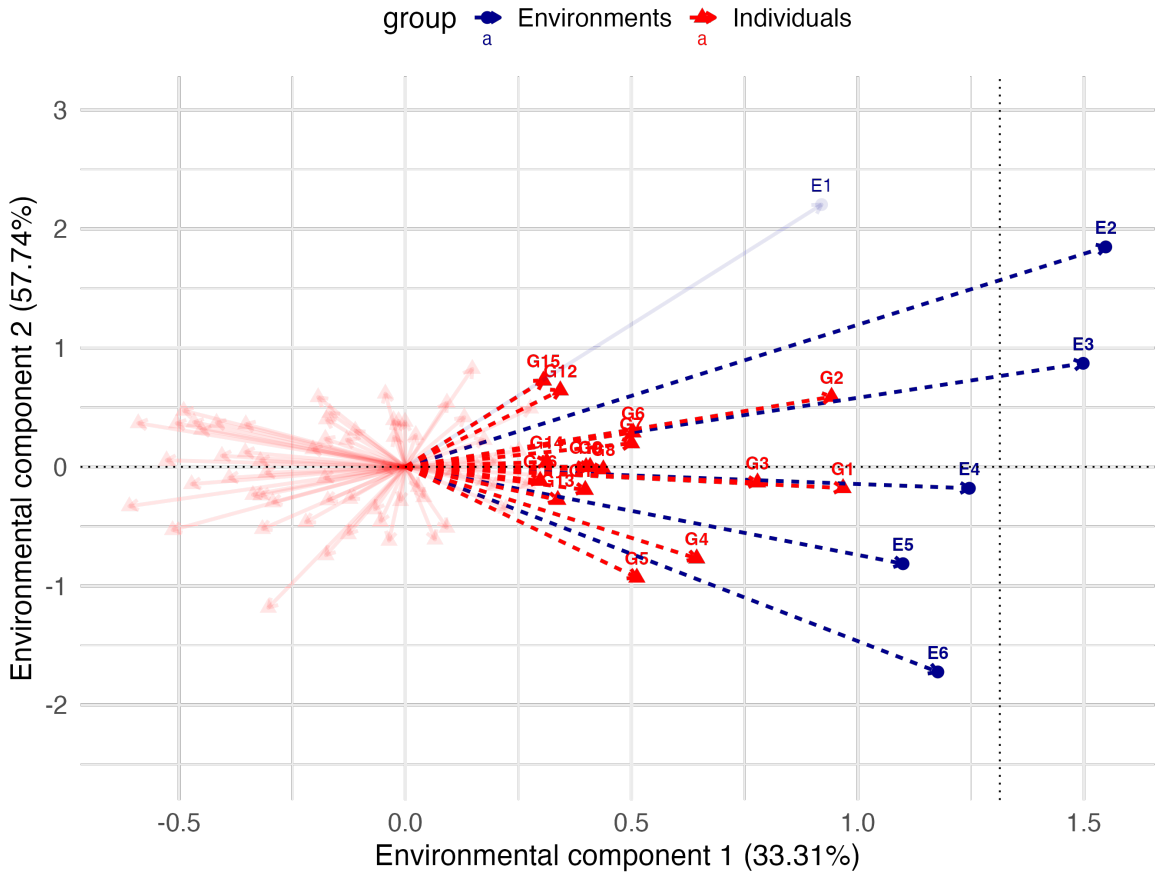

Figure S5: Informed rotation focused on average between E2-E6 (breeding objective). Biplot of loadings and scores for taurine bulls under Scenario A. The first axis captures variance correlated with the breeding objective; the second captures uncorrelated variation. Top 5 bulls shown in colours.

G4, and G5) further illustrate the differences in adaptation pattern across environments, despite the selection for an average environment.

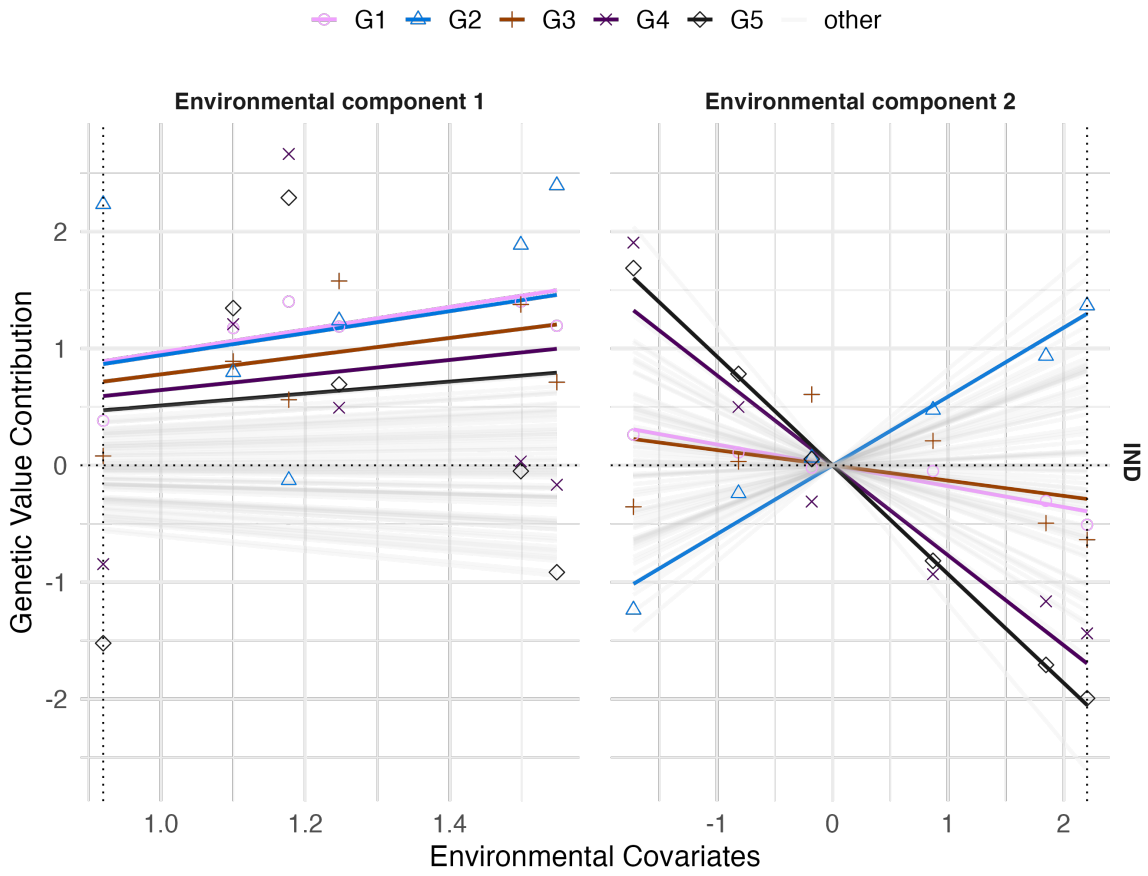

Figure S6: Reaction norms for the indicine bulls under average-focused rotation (Scenario A). Lines represent expected response across environments; coloured dots show deviations. Component 1 reflects average-correlated variation; component 2 captures uncorrelated (crossover) responses.

In the indicine programme, the informed rotation method successfully refocuses genetic variation on the relevant environmental mean, providing improved insights for multi-environment selection.

C.3 Supplementary figures

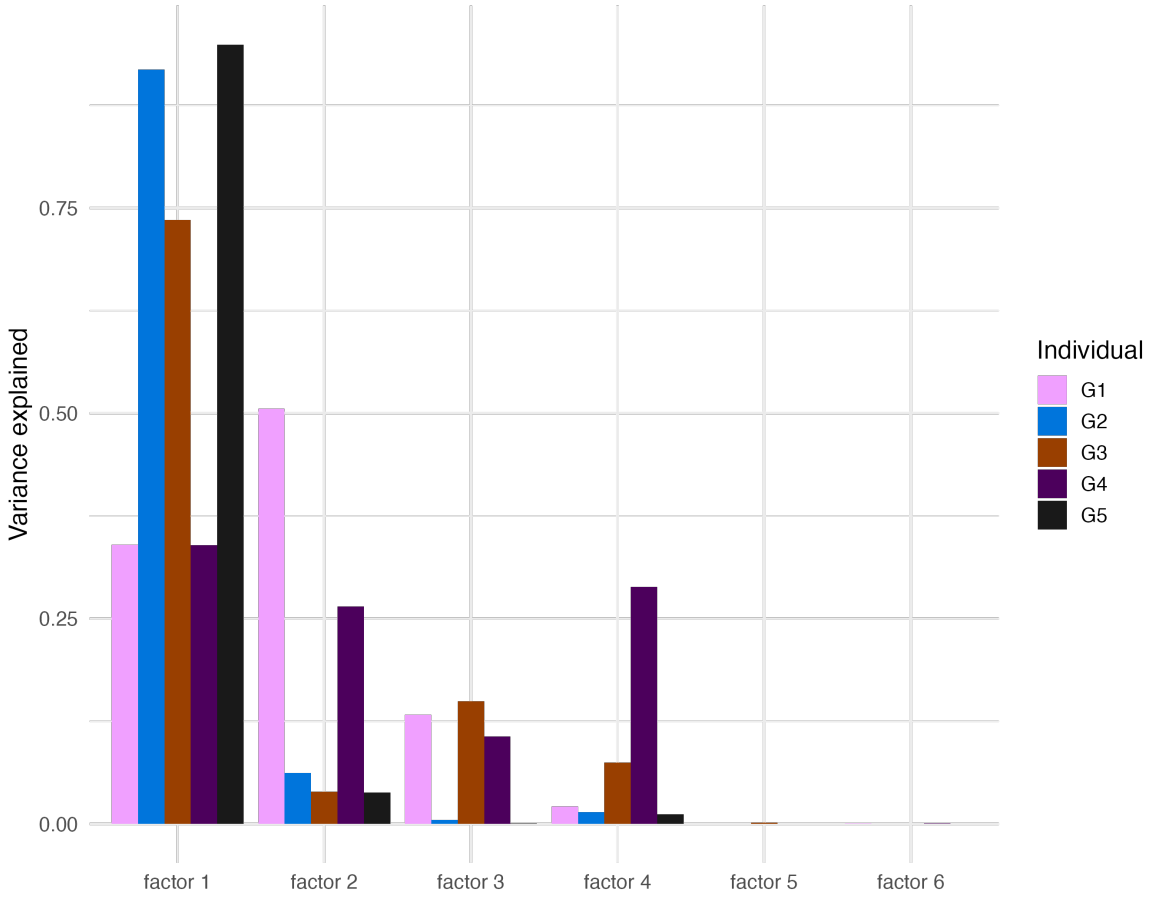

Figure S7: Factor-wise variance contributions for the top 5 taurine bulls under Scenario A. Each coloured bar represents one individual.

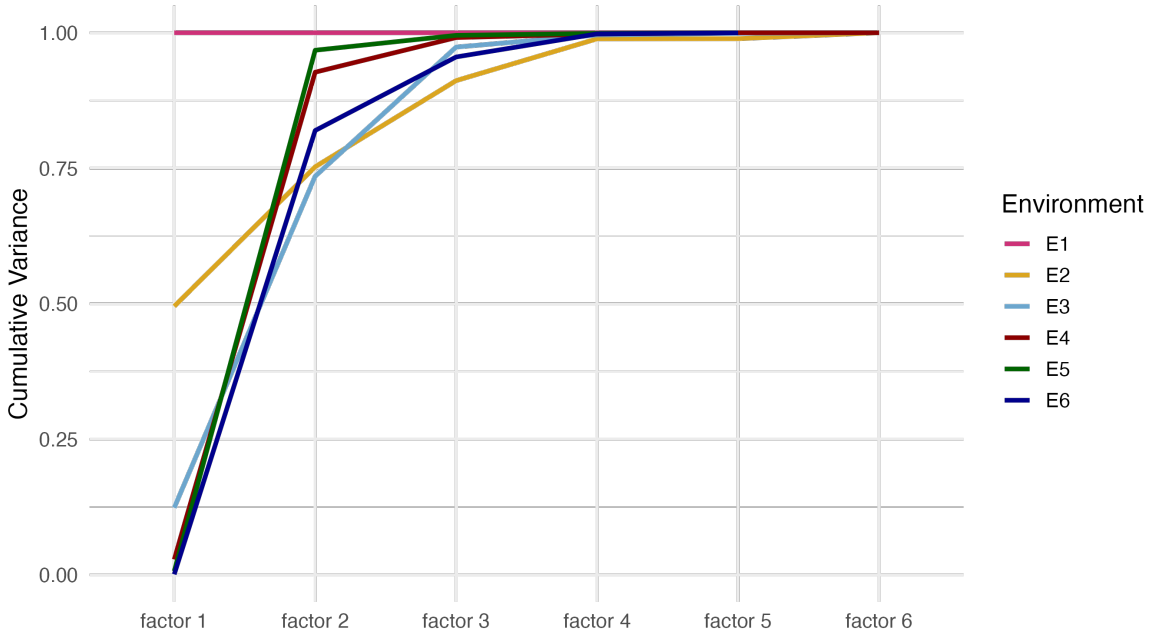

Figure S8: Cumulative variance explained by factor across environments after informed rotation in the taurine population under Scenario A (no dominance). Colours indicate individual environments; factor (environmental component) 1 explains all the variance for E1. Higher order factors better capture the variance in E2-E6.

C.4 Supplementary tables

Table S2: Proportion of variance explained in each environment by factor ($k = 6$) in the taurine (TAU), indicine (IND) and crossbreeding (CROS) programmes under four dominance scenarios (A, ADL, ADM, ADH)

BP	Scen.	factor	E1	E2	E3	E4	E5	E6	$\bar{x}_{1:6}$	$\bar{x}_{2:6}$
TAU	A	1	0.02	0.37	0.69	0.93	0.96	0.79	0.47	0.55
		2	0.25	0.44	0.29	0.05	0.03	0.16	0.19	0.19
		3	0.43	0.18	0.00	0.02	0.00	0.03	0.14	0.14
		4	0.28	0.00	0.02	0.00	0.00	0.02	0.07	0.06
		5	0.01	0.00	0.00	0.00	0.00	0.00	0.06	0.02
		6	0.01	0.01	0.00	0.00	0.00	0.00	0.06	0.03
TAU	ADL	1	0.11	0.34	0.42	0.77	0.65	0.24	0.47	0.55
		2	0.30	0.24	0.11	0.00	0.12	0.51	0.19	0.19
		3	0.03	0.38	0.06	0.11	0.03	0.05	0.14	0.14
		4	0.07	0.00	0.41	0.07	0.00	0.01	0.07	0.06
		5	0.49	0.02	0.00	0.04	0.02	0.04	0.06	0.02

Table S2: (continued)

BP	Scen.	factor	E1	E2	E3	E4	E5	E6	$\bar{x}_{1:6}$	$\bar{x}_{2:6}$
		6	0.01	0.01	0.00	0.01	0.17	0.15	0.06	0.03
		1	0.04	0.32	0.71	0.72	0.57	0.55	0.47	0.55
		2	0.05	0.29	0.18	0.02	0.00	0.37	0.19	0.19
TAU	ADM	3	0.00	0.28	0.01	0.03	0.37	0.04	0.14	0.14
		4	0.00	0.02	0.08	0.13	0.03	0.03	0.07	0.06
		5	0.00	0.09	0.02	0.10	0.03	0.01	0.06	0.02
		6	0.90	0.00	0.00	0.00	0.00	0.00	0.06	0.03
		1	0.13	0.09	0.18	0.24	0.89	0.64	0.47	0.55
		2	0.18	0.68	0.01	0.37	0.00	0.01	0.19	0.19
TAU	ADH	3	0.24	0.14	0.50	0.31	0.00	0.17	0.14	0.14
		4	0.03	0.01	0.12	0.04	0.10	0.18	0.07	0.06
		5	0.41	0.07	0.00	0.01	0.00	0.01	0.06	0.02
		6	0.02	0.01	0.19	0.04	0.00	0.00	0.06	0.03
		1	0.93	0.92	0.60	0.04	0.08	0.32	0.44	0.39
		2	0.00	0.06	0.25	0.71	0.82	0.60	0.31	0.36
IND	A	3	0.05	0.00	0.12	0.17	0.00	0.07	0.11	0.10
		4	0.02	0.02	0.03	0.08	0.09	0.02	0.07	0.07
		5	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.04
		6	0.00	0.00	0.00	0.01	0.01	0.00	0.04	0.04
		1	0.71	0.76	0.34	0.03	0.08	0.42	0.44	0.39
		2	0.01	0.02	0.24	0.82	0.58	0.27	0.31	0.36
IND	ADL	3	0.19	0.15	0.02	0.01	0.14	0.02	0.11	0.10
		4	0.04	0.05	0.23	0.04	0.08	0.14	0.07	0.07
		5	0.04	0.02	0.05	0.08	0.11	0.03	0.03	0.04
		6	0.00	0.00	0.13	0.03	0.02	0.11	0.04	0.04
		1	0.71	0.45	0.41	0.06	0.13	0.55	0.44	0.39
		2	0.01	0.02	0.30	0.60	0.42	0.18	0.31	0.36
IND	ADM	3	0.16	0.19	0.14	0.07	0.33	0.01	0.11	0.10
		4	0.04	0.00	0.12	0.21	0.07	0.05	0.07	0.07
		5	0.02	0.00	0.04	0.06	0.03	0.21	0.03	0.04
		6	0.06	0.34	0.00	0.00	0.02	0.00	0.04	0.04
		1	0.54	0.66	0.81	0.84	0.22	0.01	0.44	0.39

IND ADM

Table S2: (continued)

BP	Scen.	factor	E1	E2	E3	E4	E5	E6	$\bar{x}_{1:6}$	$\bar{x}_{2:6}$
		2	0.05	0.05	0.03	0.00	0.54	0.76	0.31	0.36
		3	0.25	0.14	0.01	0.01	0.14	0.19	0.11	0.10
		4	0.14	0.00	0.05	0.10	0.09	0.03	0.07	0.07
		5	0.00	0.01	0.09	0.04	0.00	0.00	0.03	0.04
		6	0.01	0.15	0.01	0.00	0.00	0.00	0.04	0.04
		1	0.91	0.85	0.58	0.08	0.31	0.72	0.51	0.43
		2	0.02	0.10	0.35	0.78	0.62	0.20	0.40	0.47
CROS	A	3	0.03	0.05	0.01	0.13	0.00	0.07	0.06	0.06
		4	0.04	0.00	0.05	0.00	0.05	0.01	0.02	0.03
		5	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00
		6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		1	0.95	0.86	0.17	0.01	0.25	0.46	0.51	0.43
		2	0.00	0.11	0.65	0.89	0.68	0.35	0.40	0.47
CROS	ADL	3	0.04	0.00	0.15	0.04	0.05	0.16	0.06	0.06
		4	0.01	0.03	0.02	0.05	0.02	0.03	0.02	0.03
		5	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
		6	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
		1	0.86	0.94	0.73	0.20	0.00	0.14	0.51	0.43
		2	0.03	0.00	0.16	0.68	0.92	0.77	0.40	0.47
CROS	ADM	3	0.10	0.00	0.09	0.03	0.03	0.03	0.06	0.06
		4	0.00	0.03	0.01	0.07	0.03	0.05	0.02	0.03
		5	0.00	0.01	0.00	0.01	0.02	0.01	0.00	0.00
		6	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00
		1	0.96	0.91	0.50	0.00	0.33	0.62	0.51	0.43
		2	0.02	0.07	0.37	0.84	0.64	0.29	0.40	0.47
CROS	ADH	3	0.02	0.00	0.09	0.14	0.00	0.07	0.06	0.06
		4	0.00	0.01	0.02	0.00	0.02	0.01	0.02	0.03
		5	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00
		6	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00

References

J. R. Adrion, C. B. Cole, N. Dukler, J. G. Galloway, A. L. Gladstein, et al. A community-maintained standard library of population genetic models. *eLife*, 9: e54967, 2020. doi: 10.7554/eLife.54967. URL https://doi.org/10.7554/eLife.54967.

P. Ajmone-Marsan, J. Garcia, and J. Lenstra. On the origin of cattle: How aurochs became cattle and colonized the world. *Evolutionary Antropology*, 10:148–157, 2010. doi: 10.1002/evan.20267. URL https://doi.org/10.1002/evan.20267.

H. Aliloo, J. Pryce, O. González-Recio, B. Cocks, and B. Hayes. Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. *Genetics Selection Evolution*, 48:8, 2016. doi: 10.1186/s12711-016-0186-0. URL https://doi.org/10.1186/s12711-016-0186-0.

B. Arbuckle and T. Kassebaum. Management and domestication of cattle (bos taurus) in neolithic southwest asia. *Animal Frontiers*, 11(3):10–19, May 2021. doi: 10.1093/af/vfab015. URL https://doi.org/10.1093/af/vfab015.

M. Arenas. The importance and application of the ancestral recombination graph. *Frontiers in Genetics*, 4, 2013. doi: 10.3389/fgene.2013.00206. URL https://doi.org/10.3389/fgene.2013.00206.

A. Avalos-Pacheco, M. Cronjäger, P. Jenkins, and J. Hein. An almost infinite sites model. *Theoretical Population Biology*, 160:49–61, December 2024. doi: 10.1016/j.tpb.2024.10.001. URL https://doi.org/10.1016/j.tpb.2024.10.001.

A. Bagnato and A. Rosati. From the editors—animal selection: The genomics revolution. *Animal Frontiers*, 2(1):1–2, 01 2012. doi: 10.2527/af.2011-0033. URL https://doi.org/10.2527/af.2011-0033.

D. Bates, M. Maechler, and M. Jagan. *Matrix: Sparse and Dense Matrix Classes and Methods*, 2025. URL https://CRAN.R-project.org/package=Matrix. R package version 1.7-3.

F. Baumdicker, G. Bisschop, D. Goldstein, G. Gower, A. P. Ragsdale, et al. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220(3):iyab229, 2022. doi: 10.1093/genetics/iyab229. URL https://doi.org/10.1093/genetics/iyab229.

H. C. Becker and J. Leon. Stability analysis in plant breeding. *Plant Breeding*, 101: 1–23, 1988. ISSN 0179-9541. doi: 10.1111/j.1439-0523.1988.tb00261.x. URL https://doi.org/10.1111/j.1439-0523.1988.tb00261.x.

A. Beja-Pereira, D. Caramelli, C. Lalueza-Fox, C. Vernesi, N. Ferrand, et al. The origin of european cattle: evidence from modern and ancient dna. 103(21):8113–8118, 2006. doi: 10.1073/pnas.0509210103. URL https://doi.org/10.1073/pnas.0509210103.

R. Beker. Tests for crossover genotype-environmental interactions. *Canadian Journal of Plant Science*, 68:405–410, 1988. doi: 10.4141/cjps88-051. URL https://cdnsciencepub.com/doi/pdf/10.4141/cjps88-051.

F. Betran, J. Ribaut, D. Beck, and D. De Leon. Genetic diversity, specific combining ability, and heterosis in tropical maize under stress and nonstress environments. *Crop Science*, 43:797–806, 2003. doi: 10.2135/cropsci2003.7970. URL https://doi.org/10.2135/cropsci2003.7970.

S. Bhogal and J. Brown. Dairy and products annual. Technical Report IN2024-0049, USFA Foreign Agricultural Service, October 2024. URL https://www.fas.usda.gov/data/india-dairy-and-products-annual-8. Required Report: Required - Public Distribution.

N. D. D. Board. Aproaches and experiences of nddb in development of gir, a promising indigenous milch breed, 2017. URL https://www.nddb.coop. Accessed: 2025-07-04.

P. Boettcher, J. Fatehi, and M. Schutz. Genotype × environment interactions in conventional versus pasture-based dairies in canada. *Journal of Dairy Science*, 86(1):383–389, Jan 2003. doi: 10.3168/jds.S0022-0302(03)73617-0. URL https://doi.org/10.3168/jds.S0022-0302(03)73617-0.

D. Boichard, S. Fritz, P. Croiseau, V. Ducrocq, T. Tribout, and B. Cuyabano. Erosion of estimated genomic breeding values with generations is due to long distance associations between markers and qtl. *Genetics Selection Evolution*, 57 (1), 2025. doi: 10.1186/s12711-025-00963-5. URL http://dx.doi.org/10.1186/s12711-025-00963-5.

C. Boye, S. Nirmalan, A. Ranjbaran, and F. Luca. Genotype x environment interactions in gene regulation and complex traits. *Nature Genetics*, 56(6):1057—-1068, 2024. doi: 10.1038/s41588-024-01776-w. URL http://dx.doi.org/10.1038/s41588-024-01776-w.

D. Bradley, D. MacHugh, P. Cunningham, and R. Loftus. Mitochondrial diversity and the origins of african and european cattle. *Proceedings of the National Academy of Sciences*, 93(10):5131–5135, 1996. doi: 10.1073/pnas.93.10.5131. URL https://doi.org/10.1073/pnas.93.10.5131.

L. G. Braga, F. S. Schenkel, T. C. S. Chud, J. L. Rodrigues, B. Saada, et al. Selection signatures in gir and holstein cattle. *Journal of Dairy Science*, 2024. doi: 10.3168/jds.2024-26147. URL https://doi.org/10.3168/jds.2024-26147.

V. Brajkovic, I. Pocrnic, M. Kaps, M. Špehar, V. Cubric-Curik, et al. Quantifying the effects of the mitochondrial genome on milk production traits in dairy cows: empirical results and modelling challenges. *Journal of Dairy Sciences*, 108:664–678, 2025. doi: 10.3168/jds.2024-25203. URL https://doi.org/10.3168/jds.2024-25203.

D. Brandt, C. Huber, C. Chiang, and D. Ortega-Del Vecchyo. The promise of inferring the past using the ancestral recombination graph. *Genome Biology and Evolution*, 16(2), 2024. doi: 10.1093/gbe/evae005. URL http://dx.doi.org/10.1093/gbe/evae005.

M. Brass. The domestication of livestock in the sahara. *Oxford Research Encyclopedia of African History*, 2021. doi: 10.1093/acrefore/9780190277734.013.1551. URL https://doi.org/10.1093/acrefore/9780190277734.013.1551.

L. F. Brito, N. Bedere, F. Douhard, H. R. Oliveira, M. Arnal, et al. Review: Genetic selection of high-yielding dairy cattle toward sustainable farming systems in a rapidly changing world. *Animal*, 15(1):100292, 2021. doi: 10.1016/j.animal.2021.100292. URL https://doi.org/10.1016/j.animal.2021.100292.

M. G. Bulmer. The effect of selection on genetic variability. *The American Naturalist*,

105(943):201–211, 1971. doi: 10.1086/282718. URL https://doi.org/10.1086/282718.

H. Bunning, E. Wall, M. G. G. Chagunda, G. Banos, and G. Simm. Heterosis in cattle crossbreeding schemes in tropical regions: meta-analysis of effects of breed combination, trait type, and climate on level of heterosis. *Journal of Animal Science*, 97(1):29–34, October 2018. doi: 10.1093/jas/sky406. URL https://doi.org/10.1093/jas/sky406.

A. M. Buxadera and N. Mandonnet. The importance of the genotype×environment interaction for selection and breeding programmes in tropical conditions. *CABI Reviews*, page 14 pp., 2006. doi: 10.1079/PAVSNNR20061026. URL https://doi.org/10.1079/PAVSNNR20061026.

M. P. L. Calus and R. F. Veerkamp. Estimation of environmental sensitivity of genetic merit for milk production traits. *Journal of Dairy Science*, 86(1):3756–3764, 2003. doi: 10.3168/jds.S0022-0302(03)73981-6. URL https://doi.org/10.3168/jds.S0022-0302(03)73981-6.

M. P. L. Calus, J. Vandenplas, and J. Ten Napel. Ever-growing data sets pose (new) challenges to genomic prediction models. *Journal of Animal Breeding and Genetics*, 132:407–408, 2015. doi: 10.1111/jbg.12192. URL https://doi.org/10.1111/jbg.12192.

A. Canaza-Cayo, J. Cobuci, P. Lopes, R. Torres, M. Martins, D. Daltro, and M. Silva. Genetic trend estimates for milk yield production and fertility traits of the girolando cattle in brazil. *Livestock Science*, 190:113–122, 2016. doi: 10.1016/j.livsci.2016.06.003. URL https://doi.org/10.1016/j.livsci.2016.06.003.

J. Casellas and L. Varona. Effect of mutation age on genomic predictions. *Journal of Dairy Science*, 94(8):4224—-4229, 2011. doi: 10.3168/jds.2011-4186. URL http://dx.doi.org/10.3168/jds.2011-4186.

J. Casellas, C. Esquivelzeta, and A. Legarra. Accounting for new mutations in genomic prediction models. *Journal of Dairy Science*, 96(8):5398—-5402, 2013. doi: 10.3168/jds.2012-6468. URL http://dx.doi.org/10.3168/jds.2012-6468.

J. Castresana. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, 17:540–552, 2000.

C. Castro and J. Degreenia. Dairy and products annual. Technical Report BR2024-0029, USFA Foreign Agricultural Service, October 2024. URL https://www.fas.usda.gov/data/brazil-dairy-and-products-annual-11. Required Report: Required - Public Distribution.

G. K. Chen, P. Marjoram, and J. D. Wall. Fast and flexible simulation of dna sequence data. *Genome Research*, 19(1):136–142, January 2009. doi: 10.1101/gr.083634.108. URL https://doi.org/10.1101/gr.083634.108.

S. Chen, B. Z. Lin, M. Baig, B. Mitra, R. J. Lopes, et al. Zebu cattle are an exclusive legacy of the south asia neolithic. *Molecular Biology and Evolution*, 27(1):1–6, January 2010. doi: 10.1093/molbev/msp213. URL https://doi.org/10.1093/molbev/msp213.

R. Christ, X. Wang, L. J. M. Aslett, D. Steinsaltz, and I. Hall. Clade distillation for genome-wide association studies. *GENETICS*, 2025. doi: 10.1093/genetics/iyaf158. URL http://dx.doi.org/10.1093/genetics/iyaf158.

S. A. Clark, J. M. Hickey, H. D. Daetwyler, and J. H. van der Werf. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genetics Selection Evolution*, 44(1), 2012. ISSN 1297-9686. doi: 10.1186/1297-9686-44-4. URL http://dx.doi.org/10.1186/1297-9686-44-4.

N. Clay, T. Garnett, and J. Lorimer. Dairy intensification: Drivers, impacts and alternatives. *Ambio*, 49(1):35–48, 2020. doi: 10.1007/s13280-019-01177-y. URL https://doi.org/10.1007/s13280-019-01177-y.

D. Conrad and M. Hurles. The population genetics of structural variation. *Nature Genetics*, 39(Suppl 7):S30–S36, 2007. doi: 10.1038/ng2042. URL https://doi.org/10.1038/ng2042.

J. Crossa, H. G. Gauch, and R. W. Zobel. Additive main effects and multiplicative interaction analysis of two international maize cultivar trials. *Theoretical and Applied Genetics*, 81(1):31–37, 1990. doi: 10.1007/BF00226130. URL https://doi.org/10.1007/BF00226130.

B. R. Cullis, A. B. Smith, C. P. Beeck, and W. A. Cowling. Analysis of yield and oil from a series of canola breeding trials. Part II. exploring variety by environment interaction

using factor analysis. *Genome*, 53(11):1002–1016, 2010. doi: 10.1139/G10-080. URL https://doi.org/10.1139/G10-080.

B. C. D. Cuyabano, D. Boichard, and C. Gondro. Expected values for the accuracy of predicted breeding values accounting for genetic differences between reference and target populations. *Genetics Selection Evolution*, 56(1), 2024. doi: 10.1186/s12711-024-00876-9. URL http://dx.doi.org/10.1186/s12711-024-00876-9.

P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li. Twelve years of samtools and bcftools. *GigaScience*, 10(2):giab008, February 2021. doi: 10.1093/gigascience/giab008. URL https://doi.org/10.1093/gigascience/giab008.

C. R. Darwin. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray, Albemarle street, London, 1859.

N. de Leon, J.-L. Jannink, J. W. Edwards, and S. M. Kaeppler. Introduction to a special issue on genotype by environment interaction. *Crop Science*, 56(5):2081–2089, September 2016. doi: 10.2135/cropsci2016.07.0002in. URL https://doi.org/10.2135/cropsci2016.07.0002in.

Y. Deng, R. Nielsen, and Y. S. Song. Robust and accurate bayesian inference of genome-wide genealogies for large samples. 2024. doi: 10.1101/2024.03.16.585351. URL https://doi.org/10.1101/2024.03.16.585351.

L. Di Bari, M. Bisardi, S. Cotogno, M. Weigt, and F. Zamponi. Emergent time scales of epistasis in protein evolution. *Proceedings of the National Academy of Sciences*, 121(40), Sept. 2024. ISSN 1091-6490. doi: 10.1073/pnas.2406807121. URL http://dx.doi.org/10.1073/pnas.2406807121.

G. E. Dickerson. Inbreeding and heterosis in animals. *Journal of Animal Science*, 1973 (Symposium):54–77, 1973. doi: 10.1093/ansci/1973.Symposium.54. URL https://doi.org/10.1093/ansci/1973.Symposium.54.

Y. Ding, K. Hou, Z. Xu, A. Pimplaskar, E. Petter, K. Boulier, F. Privé, B. J. Vilhjálmsson, L. M. Olde Loohuis, and B. Pasaniuc. Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature*, 618(7966):774—-781, May 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06079-4. URL http://dx.doi.org/10.1038/s41586-023-06079-4.

H. Doekes, R. Veerkamp, P. Bijma, S. Hiemstra, and J. Windig. Inbreeding depression due to recent and ancient inbreeding in dutch holstein–friesian dairy cattle. *Genetics Selection Evolution*, 51:54, 2019. doi: 10.1186/s12711-019-0497-z. URL https://doi.org/10.1186/s12711-019-0497-z.

J. P. Domínguez-Castaño, A. Legarra, I. Aguilar, Z. G. Vitezica, and L. Varona. Genomic evaluation using principal components of the genomic covariance matrix. *Genetics Selection Evolution*, 53(1):75, 2021. URL https://gsejournal.biomedcentral.com/articles/10.1186/s12711-021-00663-y.

J. Dorji, C. J. Vander Jagt, A. J. Chamberlain, B. G. Cocks, I. M. MacLeod, and H. D. Daetwyler. Recovery of mitogenomes from whole genome sequences to infer maternal diversity in 1883 modern taurine and indicine cattle. *Scientific Reports*, 12(1):5582, April 2022. doi: 10.1038/s41598-022-09427-y. URL https://doi.org/10.1038/s41598-022-09427-y.

J. Dunne, R. Evershed, M. Salque, et al. First dairying in green saharan africa in the fifth millennium bc. *Nature*, 486:390–394, 2012. doi: 10.1038/nature11186. URL https://doi.org/10.1038/nature11186.

A. Durvasula and K. E. Lohmueller. Negative selection on complex traits limits phenotype prediction accuracy between populations. *The American Journal of Human Genetics*, 108(4):620—631, 2021. doi: 10.1016/j.ajhg.2021.02.013. URL http://dx.doi.org/10.1016/j.ajhg.2021.02.013.

S. A. Eberhart and W. A. Russell. Stability parameters for comparing varieties. *Crop Science*, 6(1):36–40, 1966. doi: 10.2135/cropsci1966.0011183X000600010011x. URL https://doi.org/10.2135/cropsci1966.0011183X000600010011x.

J. Eiríksson, E. Karaman, G. Su, et al. Breed of origin of alleles and genomic predictions for crossbred dairy cows. *Genetics Selection Evolution*, 53:84, 2021. doi: 10.1186/s12711-021-00678-3. URL https://doi.org/10.1186/s12711-021-00678-3.

Embrapa Gado de Leite. Embrapa gado de leite, Accessed: 27 May 2025. URL https://www.embrapa.br/gado-de-leite.

J. Ertl, A. Legarra, Z. Vitezica, L. Varona, C. Edel, R. Emmerling, and K. Götz. Genomic analysis of dominance effects on milk production and conformation traits in fleckvieh cattle. *Genetics Selection Evolution*, 46:40, 2014. doi: 10.1186/1297-9686-46-40. URL https://doi.org/10.1186/1297-9686-46-40.

D. S. Falconer. The problem of environment and selection. *The American Naturalist*, 86:293–298, 1952. doi: 10.1086/281736. URL https://doi.org/10.1086/281736.

D. S. Falconer. Selection in different environments: Effects on environmental sensitivity (reaction norm) and on mean performance. *Genetical Research*, 56(1):57–70, 1990. doi: 10.1017/S0016672300028883. URL https://doi.org/10.1017/S0016672300028883.

D. S. Falconer and T. Mackay. *Introduction to Quantitative Genetics*. Longman, Essex, England, 4th edition, 1996.

FAO and GDP. Climate change and the global dairy cattle sector – the role of the dairy sector in a low-carbon future, 2018. URL https://www.fao.org/. Licence: CC BY-NC-SA-3.0 IGO.

FAOSTAT. Crops and livestock products, 2025. URL https://www.fao.org/faostat/en/#data/QCL. Accessed May 27, 2025.

M. Felius, M.-L. Beerling, D. S. Buchanan, B. Theunissen, P. A. Koolmees, and J. A. Lenstra. On the history of cattle genetic resources. *Diversity*, 6(4):705–750, 2014. doi: 10.3390/d6040705. URL https://doi.org/10.3390/d6040705.

J. Felsenstein. The evolutionary advantage of recombination. *Genetics*, 78(2):737–756, 1974. doi: 10.1093/genetics/78.2.737. URL https://doi.org/10.1093/genetics/78.2.737.

K. W. Finlay and G. N. Wilkinson. The analysis of adaptation in a plant-breeding programme. *Australian Journal of Agricultural Research*, 14(6):742–754, 1963. doi: 10.1071/AR9630742. URL https://doi.org/10.1071/AR9630742.

R. A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2):399—433, 1919. doi: 10.1017/s0080456800012163. URL http://dx.doi.org/10.1017/S0080456800012163.

Food and Agriculture Organization of the United Nations (FAO). *Understanding and Integrating Gender Issues into Livestock Projects and Programmes: A Checklist for Practitioners*. Rome, 2013. URL https://www.fao.org/4/i3216e/i3216e.pdf.

N. Freychet, G. Hegerl, D. Mitchell, and M. Allen. Future changes in the frequency of temperature extremes may be underestimated in tropical and subtrop-

ical regions. *Communications Earth & Environment*, 2(28), 2021. doi: 10.1038/s43247-021-00094-x. URL https://doi.org/10.1038/s43247-021-00094-x.

M. Gail and R. Simon. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, 41(2):361–372, June 1985.

C. S. Gallagher, G. S. Ginsburg, and A. Musick. Biobanking with genetics shapes precision medicine and global health. *Nature Reviews Genetics*, 26:191–202, 2025. doi: 10.1038/s41576-024-00794-y. URL https://doi.org/10.1038/s41576-024-00794-y.

Z. Gao, Y. Zhang, and P. Moorjani. Limited role of generation time changes in driving the evolution of the mutation spectrum in humans. *eLife*, 2023. doi: 10.7554/eLife.81188. URL https://doi.org/10.7554/eLife.81188.

L. A. García-Cortés and M. A. Toro. Multibreed analysis by splitting breed genetic effects. *Genetics Selection Evolution*, 38(6):601–615, 2006. doi: 10.1186/1297-9686-38-6-601. URL https://doi.org/10.1186/1297-9686-38-6-601.

A. García-Ruiz, J. B. Cole, P. M. VanRaden, G. R. Wiggans, F. J. Ruiz-López, and C. P. Van Tassell. Changes in genetic selection differentials and generation intervals in us holstein dairy cattle as a result of genomic selection. *Proceedings of the National Academy of Sciences of the United States of America*, 113(28):E3995–E4004, 2016. doi: 10.1073/pnas.1519061113. URL https://doi.org/10.1073/pnas.1519061113.

H. G. Gauch. Model selection and validation for yield trials with interaction. *Biometrics*, 44:705–715, 1988. doi: 10.2307/2531585. URL https://doi.org/10.2307/2531585.

H. G. Gauch. *Statistical Analysis of Regional Yield Trials*. Elsevier, Amsterdam, 1992. ISBN 9780444880556.

M. Gauly and S. Ammer. Review: Challenges for dairy cow production systems arising from climate changes. *Animal*, 14(S1):s196–s203, 2020. doi: 10.1017/S1751731119003239.

R. C. Gaynor, G. Gorjanc, and J. M. Hickey. Alphasimr: an r package for breeding program simulations. *G3: Genes|Genomes|Genetics*, 11(2):jkaa017, February 2021. doi: 10.1093/g3journal/jkaa017. URL https://doi.org/10.1093/g3journal/jkaa017.

N. Z. Gebrehiwot, E. M. Strucken, H. Aliloo, K. Marshal, and J. P. Gibson. The

patterns of admixture, divergence, and ancestry of african cattle populations determined from genome-wide snp data. *BMC Genomics*, 21:869, 2020. doi: 10.1186/s12864-020-07264-7. URL https://doi.org/10.1186/s12864-020-07264-7.

G. Gibson. Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 13:135–145, 2012. doi: 10.1038/nrg3118. URL https://doi.org/10.1038/nrg3118.

C. Ginja, L. Gama, O. Cortés, et al. The genetic ancestry of american creole cattle inferred from uniparental and autosomal genetic markers. *Scientific Reports*, 9:11486, 2019. doi: 10.1038/s41598-019-47636-0. URL https://doi.org/10.1038/s41598-019-47636-0.

G. Giovambattista, S.-N. Takeshima, M. V. Ripoli, Y. Matsumoto, L. A. A. Franco, et al. Characterization of bovine mhc drb3 diversity in latin american creole cattle breeds. *Gene*, 519:150–158, 2013. doi: 10.1016/j.gene.2013.01.002. URL https://doi.org/10.1016/j.gene.2013.01.002.

G. Giovambattista, S.-N. Takeshima, K. Moe, J. Pereira Rico, M. Polat, A. Loza Vega, O. Arce Cabrera, and Y. Aida. Bola-drb3 genetic diversity in highland creole cattle from bolivia. *HLA*, 96:688–696, 2020. doi: 10.1111/tan.14120. URL https://doi.org/10.1111/tan.14120.

M. Goddard. Can we make genomic selection 100 *Journal of Animal Breeding and Genetics*, 134(4):287—288, 2017. doi: 10.1111/jbg.12281. URL http://dx.doi.org/10.1111/jbg.12281.

C. Godde, D. Mason-D'Croz, D. Mayberry, P. Thornton, and M. Herrero. Impacts of climate change on the livestock food supply chain; a review of the evidence. *Global Food Security*, 28:100488, March 2021. doi: 10.1016/j.gfs.2021.100488. URL https://doi.org/10.1016/j.gfs.2021.100488.

H. F. Gollob. A statistical model which combines features of factor analytic and analysis of variance techniques. *Psychometrika*, 33(1):73–115, 1968. doi: 10.1007/BF02289676. URL https://doi.org/10.1007/BF02289676.

R. Griffiths and P. Marjoram. An ancestral recombination graph. *Progress in population genetics and human evolution*, pages 257–270, 1997.

B. Grisart, W. Coppieters, F. Farnir, L. Karim, C. Ford, et al. Positional candidate cloning of a qtl in dairy cattle: Identification of a missense mutation in the bovine

dgat1 gene with major effect on milk yield and composition. *Genome Research*, 12:222–231, 2002. doi: 10.1101/gr.224202. URL https://doi.org/10.1101/gr.224202.

G. Grolemund and W. H. Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3):1–25, 2011. URL https://www.jstatsoft.org/v40/i03/.

A. Guillenea, M. S. Lund, R. Evans, V. Boerner, and E. Karaman. A breed-of-origin of alleles model that includes crossbred data improves predictive ability for crossbred animals in a multi-breed population. *Genetics Selection Evolution*, 55 (1), 2023. doi: 10.1186/s12711-023-00806-1. URL http://dx.doi.org/10.1186/s12711-023-00806-1.

A. F. Gunnarsson, J. Zhu, B. C. Zhang, Z. Tsangalidou, A. Allmont, and P. F. Palamara. A scalable approach for genome-wide inference of ancestral recombination graphs. *bioRxiv*, 2024. doi: 10.1101/2024.08.31.610248. URL http://dx.doi.org/10.1101/2024.08.31.610248.

P. Guzmán-Luna, M. Mauricio-Iglesias, A. Flysjö, and A. Hospido. Analysing the interaction between the dairy sector and climate change from a life cycle perspective: A review. *Trends in Food Science Technology*, 126:168–179, August 2022. doi: 10.1016/j.tifs.2022.05.016. URL https://doi.org/10.1016/j.tifs.2022.05.016.

A. Götherström, C. Anderung, L. Hellborg, R. Elburg, C. Smith, D. Bradley, and H. Ellegren. Cattle domestication in the near east was followed by hybridization with aurochs bulls in europe. *Proceedings of the Royal Society B: Biological Sciences*, 272 (1579):2345–2350, 2005. doi: 10.1098/rspb.2005.3243. URL https://doi.org/10.1098/rspb.2005.3243.

D. Habier, R. L. Fernando, and J. C. M. Dekkers. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4):2389–2397, 12 2007. doi: 10.1534/genetics.107.081190. URL https://doi.org/10.1534/genetics.107.081190.

D. Habier, R. L. Fernando, and D. J. Garrick. Genomic blup decoded: A look into the black box of genomic prediction. *Genetics*, 194(3):597—-607, July 2013. doi: 10.1534/genetics.113.152207. URL http://dx.doi.org/10.1534/genetics.113.152207.

J. D. Hadfield and S. Nakagawa. General quantitative genetic methods for com-

parative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology*, 23(3):494—508, 2010. doi: 10.1111/j.1420-9101.2009.01915.x. URL http://dx.doi.org/10.1111/j.1420-9101.2009.01915.x.

O. Hanotte, D. G. Bradley, J. W. Ochieng, Y. Verjee, E. W. Hill, and J. Rege. African pastoralism: Genetic imprints of origins and migrations. *Science*, 296(5566):336–339, 2002. doi: 10.1126/science.1069878. URL https://doi.org/10.1126/science.1069878.

L. J. Harmon. *Phylogenetic Comparative Methods*. CreateSpace Independent Publishing Platform, 1.4 edition, March 2019. ISBN 978-1719584463.

K. Harris. Using enormous genealogies to map causal variants in space and time. *Nature Genetics*, 55(5):730—731, 2023. doi: 10.1038/s41588-023-01389-9. URL http://dx.doi.org/10.1038/s41588-023-01389-9.

R. Håvard, S. Martino, and M. Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009. doi: 10.1111/j.1467-9868.2008.00700.x. URL https://doi.org/10.1111/j.1467-9868.2008.00700.x.

B. Hayes, M. Carrick, P. Bowman, and M. Goddard. Genotype × environment interaction for milk production of daughters of australian dairy sires from test-day records. *Journal of Dairy Science*, 86(11):3736–3744, 2003. doi: 10.3168/jds.S0022-0302(03)73980-0. URL https://doi.org/10.3168/jds.S0022-0302(03)73980-0.

B. Hayes, H. Daetwyler, and M. Goddard. Model for genome x environment interaction: Examples in livestock. *Crop Science*, 56:1–9, 2016. doi: 10.2135/cropsci2015.07.0451. URL https://doi.org/10.2135/cropsci2015.07.0451.

B. J. Hayes and H. D. Daetwyler. 1000 bull genomes project to map simple and complex genetic traits in cattle: applications and outcomes. *Annual Review of Animal Biosciences*, 7:89–102, 2019. doi: 10.1146/annurev-animal-020518-115003. URL https://doi.org/10.1146/annurev-animal-020518-115003.

B. J. Hayes, J. Copley, E. M. Ross, S. Speight, and G. Fordyce. Multi-breed genomic evaluation for tropical beef cattle when no pedigree information is available. *Genet-*

ics, Selection, Evolution, pages 55–71, 2023. doi: 10.1186/s12711-023-00847-6. URL https://doi.org/10.1186/s12711-023-00847-6.

A. R. Hazel, B. J. Heins, and L. B. Hansen. Fertility and 305-day production of viking red-, montbéliarde-, and holstein-sired crossbred cows compared with holstein cows during their first 3 lactations in minnesota dairy herds. *Journal of Dairy Science*, 103(9):8683–8697, 2020. doi: 10.3168/jds.2020-18196. URL https://doi.org/10.3168/jds.2020-18196.

L. N. Hazel. The genetic basis for constructing selection indexes. *Genetics*, 28:476–490, 1943. doi: 10.1093/genetics/28.6.476. URL https://doi.org/10.1093/genetics/28.6.476.

C. R. Henderson. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics*, 32(1):69–83, 1976. doi: 10.2307/2529339. URL https://doi.org/10.2307/2529339.

C. R. Henderson. *Applications of Linear Models in Animal Breeding*. University of Guelph, Guelph, 3rd edition, 1984.

A. E. Hendrickson and P. O. White. Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, 17(1):65–70, 1964. doi: 10.1111/j.2044-8317.1964.tb00244.x. URL https://doi.org/10.1111/j.2044-8317.1964.tb00244.x.

E. Herrera-Luis, K. Benke, H. Volk, C. Ladd-Acosta, and G. L. Wojcik. Gene–environment interactions in human health. *Nature Reviews Genetics*, 25(11):768—784, 2024. doi: 10.1038/s41576-024-00731-z. URL http://dx.doi.org/10.1038/s41576-024-00731-z.

J. M. Hickey. Sequencing millions of animals for genomic selection 2.0. *Journal of Animal Breeding and Genetics*, 130(5):331—332, 2013. doi: 10.1111/jbg.12054. URL https://doi.org/10.1111/jbg.12054.

J. M. Hickey, S. Dreisigacker, J. Crossa, S. Hearne, R. Babu, et al. Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. *Crop Science*, 54(4):1476—-1488, July 2014. ISSN 1435-0653. doi: 10.2135/cropsci2013.03.0195. URL http://dx.doi.org/10.2135/cropsci2013.03.0195.

J. M. Hickey, T. Chiurugwi, I. J. Mackay, W. Powell, A. Eggen, et al. Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nature Genetics*, 49(9):1297–1303, 2017. doi: 10.1038/ng.3920.

W. G. Hill, M. E. Goddard, and P. M. Visscher. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics*, 4(2):e1000008, 2008. doi: 10.1371/journal.pgen.1000008. URL http://dx.doi.org/10.1371/journal.pgen.1000008.

V. Hivert, J. Sidorenko, F. Rohart, M. E. Goddard, J. Yang, N. R. Wray, L. Yengo, and P. M. Visscher. Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. *The American Journal of Human Genetics*, 108(5):786—-798, 2021. doi: 10.1016/j.ajhg.2021.02.014. URL http://dx.doi.org/10.1016/j.ajhg.2021.02.014.

R. Hofmeister, D. Ribeiro, S. Rubinacci, and O. Delaneau. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the uk biobank. *Nature Genetics*, 2023. doi: 10.1038/s41588-023-01415-w. URL https://doi.org/10.1038/s41588-023-01415-w.

Holstein Association USA, Inc. Holstein association usa, inc., 2021. URL https://www.holsteinusa.com. Accessed: 2021.

F. Hormozdiari, A. Zhu, G. Kichaev, C. J.-T. Ju, A. V. Segrè, J. W. J. Joo, H. Won, S. Sankararaman, B. Pasaniuc, S. Shifman, and E. Eskin. Widespread allelic heterogeneity in complex traits. *The American Journal of Human Genetics*, 100(5):789—-802, 2017. doi: 10.1016/j.ajhg.2017.04.005. URL http://dx.doi.org/10.1016/j.ajhg.2017.04.005.

K. Hou, Y. Ding, Z. Xu, Y. Wu, A. Bhattacharya, R. Mester, G. M. Belbin, S. Buyske, D. V. Conti, B. F. Darst, M. Fornage, C. Gignoux, X. Guo, C. Haiman, E. E. Kenny, M. Kim, C. Kooperberg, L. Lange, A. Manichaikul, K. E. North, U. Peters, L. J. Rasmussen-Torvik, S. S. Rich, J. I. Rotter, H. E. Wheeler, G. L. Wojcik, Y. Zhou, S. Sankararaman, and B. Pasaniuc. Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. *Nature Genetics*, 55(4):549—-558, 2023. doi: 10.1038/s41588-023-01338-6. URL http://dx.doi.org/10.1038/s41588-023-01338-6.

K. Hou, Z. Xu, Y. Ding, R. Mandla, Z. Shi, K. Boulier, A. Harpak, and B. Pasaniuc. Calibrated prediction intervals for polygenic scores across diverse contexts.

Nature Genetics, 56(7):1386—-1396, June 2024. ISSN 1546-1718. doi: 10.1038/s41588-024-01792-w. URL http://dx.doi.org/10.1038/s41588-024-01792-w.

S. Hu, L. A. F. Ferreira, S. Shi, G. Hellenthal, J. Marchini, D. J. Lawson, and S. R. Myers. Fine-scale population structure and widespread conservation of genetic effect sizes between human groups across traits. *Nature Genetics*, 57(2):379—-389, Feb. 2025. ISSN 1546-1718. doi: 10.1038/s41588-024-02035-8. URL http://dx.doi.org/10.1038/s41588-024-02035-8.

D. Hunde, Y. Tadesse, M. Tadesse, S. Abegaz, and T. Getachew. Community-based breeding programs can realize sustainable genetic gain and economic benefits in tropical dairy cattle systems. *Frontiers in Genetics*, 15, May 2024. doi: 10.3389/fgene.2024.1106709. URL https://doi.org/10.3389/fgene.2024.1106709.

B. H. Hunnicutt. Zebu cattle in brazil: Imported stock crossed on native—hybrids are popular with ranchers—hardy, disease-resistant and fairly good milkers—high prices paid—possibilities of interest to united states. *Journal of Heredity*, 6(5):195–201, May 1915. doi: 10.1093/oxfordjournals.jhered.a109101. URL https://doi.org/10.1093/oxfordjournals.jhered.a109101.

C. H. Hunt, B. J. Hayes, F. A. van Eeuwijk, E. S. Mace, and D. R. Jordan. Multi-environment analysis of sorghum breeding trials using additive and dominance genomic relationships. *Theoretical and Applied Genetics*, 133:1009–1018, 2020. doi: 10.1007/s00122-019-03526-7. URL https://doi.org/10.1007/s00122-019-03526-7.

A. Ignatieva, M. Favero, J. Koskela, J. Sant, and S. R. Myers. The length of haplotype blocks and signals of structural variation in reconstructed genealogies. *bioRxiv*, 2025. doi: 10.1101/2023.07.11.548567. URL https://doi.org/10.1101/2023.07.11.548567.

J. Jenko, M. C. McClure, D. Matthews, J. McClure, M. Johnsson, G. Gorjanc, and J. M. Hickey. Analysis of a large dataset reveals haplotypes carrying putatively recessive lethal and semi-lethal alleles with pleiotropic effects on economically important traits in beef cattle. *Genetics Selection Evolution*, 51:10, 2019. doi: 10.1186/s12711-019-0452-z. URL https://doi.org/10.1186/s12711-019-0452-z.

R. I. Jennrich. A simple general method for oblique rotation. *Psychometrika*, 66 (2):289–306, 2001. doi: 10.1007/BF02294844. URL https://doi.org/10.1007/BF02294844.

R. I. Jennrich and P. F. Sampson. Rotation for simple loadings. *Psychometrika*, 31 (3):313–323, 1966. doi: 10.1007/BF02289465. URL https://doi.org/10.1007/BF02289465.

M. Johnsson. The big challenge for livestock genomics is to make sequence data pay. *arXiv*, 2023a. doi: 10.48550/arXiv.2302.01140. URL https://doi.org/10.48550/arXiv.2302.01140.

M. Johnsson. Genomics in animal breeding from the perspectives of matrices and molecules. *Hereditas*, 160:20, 2023b. doi: 10.1186/s41065-023-00220-0. URL https://hereditasjournal.biomedcentral.com/articles/10.1186/s41065-023-00220-0.

A. G. Jones, R. Bürger, and S. J. Arnold. Epistasis and natural selection shape the mutational architecture of complex traits. *Nature Communications*, 5(1), 2014. doi: 10.1038/ncomms4709. URL http://dx.doi.org/10.1038/ncomms4709.

K. G. Jöreskog. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2):183–202, 1969. doi: 10.1007/BF02289343. URL https://doi.org/10.1007/BF02289343.

V. S. Junqueira, P. S. Lopes, D. Lourenco, F. F. e. Silva, and F. F. Cardoso. Applying the metafounders approach for genomic evaluation in a multibreed beef cattle population. *Frontiers in Genetics*, 11, Dec. 2020. ISSN 1664-8021. doi: 10.3389/fgene.2020.556399. URL http://dx.doi.org/10.3389/fgene.2020.556399.

L. Kachuri, N. Chatterjee, J. Hirbo, D. J. Schaid, I. Martin, et al. Principles and methods for transferring polygenic risk scores across global populations. *Nature Reviews Genetics*, 25(1):8—25, 2023. doi: 10.1038/s41576-023-00637-2. URL http://dx.doi.org/10.1038/s41576-023-00637-2.

H. F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958. doi: 10.1007/BF02289233. URL https://doi.org/10.1007/BF02289233.

M. S. Kang. Using genotype-by-environment interaction for crop cultivar development. volume 62 of *Advances in Agronomy*, pages 199–252. Academic Press, 1997. doi: https://doi.org/10.1016/S0065-2113(08)60569-6. URL https://www.sciencedirect.com/science/article/pii/S0065211308605696.

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

E. Kathambi, T. Sonstegard, and P. Larsen. Review: Cross-breeding, advanced reproductive technologies, and genetic selection in twelve dairy production systems in africa. *animal*, 19(3):101424, 2025. doi: 10.1016/j.animal.2025.101424. URL https://doi.org/10.1016/j.animal.2025.101424.

K. Katoh, K. Misawa, K. Kuma, and T. Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, 30 (14):3059–3066, 2002. doi: 10.1093/nar/gkf436. URL https://doi.org/10.1093/nar/gkf436.

P. D. Keightley and B. C. Jackson. Inferring the probability of the derived vs. the ancestral allelic state at a polymorphic site. *Genetics*, 209:897–906, 2018. doi: 10.1534/genetics.118.300853. URL https://doi.org/10.1534/genetics.118.300853.

J. Kelleher, Y. Wong, A. W. Wohns, C. Fadil, P. K. Albers, and G. McVean. Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51: 1330–1338, 2019. doi: 10.1038/s41588-019-0483-y. URL https://doi.org/10.1038/s41588-019-0483-y. Erratum in: Nat Genet. 2019 Nov;51(11):1660. doi: 10.1038/s41588-019-0523-7.

R. A. Kempton. The use of biplots in interpreting variety by environment interactions. *The Journal of Agricultural Science*, 103(1):123–135, 1984. doi: 10.1017/S0021859600043392. URL https://doi.org/10.1017/S0021859600043392.

K. Kim, T. Kwon, T. Dessie, D. Yoo, J. O. Mwai, et al. The mosaic genome of indigenous african cattle as a unique genetic resource for african pastoralism. *Nature Genetics*, 52:1099–1110, 2020. doi: 10.1038/s41588-020-0694-2. URL https://doi.org/10.1038/s41588-020-0694-2.

M. Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893–903, 1969. doi: 10.1093/genetics/61.4.893. URL https://doi.org/10.1093/genetics/61.4.893.

J. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982.

J. F. C. Kingman. On the properties of bilinear models for the balance between genetic mutation and selection. *Mathematical Proceedings of the Cambridge Philosophical Society*, 81(3):443–453, 1977. doi: 10.1017/s0305004100053512. URL http://dx.doi.org/10.1017/S0305004100053512.

R. Knight. The measurement and interpretation of genotype–environment interactions. *Euphytica*, 19:225–235, 1970. doi: 10.1007/BF00052023. URL https://doi.org/10.1007/BF00052023.

S. Kona, G. Ravikiran, P. Sasidhar, A. Sivakumar, and V. Rao. Perspectives in milk production in india. *Theriogenology*, 231:116–126, 2025. doi: 110.1016/j.theriogenology.2024.10.001. URL https://doi.org/10.1016/j.theriogenology.2024.10.001.

L. A. d. C. Lara, I. Pocrnic, T. d. P. Oliveira, I. Misztal, and D. A. L. Lourenco. Temporal and genomic analysis of additive genetic variance in breeding programmes. *Heredity*, 128:21–32, 2022. doi: 10.1038/s41437-021-00485-y. URL https://doi.org/10.1038/s41437-021-00485-y.

H. Lee, N. Pope, J. Kelleher, G. Gorjanc, and P. Ralph. Modelling complex traits with ancestral recombination graphs. *bioRxiv [Preprint]*, July 2025. doi: 10.1101/2025.07.14.664631. URL https://doi.org/10.1101/2025.07.14.664631.

A. Legarra, C. A. Garcia-Baccino, Y. C. J. Wientjes, and Z. G. Vitezica. The correlation of substitution effects across populations and generations in the presence of nonadditive functional gene action. *Genetics*, 219(4), Aug. 2021. ISSN 1943-2631. doi: 10.1093/genetics/iyab138. URL http://dx.doi.org/10.1093/genetics/iyab138.

B. Lehmann, H. Lee, L. Anderson-Trocmé, J. Kelleher, G. Gorjanc, and P. L. Ralph. On args, pedigrees, and genetic relatedness matrices. *bioRxiv [Preprint]*, 2025:03.03.641310, March 2025. doi: 10.1101/2025.03.03.641310. URL https://doi.org/10.1101/2025.03.03.641310. [Version 1].

G. Leroy, R. Baumung, P. Boettcher, B. Scherf, and I. Hoffmann. Review: Sustainability of crossbreeding in developing countries; definitely not like crossing a meadow... *Animal*, 10(2):262–273, 2016. doi: 10.1017/S175173111500213X. URL https://doi.org/10.1017/S175173111500213X.

A. Lewanski, M. Grundler, and G. Bradburd. The era of the arg: An introduction to ancestral recombination graphs and their significance in empirical evolutionary

genomics. *PLOS Genetics*, 2024. doi: 10.1371/journal.pgen.1011110. URL https://doi.org/10.1371/journal.pgen.1011110.

Y.-S. Lin, T. Tan, Y. Wang, B. Pasaniuc, A. R. Martin, and E. G. Atkinson. Differential performance of polygenic prediction across traits and populations depending on genotype discovery approach. *bioRxiv*, 2025. doi: 10.1101/2025.03.18.644029. URL http://dx.doi.org/10.1101/2025.03.18.644029.

G. Ling, D. Miller, R. Nielsen, and A. Stern. A bayesian framework for inferring the influence of sequence context on point mutations. *Molecular Biology and Evolution*, 37(3):893–903, March 2020. doi: 10.1093/molbev/msz248. URL https://doi.org/10.1093/molbev/msz248.

V. Link, J. G. Schraiber, C. Fan, B. Dinh, N. Mancuso, C. W. Chiang, and M. D. Edge. Tree-based qtl mapping with expected local genetic relatedness matrices. *The American Journal of Human Genetics*, 110(12):2077—2091, 2023. doi: 10.1016/j.ajhg.2023.10.017. URL http://dx.doi.org/10.1016/j.ajhg.2023.10.017.

L. L. Lo, R. L. Fernando, and M. Grossman. Covariances of relatives in multibreed populations: additive model. *Theoretical and Applied Genetics*, 87(4):423–430, 1993. doi: 10.1007/BF01184927. URL https://doi.org/10.1007/BF01184927.

R. Loftus, D. MacHugh, D. Bradley, P. Sharp, and P. Cunningham. Evidence for two independent domestications of cattle. *Proceedings of the National Academy of Sciences*, 91(7):2757–2761, 1994. doi: 10.1073/pnas.91.7.2757. URL https://doi.org/10.1073/pnas.91.7.2757.

M. Londoño-Gil, J. Hidalgo, A. Legarra, C. U. Magnabosco, F. Baldi, and D. Lourenco. Indirect genomic predictions for indicine cattle breeds with snd effects from a multibreed genomic evaluation. *Journal of Animal Breeding and Genetics*, 2025. doi: 10.1111/jbg.70008. URL http://dx.doi.org/10.1111/jbg.70008.

J. L. Lush, J. C. Holbert, and O. S. Willham. Genetic history of the holstein-friesian cattle in the united states. *Journal of Heredity*, 27(2):61–72, February 1936. doi: 10.1093/oxfordjournals.jhered.a104174. URL https://doi.org/10.1093/oxfordjournals.jhered.a104174.

M. Lynch. Methods for the analysis of comparative data in evolutionary biology. *Evolution*, 45(5):1065–1080, 1991. doi: 10.1111/j.1558-5646.1991.tb04375.x. URL https://doi.org/10.1111/j.1558-5646.1991.tb04375.x.

M. Lynch and W. G. Hill. Phenotypic evolution by neutral mutation. *Evolution*, 40 (5):915–935, Sept. 1986. ISSN 1558-5646. doi: 10.1111/j.1558-5646.1986.tb00561.x. URL http://dx.doi.org/10.1111/j.1558-5646.1986.tb00561.x.

L. Ma, J. R. O'Connell, P. M. VanRaden, B. Shen, A. Padhi, C. Sun, D. M. Bickhart, J. B. Cole, D. J. Null, G. E. Liu, Y. Da, and G. R. Wiggans. Cattle sex-specific recombination and genetic control from a large pedigree analysis. *PLOS Genetics*, 11(11):e1005387, 2015. doi: 10.1371/journal.pgen.1005387. URL https://doi.org/10.1371/journal.pgen.1005387.

D. MacHugh, M. Shriver, R. Loftus, P. Cunningham, and D. Bradley. Microsatellite dna variation and the evolution, domestication and phylogeography of taurine and zebu cattle (bos taurus and bos indicus). *Genetics*, 146(3):1071–1086, 1997. doi: 10.1093/genetics/146.3.1071. URL https://doi.org/10.1093/genetics/146.3.1071.

T. F. C. Mackay. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature Reviews Genetics*, 15(1):22–33, January 2014. doi: 10.1038/nrg3627. URL https://doi.org/10.1038/nrg3627.

I. M. MacLeod, D. M. Larkin, H. A. Lewin, B. J. Hayes, and M. E. Goddard. Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors. *Molecular Biology and Evolution*, 30(9):2209–2223, September 2013. doi: 10.1093/molbev/mst125. URL https://doi.org/10.1093/molbev/mst125.

F. Madalena, R. Teodoro, A. Lemos, J. Monteiro, and R. Barbos. Evaluation of strategies for crossbreeding of dairy cattle in brazil. *Journal of Dairy Science*, 73(7):1887–1901, 1990. doi: 10.3168/jds.S0022-0302(90)78869-8. URL https://doi.org/10.3168/jds.S0022-0302(90)78869-8.

A. G. Mahler. What/where is the global south? *Oxford Bibliographies in Literary and Critical Theory*, June 2017. Accessed: 9 June 2025.

A. Maiorano, D. Lourenco, S. Tsuruta, A. Ospina, N. Stafuzza, Y. Masuda, A. Vercesi Filho, J. Cyrillo, R. Curi, and J. Silva. Assessing genetic architecture and signatures of selection of dual purpose gir cattle populations using genomic information. *PLOS ONE*, 2018. doi: 10.1371/journal.pone.0200694. URL https://doi.org/10.1371/journal.pone.0200694.

A. Mäki-Tanila and W. G. Hill. Influence of gene interaction on complex trait variation with multilocus models. *Genetics*, 198(1):355—367, July 2014. ISSN 1943-2631. doi: 10.1534/genetics.114.165282. URL http://dx.doi.org/10.1534/genetics.114.165282.

J. Mandel. A new analysis of variance model for non-additive data. *Technometrics*, 13 (1):1–18, 1971. URL https://doi.org/10.1080/00401706.1971.10488751.

S. B. Manuck. The reaction norm in gene-environment interaction. *Molecular Psychiatry*, 15(9):881–882, 2009. doi: 10.1038/mp.2009.139. URL https://doi.org/10.1038/mp.2009.139. Author manuscript; available in PMC: 2011 Mar 1.

K. Marshall, G. R. Salmon, S. Tebug, J. Juga, M. MacLeod, J. Poole, I. Baltenweck, and A. Missohou. Net benefits of smallholder dairy cattle farms in senegal can be significantly increased through the use of better dairy cattle breeds and improved management practices. *Journal of Dairy Science*, 102(11):10080–10094, 2019. doi: 10.3168/jds.2019-17334. URL https://doi.org/10.3168/jds.2019-17334.

A. R. Martin, C. R. Gignoux, R. K. Walters, G. L. Wojcik, B. M. Neale, S. Gravel, M. J. Daly, C. D. Bustamante, and E. E. Kenny. Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, 100(4):635—649, 2017. doi: 10.1016/j.ajhg.2017.03.004. URL http://dx.doi.org/10.1016/j.ajhg.2017.03.004.

S. B. McClure. Domesticated animals and biodiversity: Early agriculture at the gates of europe and long-term ecological consequences. *Anthropocene*, 4:57–68, 2013. doi: 10.1016/j.ancene.2013.11.001. URL https://doi.org/10.1016/j.ancene.2013.11.001.

R. E. McDowell. Crossbreeding in tropical areas with emphasis on milk, health, and fitness. *Journal of Dairy Science*, 68(9):2418–2435, 1985. doi: 10.3168/jds.S0022-0302(85)81118-8. URL https://doi.org/10.3168/jds.S0022-0302(85)81118-8.

G. McHugo, J. Ward, S. Ng'ang'a, et al. Genome-wide local ancestry and the functional consequences of admixture in african and european cattle populations. *Heredity*, 134: 49–63, 2025. doi: 10.1038/s41437-024-00734-w. URL https://doi.org/10.1038/s41437-024-00734-w.

A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky,

K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research*, 20(9):1297–1303, September 2010. doi: 10.1101/gr.107524.110. URL https://doi.org/10.1101/gr.107524.110.

J. Merilä and B. Sheldon. Genetic architecture of fitness and nonfitness traits: empirical patterns and development of ideas. *Heredity*, 83:103—-109, 1999. doi: 10.1046/j.1365-2540.1999.00585.x. URL https://doi.org/10.1046/j.1365-2540.1999.00585.x.

T. Meuwissen and M. Goddard. Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics*, 185(2):623—631, 2010. doi: 10.1534/genetics.110.116590. URL http://dx.doi.org/10.1534/genetics.110.116590.

T. Meuwissen, B. Hayes, and M. Goddard. Genomic selection: A paradigm shift in animal breeding. *Animal Frontiers*, 6(1):6–14, January 2016. doi: 10.2527/af.2016-0002. URL https://doi.org/10.2527/af.2016-0002.

T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157:1819–1829, 2001. doi: 10.1093/genetics/157.4.1819. URL https://doi.org/10.1093/genetics/157.4.1819.

K. Meyer. Factor-analytic models for genotype × environment problems. *Genetics Selection Evolution*, 39(3):303–313, 2007. doi: 10.1051/gse:2007005. URL https://doi.org/10.1051/gse:2007005.

K. Meyer. Factor-analytic models for animal breeding data: A review. *Animal*, 3(9):1279–1288, 2009a. doi: 10.1017/S1751731109004641. URL https://doi.org/10.1017/S1751731109004641.

K. Meyer. Factor-analytic models for genotype × environment type problems and structured covariance matrices. *Genetics Selection Evolution*, 41:21, 2009b. doi: 10.1186/1297-9686-41-21. URL https://doi.org/10.1186/1297-9686-41-21.

P. Michael, C. R. de Cruz, N. Mohd Nor, S. Jamli, and Y. M. Goh. The potential of using temperate–tropical crossbreds and agricultural by-products, associated with heat stress management for dairy production in the tropics: A review. *Animals*, 12(1):1, 2022. doi: 10.3390/ani12010001. URL https://doi.org/10.3390/ani12010001.

F. Miglior, B. L. Muir, and B. J. Van Doormaal. Selection indices in holstein cattle of

various countries. *Journal of Dairy Science*, 88:1255–1263, 2005. doi: 10.3168/jds.S0022-0302(05)72792-2.

A. M. Miles, K. L. Parker Gaddis, J. B. Cole, and R. H. Fourdraine. The role of a national evaluation system in promoting dairy sustainability. *JDS Communications*, 6(3):458–463, May 2025. doi: 10.3168/jdsc.2024-0645. URL https://doi.org/10.3168/jdsc.2024-0645. EAAP/ADSA Breeding & Genetics Committee Symposium Review.

I. Misztal. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics*, pages 401–409, 2015. doi: 10.1534/genetics.115.182089. URL https://doi.org/10.1534/genetics.115.182089.

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

I. Misztal, A. Legarra, and I. Aguilar. Using recursion to compute the inverse of the genomic relationship matrix. *Journal of Dairy Science*, pages 3943–3952, 2014. doi: 10.3168/jds.2013-7752. URL https://doi.org/10.3168/jds.2013-7752.

I. Misztal, Y. Steyn, and D. Lourenco. Genomic evaluation with multibreed and crossbred data. *JDS Communications*, 3(2):156–159, Mar 2022. doi: 10.3168/jdsc.2021-0134. URL https://doi.org/10.3168/jdsc.2021-0134.

T. Miyasaka, S.-N. Takeshima, H. Sentsui, and Y. Aida. Identification and diversity of bovine major histocompatibility complex class ii haplotypes in japanese black and holstein cattle in japan. *Journal of Dairy Science*, 95(1):420–431, 2012. doi: 1doi.org/10.3168/jds.2011-4621. URL https://doi.org/10.3168/jds.2011-4621.

R. H. Moll and C. W. Stuber. Quantitative genetics–empirical results relevant to plant breeding. In N. C. Brady, editor, *Advances in Agronomy*, volume 26, pages 277–313. Academic Press, 1974. doi: 10.1016/S0065-2113(08)60801-X. URL https://doi.org/10.1016/S0065-2113(08)60801-X.

R. A. Mrode and I. Pocrnic. *Linear Models for the Prediction of the Genetic Merit of Animals*. CABI, 4th edition, 2023. doi: 10.1079/9781800620506.0000. URL http://dx.doi.org/10.1079/9781800620506.0000.

H. Mulder. Is gxe a burden or a blessing? opportunities for genomic selection and big data. *Journal of Animal Breeding and Genetics*, 134(6):435–436, 2017. doi: 10.1111/jbg.12303. URL https://doi.org/10.1111/jbg.12303.

H. A. Mulder. Genomic selection improves response to selection in resilience by exploiting genotype by environment interactions. *Frontiers in Genetics*, 7:178, 2016. doi: 10.3389/fgene.2016.00178. URL https://doi.org/10.3389/fgene.2016.00178.

H. A. Mulder and P. Bijma. Effects of genotype × environment interaction on genetic gain in breeding programs. *Journal of Animal Science*, 83(1):49–61, 2005. doi: 10.2527/2005.83149x. URL https://doi.org/10.2527/2005.83149x.

H. A. Mulder and P. Bijma. Effects of genotype × environment interaction on genetic gain in breeding programs. *Journal of Animal Science*, 89(7):2056–2069, 2011. doi: 10.2527/jas.2010-3557. URL https://doi.org/10.2527/jas.2010-3557.

M. Naval-Sánchez, L. Porto-Neto, D. Cardoso, et al. Selection signatures in tropical cattle are enriched for promoter and coding regions and reveal missense mutations in the damage response gene helb. *Genetics Selection Evolution*, 52:27, 2020. doi: 10.1186/s12711-020-00546-6. URL https://doi.org/10.1186/s12711-020-00546-6.

R. Nielsen, A. Vaughn, and Y. Deng. Inference and applications of ancestral recombination graphs. *Nature Reviews Genetics*, 26:47–58, 2025. doi: 10.1038/s41576-024-00772-4. URL https://doi.org/10.1038/s41576-024-00772-4.

M. A. Nilforooshan. Short communication: Reduced gblup equations to core animals in the algorithm for proven and young (apy). *Veterinary and Animal Science*, 23:100334, 2024. doi: 10.1016/j.vas.2024.100334. URL https://doi.org/10.1016/j.vas.2024.100334.

J. Novembre and N. H. Barton. Tread lightly interpreting polygenic tests of selection. *Genetics*, 208(4):1351—-1355, 2018. doi: 10.1534/genetics.118.300786. URL http://dx.doi.org/10.1534/genetics.118.300786.

K. Nunes et al. Admixture's impact on brazilian population evolution and health. *Science*, 388:eadl3564, 2025. doi: 10.1126/science.adl3564. URL https://doi.org/10.1126/science.adl3564.

OECD/Food and Agriculture Organization of the United Nations. *OECD-FAO Agri-*

cultural Outlook 2015. Paris, 2015. URL http://dx.doi.org/10.1787/agr_outlook-2015-en.

A. R. Ogunbawo, H. A. Mulim, G. S. Campos, A. P. Schinckel, and H. R. Oliveira. Tailoring genomic selection for bos taurus indicus: A comprehensive review of snp arrays and reference genomes. *Genes (Basel)*, 15(12):1495, November 2024. doi: 10.3390/genes15121495. URL https://doi.org/10.3390/genes15121495.

M. Oman, A. Alam, and R. W. Ness. How sequence context-dependent mutability drives mutation rate variation in the genome. *Genome Biology and Evolution*, 14(3): evac032, March 2022. doi: 10.1093/gbe/evac032. URL https://doi.org/10.1093/gbe/evac032.

C. J. O'Neill, D. L. Swain, and H. N. Kadarmideen. Evolutionary process of bos taurus cattle in favourable versus unfavourable environments and its implications for genetic selection. *Evolutionary Applications*, 3:422–433, 2010. doi: 10.1111/j.1752-4571.2010.00151.x. URL https://doi.org/10.1111/j.1752-4571.2010.00151.x.

S. Otto and B. Payseur. Crossover interference: Shedding light on the evolution of recombination. *Annual Review of Genetics*, 53:19–44, 2019. doi: 10.1146/annurev-genet-040119-093957. URL https://doi.org/10.1146/annurev-genet-040119-093957. First published as a Review in Advance on August 20, 2019.

T. Paim, E. Hay, C. Wilson, M. Thomas, L. Kuehn, S. Paiva, C. McManus, and H. Blackburn. Dynamics of genomic architecture during composite breed development in cattle. *Animal Genetics*, 51:224–234, 2020. doi: 10.1111/age.12907. URL https://doi.org/10.1111/age.12907.

J. C. d. C. Panetto, M. V. G. B. Silva, R. d. S. Verneque, M. A. Machado, A. R. Fernandes, et al. Programa nacional de melhoramento do gir leiteiro - sumário brasileiro de touros; 4ª avaliação genômica de touros - resultado do teste de progênie, May 2021.

J. C. d. C. Panetto et al., editors. *Programa Nacional de Melhoramento do Gir Leiteiro - sumário brasileiro de touros – 8a avaliação genômica de touros - resultado do teste de progênie - abril 2025*. Number 295 in Documentos / Embrapa Gado de Leite. Embrapa Gado de Leite, Juiz de Fora, 2025. URL https://girleiteiro.org.br/arquivos/2966.pdf. ISSN 1516-7453.

Y. Park, B. P. H. Metzger, and J. W. Thornton. Epistatic drift causes gradual decay of predictability in protein evolution. *Science*, 376(6595):823–830, 2022. doi: 10.1126/science.abn6895. URL https://www.science.org/doi/10.1126/science.abn6895.

J. M. Perkins and J. L. Jinks. Environmental and genotype–environmental components of variability. iii. multiple lines and crosses. *Heredity*, 23(3):339–356, 1968. doi: 10.1038/hdy.1968.35. URL https://doi.org/10.1038/hdy.1968.35.

D. Pitt, N. Sevane, E. L. Nicolazzi, D. E. MacHugh, S. D. E. Park, L. Colli, R. Martinez, M. W. Bruford, and P. Orozco-terWengel. Domestication of cattle: Two or three events? *Evolutionary Applications*, 12(1):123–136, 2018. doi: 10.1111/eva.12674. URL https://doi.org/10.1111/eva.12674.

F. Privé, H. Aschard, S. Carmi, L. Folkersen, C. Hoggart, P. F. O'Reilly, and B. J. Vilhjálmsson. Portability of 245 polygenic scores when derived from the uk biobank and applied to 9 ancestry groups from the same cohort. *The American Journal of Human Genetics*, 109(1):12—23, Jan. 2022. ISSN 0002-9297. doi: 10.1016/j.ajhg.2021.11.008. URL http://dx.doi.org/10.1016/j.ajhg.2021.11.008.

S. Qanbari, E. C. G. Pimentel, J. Tetens, G. Thaller, P. Lichtner, A. R. Sharifi, and H. Simianer. The pattern of linkage disequilibrium in german holstein cattle. *Animal Genetics*, 41(4):346—-356, 2010. doi: 10.1111/j.1365-2052.2009.02011.x. URL http://dx.doi.org/10.1111/j.1365-2052.2009.02011.x.

P. Ralph, K. Thornton, and J. Kelleher. Efficiently summarizing relationships in large samples: A general duality between statistics of genealogies and genomes. *Genetics*, 215(3):779—797, 2020. doi: 10.1534/genetics.120.303253. URL http://dx.doi.org/10.1534/genetics.120.303253.

M. D. Rasmussen, M. J. Hubisz, I. Gronau, and A. Siepel. Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, 10(5):e1004342, 2014. doi: 10.1371/journal.pgen.1004342. URL https://doi.org/10.1371/journal.pgen.1004342.

B. Raymond, A. C. Bouwman, Y. C. J. Wientjes, C. Schrooten, J. Houwing-Duistermaat, and R. F. Veerkamp. Genomic prediction for numerically small breeds, using models with pre-selected and differentially weighted markers. *Genetics Selection Evolution*, 50(1), 2018. doi: 10.1186/s12711-018-0419-5. URL http://dx.doi.org/10.1186/s12711-018-0419-5.

I. Rebollo, D. Tolhurst, J. Obšteter, J. E. Rosas, and G. Gorjanc. Leveraging ancestral recombination graphs for quantitative genetic analysis of rice yield in indica and japonica subspecies. *bioRxiv*, 2025. doi: 10.1101/2025.01.14.633033. URL http://dx.doi.org/10.1101/2025.01.14.633033.

J. Reis Filho, P. Lopes, R. Verneque, R. Torres, R. Teodoro, and P. Carneiro. Population structure of brazilian gyr dairy cattle. *Revista Brasileira de Zootecnia*, 39(12):2640–2645, 2010. doi: 10.1590/S1516-35982010001200012. URL https://doi.org/10.1590/S1516-35982010001200012.

S. Rio, T. Mary-Huard, L. Moreau, C. Bauland, C. Palaffre, D. Madur, V. Combes, and A. Charcosset. Disentangling group specific qtl allele effects from genetic background epistasis using admixed individuals in gwas: An application to maize flowering. *PLoS Genetics*, 16(3):e1008241, Mar 2020. doi: 10.1371/journal.pgen.1008241. URL https://doi.org/10.1371/journal.pgen.1008241.

A. Robertson. The sampling variance of the genetic correlation coefficient. *Biometrics*, 15(3):469–485, 1959. doi: 10.2307/2527750. URL https://doi.org/10.2307/2527750.

K. B. Rodgers, S.-S. Lee, N. Rosenbloom, A. Timmermann, G. Danabasoglu, et al. Ubiquity of human-induced changes in climate variability. *Earth System Dynamics*, 12(4):1393–1411, Dec 2021. doi: 10.5194/esd-12-1393-2021. URL https://doi.org/10.5194/esd-12-1393-2021.

R. Ros-Freixedes. The contribution of whole-genome sequence data to genome-wide association studies in livestock: Outcomes and perspectives. *Livestock Science*, 281: 105430, 2024. doi: 10.1016/j.livsci.2024.105430. URL http://dx.doi.org/10.1016/j.livsci.2024.105430.

R. Ros-Freixedes, M. Johnsson, A. Whalen, C. Y. Chen, B. D. Valente, W. O. Herring, G. Gorjanc, and J. M. Hickey. Genomic prediction with whole-genome sequence data in intensely selected pig lines. *Genetics Selection Evolution*, 54(65), 2022a. doi: 10.1186/s12711-022-00756-0. URL https://doi.org/10.1186/s12711-022-00756-0.

R. Ros-Freixedes et al. Rare and population-specific functional variation across pig lines. *Genetics Selection Evolution*, 54:47, 2022b. doi: 10.1186/s12711-022-00732-8. URL https://doi.org/10.1186/s12711-022-00732-8.

R. Roschinsky, M. Kluszczynska, J. Sölkner, R. Puskur, and M. Wurzinger. Small-

holder experiences with dairy cattle crossbreeding in the tropics: from introduction to impact. *Animal*, 9(1):150–157, 2015. doi: 10.1017/S1751731114000901.

C. Rossi, M. Sinding, V. Mullin, et al. The genomic natural history of the aurochs. *Nature*, 635:136–141, 2024. doi: 10.1038/s41586-024-08112-6. URL https://doi.org/10.1038/s41586-024-08112-6.

H. Rue and L. Held. *Gaussian Markov Random Fields*. Chapman and Hall/CRC, 2005. doi: 10.1201/9780203492024. URL http://dx.doi.org/10.1201/9780203492024.

H. Rue, A. I. Riebler, S. H. Sørbye, J. B. Illian, D. P. Simpson, and F. K. Lindgren. Bayesian computing with INLA: A review. *Annual Reviews of Statistics and Its Applications*, 4(March):395–421, 2017. URL http://arxiv.org/abs/1604.00860.

J. J. Rutledge. Greek temples, tropical kine and recombination load. *Livestock Production Science*, 68(2-3):171–179, March 2001. doi: 10.1016/S0301-6226(00)00245-1. URL https://doi.org/10.1016/S0301-6226(00)00245-1.

E. Sánchez-Molano, V. V. Kapsona, J. J. Ilska, S. Desire, J. Conington, S. Mucha, and G. Banos. Genetic analysis of novel phenotypes for farm animal resilience. *Journal of Animal Breeding and Genetics*, 2023. In press.

G. Sandler and R. York. Epistasis and deep learning in quantitative genetics. *Arcadia Science*, 2025. doi: 10.57844/arcadia-25nt-guw3. URL https://research.arcadiascience.com/pub/result-gp-deep-learning-scaling/release/1.

M. Santana Jr, R. Pereira, A. Bignardi, L. El Faro, H. Tonhat, and L. Albuquerque. History, structure, and genetic diversity of brazilian gir cattle. *Livestock Science*, 163:26–33, 2014. doi: 10.1016/j.livsci.2014.02.007. URL http://dx.doi.org/10.1016/j.livsci.2014.02.007.

C. Sartori, F. Tiezzi, N. Guzzo, E. Mancin, B. Tuliozi, and R. Mantovani. Genotype by environment interaction and selection response for milk yield traits and conformation in a local cattle breed using a reaction norm approach. *Animals*, 12(7):839, Mar 2022. doi: 10.3390/ani12070839. URL https://doi.org/10.3390/ani12070839.

C. G. Scanes. The neolithic revolution, animal domestication, and early forms of animal agriculture. In *Animals and Human Society*, pages 103–131. Elsevier, 2018. doi: 10.1016/B978-0-12-805247-1.00006-X.

L. R. Schaeffer. Application of random regression models in animal breeding. *Livestock*

Production Science, 86:35–45, 2004. doi: 10.1016/S0301-6226(03)00151-9. URL https://doi.org/10.1016/S0301-6226(03)00151-9.

J. G. Schraiber, J. P. Spence, and M. D. Edge. Estimation of demography and mutation rates from one million haploid genomes. *bioRxiv [Preprint]*, September 2024. doi: 10.1101/2024.09.18.613708. URL https://doi.org/10.1101/2024.09.18.613708. PMID: 39345369, PMCID: PMC11429810.

M. Scutari, I. Mackay, and D. Balding. Using genetic distance to infer the accuracy of genomic prediction. *PLOS Genetics*, 12(9):e1006288, Sept. 2016. ISSN 1553-7404. doi: 10.1371/journal.pgen.1006288. URL http://dx.doi.org/10.1371/journal.pgen.1006288.

M. L. Selle, I. Steinsland, F. Lindgren, V. Brajkovic, V. Cubric-Curik, and G. Gorjanc. Hierarchical modelling of haplotype effects on a phylogeny. *Frontiers in Genetics*, 11: 531218, 2021. doi: 10.3389/fgene.2020.531218. URL https://doi.org/10.3389/fgene.2020.531218.

C. A. Sevillano, J. Vandenplas, J. W. M. Bastiaansen, R. Bergsma, and M. P. L. Calus. Genomic evaluation for a three-way crossbreeding system considering breed-of-origin of alleles. *Genetics Selection Evolution*, 49(1), 2017. doi: 10.1186/s12711-017-0350-1. URL http://dx.doi.org/10.1186/s12711-017-0350-1.

M. d. Silva, M. Martins, E. Junior, J. d. C. Panetto, and M. Machado. Programa de melhoramento genético da raça girolando - avaliação genética / genômica de fêmeas, 2025. ISSN 1516-7453 / e-ISSN 2966-0866.

M. V. G. B. d. Silva et al., editors. *Programa de Melhoramento Genético da Raça Girolando - avaliação genética / genômica de fêmeas - junho 2024*. Number 288 in Documentos / Embrapa Gado de Leite. Embrapa Gado de Leite, Juiz de Fora, 2024. ISSN 1516-7453 / e-ISSN 2966-0866.

A. Sluyter. Cattle in latin american history. *Oxford Research Encyclopedia of Latin American History*, April 2023. doi: doi.org/10.1093/acrefore/9780199366439.013.1153. URL https://doi.org/10.1093/acrefore/9780199366439.013.1153.

A. B. Smith and B. R. Cullis. Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *Euphytica*, 214:143, 2018. doi: 10.1007/s10681-018-2220-5. URL https://doi.org/10.1007/s10681-018-2220-5.

A. B. Smith, A. Ganesalingam, H. Kuchel, et al. Factor analytic mixed models for the provision of grower information from national crop variety testing programs. *Theoretical and Applied Genetics*, 128:55–72, 2015. doi: 10.1007/s00122-014-2412-x. URL https://doi.org/10.1007/s00122-014-2412-x.

F. Smith. A discriminate function for plant selection. *Ann Eugen.*, 7:240–250, 1936. doi: 10.1111/j.1469-1809.1936.tb02143.x. URL https://doi.org/10.1111/j.1469-1809.1936.tb02143.x.

M. Sohail, M. J. Palma-Martínez, A. Y. Chong, et al. Mexican biobank advances population and medical genomics of diverse ancestries. *Nature*, 622:775–783, 2023. doi: 10.1038/s41586-023-06560-0. URL https://doi.org/10.1038/s41586-023-06560-0.

L. Speidel, M. Forest, S. Shi, et al. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51:1321–1329, 2019. doi: 10.1038/s41588-019-0484-x. URL https://doi.org/10.1038/s41588-019-0484-x.

A. J. Stern, P. R. Wilton, and R. Nielsen. An approximate full-likelihood method for inferring selection and allele frequency trajectories from dna sequence data. *PLoS Genetics*, 15(9):e1008384, September 2019. doi: 10.1371/journal.pgen.1008384. URL https://doi.org/10.1371/journal.pgen.1008384.

O. Syrstad. Dairy cattle cross-breeding in the tropics: Performance of secondary crossbred populations. *Livestock Production Science*, 23(1–2):97–106, October 1989. doi: 10.1016/0301-6226(89)90008-0. URL https://doi.org/10.1016/0301-6226(89)90008-0.

J. Tabet, D. Lourenco, F. Bussiman, M. Bermann, I. Misztal, P. VanRaden, Z. Vitezica, and A. Legarra. All-breed single-step genomic best linear unbiased predictor evaluations for fertility traits in us dairy cattle. *Journal of Dairy Science*, 108(1):694—-706, 2025. doi: 10.3168/jds.2024-25281. URL http://dx.doi.org/10.3168/jds.2024-25281.

S. N. Takeshima and Y. Aida. Structure, function and disease susceptibility of the bovine major histocompatibility complex. *Animal Science Journal*, 77(2):138–150, 2006. doi: 10.1111/j.1740-0929.2006.00332.x. URL https://doi.org/10.1111/j.1740-0929.2006.00332.x.

A. Talenti, J. Powell, J. D. Hemmink, et al. A cattle graph genome incorporat-

ing global breed diversity. *Nature Communications*, 13:910, 2022. doi: 10.1038/s41467-022-28605-0. URL https://doi.org/10.1038/s41467-022-28605-0.

A. Talenti, T. Wilkinson, L. J. Morrison, et al. The evolution and convergence of mutation spectra across mammals. *Communications Biology*, 8:763, 2025. doi: 10.1038/s42003-025-08181-x. URL https://doi.org/10.1038/s42003-025-08181-x.

D. Tang, J. Freudenberg, and A. Dahl. Factorizing polygenic epistasis improves prediction and uncovers biological pathways in complex traits. *The American Journal of Human Genetics*, 110(11):1875—-1887, 2023. doi: 10.1016/j.ajhg.2023.10.002. URL http://dx.doi.org/10.1016/j.ajhg.2023.10.002.

F. Tiezzi and C. Maltecca. Genotype by environment interactions in livestock farming. In R. Meyers, editor, *Encyclopedia of Sustainability Science and Technology*. Springer, New York, NY, 2022. doi: 10.1007/978-1-4939-2493-6_1115-1. URL https://doi.org/10.1007/978-1-4939-2493-6_1115-1.

D. Tolhurst. *Genomic prediction models, selection tools and association studies for genotype by environment data*. PhD thesis, University of Edinburgh, 2024. URL https://era.ed.ac.uk/handle/1842/41957.

tskit developers. *tskit: A library for working with tree sequences in Python*, 2025. URL https://tskit.dev/. Version 0.6.4.

S. Tsuruta, D. A. L. Lourenco, Y. Masuda, T. J. Lawlor, and I. Misztal. Reducing computational cost of large-scale genomic evaluation by using indirect genomic prediction. *JDS Communications*, 2(11):356—-360, 2021. doi: 10.1016/j.ajhg.2023.10.002. URL http://dx.doi.org/10.1016/j.ajhg.2023.10.002.

J. Vandenplas, M. Calus, C. Sevillano, et al. Assigning breed origin to alleles in crossbred animals. *Genetics Selection Evolution*, 48:61, 2016. doi: 10.1186/s12711-016-0240-y. URL https://doi.org/10.1186/s12711-016-0240-y.

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

L. Varona, A. Legarra, M. Toro, and Z. Vitezica. Non-additive effects in genomic selection. *Frontiers in Genetics*, 9:78, 2018. doi: 10.3389/fgene.2018.00078. URL https://doi.org/10.3389/fgene.2018.00078.

A. I. Vazquez, D. M. Bates, G. J. M. Rosa, D. Gianola, and K. A. Weigel. Technical note: An r package for fitting generalized linear mixed models in animal breeding. *Journal of Animal Science*, 88:497–504, 2010. doi: 10.2527/jas.2009-1952.

D. Vilela, João, J. Resende, J. L. Leite, E. Alves, M. Araújo, R. Gazzola, G. Luis, Baricelo, E. Carlos, F. De, and Vian. A evolução do leite no brasil em cinco décadas. *Revista de política Agrícola*, 03 2017.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, et al. Scipy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3): 261–272, 2020. doi: 10.1038/s41592-019-0686-2. URL https://doi.org/10.1038/s41592-019-0686-2.

Z. G. Vitezica, L. Varona, and A. Legarra. On the additive and dominance variance and covariance of individuals within the genomic selection framework. *Genetics*, 195 (4):1223–1230, 2013. doi: 10.1534/genetics.113.155176. URL https://doi.org/10.1534/genetics.113.155176.

P. Wainschtein, D. Jain, Z. Zheng, S. Aslibekyan, D. Becker, et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nature Genetics*, 54(3):263—273, 2022. doi: 10.1038/s41588-021-00997-7. URL http://dx.doi.org/10.1038/s41588-021-00997-7.

H. Wang, B. Su, Y. Zhang, M. Shang, S. Li, D. Xing, J. Wang, L. Bern, A. Johnson, J. Al-Armanazi, T. Hasin, D. Hettiarachchi, A. Paladines Parrales, H. Dilawar, T. J. Bruce, R. A. Dunham, and X. Wang. From heterosis to outbreeding depression: genotype-by-environment interaction shifts hybrid fitness in opposite directions. *Genetics*, 227(4):iyae090, August 2024a. doi: 10.1093/genetics/iyae090. URL https://doi.org/10.1093/genetics/iyae090.

J. Y. Wang, N. Lin, M. Zietz, J. Mares, V. M. Narasimhan, P. J. Rathouz, and A. Harpak. Three open questions in polygenic score portability. *bioRxiv*, 2024b. doi: 10.1101/2024.08.20.608703. URL http://dx.doi.org/10.1101/2024.08.20.608703.

Y. Wang, K. Tsuo, M. Kanai, B. M. Neale, and A. R. Martin. Challenges and opportunities for developing more generalizable polygenic risk scores. *Annual Review of Biomedical Data Science*, 5(1):293–320, 2022. doi: 10.1146/annurev-biodatasci-111721-074830. URL http://dx.doi.org/10.1146/annurev-biodatasci-111721-074830.

J. A. Ward, G. P. McHugo, M. J. Dover, T. J. Hall, S. I. Ng'ang'a, et al. Genome-wide local ancestry and evidence for mitonuclear coadaptation in african hybrid cattle populations.

D. Waters, J. van der Werf, H. Robinson, et al. Partitioning the forms of genotype-by-environment interaction in the reaction norm analysis of stability. *Theoretical and Applied Genetics*, 136:99, 2023. doi: 10.1007/s00122-023-04319-9. URL https://doi.org/10.1007/s00122-023-04319-9.

K. A. Weigel, P. M. VanRaden, H. D. Norman, and H. Grosu. A 100-year review: Methods and impact of genetic selection in dairy cattle—from daughter–dam comparisons to deep learning algorithms. *Journal of Dairy Science*, 100(12):10234–10250, 2017. doi: 10.3168/jds.2017-12954. URL https://doi.org/10.3168/jds.2017-12954.

C. R. Werner, D. C. Gemenet, and D. J. Tolhurst. Fieldsimr: an r package for simulating plot data in multi-environment field trials. *Frontiers in Plant Science*, 15: 1330574, 2024. doi: 10.3389/fpls.2024.1330574. URL https://doi.org/10.3389/fpls.2024.1330574. Sec. Plant Breeding, Published 04 April 2024.

K. A. Wetterstrand. Dna sequencing costs: Data from the nhgri genome sequencing program (gsp). https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data, 2023. Accessed: 21/05/2025.

H. Wickham, M. Averick, J. Bryan, W. Chang, L. D'Agostino McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. Lin Pedersen, E. Miller, S. Milton Bache, K. Müller, J. Ooms, D. Robinson, D. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

M. Wicki, A. Legarra, and J. Raoul. Study of genetic progress in the context of disconnection between two originally connected populations. *Journal of Animal Breeding and Genetics*, June 2025. ISSN 1439-0388. doi: 10.1111/jbg.12946. URL http://dx.doi.org/10.1111/jbg.12946.

Y. C. J. Wientjes, P. Bijma, M. P. L. Calus, B. J. Zwaan, Z. G. Vitezica, and J. van den Heuvel. The long-term effects of genomic selection: 1. response to selection, additive genetic variance, and genetic architecture. *Genetics Selection Evolution*, 54(1), Mar. 2022. ISSN 1297-9686. doi: 10.1186/s12711-022-00709-7. URL http://dx.doi.org/10.1186/s12711-022-00709-7.

G. R. Wiggans, J. B. Cole, S. M. Hubbard, and T. S. Sonstegard. Genomic selection in dairy cattle: The usda experience. *Annual Review of Animal Biosciences*, 5:309–327, 2017. doi: 10.1146/annurev-animal-021815-111422. URL https://doi.org/10.1146/annurev-animal-021815-111422.

R. L. Willham. Genetic improvement of beef cattle in the united states: Cattle, people and their interaction. *Journal of Animal Science*, 54(3):659–666, March 1982. doi: 10.2527/jas1982.543659x. URL https://doi.org/10.2527/jas1982.543659x.

R. William and E. Pollak. Theory of heterosis. *Journal of Dairy Science*, 68(9): 2411–2417, September 1985. doi: 10.3168/jds.S0022-0302(85)81117-6. URL https://doi.org/10.3168/jds.S0022-0302(85)81117-6.

A. Winter, W. Krämer, F. A. O. Werner, S. Kollers, S. Kata, G. Durstewitz, J. Buitkamp, J. E. Womack, G. Thaller, and R. Fries. Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-coa:diacylglycerol acyltransferase (dgat1) with variation at a quantitative trait locus for milk fat content. *Proceedings of the National Academy of Sciences*, 99(14):9300–9305, 2002. doi: 10.1073/pnas.142293799. URL https://doi.org/10.1073/pnas.142293799.

A. W. Wohns, A. Wong, B. Jeffery, A. Akbari, S. Mallick, R. Pinhasi, N. Patterson, D. Reich, J. Kelleher, and G. McVean. A unified genealogy of modern and ancient genomes. *Science*, 375:eabi8264, 2022. doi: 10.1126/science.abi8264. URL https://doi.org/10.1126/science.abi8264.

Y. Wong, A. Ignatieva, J. Koskela, G. Gorjanc, A. W. Wohns, and J. Kelleher. A general and efficient representation of ancestral recombination graphs. *Genetics*, 228 (1):iyae100, September 2024. doi: 10.1093/genetics/iyae100. URL https://doi.org/10.1093/genetics/iyae100.

S. Wright. Systems of mating. i. the biometric relations between parent and offspring. *Genetics*, 6(2):111–123, March 1921. doi: 10.1093/genetics/6.2.111. URL https://doi.org/10.1093/genetics/6.2.111.

X. Xia, K. Qu, Y. Wang, et al. Global dispersal and adaptive evolution of domestic cattle: a genomic perspective. *Stress Biology*, 3:8, 2023. doi: 10.1007/s44154-023-00085-2. URL https://doi.org/10.1007/s44154-023-00085-2.

S. Yair and G. Coop. Population differentiation of polygenic score predictions under stabilizing selection. *Philosophical Transactions of the Royal Society B: Biological*

Sciences, 377(1852), 2022. doi: 10.1098/rstb.2020.0416. URL http://dx.doi.org/10.1098/rstb.2020.0416.

W. Yan and M. S. Kang. *GGE Biplot Analysis: A Graphical Tool for Breeders, Geneticists, and Agronomists*. CRC Press, Boca Raton, 2003. ISBN 978-0849313380. doi: 10.1201/9781420040371. URL https://doi.org/10.1201/9781420040371.

W. Yan, L. A. Hunt, Q. Sheng, and Z. Szlavnics. Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Science*, 40(3):597–605, 2000. doi: 10.2135/cropsci2000.403597x. URL https://doi.org/10.2135/cropsci2000.403597x.

A. I. Young. Discovering missing heritability in whole-genome sequencing data. *Nature Genetics*, 54(3):224—226, 2022. doi: 10.1038/s41588-022-01012-3. URL http://dx.doi.org/10.1038/s41588-022-01012-3.

Z. B. Zeng and C. C. Cockerham. Mutation models and quantitative genetic variation. *Genetics*, 133(3):729—-736, 1993. doi: 10.1093/genetics/133.3.729. URL http://dx.doi.org/10.1093/genetics/133.3.729.

B. Zhang, A. Biddanda, Gunnarsson, et al. Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. *Nature Genetics*, 55:768–776, 2023. doi: 10.1038/s41588-023-01379-x.

K. Zhang, A. Lenstra, S. Zhang, W. Liu, and J. Liu. Evolution and domestication of the bovini species. *Animal Genetics*, 51:637–657, 2020. doi: 10.1111/age.12974. URL https://doi.org/10.1111/age.12974.

J. Zhu, G. Kalantzis, A. Pazokitoroudi, A. F. Gunnarsson, H. Loya, H. Chen, S. Sankararaman, and P. F. Palamara. Fast variance component analysis using large-scale ancestral recombination graphs. *bioRxiv*, 2024. doi: 10.1101/2024.08.31.610262. URL https://doi.org/10.1101/2024.08.31.610262.

R. W. Zobel, M. J. Wright, and H. G. Gauch. Statistical analysis of a yield trial. *Agronomy Journal*, 80(3):388–393, 1988. doi: 10.2134/agronj1988.00021962008000030002x. URL https://doi.org/10.2134/agronj1988.00021962008000030002x.