TRACING CAT DOMESTICATION THROUGH POPULATION GENETICS AND CAPTURING GENOTYPE-BY-ENVIRONMENT INTERACTIONS IN US BEEF CATTLE GENOMIC PREDICTIONS

A Dissertation

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

SARA MICHELLE NILSON

Dr. Jared Decker, Dissertation Supervisor

JULY 2022

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

TRACING CAT DOMESTICATION THROUGH POPULATION GENETICS AND CAPTURING GENOTYPE-BY-ENVIRONMENT INTERACTIONS IN US BEEF CATTLE GENOMIC PREDICTIONS

presented by Sara Nilson,

A candidate for the degree of doctor of philosophy,

And hereby certify that, in their opinion, it is worthy of acceptance.

_____

Professor Jared Decker

_____

Professor Robert Schnabel

_____

Professor Jerry Taylor

_____

Professor Leslie Lyons

_____

Professor Elizabeth King

ACKNOWLEDGEMENTS

I would like to acknowledge my committee members first and foremost. Dr. Jared Decker for sticking with me for all these years and helping me develop my skill set, and intellectual knowledge as a scientist. Dr. Robert Schnabel for helping me improve my computational abilities and being patient when I stressed out the computing cluster system when I didn't know what I was doing. Dr. Jerry Taylor for inspiring me during my first semester to pursue my research with gusto and keep on learning for the enjoyment of answering my own questions. Dr. Leslie Lyons for bringing me into the world of cats that I would not have had the chance to explore otherwise. She helped me follow an interest in conservation and population genomics with a species that is near and dear to my heart. Dr. Elizabeth King for being so patient and understanding, a shoulder to cry on, and expanding my interests in evolutionary genomics. The conferences we attended together helped me think of my research in a different context and allowed me to make my research applicable on a larger scale.

I would like to acknowledge my fellow graduate students, postdoctoral colleges, and collaborators. First up Jenna Kalleberg, I can't express my love and appreciation for you with just a few words. You helped me out too many times to count, and I am eternally grateful for having you as a colleague and friend. Dr. Harly Durbin and Dr. Troy Rowan who both were discussion and bouncing boards for research and questions during our shared time together. Dr. Camila Braz who assisted me many times in trying to figure out how to utilize software to analyze my data and understand my results.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

LIST OF TABLES

Chapter 2

Chapter 3

LIST OF SUPPLEMENTARY AND ADDITIONAL FILES

Chapter 1

Chapter 2

LIST OF ABBREVIATIONS

BW = birth weight

WW = weaning weight

YW = yearling weight

CG = contemporary group

GxE = genotype x environment

EBV = estimated breeding value

EBV_A = additive estimated breeding value

D_{GxE} = genotype x environment deviation

EBV_{Total} = total estimated breeding value

HP = high plains

FB = fescue belt

UMN = upper midwest & northeast

SNP = single nucleotide polymorphism

TRACING CAT DOMESTICATION THROUGH POPULATION GENETICS AND CAPTURING GENOTYPE-BY-ENVIRONMENT INTERACTIONS IN US BEEF CATTLE GENOMIC PREDICTIONS

Sara M. Nilson

Dr. Jared Decker, Dissertation Supervisor

ABSTRACT

Cat domestication initiated as a symbiotic relationship between wildcats and the peoples of developing agrarian societies in the Fertile Crescent. To refine the sites of cat domestication, over 1,000 random-bred cats of primarily Eurasian descent were genotyped. The overall cat population structure suggested a single worldwide population with significant isolation by distance of peripheral subpopulations with decreased heterozygosity as genetic distance from the proposed cat progenitor's (*F.s. lybica*) natural habitat increased. Domestic cat origins are focused in the eastern Mediterranean Basin, spreading to nearby islands, down the Levantine coast and into the Nile Valley.

Climate change is driving the need for incorporating genotype-by-environment interactions in beef cattle genomic prediction models as animals frequently re-rank across environments. For United States Gelbvieh and Red Angus beef cattle, genotype-by-environmental inclusive models were compared to the current national genomic evaluation. Genotype-by-environment effects contributed ~3%-11% of the phenotypic variation to growth traits. Maternal and direct genotype-by-environment effects varied across growth traits. With slightly higher accuracies, the current national genomic evaluation models tend to outperform the genotype-by-environment models.

CHAPTER 1

Genetics of random-bred cats support the cradle of cat domestication in the Near East

Sara M. Nilson[1], Barbara Gandolfi[2], Robert A. Grahn[2], Jennifer D. Kurushima[2], Monika J. Lipinski[2], Ettore Randi[3], Nashwa E. Waly[4], Carlos Driscoll[5], Hugo Murua Escobar[6], Robert K. Schuster[7], Soichi Maruyama[8], Norma Labarthe[9,10], Bruno B. Chomel[2], Sankar Kumar Ghosh[11], Haydar Ozpinar[12], Hyung Chul Rah[13], Javier Millàn[14,15,16], Flavya Mendes-de-Almeida[10], Julia K. Levy[17], Elke Heitz[18], Margie A. Scherk[19], Paulo C. Alves[20,21], Jared E. Decker[1,22], Leslie A. Lyons[2,23]

[1]Division of Animal Sciences, University of Missouri, Columbia, MO 65211, USA

[2]Department of Health & Reproduction, School of Veterinary Medicine, University of California, Davis, CA 95616, USA

[3]Department of Chemistry and Bioscience, Aalborg University, Fredrik Bajers Vej 7H, 9220 Aalborg Øst, Denmark

[4]Department of Animal Medicine, Faculty of Veterinary Medicine, Assuit University, 71526, Assiut, Egypt

[5]Galton Corporation, Frederick, MD 21701, USA

[6]Clinic for Hematology, Oncology and Palliative Care, University Medical Center Rostock, 18055 Rostock, Germany

[7]Central Veterinary Research Laboratory, Dubai, United Arab Emirates

[8]Laboratory of Veterinary Public Health, Nihon University, 1866 Kameino, Fujisawa, Kanagawa, 252-8510, Japan

[9]Programa de Bioética, Ética Aplicada e Saúde Coletiva, Fundação Oswaldo Cruz, Rio de Janeiro, RJ, Brazil. Fundação Oswaldo Cruz, Av. Brazil 4365, Rio de Janeiro, RJ, 21040-360, Brazil.

[10]Programa de Pós-Graduação em Medicina Veterinária - Clínica e Reprodução Animal, Faculdade de Veterinária, Universidade Federal Fluminense, Rua Vital Brazil Filho 64, Niterói, RJ, 24230-340, Brazil

[11]Department of Biotechnology, Assam University, Silchar, 788011, India

[12]Graduate School of Health Sciences, Istanbul Gedik University, 34876 İstanbul, Turkey

[13]Research Institute of Veterinary Medicine, College of Veterinary Medicine, Chungbuk National University, Cheongju 28644, South Korea

[14]Instituto Agroalimentario de Aragón-IA2 (Universidad de Zaragoza-CITA), Miguel Servet 177, 50013 Zaragoza, Spain

[15]Fundación ARAID, Avda. de Ranillas, 50018 Zaragoza, Spain

[16]Facultad de Ciencias de la Vida, Universidad Andres Bello, Santiago, Chile.

[17]Maddie's Shelter Medicine Program, College of Veterinary Medicine, University of Florida, Gainesville, FL, 32608, USA

[18]Al Qurum Vet Clinic, Muscat, Oman.

[19]CatsINK Vancouver, British Columbia V5N 4Z4, Canada

[20]CIBIO/InBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos/InBIO Associate Lab & Faculdade de Ciências, Universidade do Porto, Campus e Vairão, 4485–661 Vila do Conde, Portugal

[21]Wildlife Biology Program, University of Montana, Missoula, Montana, 59812 USA

[22]Institute for Data Science and Informatics, University of Missouri, Columbia, MO 65211, USA

[23]Department of Veterinary Medicine & Surgery, College of Veterinary Medicine, University of Missouri, Columbia, MO 65211, USA

Abstract

Cat domestication initiated as a symbiotic relationship between wildcats (*Felis silvestris* subspecies) and the peoples of developing agrarian societies in the Fertile Crescent. As humans transitioned from hunter-gatherers to farmers ~12,000 years ago, bold wildcats likely capitalized on increased prey density (i.e., rodents). Humans benefited from the cat's predation on these vermin. To refine the sites of cat domestication, over 1,000 random-bred cats of primarily Eurasian descent were genotyped for single nucleotide variants and short tandem repeats. The overall cat population structure suggested a single worldwide population with significant isolation by distance of peripheral subpopulations. The cat population heterozygosity decreased as genetic distance from the proposed cat progenitor's (*F.s. lybica*) natural habitat increased. Domestic cat origins are focused in the eastern Mediterranean Basin, spreading to nearby islands, down the Levantine coast and into the Nile Valley. Cat population diversity supports the migration patterns of humans and other symbiotic species.

Introduction

The domestication and the geographical origins of the household cat (*Felis silvestris catus; Felis catus* (Kitchener *et al.*, 2017)) has been partially reconstructed from archaeological discoveries, artistic works, cultural changes, and genetics of ancient and modern felids. The cat's domestication process likely initiated ~12,000 years ago in the Fertile Crescent with initial contact between *Felis silvestris lybica* and farmers. Grain stores and refuse from developing societies attracted mice which led to a synanthropic trinity between humans, rodents, and felids (Vigne *et al.*, 2004; Driscoll *et al.*, 2007; Faure and Kitchener, 2009; Zeder, 2012; Ottoni *et al.*, 2017; Cucchi *et al.*, 2020). Feline remains, buried alongside human remains, were discovered at an archeological site dating to ~9,500 years ago, suggesting humans had formed a relationship with cats and transported cats to Cyprus (Vigne *et al.*, 2004, 2012). The earliest remains of suggested tamed cats in Egypt date to the fourth millennium BC (Baldwin, 1975; Málek, 1993; Van Neer *et al.*, 2014) and suggest felines became integral to Egyptian culture, culminating in thousands of mummified cats as votive offerings (Baldwin, 1975; Faure and Kitchener, 2009; Kurushima *et al.*, 2012; Baca *et al.*, 2018). Beginning in the first millennium BC, progeny of the Egyptian tamed cats were spread through trade and maritime routes by Phoenician, Carthaginian, Greek, Etruscan "cat-thief" traders and later by the Romans (Baldwin, 1975; Faure and Kitchener, 2009; Ottoni *et al.*, 2017).

The first occurrence of the mitochondrial haplotype A* of a *F.s. lybica/catus* species was reported in Bulgaria at ~6,400 years ago that is prior to the occurrence in Poland about 3,400 - 2,500 years ago, thereby, extending *F.s. lybica/catus* into a shared niche with *F.s. silvestris* from Anatolia to Eastern Europe (Krajcarz *et al.*, 2016, 2020;

Ottoni *et al.*, 2017; Baca *et al.*, 2018). Archeological evidence suggests the domestication process of *F.s. lybica* individuals initiated in the Near East with agrarian societal development within the Fertile Crescent and the Levant, intensified in Egypt along with cultural worships, leading to human migration and trade facilitating the domesticated feline dispora (Vigne *et al.*, 2004; Faure and Kitchener, 2009; Van Neer *et al.*, 2014; Ottoni *et al.*, 2017).

To genetically assess wildcats, feral domestic, and fancy-breed domestic cat relationships, a phylogenetic study was conducted with mitochondrial DNA (mtDNA) sequences of 2,604 base pairs from *ND5* and *ND6*, and 36 short tandem repeats (STR) genotypes (Driscoll *et al.*, 2007). A singular domestication origin in the Near East, arising from *F.s. lybica* was suggested, however, a limited sampling of wildcat subspecies was available. An expanded study of random-bred, domestic breeds, and wildcats with STR data reconfirmed the most likely origin of domestication was the Mediterranean Basin; however, four significant genetic distinctions were identified amongst 13 Eurasia cat populations (Lipinski *et al.*, 2008) based on allele frequencies and Bayesian clustering, particularly for the Far eastern, Mediterranean, Western European and Kenyan cats. Studies of the mtDNA control region variation in random-bred cats have also supported four to five major cat lineages with 12 common mitotypes representing maternal lineage diversity (Grahn *et al.*, 2011). A mtDNA study of mainly ancient and some modern felid samples from Europe, Africa, and Asia also traced modern felines to multiple *F.s. lybica* lineages within the Fertile Crescent (Ottoni *et al.*, 2017). While these previous studies all support the domesticated *F.s. catus* arose from *F.s. lybica* originating in the Near East, cats from other Eurasia regions of early

agricultural development, including the Near and Middle East as well as the Indus Valley of Pakistan, have not been examined in the context of contributing to feline domestication, which may account for the significant genetic distinction between Eastern and Western cat populations found in additional studies (Lipinski *et al.*, 2008). Recent studies of Chinese random-bred cats and the local wildcat species/subspecies (*F.s. bieti*) suggests the noted introgression of this wildcat with random-bred cats in China does not explain the distinctive genetics of Far Eastern and Western European random-bred cats; further, the agricultural center near the middle Yangtze and Yellow Rivers is likely not a second domestication site for cats (Yu *et al.*, 2021).

Although European colonization occurred only a few hundred years ago, regional cat populations tend to represent the initial domesticates of colonization and not unique or highly admixed populations, such as the cats in Australia (Spencer *et al.*, 2015; Koch *et al.*, 2016) as well as, North America and Nairobi, Kenya that are both genetically most similar to cats of Western Europe (Lipinski *et al.*, 2008). Interestingly, cats of Madagascar suggest genetic similarity with cats from the Arabia Sea trade routes, namely the Kenyan islands of Lamu and Pate, Oman, Kuwait and Iran and not cats imported by more recent colonists from France (Sauther *et al.*, 2020), further suggesting demographic stasis and the original influx of cats to a region may have the strongest influence on genetic signatures, rather than more recent migrants. Ancient DNA studies often suggest the converse; modern populations have no power to infer the dynamics of temporal populations movements (see reviews, (Freedman and Wayne, 2017; Frantz *et al.*, 2020)).

Cat domestication is likely commensal with agricultural development, and with the onset of the Holocene ~10,000 years ago agriculture developed independently at

perhaps several different global regions: the Near East likely the earliest, followed closely by agricultural sites in China, Southeast Asia and later in the Americas (Bellwood, 2005). The advent of agriculture altered human culture from nomadic hunter-gatherers to more sedentary lifestyles, leading to the establishment of increasingly larger settlements. Archeological discoveries of human remains and artifacts in the Near East and the middle Yangtze and Yellow Rivers in China indicate the earliest emergence of complex civilizations (Baldwin, 1975; Bar-Yosef, 1998; Hu *et al.*, 2014). The Indus Valley of modern-day Pakistan is also argued as a historical center for agricultural development (Bellwood, 2005).This current investigation of random-bred cats focused on population sampling near regions of early human agricultural developments, with extensive representation from the Near/Middle East, Pakistan, and near the Yellow River in China, with the addition of populations across Eurasia, from Southeast Asia to Great Britain to clarify historical cat population dynamics.

Random-bred cats represent an intermediate step in cat domestication, between wildcats and highly selected cat breeds, and since cats have and continue to perform their key role of vermin control without human assistances, random-bred cats may have escaped intense selective pressures due to breed formation, such as the strong selection pressure for particular phenotypes (Kurushima *et al.*, 2013). While modern populations only represent the latest epoch of migration and admixture (Pickrell and Reich, 2014), random-bred cats likely represent clearer patterns of historical diversity than fancy-breed cats. The historical time period reflected by random-bred cat genetic diversity is unknown and likely variable. This study investigated the genetic diversity of modern cat populations to determine if current genetic distinctions are discrete, suggesting possible

secondary genetic progenitors, or a continuum of diversity from a population center and due to isolation by distance. The geographical origins of cat domestication should be near the centers of cat genetic diversity.

Materials and methods

Sample collection

Samples were collected via buccal (cheek) swabs, FTA Cards (Whatman International Ltd.), gonads from neuter and spay clinics, and donated EDTA whole blood samples. DNA isolations were conducted following the manufacturer's protocol using the method appropriate for the sample, including QIAamp DNA blood mini kits, Qiagen DNA Easy kits (Qiagen, Valencia, CA, USA), organic extractions or methods for FTA card blood spots (Gandolfi *et al.*, 2016). Samples were amplified using whole genome amplification (REPLI-g Mini Kit, Qiagen) when DNA quantity was insufficient.

Cat samples (n = 564) from a previous study included random-bred cats from 17 locations (Supplementary Table 1,2)(Lipinski *et al.*, 2008). Cats previously labeled in the (Lipinski *et al.*, 2008) study as from Singapore were actually from Taiwan. Four African wildcat samples (*F.s. lybica*) were collected as part of other studies from the Western Sahara, Morocco, Tunisia and Mauritania, and provided as extracted and whole genome amplified DNA (Randi *et al.*, 2001; Lecis *et al.*, 2006; Oliveira *et al.*, 2015). The STR data for the cats from Madagascar (n = 27) has been previously published (Sauther *et al.*, 2020). Domestic cat samples from Portugal and Italy have been previously analyzed, but new data was generated for this study (Lecis *et al.*, 2006; Oliveira *et al.*, 2008a, b). Additional random-bred cat samples were collected from 30 new countries and additional population locations within several countries including Brazil, China, Egypt, Italy,

Kenya, South Korea, and the USA (Supplementary Table 1, 2). For the STR analyses, 1,857 random-bred cats and the four African wildcats were genotyped. For the SNP analyses, 969 random-bred cats were genotyped. In addition, the same four African wildcats and 10 cats collected as roadkill from Spain, wildcat hybrids, were included in the SNP dataset (Oliveira *et al.,* 2008a; Oliveira *et al.,* 2015).

Genotyping

Thirty-six autosomal STRs were genotyped following the PCR and analysis procedures in a previous study (Lipinski *et al.*, 2008) (Supplementary Table 3). Unlinked non-coding autosomal SNPs (n=132) were selected to represent all autosomes from the 1.9x coverage cat genomic sequence, which were defined by one Abyssinian cat (Pontius *et al.*, 2007). The SNPs have been remapped to cat genome assembly Felis Catus 9.0 (Buckley *et al.*, 2020). Primers were designed with the VeraCode Assay Designer software (Illumina Inc., San Diego, CA, USA). The SNPs had a Ranking Score of 0.75 or higher (with a mean design score of 0.95) and a Gen Train Score of > 0.55 (Supplementary Table 4). Golden Gate Assay amplification and BeadXpress reads were performed per the manufacturer's protocol (Illumina Inc.) on 50-500ng of DNA or whole genome amplified product. BeadStudio software v. 3.1.3.0 with the Genotyping module v. 3.2.23 (llumina Inc.) was used to analyze the data. In PLINK v1.9, quality control for minor allele frequency was set at 0.005 and genotype call rate set at 0.8 (Chang *et al.*, 2015). The genotyping data for the project are presented in Supplementary Files 1 and 2.

Principal Component Analysis

To project the genetic similarities among individuals, principal component analysis (PCA) was performed for the SNP data with the smartpca program from the

EIGENSOFT package (Patterson *et al.*, 2006). To determine the potential effect of population size, 4 different PCAs were generated by grouping individuals by 1) sample location, 2) country, 3) sample location with populations randomly reduced to a maximum of 25 individuals, and 4) country with populations randomly reduced to a maximum of 40 individuals. Due to minimal visual differences, all further SNP analyses were conducted on the country grouped data with populations randomly reduced to a maximum of 40 individuals per country, for a total of 969 random-bred felines in the dataset. A PCA of the STR data set was conducted with the R package adegenet v2.1.1 (Jombart, 2008).

Population Structure

A variational Bayesian framework, fastSTRUCTURE, estimates the admixture proportions of individuals when given *K* populations (Raj *et al.*, 2014). The fastSTRUCTURE software proposes two metrics to select and identify *K*: the *K* that maximizes the log-marginal likelihood lower bound of the dataset (K*) and the minimum *K* that accounts for a cumulative ancestry of 99.99% (KC). fastSTRUCTURE was run independently for a *K* of 1 to 20 for the SNP data. As fastSTRUCTURE is specific to biallelic data, a Bayesian clustering method, STRUCTURE v2.3.4, was utilized for STR analyses for jointly inferring the *K* populations represented and probabilistically assigning each individual to one or more populations (Pritchard *et al.*, 2000). Overall, STRUCTURE was run from a *K* of 1 to 35 with each independent *K* run 20 times. Runs consisted of a 50,000 burn-in period with 50,000 MCMC replications, and the results were averaged with CLUMPP v1.1.2 (Jakobsson and Rosenberg, 2007). Averaged results were calculated only for *K* of 1 to 5 as higher *K* values did not converge, most likely due

to little population structure differences. The ΔK distribution was calculated following the process implemented by (Evanno *et al.*, 2005) to determine an optimal value of *K*.

Admixture

To identify admixture and support fastSTRUCTURE and STRUCTURE observations, f_3 statistics were calculated among all sample location populations (significant results presented in Supplementary Table 5) for the SNP dataset, excluding small populations and those from the Americas and Australia, with the *threepop* component of the TreeMix program (Pickrell and Pritchard, 2012).

Isolation by Distance

To formally test for isolation by distance at the finest geographical scale, SNP populations were reclassified back to their sample location labels to achieve fine-scale results (Supplementary Table 6, 7). Due to potential bias, sample locations with less than five individuals were removed from further analyses. In addition, sample locations from the Americas and Australia were excluded due to strong evidence supporting European ancestry and geographic distance being exaggerated due to human-mediated migration. The remaining sample location populations had f_3 statistics calculated and those populations with significant values were removed to reduce noise generated by admixture possibly due to migration events (see Admixture). Removal of admixed locations was done to strengthen the relationship between modern samples and ancient processes by removing more recent admixture events. For the SNP data, 24 sample location populations were analyzed, not including the *F.s. lybica* and the wildcat hybrid populations. For the STR data, the same individuals from the SNP dataset were used resulting in 22 populations. The two populations lost due to no STR genotypes were from

Spain and Portugal. Isolation by distance was tested with a Mantel test between calculated matrices of geographical distances (geodesic in meters from latitude and longitude coordinates) and Cavalli-Sforza and Edwards chord genetic distances with the adegenet and geodist R packages (Jombart, 2008; Karney, 2013; Séré *et al.*, 2017). Cavalli-Sforza and Edwards chord genetic distances were used as it was previously shown to be a more powerful approach for isolation by distance (Séré *et al.*, 2017). The Mantel test results are calculated with a Monte-Carlo test with 999 replicates; the final reported correlation and p-value are the average of 1,000 independent Monte-Carlo tests. To further explore expansion and migration patterns, isolation by distance was calculated among all of the samples collected in the contiguous United States of America for both data types.

Genetic Diversity

Observed and expected heterozygosities were calculated for the SNP and STR populations used in the isolation by distance analyses, and for all sample locations with the adegenet R package (Jombart, 2008). F-statistics were calculated for the SNP and STR random-bred cat data on a worldwide population level with the hierfstat R package (Goudet, 2005). In addition, F_{IS} statistics were calculated for all sample locations with the equation: $F_{IS} = 1 - (H_{obs} / H_{exp})$ (Supplementary Table 8).

Results

The genotyped cat samples consisted of 1,987 random-bred cats (*F.s. catus*), four African wildcats (*F.s. lybica*), and 10 hybrids of domestic and European wildcats (*F.s. silvestris*)(Oliveira *et al.*, 2015). Random-bred cats (n = 839) genotyped for both SNPs and STRs, 1,018 cats were genotyped for STRs only and 130 cats were genotyped for

SNPs only. The four African wildcats were genotyped for both SNPs and STRs, while the 10 putative hybrids of domestic and European wildcats were only genotyped for the SNPs. The random-bred cats represent over 40 countries including over 85 sampling sites (Supplementary Table 1, 2). The distribution of the populations and marker types is depicted in Supplementary Figure 1. A majority of sampling was focused on the European and Asian continents, particularly the Near East region.

The principal component analyses (PCA) of the SNP and STR data sets have similar patterns (Figure 1). Principal component 1 (PC1) forms a cline of felines from Asia and the Middle East (negative values) to Europe and the Americas (positive values) with felines from Africa and the Near East central to the peripheral populations. Principal component 2 (PC2) highlights differences between Asian cats from the Near East (Cyprus, Israel, Egypt, Jordan, Lebanon, Greece), the Middle East (Bahrain, Iran, Iraq, Kuwait, Oman, Pakistan, UAE), and African cats (Tunisia, Kenya, Madagascar). In both data sets, the four African wildcats (*F.s. lybica*), which are considered the progenitor sub-species for the domestic cat, are positioned mainly with the felines from the Near East, near the center of the PCA space. However, the wildcat hybrids in the SNP data set cluster peripherally from the random-bred felines, and appear more closely related to the felines from Western Europe, including cats from the Americas, which could be due to introgression among the populations.

The San Marcos Island (Baja California Sur, Mexico) population in the STR data set diverges from the European and American felines, reflective of a small, isolated island population. Both PCA reflect genetic divergence due to geographic separation, but the populations form a cline rather than clear geographical clusters. Southeastern and

East Asian cats are located at one periphery of the distribution, as are the Near Eastern and Mediterranean cats in another, with the Western European cats in the third. Only 8 - 9% of the total genetic variability could be attributed to differences among the cat populations (STR F_{ST} = 0.078; SNP F_{ST} = 0.088). On average, the local populations had a deficit of heterozygotes of 6-9% (STR F_{IS} = 0.088; SNP F_{IS} = 0.063) whereas the total worldwide random-bred population had a deficit of heterozygotes of 15-16% (STR F_{IT} = 0.159; SNP F_{IT} = 0.146).

Population structure was estimated across both data sets to gain insight into the admixture of the current random-bred felines. For the SNP data, a *K* of 1 explains 99.99% of the variation in the data set (KCstatistic, (Raj *et al.*, 2014)) suggesting the worldwide random-bred feline populations do not form genetically distinct clusters, even though the cats have been geographically separated. A *K* of 2 maximizes the log-marginal likelihood lower bound (K*statistic, (Raj *et al.*, 2014) separating felines between Western European ancestry, and Asian/Middle Eastern/Mediterranean ancestry (Supplementary Figure 2). Felines from Africa (Nairobi, Kenya, and Tunis, Tunisia) and Western Europe share genetic similarity between the two ancestry assignments. Cats in the Americas have a genetic profile typical of Western European cats. The African cats from the eastern islands of Kenya, Lamu and Pate, share genetic similarity with cats from the Middle East and the Eastern Mediterranean. These similarities are maintained through higher levels of sub-structure. The *K* of 2 ancestry pattern reflects population positionings in the PCA along PC1. As the *K* increases to 3 and 4, the Asian felines reflect PC2, and the sub-regional distinction appears with East Asia and Southeast Asia (Figures 1, 2 and Supplementary Figures 3, 4). As *K* increases up to 5, the island population of San Marcos

appears distinct, with additional sub-regional assignments within Western Europe, Mediterranean, Near/Middle East, and East Asia (Supplementary Figure 4). Countries such as India and Sri Lanka appear to be highly admixed. The four African wildcats are similar to a typical Western European population and the putative hybrids of domestic and European wildcats are a more cohesive grouping in which Western European cats share ancestry.

Similar population structuring is depicted by the STR analyses. For the STRs, the modal value of the $\mathbf{\Delta}K$ distribution was at $K = 2$. The ancestral populations were split between Western Europe versus Middle East/Asia, which is concordant with the SNP data (Supplementary Figure 5). Random-bred cats from Africa (South Africa, Nairobi, Kenya, and Tunis, Tunisia) and the Near East were mixed almost equally between these two ancestry assignments. As K increases, more geographical separation is depicted: K of 3 distinguishes Asian cats from the Mediterranean / Near / Middle Eastern cats, a K of 4 separates the Mediterranean/Near East cats from the Middle East felids, and a K of 5 brings out the island population from San Marcos (Figure 3 and Supplementary Figures 6, 7). Overall, the population structure between the SNPs and STRs are concordant and consistent with the patterns observed in the PCA (Engelhardt and Stephens, 2010), supporting the inference that the worldwide random-bred subpopulations are a single population with genetic differentiation due to separation by geographic distance . The most observable difference between SNPs and STRs is the SNPs differentiate Southeast and Eastern Asian cats at $K = 4$ while STRs maintain the Asian cats as a stronger cluster and differentiate Mediterranean / Near Eastern cats from the cats of the Middle East (Supplementary Figures 3, 6).

The allele frequencies of the cat populations were analyzed to calculate f_3 statistics with corresponding z-scores to evaluate possible admixture (Supplementary Table 5)(Reich *et al.*, 2009). There are 234 of 51,888 comparisons with a z-score ≤ -2 (0.45%), supporting admixture within the 22 target populations. The sample population from Lahore, Pakistan has the lowest z-score of -4.9 and 56 significant z-scores with other populations that are highly indicative of admixture. Populations most frequently contributing to significant admixture as parent (i.e., donor) populations include: Thailand, Vietnam, the wildcat hybrids, and Asyut, Egypt.

Since population structure analyses suggest a single population with possible differentiation due to geographic separation, isolation by distance was formally tested among the sample location populations in Europe, Africa, Near East, Middle East, and Asia for which evidence of admixture was not observed from f_3 statistics (see Admixture and Supplementary Table 5). When the population pairwise geographic distances are plotted against the Cavalli-Sforza and Edwards chord genetic distances (Séré *et al.*, 2017), a clear trend is observed; as the geographic distance increases between populations the genetic distance also increases (Figure 4 and Supplementary Table 6, 7). The Mantel test between distance matrices resulted in a positive correlation of 0.447 with a p-value of 0.001 for the SNP data, and a positive correlation of 0.302 with a p-value of 0.0076 for the STR data. When the admixed populations were included in a separate Mantel test for isolation by distance, the SNP data had a positive correlation of 0.369 with a p-value of 0.001, and the STR data had a positive correlation of 0.23 with a p-value of 0.0025. The ~10% decrease in correlations between genetic and geographic distance when the admixed populations were included could be due to the increased

genetic noise from migrants. Conversely, isolation by distance analyses were not significant (SNP p-value =0.871; STR p-value = 0.405) for random-bred cats in the contiguous United States of America, suggesting multiple importations of felines into the USA and little geographical structure in the genomic data.

Based on the significant isolation by distance, observed and expected heterozygosities were calculated for each sample site. When the observed heterozygosity is plotted against the genetic distance from the domestic progenitor, *F.s. lybica*, a negative relationship is identified; as the genetic distance from *F.s. lybica* increases the observed heterozygosity decreases (Figure 5). There is a negative correlation for the SNP data of -0.57 with a p-value of 0.0034 while the STR correlation is -0.33 with a p-value of 0.13. To explore the geographic and observed heterozygosity relationship further, the populations were plotted on a map to identify an epicenter of high diversity that decreases outwards in a radial fashion as expected from a center of domestication (Figure 6 and Supplementary Table 8). The centers of diversity are focused in the Mediterranean side of the Fertile Crescent, including the Levant and expanding into the Nile Valley and Mesopotamia. Cat populations with high heterozygosity are also identified in Agra, India, Sri Lanka, and the island population of Majorca, Spain (Supplementary Table 8).

Discussion

Throughout the world, the domestic cat is a beloved and charismatic companion animal. Although as popular of a pet as the domestic dog, the origins of the domestic cat are less studied. Random-bred cats (i.e., feral, moggie, alley, house, community, street or barn cats) remain a behaviorally semi-domesticated species that can quickly revert to a wild state. While they have a low survival rate in the wild, their high reproductive

capacity increases population size (Nutter *et al.*, 2004). As apex predators, this reversion capability has often been exploited to eradicate invasive animals from island populations, whereas later, the cats themselves became invasive alien species (Rendall *et al.*, 2021; Plein *et al.*, 2022).

Here, the random-bred cats of the study represent semi-domesticated animals that lie somewhere between "habituation" and "commercial breeds and pets" on the commensal domestication trajectory (Zeder, 2012; Larson and Burger, 2013). For cats, human assistance is not necessarily required for mating, shelter, safety or the procurement of food (Driscoll *et al.*, 2009). The cat's semi-domesticated behavioral state is consistent with weaker human-influenced artificial selection pressures on the species. Although cats may have been domesticated at approximately the same time as many agricultural species, ~8,000 - 10,000 years ago, cats have scavenged refuge and curbed vermin populations during their symbiotic relationship with humans (Clutton-Brock, 1988). Therefore, for the past several thousand years, cats have not been transformed drastically in form or function, unlike dogs and economically important species. Only for the past ~200 years, cat breeds, not random-bred cats, have been selected for mainly monogenic aesthetic traits undergoing novelty selection on a small number of loci and likely a small portion of the genome. Minor structural differences and no functional behavioral differences were present in cats when the first cat show took place in 1871 ('The Cat-Show', 1871). The semi-domesticated nature of random-bred cats makes them an excellent resource to understand cat population origins, domestication, and dispersal.

Single nucleotide polymorphism (SNP) and STR genotypes of an extended and fine-scale sampling of Eurasian cats demonstrated domestication most likely occurred in

the concentrated region of the Fertile Crescent. The focused sampling plan was to test alternative hypotheses of multiple domestication centers in 1) Near East, 2) China, and 3) Southeast Asia against the null hypothesis of a single domestication center. This focused sampling could also identify distinct populations indicative of admixture with wild relatives. However, despite this intensive sampling, only the Near East is suggested as a site of cat domestication indicating a pattern of dispersal outwards from regions like the Levant and the Nile Valley, while elsewhere in the world lacks this pattern (Vigne *et al.*, 2004; Driscoll *et al.*, 2007; Lipinski *et al.*, 2008).

For other domesticated species, isolation by distance testing and genetic diversity measurements reveal a pattern of expansion from the domesticated founders (Ramachandran *et al.*, 2005; Scheu *et al.*, 2015; Malomane *et al.*, 2020). Previous genetic studies examined the extremes of the geographical locations while the current research included bridging populations, which revealed the structure of worldwide random-bred populations is nearly a panmictic population with evidence of isolation by distance at the peripheries of their migration (Lipinski *et al.*, 2008). As found for human populations (Barbujani *et al.*, 1997), a majority of genetic diversity is explained within populations and distinctions can be observed only at the peripheries of migration patterns and do not account for the vast genetic diversity of cats. This pattern of isolation by distance, with highest levels of diversity near sites of domestication is observed in other species. Chickens, like cats, dispersed from a domestication center by human-mediated migration and the majority of genetic diversity variation is explained by genetic distance to the wild populations (Malomane *et al.*, 2020). Village dogs, like random-bred cats, are considered to be free-breeding with minimal admixture due to isolation and have escaped human-

mediated inbreeding (Shannon *et al.*, 2015). Although the location of dog domestication is disputed (Bergström *et al.*, 2020), genetic signatures have been used to infer a Central Asia domestication of dogs. Patterns of short-range linkage disequilibrium decay were found to be lowest in village dog populations from Central Asia with rates rising as geographical distance increased (Shannon *et al.*, 2015). After filtering admixed populations to remove the most recent epoch of admixture and improve the fit between modern samples and ancient ancestry patterns, the random-bred cat results suggest a similar pattern: genetic diversity is higher in populations located where the progenitor species began to interact with humans resulting in a shorter genetic distance and heterozygosity decreasing as geographic distance increases out from this origin. Studies using ancient DNA of domesticated cats may reveal a more complicated process (MacHugh *et al.*, 2017), but the pattern revealed in random-bred cats is striking and agrees with archeological evidence. Unlike many domestication studies that must use modern breeds for comparisons, these random-bred cats have likely had less selection and lower founder effects and lower genetic loss by drift since cats are under fewer constraints by humans.

The cat diaspora is relatively more recent than for humans or canines. As European maritime exploration to conquer and settle new lands increased, felines were brought on ships for trade and to safeguard food and wares from rodents (Faure and Kitchener, 2009). Migration of cats rose with imperialism exploration and colonization, which increased the numbers of ships traveling to the Americas. The data suggests cats in distant areas from the Near East, including Australia, the Americas, and colonial regions such as Tunisia and mainland Kenya, are close derivatives of Western European cats,

reflecting western European colonization. The admixed genetics from Western Europe and the Near East cats were subsequently spread to Portuguese colonies in the Americas (Ruiz-Garcia *et al.*, 2005). Although wild felids migrated to the Americas across ancient land bridges and small felids of domestic cat size have been present in South America for millions of years (Johnson *et al.*, 2006; Li *et al.*, 2016), domestic cats only populated the Americas with the arrival of Europeans in the 1500's. This work reinforces domestic felines from the Americas are closely related to those from Europe suggesting insufficient time for drift or selection to cause genetic distinction (Lipinski *et al.*, 2008).

Cats migrated to Europe and to the east of the Fertile Crescent along with agricultural development and trade (Ottoni *et al.*, 2017; Baca *et al.*, 2018). Pakistan felines tend to have more European influence than other countries in the Middle East, possibly due to the influence and control of the British East India Company in Southern Asia, which is supported by several significant f_3 statistics with a contributing population from Europe. Kuwait felines have a higher percentage of Near Eastern ancestry resulting from the location of the country being a center of land and sea trade routes, and a major oil producer resulting in the influx of foreign workers from nearby countries (Shah and Al-Qudsi, 1989). India and Sri Lanka both have ancestry admixture from many populations and higher observed heterozygosity attesting to the large amounts of movement of traders due to land and maritime Silk Road routes. Being able to trace these human and cat migration patterns through genetics speaks to the diversity and depth of this sample population reinforcing our ability to narrow the origin of domestication.

European wildcats (*F.s. silvestris)* have many studies focused on the concern of introgression with free-roaming or partially-free-roaming random-bred cats (Beaumont *et*

al., 2001; Witzenberger and Hochkirch, 2014; Oliveira *et al.*, 2015; Koch *et al.*, 2016; Mattucci *et al.*, 2019; Quilodrán *et al.*, 2020). A sample of 130 European wildcat samples were initially collected as unknown wildcats, and some of these samples were later suggested as hybrids with domestic cat introgression, (Oliveira *et al.*, 2015). Hence, the clustering of the 10 wildcat hybrid felines on the periphery of the Western European cats is expected. The 10 wildcat hybrids included in this study have little to no random-bred ancestry in our fastSTRUCTURE analysis and produce no significant f_3 statistics, due to the lack of a *F. s. silvestris* reference population. However, the genotypes from these hybrids suggest that European wildcat influence is pervasive throughout populations in Europe but also can be tracked through genetics of populations in the New World like those in the Americas. Recently, an investigation of cats from China, including a sampling of the Chinese wildcat (*F.s. bieti*), suggested some gene flow between this wild species and domestic cats, but not sufficiently to explain the genetic difference between Far Eastern and Western domestic cats. Although a few Asian wildcats (*F.s. ornata*) were included in the Chinese study, the four cats were sampled from one site and specimens from wildcats from the Near East and the Indus Valley were not available (Yu *et al.*, 2021). Thus, further studies are needed to evaluate the complexity of domestication and the influence of admixture with wild populations on modern domestics (Larson and Burger, 2013).

Although cats and agricultural species serve very different purposes to humans, the geographic patterns of admixture in cats are a near perfect reflection of admixture and migration in cattle populations, such as along the Silk Road and in the Americas (Decker *et al.*, 2014). Along with archeological and genetic data, even the cat's prey, house mice,

have also represented bio-proxies for human migration patterns (Rajabi-Maham *et al.*, 2008; Jones *et al.*, 2013; Cucchi *et al.*, 2020; Li *et al.*, 2020).

Overall, worldwide random-bred feline populations exhibit low levels of genetic differentiation even when geographically separated; however, populations on the peripheries of migration can be genetically differentiated. Populations were significantly isolated by distance; between populations the genetic distance increased as the geographic distance increased. Observed heterozygosity was higher in populations located near the Mediterranean Basin of the Fertile Crescent where archeological evidence points towards the first human-cat interactions. Additionally, these populations have a shorter genetic distance to the progenitor species *F.s. lybica*. The origin of domestication for *F.s. catus* is suggested as the coastal regions of the Mediterranean Basin of the Fertile Crescent where cats have high observed heterozygosity and a short genetic distance to the progenitor subspecies. As highly agrarian societies developed, domesticated cats then spread down into the Nile Valley where cultural integration of felines into society slightly decreased heterozygosity and increased the genetic distance from the initial founders. Mummified Egyptian cats have control region mtDNA mitotypes specific to the mitotype G of contemporary Egyptian cats and a mitotype D highly common in Near and Middle Eastern populations, but one mummified cat also had a common mitotype C that has worldwide distribution (Kurushima *et al.*, 2012), perhaps supported by the Egyptian domestication origin suggested by ancient DNA studies (Ottoni *et al.*, 2017). The slightly lower diversity could be an influence of ancient cultural selections. Further studies on ancient, regional wildcat populations would further decipher cat origins. Cats likely spread through-out Eurasia as agricultural development

spread, causing isolation by distance. Once larger sea bearing vessels facilitated the trade of goods and stores, cat migrations reached more distant ports, including the Americas and Australia in the 1500's. Modern transport of pets has and will continue to increase admixture around the world, however, cat populations in the Americas, Australia, and Madagascar seem to represent the cats of human colonists, where indigenous cats, including wildcats do not exist. Even the cats of mainland Kenya and the eastern coastal Kenyan islands have genetic signatures similar to Western Europe and the Arabian Sea, respectively. While these results are supported by large sample sizes, denser genotypes of these populations would allow for additional methodologies including linkage disequilibrium analyses, which could lead to even further clarification of the center of cat domestication.

This study infers the relationships, dispersal, admixture, and genetic distances among worldwide random-bred cats from patterns of genetic polymorphism, which were unlinked and randomly identified and assumed to be neutral. Population bottlenecks and effective population sizes cannot be evaluated in the current study. Additional studies including data from various wildcat species/subspecies, particularly *F.s. ornata* from the Iraq, Iran, the Indus Valley regions, and Northwestern India could further explain the genetic variation seen in cat populations. Genetic and archeological studies form pre-farming cats would be an important addition in further clarifying the cat domestication process. The patterns of genetic diversity and differentiation observed in worldwide random-bred cats parallel those of other species, especially humans once they become farmers, suggesting human history is written in the DNA of domesticated species.

References

Baca M, Popović D, Panagiotopoulou H, Marciszak A, Krajcarz M, Krajcarz MT, *et al.* (2018). Human-mediated dispersal of cats in the Neolithic Central Europe. *Heredity* **121**: 557–563.

Baldwin JA (1975). Notes and Speculations on the Domestication of the Cat in Egypt. *Anthropos* **70**: 428–448.

Barbujani G, Magagni A, Minch E, Luca Cavalli-Sforza L (1997). An apportionment of human DNA diversity. *Proc Natl Acad Sci U S A* **94**: 4516–4519.

Bar-Yosef O (1998). The Natufian Culture in the Levant, Threshold to the Origins of Agriculture. *Evol Anthropol* **6**: 159–177.

Beaumont M, Barratt EM, Gottelli D, Kitchener AC, Daniels MJ, Pritchard JK, *et al.* (2001). Genetic diversity and introgression in the Scottish wildcat. *Mol Ecol* **10**: 319–336.

Bellwood P (2005). First Farmers: The Origins of Agricultural Societies. In: *First Farmers: The Origins of Agricultural Societies*, unknown.

Bergström A, Frantz L, Schmidt R, Ersmark E, Lebrasseur O, Girdland-Flink L, *et al.* (2020). Origins and genetic legacy of prehistoric dogs. *Science* **370**: 557–564.

Buckley RM, Davis BW, Brashear WA, Farias FHG, Kuroki K, Graves T, *et al.* (2020). A new domestic cat genome assembly based on long sequence reads empowers

feline genomic medicine and identifies a novel gene for dwarfism. *PLoS Genet* **16**: e1008926.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**: 7.

Clutton-Brock J (1988). *A Natural History of Domesticated Mammals*. Cambridge University Press.

Cucchi T, Papayianni K, Cersoy S, Aznar-Cormano L, Zazzo A, Debruyne R, *et al.* (2020). Tracking the Near Eastern origins and European dispersal of the western house mouse. *Sci Rep* **10**: 8276.

Decker JE, McKay SD, Rolf MM, Kim J, Molina Alcalá A, Sonstegard TS, *et al.* (2014). Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. *PLoS Genet* **10**: e1004254.

Driscoll CA, Macdonald DW, O'Brien SJ (2009). From wild animals to domestic pets, an evolutionary view of domestication. *Proc Natl Acad Sci U S A* **106 Suppl 1**: 9971–9978.

Driscoll CA, Menotti-Raymond M, Roca AL, Hupe K, Johnson WE, Geffen E, *et al.* (2007). The Near Eastern origin of cat domestication. *Science* **317**: 519–523.

Engelhardt BE, Stephens M (2010). Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genet* **6**: e1001117.

Evanno G, Regnaut S, Goudet J (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* **14**: 2611–2620.

Faure E, Kitchener AC (2009). An Archaeological and Historical Review of the Relationships between Felids and People. *Anthrozoös* **22**: 221–238.

Frantz LAF, Bradley DG, Larson G, Orlando L (2020). Animal domestication in the era of ancient genomics. *Nat Rev Genet* **21**: 449–460.

Freedman AH, Wayne RK (2017). Deciphering the Origin of Dogs: From Fossils to Genomes. *Annu Rev Anim Biosci* **5**: 281–307.

Gandolfi B, Grahn RA, Gustafson NA, Proverbio D, Spada E, Adhikari B, *et al.* (2016). A Novel Variant in CMAH Is Associated with Blood Type AB in Ragdoll Cats. *PLoS One* **11**: e0154973.

Goudet J (2005). hierfstat, a package for r to compute and test hierarchical F-statistics. *Mol Ecol Notes* **5**: 184–186.

Grahn RA, Kurushima JD, Billings NC, Grahn JC, Halverson JL, Hammer E, *et al.* (2011). Feline non-repetitive mitochondrial DNA control region database for forensic evidence. *Forensic Sci Int Genet* **5**: 33–42.

Hu Y, Hu S, Wang W, Wu X, Marshall FB, Chen X, *et al.* (2014). Earliest evidence for commensal processes of cat domestication. *Proc Natl Acad Sci U S A* **111**: 116–120.

Jakobsson M, Rosenberg NA (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**: 1801–1806.

Johnson WE, Eizirik E, Pecon-Slattery J, Murphy WJ, Antunes A, Teeling E, *et al.* (2006). The late Miocene radiation of modern Felidae: a genetic assessment. *Science* **311**: 73–77.

Jombart T (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**: 1403–1405.

Jones EP, Eager HM, Gabriel SI, Jóhannesdóttir F, Searle JB (2013). Genetic tracking of mice and other bioproxies to infer human history. *Trends Genet* **29**: 298–308.

Karney CFF (2013). Algorithms for geodesics. *J Geodesy* **87**: 43–55.

Kitchener AC, Breitenmoser-Würsten C, Eizirik E, Gentry A, Werdelin L, Wilting A, *et al.* (2017). A revised taxonomy of the Felidae : The final report of the Cat Classification Task Force of the IUCN Cat Specialist Group. : 32616.

Koch K, Algar D, Schwenk K (2016). Feral Cat Globetrotters: genetic traces of historical human-mediated dispersal. *Ecol Evol* **6**: 5321–5332.

Krajcarz M, Krajcarz MT, Baca M, Baumann C, Van Neer W, Popović D, *et al.* (2020). Ancestors of domestic cats in Neolithic Central Europe: Isotopic evidence of a synanthropic diet. *Proc Natl Acad Sci U S A* **117**: 17710–17719.

Krajcarz M, Makowiecki D, Krajcarz MT, Masłowska A, Baca M, Panagiotopoulou H, *et al.* (2016). On the trail of the oldest domestic cat in Poland. An insight from morphometry, ancient DNA and radiocarbon dating: On the trail of the oldest domestic cat in Poland. *Int J Osteoarchaeol* **26**: 912–919.

Kurushima JD, Ikram S, Knudsen J, Bleiberg E, Grahn RA, Lyons LA (2012). Cats of the Pharaohs: Genetic Comparison of Egyptian Cat Mummies to their Feline Contemporaries. *J Archaeol Sci* **39**: 3217–3223.

Kurushima JD, Lipinski MJ, Gandolfi B, Froenicke L, Grahn JC, Grahn RA, *et al.* (2013). Variation of cats under domestication: genetic assignment of domestic cats to breeds and worldwide random-bred populations. *Anim Genet* **44**: 311–324.

Larson G, Burger J (2013). A population genetics view of animal domestication. *Trends Genet* **29**: 197–205.

Lecis R, Pierpaoli M, Birò ZS, Szemethy L, Ragni B, Vercillo F, *et al.* (2006). Bayesian analyses of admixture in wild and domestic cats (Felis silvestris) using linked microsatellite loci. *Mol Ecol* **15**: 119–131.

Li G, Davis BW, Eizirik E, Murphy WJ (2016). Phylogenomic evidence for ancient hybridization in the genomes of living cats (Felidae). *Genome Res* **26**: 1–11.

Li Y, Fujiwara K, Osada N, Kawai Y, Takada T, Kryukov AP, *et al.* (2020). House mouse Mus musculus dispersal in East Eurasia inferred from 98 newly determined complete mitochondrial genome sequences. *Heredity* **126**: 132–147.

Lipinski MJ, Froenicke L, Baysac KC, Billings NC, Leutenegger CM, Levy AM, *et al.* (2008). The ascent of cat breeds: genetic evaluations of breeds and worldwide random-bred populations. *Genomics* **91**: 12–21.

MacHugh DE, Larson G, Orlando L (2017). Taming the Past: Ancient DNA and the Study of Animal Domestication. *Annu Rev Anim Biosci* **5**: 329–351.

Málek J (1993). *The cat in ancient Egypt*. Published by British Museum Press for the Trustees of the British Museum.

Malomane DK, Weigend S, Schmitt AO, Weigend A, Reimer C, Simianer H (2020). Genetic diversity in global chicken breeds as a function of genetic distance to the wild populations. *bioRxiv*: 2020.01.29.924696.

Mattucci F, Galaverni M, Lyons LA, Alves PC, Randi E, Velli E, *et al.* (2019). Genomic approaches to identify hybrids and estimate admixture times in European wildcat populations. *Sci Rep* **9**: 11612.

Nutter FB, Levine JF, Stoskopf MK (2004). Reproductive capacity of free-roaming domestic cats and kitten survival rate. *J Am Vet Med Assoc* **225**: 1399–1402.

Oliveira R, Godinho R, Randi E, Alves PC (2008). Hybridization versus conservation: are domestic cats threatening the genetic integrity of wildcats (Felis silvestris

silvestris) in Iberian Peninsula? *Philos Trans R Soc Lond B Biol Sci* **363**: 2953–2961.

Oliveira R, Godinho R, Randi E, Ferrand N, Alves PC (2008). Molecular analysis of hybridisation between wild and domestic cats (Felis silvestris) in Portugal: implications for conservation. *Conserv Genet* **9**: 1–11.

Oliveira R, Randi E, Mattucci F, Kurushima JD, Lyons LA, Alves PC (2015). Toward a genome-wide approach for detecting hybrids: informative SNPs to detect introgression between domestic cats and European wildcats (Felis silvestris). *Heredity* **115**: 195–205.

Ottoni C, Van Neer W, De Cupere B, Daligault J, Guimaraes S, Peters J, *et al.* (2017). The palaeogenetics of cat dispersal in the ancient world. *Nature Ecology &Amp; Evolution* **1**: 0139.

Patterson N, Price AL, Reich D (2006). Population structure and eigenanalysis. *PLoS Genet* **2**: e190.

Pickrell JK, Pritchard JK (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* **8**: e1002967.

Pickrell JK, Reich D (2014). Toward a new history and geography of human genes informed by ancient DNA. *Trends Genet* **30**: 377–389.

Plein M, O'Brien KIR, Holdenb MH, Adamsa MP, Baker CM, Bean NG, *et al.* (2022). Modeling total predation to avoid perverse outcomes from cat control in a data-poor island ecosystem. *Conserv Biol*.

Pontius JU, Mullikin JC, Smith DR, Agencourt Sequencing Team, Lindblad-Toh K, Gnerre S, *et al.* (2007). Initial sequence and comparative analysis of the cat genome. *Genome Res* **17**: 1675–1689.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Quilodrán CS, Nussberger B, Macdonald DW, Montoya-Burgos JI, Currat M (2020). Projecting introgression from domestic cats into European wildcats in the Swiss Jura. *Evol Appl* **13**: 2101–2112.

Rajabi-Maham H, Orth A, Bonhomme F (2008). Phylogeography and postglacial expansion of Mus musculus domesticus inferred from mitochondrial DNA coalescent, from Iran to Europe. *Mol Ecol* **17**: 627–641.

Raj A, Stephens M, Pritchard JK (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**: 573–589.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A* **102**: 15942–15947.

Randi E, Pierpaoli M, Beaumont M, Ragni B, Sforzi A (2001). Genetic identification of wild and domestic cats (Felis silvestris) and their hybrids using Bayesian clustering methods. *Mol Biol Evol* **18**: 1679–1693.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009). Reconstructing Indian population history. *Nature* **461**: 489–494.

Rendall AR, Sutherland DR, Baker CM, Raymond B, Cooke R, White JG (2021). Managing ecosystems in a sea of uncertainty: invasive species management and assisted colonizations. *Ecol Appl* **31**: e02306.

Ruiz-Garcia M, Alvarez D, Shostell JM (2005). Population genetic analysis of cat populations from Mexico, Colombia, Bolivia, and the Dominican Republic: identification of different gene pools in Latin America. *J Genet* **84**: 147–171.

Sauther ML, Bertolini F, Dollar LJ, Pomerantz J, Alves PC, Gandolfi B, *et al.* (2020). Taxonomic identification of Madagascar's free-ranging 'forest cats'. *Conserv Genet* **21**: 443–451.

Scheu A, Powell A, Bollongino R, Vigne J-D, Tresset A, Çakırlar C, *et al.* (2015). The genetic prehistory of domesticated cattle from their origin to the spread across Europe. *BMC Genet* **16**: 54.

Séré M, Thévenon S, Belem AMG, De Meeûs T (2017). Comparison of different genetic distances to test isolation by distance between populations. *Heredity* **119**: 55–63.

Shah NM, Al-Qudsi SS (1989). The Changing Characteristics of Migrant Workers in Kuwait. *Int J Middle East Stud* **21**: 31–55.

Shannon LM, Boyko RH, Castelhano M, Corey E, Hayward JJ, McLean C, *et al.* (2015). Genetic structure in village dogs reveals a Central Asian domestication origin. *Proc Natl Acad Sci U S A* **112**: 13639–13644.

Spencer PBS, Yurchenko AA, David VA, Scott R, Koepfli K-P, Driscoll C, *et al.* (2015). The Population Origins and Expansion of Feral Cats in Australia. *J Hered* **107**: 104–114.

The Cat-Show (1871). *Penny Illustrated Paper, The Naturalist*: 511.

Van Neer W, Linseele V, Friedman R, De Cupere B (2014). More evidence for cat taming at the Predynastic elite cemetery of Hierakonpolis (Upper Egypt). *J Archaeol Sci* **45**: 103–111.

Vigne J-D, Briois F, Zazzo A, Willcox G, Cucchi T, Thiébault S, *et al.* (2012). First wave of cultivators spread to Cyprus at least 10,600 y ago. *Proc Natl Acad Sci U S A* **109**: 8445–8449.

Vigne J-D, Guilaine J, Debue K, Haye L, Gérard P (2004). Early taming of the cat in Cyprus. *Science* **304**: 259.

Witzenberger KA, Hochkirch A (2014). The genetic integrity of the *ex situ* population of the European wildcat (Felis silvestris silvestris) is seriously threatened by introgression from domestic cats (Felis silvestris catus). *PLoS One* **9**: e106083.

Yu H, Xing Y-T, Meng H, He B, Li W-J, Qi X-Z, *et al.* (2021). Genomic evidence for the Chinese mountain cat as a wildcat conspecific (*Felis silvestris bieti*) and its introgression to domestic cats. *Science Advances* **7**: eabg0221.

Zeder MA (2012). The Domestication of Animals. *J Anthropol Res* **68**: 161–190.

Bellwood P, Gamble C, Le Blanc SA, Pluciennik M, Richards M, Terrell JE. First Farmers: the Origins of Agricultural Societies. Blackwell, 2005; Cambridge Archaeological Journal. 2007 Feb;17(1):87-109.

Koch K, Algar D, Searle JB, Pfenninger M, Schwenk K. A voyage to Terra Australis: human-mediated dispersal of cats. BMC Evol Biol. 2015 Dec 4;15:262.

Kurushima JD. Genetics. University of California, Davis, ProQuest Dissertations and Theses; 2011. Genetic Analysis of Domestication Patterns in the Cat (Felis catus): Worldwide Population Structure, and Human-mediated Breeding Patterns Both Modern and Ancient. PhD dissertation; p. 148. (Publication No. AAT 11271.)

Spencer PB, Yurchenko AA, David VA, Scott R, Koepfli KP, Driscoll C, O'Brien SJ, Menotti-Raymond M. The Population Origins and Expansion of Feral Cats in Australia. J Hered. 2016 Mar;107(2):104-14.

Todd, NB. (1977) Cats and Commerce. Scientific American 237, 5:100-107.

Figures

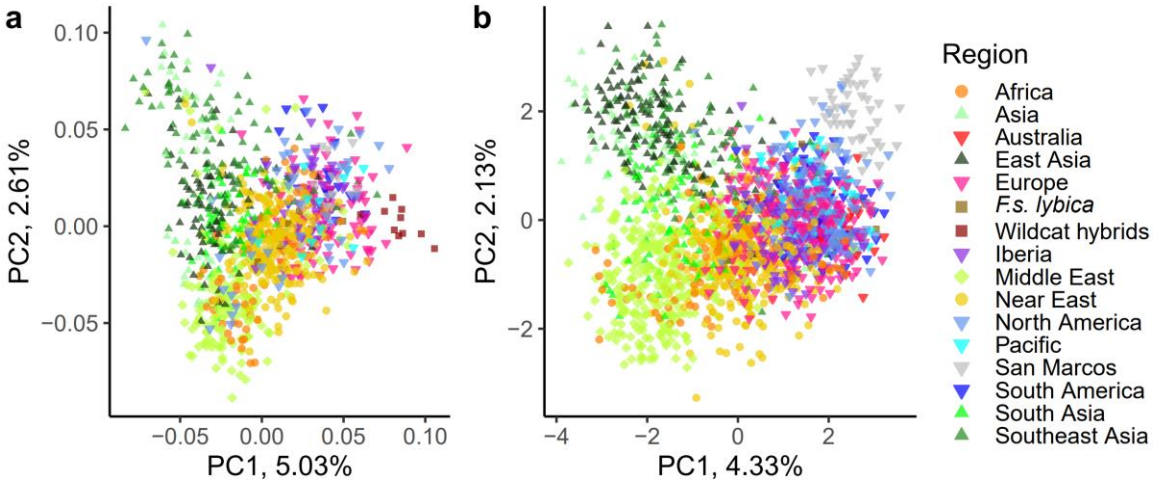


Figure 1. Principal component analyses (PCA) of genetic variation in random-bred and wildcat felines. a PCA plot of SNP data (N = 983). **b** PCA plot of STR data (N = 1,861). A single point represents an individual, the shape represents a geographic region, the color represents a geographic sub-region. The two wildcat populations are denoted by squares of different colors. Middle Eastern, South Asia, and Western European cats form the peripheral subpopulations of random-bred cats. The wildcat hybrids and the island population of San Marcos, Baja California, are additional peripheral populations for **a** and **b,** respectively.

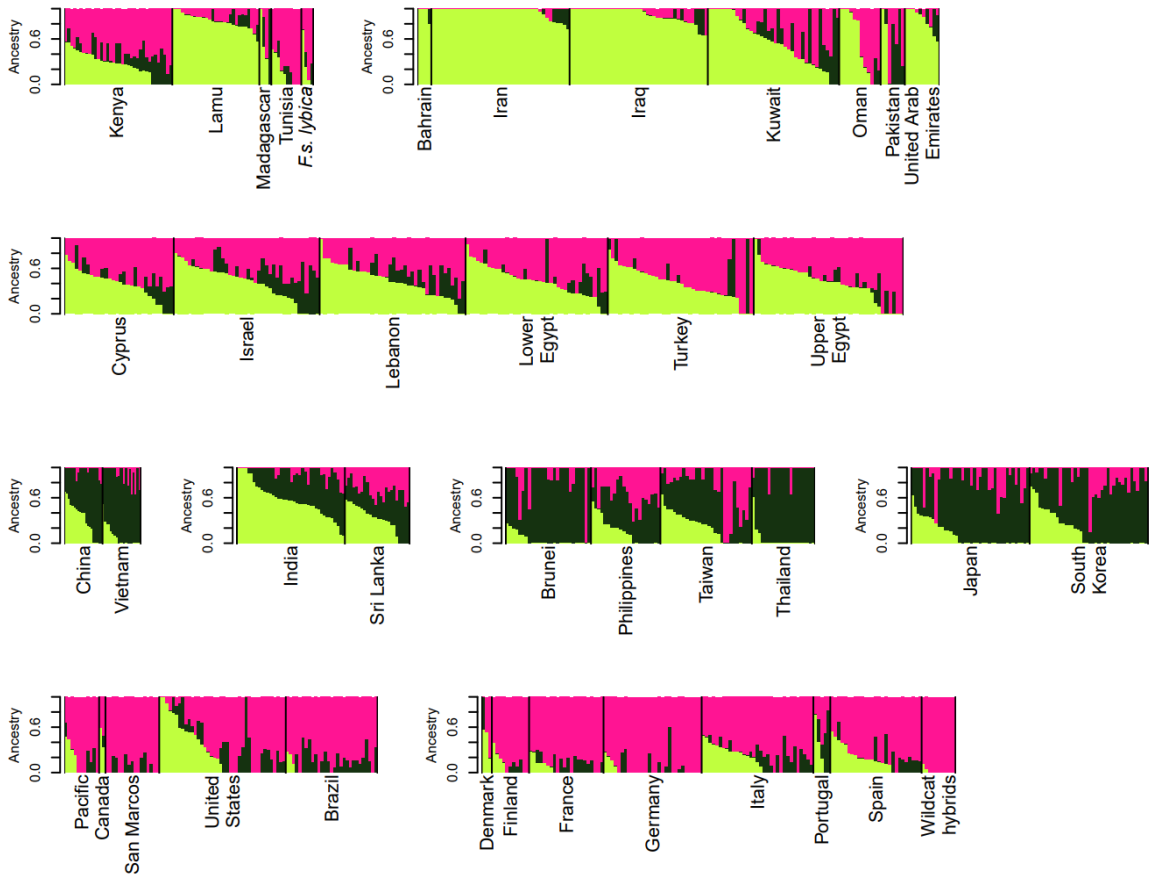


Figure 2. Random-bred cat population SNP fastSTRUCTURE plot of *K* = 3.

Population contributions are represented by different colors, individual vertical bars represent an individual, and populations are separated by black lines.

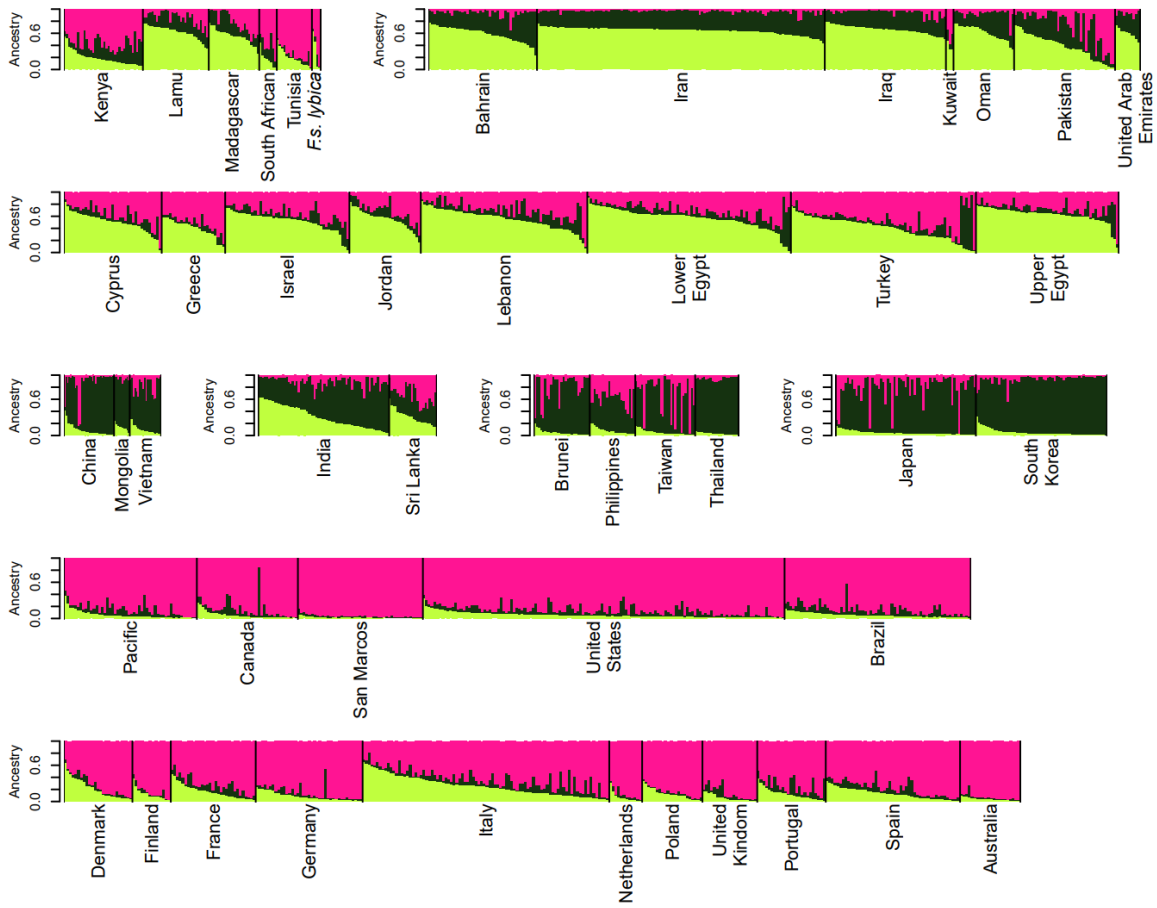


Figure 3. Random-bred cat population STR STRUCTURE plot of *K* = 3. Population contributions are represented by different colors, individual vertical bars represent an individual, and populations are separated by black lines.

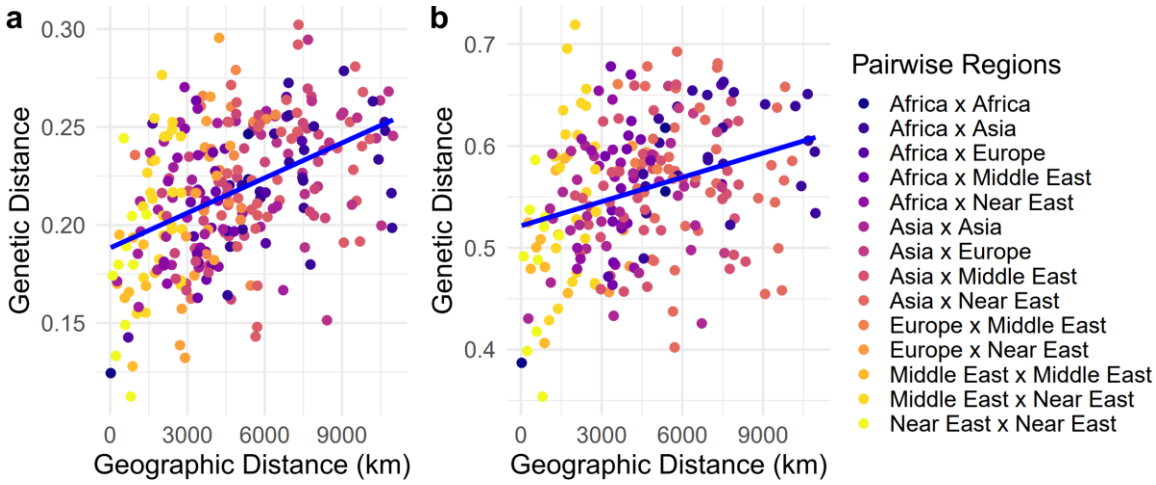


Figure 4. Comparison of geographic distance and genetic distance of random-bred cat populations. a Plot of SNP data with 24 sample locations with a regression line indicating a correlation of 0.447 with a p-value of 0.001. **b** Plot of STR data with 22 sample locations with a regression line indicating a correlation of 0.302 with a p-value of 0.0076. Each point represents an individual pairwise comparison of sample location populations.

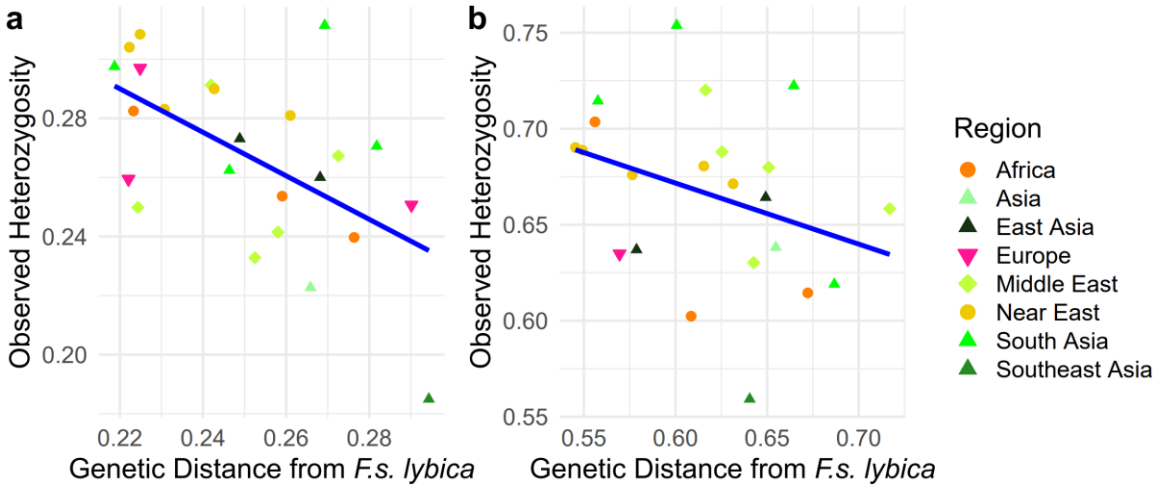


Figure 5. Comparison of genetic distance from *F.s. lybica* and observed heterozygosity. A Plot of SNP data with a regression line indicating a correlation of -0.57 with a p-value of 0.0034. **B** Plot of STR data with a regression line indicating a correlation of -0.33 with a p-value of 0.13. Each point represents a sample location population, shape represents a geographical region, and color represents a geographical sub-region.

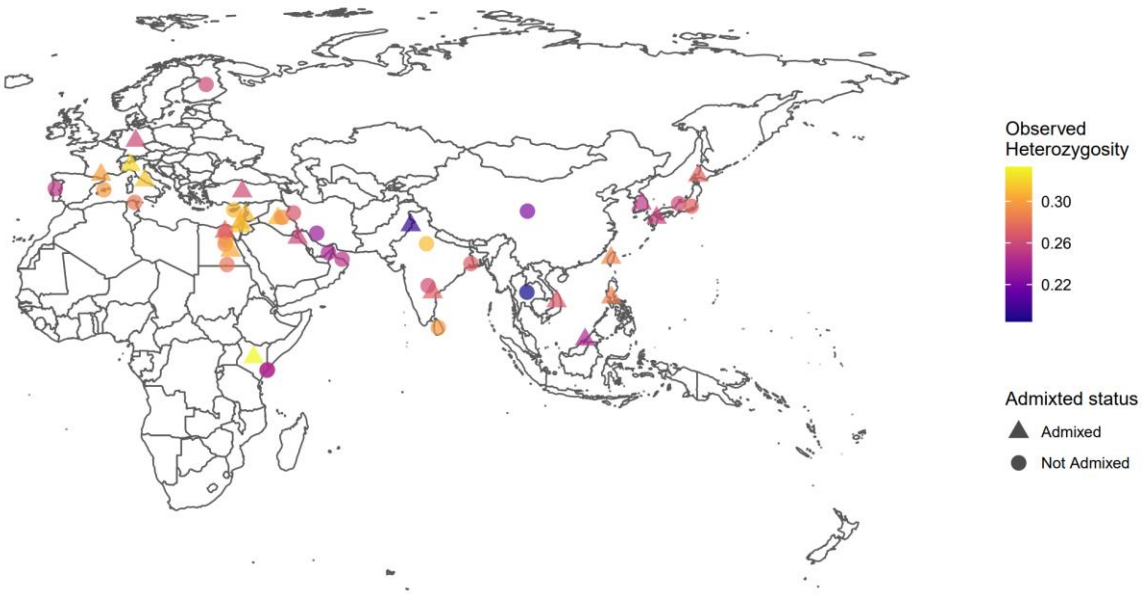
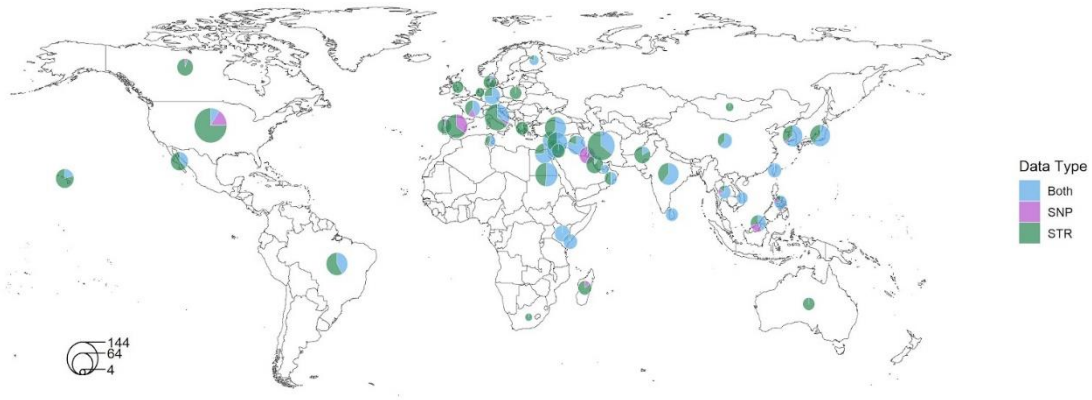
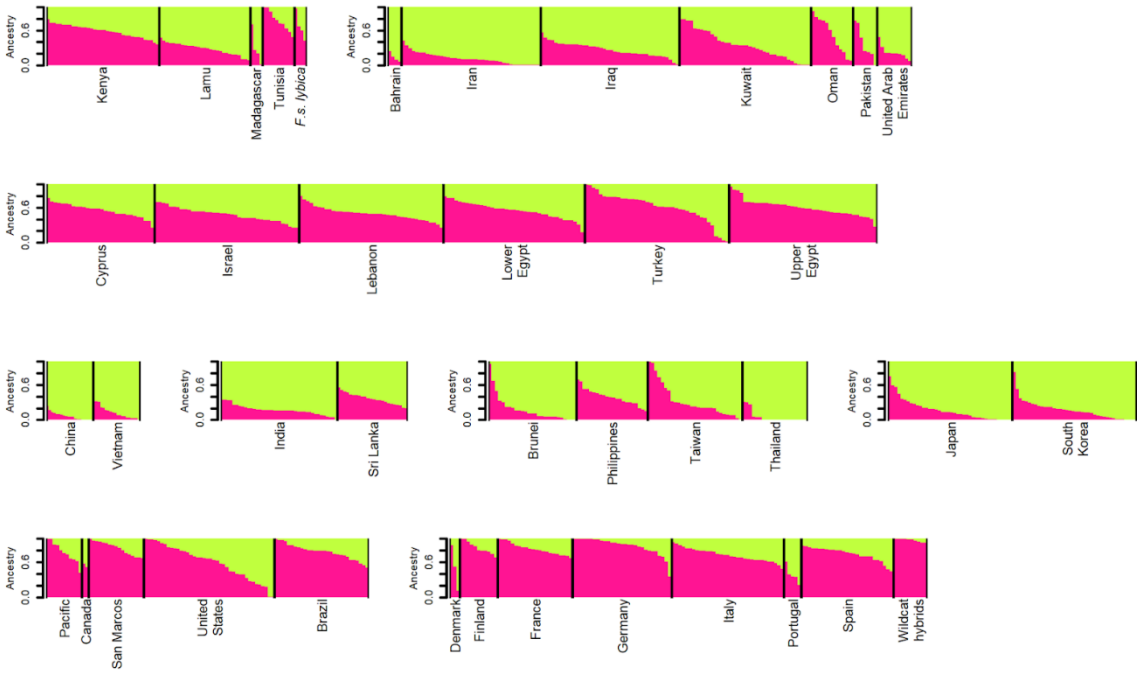


Figure 6. Observed heterozygosity by sample location of SNP data for random-bred cat populations of Eurasia. Each point represents a sample location population with the color showing the calculated observed heterozygosity. The triangle shape indicates an admixed population with a significant f_3 statistic, and the circle shape represents non-admixed populations. Populations of yellow and light orange shades are focused in the Near East and Mediterranean Basin.

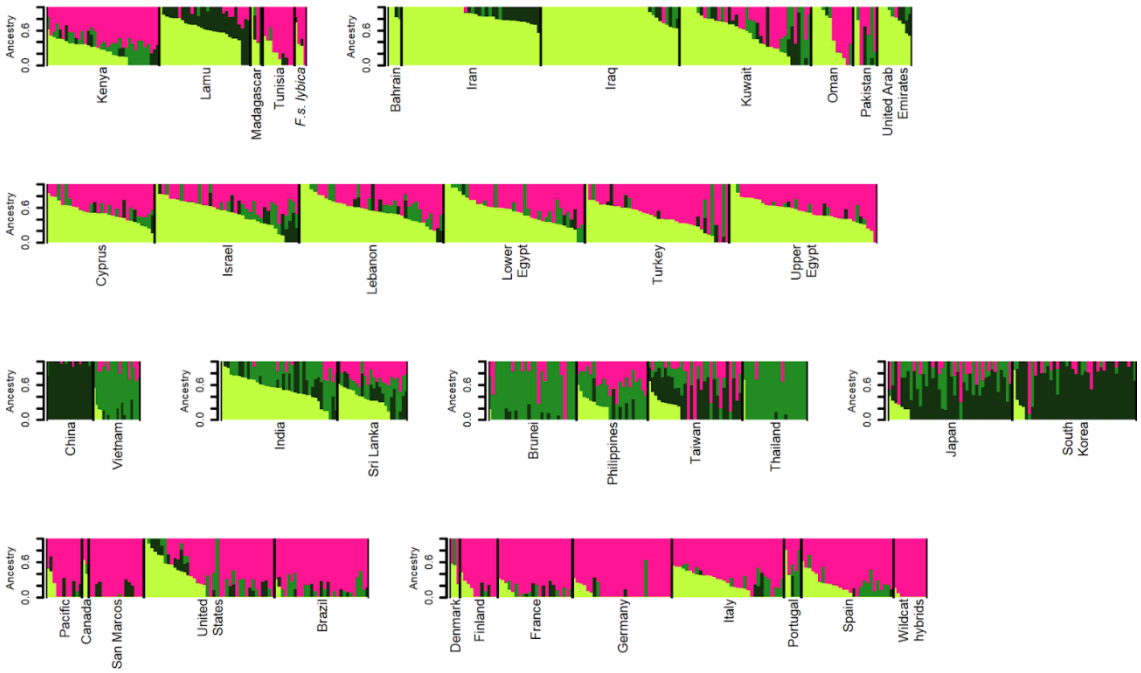
Supplementary Figures



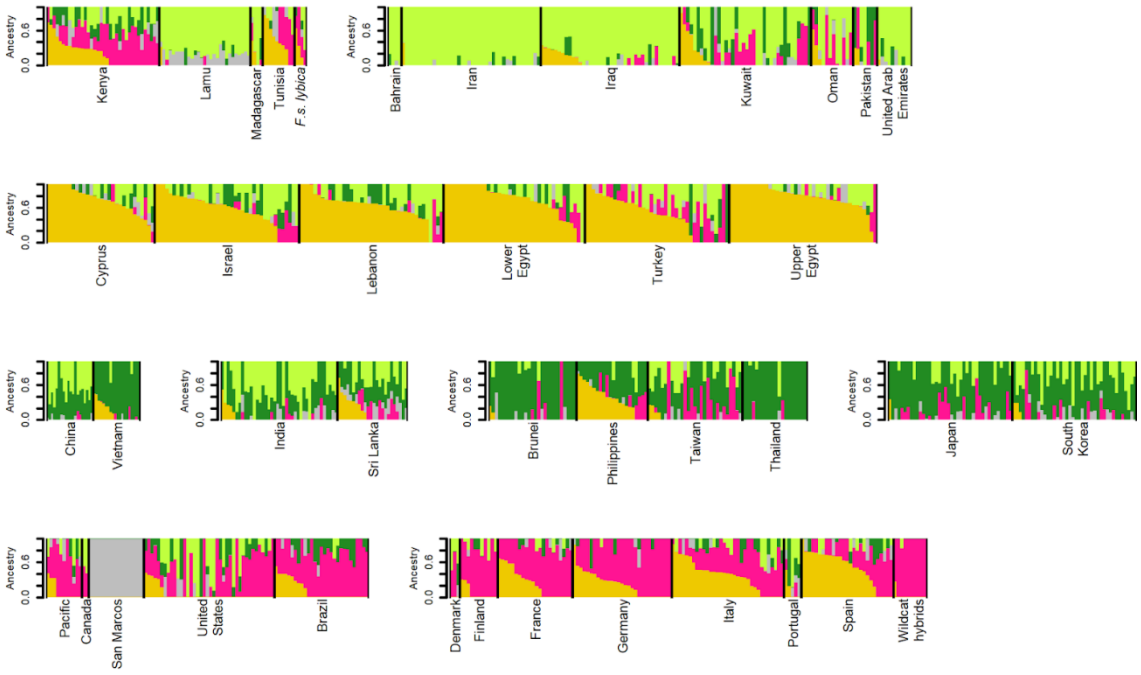
Supplementary Figure 1. Location and data type of sample populations. Each color corresponds to if an individual has both SNP and STR genotypes, only SNP genotypes, or only STR genotypes. The size of the pie is proportional to the number of individuals sampled in that location.



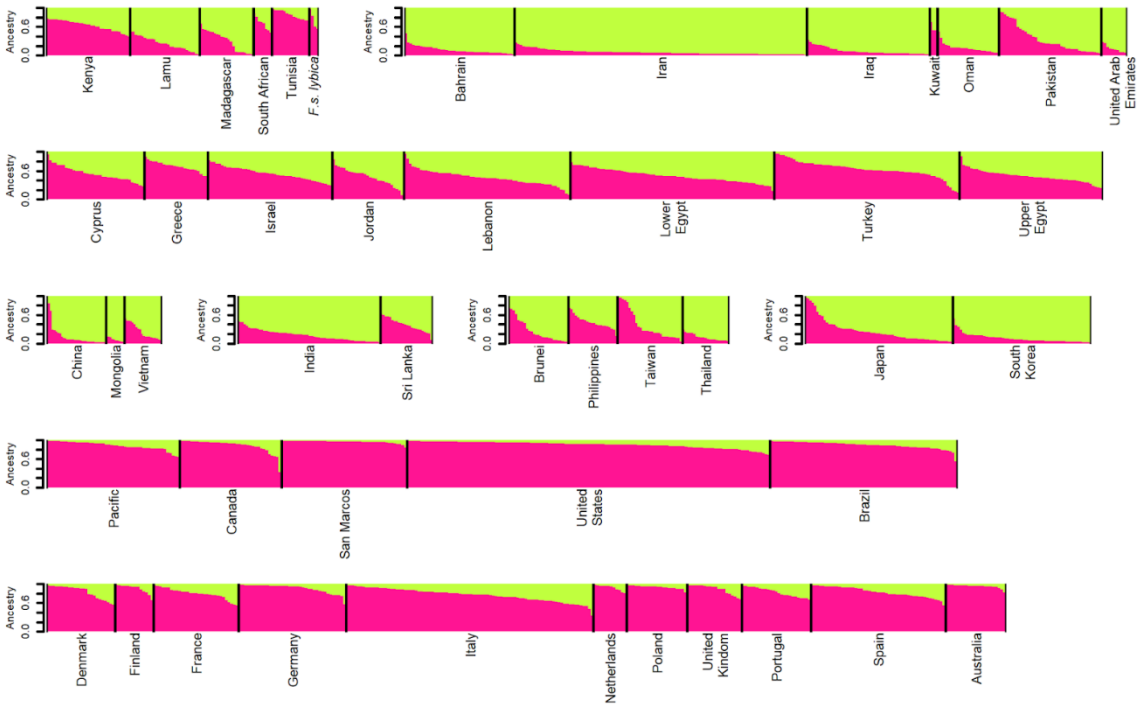
Supplementary Figure 2. SNP fastSTRUCTURE plot of *K* = 2. Population contributions are represented by different colors, individual vertical bars represent an individual, and populations are separated by black lines.



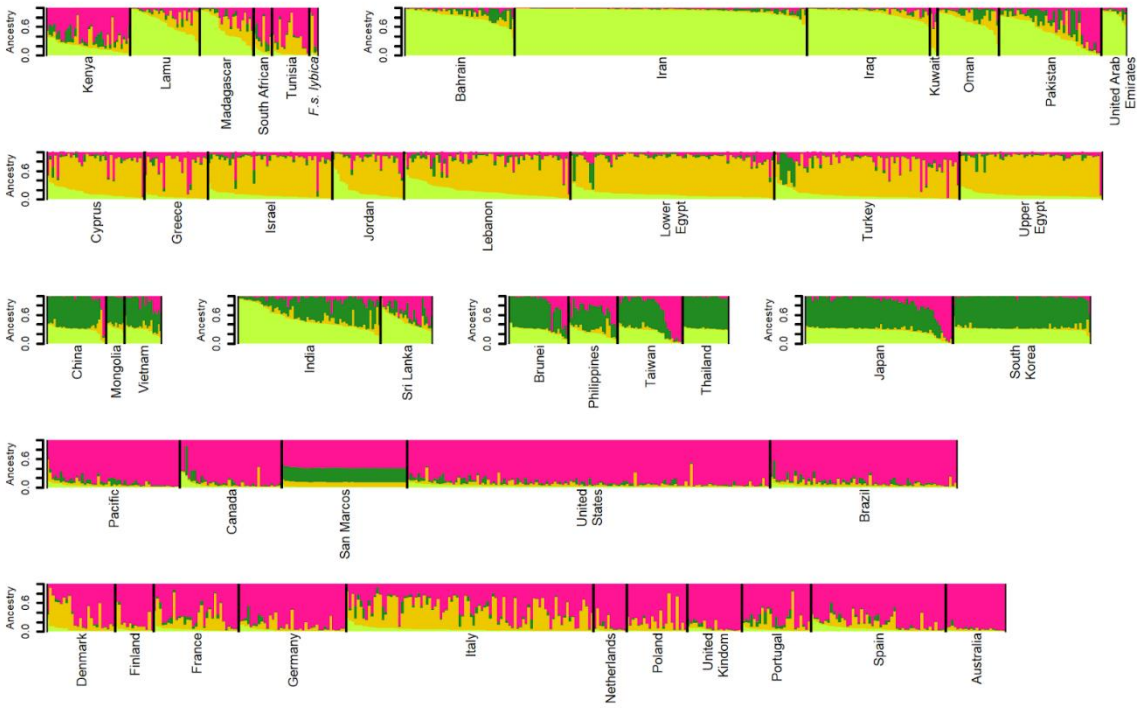
Supplementary Figure 3. SNP fastSTRUCTURE plot of *K* = 4. Population contributions are represented by different colors, individual vertical bars represent an individual, and populations are separated by black lines.



Supplementary Figure 4. SNP fastSTRUCTURE plot of *K* = 5. Population contributions are represented by different colors, individual vertical bars represent an individual, and populations are separated by black lines.



Supplementary Figure 5. STR STRUCTURE plot of *K* = 2. Population contributions are represented by different colors, individual vertical bars represent an individual, and populations are separated by black lines.



Supplementary Figure 6. STR STRUCTURE plot of *K* = 4. Population contributions are represented by different colors, individual vertical bars represent an individual, and populations are separated by black lines.

Supplementary Tables

Available upon request by Dr. Jared Decker at deckerje@missouri.edu

CHAPTER 2

Prediction Accuracy for Genotype-by-Environment Models Applied to Growth Traits of US Gelbvieh Beef Cattle

Sara Nilson[1], Troy Rowan[1,2,3,4], Robert Schnabel[1,5], Jared Decker[1,5]*

[1]Division of Animal Sciences, University of Missouri, Columbia, MO, 65211, USA

[2]Genetics Area Program, University of Missouri, Columbia, MO, 65211, USA

[3]Department of Animal Science, University of Tennessee, Knoxville, TN, 37996, USA

[4]College of Veterinary Medicine, Large Animal Clinical Science, University of Tennessee, Knoxville, TN, 37996, USA

[5]Institute for Data Science and Informatics, University of Missouri, Columbia, MO, 65211, USA

Abstract

Background

Climate change and the growing human population is driving the need for sustainable agriculture. In the beef industry, we need selection for environmental resiliency to increase production by improving resource efficiency and maintaining animal welfare. Exposed to the elements during their entire life, beef cattle must be identified and selected for enhanced yield in the context of their production environment. Cattle frequently re-rank for their genetic potential across environments due to genotype x

environment interactions which can lead to unintended decreases in production. Incorporation of interactions in genomic prediction models on a national level has yet to be considered even though the potential positive impacts have been acknowledged.

Results

A model including a genotype x environment interaction out performed a bivariate genotype x environment model and the currently utilized national evaluation model in terms of accuracy. This model provided clarity by separating the phenotypic variance of an individual into additive genetic and genotype x environment components. The genotype x environment model estimated that ~3%-12% of the variance in the production traits of birth weight, weaning weight, and yearling weight was due to the environment. By accounting for the genotype x environment interaction, additive genetic prediction accuracies tended to rise over the accuracy of prediction achieved by using the currently utilized national model. Additionally, a selection informative genotype x environment deviation was estimated conveying the expected difference of the observed phenotype due to the environment. Minimal reranking of animals was observed between the national model and the genotype x environment model.

Conclusions

The beef industry should capitalize on the increased accuracy of additive genetic genomic predictions for production traits by accounting for genotype x environment interactions. Breeders and producers will be able to identify and select animals who are best genetically suited to production in their environment, and with the added assistance of the estimated genotype x environment deviation conduct genetic improvement at a faster rate.

There is potential for enhancing prediction accuracy through defining and refinement of genetic interactions not only with the environment but also with management practices. By adapting genomic prediction advances with the inclusions of genotype x environment interactions, the industry will reach sustainable intensification goals improving food security for future generations.

Background

The growing human population and climate change have brought the importance of food security to the forefront of awareness creating a push for sustainable intensification, especially in animal production(1–5). Sustainable intensification with regards to livestock production means the improvement of production to minimize the yield gap, improve resource efficiency, and enhanced animal welfare(3,5,6). Beef cattle create efficiency by occupying land unsuitable for agricultural crops and by transforming low-quality food stuffs to high-quality protein. As cattle are one of the last agricultural species living their entire lives outside, they are exposed to a vast spectrum of environmental pressures and variation under which they are expected to perform. Cattle may be losing their adaptation to specific environments through the widespread use of artificial insemination and the shipping of semen globally(7). Consequently, the development of tools to identify and select cattle that are suited for improved production in a specific environment has become increasingly important to the beef industry(7). By taking into account the combined additive genetic potential and environment-specific genetic deviation of individuals through inclusion of genotype x environment interactions (GxE) in prediction models,

producers could make more informed selection decisions given their location and climate allowing for an enhancement in the sustainability of production.

GxE refers to when an individual's genotypic value for a trait varies according to the environment in which the trait is expressed, resulting in a change in the observed phenotype. This phenomenon is frequently described as phenotypic plasticity or environment sensitivity. Phenotypic response to the environment has been explored in cattle for health, production traits, and adaptation to environmental stressors which all have implications for improving sustainability and welfare (e.g., tick resistance, body weight, and production in tropical climates)(8–12). Frequently utilized methods in cattle examine the GxE response by estimating the genetic value of an individual across an environmental gradient pinpointing the optimal environmental conditions under which the individual's genetic potential is maximized(11–17). Unfortunately, these methods are usually limited to a single environmental variable which is considered responsible for the most environmentally sensitive component of phenotypic variation, while in reality, the environment is a complex system of interactions. Additionally, these methods often lack clarity when the genetic value of an individual is reported an average effect across environments and the GxE deviation from additive genetic merit is ignored. While many studies in cattle have concluded that GxE effects should be included in predictions and used in making selection decisions, no beef cattle breed association in the United States has adopted the application of GxE deviations from the average in their predictions to date(18–20).

By analyzing data for the Gelbvieh beef cattle breed which is widely geographically distributed across the United States, this study applied a GxE model that was not

restricted to a single environmental variable. To clarify the relationship between the genotype and the environment, the GxE model estimates and partitions the variance contributions to the phenotype separately into the additive genetic merit and the GxE interaction. Additionally, we compared the accuracy of predictions from the GxE model to those produced using the current national evaluation model and a bivariate GxE model. We show that the inclusion of the GxE interaction in genomic predictions can profoundly impact the prediction of additive genetic merit for production traits. Consequently the use of GxE models will allow more informed selection decisions within the beef industry enhancing food security for future generations.

Methods

Data

Phenotypic and genotype data were provided by the American Gelbvieh Association for individuals born between 1972 and 2016. The analyzed phenotypes included birth weight (BW), 205-day adjusted weaning weight (WW), and 365-day adjusted yearling weight (YW). Records with a BW of 0lbs, WW greater than 1100lbs, YW greater than 1700lbs, and when the recorded WW was greater than the YW were all set to missing. All analyses were restricted to the 12,561 individuals with genotypes in the post-filter dataset regardless of the presence of missing phenotypes.

Genotypes and Imputation

Genotyped loci varied according to the utilized assays which had differing single nucleotide polymorphism (SNP) densities and included the GeneSeek GGP-LDv3, GeneSeek GGP-LDv4, GeneSeek GGP-90KT, GeneSeek GGP-HDv3, GeneSeek GGP-F250, Illumina BovineSNP50, and Illumina BovineHD. Genotypes were imputed to the union set of ~830k autosomal SNPs using the pipeline described by Rowan *et al.* (2021). Briefly, the process included: genotype quality control performed in PLINK (v1.9), referenced-based phasing with Eagle (v2.4), and imputation with Minimac3 (v2.0.1)(22–25). SNP coordinates were from the ARS-UCD1.2 bovine reference genome(26). After filtering for minor allele frequency greater than 0.01 in PLINK (v1.9), there were 715,397 SNPs available for analysis(22,23).

Ecoregion Definition

To account for the multivariable complexity of the environment, we utilized the nine discrete ecoregions described by Rowan *et al.* (2021). These were defined using k-means clustering of the 30-year normals for environmental variables: mean temperature, precipitation (mm/year), and elevation (m above sea level)(7). Individuals were assigned to ecoregions using breeder supplied farm zip codes. Animals located in zip codes with multiple ecoregion assignments had their assigned ecoregion set to missing. Ecoregions were categorized as Desert (DT), Southeast (SE), High Plains (HP), Rainforest, Arid Prairie, Foothills, Forested Mountains (FM), Fescue Belt (FB), and the Upper Midwest and Northeast (UMN)(7). Ecoregions with less than 100 individuals with phenotypes

were not analyzed to minimize estimation sampling variance, which excluded: Rainforest, Arid Prairie, and the Foothills.

Pre-Correcting of Phenotypes

Phenotypes were pre-corrected for contemporary group (CG) and sex. Contemporary groups were defined as birth year, birth season (Spring or Fall), breeder zip code, and additional CG information provided by the American Gelbvieh Association (AGA). Breeder zip code was utilized as a proxy for herd identification due to the unavailability of a provided herd identifier. All three phenotypes were adjusted using two separate analyses with sex as a fixed effect and CG as a fixed or random effect due to 80.1% of the CGs having less than 5 individuals for BW, 80.6% for WW, and 81.7% for YW. When CG assignment was missing or a CG had less than 5 individuals, a 'breed average' CG level was assigned when CG was fit as a fixed effect. This 'breed average' level was constrained to zero during CG effect estimation and when used in phenotype adjustment. When CG was fit as a random effect and CG information was missing, the CG level was defined as missing. Sex was defined as male, female, or unknown with the unknown level being constrained to zero during phenotype adjustement. For BW and WW, an additional random maternal effect was included in both adjustment models and was defined as the dam identification number. Fixed effects were estimated with the '--reml-est-fix' option while random effects were predicted using best linear unbiased prediction with the '--reml-pred-rand' option in GCTA with a single-trait animal model while controlling for population structure by including a genomic relationship matrix (GRM) created with the '--make-grm-alg 1' option(27). Random effects were included in the models by first creating an incidence matrix of the effects levels and then multiplying it by its transpose

to obtain a relationship matrix for the effect for the 12,561 genotyped individuals. Adjusted phenotypes with CG estimated either as fixed or random effects were utilized in all downstream analyses of BW, WW, and YW.

Variance Component Estimation

Variance components were estimated for three types of models including: the national, a univariate GxE, and a GxE bivariate. Variance components for the national models were estimated for each set of adjusted BW, WW, and YW phenotypes by restricted maximum likelihood in GCTA with the following univariate linear mixed model:

$$y^* = \mu + Z_A a_A + e$$

where $\boldsymbol{y}^*$is a vector of adjusted phenotypes, $\boldsymbol{\mu}$ is the overall mean, $Z_A$is the incidence matrix relating the adjusted phenotypes to the random additive genetic effects, $\boldsymbol{a_A} \sim$ N(0, $\boldsymbol{G}\sigma_a^2$) is a vector of random additive genetic effects, and $\boldsymbol{e} \sim$ N(0, $\boldsymbol{I}\sigma_e^2$) is a vector of random residuals. $\boldsymbol{G}$ is the GRM and $\boldsymbol{I}$ is an identity matrix. Variance components for the bivariate models were estimated using data for the three largest ecoregions for the adjusted BWs using GCTA with the following model:

$$Y^* = Xb + Z_A a_A + e$$

where $\boldsymbol{Y}^*$is the *n x 2* matrix of adjusted phenotypes for two ecoregions, X is the incidence matrix relating overall means to ecoregion phenotypes, b is the vector containing the overall means, $Z_A$is an incidence matrix relating the adjusted phenotypes to the random additive genetic effects, $\boldsymbol{a_A} \sim$ N(0, $\begin{bmatrix} \boldsymbol{G}\sigma_{a,t1}^2 & rG \\ rG & \boldsymbol{G}\sigma_{a,t2}^2 \end{bmatrix}$) are the random additive genetic

effects, and $\boldsymbol{e} \sim N(0, \begin{bmatrix} \boldsymbol{G}\sigma^2_{e,t1} & 0 \\ 0 & \boldsymbol{G}\sigma^2_{e,t2} \end{bmatrix})$ are the random residuals. Variance components for the GxE model were estimated for each set of adjusted BW, WW, and YW phenotypes using the following model:

$$y^* = \mu + X_E b_E + Z_A a_A + Z_{GxE} a_{GxE} + e$$

where $\boldsymbol{y}^*$is the vector of adjusted phenotypes, $\boldsymbol{\mu}$ is the overall mean, b_E is a vector of ecoregion environment effects, X_E is the incidence matrix relating the adjusted phenotypes to the ecoregion environment effects, Z_A is the incidence matrix relating the adjusted phenotypes to the random additive genetic effects, $\boldsymbol{a_A} \sim N(0, \boldsymbol{G}\sigma_a^2)$ are the random additive genetic effects, Z_{GxE} is an incidence matrix relating the adjusted phenotypes to the random GxE effects , $\boldsymbol{a_{GxE}} \sim N(0, \boldsymbol{G_{GxE}}\sigma_{GxE}^2)$ is a vector of random GxE effects, and $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$ is the vector of random residuals. $\boldsymbol{G_{GxE}}$ is $\boldsymbol{G_{[i,j]}}$ when individuals $\boldsymbol{i}$ and $\boldsymbol{j}$ are from the same ecoregion and otherwise zero(28). Estimated variance components for all three models applied to both sets of adjusted phenotypes were utilized for downstream analyses.

Estimating Breeding Values and Validation Set Determination

Breeding values were estimated (EBV) for individuals with all models: the national model, the GxE model, and three bivariate models when CG were fit as either a fixed or random effect using blupf90 from the BLUPF90 suite of programs(29). All three adjusted BW, WW, and YW phenotypes were analyzed with the national and GxE model resulting in six analyses for each model for the two sets of adjusted phenotypes. However, the three bivariate models were used only to analyze the adjusted BWs resulting in six

analyses. All models were analyzed as genomic best linear unbiased predictions (only included genomic relationships, not pedigree relationships). To ensure that $\boldsymbol{G}$ was nonsingular, $\boldsymbol{G}$ was blended with $\boldsymbol{A}$ as $0.99\boldsymbol{G} + 0.01\boldsymbol{A}$ where $\boldsymbol{A}$ was the pedigree relationship matrix for these animals which was set to an $\boldsymbol{I}$ matrix since no pedigree information was used.

The LR Method was used to evaluate prediction performance(30,31). The EBVs estimated under the three types of models used the whole phenotypic dataset and a partial dataset where the youngest 10% of individuals identified by their birth date had their phenotypes set to missing to create a validation set that could be used to evaluate model bias and determine prediction accuracy. The youngest 10% of animals represent the animals which are selection candidates in the next generation. The validation set of individuals were ecoregion distributed in which the youngest 10% of individuals within each ecoregion had their phenotype set to missing (Figure 1). The ecoregion distributed validation set was proposed to test the robustness of the predicted EBVs as cattle are not evenly spatially distributed across environments, and a more densely populated ecoregion could dominate the validation set which may not be genetically representative of the next generation of selection candidates industry wide. The bivariate models analyzed adjusted phenotypes within pairs of ecoregions, their validation set included the youngest 10% of individuals within each analyzed ecoregion which overlapped with individuals in the ecoregion distributed validation set.

Measures of Prediction Accuracy

Prediction accuracies were estimated by:

$$\widehat{acc}_{LR} = \sqrt{\frac{cov(\hat{u}_w, \hat{u}_p)}{(1 - \underline{F})\hat{\sigma}_u^2}}$$

where $\hat{u}_w$is the vector of EBVs of the validation set of individuals for the whole dataset, $\hat{u}_p$is the vector of predicted EBVs of the validation set of individuals for the partial dataset, $\underline{F}$ is the average inbreeding coefficient of the validation individuals, and $\hat{\sigma}_u^2$is the estimated additive genetic variance of the trait(30,31). The inbreeding coefficient was calculated in PLINK using the '--ibc' command(22,23). When estimating accuracy for the GxE models the formulae were unchanged for the additive genetic component (EBV_A), but differed when calculated for the GxE deviation (D_{GxE}) and the total combined EBV (EBV_{Total}= EBV_A + D_{GxE}). For the accuracy calculations for the D_{GxE}: $\hat{\sigma}_u^2$is replaced with $\hat{\sigma}_{GxE}^2$which is the estimated variance component for the D_{GxE}. For the accuracy calculations for EBV_{Total}: $\hat{\sigma}_u^2$ is replaced with $\hat{\sigma}_{Total}^2 = \hat{\sigma}_u^2 + \hat{\sigma}_{GxE}^2$.

Comparing Estimated Breeding Values Across Models

To compare EBVs across models, the validation sets' EBVs were plotted against each other in R (v3.4.3) and the slope of the line of best fit was estimated using the lm() function from the stats package (v3.4.3) to ascertain if the slope deviated from unity indicating prediction bias(32). Pearson and Spearman correlations were calculated among the validation individuals' EBVs between models in R (v3.4.3) with the cor() function from the stats package (v3.4.3) to estimate the strength of the linear relationship and the reranking of individuals(32).

Results

Variance Component Estimation

After quality control there were 12,561 genotyped individuals; Table 1 shows the distribution of individuals across phenotypes and ecoregions. The estimated variance contribution of CG to the phenotype when fitted as a random effect was 24.8%, 36.02%, and 42.06% for BW, WW, and YW, respectively. Birth and weaning weight were additionally adjusted for a random maternal effect which was estimated to account for: 15.1% and 10.1% of the BW phenotypic variance, and 4.86% and 3.08% of the WW phenotypic variance when CG was fit as a fixed or random effect, respectively. After adjusting for sex and management differences captured by the CG, the additive genetic variance components were estimated for the two sets of adjusted phenotypes.

First, variance components were estimated with a model similar to the currently utilized national genetic evaluation model which does not take into account a GxE effect. When CG were fit as a fixed effect, the heritabilities were estimated as 37%, 24.6%, and 37.49% for BW, WW, and YW, respectively. When CG was fitted as a random effect, the heritabilities were estimated as 39%, 28.93%, and 43.81% for BW, WW, and YW, respectively. These results suggest that for a national model management practices, environmental differences, or other genetic types besides additive may have a larger impact on the observed variance of WW as compared to the estimations for BW and YW.

Additive genetic variance components were then estimated with bivariate models for BW only for the three largest ecoregions: High Plains and Fescue Belt, High Plains and Upper Midwest & Northeast, and Fescue Belt and Upper Midwest & Northeast (Table 2). The

estimated genetic correlations for these models when CG was fitted as a fixed effect are: 0.95 for HP and FB, 0.98 for HP and UMN, and 0.83 for FB and UMN. When CG were fit as a random effect the estimated genetic correlations are: 1.0 for HP and FB, 0.98 for HP and UMN, and 0.94 for FB and UMN.

Finally, the additive genetic and D_{GxE} variance components were estimated for the GxE model when CG were fit as either a fixed or random effect (Table 3). The variance estimate for the D_{GxE} effect ranges from 5%-12% across life stage weights when the CG were fit as a fixed effect, and when CG were fit as a random effect the D_{GxE} effect variance estimates range from 3%-6% (Table 3).

Accuracy of Estimated Breeding Values

The accuracy of the predicted EBVs generated by each model were estimated with the LR method ($\widehat{acc}_{LR}$) which compares the EBV from a partial dataset to EBV from the whole dataset. The validation set of individuals were ecoregion distributed to test the robustness of the EBVs and the D_{GxE} due to an uneven distribution of the youngest animals across environments (Figure 1). For BW, the national model had a $\widehat{acc}_{LR}$~ 74% across the two-ways of adjusting for CG suggesting minimal differences between the adjustment methods (Table 4). For WW and YW, the national model had a slightly higher accuracy for when the CG were fit as random. Overall the national models' accuracies tended to be higher when CG were fit as random which could be due to the recovery of additional information by the inclusion of more CG levels.

For the bivariate model High Plains (validation n = 333) and Fescue Belt (validation n = 394), $\widehat{acc}_{LR}$ tended to increase slightly when CG were fit as random (Table 5). The second bivariate model of High Plains (n = 333) and Upper Midwest & Northeast (n = 100) had similar accuracies for CG adjustments, but the most noticeable difference in accuracy is when EBVs are predicted across ecoregions (Table 5). When trained in the larger ecoregion, High Plains, and validated in the smaller ecoregion, Upper Midwest & Northeast, High Plains animals EBVs prediction accuracy is 68%. When trained in the smaller ecoregion, Upper Midwest & Northeast, and validated in the larger ecoregion, High Plains, the Upper Midwest & Northeast individuals EBVs accuracy drops to ~41%-42%. This drop in accuracy appears to be a function of difference in sample size (HP, n = 3328; UMN, n = 999) since these two regions have a high estimated genetic correlation, 0.98, which would indicate a more similar environment, potentially similar GxE effects, and less reranking of individuals. Lastly, when Fescue Belt (n = 394) and Upper Midwest & Northeast (n = 100) were modeled the $\widehat{acc}_{LR}$were higher when the CG were fit as random. A similar decrease in accuracy was observed for predicting across ecoregions; trained in the larger ecoregion Fescue Belt, and validated in the smaller Upper Midwest & Northeast, animals from the Fescue Belt EBVs had higher accuracy (68% CG fixed, 73% CG random) while training in the smaller ecoregion Upper Midwest & Northeast and validating in the larger ecoregion Fescue Belt, the Upper Midwest & Northeast EBVs were less accurate (43% CG fixed and random). Collectively, the the three BW bivariate models follow the trend of a slightly higher $\widehat{acc}_{LR}$ when CG were fit as random, and a noticeable impact on accuracy when predicting across ecoregions due to differences in sample size.

The GxE models had accuracies calculated for the EBV_A which are directly comparable to the EBVs calculated by the other models, but accuracies were also calculated for the estimated D_{GxE} and EBV_{Total} to examine the full impact of including a GxE effect in genomic prediction. For weaning weight and yearling weight, the accuracies for the EBV_A and EBV_{Total}were higher when the CG were fit as a random effect (Table 6). The $\widehat{acc}_{LR}$of the D_{GxE} remained fairly consistent across the three phenotypes ranging from 11.7%-28% and usually was higher when CG were fit as a random effect. Ranking the types of models in terms of accuracies, the bivariate GxE model consistently has lower accuracy than the other two models. Compared to the national model, the GxE model had comparable accuracies for the predicting EBV_A of BW, slightly lower (~2%) for WW, and higher for YW (~10%).

Comparing Estimated Breeding Values Across Models

To compare and contrast the models beyond accuray, we estimated Pearson correlations, Spearman correlations, and the slopes of lines of best fit among the predicted EBVs from the partial datasets. Comparison statistics were not evaluated for the bivariate models due to decreased accuracy when compared to the other two models, the strong influence of sample size, and the need for multiple pairwise models to be analyzed to account for all environmental comparisons. Even though both the bivariate and GxE models are taking into account environmental effects, the univariate GxE model excelled in comparison due to the structure of jointly analyzing all individuals, phenotypes, and environments simultaneously. The national EBVs were compared to the GxE EBV_A and EBV_{Total} for

BW, WW, and YW when CG were fit as a random and fixed effect. When comparing the EBV_A from the GxE to the national EBVs, Pearson and Spearman correlations are consistently high for BW and WW with decreases for YW when the CG was fit as a random effect (Figure 2; see Additional File 1, Table S1). Correlations tend to be closer to 1 when the CG were fit as a fixed effect for BW and YW, and a random effect for WW. Slopes are closer to 1 for WW and YW when CG were fit as random while BW were closer when CG were fit as a fixed effect (Figure 2; see Additional File 1, Table S1). When comparing the GxE models' EBV_{Total} to the national EBVs the slopes follow the same trend as the EBV_A comparison (Figure 3; see Additional File 1, Table S1).

Altogether when comparing the GxE and national model, the GxE models' EBV_A were more similar to the national EBVs than the EBV_{Total} as indicated by higher correlations and less reranking of individuals which would be due to the inclusion of the D_{GxE} (see Additional File 1, Table S1). Birth weight differences between the national and GxE models were small in terms of correlations, but when CG were fit as a fixed effect the slopes indicated closer unity. On the other hand, WW tended to have higher correlations and slopes trending closer to 1 when CG were fit as a random. Yearling weight had higher correlations when the CG were fit as a fixed effect, but the slopes were closer to 1 when the CG were fit as a random effect. The GxE model had higher accuracy of prediction for the EBV_A only for YW when CG were fit as a random effect and was comparable in accuracy for BW when CG were fit as a fixed effect. Both models tended to have greater EBV_A prediction accuracy when CG were fit as a random effect than a fixed effect (see Additional File 1, Table S1). Due to the decrease in accuracy when the

D_{GxE} was included in the EBV_{Total} and the reduction of correlations when compared back to the national EBVs, the EBV_A and EBV_{Total} should be considered separately for making selection decisions since they are informative on different types of information, solely the additive genetic potential and the potential with deviation due to GxE, respectively.

Discussion

The US beef industry relies on the continuous improvement of genetics, management practices, and favorable environmental conditions to increase yield as cattle are one of the last livestock species that lives and produces outside exposed to the elements. Variation in the phenotype is explained by the combined effects of genetics and the environment, yet, to date, no national beef cattle evaluation in the United States considers the effect of GxE, but rather simply focuses on predicting the average additive genetic merit of animals across all environments. Here, we illustrate that by including the GxE effect in genomic prediction models, future genomic predictions could improve the identification and selection of animals best suited to their environment thus in turn increasing sustainability of the beef industry.

While accounting for the environment in genomic prediction models may seem straightforward, there are many different analytical methods that can be used for including the interaction between genetics and environment. (33) described a method addressing selection with GxE for when only two environments were considered in which a measured phenotype could be considered to be two different characters, the two

environments are considered as treatments, and the GxE effect is formulated in terms of the genetic correlation between the characters. Even though they were defined using several environmental variables, our ecoregions were discrete allowing us to fit bivariate models for each phenotype in each ecoregion. For the numerically largest of the ecoregions that we analyzed, genetic correlations were higher between ecoregions when the CG were fit as a random effect and lower when the CG was fit as a fixed effect which could be due to greater prediction error variance or bias(34). Even with high genetic correlations which would be indicative of more similar environments reducing the potential for the reranking of individuals, we observed low to moderate accuracies for EBV prediction within and across ecoregions. These across environment accuracies were influenced by differences in sample size between the ecoregions with numerically smaller ecoregions predicting EBVs for numerically larger ecoregions with reduced accuracy. However, ecoregions with larger sample sizes could predict ecoregion-specific breeding values for animals from other ecoregions with moderate accuracy. Additionally, we demonstrated that the bivariate models had lower prediction accuracy compared to the currently utilized univariate national model that does not account for GxE interactions. Even though the bivariate models model these interactions and allow for EBV prediction across environments, we do not recommend them for implementation in industry as beef cattle are not uniformly distributed across environments causing low accuracy for predictions and the limitation to two environments requires multiple pairwise models to be analyzed. Furthermore, these models frequently suffered from poor convergence. Unlike plants where genotypes can be replicated across environments, the data structure

of animal genotypes only having observations in one environment made these models difficult for cattle populations.

Exploring other GxE model types, random regression and reaction norm models have been analyzed in beef and dairy cattle to evaluate the impact of a continuous environmental gradient on several traits such as weight, gain, stayability, and reproductive traits(11,12,15–17,35,36). The appeal of these models is the ability to model higher order interactions with the environmental variable allowing changes in the performance or plasticity of a genotype as the environment shifts along the gradient(35,37). Even though these methods have the ability to predict EBVs along the extent of an environmental gradient, there are shortcomings associated with the approach. The first is that the environmental gradient is often limited to a single environmental factor that must be chosen to represent the most important of the environmental effects, when in reality the environment is a complex system of interacting factors that contribute to phenotypic variation. Some of the methods that are currently utilized to overcome this limitation include the use of a climate index, estimating year effects, and estimating CG effects for the environmental gradient(11,12,15–17,35). The second problem arises when estimates of CG effects, usually from a related or indicator trait, are utilized to represent the environmental gradient. Contemporary groups are usually defined using management, year, and farm or herd information which all tie an animal to a specific location and climate at a specific point in time which captures the environment an animal experiences(11,16). The issue arises when CG for the trait of interest are included in the model; while the CG are technically defined for different traits, cattle are usually stationary within an environment and handled as a cohort so the environmental and

management effects that an animal experiences can overlap across measured traits which are then potentially being accounted for twice during genetic value estimation(12,15). The third limitation of these types of models is that while genetic values for individuals are obtained for the entire environmental gradient which can measure plastic response and pinpoint favorable environmental conditions that would maximize the phenotypic potential, the specific impact of GxE on the phenotypic variance is not estimated separately from the other genetic variance components. Since we wanted to quantify the proportion of phenotypic variance due to the GxE effect alone for consideration in selection decisions and include multivariable environmental conditions to better represent the experienced environment, random regression and reaction norm models were not evaluated in this study.

The univariate GxE model utilized in this study allowed the ability to take into account long-term multivariable environmental information through our definition of ecoregions while directly estimating the variance explained from the SNPs and the environmental measures by calculating shared genetic relationships among individuals within shared environments(27,28). Leveraging data for a common beef cattle breed, Gelbvieh, that is utilized in production throughout the United States, the influence of GxE was estimated to account for 3%-12% of the variation in body weight across life stages measured as BW, WW, and YW (Table 3). This variation can have a significant impact on the beef industry when an animal could have a large negative or positive GxE effect depending on the environment in which they live; this interaction should be considered when deciding breeding decisions for future generations. Moreover, we found that when the GxE effect

was included, the accuracy for predicting the additive genetic component increased by ~10% for YW when the CG were fit as a random effect, comparable accuracies for BW when CG were fit as a fixed effect, and ~2% decreases for WW regardless of how CG were fit (Table 4, 6). While expanding the potential for GxE predictions, a drawback of this method is that the estimated GxE component, D_{GxE}, is specific to the environmental conditions in which the animal was raised and there is not a predicted D_{GxE} for other environments. We propose that this D_{GxE} could be used separately from the EBV_A as an additional selection criteria for culling animals with a strongly negative GxE for their current environment. These results indicate that the beef industry should be taking advantage of GxE in their prediction models for production traits; by including GxE, accuracy of predicting the additive genetic component has the potential to increase and the quantified impact of the environment will guide breeders and producers in identifying and selecting animals that will have improved productivity in their environments.

An internal factor to consider is how to include CG in the GxE models as our results reveal a sensitivity in prediction accuracy due to CG being included as either a fixed or random effect. Traditionally, in cattle evaluations CG are treated as a fixed effect since management practices are considered to be nonrandom systematic effects(34,38,39). While the differences among levels of a fixed effect can be estimated, many individuals will be needed within a level to handle the degree of freedom requirements which can be challenging as most CG levels in the beef industry are small partially due to choice reporting of favored animals and that ~90% of beef farms in the US have less than 100 cows(34,40,41). When CG were included as a random effect there is an increase in the

potential to introduce or reduce bias depending on the definition of contemporary group and nonrandom associations, but there is limited impact of small CG sizes(34,38,39). This study fit CG both ways to determine the impact on accuracy as for our data ~80% of the levels of CG had less than 5 individuals. For the GxE model, accuracies tended to be higher for the EBV_A and D_{GxE} when the CG was fit as a random effect (Table 3, 6)(34,38). While adjusting for CG effects attempts to control for systematic effects due to cohorts, farm/herd, and management practices, CG does tie an individual to a specific climate and location. As a proxy for herd identification, we included breeder zip codes in the CG formation which specifically ties those animals to a physical location and environment, not just management practices. A possible reason why there was decrease in the estimated variance and accuracy of EVB_A and D_{GxE} for when CG were fit as a fixed effect could be due to the more aggressive removal of variation which would have been attributed to the D_{GxE}. Additionally, on average, when CG were fit as a fixed effect ~80% of the CG levels per ecoregion were not included due to membership being less than 5 animals which is a sizable loss of variation tied to those ecoregions and the corresponding D_{GxE}. These differences result in a tradeoff between estimation and accuracy and lead us to believe that our GxE effect variance estimates are more representative of a lower (CG random) and upper (CG fixed) bound for the true amount of variation influence on the phenotype due to the environment. The methodology for including a GxE effect in relation to controlling for systematic management effects with CG will need to be explored more in depth as the method of choice may be dependent on the amount and types of data available for study.

As CG definitions can conflate environment and management effects, a method that could be explored in the future of beef cattle genomic prediction under a GxE model will be to further separate the contributing components to a phenotype into a genotype x environment x management (GxExM) effect. Management practices are flexible and quicker in response to short term or infrequent negative environmental effects; while genetic selection can shift a population towards a more favorable response in an environment, this is generation length dependent which may not keep pace with a highly variable climate. (42) identified three points of consideration for future inclusion and application of GxExM which are: there must be positive repeatable GxM contributions that can be tied to specific practices, mechanisms for detection and selection of the GxM benefits need to be included in the species improvement process, and finally, farmers must include these positive GxM practices in their production systems. Simulations for canola yield have shown that through changing plant density, sowing date, or irrigation that these management practices can narrow the yield gap across different regions in China in response to the differences in seasonal rainfall(43). Maize in Ethiopia saw the potential to increase yield by 48% due to a change in plant density for a specific genotype when taking genotype x management effects into account(44). Accounting for and including GxExM has been recognized in plant breeding and production for more than half a century, but currently less than 1% of US farmers utilize prediction methods for management practice decisions even though they have the potential to maximize phenotypes that would otherwise not be observed(45,46). Utilizing GxExM in predictions has been recognized by researchers in the plant industry and implementation of GxExM

in beef cattle could have resounding beneficial effects, yet the adoption could be slow due to socio-economic barriers of limited exposure and education, perception of the technology and benefits, or associated costs(47). Additionally, the industry would need to cooperate as a whole to define, identify, and record different types of management practices that could be accounted for in GxExM. The limitations are a daunting hurdle to overcome, but the potential for improving sustainable intensification in the beef cattle industry through GxE and GxExM is evident.

Conclusions

While bivariate GxE models allow for EBV prediction across environments, their accuracy is highly influenced by sample size often performing worse than the current national model that does not account for GxE effects. Therefore, bivariate GxE models are not recommended for implementation in the US beef cattle industry. We used a method to include a compound GxE effect in genomic prediction that was comparable in accuracy for predicting the direct genetic component of BW, and increased the accuracy of predicting the additive genetic component of YW. Accuracy of the EBV_A for YW rose 11% in the GxE model over the standard national model currently implemented by industry which does not consider GxE effects. We quantified the amount of variance that the GxE effect contributes, ranging from ~3%-12%, to the total phenotypic variation across production life stages. Accounting for CG as a random effect improved prediction accuracy for WW and YW due to management practices being tied to a specific location

and climate. Prediction accuracies for GxE inclusive models could be further improved by expanding the number of environmental variables, refining the definition of CG, and separating the effects of environment and management. In short, for the Gelbvieh breed of beef cattle in the US, our results support the inclusion of GxE effects in genomic prediction for accurate EBV_A predictions for especially for YW and the new supplemental selection information in the form of a D_{GxE} for the current environment for improving sustainable production.

References

1. Godfray HCJ, Garnett T. Food security and sustainable intensification. Philos Trans R Soc Lond B Biol Sci. 2014 Apr 5;369(1639):20120273.

2. Pretty J, Bharucha ZP. Sustainable intensification in agricultural systems. Ann Bot. 2014 Dec;114(8):1571–96.

3. Gamborg C, Sandøe P. Sustainability in farm animal breeding: a review. Livestock Production Science. 2005 Mar 1;92(3):221–31.

4. Godfray HCJ, Beddington JR, Crute IR, Haddad L, Lawrence D, Muir JF, et al. Food security: the challenge of feeding 9 billion people. Science. 2010 Feb 12;327(5967):812–8.

5. Place SE, Mitloehner FM. The nexus of environmental quality and livestock welfare. Annu Rev Anim Biosci. 2014 Feb;2:555–69.

6. Rauw WM, Gomez-Raya L. Genotype by environment interaction and breeding for robustness in livestock. Front Genet. 2015 Oct 20;6:310.

7. Rowan TN, Durbin HJ, Seabury CM, Schnabel RD, Decker JE. Powerful detection of polygenic selection and evidence of environmental adaptation in US beef cattle. PLoS Genet. 2021 Jul;17(7):e1009652.

8. Piccoli ML, Brito LF, Braccini J, Oliveira HR, Cardoso FF, Roso VM, et al. Comparison of genomic prediction methods for evaluation of adaptation and productive efficiency traits in Braford and Hereford cattle. Livest Sci. 2020 Jan 1;231:103864.

9. Mota RR, Tempelman RJ, Lopes PS, Aguilar I, Silva FF, Cardoso FF. Genotype by environment interaction for tick resistance of Hereford and Braford beef cattle using reaction norm models. Genet Sel Evol. 2016 Jan 14;48:3.

10. Carvalheiro R, Costilla R, Neves HHR, Albuquerque LG, Moore S, Hayes BJ. Unraveling genetic sensitivity of beef cattle to environmental variation under tropical conditions. Genet Sel Evol. 2019 Jun 20;51(1):29.

11. Bignardi AB, El Faro L, Pereira RJ, Ayres DR, Machado PF, de Albuquerque LG, et al. Reaction norm model to describe environmental sensitivity across first lactation in dairy cattle under tropical conditions. Trop Anim Health Prod. 2015 Oct;47(7):1405–10.

12. Oliveira DP, Lourenco DAL, Tsuruta S, Misztal I, Santos DJA, de Araújo Neto FR, et al. Reaction norm for yearling weight in beef cattle using single-step genomic

evaluation. J Anim Sci. 2018 Feb 15;96(1):27–34.

13. Kolmodin R, Strandberg E, Madsen P, Jensen J, Jorjani H. Genotype by Environment Interaction in Nordic Dairy Cattle Studied Using Reaction Norms. Acta Agric Scand A Anim Sci. 2002 Jan 1;52(1):11–24.

14. Calus MPL, Veerkamp RF. Estimation of environmental sensitivity of genetic merit for milk production traits using a random regression model. J Dairy Sci. 2003 Nov;86(11):3756–64.

15. Mota LFM, Fernandes GA Jr, Herrera AC, Scalez DCB, Espigolan R, Magalhães AFB, et al. Genomic reaction norm models exploiting genotype × environment interaction on sexual precocity indicator traits in Nellore cattle. Anim Genet. 2020 Mar;51(2):210–23.

16. Mattar M, Silva LOC, Alencar MM, Cardoso FF. Genotype × environment interaction for long-yearling weight in Canchim cattle quantified by reaction norm analysis. J Anim Sci. 2011 Aug;89(8):2349–55.

17. MacNeil MD, Cardoso FF, Hay E. Genotype by environment interaction effects in genetic evaluation of preweaning gain for Line 1 Hereford cattle from Miles City, Montana. J Anim Sci. 2017 Sep;95(9):3833–8.

18. Butts WT, Koger M, Pahnish OF, Burns WC, Warwick EJ. Performance of two lines of Hereford cattle in two environments. J Anim Sci. 1971 Nov;33(5):923–32.

19. Chiaia HLJ, de Lemos MVA, Venturini GC, Aboujaoude C, Berton MP, Feitosa FB,

et al. Genotype × environment interaction for age at first calving, scrotal circumference, and yearling weight in Nellore cattle using reaction norms in multitrait random regression models. J Anim Sci. 2015 Apr;93(4):1503–10.

20. Tiezzi F, Gaddis KP, Clay JS, Maltecca C. Accounting for genotype by environment interaction in genomic predictions for US Holstein dairy cattle. IB [Internet]. 2015 Aug 11 [cited 2022 Jul 1];(49). Available from: https://journal.interbull.org/index.php/ib/article/view/1612

21. Rowan TN, Hoff JL, Crum TE, Taylor JF, Schnabel RD, Decker JE. A multi-breed reference panel and additional rare variants maximize imputation accuracy in cattle. Genet Sel Evol. 2019 Dec 26;51(1):77.

22. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.

23. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015 Feb 25;4:7.

24. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016 Oct;48(10):1284–7.

25. Loh PR, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016 Nov;48(11):1443–8.

26. Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience [Internet]. 2020 Mar 1;9(3). Available from: http://dx.doi.org/10.1093/gigascience/giaa021

27. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011 Jan 7;88(1):76–82.

28. Vinkhuyzen AAE, Wray NR. Novel directions for G × E analysis in psychiatry. Epidemiol Psychiatr Sci. 2015 Feb;24(1):12–9.

29. Misztal I, Tsuruta S, Lourenco D. BLUPF90 family of programs [Internet]. 2014 [cited 2021 Apr 20]. Available from: http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all2.pdf

30. Legarra A, Reverter A. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. Genet Sel Evol. 2018 Nov 6;50(1):53.

31. Macedo FL, Reverter A, Legarra A. Behavior of the Linear Regression method to estimate bias and accuracies with correct and incorrect genetic evaluation models. J Dairy Sci. 2020 Jan;103(1):529–44.

32. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2017. Available from: https://www.R-project.org/

33. Falconer DS. The Problem of Environment and Selection. Am Nat. 1952;86(830):293–8.

34. Visscher PM, Goddard ME. Fixed and Random Contemporary Groups. J Dairy Sci. 1993 May 1;76(5):1444–54.

35. Fennewald DJ, Weaber RL, Lamberson WR. Genotype by environment interaction for stayability of Red Angus in the United States. J Anim Sci. 2018 Mar 6;96(2):422–9.

36. Tiezzi F, de Los Campos G, Parker Gaddis KL, Maltecca C. Genotype by environment (climate) interaction improves genomic prediction for production traits in US Holstein cattle. J Dairy Sci. 2017 Mar;100(3):2042–56.

37. Yang W, Chen C, Steibel JP, Ernst CW, Bates RO, Zhou L, et al. A comparison of alternative random regression and reaction norm models for whole genome predictions. J Anim Sci. 2015 Jun;93(6):2678–92.

38. Ugarte E, Alenda R, Carabaño MJ. Fixed or Random Contemporary Groups in Genetic Evaluations. J Dairy Sci. 1992 Jan 1;75(1):269–78.

39. Van Vleck LD. Contemporary Groups for Genetic Evaluations. J Dairy Sci. 1987 Nov 1;70(11):2456–64.

40. Harrison XA, Donaldson L, Correa-Cano ME, Evans J, Fisher DN, Goodwin CED, et al. A brief introduction to mixed effects modelling and multi-model inference in ecology. PeerJ. 2018 May 23;6:e4794.

41. USDA National Agricultural Statistics Service. 2017 Census of Agriculture [Internet]. Available from: www.nass.usda.gov/AgCensus

42. Cooper M, Voss-Fels KP, Messina CD, Tang T, Hammer GL. Tackling G × E × M interactions to close on-farm yield-gaps: creating novel pathways for crop improvement by predicting contributions of genetics and management to crop productivity. Theor Appl Genet. 2021 Jun;134(6):1625–44.

43. He D, Wang E, Wang J, Lilley JM. Genotype×environment×management interactions of canola across China: A simulation study. Agric For Meteorol. 2017 Dec 15;247:424–33.

44. Seyoum S, Rachaputi R, Fekybelu S, Chauhan Y, Prasanna B. Exploiting genotype x environment x management interactions to enhance maize productivity in Ethiopia. Eur J Agron. 2019 Feb 1;103:165–74.

45. Messina CD, Cooper M, Hammer GL, Berning D, Ciampitti I, Clark R, et al. Two decades of creating drought tolerant maize and underpinning prediction technologies in the US corn-belt: Review and perspectives on the future of crop design [Internet]. bioRxiv. 2020 [cited 2021 Sep 12]. p. 2020.10.29.361337. Available from: https://www.biorxiv.org/content/10.1101/2020.10.29.361337v1.full

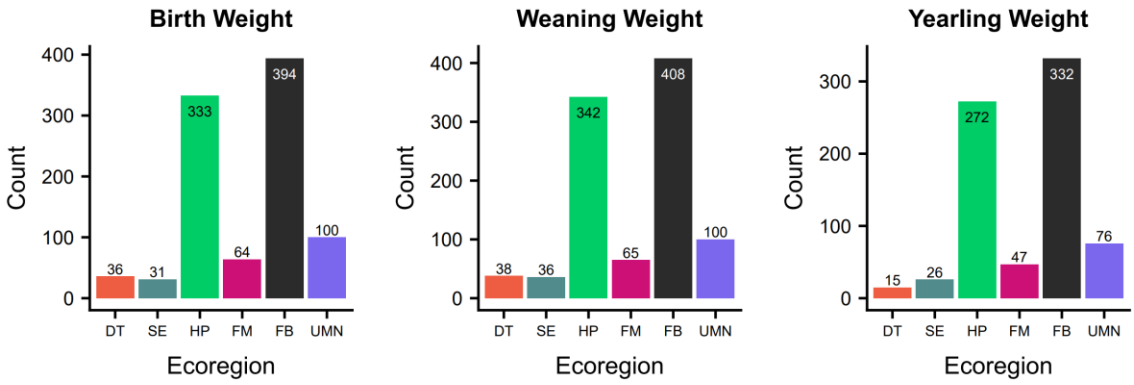
46. Jones JW, Antle JM, Basso B, Boote KJ, Conant RT, Foster I, et al. Brief history of agricultural systems modeling. Agric Syst. 2017 Jul;155:240–54.

47. Beres BL, Hatfield JL, Kirkegaard JA, Eigenbrode SD, Pan WL, Lollato RP, et al. Toward a Better Understanding of Genotype × Environment × Management

Interactions-A Global Wheat Initiative Agronomic Research Strategy. Front Plant Sci. 2020 Jun 16;11:828.

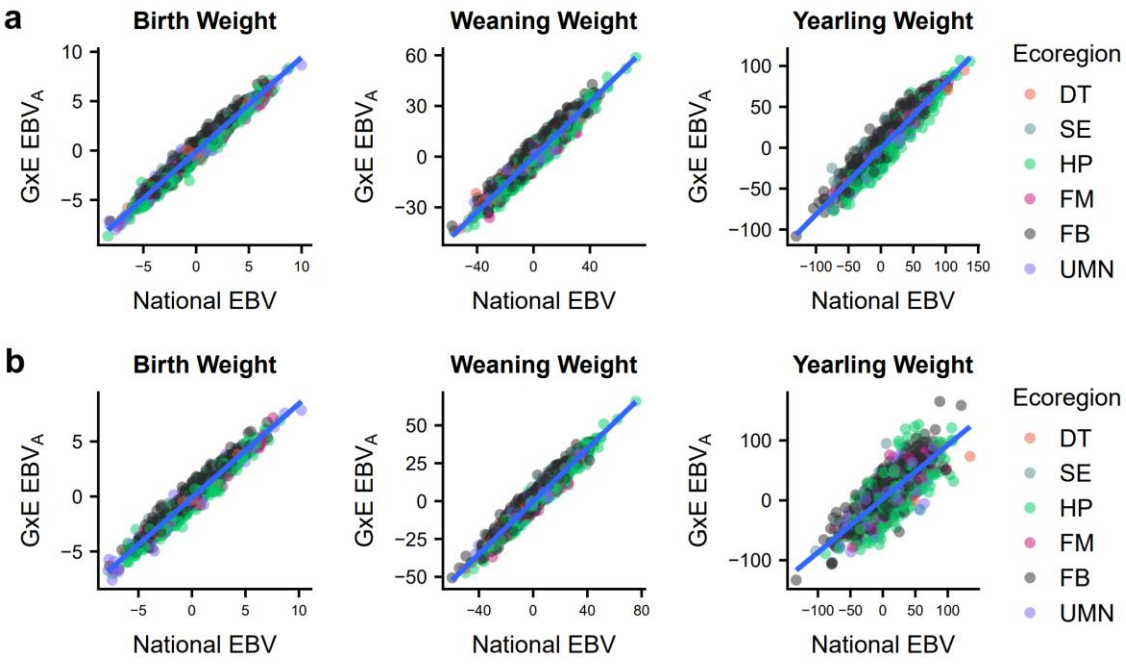
Figures

Figure 1 Ecoregion distribution of individuals in the validations set



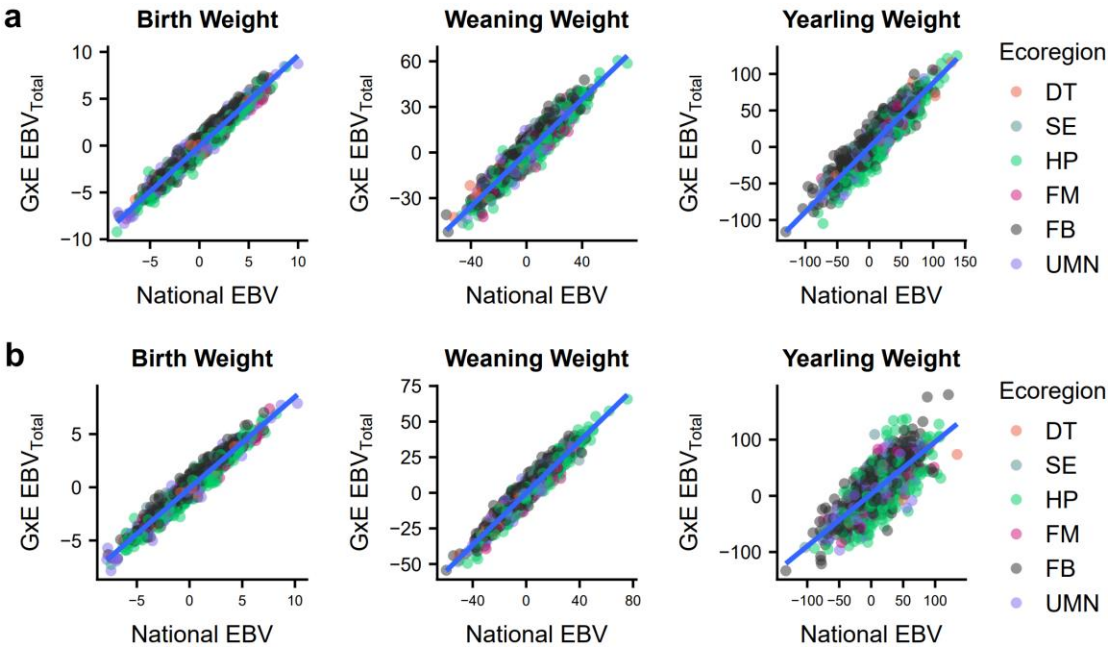
The distribution of individuals across ecoregions in the ecoregion distributed validation set for birth weight (n = 958), weaning weight (n = 989), and yearling weight (n = 768). Ecoregions are designated as the following: DT is Desert, SE is Southeast, HP is High plains, FM is Forested Mountains, FB is Fescue Belt, and UMN is Upper Midwest & Northeast.

Figure 2 National EBVs and GxE EBV$_A$ for the ecoregion distributed validation set



The national estimated breeding values (EBV) and the genotype x environment (GxE) additive estimated breeding values (EBV$_A$) for the ecoregion distributed validation set of individuals when **a,** the contemporary group is fit as a fixed effect and **b,** when the contemporary group is fit as a random effect with lines of best fit in blue. Each dot represents an individual and the color corresponds to the ecoregion the animal originates from. Ecoregions are designated as the following: DT is Desert, SE is Southeast, HP is High plains, FM is Forested Mountains, FB is Fescue Belt, and UMN is Upper Midwest & Northeast.

Figure 3 National EBVs and GxE EBV_{Total} for the ecoregion distributed validation set



The national estimated breeding values (EBV) and the genotype x environment (GxE) total estimated breeding values (EBV_{Total}) for the ecoregion distributed validation set of individuals when **a,** the contemporary group is fit as a fixed effect and **b,** when the contemporary group is fit as a random effect with lines of best fit in blue. Each dot represents an individual and the color corresponds to the ecoregion the animal originates from. Ecoregions are designated as the following: DT is Desert, SE is Southeast, HP is High plains, FM is Forested Mountains, FB is Fescue Belt, and UMN is Upper Midwest & Northeast.

Tables

Table 1 Distribution of individuals with genotypes and phenotypes across ecoregions

Phenotype	DT	SE	HP	FM	FB	UMN	No Eco	No Pheno	Total
BW	362	311	3328	641	3937	999	809	2174	12561
WW	376	356	3420	649	4081	996	806	1877	12561
YW	154	259	2720	474	3316	756	613	4269	12561

There were 12561 individuals with genotypes with up to the three phenotypes, birth weight (BW), weaning weight (WW), and yearling weight (YW), and six ecoregions. Ecoregions are designated as the following: DT is Desert, SE is Southeast, HP is High Plains, FM is Forested Mountains, FB is Fescue Belt, and UMN is Upper Midwest & Northeast. Individuals included in the No Eco (No Ecoregion) column have reported phenotypes but no ecoregion designation, while those in No Pheno (No Phenotype) have no phenotype or ecoregion identified.

Table 2 Additive genetic variance estimates for the bivariate models for birth weight

Bivariate model	Fixed	Random
High Plains and Fescue Belt		
High Plains	38%	37%
Fescue Belt	35%	36%
High Plains and Upper Midwest &Northeast		
High Plains	40%	41%
Upper Midwest and Northeast	47%	46%
Fescue Belt and Upper Midwest &Northeast		
Fescue Belt	36%	38%
Upper Midwest &Northeast	48%	44%

Additive genetic variance estimates expressed as a percentage of phenotypic variance contributed for birth weight when contemporary group was fit as a fixed or random effect. Ecoregions are designated as following: HP is High Plains, FB is Fescue Belt, and UMN is Upper Midwest & Northeast.

Table 3 Variance component estimates for the GxE models and three phenotypes

Phenotype	Fixed	Random
Birth Weight		
Additive	35%	38%
GxE	5%	3%
Weaning Weight		
Additive	18.85%	24.77%
GxE	8.87%	6.16%
Yearling Weight		
Additive	30.28%	39.84%
GxE	11.96%	5.16%

Additive genetic and genotype x environment (GxE) variance component estimates expressed as a percentage of phenotypic variance for: birth weight (BW), weaning weight (WW), and yearling weight (YW) when the contemporary group was fit as a fixed or random effect.

Table 4 Accuracies for the national models

Phenotype	Fixed	Random
Birth Weight	73.88%	73.97%
Weaning Weight	65.11%	68.36%
Yearling Weight	58.26%	62.87%

Accuracies estimated for the national evaluation models for birth weight (BW), weaning weight (WW), and yearling weight (YW) when contemporary group was fit as a fixed or random effect.

Table 5 Accuracies for the three bivariate models for birth weight

Bivariate	Animals Ecoregion	Ecoregion of Prediction	Fixed	Random
HP & FB	HP	HP	65.34%	64.52%
	HP	FB	53.16%	53.90%
	FB	HP	57.39%	59.30%
	FB	FB	59.49%	60.90%
HP & UMN	HP	HP	6436%	62.46%
	HP	UMN	68.12%	68.18%
	UMN	HP	41.67%	40.58%
	UMN	UMN	63.48%	64.48%
FB & UMN	FB	FB	59.04%	60.12%
	FB	UMN	67.64%	72.56%
	UMN	FB	46.78%	46.96%
	UMN	UMN	62.65%	64.76%

The accuracies calculated for when the contemporary group was fit as a fixed or random effect for the three bivariate models for birth weight for the ecoregion distributed validation set for: High Plains (HP) & Fescue Belt (FB), HP & Upper Midwest & Northeast (UMN), and FB & UMN. Animal's Ecoregion is defined as the ecoregion which has the adjusted phenotype that the individual resides in while Ecoregion of Prediction is the ecoregion the EBV is predicted for.

Table 6 Accuracies for the GxE models

Phenotype	Fixed	Random
Birth Weight		
EBV_A	73.31%	66.27%
D_{GxE}	16.89%	11.728%
EBV_{Total}	69.66%	64.41%
Weaning Weight		
EBV_A	63.11%	66.42%
D_{GxE}	20.45%	19.15%
EBV_{Total}	57.00%	62.55%
Yearling Weight		
EBV_A	55.65%	73.51%
D_{GxE}	22.68%	28.64%
EBV_{Total}	51.99%	74.78%

Accuracies for the genotype x environment models (GxE) when contemporary group was fit as a fixed or random effect for birth weight, weaning weight, and yearling weight. Accuracies were calculated for the additive (EBV_A), GxE deviation (D_{GxE}), and the total genetic merit (EBV_{Total}).

Additional files

Additional file 1 Table S1

Format: pdf

Title: Comparison statistics between GxE and national models' EBVs with the ecoregion distributed validation set

Description: The calculated statistics, slope of the line of best fit, Pearson correlation, and Spearman correlation, for comparing the genotype x environment (GxE) additive estimated breeding values (EBV_A) and total estimated breeding values (EBV_{Total}) to the national EBVs with the ecoregion distributed validation set when the contemporary group is fit as a fixed or random effect for birth weight, weaning weight, and yearling weight.

Available upon request by Dr. Jared Decker at deckerje@missouri.edu

CHAPTER 3

Direct and maternal genotype-by-environment effects in US Red Angus genomic predictions for growth traits

Sara Nilson[1], Troy Rowan[1,2,3,4], Robert Schnabel[1,5], Jared Decker[1,5]*

[1]Division of Animal Sciences, University of Missouri, Columbia, MO, 65211, USA

[2]Genetics Area Program, University of Missouri, Columbia, MO, 65211, USA

[3]Department of Animal Science, University of Tennessee, Knoxville, TN, 37996, USA

[4]College of Veterinary Medicine, Large Animal Clinical Science, University of Tennessee, Knoxville, TN, 37996, USA

[5]Institute for Data Science and Informatics, University of Missouri, Columbia, MO, 65211, USA

Abstract

Background

Climate changes pressure the US beef cattle production to increase sustainability as environmental effects are beginning to shift. Frequently, cattle re-rank in genetic potential across environments due to genotype-by-environment interactions and will do so more drastically as unsteady climate changes become the new normal. As the environment a dam creates for her progeny is affected by her own environment, maternal genotype-by-environmental effects need to be considered in addition to direct genotype-by-environment effects on the growth and production of cattle. By estimating the

variance these effects contribute to the birth weight of an animal and incorporating them into genomic prediction and selection practices, breeders and producers will be able to achieve sustainability goals even as the environment varies.

Results

Maternal and direct genotype-by-environment variance components were estimated for birth weight using a genotype-by-environment model. With the full single nucleotide polymorphism set, genotype-by-environment models were close in predictive accuracy to the currently implemented national genomic evaluation model for the additive and maternal genetic effects. When the genotype-by-environment model was combined with a selection and environment associated enriched single nucleotide polymorphism sets, the estimated direct genotype-by-environment effect contribution to birth weight increases from ~3% to ~42%. Similarly, the estimated maternal genotype-by-environment effect estimate increased from 0.14% to 1.95% as compared to the full SNP set. With the focused single nucleotide polymorphism set being utilized for the genotype-by-environment model, the national model has higher prediction accuracy for the additive genetic effect while the genotype-by-environment model tends to have higher accuracy for predicting the maternal genetic effect.

Conclusions

Genotype-by-environment models can partition the phenotypic variance into maternal and direct genotype-by-environment effects for birth weight. While the estimated contribution of these effects can be large, the tradeoff of a lower prediction accuracy, 3%-10% difference, for the additive genetic effect needs to be considered. On the other hand, the prediction accuracy for the maternal genetic effect was slightly increased, 1%-3%,

with a genotype-by-environment model as compared to the current national genomic evaluation model. Genotype-by-environment models are currently not suggested for implementation in the US beef cattle industry due to decreased prediction accuracy and as environmental effects could potentially be countered through improved management practices.

Background

Genotype-by-environment (GxE) interactions have been widely acknowledged as a source of variation affecting the phenotypes of cattle which have the potential to improve upon current genomic prediction and selection. Models accounting for GxE have been analyzed mainly for growth traits, reproductive traits, and milk production traits (1–6). GxE can be modeled across many environments with multiple methods often resulting in varying effect estimates, accuracies, and recommendations. One methodological difference is the definition of environment; some define environment as a function of contemporary groups (6–9), herd-year averages (10), climate factors (5,11–13), or a mixture of environment and management classifications (2–4) to list a few. Due to these differences in definition, modeling, and predictive accuracy fluctuations, no GxE models have been officially adopted by any national cattle genomic evaluations to date.

Even though GxE interactions have been included in genomic prediction models before, they have not previously been partitioned out into maternal and direct GxE effects for prediction even though the influences of fetal genetics and uterine environment is widely acknowledged (14–18). Maternal and direct GxE effects are thought to fluctuate in the

amount they contribute to the phenotype across life stages. For example, maternal GxE effects would potentially have a larger contribution earlier in life affecting birth and weaning weights while direct GxE effects would have a larger effect later in life affecting yearling weights. By separately estimating maternal and direct GxE effects, management practices could be adjusted to provide additional support, e.g. nutritional supplementation during and post pregnancy affecting pre and postnatal growth. A few studies have examined the effects of heat stress relating to a dam's reproductive traits and milk yield which in turn have an effect on their calves, but these studies are usually limited to phenotypes measured on the dams instead of on both dam and progeny (14,19–22). Incorporating additional information in the selection of genomic information for predictions holds promise by reducing noise in estimations and enriching for pertinent information related to the desired trait of interest. Biological knowledge through gene ontology, association studies, quantitative trait loci, and other types of omic data have been explored in the goal of SNP or genomic region inclusion for genomic prediction in cattle (23–27). These methods have been applied to many traits including but not limited to milk yield traits, reproductive traits, disease risk, and profit indexes (23–27). For some of the studies, a focused SNP set was able to improve prediction accuracy, reliabilities, and able to explain more of the variation in the traits of interest (23–26). To our knowledge this enrichment method through preselected genomic data has not been applied to the estimation and prediction of GxE effects in cattle.

Here we explore GxE models that partition GxE effects into direct and maternal components to estimate the separate contributions to the variance of birth weight (BW). By applying these models to the Red Angus breed in the US, GxE effects will be captured

across many different types of environments as the breed is widely distributed. Additionally we will test the effect of enriched SNP subsets on the additive and GxE effect estimation and prediction accuracy. By employing two different SNP subsets, a selection focused and an environment focused, we hope to increase the orthogonality of our variance component estimates.

Methods

Data

Phenotype and genotype data were provided by the Red Angus Association of America. The phenotype analyzed was birth weight which had been pre-adjusted for age of dam. Contemporary groups were formed to breed organization specification with groups numbering less than five animals removed. Post filtering, the dataset consisted of 19,739 individuals with phenotypes and genotypes and 3,284 additional dam and sire genotypes. As previously described by Rowan *et al.* (2021), nine discrete ecoregions were utilized to capture the complex environments across the United States. Animals were assigned to an ecoregion based on breeder zip code; those with multiple ecoregion assignments had their ecoregion set to missing.

Genotypes and Imputation

Genotyped loci varied according to the utilized assays including the GeneSeek GGP-LDv3, GeneSeek GGP-LDv4, GeneSeek GGP-90KT, GeneSeek GGP-HDv3, GeneSeek GGP-F250, Illumina BovineSNP50, and Illumina BovineHD. Genotypes were imputed to a union set of ~850k autosomal SNPs with coordinates from the ARS-UCD1.2 bovine reference genome (28) following the method described by Rowan *et al.* (2019). The

process includes: genotype quality control performed in PLINK (v1.9) , referenced-based phasing with Eagle (v2.4), and imputation with Minimac3 (v2.0.1)(30–33). After filtering for minor allele frequency greater than 0.01 in PLINK (v1.9), there were 649,114 SNPs available for analysis. Additional SNP subsets were identified by rank of significance in previous Generation Proxy Selection Mapping (GPSM) and environmental Genome-Wide Association Studies (envGWAS) analyzes in the Red Angus and Simmental breeds as their genetic evaluations are jointly analyzed in industry (12,34). Briefly, GPSM identifies SNPs under polygenic selection while envGWAS captures SNPs associated with local adaptation. The top 20,000 and 40,000 SNPs were selected from each analysis corresponding to $4N_eL$ estimates for when $N_e \approx 150$ or $N_e \approx 350$ and $L = 30$ (35,36). Top GPSM associated SNPs were used to construct additive genomic relationship matrices and top envGWAS SNPs were used to construct the GxE relationship matrices. Overall, three SNP sets were analyzed: the full ~650k, 40k, and 20k.

Variance Component Estimation

Birth weight variance components were estimated with a national model with the full SNP set and a GxE model with the three SNP sets. Variance components for the national model were estimated by average-information restricted maximum likelihood in airemlf90 with the following univariate linear mixed model:

$$y = X_C b_C + Z_A u_A + Z_M u_M + e$$

where y is a vector of phenotypes; b_C is a vector of contemporary group effects; u_A is a vector of random additive genetic effects and u_M is a vector of random maternal genetic effects where $var\begin{bmatrix} u_A \\ u_M \end{bmatrix} = \begin{bmatrix} \boldsymbol{H}\sigma_A^2 & \boldsymbol{H}cov_{A,M} \\ \boldsymbol{H}cov_{M,A} & \boldsymbol{H}\sigma_M^2 \end{bmatrix}$; $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$ is a vector of random

residuals; X_C, Z_A and Z_M are incidence matrices relating y to b_C , u_A , and u_M , respectively. ***I*** is an identity matrix. ***H*** is the blended relationship matrix of the VanRaden genomic relationship matrix ***G*** and the subset of the numerator relationship matrix $\boldsymbol{A_{22}}$ (37–39). Variance components for the GxE model were estimated by average-information restricted maximum likelihood in airemlf90 with the following univariate linear mixed model:

$$y = X_C b_C + X_E b_E + Z_A u_A + Z_M u_M + Z_{DGxE} u_{DGxE} + Z_{MGxE} u_{MGxE} + e$$

where y is a vector of phenotypes; b_C is a vector of contemporary group effects; b_E is a vector of ecoregion environment effects; u_A is a vector of random additive genetic effects and u_M is a vector of random maternal genetic effects where $var\begin{bmatrix} u_A \\ u_M \end{bmatrix} =$

$\begin{bmatrix} \boldsymbol{H}\sigma_A^2 & \boldsymbol{H}cov_{A,M} \\ \boldsymbol{H}cov_{M,A} & \boldsymbol{H}\sigma_M^2 \end{bmatrix}$;

; u_{DGxE} ~ N(0,$\boldsymbol{H_{GxE}}\sigma_{\mathrm{DGxE}}^2$) is a vector of random direct GxE effects; u_{MGxE} ~ N(0,$\boldsymbol{H_{GxE}}\sigma_{\mathrm{MGxE}}^2$) is a vector of random maternal GxE effects; $\boldsymbol{e}$ ~ N(0, $\boldsymbol{I}\sigma_e^2$) is a vector of random residuals; X_C , X_E , Z_A , Z_M , Z_{DGxE} , and Z_{MGxE} are incidence matrices relating y to b_C , b_E , u_A , u_M , u_{DGxE} , u_{MGxE} , respectively. ***I*** is an identity matrix. ***H*** is the blended relationship matrix of the VanRaden genomic relationship matrix ***G*** and the subset of the numerator relationship matrix $\boldsymbol{A_{22}}$ (37–39). $\boldsymbol{H_{GxE}}$ is $\boldsymbol{H_{[i,j]}}$ when individuals are from the same ecoregion and otherwise is 0 (40). The estimated variance components from these models were utilized in downstream predictions for calculating individuals estimated breeding values (EBVs).

Validation and Accuracy

To validate the genomic prediction models the youngest year of individuals was identified, 2019, and grouped by their contemporary group level. Contemporary groups spanning the calendar year were removed. The remaining contemporary groups (n of animals = 3699, 18.7% of N) were randomly assigned to one of three validation sets, with the resulting number of individuals per validation set: n_1 = 1,213 (6%), n_2 = 1,109 (5.6%), and n_3 = 1,377 (7%).

After calculating EBVs with the entire dataset (whole), the three validation sets of individuals had their phenotypes set to missing iteratively (partial) in order to calculate accuracy of the EBVs. Prediction accuracy was estimated with the following:

$$\widehat{acc}_{LR} = \sqrt{\frac{cov(\hat{u}_w, \hat{u}_p)}{(1-\bar{F})\hat{\sigma}_u^2}}$$

where $\widehat{u_w}$ is the vector of EBVs of the validation set of individuals for the whole dataset, $\widehat{u_p}$ is the vector of predicted EBVs of the validation set of individuals for the partial dataset, $\bar{F}$ is the average inbreeding coefficient of the validation individuals, and $\widehat{\sigma_u^2}$ is the corresponding estimated variance of the effect for the trait (41,42). The inbreeding coefficients were calculated in PLINK(v1.9) with the '--ibc' command after the full set of SNPs (649,114) were pruned for linkage disequilibrium with '--indep-pairwise', a window size of 2,000kb, a set size of 1,000 SNPs, and a r^2 = 0.8 resulting in 200,599 SNPs utilized in estimation (30,31). Estimated breeding values were compared across the partial and full datasets by plotting the validation sets' EBVs against each other in R(v3.4.3) with the slope of the line of best fit being estimated with the stats package 'lm()' function to estimate over or under dispersion (43).

Model Comparison

Estimated breeding values were compared across the models by plotting the validation sets' EBVs against each other in R(v3.4.3) with the slope of the line of best fit being estimated with the stats package 'lm()' function to visualize differences between the models (43). Pearson and Spearman correlations were calculated among the validations' EBVs to measure the linear correlation and reranking of individuals among models.

Results

Variance Component Estimation

Variance components were estimated for birth weight for the national model and GxE model with the full SNP set, and for the GxE model with the 40k and 20k SNP set (Table 1).There are marginal differences in estimated variances for the shared components between the national and GxE model when the full SNP set is utilized. As the SNP sets are enriched for SNPs with evidence of directional selection or local adaptation for the GxE model, the direct and maternal additive genetic variance estimates decrease while the direct and maternal GxE variances estimates increase considerably. Similarly, the narrow-sense heritabilities of the direct and maternal genetic effects are close between the national and GxE models when the full SNP set is utilized, but when the SNPs are reduced to the targeted sets to those associated with directional selection the estimates decrease (Table 2). The PVE for the maternal genotype-by-environment effect slightly increases as the SNPs are reduced to those associated with environmental selection, while the PVE for the direct genotype-by-environment effect greatly increases when the SNPs

are reduced. In addition, the GxE models estimated ecoregions fixed effect solutions vary little across the SNP sets while consistent in direction (Table 3).

Model Accuracies

The accuracies of each models predicted EBVs were calculated as $\widehat{acc}_{LR}$ which examines the relationship between the validation individuals' EBVs from when the whole dataset was utilized and when their phenotypes were set to missing (partial) (41,42). The accuracy of the additive genetic EBVs decreased from the national to the GxE models and as the SNP set was reduced in size (Table 4). There was not a statistically significant difference from the national or GxE models with the full SNP data (t = 2.29, p-value = 0.09). The national model was more accurate than the 40k model (t = 3.95, p-value = 0.02) and the 20k model (t = 4.73, p-value = 0.01). Accuracies of the maternal genetic EBVs were very similar between the national and the full SNP GxE model, but the GxE model accurcies increased as the SNP set decreased in size (Table 4). The accuracies for the 20k SNP GxE model were significantly larger than the accuracies for the national model (t = -3.67, p-value = 0.02).

Estimated Breeding Value Comparisons

Within the models, whole and partial EBVs for the validation sets were plotted against each other to visualize dispersion in addition to calculating pearson and spearman correlations. For the national model, the maternal genetic effect has slightly lower correlations with a bit more reranking of individuals as compared to the direct genetic effect (Table 5; Figure 1). Slopes follow a similar trend with the direct genetic effects' slope being closer to 1 at 0.996 and the maternal genetic effect's average slope is slightly

decreased, though not significantly different from 1 (t = -1.24, p-value = 0.34), representing a slightly more dispersed partial EBV. The GxE model with the full set of SNPs maternal genetic EBVs are highly correlated with minimal reranking of individuals, but the direct genetic EBVs have lower correlations in comparison (Table 6; Figure 2). Despite these lower correlations, the average slopes for both additive genetic effects are very close to 1, 1.003 for direct and 0.996 for maternal. When the SNP set is dropped to 40k and 20k for the GxE model, the correlations follow the same trend as the full SNP set with maternal genetic effects having higher correlations than the direct genetic effects with the average slopes both being close to 1(Tabel 7, 8; Figure 3,4).

To compare and contrast the national model and the GxE models, predicted partial EBVs for the validation sets were plotted against each other to estimate the slopes of the lines of best fit, and Pearson and Spearman correlations were calculated. When the national model is compared to each of the GxE models, the predicted EBV_A are moderate to highly correlated with some reranking with both correlation statistics decreasing as the SNP set is reduced in size for the GxE models (Table 9; Figure 5a; Figure 6a; Figure 7a). However, the slopes on average are estimated close to or are 1 indicating minimal prediction bias between the predicted EBV_A (Tabel 9; Figure 5a; Figure 6a; Figure 7a). The EBV_M comparisons follow the same trends as the EBV_A; the correlation statistics decrease as the SNP set decreases for the GxE model (Table 10; Figure 5b; Figure 6b; Figure 7b). Additionally, the slopes follow this decreasing trend as the SNP set decreases for the GxE model which indicates a higher EBV_M estimate for the GxE models than the national model (Table 10; Figure 5b; Figure 6b; Figure 7b).

Discussion

The results presented here demonstrate the ability of a GxE model to partition environmental interactions into maternal and direct components while being comparable in prediction accuracy for the direct and maternal genetic effects to a model which does not account for GxE. A selection and environmental associated targeted SNP set did not improve the prediction accuracy of the direct genetic effect, but did slightly raise the accuracy of the maternal genetic effect. These results highlight several areas of discussion: direct and maternal GxE effects, and SNP selection.

While GxE interactions and quantitative trait loci (QTL) have been previously identified for birth weight in cattle (11–13,15,16,21,44), more work needs to be done as there is not always a clear separation between direct GxE effects of the fetus' genetics responding to the environment and the maternal GxE effects (13,14,16,45–47). First with the GxE models presented here, the estimated effect on birth weight from direct GxE effects was 3% with the full SNP set and ~42% with an environmental associated enriched SNP set (Table 2). Total GxE effect estimates on birth weight in cattle have previously been estimated at 10% (13) and 3%-5% (Nilson et al. 2022), and as GxE effects have not previously been quantified separately as direct and maternal to the authors' knowledge, these direct GxE estimates are surprisingly high. These suggest for BW at least, the fetal genetics interacting with the environment plays a larger role in phenotypic variance than previously estimated. With the increased estimate of variance due to direct GxE, from ~3%-42% as the SNP sets are reduced in size, the heritability and prediction accuracy of

the direct genetic EBV decreases (Tables 2, 4). These coincide with an increase in the standard errors for the direct GxE variance estimates and an increase in the range of estimated direct GxE effects for animals. The direct GxE variance estimates potentially could have been inflated due to the repeated use of the same environmental variables; they were utilized in envGWAS to detect SNPs interacting with the environment, to create the ecoregions, and then the ecoregions were utilized for the environment in our models (12). In relation to the ecoregions, animals in the High Plains (HP) had the largest average BWs and animals in the Arid Prairie (AP) had the lightest average BWs (Table 3). There was a range in the effects of other regions, but most of these ecoregion's confidence intervals overlapped zero. This suggests that temperature, amount of rain, and thus amount of forage likely has the largest impact on cow and calf performance. Maternal effects in cattle have been proved to have large impacts on the life of their progeny due to their additive genetics contribution and environment. When the calf is developing *in utero*, the environment a dam creates for her offspring is in part a function of her own environment, this results in maternal GxE effects affecting her progeny.Genotype-by-environment variance contribution to birth weight estimates are rarely reported in cattle, one being estimated at 10% (13); this is due to many studies identifying if GxE interactions exist through genetic correlations (11,16,21,44,47). The GxE model presented here estimated the variance component of maternal GxE (0.14%-1.95%) effects on birth weight (Table 2). Maternal GxE effects have been observed due to differences in nutrition (15–17), heat stress (17,19,20,48), and reproductive traits (10,14,49). Even though the maternal and direct GxE estimates presented here are

in contrast with thoughts that the uterine environment may play a larger role than fetal genetics on birth weight (14), they are supported by others who suggest that fetal genotype determines the max potential for growth which is then limited by the maternal environment (17,49). These separate estimates give insight into the GxE effects on birth weight as the maternal GxE contribution is low which could indicate a rather plastic uterine environment in response to external environmental pressures while the main phenotypic differences are due to how the fetus' genetics are responding to the external environment or direct GxE effects. Maternal GxE effects may have been estimated to have a smaller impact on birth weight here, but they hold great value in genomic prediction and selection as an optimal maternal environment will allow for a calf to maximize upon its' additive genetic potential regardless of external environmental pressures. Additionally, this maternal GxE effect may have a large impact on a trait with stronger maternal influence like weaning weight.

Lastly, the number of SNPs utilized for the GxE models were reduced to a total of 40k or 20k previously identified significantly associated SNPs with selection to estimate the additive direct and maternal genetic effects or SNPs significantly associated with the environment to estimate the direct and maternal GxE effects (12). Our intention was to reduce noise caused by the large amount of estimation for SNPs potentially not associated with the trait and enrich the amount of relevant information pertinent to the traits analyzed (50–52), while not double counting birth weight phenotype data. Incorporation of biological information into genomic prediction is not new whether this be through gene ontology, annotation, identified SNPs that explain a percentage of the phenotypic variance of a trait, marker weighting, etc. (23,24,50–56). These methods hold

promise as targeted SNP sets have been evaluated to hold comparable or improved predictive ability to their denser counterparts (23,24,26,57), but accuracy is not always increased which could be due to the exclusion of regulatory elements, polygenic effects, and genetic architecture (25,52–54). Our selection enriched SNP set for the GxE model decreased predictive accuracy of the additive direct genetic effect by 5% for the 40k, and 9% for the 20k as compared to the full SNP set, but increased the accuracy of the maternal genetic effect by 1% and 3% for the 40k and 20k sets respectively (Table 4). The increase in the maternal genetic effect accuracy may be capturing the high quality of maternal characteristics that the Red Angus breed is known for and speaks to the findings of Rowan *et al.* (2021) where selection on fertility related traits were enriched in the Red Angus breed. When compared to the full SNP set, the environmental associated enriched SNP sets increased the estimated PVE by the direct GxE effect from ~3% to 42% and the maternal GxE effect from 0.14% to 1.95% (Table 2). While these GxE variation findings are exciting, a major limitation of this study is how the GxE effects were fit in our model. By setting discrete contrasts among the defined ecoregions which can encompass shared or similar environmental features, we are not capturing the shared GxE effects across ecoregions which may exist due to widespread gene flow through the adoption of artificial insemination (12,58). When fitting multiple genetic effects, orthogonal estimates of variances is a concern (59). Using different SNPs to calculate the additive relationship matrix and the GxE relationship matrix may have increased the orthogonality of the variance components and led to vastly different direct GxE variance estimates between the 40k and 20k SNP sets and the full SNP data. Additionally by changing the method of

parameterization of the GxE effects, the definition of environment, and how they are fit in a model, these variance component estimates may change in magnitude and not reflect the true genetic architecture underlying BW (52,60). In short for the results presented here utilizing two targeted SNP sets for a GxE model, a selection enriched set for the direct and maternal genetic effects and an environmental associated enriched set for the direct and maternal GxE effects, decreased the accuracy of direct genetic prediction and slightly increased the accuracy of maternal genetic prediction. While not conveying a clear predictive advantage overall, a focused SNP set needs further investigation in the application towards estimating and predicting effects in a GxE framework.

Conclusions

At this point in time no US beef cattle genomic evaluations account for GxE effects, instead they simply estimate additive genetic effects across all environments. Here we evaluated a GxE model that partitioned GxE effects into maternal and direct as these were hypothesized to contribute in different magnitudes across life stages and could have implications for selection and management practices. Additionally by enriching our SNP sets for selection and environmental association, we narrowed our genotypes to the top 40k or 20k significant SNPs previously identified by GPSM and envGWAS analyses (12). For BW, the evaluated national model accurately predicted the direct genetic effect 3%-10% higher than the GxE models, and was comparable in accuracy for the maternal genetic effect (Table 4). Additionally, there was minimal reranking of individuals between the national and GxE model when the full SNP set was utilized with spearman

correlations averaging 0.945 for the direct genetic EBV and 0.94 for the maternal genetic EBV (Tables 9, 10). While the direct GxE effect estimates were surprisingly high, ~41%, when an enriched SNP set was utilized, the trade off in predictive accuracy was not appealing to achieve a GxE variance component estimation. Overall, there is no clear advantage the GxE model or the focused SNP set has in comparison to the currently utilized national evaluation model for BW. Therefore at this time GxE models and reduced SNP sets are not recommended for implementation and further research will be needed to improve upon them in terms of definition, parameterization, before being considered for prediction.

References

1. Kolmodin R, Strandberg E, Madsen P, Jensen J, Jorjani H. Genotype by Environment Interaction in Nordic Dairy Cattle Studied Using Reaction Norms. Acta Agric Scand A Anim Sci. 2002 Jan 1;52(1):11–24.

2. Sartori C, Tiezzi F, Guzzo N, Mancin E, Tuliozi B, Mantovani R. Genotype by Environment Interaction and Selection Response for Milk Yield Traits and Conformation in a Local Cattle Breed Using a Reaction Norm Approach. Animals (Basel) [Internet]. 2022 Mar 26;12(7). Available from: http://dx.doi.org/10.3390/ani12070839

3. Tiezzi F, Gaddis KP, Clay JS, Maltecca C. Accounting for genotype by environment interaction in genomic predictions for US Holstein dairy cattle. IB [Internet]. 2015 Aug 11 [cited 2022 Jul 1];(49). Available from: https://journal.interbull.org/index.php/ib/article/view/1612

4. Schmid M, Imort-Just A, Emmerling R, Fuerst C, Hamann H, Bennewitz J. Genotype-by-environment interactions at the trait level and total merit index level for milk production and functional traits in Brown Swiss cattle. Animal. 2021 Jan;15(1):100052.

5. Santana ML Jr, Pedrosa VB, Groeneveld E, Ferraz JBS, Cardoso FF, Eler JP. Genotype x environment interaction for growth and reproduction traits of composite beef cattle in Brazil. In: CD-ROM Communication 0162 in Proceedings of the 9th World Congress, Leipzig [Internet]. researchgate.net; 2010. Available from: https://www.researchgate.net/profile/Mario_Santana4/publication/265741003_Genotype_X_Environment_Interaction_For_Growth_And_Reproduction_Traits_Of_Composite_Beef_Cattle_In_Brazil/links/559bd72008aee2c16df02475.pdf

6. Chiaia HLJ, de Lemos MVA, Venturini GC, Aboujaoude C, Berton MP, Feitosa FB, et al. Genotype × environment interaction for age at first calving, scrotal circumference, and yearling weight in Nellore cattle using reaction norms in multitrait random regression models. J Anim Sci. 2015 Apr;93(4):1503–10.

7. Oliveira MM, Santana ML, Cardoso FF. Multiple-breed reaction norm animal model accounting for robustness and heteroskedastic in a Nelore-Angus crossed population. Animal. 2016 Jul;10(7):1093–100.

8. Mattar M, Silva LOC, Alencar MM, Cardoso FF. Genotype × environment interaction for long-yearling weight in Canchim cattle quantified by reaction norm analysis. J Anim Sci. 2011 Aug;89(8):2349–55.

9. Mota LFM, Fernandes GA Jr, Herrera AC, Scalez DCB, Espigolan R, Magalhães AFB, et al. Genomic reaction norm models exploiting genotype × environment interaction on sexual precocity indicator traits in Nellore cattle. Anim Genet. 2020 Mar;51(2):210–23.

10. Strandberg, Kolmodin, Madsen. Genotype by environment interaction in Nordic dairy cattle studied by use of reaction norms. Interbull [Internet]. 2000; Available from: https://journal.interbull.org/index.php/ib/article/download/823/823

11. Fennewald DJ, Weaber RL, Lamberson WR. Genotype by environment interactions for growth in Red Angus. J Anim Sci. 2017 Feb;95(2):538–44.

12. Rowan TN, Durbin HJ, Seabury CM, Schnabel RD, Decker JE. Powerful detection of polygenic selection and evidence of environmental adaptation in US beef cattle. PLoS Genet. 2021 Jul;17(7):e1009652.

13. Braz CU, Rowan TN, Schnabel RD, Decker JE. Genome-wide association analyses identify genotype-by-environment interactions of growth traits in Simmental cattle. Sci Rep. 2021 Jun 25;11(1):13335.

14. Sharma RK, Blair HT, Jenkinson CMC, Kenyon PR, Cockrem JF, Parkinson TJ. Uterine environment as a regulator of birth weight and body dimensions of newborn lambs. J Anim Sci. 2012 Apr;90(4):1338–48.

15. González-Recio O, Ugarte E, Bach A. Trans-generational effect of maternal lactation during pregnancy: a Holstein cow model. PLoS One. 2012 Dec 20;7(12):e51816.

16. Hay EH, Roberts A. Genomic evaluation of genotype by prenatal nutritional environment interaction for maternal traits in a composite beef cattle breed. Livest Sci. 2019 Nov 1;229:118–25.

17. Ferrell CL. Factors Influencing Fetal Growth and Birth Weight in Cattle. https://digitalcommons.unl.edu › hruskareportshttps://digitalcommons.unl.edu › hruskareports [Internet]. 1993 [cited 2022 Jun 30]; Available from: https://digitalcommons.unl.edu/hruskareports/132/

18. Garrett JE, Geisert RD, Zavy MT, Morgan GL. Evidence for maternal regulation of early conceptus growth and development in beef cattle. J Reprod Fertil. 1988 Nov;84(2):437–46.

19. Halli K, Vanvanhossou SF, Bohlouli M, König S, Yin T. Identification of candidate genes on the basis of SNP by time-lagged heat stress interactions for milk production traits in German Holstein cattle. PLoS One. 2021 Oct 14;16(10):e0258216.

20. Collier RJ, Dahl GE, VanBaale MJ. Major advances associated with environmental effects on dairy cattle. J Dairy Sci. 2006 Apr;89(4):1244–53.

21. Fisher LJ, Williams CJ. Effect of Environmental Factors and Fetal and Maternal Genotype on Gestation Length and Birth Weight of Holstein Calves1. J Dairy Sci. 1978 Oct 1;61(10):1462–7.

22. Yin T, König S. Genetic parameters for body weight from birth to calving and associations between weights with test-day, health, and female fertility traits. J Dairy Sci. 2018 Mar 1;101(3):2158–70.

23. Moser G, Khatkar MS, Hayes BJ, Raadsma HW. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet Sel Evol. 2010 Oct 16;42:37.

24. Weigel KA, de los Campos G, González-Recio O, Naya H, Wu XL, Long N, et al. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J Dairy Sci. 2009 Oct;92(10):5248–57.

25. Veerkamp RF, Bouwman AC, Schrooten C, Calus MPL. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein–Friesian cattle. Genet Sel Evol. 2016 Dec 1;48(1):95.

26. VanRaden PM, Tooker ME, O'Connell JR, Cole JB, Bickhart DM. Selecting sequence variants to improve genomic predictions for dairy cattle. Genet Sel Evol. 2017 Mar 7;49(1):32.

27. Hoff JL, Decker JE, Schnabel RD, Seabury CM, Neibergs HL, Taylor JF. QTL-mapping and genomic prediction for bovine respiratory disease in U.S. Holsteins using sequence imputation and feature selection. BMC Genomics. 2019 Jul 5;20(1):555.

28. Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience [Internet]. 2020 Mar 1;9(3). Available from: http://dx.doi.org/10.1093/gigascience/giaa021

29. Rowan TN, Hoff JL, Crum TE, Taylor JF, Schnabel RD, Decker JE. A multi-breed reference panel and additional rare variants maximize imputation accuracy in cattle. Genet Sel Evol. 2019 Dec 26;51(1):77.

30. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.

31. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015 Feb 25;4:7.

32. Loh PR, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016 Nov;48(11):1443–8.

33. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016 Oct;48(10):1284–7.

34. Decker JE, Vasco DA, McKay SD, McClure MC, Rolf MM, Kim J, et al. A novel analytical method, Birth Date Selection Mapping, detects response of the Angus (Bos taurus) genome to selection on complex traits. BMC Genomics. 2012 Nov 9;13:606.

35. Pocrnic I, Lourenco DAL, Masuda Y, Legarra A, Misztal I. The Dimensionality of Genomic Information and Its Effect on Genomic Prediction. Genetics. 2016 May;203(1):573–81.

36. Kemper KE, Goddard ME. Understanding and predicting complex traits: knowledge from cattle. Hum Mol Genet. 2012 Oct 15;21(R1):R45–51.

37. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008 Nov;91(11):4414–23.

38. Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J Dairy Sci. 2010 Feb;93(2):743–52.

39. Aguilar I, Misztal I, Legarra A, Tsuruta S. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. J Anim Breed Genet. 2011 Dec;128(6):422–8.

40. Vinkhuyzen AAE, Wray NR. Novel directions for G × E analysis in psychiatry. Epidemiol Psychiatr Sci. 2015 Feb;24(1):12–9.

41. Legarra A, Reverter A. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. Genet Sel Evol. 2018 Nov 6;50(1):53.

42. Macedo FL, Reverter A, Legarra A. Behavior of the Linear Regression method to estimate bias and accuracies with correct and incorrect genetic evaluation models. J Dairy Sci. 2020 Jan;103(1):529–44.

43. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2017. Available from: https://www.R-project.org/

44. Lopez BI, Santiago KG, Seo K, Jeong T, Park JE, Chai HH, et al. Genetic Parameters of Birth Weight and Weaning Weight and Their Relationship with Gestation Length and Age at First Calving in Hanwoo (Bos taurus coreanae). Animals (Basel) [Internet]. 2020 Jun 23;10(6). Available from: http://dx.doi.org/10.3390/ani10061083

45. Beaumont RN, Warrington NM, Cavadino A, Tyrrell J, Nodzenski M, Horikoshi M, et al. Genome-wide association study of offspring birth weight in 86 577 women identifies five novel loci and highlights maternal genetic effects that are independent of fetal genetics. Hum Mol Genet. 2018 Feb 15;27(4):742–56.

46. Warrington NM, Beaumont RN, Horikoshi M, Day FR, Helgeland Ø, Laurin C, et al. Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. Nat Genet. 2019 May;51(5):804–14.

47. Wang X, Lan X, Radunz AE, Khatib H. Maternal nutrition during pregnancy is associated with differential expression of imprinted genes and DNA methyltranfereases in muscle of beef cattle offspring. J Anim Sci. 2015 Jan;93(1):35–40.

48. Collier RJ, Baumgard LH, Zimbelman RB, Xiao Y. Heat stress: physiology of acclimation and adaptation. Anim Front. 2019 Jan;9(1):12–9.

49. Bourdon RM, Brinks JS. Genetic, environmental and phenotypic relationships among gestation length, birth weight, growth traits and age at first calving in beef cattle. J Anim Sci. 1982 Sep;55(3):543–53.

50. Zhang Z, Ober U, Erbe M, Zhang H, Gao N, He J, et al. Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. PLoS One. 2014 Mar 24;9(3):e93017.

51. Zhang Z, Erbe M, He J, Ober U, Gao N, Zhang H, et al. Accuracy of whole-genome prediction using a genetic architecture-enhanced variance-covariance matrix. G3 . 2015 Feb 9;5(4):615–27.

52. Morgante F, Huang W, Maltecca C, Mackay TFC. Effect of genetic architecture on the prediction accuracy of quantitative traits in samples of unrelated individuals. Heredity . 2018 Feb 10;120(6):500–14.

53. Carvalho FE, Espigolan R, Berton MP, Neto JBS, Silva RP, Grigoletto L, et al. Genome-wide association study and predictive ability for growth traits in Nellore cattle. Livest Sci. 2020 Jan 1;231:103861.

54. Abdollahi-Arpanahi R, Morota G, Peñagaricano F. Predicting bull fertility using genomic data and biological information. J Dairy Sci. 2017 Dec 1;100(12):9656–66.

55. MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, Chamberlain AJ, et al. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics. 2016 Feb 27;17:144.

56. Edwards SM, Sørensen IF, Sarup P, Mackay TFC, Sørensen P. Genomic Prediction for Quantitative Traits Is Improved by Mapping Variants to Gene Ontology Categories in Drosophila melanogaster. Genetics. 2016 Aug;203(4):1871–83.

57. Sarup P, Jensen J, Ostersen T, Henryon M, Sørensen P. Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs. BMC Genet. 2016 Jan 5;17:11.

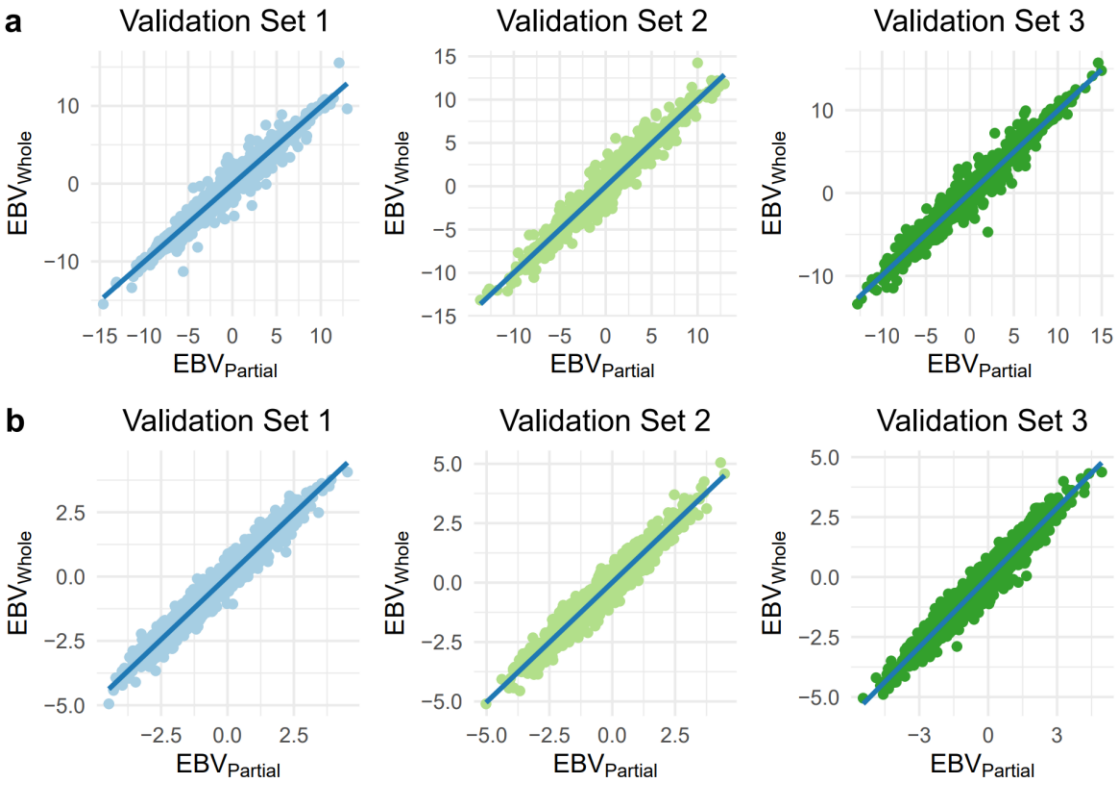
58. Lenormand T. Gene flow and the limits to natural selection. Trends Ecol Evol. 2002 Apr 1;17(4):183–9.

59. Vitezica ZG, Legarra A, Toro MA, Varona L. Orthogonal Estimates of Variances for Additive, Dominance, and Epistatic Effects in Populations. Genetics. 2017 Jul;206(3):1297–307.

60. Huang W, Mackay TFC. The Genetic Architecture of Quantitative Traits Cannot Be Inferred from Variance Component Analysis. PLoS Genet. 2016 Nov;12(11):e1006421.

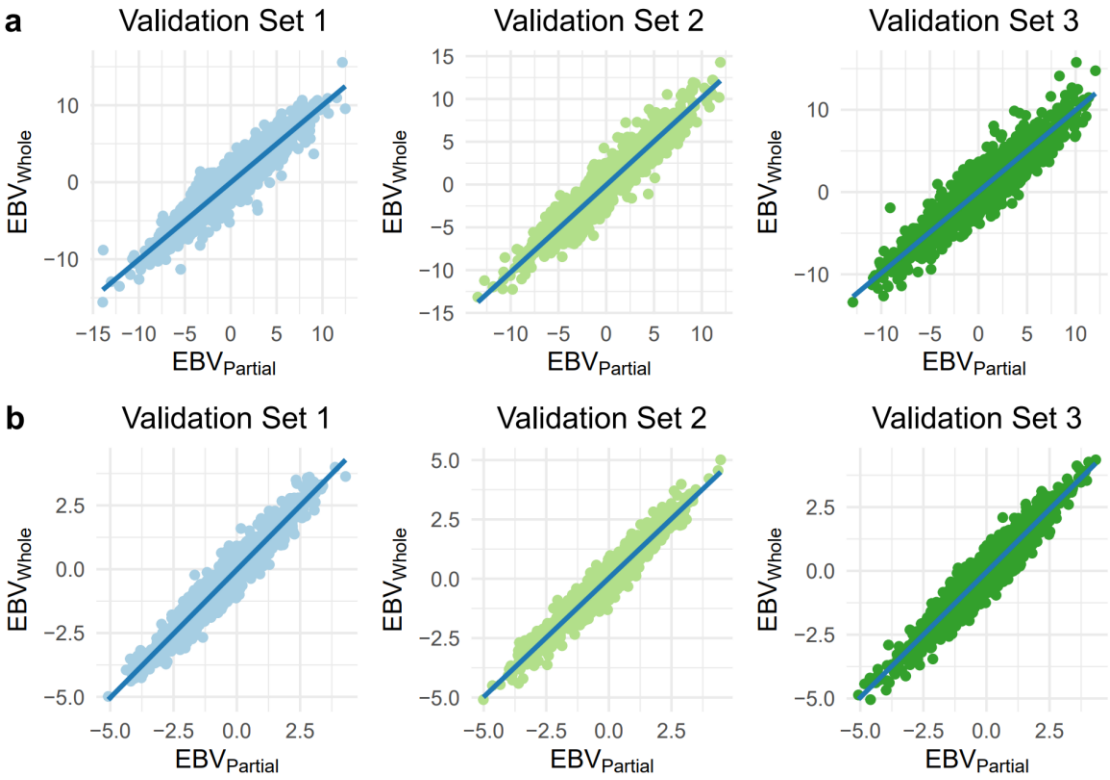
Figures

Figure 1 National whole and partial estimated direct and maternal genetic breeding values



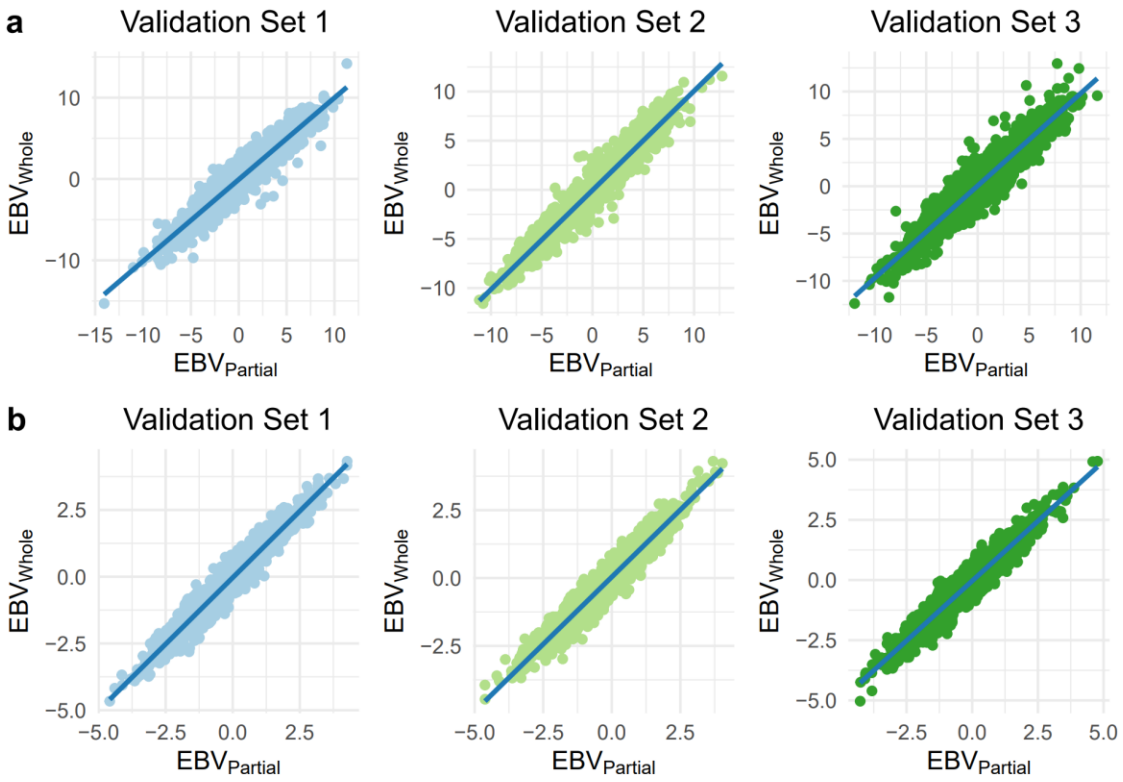
The national models' whole (y-axis) estimated breeding values for the **a** direct genetic effect and **b** the maternal genetic effect are plotted against the national partials' (x-axis) estimated breeding values for the three validation sets.

Figure 2 GxE Full SNP whole and partial estimated direct and maternal genetic breeding values



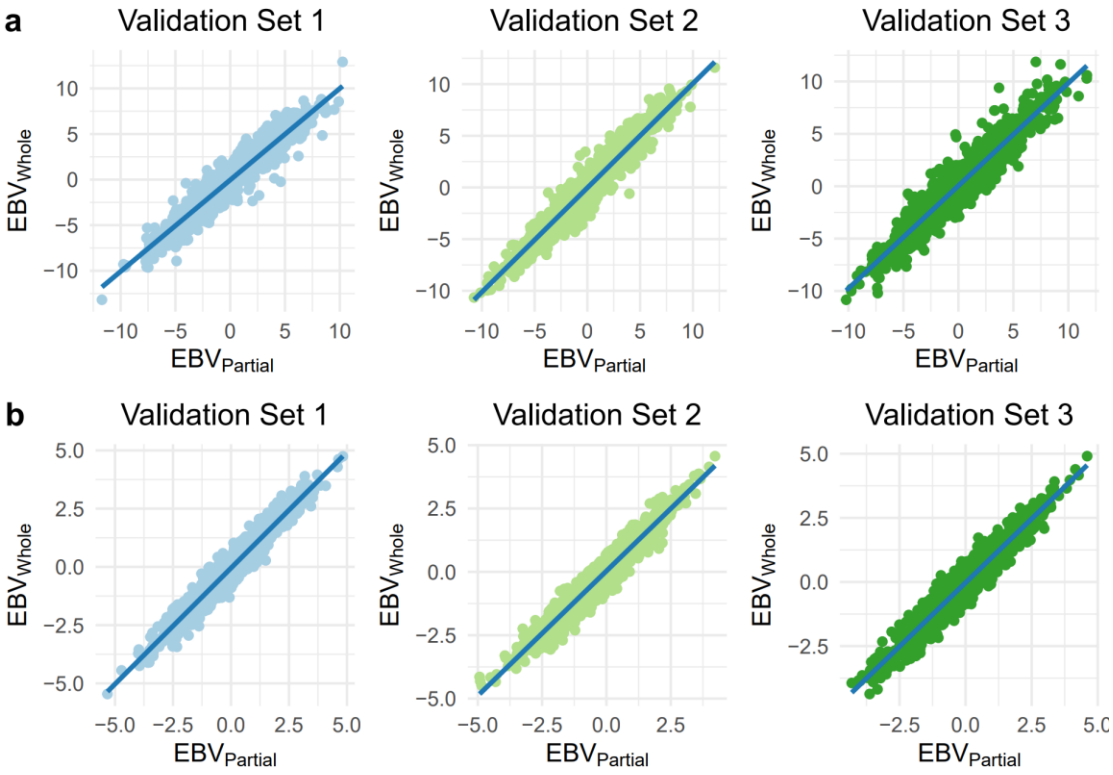
The genotype-by-environment full SNP models' whole (y-axis) estimated breeding values for the **a** direct genetic effect and **b** the maternal genetic effect are plotted against the national partials' (x-axis) estimated breeding values for the three validation sets.

Figure 3 GxE 40k SNP whole and partial estimated direct and maternal genetic breeding values



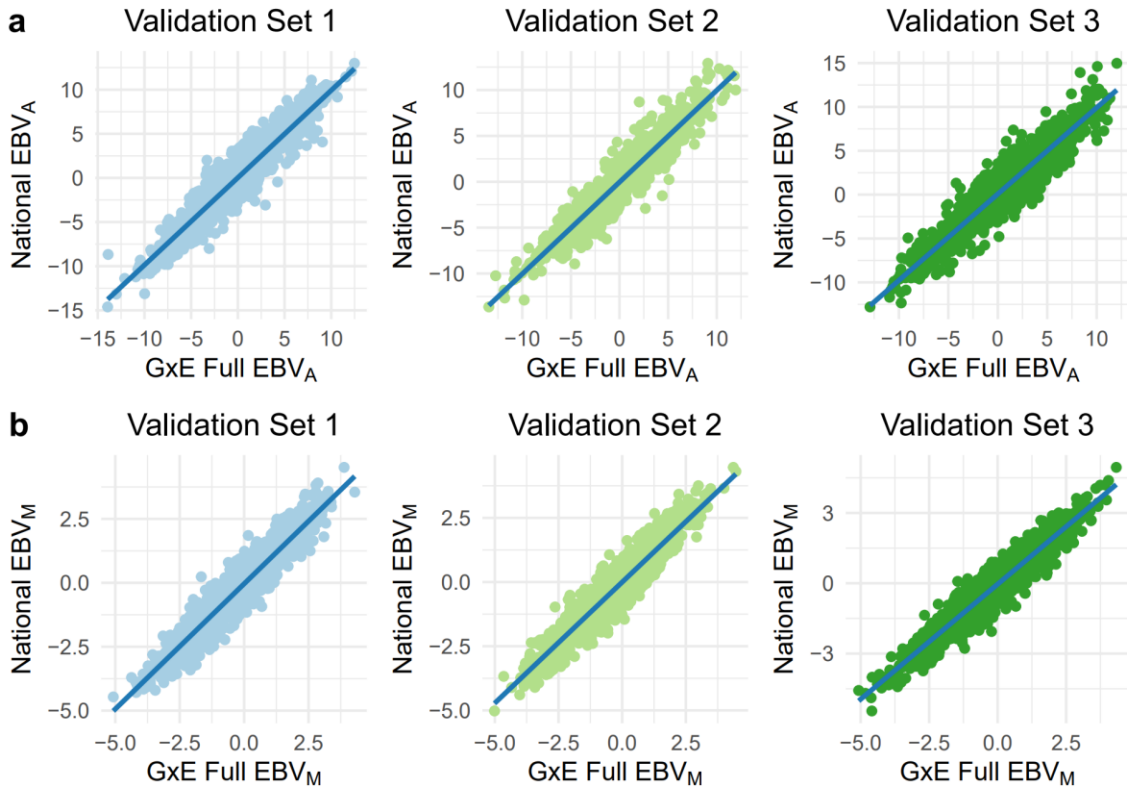
The genotype-by-environment 40k SNP models' whole (y-axis) estimated breeding values for the **a** direct genetic effect and **b** the maternal genetic effect are plotted against the national partials' (x-axis) estimated breeding values for the three validation sets.

Figure 4 GxE 20k SNP whole and partial estimated direct and maternal genetic breeding values



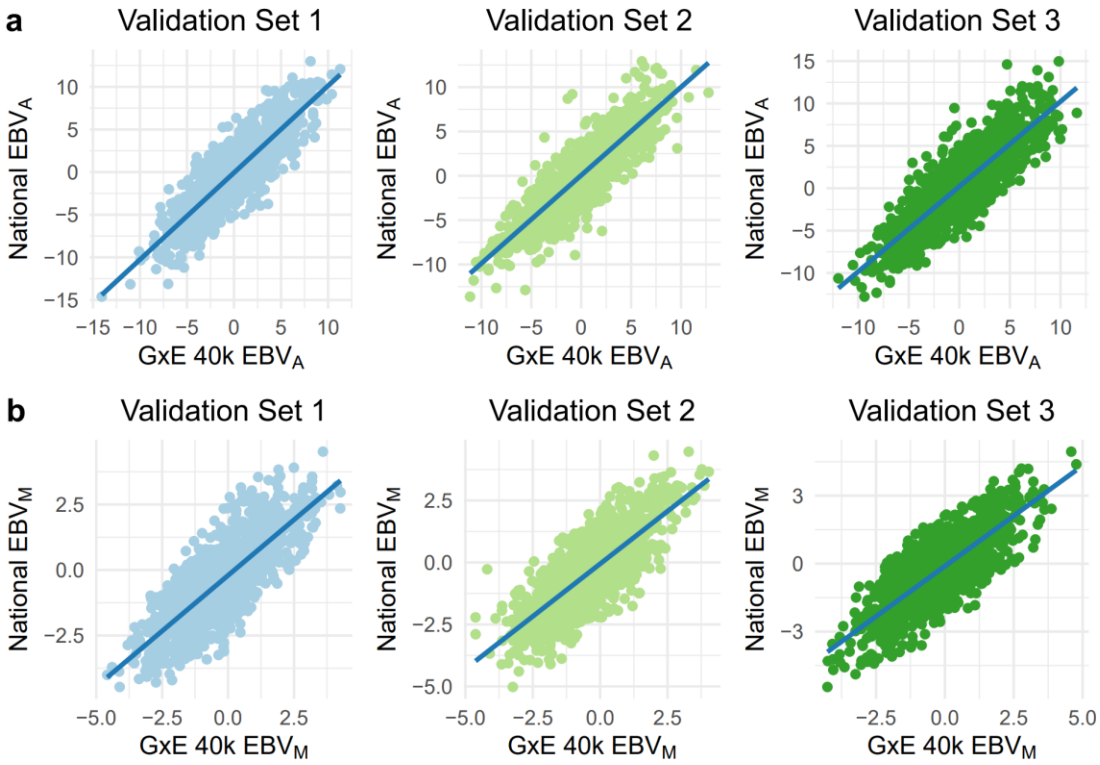
The genotype-by-environment 20k SNP models' whole (y-axis) estimated breeding values for the **a** direct genetic effect and **b** the maternal genetic effect are plotted against the national partials' (x-axis) estimated breeding values for the three validation sets.

Figure 5 National and genotype-by-environment full SNP set models' estimated direct and maternal genetic breeding values



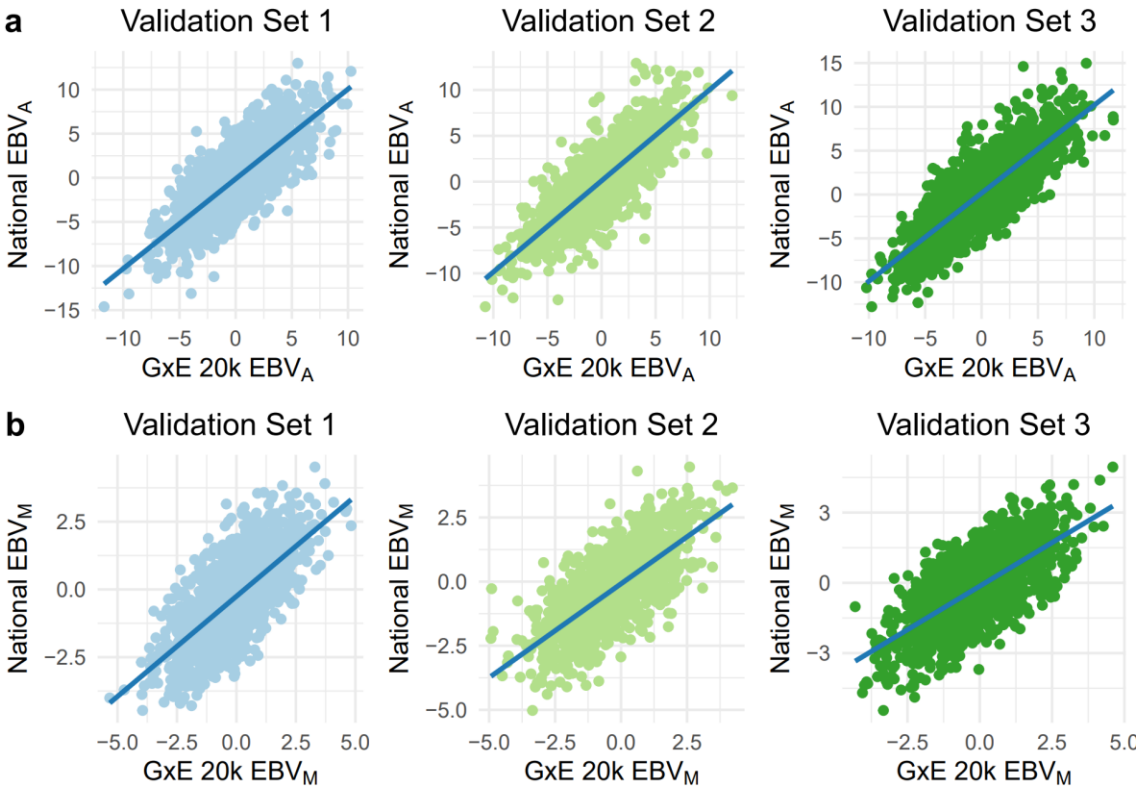
The national models' (y-axis) estimated breeding values for the **a** direct genetic effect and **b** the maternal genetic effect are plotted against the full single nucleotide snp set genotype-by-environment models' (x-axis) corresponding estimated breeding values for the three validation sets.

Figure 6 National and genotype-by-environment 40k SNP set models' estimated direct and maternal genetic breeding values



The national models' (y-axis) estimated breeding values for the **a** direct genetic effect and **b** the maternal genetic effect are plotted against the 40k single nucleotide snp set genotype-by-environment models' (x-axis) corresponding estimated breeding values for the three validation sets.

Figure 7 National and genotype-by-environment 20k SNP set models' estimated direct and maternal genetic breeding values



The national models' (y-axis) estimated breeding values for the **a** direct genetic effect and **b** the maternal genetic effect are plotted against the 20k single nucleotide snp set genotype-by-environment models' (x-axis) corresponding estimated breeding values for the three validation sets.

Tables

Table 1 Estimated variance components for birth weight

Model	SNP	$\widehat{\sigma_A^2}$	$\widehat{\sigma_M^2}$	cov(A,M)	$rG_{A,M}$	$\widehat{\sigma_{DGxE}^2}$	$\widehat{\sigma_{MGxE}^2}$	$\widehat{\sigma_e^2}$
National	**Full**	29.33	10.96	-4.71	-0.26	-	-	39.27
GxE	**Full**	29.35	10.92	-4.73	-0.26	2.59	0.12	39.04
GxE	**40k**	26.58	9.14	-3.60	-0.23	59.28	1.39	43.18
GxE	**20k**	25.04	8.75	-4.05	-0.27	59.02	2.78	46.85

The estimated variance components for birth weight across the national and genotype-by-environment (GxE) models with the full SNP set and targeted SNP sets of 40,000 and 20,000. The estimated components are the direct genetic variance ($\widehat{\sigma_A^2}$ **)**, maternal genetic variance ($\widehat{\sigma_M^2}$ **)**, the covariance between the direct and maternal genetic (cov(A,M)), the genetic correlation between the direct and maternal genetic ($rG_{A,M}$), the direct GxE variance ($\widehat{\sigma_{DGxE}^2}$ **)**, the maternal GxE variance ($\widehat{\sigma_{MGxE}^2}$ **)** and the residual variance ($\widehat{\sigma_e^2}$ **)**. Missing values are represented by a dash due to those effects not being included in the model.

Table 2 Narrow-sense heritabilities and percent variance explained for random effects

Model	SNP	$\widehat{h^2}$ of A	$\widehat{h^2}$ of M	PVE of DGxE	PVE of MGxE
National	**Full**	0.368	0.138	-	-
GxE	**Full**	0.358	0.133	0.032	0.0014
GxE	**40k**	0.190	0.065	0.425	0.0099
GxE	**20k**	0.176	0.061	0.414	0.0195

The estimated narrow-sense heritabilities ($\widehat{h^2}$) for direct (A) and maternal genetic effects (M), and the percent variances explained (PVE) for the direct genotype-by-environment effect (DGxE) and the maternal genotype-by-environment effect (MGxE) for the national model, the genotype-by-environment full SNP model, the genotype-by-environment 40k SNP model, and the genotype-by-environment 20k SNP model. Missing values are represented by a dash due to those effects not being included in the model.

Table 3 Ecoregion fixed effect solutions for genotype-by-environment models

SNP	DT	SE	HP	AP	FH	FM	FB	UMN
Full	0.197 (1.427)	-0.783 (1.011)	1.344 (0.473)	-2.194 (1.104)	0.000 (0.000)	0.828 (0.518)	-1.237 (0.676)	-0.621 (0.849)
40k	0.448 (1.460)	-0.689 (1.037)	1.543 (0.485)	-2.580 (1.131)	0.000 (0.000)	0.783 (0.531)	-1.040 (0.691)	-0.813 (0.870)
20k	0.358 (1.488)	-0.955 (1.056)	1.703 (0.493)	-2.534 (1.151)	0.000 (0.000)	0.932 (0.541)	-0.997 (0.702)	-0.764 (0.885)

Estimated fixed effect solutions in the genotype-by-environment models for ecoregions: Desert (DT, n = 725), Southeast (SE, n = 959), High Plains (HP, n = 6437), Arid Prairie (AP, n = 1011), Foothills (FH, n = 19), Forested Mountains (FM, n = 3229), Fescue Belt (FB, n = 2889), and the Upper Midwest and Northeast (UMN, n = 1927). Standard error of the solution is below the estimate within parentheses.

Table 4 Average accuracies of the additive genetic estimated breeding values

Estimated Breeding Value	National, Full	GxE, Full	GxE, 40k	GxE, 20k
Direct	0.80 (0.012)	0.77 (0.009)	0.72 (0.015)	0.70 (0.019)
Maternal	0.46 (0.006)	0.46 (0.009)	0.47 (0.007)	0.49 (0.006)

Average accuracies of the additive direct and maternal genetic estimated breeding values for the national model, the genotype-by-environment full SNP model, the genotype-by-environment 40k SNP model, and the genotype-by-environment 20k SNP model across the three validation sets. Standard errors are reported within parentheses.

Table 5 Average comparison statistics among National whole and partial additive genetic estimated breeding values

Estimated Breeding Value	Pearson	Spearman	Slope
Direct	0.980 (0.001)	0.978 (0.001)	0.996 (0.001)
Maternal	0.973 (0.002)	0.969 (0.003)	0.985 (0.012)

The average Pearson correlations, Spearman correlations, and slopes of the line of best fit among the national models' whole and partial estimated breeding values for the direct and maternal genetic effects across the three validation sets. Standard errors are reported within parentheses.

Table 6 Average comparison statistics among GxE Full SNP whole and partial additive genetic estimated breeding values

Estimated Breeding Value	Pearson	Spearman	Slope
Direct	0.944 (0.005)	0.939 (0.006)	1.003 (0.009)
Maternal	0.966 (0.004)	0.960 (0.006)	0.996 (0.006)

The average Pearson correlations, Spearman correlations, and slopes of the line of best fit among the genotype-by-environment, full SNP set, models' whole and partial estimated breeding values for the direct and maternal genetic effects across the three validation sets. Standard errors are reported within parentheses.

Table 7 Average comparison statistics among GxE 40k SNP whole and partial additive genetic estimated breeding values

Estimated Breeding Value	Pearson	Spearman	Slope
Direct	0.950 (0.006)	0.945 (0.007)	0.997 (0.011)
Maternal	0.967 (0.003)	0.964 (0.003)	0.993 (0.002)

The average Pearson correlations, Spearman correlations, and slopes of the line of best fit among the genotype-by-environment, 40k SNP set, models' whole and partial estimated breeding values for the direct and maternal genetic effects across the three validation sets. Standard errors are reported within parentheses.

Table 8 Average comparison statistics among GxE 20k SNP whole and partial additive genetic estimated breeding values

Estimated Breeding Value	Pearson	Spearman	Slope
Direct	0.957 (0.006)	0.952 (0.005)	0.997 (0.009)
Maternal	0.970 (0.001)	0.967 (0.002)	0.992 (0.004)

The average Pearson correlations, Spearman correlations, and slopes of the line of best fit among the genotype-by-environment, 20k SNP set, models' whole and partial estimated breeding values for the direct and maternal genetic effects across the three validation sets. Standard errors are reported within parentheses.

Table 9 Average comparison statistics among direct genetic estimated breeding values when compared to the national model

Model	SNP	Pearson	Spearman	Slopes
GxE	**Full**	0.948 (0.002)	0.945 (0.004)	0.991 (0.005)
GxE	**40k**	0.865 (0.005)	0.856 (0.008)	1.004 (0.008)
GxE	**20k**	0.808 (0.007)	0.798 (0.008)	1.004 (0.007)

The average Pearson correlations, Spearman correlations, and slopes for the lines of best fit among partial direct genetic estimated breeding values when the genotype-by-environment models are compared to the national model across the three validation sets. Standard errors are reported within parentheses.

Table 10 Average comparison statistics among maternal genetic estimated breeding values when compared to the national model

Model	SNP	Pearson	Spearman	Slopes
GxE	**Full**	0.947 (0.002)	0.940 (0.001)	0.968 (0.013)
GxE	**40k**	0.790 (0.002)	0.774 (0.005)	0.862 (0.013)
GxE	**20k**	0.701 (0.006)	0.685 (0.012)	0.738 (0.003)

The average Pearson correlations, Spearman correlations, and slopes for the lines of best fit among partial maternal genetic estimated breeding values when the genotype-by-environment models are compared to the national model across the three validation sets. Standard errors are reported within parentheses.

VITA

Originally from Texas, Sara grew up on a ranch with Maine Anjou and Shorthorn beef cattle. She showed these cattle as a part of her school's Future Farmers of America organization starting at the age of 9. As a member of the FFA, she participated in judging and quiz bowl contests throughout high school. Exposure to ranch life and with a love for animals, she decided to follow her grandfather and pursue a career in veterinary medicine. After graduating in the top ten percent of her class in May 2010, she attended Tarleton State University as a pre-veterinary student. Two years later, she transferred to Oklahoma State University to finish her Bachelor's degree in May 2014. Wanting to pursue research with a newfound interest in genetics, she got accepted in a Mater's program at the University of Nebraska-Lincoln to study host genomics and virus interactions. After completing her thesis, she wanted to improve her skill set to encompass data generation and analysis, so she applied to a doctoral program at the University of Missouri. While in Missouri, she researched population genomics with the domestic cat which helped further develop her passion for conservation. Additionally, she researched local adaptation in beef cattle applied to genomic predictions with the hopes of matching cattle to their environments. All throughout her education she gained an interest in science communication and outreach, especially through photography. During her spare time, she tries to go hiking with her dog and take photos of nature. Currently she hopes to move more into conservation genomics to help preserve endangered species for future generations while incorporating her love of photography.